arXiv:1506.02699v1 [stat.ML] 8 Jun 2015

Community detection in multi-relational data with restricted multi-layer stochastic blockmodel*

Subhadeep Paul and Yuguo Chen

Department of Statistics
University of Illinois at Urbana-Champaign
Champaign, IL 61820
USA

Abstract: In recent years there has been an increased interest in statistical analysis of data with multiple types of relations among a set of entities, mainly driven by applications in biology, social sciences, e-commerce and marketing. Such multi-relational data can be represented as multi-layer graphs where multiple types of edges represent the relations and the set of vertices/nodes represents the entities. An important learning goal in such networks is to detect an underlying set of communities leveraging information from all the layers. For community detection in multi-layer graphs, we consider a random graph model, multi-layer stochastic blockmodel (MLSBM), which is an extension of the stochastic block model. In this connection we also propose a model with a restricted parameter space, restricted multi-layer stochastic blockmodel (RMLSBM), for applications where either the network layers are sparse or the number of communities is large or both. We derive consistency results for community assignments through both methods where MLSBM is assumed to be the true model, and either the number of nodes or the number of types of edges or both grow. We compare the two methods both in terms of performance in simulation and asymptotic performance under different asymptotic setups. We establish the advantage of RMLSBM over MLSBM when either the growth rate of the number of communities is high or the growth rate of the average degree of the component graphs in the multi-graph is low. To solve the computationally challenging problem of community assignment through maximum likelihood estimation, we derive a variational EM algorithm. The simulation studies and real data applications confirm the superior performance of the multi-layer approaches in comparison to independent modeling of the layers or majority voting.

Keywords and phrases: Consistency, community detection, multi-layer networks, maximum likelihood inference, restricted maximum likelihood inference, stochastic blockmodel.

1. Introduction

Over the last decade, relational data has become ubiquitous in all forms of human activities. In many applications of statistics and machine learning, one encounters relational data where the entities are represented as nodes or vertices and the relations or interactions between the entities as edges of a graph. Applications of such graphs or networks include many information systems such as social networks, World Wide Web, user information databases in e-commerce, metabolic networks, gene regulatory networks, protein-protein interaction networks and food web.

In majority of the cases dealt with in the literature, the relations are assumed to be of the same type such as web page linkage, friendship, co-authorship and protein-protein interaction. However in modern complex relational databases and networks, we often have information regarding relationships of multiple types among the nodes. For example, in the context of internet services a set of users may be connected through email, messaging, social media, etc., each one of them creating one layer or type of the user-user interaction network ([26]). Similarly, users in a social network can have "friendship", "mentions", "following", etc. ([12]) or researchers in academia may have co-

*Supported in part by NSF Grant DMS-1406455.

authorship, citations, title/abstract similarity, etc., as different types of relations among themselves. In genomics data, cellular components can have different aspects of interactions among them, e.g., protein-protein physical interactions and gene co-expressions ([22]). Such multi-relational data can be represented as multi-layer graphs where multiple types of edges represent the relations and the set of vertices/nodes represents the entities ([17]).

One of the most important and widely investigated learning goals in an information network is clustering the entities on the basis of the relationships between them into densely connected subsets called "communities". In contrast to the idealistic random graph model proposed in [10], where edges between nodes are formed completely at random and with equal probability, real world networks exhibit many interesting properties like community structure, degree heterogeneity, etc. From a probabilistic point of view, communities can be thought of as groups of vertices which are more likely to be connected to each other compared to the rest of the graph, i.e., the probability of having an edge between two vertices belonging to the same group is higher than that of having an edge between vertices belonging to different communities. Consequently we would observe the number of intra community edges to be higher than inter community edges.

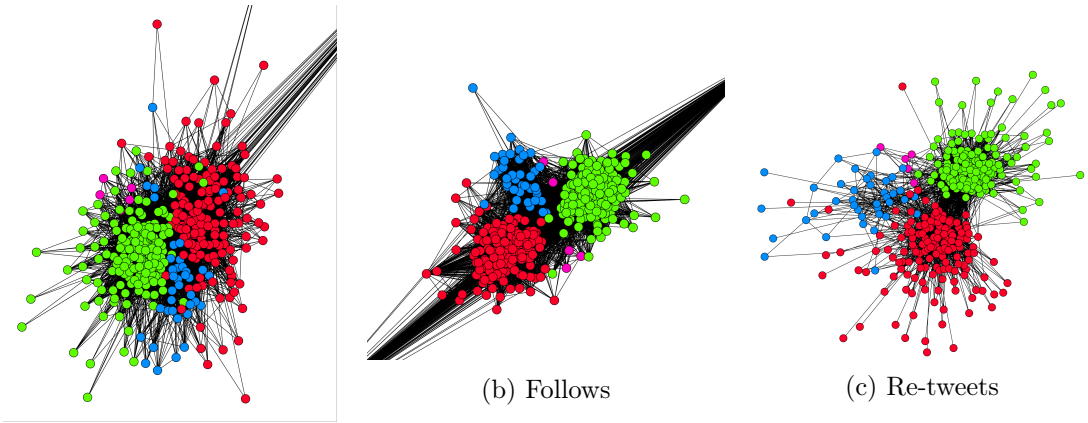
Many researchers have proposed methods and algorithms for community detection in networks. Such methods can broadly be divided into three categories: methods based on probabilistic models, methods based on the maximization of a global objective function and those based on spectral or matrix factorization of the adjacency matrix or the Laplacian matrix. The stochastic blockmodel ([16], [25]) is a statistical model for random graphs with a natural community structure. It is one of a large class of statistical models described in the literature for community detection in complex networks, which includes the latent variable ([14]) and latent space models ([15]), the degree corrected blockmodel ([18], [33]) and the mixed membership blockmodel ([1]). Various likelihood maximization based inference strategies have been proposed in the literature to simultaneously infer the block assignments and the parameters in the stochastic blockmodel, e.g., profile likelihood maximization ([2]), maximizing the conditional likelihood ([6]), and variational EM under mixture model settings ([8]). Other strategies involve Bayesian inference using Gibbs sampling or variational methods ([20]) and optimizing a modularity function over all possible partitions of the graph ([23]). See [11] for a detailed review of statistical inference in networks.

Several authors have also studied the conditions required on the growth of the number of communities and the degree density of networks for the estimation strategies to be consistent. Bickel and Chen [2] and Zhao et al. [33] studied the conditions for community detection through modularity maximization under the stochastic blockmodel and the degree corrected stochastic blockmodel respectively. Choi et al. [6] laid down the conditions necessary for the consistency of maximum likelihood estimation under the stochastic blockmodel. This work was extended by Rohe et al. [28] with a regularized estimator to high dimensional settings where the number of communities grows roughly as fast as the number of nodes. Celisse et al. [5] derived consistency and Bickel et al. [3] derived asymptotic normality of the maximum likelihood estimators and their variational approximations in the mixture model settings.

In this paper our primary focus is on the problem of detecting an underlying community structure in multi-layer networks. We assume that such networks have an implicit community structure and different observed layers manifest that underlying structure with varying amount of information and noise. As an example of a network, where such an assumption is reasonable, we analyze a twitter network of British MPs where the underlying communities are based on their party memberships and the three observed layers, "mentions", "follows" and "'re-tweets" manifest that structure in varying proportions. In such cases the multi-layer graph is a more accurate representation of the underlying similarity of the objects and each layer can provide only a "partial" information about the data ([27]). The goal in such cases would be to correctly identify the underlying set of communities combining

information from all three layers.

Earlier approaches towards multi-relational data or multi-layer graph clustering suffer from the deficiency that they either cluster each graph independently and combine the results, or aggregate the graphs and cluster the aggregated graph. These approaches fail to take into account the dependency among the different layers, in particular the correlation among different types of edges that share the same pair of nodes. Moreover, the multiple network layers can have different characteristics in terms of sparsity and noise. Some layers may be dense but may carry little worthwhile information, whereas some layers may be extremely sparse but may carry valuable information. The aggregation process of graphs could lose the intrinsic heterogeneity of the network layers. Here we attempt to address the problem of how to efficiently cluster the nodes or entities in a network taking into account all types of layers or relations among them. Several approaches have been recently proposed in the literature for this purpose. Among them are approaches based on collective or joint matrix factorization ([24], [31], [27]), non-parametric Bayesian models and latent factor models ([17]), extensions of spectral clustering ([9]) and modularity ([21]) to multi-layer graphs. However there is a lack of statistical analysis of the properties of those methods.



(a) Mention

(b) Follows

(c) Re-tweets

Fig 1: A 3-layer twitter network of British MPs. The nodes are colored according to an underlying community structure: the party memberships.

For community detection in multi-layer networks, we consider a natural extension of the standard stochastic blockmodel to multi-layer settings that we will call "multi-layer stochastic blockmodel" (MLSBM). This model, also considered in [13] as "multi-graph SBM", is in the spirit of multi-relational models described in [16], [32] and [19]. The authors in [13] prove the consistency of the maximum likelihood estimates in this model when the number of relations grows. They keep the number of nodes (and hence the number of communities) fixed. However, as we will see later in both the asymptotic analysis and simulation studies that this model does not perform very well when either the number of communities grows fast or the network layers are sparse on average. Hence, we propose a restricted version of this model through restrictions on the parameter space which is capable of handling networks with a large number of communities. We call this model "restricted multi-layer stochastic blockmodel" (RMLSBM). We derive conditions on the growth of the number of communities and the average edge density of the networks under which the maximum likelihood estimate of the class assignment vector is consistent (in the sense that the proportion of misclassified nodes tends to 0 as the number of nodes, and possibly the number of relations as well, grows). To

compute the unknown class assignments and block model parameters simultaneously, we follow [8] and propose a variational estimation strategy.

The rest of the paper is organized as follows. Section 2 extends the stochastic blockmodel to multi-layer settings and defines the two models, MLSBM and RMLSBM. Section 3 settles the consistency of the community assignments through maximum likelihood estimation in the two models when the true data generating model is MLSBM. Section 4 describes two estimation strategies for the two models. Section 5 describes the results of a simulation study to validate the theoretical results. Section 6 presents the application of the methods to the Twitter UK politics data set. Section 7 gives concluding remarks.

2. Extension of blockmodels to multi-layer settings

We consider an undirected multi-layer graph $G = \{V, E\}$, where the vertex set V consists of N vertices and the edge set E consists of edges of M different types representing different relations. We can view the multi-graph as a graph with vector valued edge information, i.e., the adjacency matrix A consists of elements A_{ij}, who are themselves M dimensional vectors: $A_{ij} = \{A_{ij}^{(1)}, A_{ij}^{(2)}, \ldots, A_{ij}^{(M)}\}$. An alternative way to approach the problem is to view the multi-graph as a collection of M, $N \times N$ adjacency matrices $\{A^{(1)}, A^{(2)}, \ldots, A^{(M)}\}$, each corresponding to one particular type of relation. The rest of the set up is similar to the regular stochastic block model for one-layer case with K blocks ([25]). We assume the number of communities K is known. Let $z = \{z_1, z_2, \ldots, z_N\}$ be the community indicator vector for the N nodes, such that each z_i takes exactly one value from the set $\{1, \ldots, K\}$ and $z_i = q$ if and only if node i belongs to community q. Conditional on the community indicator vector z, the edges are formed independently as Bernoulli random variables with probabilities depending only on the community assignments and the type of edges.

We expect the random variables representing the different types of edges between a pair of nodes to be correlated, and our aim here is to take that correlation into account while clustering the nodes. However we do not explicitly model this correlation among the edges of different types between nodes, i.e., conditional on the community assignments we still assume them to be independent random variables. Instead we induce dependence among the edges of different types through the latent variables so that unconditionally the edges are dependent. In particular, we model the probability of an edge as dependent on both the community to which its nodes belong as well as what type of edge it is. For this purpose, in what follows we describe the two extensions of the standard stochastic block model to multi-layer settings.

Each node i of the network is independently assigned a cluster/community denoted by its latent variable z_i, with $z_i = q$ for exactly one $q \in \{1, \ldots, K\}$. Except for the estimation algorithm, the model is always represented as a conditional block model and z is assumed to be a fixed unknown parameter of the model and need to be estimated from data. Conditioned on the community assignments of the nodes z_i and z_j, the edges are formed independently following Bernoulli distribution

$$A_{ij}^{(m)} | (z_i = q, z_j = l) \sim Bernoulli(P_{ql}^{(m)}).$$

The first model assigns a separate probability for the mth type of edge between nodes belonging to the qth and the lth community independent of all other edges. We call this model the "multi-layer stochastic blockmodel" (MLSBM). The probability of an mth type of edge between nodes i and j belonging to communities q and l respectively can be written as

$$P_{ij}^{(m)} = \pi_{z_i z_j}^{(m)} = \pi_{ql}^{(m)}, \quad i, j \in \{1, \ldots, N\},\ m \in \{1, \ldots, M\},\ q, l \in \{1, \ldots, K\}.$$

The set of parameters for the model, $\pi = \{\pi_{ql}^{(m)};\ q \le l,\ q,l \in \{1,\dots,K\},\ m \in \{1,\dots,M\}\}$ has $K(K+1)M/2$ elements. This model is "saturated" in the sense that we have a different parameter for each of the different types of edges between nodes belonging to different communities. Denote the range of this parameter set or array as $\Pi = \{\pi \in [0,1]^{K(K+1)M/2}\}$. When K grows with N and/or M, the number of parameters to be estimated in this model grows linearly both with N and M and quickly becomes large. Hence the MLE performs poorly especially when the individual network layers are spare. This problem does not arise in the asymptotic settings where only M grows and K remains fixed. However, it has been empirically shown that in most real world networks the average cluster size does not grow with the size of the network and consequently, K grows with N. Hence this settings is rather unrealistic. This motivates us to propose the second related model whose number of parameters grows much slowly compared to MLSBM.

The second model assumes the probability of the mth type of edge appearing between nodes i and j is governed by two factors: the first one being the community assignment of the two nodes and the second one being the type of edge. Hence the model has two sets of parameters: a $K \times K$ parameter matrix $\pi_{K\times K}$ corresponding to the community structure, and an $M \times 1$ vector $\beta_{M\times 1}$ which contains the parameters for different types of edges. We call this model the restricted multi-layer stochastic blockmodel (RMLSBM).

Notice that in the second model, if the edges were all of the same type, we would just have $\beta_m = \beta$ for all $m \in \{1,\dots,M\}$ and then we will recover the standard stochastic blockmodel, with probabilities of edges determined solely by the community assignments. On the other hand, if we did not have a community structure, but M types of edges, then π_{ql} would be identical for all communities q, l and the probability of an edge between nodes i and j will solely be determined by the type of edge. This model can retrieve information from sparse but highly informative edge types as the sparsity of the network layers will be captured in the β_m parameters. Hence, although we assume the edges to be conditionally independent, this model induces two types of correlations unconditionally — among the edges of the same type and among the edges that share nodes of the same community.

The probability $P_{ij}^{(m)}$ in RMLSBM , which denotes the probability of an mth type of edge between nodes i and j belonging to communities q and l respectively, can be modeled in the following way with the logit link function

$$\text{logit}(P_{ij}^{(m)}) = \pi_{ql} + \beta_m,\ \ i,j \in \{1,\dots,N\},\ m \in \{1,\dots,M\},\ q,l \in \{1,\dots,K\}.$$

This model has $K(K+1)/2 + M$ parameters for an undirected graph. Hence, when both K^2 and M grows, the growth rate in the number of parameters for this model is same as the maximum of the growth rates. In comparison, the number of parameters in MLSBM would grow linearly with both K^2 and M. This makes the maximum likelihood estimator in RMLSBM a regularized estimator. However, for the model to be identifiable, we require the parameters β_m to satisfy the condition $\sum_m \beta_m = 0$. Hence we have one less free parameter. Denote the set of parameters for RMLSBM as $\pi^R = \{(\pi_{ql}, \beta_m) :\ q \le l,\ q,l \in \{1,\dots,K\},\ m \in \{1,\dots,M\}\}$ and its range as $\Pi^R = \{\pi^R \in \mathcal{R}^{K(K+1)/2+M},\ \sum_m \beta_m = 0\}$. To prove the consistency of maximum likelihood estimation under MLSBM, we assume $\pi_{ql}, \beta_m \in (-C\log(MN^2),\ C\log(MN^2))$ for some constant $C > 0$. This condition ensures that π_{ql} and β_m are bounded away from $\pm\infty$.

3. Consistency

In this section, we discuss the consistency of maximum likelihood estimation of the proposed models under three asymptotic regimes with varying conditions imposed on the growth of the number of

communities (K) and the expected total number of edges of the multi-layer graph (L). We first define a one to one transformation of the parameters of RMLSBM as

$$\phi_{ql}^{(m)} = \text{logit}^{-1}(\pi_{ql} + \beta_m) = \frac{\exp(\pi_{ql} + \beta_m)}{1 + \exp(\pi_{ql} + \beta_m)}. \tag{3.1}$$

Now we assume that the data are generated from the more general model MLSBM and view RMLSBM as an MLSBM with the following restrictions on the parameters:

$$\begin{aligned}\Phi = \{\phi \in [0,1]^{K(K+1)M/2} :\ & \phi_{ql}^{(m)} = \text{logit}^{-1}(\pi_{ql} + \beta_m), \\ & \pi_{ql}, \beta_m \in (-C\log(MN^2),\ C\log(MN^2))\}.\end{aligned} \tag{3.2}$$

This way the maximum likelihood estimate (MLE) in RMLSBM can be thought of as a restricted maximum likelihood estimate (RMLE) of MLSBM.

Our aim is to investigate the consistency of both the MLE and the RMLE under three asymptotic regimes where we let either the number of nodes (N) or the number of types of edges (M) or both to grow. This setup is quite appropriate for modern day multi-layer networks, where data collection increases both in terms of new entities as well as new features or layers getting added to the database. Consequently methods are being sought which would be consistent in such situations. Although some results for the MLE were obtained in [13] under the settings when M grows, but N and consequently K remains fixed, we still include it in our asymptotic analysis since we need results for the growing N and K case for asymptotic comparison with regularized estimator. The different asymptotic setups we consider under the three regimes of growth in N and M are described below.

1. As both M and N grow, let $K = O(N^{1/2})$ and $L = \omega(MN(\log N)^{3+\delta})$ for some $\delta > 0$ for the MLE, while $K = O((MN)^{1/2-\epsilon})$ and $L = \omega(MN(\log N)^{3+\delta})$ with $\epsilon, \delta > 0$ for the RMLE. For the RMLE, we further require that $M = O(N)$ so that K does not exceed N.
2. As N grows, M either is fixed or grows slower than N, i.e., either M is $o(1)$, or $M \to \infty$ and $M = o(N)$. In this regime, let $K = O(N^{1/2})$, $L = \omega(N(\log N)^{3+\delta})$ for some $\delta > 0$ for the RMLE.
3. As both $N \to \infty$ and $M \to \infty$ with M growing faster than N, i.e., $M = \omega(N)$, for RMLE we consider two related setups: (a) $K = O(\frac{N}{\log M \log N})$, $L = \omega(MN(\log N)^{1+\delta})$ for some $\delta > 0$; and (b) $K = O(N^{1/2})$, L is either $\omega(M(\log M)^{2+\delta}(\log N)^{1+\delta})$ for some $\delta > 0$ if $(\log M)^{2+\delta} = O(N)$, or $\omega(MN(\log N)^{1+\delta})$ for some $\delta > 0$ otherwise. In setting (a), we further require $\log M$ to grow slower than N for the growth of K to be meaningful. Also, in that setup if $\log M$ grows at the same rate as $(\log N)^{\beta}$ for some $\beta > 0$, the number of communities grows almost as fast as the number of nodes except for the log terms and is "highest dimensional" in the sense of [28].

Note that the first regime assumes no relation between the growth rates of N and M, while the next two regimes assume certain relations between the two growth rates. So the last two regimes can be thought of as special cases of the first one in terms of the growth rates of N and M. Naturally we expect some relaxation in the required growth conditions on K and L in the last two regimes. The asymptotic setups described above reflect this relaxation for the RMLE. However no such relaxation is possible for the MLE. Hence we will prove that MLE in MLSBM is consistent under the first asymptotic regime, whereas MLE in RMLSBM (i.e., the RMLE of MLSBM under the restrictions defined by Equation (3.2)) is consistent under all three asymptotic regimes. The MLSBM, despite being intuitively the simplest extension, does not perform as well as the RMLSBM for community detection in multi-relational networks if the networks are sparse at an average or contain a large number of communities.

3.1. Preliminaries

Since in this paper our primary interest is in modeling multi-layer networks where layers are sparse on an average, we require the true MLSBM model probabilities $\pi_{ql}^{(m)}$ to satisfy certain sparsity conditions. As [33] pointed out, if the block model probabilities remain fixed as N increases, then the network will be unrealistically dense. In this connection it is worth noting that [30] let the probabilities remain fixed and as a result the networks considered there have linearly increasing average degree, while both [2] and [6] considered networks with poly-logarithmically increasing average degree and hence gradually decaying probabilities. Here to keep the network sparse, we scale down the block model probabilities accordingly as N increases.

We introduce a new notation L' to denote the quantity inside the asymptotic notation ω in the growth rate of L under different asymptotic setups. As an example, consider the case when $L = \omega(MN(\log N)^{3+\delta})$, then $L' = MN(\log N)^{3+\delta}$. Hence L' can be viewed as the minimum rate at which L is required to grow under a particular asymptotic setup. The blockmodel parameters are restricted to have an upper bound that decreases with increasing N except for a small finite set indexed by the triplet $Q = \{q, l, m\}$ such that the expected number of edges in the set $|E_Q| = o\left(\frac{L'}{\log(MN^2)}\right)$. For the set Q we can have $\frac{1}{MN^2} \leq \pi_{ql}^{(m)} \leq 1 - \frac{1}{MN^2}$. For all $\{q, l, m\} \notin Q$, the parameters are restricted in the following way

$$\pi_{ql}^{(m)} \in \left(\frac{1}{MN^2},\ C\frac{L'}{MN^2(\log M \log N)^{2+\delta}}\right), \tag{3.3}$$

for some $\delta > 0$ and some constant C, so that the upper bound is determined by the expected density of the network. The exact upper bound is determined by L' and consequently, by the growth rate of L and varies under the different asymptotic assumptions.

For any arbitrary partition z of the entities in the graph, the log likelihood of the set of M adjacency matrices $A = \{A^{(1)}, \ldots, A^{(M)}\}$ under the MLSBM with parameters $\pi = \{\pi_{ql}^{(m)}\}$ is

$$l(A; z, \pi) = \sum_{m=1}^{M} \sum_{i<j} \{A_{ij}^{(m)} \log\ \pi_{z_i z_j}^{(m)} + (1 - A_{ij}^{(m)}) \log\ (1 - \pi_{z_i z_j}^{(m)})\}. \tag{3.4}$$

Note that for an undirected graph with no self-loops, both $A^{(m)}$ and $\pi^{(m)}$, $m = 1, \ldots, M$, are symmetric matrices in $\{0, 1\}^{N \times N}$ and $[0, 1]^{K \times K}$ respectively. The Bernoulli parameters $\pi_{z_i z_j}^{(m)}$ depend both on the class assignment z and the type of relation m. For a fixed class assignment z, let N_q denote the number of nodes assigned to class q, and n_{ql} denote the maximum number of possible edges between classes q and l. So we have $n_{ql} = N_q N_l$ and $n_{qq} = \binom{N_q}{2}$. For an arbitrary partition z, the MLE of $\pi_{(z)}$ is

$$\hat{\pi}_{(z)ql}^{(m)} = \frac{1}{n_{ql}} \sum_{i<j} A_{ij}^{(m)} 1\{z_i = q, z_j = l\},\ m = 1, \ldots, M,\ q, l = 1, \ldots, K, \tag{3.5}$$

where $1\{\cdot\}$ is the indicator function. Note that for a fixed partition z, the denominator n_{ql} in the MLE $\hat{\pi}_{(z)ql}^{(m)}$ is the same for all edge types m.

Now we define the expectation of $\hat{\pi}_{(z)}$ as $\bar{\pi}_{(z)}$ and that of $l(A; z, \pi)$ as $\bar{l}_P(z, \pi)$ under the independent Bernoulli($P_{ij}^{(m)}$) model. Then we have

$$\bar{\pi}_{(z)ql}^{(m)} = \frac{1}{n_{ql}} \sum_{i<j} P_{ij}^{(m)} 1\{z_i = q, z_j = l\},\ m = 1, \ldots, M,\ q, l = 1, \ldots, K, \tag{3.6}$$

$$\bar{l}_P(z,\pi) = \sum_{m=1}^{M}\sum_{i<j}\{P_{ij}^{(m)}\log\ \pi_{z_iz_j}^{(m)} + (1-P_{ij}^{(m)})\log\ (1-\pi_{z_iz_j}^{(m)})\}. \tag{3.7}$$

Clearly for a given z, $\hat{\pi}_{(z)}$ and $\bar{\pi}_{(z)}$ are the maximizers of the functions $l(A;z,\pi)$ and $\bar{l}_P(z,\pi)$ respectively, and we let $l(A;z)$ and $\bar{l}_P(z)$ denote the corresponding maximum values.

We extend Lemma 1 of [6] to multi-layer settings as follows:

$$\begin{aligned} l(A;z) - \bar{l}_P(z) =& \sum_m\sum_{i<j}\left\{A_{ij}^{(m)}\log\left(\frac{\hat{\pi}_{z_iz_j}^{(m)}}{\bar{\pi}_{z_iz_j}^{(m)}}\right) + (1-A_{ij}^{(m)})\log\left(\frac{1-\hat{\pi}_{z_iz_j}^{(m)}}{1-\bar{\pi}_{z_iz_j}^{(m)}}\right)\right\} \\ &+ X - E(X) \\ =& \sum_m\sum_{q\le l} n_{ql}D(\hat{\pi}_{(z)ql}^{(m)}||\bar{\pi}_{(z)ql}^{(m)}) + X - E(X), \end{aligned} \tag{3.8}$$

where

$$X = \sum_{m=1}^{M}\sum_{i<j}A_{ij}^{(m)}\log\left(\frac{\bar{\pi}_{z_iz_j}^{(m)}}{1-\bar{\pi}_{z_iz_j}^{(m)}}\right). \tag{3.9}$$

Here $D(a||b)$ is the Kullback-Liebler divergence between two Bernoulli random variables with parameters a and b respectively. This equation decomposes the difference between the maximized likelihood and its expected value in terms of $\hat{\pi}_{(z)}$ and $\bar{\pi}_{(z)}$ for a given class assignment vector z.

Next we turn our attention to RMLSBM. As mentioned before, we consider RMLSBM as a restricted version of MLSBM, and the MLE of RMLSBM can be viewed as a RMLE of MLSBM under the restrictions. Given a class assignment z, the RMLE $\hat{\pi}_{z_iz_j}^{(m)R} = \{\hat{\pi}_{(z)ql}, \hat{\beta}_{(z)m}\}$ is the maximizer of $l^R(A;z,\pi^R)$, the multi-layer block model log likelihood within the restricted parameter space. Substituting the estimated parameters in the likelihood function gives $l^R(A;z)$, the maximum of the likelihood function within the restricted parameter space. However, no closed form solution exists for the RMLE. Instead we have the following $M + K(K+1)/2$ estimating equations:

$$\frac{\partial}{\partial\beta_m} := \sum_{i<j}\left(A_{ij}^{(m)} - \frac{\exp(\hat{\pi}_{z_iz_j}+\hat{\beta}_m)}{1+\exp(\hat{\pi}_{z_iz_j}+\hat{\beta}_m)}\right), \tag{3.10}$$

$$\frac{\partial}{\partial\pi_{z_iz_j}} := \sum_{i<j}\sum_m\left(A_{ij}^{(m)} - \frac{\exp(\hat{\pi}_{z_iz_j}+\hat{\beta}_m)}{1+\exp(\hat{\pi}_{z_iz_j}+\hat{\beta}_m)}\right). \tag{3.11}$$

One of the equations is redundant since if we add the equations in (3.10), the resulting equation is identical to the sum of the equations in (3.11).

Now we use the transformation defined by ϕ in Equation (3.1). The likelihood with respect to the new parameters can be represented as

$$l^R(A;z,\phi) = \sum_{m=1}^{M}\sum_{i<j}\{A_{ij}^{(m)}\log\ \phi_{z_iz_j}^{(m)} + (1-A_{ij}^{(m)})\log\ (1-\phi_{z_iz_j}^{(m)})\}, \tag{3.12}$$

and the estimating equations in (3.10) and (3.11) can be written as

$$\begin{aligned} \frac{1}{N(N+1)/2}\sum_{q\le l} n_{ql}\hat{\phi}_{(z)ql}^{(m)} &= \frac{1}{N(N+1)/2}\sum_{q\le l}\sum_{i<j}A_{ij}^{(m)}1\{z_i=q, z_j=l\} \\ &= \frac{1}{N(N+1)/2}\sum_{i<j}A_{ij}^{(m)}, \quad m=1,\dots,M, \end{aligned} \tag{3.13}$$

$$\frac{1}{M}\sum_m \hat{\phi}^{(m)}_{(z)ql} = \frac{1}{Mn_{ql}}\sum_m\sum_{i<j} A^{(m)}_{ij} 1\{z_i = q, z_j = l\},\ q \le l \in \{1,\dots,K\}. \tag{3.14}$$

Together the right hand sides of these equations are the complete and sufficient statistics for the model. Hence we have $K(K+1)/2 + M - 1$ independent equations which will together determine the MLE of $K(K+1)/2 + M - 1$ free parameters in the set $\pi^R_{(z)}$. Here it is understood that the estimation procedure ensures that the finiteness condition of π_{ql} and β_m are respected possibly by restricting $\pi_{ql}, \beta_m \in (-C\log(MN^2),\ C\log(MN^2))$. By the functional invariance property of the MLE, $\hat{\phi}^{(m)}_{(z)ql} = \frac{\exp(\hat{\pi}_{ql}+\hat{\beta}_m)}{1+\exp(\hat{\pi}_{ql}+\hat{\beta}_m)}$ is the MLE of $\phi^{(m)}_{(z)ql}$. Note that the minimum value any $\hat{\phi}^{(m)}_{(z)ql}$ can take due to the imposed boundedness constraint is $1/MN^2$. This value is sufficiently small so that none of the partial sums in the left hand side of Equations (3.13) and (3.14) exceeds 1.

As before we define expectations of $\hat{\phi}_z$ as $\bar{\phi}_z$ and that of $l^R(A; z, \phi)$ as $\bar{l}^R_P(z, \phi)$ under the independent Bernoulli($P^{(m)}_{ij}$) model. Then,

$$\bar{l}^R_P(z,\phi) = \sum_{m=1}^{M}\sum_{i<j}\{P^{(m)}_{ij}\log(\bar{\phi}^{(m)}_{z_iz_j}) + (1-P^{(m)}_{ij})\log(1-\bar{\phi}^{(m)}_{z_iz_j})\}. \tag{3.15}$$

For a given class assignment z, $\hat{\phi}_z$ and $\bar{\phi}_z$ are the maximizers of the functions $l^R(A; z, \phi)$ and $\bar{l}^R_P(z, \phi)$ respectively, and we let $l^R(A; z)$ and $\bar{l}^R_P(z)$ denote the corresponding maximum values. The difference between the maximized values of the observed and expected likelihood can be decomposed in two parts similar to Equation (3.8) as follows

$$\begin{aligned}
&l^R(A;z) - \bar{l}^R_P(z)\\
&= \sum_m\sum_{i<j}\left\{A^{(m)}_{ij}\log\left(\frac{\hat{\phi}^{(m)}_{z_iz_j}}{\bar{\phi}^{(m)}_{z_iz_j}}\right) + (1-A^{(m)}_{ij})\log\left(\frac{1-\hat{\phi}^{(m)}_{z_iz_j}}{1-\bar{\phi}^{(m)}_{z_iz_j}}\right)\right\} + X - E(X)\\
&= \sum_m\sum_{i<j}\left\{A^{(m)}_{ij}(\hat{\pi}_{ql}+\hat{\beta}_m-\bar{\pi}_{ql}-\bar{\beta}_m) - \log\left(\frac{1+\exp(\hat{\pi}_{ql}+\hat{\beta}_m}{1+\exp(\bar{\pi}_{ql}+\bar{\beta}_m)}\right)\right\} + X - E(X)\\
&= \sum_{q\le l}(\hat{\pi}_{ql}-\bar{\pi}_{ql})\sum_m\sum_{i<j}A^{(m)}_{ij}1\{z_i=q,z_j=l\} + \sum_m(\hat{\beta}_m-\bar{\beta}_m)\sum_{i<j}A^{(m)}_{ij}\\
&\quad - \sum_m\sum_{q\le l}n_{ql}\log\left(\frac{1+\exp(\hat{\pi}_{ql}+\hat{\beta}_m}{1+\exp(\bar{\pi}_{ql}+\bar{\beta}_m)}\right) + X - E(X)\\
&= \sum_{q\le l}(\hat{\pi}_{ql}-\bar{\pi}_{ql})n_{ql}\sum_m\hat{\phi}^{(m)}_{(z)ql} + \sum_m(\hat{\beta}_m-\bar{\beta}_m)\sum_{q\le l}n_{ql}\hat{\phi}^{(m)}_{(z)ql}\\
&\quad - \sum_m\sum_{q\le l}n_{ql}\log\left(\frac{1+\exp(\hat{\pi}_{ql}+\hat{\beta}_m)}{1+\exp(\bar{\pi}_{ql}+\bar{\beta}_m)}\right) + X - E(X)\\
&= \sum_m\sum_{q\le l}n_{ql}\left\{\hat{\phi}^{(m)}_{(z)ql}\log\left(\frac{\hat{\phi}^{(m)}_{(z)ql}}{\bar{\phi}^{(m)}_{(z)ql}}\right) + (1-\hat{\phi}^{m}_{(z)ql})\log\left(\frac{1-\hat{\phi}^{(m)}_{(z)ql}}{1-\bar{\phi}^{(m)}_{(z)ql}}\right)\right\} + X - E(X)\\
&= \sum_m\sum_{q\le l}n_{ql}D\left(\hat{\phi}^{(m)}_{(z)ql}\ ||\ \bar{\phi}^{(m)}_{(z)ql}\right) + X - E(X),
\end{aligned} \tag{3.16}$$

where as before,

$$X = \sum_{m=1}^{M} \sum_{i<j} A_{ij}^{(m)} \log \left(\frac{\bar{\phi}_{z_i z_j}^{(m)}}{1 - \bar{\phi}_{z_i z_j}^{(m)}} \right). \tag{3.17}$$

Since the maximum of unrestricted likelihood would be at least as large as the maximum of restricted likelihood, we have $l(A; z) \geq l^R(A; z)$ and $\bar{l}_P(z) \geq \bar{l}_P^R(z)$ for all z.

Now let $\bar{z}$ denote the true partition. Further let $\hat{z}$ and $\hat{z}^R$ denote the maximum likelihood estimates of $\bar{z}$ under the two models MLSBM and RMLSBM respectively, i.e.,

$$\hat{z} = \arg\max_z l(A, z). \tag{3.18}$$

$$\hat{z}^R = \arg\max_z l^R(A, z). \tag{3.19}$$

3.2. Main results

We give several theorems in this subsection as we develop towards our main result. These theorems provide insights into the conditions required under the three asymptotic regimes discussed in the beginning of Section 3, which in turn provide comparison between the asymptotic behavior of the two models MLSBM and RMLSBM. Before we state the theorems, we need the following lemma that bounds the size of the set of possible estimated values of the parameters of MLSBM and RMLSBM. All the proofs are given in the Appendix A.

Lemma 1. *For a fixed z, let $\hat{\pi}_{(z)} = \{\hat{\pi}_{(z)ql}^{(m)};\ q, l \in \{1, \ldots, K\},\ m \in \{1, \ldots, M\}\}$ denote the MLE of the parameters of MLSBM, and let $\hat{\pi}_{(z)}^R = \{(\hat{\pi}_{(z)ql}, \hat{\beta}_{(z)m});\ q \leq l,\ q, l \in \{1, \ldots, K\},\ m \in \{1, \ldots, M\}\}$ be the MLE of the parameters of RMLSBM. Then for any z, we have the size of the set of all possible values that $\hat{\pi}_{(z)}$ can take as*

$$|\hat{\Pi}_{(z)}| \leq \left(\frac{N}{K} + 1 \right)^{MK(K+1)},$$

and that $\hat{\pi}_{(z)}^R$ can take as

$$|\hat{\Pi}_{(z)}^R| \leq \left(M^{1/2} \left(\frac{N}{K} + 1 \right) \right)^{K^2+K} \left(\frac{N(N+1)}{2} + 1 \right)^M,$$

where $\hat{\Pi}_{(z)}$ and $\hat{\Pi}_{(z)}^R$ denote the range of $\hat{\pi}_{(z)}$ and $\hat{\pi}_{(z)}^R$ respectively for a fixed z.

The next three theorems bound the difference in the maximized log likelihood and its expected value for both MLSBM and RMLSBM as defined in Equations (3.8) and (3.16).

Theorem 1. *Suppose a MLSBM and a RMLSBM, both with K classes and M layers, are fitted to the graph with adjacency matrix $\{A_{ij}\}_{i<j} = \{A_{ij}^{(1)}, \ldots, A_{ij}^{(M)}\}_{i<j},\ i, j = 1, \ldots, N$, where $A_{ij}^{(m)}$ are independent Bernoulli($P_{ij}^{(m)}$) trials. For any class assignment z, suppose the estimate $\hat{\pi}_{(z)}$ maximizes the multi-layer block model likelihood $l(A; z, \pi)$ and the estimate $\hat{\phi}_{(z)}$ maximizes the likelihood from the restricted model, i.e., the multi-layer block model likelihood $l^R(A; z, \phi)$ under the restricted parameter space defined by Π^R. Then for any $\epsilon > 0$,*

$$P\left(\max_z \sum_{q \leq l} n_{ql} \sum_m D\left(\hat{\pi}_{(z)ql}^{(m)} \,||\, \bar{\pi}_{(z)ql}^{(m)} \right) \geq \epsilon \right) \tag{3.20}$$
$$\leq \exp\left(N \log K + M(K^2 + K) \log \left(\frac{N}{K} + 1 \right) - \epsilon \right),$$

$$P\left(\max_z\left\{\sum_m \frac{N(N+1)}{2} D\left(\frac{\sum_{q\le l} n_{ql}\hat{\phi}^{(m)}_{(z)ql}}{N(N+1)/2} \,\Big\|\, \frac{\sum_{q\le l} n_{ql}\bar{\phi}^{(m)}_{(z)ql}}{N(N+1)/2}\right)\right\} \ge \epsilon\right) \tag{3.21}$$
$$\le \exp\left(N\log K + (K^2+K)\log\left(\frac{NM^{1/2}}{K}+1\right) + M\log\left(\frac{N(N+1)}{2}+1\right) - \epsilon\right),$$

$$P\left(\max_z\left\{\sum_{q\le l} M n_{ql} D\left(\frac{1}{M}\sum_m \hat{\phi}^{(m)}_{ql} \,\Big\|\, \frac{1}{M}\sum_m \bar{\phi}^{(m)}_{ql}\right)\right\} \ge \epsilon\right) \tag{3.22}$$
$$\le \exp\left(N\log K + (K^2+K)\log\left(\frac{NM^{1/2}}{K}+1\right) + M\log\left(\frac{N(N+1)}{2}+1\right) - \epsilon\right).$$

The first result (3.20) provides a bound for the first part of the right hand side of Equation (3.8) for MLSBM. The results (3.21) and (3.22) provide a bound that will be used in Theorem 3 to bound the first part of the corresponding likelihood decomposition for RMLSBM in Equation (3.16). In the proofs of the next two theorems, we first bound the second part of Equations (3.8) and (3.16), and then combine the results to provide a bound for the difference between the log likelihood and its expected value under any arbitrary partition z for MLSBM and RMLSBM respectively.

Theorem 2. *Suppose a MLSBM with K classes and M layers is fitted to the graph whose edges $A^{(m)}_{ij}$ are independent Bernoulli($P^{(m)}_{ij}$) trials. If we further assume that (i) $\frac{1}{MN^2} \le P^{(m)}_{ij} \le 1 - \frac{1}{MN^2}$ for all $i<j$, (ii) $K = O(N^{1/2})$, and (iii) the total expected number of edges of the entire multi-layer graph $L = \sum_m \sum_{i<j} E(A^{(m)}_{ij})$ is $\omega(MN(\log N)^{3+\delta})$ for some $\delta > 0$ as both M and N grow, then*

$$\max_z |l(A;z) - \bar{l}_P(z)| = o_P(L).$$

The result of this theorem holds under the given conditions irrespective of the relationship between the growth rates of M and N. We state the result under the first asymptotic regime mentioned at the beginning of Section 3 since we do not get any relaxation in the assumption regarding the total expected number of edges if we assume certain relations between the growth rates of M and N.

The next theorem states that the restricted likelihood in RMLSBM is also asymptotically well behaved under five independent sets of conditions corresponding to the three asymptotic regimes discussed at the beginning of Section 3. The first two sets of conditions correspond to regime 1, the third set of conditions corresponds to regime 2, and the last two sets of conditions correspond to regime 3.

Theorem 3. *Assume that a RMLSBM with K classes and M layers is fitted to the graph whose edges $A^{(m)}_{ij}$ are independent Bernoulli($P^{(m)}_{ij}$) trials. If we further assume any of the following five sets of conditions with respect to the growth of the properties of the model under different asymptotic settings:*

(i) both M and N grow, $K = O(N^{1/2})$, $\frac{1}{MN^2} \le P^{(m)}_{ij} \le C\frac{\log N}{N(\log M)^{2+\delta}}$ for all $i<j$, where C is a constant, and the total expected number of edges of the entire multi-layer graph $L = \omega(MN(\log N)^{3+\delta})$ for some $\delta > 0$;

(ii) both M and N grow but $M = O(N)$, $K = O((MN)^{1/2-\epsilon})$ for some $\epsilon > 0$, $\frac{1}{MN^2} \le P^{(m)}_{ij} \le C\frac{\log N}{N(\log M)^{2+\delta}}$ for all $i<j$, where C is a constant, and the total expected number of edges of the entire multi-layer graph $L = \omega(MN(\log N)^{3+\delta})$ for some $\delta > 0$;

(iii) M is either a constant or grows slower than N, i.e., $M = o(N)$, $K = O(N^{1/2})$, $\frac{1}{MN^2} \leq P_{ij}^{(m)} \leq C\frac{\log N}{MN(\log M)^{2+\delta}}$ for all $i < j$, where C is a constant, and the total expected number of edges of the entire multi-layer graph L is $\omega(N(\log N)^{3+\delta})$ for some $\delta > 0$;

(iv) M grows and N is either a constant or grows slower than M, i.e., $M = \omega(N)$, $K = O(\frac{N}{\log N \log M})$, $\frac{1}{MN^2} \leq P_{ij}^{(m)} \leq C\frac{1}{N\log N(\log M)^{2+\delta}}$ for all $i < j$, where C is a constant, and the total expected number of edges of the entire multi-layer graph $L = \omega(MN(\log N)^{1+\delta})$ for some $\delta > 0$;

(v) M grows and N is either a constant or grows slower than M, i.e., $M = \omega(N)$, $K = O(N^{1/2})$, $\frac{1}{MN^2} \leq P_{ij}^{(m)} \leq \min\left(C\frac{1}{N^2\log N}, C\frac{1}{N\log N(\log M)^{2+\delta}}\right)$ for all $i < j$, where C is a constant, and the total expected number of edges of the entire multi-layer graph L is larger than the the smaller of $M(\log M)^{2+\delta}(\log N)^{1+\delta}$ and $MN(\log N)^{1+\delta}$ for some $\delta > 0$;

then,

$$\max_z |l^R(A; z) - \bar{l}_P^R(z)| = o_P(L).$$

It is clear from Theorem 2 and Theorem 3 that in RMLSBM, the bound on the likelihood can be established both for relatively milder conditions on the expected total number of edges and relatively faster growth conditions on the number of communities. As we will see in Theorem 5 and the discussion following it, this enables RMLSBM to be a more attractive model for community detection either when the number of communities is large or when we have relatively sparser graphs.

Now we are ready to state our main results which show that when the true data generating process is a K-class MLSBM, the fraction of nodes misclustered by the maximum likelihood estimates and the restricted maximum likelihood estimates converge to zero under different asymptotic regimes. We define the number of "misclustered" nodes $N_e(\hat{z})$ as the number of incorrect class assignments under $\hat{z}$, counted for every node whose true class under $\bar{z}$ is not in the majority within its estimated class under $\hat{z}$ ([6]).

The previous results (Theorems 1, 2, 3) hold for any $P_{ij}^{(m)}$ whenever they are bounded as described in the theorems. Now we assume further structure on the probabilities, namely a MLSBM. Denote the true partition as $\bar{z}$, and under the true partition, let the true block model parameter array be $\bar{\pi}$. Hence, under MLSBM we have

$$P_{ij}^{(m)} = \bar{\pi}_{\bar{z}_i\bar{z}_j}^{(m)}.$$

Consequently, $\bar{l}_P(\bar{z}, \pi)$ from Equation (3.7) is maximized by the true model parameter $\bar{\pi}$, and we have the maximized expected likelihood as

$$\bar{l}_P(\bar{z}) = \sum_{m=1}^{M}\sum_{q\leq l} n_{ql}\{\bar{\pi}_{ql}^{(m)}\log\bar{\pi}_{ql}^{(m)} + (1-\bar{\pi}_{ql}^{(m)})\log(1-\bar{\pi}_{ql}^{(m)})\}. \tag{3.23}$$

On the other hand, the expected restricted likelihood is maximized by the parameter array $\bar{\pi}^R$ under the restricted parameter space of RMLSBM. Note that this is different from the true model parameter array $\bar{\pi}$ due to the restrictions imposed on the parameter space. Using the transformation

introduced in Equation (3.1), the maximized expected restricted likelihood is

$$\begin{aligned}\bar{l}_P^R(\bar{z}) &= \sum_{m=1}^{M}\sum_{i<j}\{P_{ij}^{(m)}\log\bar{\phi}_{\bar{z}_i\bar{z}_j}^{(m)} + (1-P_{ij}^{(m)})\log(1-\bar{\phi}_{\bar{z}_i\bar{z}_j}^{(m)})\} \\ &= \sum_{m=1}^{M}\sum_{i<j}\{\bar{\pi}_{\bar{z}_i\bar{z}_j}^{(m)}\log\bar{\phi}_{\bar{z}_i\bar{z}_j}^{(m)} + (1-\bar{\pi}_{\bar{z}_i\bar{z}_j}^{(m)})\log(1-\bar{\phi}_{\bar{z}_i\bar{z}_j}^{(m)})\} \\ &= \sum_{m=1}^{M}\sum_{q\le l} n_{ql}\{\bar{\pi}_{ql}^{(m)}\log\bar{\phi}_{ql}^{(m)} + (1-\bar{\pi}_{ql}^{(m)})\log(1-\bar{\phi}_{ql}^{(m)})\}. \end{aligned} \tag{3.24}$$

For MLSBM, if the conclusion $\max_z |l(A;z) - \bar{l}_P(z)| = o_P(L)$ of Theorem 2 holds, the data are generated according to a K-class blockmodel with membership vector $\bar{z}$ and probability matrix $\bar{\pi}$, and the maximum-likelihood K-class blockmodel class assignment estimator is $\hat{z}$, then it is easy to see

$$\begin{aligned}\bar{l}_P(\bar{z}) - \bar{l}_P(\hat{z}) &\le \bar{l}_P(\bar{z}) - \bar{l}_P(\hat{z}) + l(A,\hat{z}) - l(A,\bar{z}) \\ &\le |\bar{l}_P(\bar{z}) - l(A,\bar{z})| + |\bar{l}_P(\hat{z}) - l(A,\hat{z})| = o_P(L).\end{aligned} \tag{3.25}$$

Note that the terms $\bar{l}_P(\bar{z}) - \bar{l}_P(\hat{z})$ and $l(A,\hat{z}) - l(A,\bar{z})$ are positive quantities as mentioned earlier. The next theorem relates Equation (3.25) with the fraction of misclustered nodes $N_e(\hat{z})$ and the expected total number of edges L to establish a bound for the misclustering rate.

Theorem 4. *Suppose the data are generated according to a K-class MLSBM with membership vector $\bar{z}$ and parameter array $\bar{\pi}$, the conclusion of Theorem 2 holds, and the following conditions hold with respect to the model sequence: for all blockmodel classes $q = 1, \ldots, K$, class size N_q grows as $s = \min_q\{N_q\} = \Omega(N/K)$, and over all distinct class pairs (q,l) and all classes $c \ne \{q,l\}$,*

$$\begin{aligned}&\min_{q,l}\min_{m}\max_{c}\left\{D\left(\bar{\pi}_{qc}^{(m)}\,\middle\|\,\frac{\bar{\pi}_{qc}^{(m)}+\bar{\pi}_{lc}^{(m)}}{2}\right) + D\left(\bar{\pi}_{lc}^{(m)}\,\middle\|\,\frac{\bar{\pi}_{qc}^{(m)}+\bar{\pi}_{lc}^{(m)}}{2}\right)\right\} \\ &\qquad = \Omega\left(\frac{LK}{MN^2}\right),\end{aligned} \tag{3.26}$$

then

$$N_e(\hat{z}) = o_P(N). \tag{3.27}$$

Note that condition (3.26) is very similar to condition (ii) of Theorem 3 in [6] with the total number of edges for the single layer case being replaced by the average number of edges L/M in each layer for the multi-graph. This ensures that any two rows in any of the layer matrices $\bar{\pi}^{(m)}$ of $\bar{\pi}$ differ in at least one entry by at least a constant times $\frac{LK}{MN^2}$. Also, when we take into account the asymptotic conditions required on the growth of K and L for the result of Theorem 2 to hold, i.e., $K = O(N^{1/2})$ and $L = \omega(MN(\log N)^{3+\delta})$ with M and N both growing, then we have $\frac{LK}{MN^2} = \omega\left(\frac{(\log N)^{3+\delta}}{N^{1/2}}\right)$. As argued in [6], if L is close to its least possible rate of growth, $\frac{LK}{MN^2}$ goes to 0 for large N and the condition is not too prohibitive. For example, if $L = MN(\log N)^{\beta}$ with $\beta > 4$, then $(\log N)^{\beta} = o(N^{1/2})$, so $\frac{LK}{MN^2}$ goes to 0 and the condition is not overly restrictive.

We state the corresponding conclusion for the restricted likelihood estimation (for RMLSBM) over two lemmas and one theorem. The first lemma bounds the difference between the maximized expected likelihoods from the unrestricted and the restricted models under the true partition. The

second lemma uses this result along with the result of Theorem 3 to bound the difference between the maximized expected likelihood for the restricted model under the RMLE and the maximized expected likelihood for the unrestricted model under the true partition.

Lemma 2. *Under the true partition $\bar{z}$, if any of the five sets of conditions in Theorem 3 on the growth of multi-layer blockmodel parameters holds, then $\bar{l}_P(\bar{z}) - \bar{l}_P^R(\bar{z}) = o_P(L)$, where L is the expected number of edges in the multi-layer graph under the corresponding set of conditions.*

Lemma 3. *Under the true partition $\bar{z}$ and the RMLE of the partition $\hat{z}^R$ (i.e., the MLE in the restricted model RMLSBM), we have $\bar{l}_P(\bar{z}) - \bar{l}_P^R(\hat{z}^R) = o_P(L)$ whenever the conclusion of Theorem 3 holds.*

Now we are ready to show that the class membership assignment vector estimated through the maximum likelihood estimation in the restricted model RMLSBM is consistent under data generated from the MLSBM.

Theorem 5. *Suppose the data are generated according to a K-class MLSBM with membership vector $\bar{z}$ and parameter array $\bar{\pi}$, the conclusion of Lemma 3 holds, and the following conditions hold with respect to the model sequence: for all blockmodel classes $q = 1, \ldots, K$, class size N_q grows as $s = \min_q\{N_q\} = \Omega(N/K)$, and over all distinct class pairs (q, l) and all classes $c \neq \{q, l\}$,*

$$\min_{q,l} \min_{m} \max_{c} \left\{ D\left(\bar{\pi}_{qc}^{(m)} \,\middle\|\, \frac{\bar{\pi}_{qc}^{(m)} + \bar{\pi}_{lc}^{(m)}}{2}\right) + D\left(\bar{\pi}_{lc}^{(m)} \,\middle\|\, \frac{\bar{\pi}_{qc}^{(m)} + \bar{\pi}_{lc}^{(m)}}{2}\right) \right\} = \Omega(g), \tag{3.28}$$

then under any of the five sets of growth conditions in Theorem 3, we have

$$N_e(\hat{z}^R) = o_P(h). \tag{3.29}$$

Here g in condition (3.28) and the growth rate h depend on the asymptotic conditions imposed on K and L. The growth rate h can be determined from g by the relationship $h = \frac{KL}{MNg}$. In particular, (i) when $K = O(N^{1/2})$, $L = \omega(MN(\log N)^{3+\delta})$ with M and N both growing arbitrarily, then we have $g = \frac{LK}{MN^2} = \omega\left(\frac{(\log N)^{3+\delta}}{N^{1/2}}\right)$ and $h = N$; (ii) when $K = O((MN)^{1/2-\epsilon})$, $L = \omega(MN(\log N)^{3+\delta})$ with M and N both growing so that $M = O(N)$, then we have $g = \frac{LK}{MN^2} = \omega\left((\frac{M}{N})^{1/2}\right)$ and $h = N$; (iii) when $K = O(N^{1/2})$, $L = \omega(N(\log N)^{3+\delta})$ and $M = o(N)$, then we have $g = \frac{LK}{N^2} = \omega\left(\frac{(\log N)^{3+\delta}}{N^{1/2}}\right)$ and $h = N/M$; (iv) when $K = O(N^{1-\epsilon}/\log M)$, $L = \omega(MN(\log N)^{1+\delta}$ and $M = \omega(N)$, then we have $g = \frac{LK}{MN^2} = \omega\left(\frac{1}{\log M}\right)$ and $h = N$; (v) when $K = O(N^{1/2})$, L is $\omega(MN(\log N)^{1+\delta})$ if $N < (\log M)^{2+\delta}$ or $\omega(M(\log M)^{2+\delta}(\log N)^{1+\delta})$ if $N > (\log M)^{2+\delta}$ and $M = \omega(N)$, then we have $g = \frac{LK}{MN^2} = \omega\left(\frac{(\log N)^{1+\delta}}{N^{1/2}}\right)$ or $g = \frac{LK}{MN^2} = \omega\left(\frac{(\log M)^{2+\delta}(\log N)^{1+\delta}}{N^{3/2}}\right)$ and $h = N$.

Note that in Theorem 5, we have used generic notations g and h to denote functions of the network properties such as N, K and L. The functions g and h vary across asymptotic setups as well as across the models. This is so because the regularity condition (3.28) on the difference among the elements of block model probability matrices should be as less prohibitive as possible. Note that in our results, we have chosen g in such a way that if L is close to its least possible rate of growth, then g asymptotically decays to 0 under the assumed asymptotic setup. This ensures that our condition (3.28) is not overly restrictive. It also enables us to understand and contrast the asymptotic behavior of the models from a unified point of view.

3.3. Sparse networks

The results of all previous theorems imply that for sparse multi-layer networks, consistency can be achieved with a large number of relatively sparser graphs as long as they together satisfy the edge density requirement. In the case when M grows slower than N, in MLSBM we do not get any relaxation in the required growth condition on the total expected number of edges from all the graph layers combined, and it remains $\omega(MN(\log N)^{3+\delta})$ for $K = O(N^{1/2})$. However in RMLSBM we only require the total expected number of edges from all layers to be $\omega(N(\log N)^{3+\delta})$ for $K = O(N^{1/2})$ (Condition (iii) of Theorem 3). This implies that we only require the expected number of edges per layer to be $\omega(N(\log N)^{3+\delta}/M)$ on average. For perspective, if M grows faster than $(\log N)^{3+\delta}$, then the average number of edges per layer needs to grow only at $O(N)$, which is the sparse bounded degree regime. This case is extremely challenging for single layer networks. In comparison, the consistency of the MLE in MLSBM requires the average expected number of edges per layer to be $\omega(N(\log N)^{3+\delta})$ ([6]) and hence the average degree per layer must grow atleast as $(\log N)^{3+\delta}$. Thus consistency can be achieved with a large number of relatively sparse layers. This is particularly important as most modern applications of community detection in multi-layer graph fall under this asymptotic scenario.

3.4. A Large number of communities

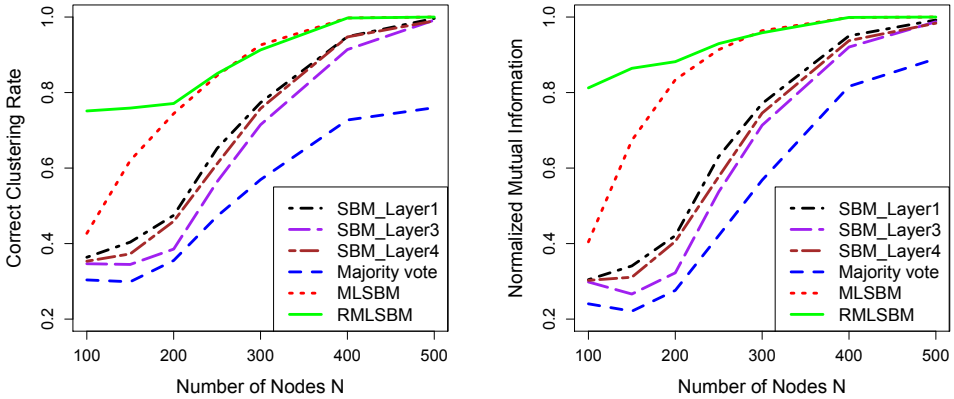
Under MLSBM, consistent community detection is possible when the number of communities grows as $K = O(N^{1/2})$ and the total expected number of edges is $\omega(MN(\log N)^{3+\delta})$ as both M and N grow. However, if we assume $K = O((MN)^{1/2-\epsilon})$ for some $\epsilon > 0$, then we require the total expected number of edges to be $\omega(M^2N(\log N)^{3+\delta})$ which is unrealistically dense. On the other hand, under RMLSBM consistent estimation is possible with comparable edge density even when the number of communities grows faster, either as $K = O((MN)^{1/2-\epsilon})$ when both M and N grow but $M = O(N)$, or as $K = O(\frac{N}{\log M \log N})$ when N grows slower than M (Conditions (ii) and (iv) of Theorem 3). Hence the restricted model is advantageous for community detection in networks with a large number of communities.

4. Estimation using mixture model approach

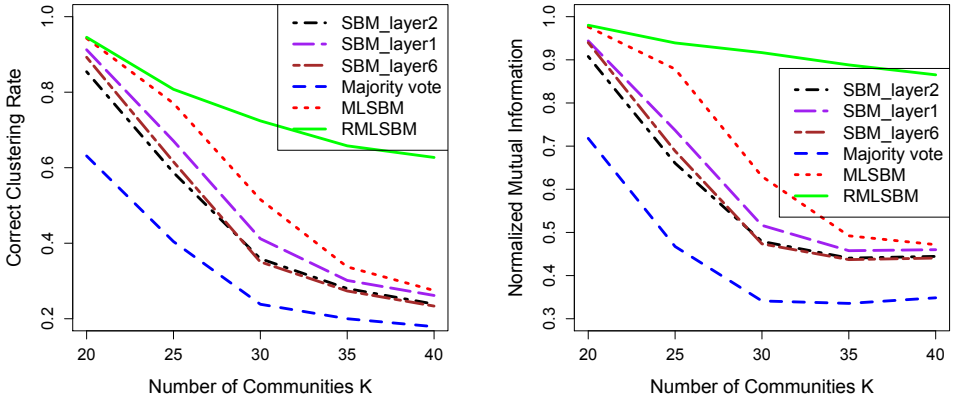
Simultaneous maximum likelihood estimation of parameters and class assignments in the stochastic blockmodel is a difficult problem ([25], [6], [28]). The same difficulties remain in the MLSBM and its restricted version. Consequently, to obtain an estimation algorithm here, we view the MLSBM as a mixture model with discrete latent variables Z. In this case, Z_i is a missing random variable that follows a multinomial distribution with K parameters: $Z_i \sim Mult(1, \alpha = (\alpha_1, \alpha_2, \ldots, \alpha_K))$. We follow the framework laid out by [8] to simultaneously estimate the conditional blockmodel parameters and the class assignments with variational EM technique. The derivations for MLSBM are straightforward extensions of the corresponding formula in [8] and are omitted in this paper while the update rules for RMLSBM have been derived in the Appendix B. The update steps for MLSBM and RMLSBM are also provided in the Appendix B under Algorithm 1 and Algorithm 2 respectively.

5. Simulation results

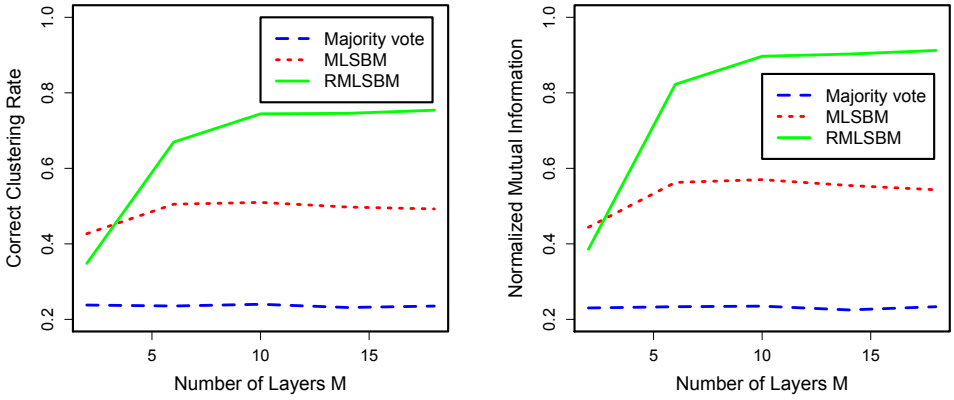
In this section we numerically test the asymptotic results and compare the performance of the models through a small simulation study. We generate data from the more general model, MLSBM. We then



(a)



(b)



(c)

Fig 2: Comparison of the performance of various methods for three simulation settings in terms of CCR and NMI: (a) fixed $K = 10$ and $M = 5$ while N increases from 100 to 500; (b) fixed $N = 400$ and $M = 6$ while K increases from 20 to 40; (c) fixed $N = 200$ and $K = 15$ while M increases from 2 to 18. All CCR and NMI values are averaged over 50 simulations.

compare the relative performance of the two multi-layer methods between themselves as well as with single layer methods and baseline methods such as majority voting. The comparison is done under various settings on the number of nodes N, the number of communities K, the number of types of relations M, and the expected total number of edges L. For computing the majority voting solution, the cluster labels of different single layer algorithms are aligned with each other by solving the Linear Sum Assignment Problem.

Since the true class labels of the nodes are known in simulated data, we compare the class assignments from different methods with the true labels. We use correct clustering rate (CCR) and normalized mutual information (NMI) as measures of similarity between partitions. The CCR counts the fraction of nodes whose cluster assignment matches the true class label (as determined by the true class label of the majority of nodes in that cluster). The higher the CCR, the better the performance of the clustering method. However note that this definition of "correct clustering" does not involve optimal assignment of the cluster labels to class labels and hence is different from the usual correct classification rate used in supervised learning. On the other hand, the NMI is an information theoretic measure of the mutual dependence or similarity of two random variables. The NMI takes values in the range of 0 to 1, with 0 indicating random cluster assignment with respect to the true class labels, and 1 indicating perfect match between the true and assigned clusters. If NMI is 0, it means even though the cluster assignment was not completely random and done according to some algorithm, the solution presents no information regarding the true class labels.

In all the simulation studies we repeat the experiments 50 times and take the average of our measures across them. We first generate the node labels independently from a multinomial distribution with probabilities $P(Z_i = k) = \alpha_k$. Then we generate the data using the node labels and M connectivity matrices, all of which give larger probability to connections within groups in comparison to the connections between groups. The general structure of the connectivity matrix is

$$P_{K\times K} = \lambda I_K + \epsilon 1_{K\times K} - \epsilon I_K, \tag{5.1}$$

so that all the diagonal elements are λ and all the off diagonal elements are ϵ. The parameters λ and ϵ differ among simulation studies as well as among layers in a particular simulation.

5.1. Fixed K and M while N increases

In this simulation, we take $M = 5$ types of edges or network layers, each with a separate connectivity matrix inducing a different network. We keep the number of communities K fixed at 10 and vary the number of nodes N from 100 to 500. The aim of this study is to compare the two multi-layer methods with the single layer methods and baseline methods in terms of the number of nodes required to achieve a consistent estimation of community assignment with moderately low number of communities. The values of the parameters of the connectivity matrix, λ and ϵ, were chosen with the aim of having some variation among the layers in terms of edge density while keeping the ratio between λ and ϵ at roughly $4 : 1$. Figure 2(a) displays the results from this study. Clearly the two multi-layer methods reach CCR and NMI of close to 1 faster than the single layer ones as the number of nodes increases. The accuracy of majority voting behaves similarly to the single layer ones. Moreover, for a small number of nodes, the RMLSBM performs better than all the other methods considered.

5.2. Fixed N and M while K increases

In this simulation, we test the performance of the two multi-layer methods against the single layer and baseline methods with increasing number of communities. We fix the number of nodes N and

the number of layers M at 400 and 6 respectively, while we let K increase from 20 to 40 in steps of 5. The results from this simulation study are displayed in Figure 2(b). Whereas the accuracy of community detection in all the single-layer methods and the baseline method decreases rapidly with increasing number of communities, the multi-layer methods explored here, especially the RMLSBM, perform well even with a large number of communities. Between RMLSBM and MLSBM, RMLSBM clearly outperforms MLSBM as the number of communities grows. In terms of NMI, while the score for MLSBM, majority voting and the single layer SBMs reduce below 0.5, it settles to a value close to 0.9 for RMLSBM as the number of communities increases to 40. Similarly, the CCR for RMLSBM settles around 0.7 while for all other methods described here it decreases below 0.4.

5.3. Fixed N and K while M increases

In this simulation, we keep the number of nodes N and the number of communities K fixed at 200 and 15 respectively, while we increase the number of layers M gradually from 2 to 20. Each layer of the multi-layer network was generated from a K-class stochastic block model with the connectivity matrix given in Equation (5.1) with parameters $\epsilon = 0.15 \pm U(-0.04,\ 0.04)$ and $\lambda = 3\epsilon$. Here $U(a, b)$ is a random number from the uniform distribution between a and b. We compare the performance of MLSBM and RMLSBM with majority voting in terms of the accuracy of community detection. The curves for majority votes in Figure 2(c) remain almost flat with increasing number of layers, indicating that the accuracy of community detection does not improve with more layers. For MLSBM, the accuracy increases initially, however the improvement quickly stops being significant and both the curves flatten with increasing layers. For RMLSBM, however, the accuracy of community detection increases with the number of layers and its CCR and NMI values are much higher than the other two methods.

The three studies clearly point out the advantages of the multi-layer methods over the single layer ones and the baseline one, as well as the relative advantage of RMLSBM over MLSBM within the scope of the simulations.

6. Twitter UK politics dataset

In this section we test our method on a real dataset on interactions between British Members of Parliament (MPs) in the social networking site Twitter curated by [12]. Although the original dataset consists of 419 nodes we only considered the largest subset that is connected across all layers for our analysis. Hence our multi-layer network consisted of 381 nodes. The different layers of network we have correspond to three direct relations: "mentions", "follows" and "retweets", and three derived relations, "mentioned by the same person (co-mentions)", "followed by the same person (co-follow)", and "retweeted by the same person (co-retweets)". All relations are assumed to be binary by assigning one if the relation is true for at least one case (e.g., if at least one person follows both MP i and MP j, then the relation "co-follow" between the two MPs is true). All the relations individually can be represented as graphs. For the graphs with direct relations, "mentions", "follows", and "retweets", a directed edge from node i to node j implies that MP i mentioned, followed or retweeted respectively MP j at least once in his/her tweets. We converted all directed edges into undirected edges for this analysis. Average degrees of nodes in different network layers are presented in Table 1. Note that among the direct layers, "follows" is relatively dense compared to "mentions" and "retweets", while the derived networks are overall much denser compared to the direct ones.

The goal here is to cluster the MPs into communities based on the information about their twitter activities. The ground truth communities are known to be consisting of five communities corresponding to the political affiliations of the MPs: Conservatives, Labour, Liberal Democrats,

Table 1
Average degrees of nodes in different network layers for Twitter UK politics dataset

Mentions	Follows	Retweets	Co-mention	Co-follow	Co-retweet
58.48	98.34	31.88	361.51	297.21	147.56

Table 2
Ground truth community sizes

Conservative	Labour	Liberal Democrat	SNP	other
152	178	39	5	7

SNP and Other, and the sizes of the five communities are given in Table 2. The clustering quality is assessed through NMI and CCR as before. As the optimization methods used here are sensitive to the starting values, we use several starting values and report the result corresponding to the highest score of the objective function.

Part (a) of Table 3 reports the performance of the algorithm for the six individual layers considered. Note that the performance of the derived networks is worse compared to the direct ones despite being denser. Clearly the signal in favor of the ground truth is stronger in the "direct networks" compared to the "derived networks". The performance of majority vote, MLSBM and RMLSBM on multi-layer networks constructed from the three direct layers and all layers together are given in part (b) of Table 3. In both cases the multi-layer methods outperform majority voting, and between the two multi-layer methods, RMLSBM outperforms MLSBM. From the results for direct networks, we note that the performance of multi-layer models is not affected by inclusion of relatively sparse networks ("mentions", "retweets") and multi-layer models perform better than the densest layer ("follows"), as long as all the signal strength is high. However the performance deteriorates as the signal quality becomes bad with the inclusion of poor performing derived networks. RMLSBM is more robust towards such layers with poor signal compared to MLSBM.

Table 3
The NMI and CCR for Twitter UK politics data

Measure	Mentions	Follows	Retweets	Co-mentions	Co-follows	Co-retweets
NMI	0.4522	0.5992	0.4610	0.3449	0.2520	0.4009
CCR	0.8182	0.9022	0.7926	0.7565	0.7053	0.8136

(a) Individual network layers

	NMI			CCR		
	Majority	MLSBM	RMLSBM	Majority	MLSBM	RMLSBM
Direct networks	0.5213	0.6764	0.6821	0.8477	0.9527	0.9553
All networks	0.3825	0.5428	0.6250	0.7217	0.8393	0.9107

(b) Combined network layers

7. Discussions

In this paper we extended the stochastic block model to the multi-layer settings with two related models, MLSBM and its restricted version RMLSBM. We used these models for community detection in multi-layer networks through maximum likelihood estimation. We solved the challenging problem of simultaneous inference of model parameters and latent community assignments with variational EM algorithm combined with gradient descent. The maximum likelihood estimates of both models are consistent under data generated from the more general model MLSBM with suitable conditions on the growth rate of the number of communities, the number of types of edges/layers, and the total number of edges of the entire multi-layer graph.

Extensive simulation studies confirmed the superiority of the proposed methods over the baseline methods (majority voting) and methods for single-layer networks. In the simulation studies, the RMLSBM almost always outperforms the single-layer methods as well as the majority voting and MLSBM, when either the number of communities is large or the graph layers are relatively sparse. This includes the case when the individual layers have bounded average degree, which is an extremely challenging case for single layer networks. We would like to emphasize that handling the bounded degree case would not be possible with the usual MLSBM extension. The observations of this paper are in line with previous work on regularization, especially in regression models where a parsimonious model with similar accuracy is preferred over a model with a large number of parameters. The RMLSBM approximates the MLSBM quite well with fewer parameters for most multi-layer networks with a large number of communities. Hence in small networks or relatively sparse ones the RMLSBM outperforms the MLSBM.

APPENDIX A

For brevity of notation henceforth in the Appendix, we remove the subscript (z) from $\pi_{(z)}$, $\pi^R_{(z)}$ and $\phi_{(z)}$, denoting the set of parameters of MLSBM, RMLSBM and the transformation of the set of parameters of RMLSBM respectively for a fixed z. We also remove the subscript (z) from $\hat{\Pi}_{(z)}$ and $\hat{\Pi}^R_{(z)}$.

Proof of Lemma 1

We first determine the size of the set of all possible values that the MLE of the parameter array π can take in the MLSBM. Notice that from Equation (3.5) the estimate $\hat{\pi}^{(m)}$ of the parameter matrix for any layer m can take any of the $\prod_{q\leq l}(n_{ql}+1)$ values, since its $K(K+1)/2$ upper diagonal components ($\hat{\pi}^{(m)}_{ql}$, $q \leq l$, $q, l \in \{1, \ldots, K\}$) can take any of the $n_{ql}+1$ values in the set $\{0, 1/n_{ql}, \ldots, 1\}$ independently. Hence, $|\hat{\Pi}| = \prod_m \prod_{q\leq l}(n_{ql}+1)$. However this is subject to the constraint that $\sum_{q\leq l} n_{ql} = \binom{N}{2}$. This implies that $|\hat{\Pi}|$ is a product of $\binom{K+1}{2}$ positive terms whose sum is fixed. So $|\hat{\Pi}|$ is maximized when the terms are all equal, i.e., $n_{ql} = \binom{N}{2} \Big/ \binom{K+1}{2}$ uniformly across all m. Hence we have the following inequality

$$
\begin{aligned}
|\hat{\Pi}| &\leq \left(\binom{N}{2} \Big/ \binom{K+1}{2} + 1 \right)^{MK(K+1)/2} \\
&< \left(\frac{N^2}{K^2} + 1 \right)^{MK(K+1)/2} < \left(\frac{N}{K} + 1 \right)^{MK(K+1)} .
\end{aligned}
$$

Now we turn our attention to the set of values the MLE of the parameter array in RMLSBM can take. Note that Equations (3.13) and (3.14) together represent $K(K+1)/2+M$ equations involving partial sums of the MLEs of the $K(K+1)/2+M$ elements in the parameter array π^R (although the equations are written in terms of the transformation ϕ for convenience, they actually represent the same equations as Equations (3.10) and (3.11). The right hand side of the equations together are the sufficient statistics under the RMLSBM. Note that due to the identifiablility constraint, we have only $K(K+1)/2+M-1$ free parameters. On the other hand, one of the equations in the set of equations is also redundant, since adding together the first M equations represented by Equation (3.13) and adding the remaining $K(K+1)/2$ equations represented by Equation (3.14) yield the

same equation and hence there is one linear dependence. This set of equations determines the MLE of π^R. Hence the size of the set of all distinct solutions $\hat{\pi}^R$ is at most the number of possible sets of system of equations. To determine the later, we notice that the right hand side of each of the first set of M equations can take $N(N+1)/2+1$ values from the set $\{0, 2/[N(N+1)], \ldots, 1\}$, while the right hand side of each of the next set of $K(K+1)/2$ equations can take $Mn_{ql}+1$ values from the set $\{0, 1/(Mn_{ql}), \ldots, 1\}$. So the size of the set of possible values the estimated parameter array $\hat{\pi}^R$ can take is

$$|\hat{\Pi}^R| \leq \prod_{q \leq l}(Mn_{ql}+1) \prod_{m=1}^{M} \left(\frac{N(N+1)}{2} + 1 \right).$$

The first term is maximized as before when all the n_{ql}'s are equal, i.e., $n_{ql} = \binom{N}{2} \Big/ \binom{K+1}{2}$. The second term is a fixed quantity. So we have

$$\begin{aligned}
|\hat{\Pi}^R| &\leq \left(M \binom{N}{2} \Big/ \binom{K+1}{2} + 1 \right)^{K(K+1)/2} \left(\frac{N(N+1)}{2} + 1 \right)^M \\
&\leq \left(M \frac{N^2}{K^2} + 1 \right)^{K(K+1)/2} \left(\frac{N(N+1)}{2} + 1 \right)^M \\
&\leq \left(M^{1/2} \frac{N}{K} + 1 \right)^{K(K+1)} \left(\frac{N(N+1)}{2} + 1 \right)^M.
\end{aligned}$$

Lastly notice that the transformation defined by Equation (3.1) is an onto function but not necessarily one-to-one, so one or more parameter arrays π^R map to one ϕ. Hence for every estimate $\hat{\phi}$ there exists a corresponding estimate array $\hat{\pi}^R$. Therefore we have

$$|\hat{\Phi}| \leq |\hat{\Pi}^R| \leq \left(M^{1/2} \frac{N}{K} + 1 \right)^{K(K+1)} \left(\frac{N(N+1)}{2} + 1 \right)^M.$$

Proof of Theorem 1

The proof closely follows the proof of Theorem 1 in [6]. Following the arguments in [6], we first notice that for a fixed z, each estimate $\hat{\pi}_{ql}^{(m)}$ is a sum of n_{ql} independent Bernoulli random variables with mean $\bar{\pi}_{ql}^{(m)}$. Hence the probability that $\hat{\pi}_{ql}^{(m)} = \nu$, where $\nu \in \{0, 1/n_{ql}, \ldots, 1\}$ can be bounded as

$$P(\hat{\pi}_{ql}^{(m)} = \nu) \leq \exp\left(-n_{ql} D(\nu \mid\mid \bar{\pi}_{ql}^{(m)}) \right),$$

and by the independence of $A_{ij}^{(m)}$, the bound on the probability of any realization $\hat{\pi}$ is

$$P(\hat{\pi}) \leq \exp\left(-\sum_{q \leq l} n_{ql} \sum_m D(\hat{\pi}_{ql}^{(m)} \mid\mid \bar{\pi}_{ql}^{(m)}) \right).$$

Recall $\hat{\Pi}$ denotes the set of values the estimate array $\hat{\pi}$ can take for a fixed class assignment z. In Lemma 1, we have bounded the size of this set as $|\hat{\Pi}| \leq \left(\frac{N}{K}+1\right)^{MK(K+1)}$. Now we consider the event that $\sum_{q \leq l} n_{ql} \sum_m D(\hat{\pi}_{ql}^{(m)} \mid\mid \bar{\pi}_{ql}^{(m)})$ is at least as large as some $\epsilon > 0$, and derive an upper

bound for its probability of occurrence:

$$P(\hat{\Pi}_\epsilon) = P\left(\hat{\pi} \in \hat{\Pi};\ \sum_{q\le l} n_{ql} \sum_m D(\hat{\pi}_{ql}^{(m)} \ || \ \bar{\pi}_{ql}^{(m)}) \ge \epsilon\right) = \sum_{\hat{\pi}\in\hat{\Pi}_\epsilon} P(\hat{\pi})$$

$$\le \sum_{\hat{\pi}\in\hat{\Pi}_\epsilon} \exp\left(-\sum_{q\le l} n_{ql} \sum_m D(\hat{\pi}_{ql}^{(m)} \ || \ \bar{\pi}_{ql}^{(m)})\right) \le \sum_{\hat{\pi}\in\hat{\Pi}_\epsilon} \exp(-\epsilon)$$

$$= |\hat{\Pi}_\epsilon| \exp(-\epsilon) \le |\hat{\Pi}| \exp(-\epsilon) \le \left(\frac{N}{K}+1\right)^{MK(K+1)} \exp(-\epsilon)$$

Hence for all $\epsilon > 0$, we have over all K^N possible class assignments z ,

$$P\left(\max_z \sum_{q\le l} n_{ql} \sum_m D(\hat{\pi}_{ql}^{(m)} \ || \ \bar{\pi}_{ql}^{(m)}) \ge \epsilon\right) \le P\left(\bigcup_z \left\{\sum_{q\le l} n_{ql} \sum_m D(\hat{\pi}_{ql}^{(m)} \ || \ \bar{\pi}_{ql}^{(m)}) \ge \epsilon\right\}\right)$$

$$\le K^N \exp\left(MK(K+1)\log\left(\frac{N}{K}+1\right) - \epsilon\right) \le \exp\left(N\log K + M(K^2+K)\log\left(\frac{N}{K}+1\right) - \epsilon\right).$$

The proof for the restricted case, although follows the same structure as before, is more involved as we need to deal with estimating equations instead of closed form solutions. Note that for a fixed z, the left hand side of each of the M estimating equations in (3.13) is $\frac{1}{N(N+1)/2}\sum_{q\le l} n_{ql}\hat{\phi}_{ql}^{(m)}$, which is a sum of $N(N+1)/2$ independent Bernoulli random variables with mean $\frac{1}{N(N+1)/2}\sum_{q\le l} n_{ql}\bar{\phi}_{ql}^{(m)}$ respectively. Hence the probability that
$\frac{1}{N(N+1)/2}\sum_{q\le l} n_{ql}\hat{\phi}_{ql}^{(m)} = \nu_m$, where $\nu_m \in \{0, 2/[N(N+1)], \ldots, 1\}$ can be bounded as

$$P\left(\frac{\sum_{q\le l} n_{ql}\hat{\phi}_{ql}^{(m)}}{N(N+1)/2} = \nu_m\right) \le \exp\left(-\frac{N(N+1)}{2} D\left(\nu_m \ \Big|\Big| \ \frac{\sum_{q\le l} n_{ql}\bar{\phi}_{ql}^{(m)}}{N(N+1)/2}\right)\right),$$

for $m \in \{1, \ldots, M\}$.

Similarly the left hand side of each of the $K(K+1)/2$ estimating equations in (3.14) is $\frac{1}{M}\sum_m \hat{\phi}_{ql}^{(m)}$, which is a sum of Mn_{ql} independent Bernoulli random variables with mean $\frac{1}{M}\sum_m \bar{\phi}_{ql}^{(m)}$. Hence the probability that
$\frac{1}{M}\sum_m \hat{\phi}_{ql}^{(m)} = \nu_{ql}$, where $\nu_{ql} \in \{0, 1/(Mn_{ql}), \ldots, 1\}$ can be bounded as

$$P\left(\frac{1}{M}\sum_m \hat{\phi}_{ql}^{(m)} = \nu_{ql}\right) \le \exp\left(-Mn_{ql} D\left(\nu_{ql} \ \Big|\Big| \ \frac{1}{M}\sum_m \bar{\phi}_{ql}^{(m)}\right)\right),$$

for $q \le l$, $q, l \in \{1, \ldots, K\}$.

Now since these $K(K+1)/2 + M$ estimating equations together determine the MLE $\hat{\pi}^R$ of RMLSBM, the probability of any realization of $\hat{\pi}^R$ is bounded by the joint probability of the occurrence of the estimating equations. Note that although the equations within the two sets ((3.13) and (3.14)) are independent of each other, the two sets of equations are not independent of each other.

Hence because of the inequalities that $P(A \cap B) \leq P(A)$ and $P(A \cap B) \leq P(B)$, we have

$$P(\hat{\pi}^R) \leq \prod_m P\left(\frac{1}{N(N+1)/2}\sum_{q\leq l} n_{ql}\hat{\phi}_{ql}^{(m)}\right) \tag{7.1}$$
$$\leq \exp\left(-\sum_m \frac{N(N+1)}{2} D\left(\frac{\sum_{q\leq l} n_{ql}\hat{\phi}_{ql}^{(m)}}{N(N+1)/2} \,\Big\|\, \frac{\sum_{q\leq l} n_{ql}\bar{\phi}_{ql}^{(m)}}{N(N+1)/2}\right)\right),$$

and

$$P(\hat{\pi}^R) \leq \prod_{q\leq l} P\left(\frac{1}{M}\sum_m \hat{\phi}_{ql}^{(m)}\right) \tag{7.2}$$
$$\leq \exp\left(-\sum_{q\leq l} M n_{ql} D\left(\frac{1}{M}\sum_m \hat{\phi}_{ql}^{(m)} \,\Big\|\, \frac{1}{M}\sum_m \bar{\phi}_{ql}^{(m)}\right)\right).$$

For brevity, we call the right hand sides of Equations (7.1) and (7.2) as $\exp(-E_1)$ and $\exp(-E_2)$ respectively. From Lemma 1, we have the size of set of all possible values $\hat{\pi}^R$ can take

$$|\hat{\Pi}^R| \leq \left(M^{1/2}\frac{N}{K}+1\right)^{K(K+1)}\left(\frac{N(N+1)}{2}+1\right)^M.$$

Now we consider the event that E_i is at least as large as some $\epsilon > 0$ for $i = 1, 2$ respectively.

$$P(\hat{\Pi}_\epsilon^R) = P(\hat{\pi}^R \in \hat{\Pi}^R; E_i \geq \epsilon) = \sum_{\hat{\pi}^R \in \hat{\Pi}_\epsilon^R} P(\hat{\pi}^R) \leq \sum_{\hat{\pi}^R \in \hat{\Pi}_\epsilon^R} \exp(-E_i)$$
$$\leq |\hat{\Pi}^R|\exp(-\epsilon) \leq \left(M^{1/2}\frac{N}{K}+1\right)^{K(K+1)}\left(\frac{N(N+1)}{2}+1\right)^M \exp(-\epsilon).$$

Hence for all $\epsilon > 0$, we have over all K^N possible class assignments z,

$$P\left(\max_z\left\{\sum_m \frac{N(N+1)}{2} D\left(\frac{\sum_{q\leq l} n_{ql}\hat{\phi}_{ql}^{(m)}}{N(N+1)/2} \,\Big\|\, \frac{\sum_{q\leq l} n_{ql}\bar{\phi}_{ql}^{(m)}}{N(N+1)/2}\right)\right\} \geq \epsilon\right)$$
$$\leq \exp\left(N\log K + (K^2+K)\log\left(M^{1/2}\frac{N}{K}+1\right) + M\log\left(\frac{N(N+1)}{2}+1\right) - \epsilon\right),$$

and

$$P\left(\max_z\left\{\sum_{q\leq l} M n_{ql} D\left(\frac{1}{M}\sum_m \hat{\phi}_{ql}^{(m)} \,\Big\|\, \frac{1}{M}\sum_m \bar{\phi}_{ql}^{(m)}\right)\right\} \geq \epsilon\right)$$
$$\leq \exp\left(N\log K + (K^2+K)\log\left(M^{1/2}\frac{N}{K}+1\right) + M\log\left(\frac{N(N+1)}{2}+1\right) - \epsilon\right).$$

Proof of Theorem 2

First we note that X, as defined in Equation (3.9), is a sum of bounded independent random variables, because each element $X_{ij}^{(m)}$ in the sum is bounded by $C = 2\log(\sqrt{M}N)$ in absolute value.

So we can use a Bernstein type inequality for sums of bounded independent random variables ([7]) to obtain

$$P(|X - E(X)| > \epsilon) \leq \exp\left(-\frac{\epsilon^2}{2\sum_m\sum_{i<j} E[X_{ij}^{(m)2}] + \frac{2}{3}\epsilon C}\right)$$
$$\leq \exp\left(-\frac{\epsilon^2}{8L\log^2(\sqrt{M}N) + \frac{4}{3}\epsilon\log(\sqrt{M}N)}\right),$$

since $\sum_m\sum_{i<j} E[X_{ij}^{(m)2}] = \sum_m\sum_{i<j} P_{ij}^{(m)}\log^2(\bar{\pi}_{ql}^{(m)}/(1-\bar{\pi}_{ql}^{(m)})) < 4L\log^2(\sqrt{M}N)$. Combining this inequality with the result in Theorem 1, we have over all possible K^N class assignments z,

$$\max_z P(|l(A;z) - \bar{l}_P(z)| > 2\epsilon L)$$
$$\leq \max_z\left(P\left(\sum_{q\leq l} n_{ql}\sum_m D(\hat{\pi}_{ql}^{(m)} \mid\mid \bar{\pi}_{ql}^{(m)}) > \epsilon L\right) + P\left(|X - E(X)| > \epsilon L\right)\right)$$
$$\leq \exp\left(N\log K + M(K^2+K)\log\left(\frac{N}{K}+1\right) - \epsilon L\right)$$
$$+ \exp\left(N\log K - \frac{\epsilon^2 L}{8\log^2(\sqrt{M}N) + \frac{4}{3}\epsilon\log(\sqrt{M}N)}\right),$$

which goes to zero asymptotically as N grows under the growth conditions mentioned on K and L. So we have

$$\max_z |l(A;z) - \bar{l}_P(z)| = o_P(L).$$

Proof of Theorem 3

The proof for the RMLSBM will be a slight modification of the earlier proof for MLSBM. As before we need to bound the two terms in the decomposition of the difference between maximized likelihood and its expected value defined in Equation (3.16). For that we write the first part in the right hand side of (3.16), which we call E_3 here for brevity, in terms of the quantities we have already bounded in Theorem 1. We begin by noticing that, since the Kullback-Liebler divergence $D(a||b)$ is convex, we can use a reverse of Jensen's inequality ([29], [4]) to write

$$\sum_{q\leq l} n_{ql} D\left(\hat{\phi}_{ql}^{(m)} \mid\mid \bar{\phi}_{ql}^{(m)}\right) \leq \frac{N(N+1)}{2} D\left(\frac{\sum_{q\leq l} n_{ql}\hat{\phi}_{ql}^{(m)}}{N(N+1)/2} \,\Big|\Big|\, \frac{\sum_{q\leq l} n_{ql}\bar{\phi}_{ql}^{(m)}}{N(N+1)/2}\right) + \log(MN^2),$$

and

$$\sum_m n_{ql} D\left(\hat{\phi}_{ql}^{(m)} \mid\mid \bar{\phi}_{ql}^{(m)}\right) \leq M n_{ql} D\left(\frac{1}{M}\sum_m \hat{\phi}_{ql}^{(m)} \,\Big|\Big|\, \frac{1}{M}\sum_m \bar{\phi}_{ql}^{(m)}\right) + \log(MN^2).$$

To derive the inequality, we used $-\log(\hat{\phi}_{ql}^{(m)}/\bar{\phi}_{ql}^{(m)})$ as our convex function of $\hat{\phi}_{ql}^{(m)}/\bar{\phi}_{ql}^{(m)}$ on the interval $[1/(MN^2), 1-1/(MN^2)]$ to obtain a reverse of the "log-sum inequality". Summing the two inequalities over m and q, l respectively, we have

$$E_3 \leq 2\sum_m \frac{N(N+1)}{2} D\left(\frac{\sum_{q\leq l} n_{ql}\hat{\phi}_{ql}^{(m)}}{N(N+1)/2} \,\Big|\Big|\, \frac{\sum_{q\leq l} n_{ql}\bar{\phi}_{ql}^{(m)}}{N(N+1)/2}\right) + o(M(\log(\sqrt{M}N))^{1+\delta}),$$

and

$$E_3 \leq 2\sum_{q\leq l} Mn_{ql} D\left(\frac{1}{M}\sum_m \hat{\phi}_{ql}^{(m)} \,\Big|\Big|\, \frac{1}{M}\sum_m \bar{\phi}_{ql}^{(m)}\right) + o(K^2(\log(\sqrt{M}N))^{1+\delta}).$$

Hence E_3 is bounded by the minimum of the above two upper bounds. Since the first part in the right hand side of the above two inequalities is bounded by the same quantity, we will take the inequality for which the second part is smaller. Under the conditions on the growth of L in the theorem, the minimum of the two second parts is $o(L)$. Consequently,

$$\begin{aligned}
&\max_z P(|l^R(A;z) - \bar{l}_P^R(z)| > 2\epsilon L) \\
&\leq \exp\left(N\log K + (K^2+K)\log\left(M^{1/2}\frac{N}{K}+1\right) + M\log\left(\frac{N(N+1)}{2}+1\right) - \epsilon L\right) \\
&+ \exp\left(N\log K - \frac{\epsilon^2 L}{8\log^2(\sqrt{M}N) + \frac{4}{3}\epsilon\log N}\right),
\end{aligned}$$

so under the growth conditions mentioned under different asymptotic settings,

$$\max_z |l^R(A;z) - \bar{l}_P^R(z)| = o_P(L).$$

Proof of Lemma 2

For large N, subtracting Equation (3.24) from Equation (3.23) we have

$$\begin{aligned}
&\bar{l}_P(\bar{z}) - \bar{l}_P^R(\bar{z}) \\
=&\sum_{q\leq l} n_{ql} \sum_m D(\bar{\pi}_{ql}^{(m)} || \bar{\phi}_{ql}^{(m)}) \\
\leq&|E_Q|\log(MN^2) + \left(\frac{MN(N+1)}{2} - |E_Q|\right) C_1 \frac{L'}{MN^2(\log M)^{1+\delta}(\log N)^{2+\delta}} \\
&\log\left(\frac{C_1 L'/(MN^2(\log M)^{1+\delta}(\log N)^{2+\delta})}{1/MN^2}\right) \\
=&o_P(L') + \frac{C_1 L'}{(\log M)^{1+\delta}(\log N)^{2+\delta}} \log\left(\frac{C_1 L'}{(\log M)^{1+\delta}(\log N)^{2+\delta}}\right) \\
=&o_P(L') + o_P(L')\log\left(\frac{C_1 L'}{(\log M)^{1+\delta}(\log N)^{2+\delta}}\right) \Big/ [(\log M)^{1+\delta}(\log N)^{1+\delta}] \\
=&o_P(L') + o_P(L')R \\
=&o_P(L),
\end{aligned}$$

where C_1 is a constant and $R = \log\left(\frac{C_1 L'}{(\log M)^{1+\delta}(\log N)^{2+\delta}}\right) \Big/ [(\log M \log N)^{1+\delta}]$. The inequality in step 2 comes from the upper bound on $D(p||q)$ which can be derived as follows. Without loss of generality, we can assume that $p > q$ and $D(p||q) \leq p\log\frac{p}{q} \leq p_{\max}\log\frac{p_{\max}}{q_{\min}}$. Next we replace $p_{\max}$ and $q_{\min}$ by the assumption on the lower and upper bounds of the restricted block model probabilities given in Equation (3.3).

Now to complete the proof, we only need to verify that under the five sets of conditions in Theorem 3, the term R in the right hand side of the above derivation is $o(1)$. Under the first two

sets of conditions, $L' = MN(\log N)^{3+\delta}$ and consequently $R = \frac{\log(MN\log N/(\log M)^{1+\delta})}{(\log M \log N)^{1+\delta}} = o(1)$. Under the third set of conditions, $L' = N(\log N)^{3+\delta}$ and hence $R = \frac{\log(N\log N/(\log M)^{1+\delta})}{(\log M \log N)^{1+\delta}} = o(1)$. Finally under the last two sets of conditions, if $L' = MN(\log N)^{1+\delta}$ then $R = \frac{\log(MN/(\log M)^{1+\delta})}{(\log M \log N)^{1+\delta}} = o(1)$, and if $L' = M(\log M)^{2+\delta}(\log N)^{1+\delta}$ then $R = \frac{\log(M(\log M)^{1+\delta})}{(\log M \log N)^{1+\delta}} = o(1)$.

Proof of Lemma 3

Note that $\bar{l}_P(\hat{z}^R) \geq \bar{l}_P^R(\hat{z}^R)$ since the maximum of the unrestricted likelihood $\bar{l}_P(z)$ is uniformly larger than or equal to the maximum of the restricted likelihood $\bar{l}_P^R(z)$ for all z. Moreover, $\bar{z}$ maximizes $\bar{l}_P(\cdot)$ and hence $\bar{l}_P(\bar{z}) - \bar{l}_P^R(\hat{z}^R) \geq 0$. Notice that $l^R(A, \hat{z}^R) - l^R(A, \bar{z})$ is positive since the observed restricted likelihood is maximized at $\hat{z}^R$. So we have

$$\begin{aligned}
\bar{l}_P(\bar{z}) - \bar{l}_P^R(\hat{z}^R) &\leq \bar{l}_P(\bar{z}) - \bar{l}_P^R(\hat{z}^R) + l^R(A, \hat{z}^R) - l^R(A, \bar{z}) \\
&\leq |\bar{l}_P(\bar{z}) - l^R(A, \bar{z})| + |\bar{l}_P^R(\hat{z}^R) - l^R(A, \hat{z}^R)| \\
&\leq |\bar{l}_P(\bar{z}) - \bar{l}_P^R(\bar{z})| + |\bar{l}_P^R(\bar{z}) - l^R(A, \bar{z})| + |\bar{l}_P^R(\hat{z}^R) - l^R(A; \hat{z}^R)| \\
&= o_P(L),
\end{aligned}$$

by Lemma 2 and Theorem 3.

Proof of Theorem 4

The proof requires the concepts of partition and refinement as laid out in [6]. We briefly review the concepts here and apply them to the MLE of MLSBM and its regularized version in RMLSBM. Let $[N]$ denote the set of integers $\{1, 2, \ldots, N\}$. Any multi-layer blockmodel induces a partition of the M upper triangular probability matrices. Formally we define a partition of $\{P_{ij}^{(m)}\}_{i<j}$ into U subsets $\{S_1, \ldots, S_U\}$ by the following mapping

$$\Theta : (i, j)_{i\in[N],\ j\in[N],\ i<j} \to [U].$$

Note that the partitions induced on all M probability matrices are the same, since the partition is a function only of the indices and not of the type of edges. There exists a bijection between the set $[U]$ and the upper triangular part of the parameter matrices of MLSBM, so we can write $\pi_{\Theta(i,j)} = \pi_{z_i z_j}$.

In MLSBM, for a general partition, we define $S_u = \{(i, j) : \ \Theta(i, j) = u,\ i < j\}$ and $\bar{\pi}_u = |S_u|^{-1} \sum_m \sum_{\Theta(i,j)=u, i<j} P_{ij}^{(m)}$, so that we can define the log likelihood under this partition as

$$\bar{l}_P^*(\Theta) = \sum_{m=1}^{M} \sum_{i<j} \{P_{ij}^{(m)} \log\ \bar{\pi}_{\Theta(i,j)}^{(m)} + (1 - P_{ij}^{(m)}) \log\ (1 - \bar{\pi}_{\Theta(i,j)}^{(m)})\}.$$

It is easy to see that $\bar{l}_P^*(\Theta^z) = \bar{l}_P(z)$, where Θ^z is the partition corresponding to block model assignment z. A refinement Θ' of partition Θ further subdivides the partitions in Θ into subgroups or sub-partitions so that $\Theta'(i_1, j_1)_{i_1<j_1} = \Theta'(i_2, j_2)_{i_2<j_2} \Rightarrow \Theta(i_1, j_1)_{i_1<j_1} = \Theta(i_2, j_2)_{i_2<j_2}$. From Lemma A2 of [6], it can be easily obtained

$$\bar{l}_P^*(\Theta) \leq \bar{l}_P^*(\Theta').$$

One such refinement is constructed in the following way ([6]). We consider a K class MLSBM with membership vector $\bar{z}$ and let Θ^z denote a partition of $\{P_{ij}^{(m)}\}_{i<j}$ for any z. Now, for a given membership class under z, partition the corresponding set of nodes into subclasses according to the true class assignment $\bar{z}$ of each node. Then remove one node from each of the two largest subclasses so obtained, and group them together as a pair; continue this pairing process until no more than one nonempty subclass remains. If pair (i,j) is chosen from the above procedure, then $z_i = z_j$ and $\bar{z}_i \neq \bar{z}_j$. Define C_1 as the number of (i,j) pairs selected by the above method. Since at least one of i or j is misclustered, we have $N_e(z)/2 \leq C_1 \leq N_e(z)$.

Next, for each C_1 pairs find all other distinct indices k for which condition (3.26) of the theorem is satisfied. Let C_2 denote the total number of distinct triples that can be formed in this manner. For each of the C_2 such triples (i,j,k), we remove P_{ik} and P_{jk} from their previous subset assignment under Θ^z and place them in a new distinct two element subset. This partition so created is a refinement of the original partition Θ^z, and we call this refined partition Θ'^z. The condition (3.26) of the theorem implies that for each pair of classes (q,l), there exists at least one class c that satisfies,

$$D\left(\bar{\pi}_{qc}^{(m)} \,\Big\|\, \frac{\bar{\pi}_{qc}^{(m)} + \bar{\pi}_{lc}^{(m)}}{2}\right) + D\left(\bar{\pi}_{lc}^{(m)} \,\Big\|\, \frac{\bar{\pi}_{qc}^{(m)} + \bar{\pi}_{lc}^{(m)}}{2}\right) \geq \frac{LK}{MN^2}. \tag{7.3}$$

Consequently for any of the C_1 pairs of nodes under the true partition, we obtain triples at least as large as the cardinality of the smallest class. Hence C_2 is at least as large as $C_1 s$, where s the size of the smallest class. Now as per assumption, $s = \Omega(N/K)$. Hence we can bound the difference in the likelihood:

$$\begin{aligned}\bar{l}_P(\bar{z}) - \bar{l}_P^*(\Theta'^z) &= \sum_m \sum_{i<j} D\left(P_{ij}^{(m)} \,\|\, \pi_{\Theta'^z(i,j)}^{(m)}\right) = C_2 M \Omega\left(\frac{LK}{MN^2}\right) \\ &= C_1 M \Omega\left(\frac{N}{K}\frac{LK}{MN^2}\right) = \frac{N_e(z)}{2}\Omega(L)\frac{MNKL}{KLMN^2} = \frac{N_e(z)}{N}\Omega(L).\end{aligned}$$

Since the above procedure is valid for any class assignment vector z, we can apply it for the maximum likelihood estimate $\hat{z}$ as well. Note that $\hat{z}$ induces partition $\Theta^{\hat{z}}$ of the probability matrices $\{P_{ij}^{(m)}\}_{i<j,\ m=\{1,\ldots,M\}}$ and its refinement $\Theta'^{\hat{z}}$ increases the likelihood, i.e., $\bar{l}_P^*(\Theta^{\hat{z}}) \leq \bar{l}_P^*(\Theta'^{\hat{z}})$. Also we have $\bar{l}_P^*(\Theta^{\hat{z}}) = \bar{l}_P(\hat{z})$. Consequently we have,

$$\bar{l}_P(\bar{z}) - \bar{l}_P(\hat{z}) \geq \bar{l}_P(\bar{z}) - \bar{l}_P^*(\Theta'^{\hat{z}}) = \frac{N_e(\hat{z})}{N}\Omega(L).$$

Combining this with the result from Equation (3.25), we have

$$N_e(\hat{z}) = o_P(N).$$

Proof of Theorem 5

To prove the corresponding result for RMLSBM, we define regularized partition Θ^R of the matrices of probabilities between nodes $P_{ij}^{(m)}$, computed according to the restricted model RMLSBM and its refinement Θ'^R in exactly the same way. We further define the corresponding restricted log likelihood associated with this partition Θ^R as $\bar{l}_P^{*R}(\Theta^R)$. For convenience we again resort to the transformation defined by Equation (3.1)

$$\bar{l}_P^{*R}(\Theta^R) = \sum_{m=1}^{M} \sum_{i<j} \{P_{ij}^{(m)} \log \bar{\phi}_{\Theta^R(i,j)}^{(m)} + (1 - P_{ij}^{(m)}) \log(1 - \bar{\phi}_{\Theta^R(i,j)}^{(m)})\}.$$

For any membership assignment z^R from the RMLSBM, let $\bar{l}_P^{*R}(\Theta_{z^R}^R)$ be the corresponding partition of $P_{ij}^{(m)}$. It follows from this definition that $\bar{l}_P^{*R}(\Theta_{z^R}^R) = \bar{l}_P^R(z^R)$. Hence we have

$$\bar{l_P}(\bar{z}) - \bar{l}_P^{*R}(\Theta_{z^R}^R) = \sum_m \sum_{i<j} D\left(P_{ij}^{(m)} \mid\mid \bar{\phi}^{(m)}_{\Theta'^R_{z^R}(i,j)}\right) = C_2 M\Omega(g) = C_1 M\Omega\left(\frac{N}{K}g\right)$$

$$= \frac{N_e(z^R)}{2}\Omega(L)\frac{MN}{KL}g = \frac{N_e(z^R)}{h}\Omega(L).$$

Now we specialize to $\hat{z}^R$. Since Θ'^R is a refinement of Θ^R, it increases the restricted likelihood, i.e., $\bar{l}_P^{*R}(\Theta'^R_{\hat{z}^R}) \geq \bar{l}_P^{*R}(\Theta_{\hat{z}^R}^R)$. Using this and the fact that $\bar{l}_P^{*R}(\Theta_{\hat{z}^R}^R) = \bar{l}_P^R(\hat{z}^R)$, we have

$$\bar{l}_P(\bar{z}) - \bar{l}_P^R(\hat{z}^R) \geq \bar{l}_P(\bar{z}) - \bar{l}_P^{*R}(\Theta'^R_{\hat{z}^R}) = \frac{N_e(\hat{z}^R)}{h}\Omega(L).$$

The left hand side is $o(L)$ by Lemma 3, and hence,

$$N_e(\hat{z}^R) = o_P(h).$$

APPENDIX B

Derivation of variational inference for RMLSBM

We derive the update rules for RMLSBM. Note that for the restricted model, the complete data log likelihood is given by

$$\begin{aligned} l(A,Z) &= l(A|Z) + l(Z) \\ &= \sum_i \sum_q Z_{iq}\alpha_q + \frac{1}{2}\sum_{i\neq j}\sum_{q,l}\sum_m Z_{iq}Z_{jl}\{A_{ij}^{(m)}(\hat{\pi}_{ql} + \hat{\beta}_m) \\ &\quad - \log(1+\exp(\hat{\pi}_{ql} + \hat{\beta}_m)\}. \end{aligned}$$

The likelihood of the observed data can be obtained by summing the complete data likelihood over all possible values of the unobserved missing class assignment labels Z. However, note that the number of all possible assignments grows exponentially as K^N, and the sum quickly becomes computationally intractable even for moderate N. Hence instead we use the EM algorithm for mixture models, where the unobserved class assignments are treated as missing values. However one needs to compute the conditional distribution of the missing values (class assignments here) given the observed data, i.e., $P(Z|A)$. Unfortunately, as argued by [8], $P(Z|A)$ is itself intractable, since the probability of the latent class assignments of a node depends not only on the observed edges connected to that node, but also on the connectivity pattern of the whole network.

The variational approximation concentrates the search for optimal class assignments to a smaller set by assuming that the class assignments follow a multinomial distribution with parameters known as variational parameters. It aims at maximizing an expression containing the log likelihood and the negative of the Kullback-Liebler (KL) divergence between the true probability distribution of $P(Z|A)$ and its variational approximation $R_A(\cdot)$. If the approximation to the distribution coincides with the distribution, then the KL divergence is zero and the variational approximation is the same as the regular EM. So the new objective function to be optimized as a lower bound of $l(A)$ is

$$J(R_A) = \log l(A) - KL[R_A(\cdot),\ P(\cdot|A)].$$

Algorithm 1: Variational EM algorithm for MLSBM

while *either convergence criterion on parameters not met or* $t < t_{max}$ **do**
  `// E-step: Compute variational estimates` $\tau = \{\tau_{iq}\}$
  while *either convergence criteria on* τ *are not met or* $s < s_{max}$ **do**
    for $i \leftarrow \{1, 2, \ldots, N\}$ **do**
      for $q \leftarrow \{1, 2, \ldots, K\}$ **do**
        $\hat{\tau}_{iq}^{(s+1)} = \exp[\hat{\alpha}_q^{(t)} \sum_{i<j} \sum_{l} \sum_{m} \hat{\tau}_{jl}^{(s)} \{A_{ij}^{(m)} \hat{\pi}_{qlm}^{(t)} + (1 - A_{ij}^{(m)})(1 - \hat{\pi}_{qlm}^{(t)})\}]$
        $s = s + 1$
      end
    end
  end
  $\hat{\tau}_{iq}^{(t+1)} = \hat{\tau}_{iq}^{(t+1)} / \sum_{q=1}^{K} \hat{\tau}_{iq}^{(t+1)}$
  `// M-step: Estimate the parameters`
  for $q \leftarrow 1$ *to* K **do**
    $\hat{\alpha}_q^{(t+1)} = \frac{1}{N} \sum_{i=1}^{N} \hat{\tau}_{iq}^{(t+1)}$
    for $m \leftarrow 1$ *to* M **do**
      for $l \leftarrow 1$ *to* K **do**
        $\hat{\pi}_{qlm}^{(t+1)} = \frac{\sum_{i<j} \hat{\tau}_{iq}^{(t+1)} \hat{\tau}_{jl}^{(t+1)} A_{ij}^{(m)}}{\sum_{i<j} \hat{\tau}_{iq}^{(t+1)} \hat{\tau}_{jl}^{(t+1)}}$
      end
    end
  end
  $t = t + 1$
end

Algorithm 2: Variational EM algorithm for RMLSBM

while *either convergence criteria on parameters are not met or* $t < t_{max}$ **do**
  `// E-Step: Compute variational estimates` $\tau = \{\tau_{iq}\}$
  while *either convergence criteria on* τ *are not met or* $s < s_{max}$ **do**
    for $i \leftarrow \{1, 2, \ldots, N\}$ **do**
      for $q \leftarrow \{1, 2, \ldots, K\}$ **do**
        $\hat{\tau}_{iq}^{(s+1)} = \exp[\hat{\alpha}_q^{(t)} \sum_{i<j} \sum_{l} \sum_{m} \hat{\tau}_{jl}^{(s)} \{A_{ij}^{(m)} (\hat{\pi}_{ql}^{(t)} + \hat{\beta}_m^{(t)}) - \log(1 + \exp(\hat{\pi}_{ql}^{(t)} + \hat{\beta}_m^{(t)}))\}]$
        $s = s + 1$
      end
    end
  end
  `// Normalize the variational estimates so that they sum to 1 for each` i
  $\hat{\tau}_{iq}^{(t+1)} = \hat{\tau}_{iq}^{(t+1)} / \sum_{q=1}^{K} \hat{\tau}_{iq}^{(t+1)}$
  `// M-step: Estimate the parameters`
  for $q \leftarrow 1$ *to* K **do**
    $\hat{\alpha}_q^{(t+1)} = \frac{1}{N} \sum_{i=1}^{N} \hat{\tau}_{iq}^{(t+1)}$
  end
  `// Use BFGS optimization method to find the parameters`
  $(\hat{\pi}^{(t+1)}, \hat{\beta}^{(t+1)}) = \arg\max_{\pi, \beta} J(\pi, \beta)$
  $t = t + 1$
end

Here we constraint R_A to have the following form of the product of multinomial densities

$$R_A(Z) = \prod_i \prod_q \tau_{iq}^{Z_{iq}}.$$

The variational distribution $R_A(Z)$ has the interpretation of being an approximation of $P(Z|A)$.

In the E step of the following variational EM algorithm, we compute the variational approximation estimates of the probabilities of class assignments for each node. Given the model parameters α, π, β, the variational parameters τ can be computed by minimizing the function

$$\begin{aligned} J(R_A) = \sum_i \sum_q \tau_{iq} \log(\alpha_q) + \frac{1}{2} \sum_{i \neq j} \sum_{q,l} \sum_m Z_{iq} Z_{jl} \{ A_{ij}^{(m)} (\hat{\pi}_{ql} + \hat{\beta}_m) \\ - \log(1 + \exp(\hat{\pi}_{ql} + \hat{\beta}_m)\} - \sum_i \sum_q \tau_{iq} \log(\tau_{iq}) \end{aligned} \tag{7.4}$$

with the constraint that $\sum_q \tau_{iq} = 1$ for all i. The solution for the $(t+1)$th EM step can be readily obtained as

$$\hat{\tau}_{iq}^{(t+1)} = \exp\left[\hat{\alpha}_q^{(t)} \sum_{i<j} \sum_l \sum_m \hat{\tau}_{jl}^{(t)} \{ A_{ij}^{(m)} (\hat{\pi}_{ql}^{(t)} + \hat{\beta}_m^{(t)}) \log(1 + \exp(\hat{\pi}_{ql}^{(t)} + \hat{\beta}_m^{(t)}))\} \right].$$

In the M step we estimate the parameters of the model by maximizing the approximate likelihood. Since we do not have a closed form solution for the parameters π and β, we use a gradient descent algorithm (BFGS optimization algorithm) to simultaneously optimize the objective function with respect to all the parameters. The gradients of the objective function with respect to π and β are

$$\frac{\partial}{\partial \beta_m^{(t)}} := \sum_{i \neq j} \sum_{q,l} \hat{\tau}_{iq}^{(t)} \hat{\tau}_{jl}^{(t)} \left(A_{ij}^{(m)} - \frac{\exp(\hat{\pi}_{ql}^{(t)} + \hat{\beta}_m^{(t)})}{1 + \exp(\hat{\pi}_{ql}^{(t)} + \hat{\beta}_m^{(t)})} \right), \tag{7.5}$$

$$\frac{\partial}{\partial \pi_{ql}^{(t)}} := \sum_{i \neq j} \sum_m \hat{\tau}_{iq}^{(t)} \hat{\tau}_{jl}^{(t)} \left(A_{ij}^{(m)} - \frac{\exp(\hat{\pi}_{ql}^{(t)} + \hat{\beta}_m^{(t)})}{1 + \exp(\hat{\pi}_{ql}^{(t)} + \hat{\beta}_m^{(t)})} \right). \tag{7.6}$$

The two algorithms corresponding to the two models are described in Algorithm 1 and Algorithm 2 respectively.

Bibliography

[1] Airoldi, E. M., Blei, D. M., Fienberg, S. E. and Xing, E. P. (2008). Mixed membership stochastic blockmodels. *J. Mach. Learn. Res.* **9** 1981–2014.

[2] Bickel, P. J. and Chen, A. (2009). A nonparametric view of network models and Newman–Girvan and other modularities. *Proceedings of the National Academy of Sciences* **106** 21068–21073.

[3] Bickel, P. J., Choi, D., Chang, X. and Zhang, H. (2013). Asymptotic normality of maximum likelihood and its variational approximation for stochastic blockmodels. *Ann. Statist* **41** 1922-1943.

[4] Budimir, I., Dragomir, S. S. and Pecaric, J. (2001). Further reverse results for Jensen's discrete inequality and applications in information theory. *J. Inequal. Pure Appl. Math.* **2** 5.

[5] Celisse, A., Daudin, J. J. and Pierre, L. (2012). Consistency of maximum-likelihood and variational estimators in the stochastic block model. *Electronic Journal of Statistics* **6** 1847-1899.

[6] Choi, D. S., Wolfe, P. J. and Airoldi, E. M. (2012). Stochastic blockmodels with a growing number of classes. *Biometrika* **99** 273-284.

[7] Chung, F. and Lu, L. (2006). *Complex graphs and Networks*. American mathematical society.

[8] Daudin, J. J., Picard, F. and Robin, S. (2008). A mixture model for random graphs. *Stat Comput* **18** 173-183.

[9] Dong, X., Frossard, P., Vandergheynst, P. and Nefedov, N. (2012). Clustering with multi-layer graphs: A spectral perspective. *IEEE Transactions on Signal Processing* **60** 5820–5831.

[10] Erdos, P. and Renyi, A. (1960). On the evolution of random graphs. *Publication of the Mathematical Institute of the Hungarian Academy of Sciences* **5** 17-61.

[11] Goldenberg, A., Zheng, A. X., Fienberg, S. E. and Airoldi, E. M. (2010). A survey of statistical network models. *Foundations and Trends in Machine Learning* **2** 129-233.

[12] Greene, D. and Cunningham, P. (2013). Producing a unified graph representation from multiple social network views. *ACM Web Science* **2** 129-233.

[13] Han, Q., Xu, K. S. and Airoldi, E. M. (2014). Consistent estimation of dynamic and multi-layer Networks. *arXiv preprint arXiv:1410.8597*.

[14] Handcock, M. S., Raftery, A. E. and Tantrum, J. M. (2007). Model-based clustering for social networks. *J. Roy. Statist. Soc. Ser. A* **170** 301-354.

[15] Hoff, P. D., Raftery, A. E. and Handcock, M. S. (2002). Latent space approaches to social network analysis. *J. Amer. Statist. Assoc.* **97** 1090-1098.

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

[17] Jenatton, R., Le Roux, N., Bordes, A. and Obozinski, G. (2012). A latent factor model for highly multi-relational data. *Advances in Neural Information Processing Systems* 3167–3175.

[18] Karrer, B. and Newman, M. E. J. (2011). Stochastic blockmodels and community structure in networks. *Phys. Rev. E.* **83** 016107.

[19] Kemp, C., Tenenbaum, J. B., Griffiths, T. L., Yamada, T. and Ueda, N. (2006). Learning systems of concepts with an infinite relational model. In *Proceedings of the 21st National Conference on Artificial Intelligence* **1** 381–388.

[20] Latouche, P., Birmele, E. and Ambroise, C. (2011). Overlapping stochastic block models with application to the French political blogosphere. *Ann. Appl. Stat.* **5** 309-336.

[21] Mucha, P. J., Richardson, T., Macon, K., Porter, M. A. and Onnela, J. P. (2010). Community structure in time-dependent, multiscale, and multiplex networks. *Science* **328** 876–878.

[22] Narayanan, M., Vetta, A., Schadt, E. and Zhu, J. (2010). Simultaneous clustering of multiple gene expression and physical interaction datasets. *PLoS. Comp. Bio.* **6** e1000742.

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

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

[25] Nowicki, K. and Snijders, T. (2001). Estimation and prediction for stochastic block structures. *J. Am. Stat. Assoc.* **96** 1077-1087.

[26] Papalexakis, E. E., Akoglu, L. and Ience, D. (2013). Do more views of a graph help? Community detection and clustering in multi-graphs. In *Proceedings of the 16th International Conference on Information Fusion* 899–905.

[27] Rocklin, M. and Pinar, A. (2011). Latent clustering on graphs with multiple edge types. In *Algorithms and Models for the Web Graph* 38–49. Springer.

[28] Rohe, K., Qin, T. and Fan, H. (2012). The highest dimensional stochastic blockmodel with a regularized estimator. *Statistica Sinica* **39** 1878-1915.

[29] Simic, S. (2009). On an upper bound for Jensen's inequality. *Journal of Inequalities in Pure and Applied Mathematics* **10** 60.

[30] Snijders, T. A. B. and Nowicki, K. (1997). Estimation and prediction for stochastic blockmodels for graphs with latent block structure. *Journal of Classificaion* **14** 75-100.

[31] Tang, W., Lu, Z. and Dhillon, I. S. (2009). Clustering with multiple graphs. In *Proceedings of the 9th IEEE International Conference on Data Mining* 1016–1021.

[32] Taskar, B., Segal, E. and Koller, D. (2001). Probabilistic classification and clustering in relational data. In *Proceedings of the 17th International Joint Conference on Artificial Intelligence* 870–876.

[33] Zhao, Y., Levina, E. and Zhu, J. (2012). Consistency of community detection in networks under degree-corrected stochastic block models. *Ann. Statist* **40** 2266-2292.