Journal of Machine Learning Research 1 (2014) 2239-2312 Submitted 7/13; Revised 11/13; Published 10/00

A Tensor Approach to Learning Mixed Membership Community Models

Animashree Anandkumar A.ANANDKUMAR@UCI.EDU
Department of Electrical Engineering & Computer Science,
University of California Irvine,
Irvine, CA 92697, USA

Rong Ge RONGGE@MICROSOFT.COM
Microsoft Research
One Memorial Drive,
Cambridge MA 02142, USA

Daniel Hsu DJHSU@CS.COLUMBIA.EDU
Department of Computer Science,
Columbia University
116th Street and Broadway,
New York, NY 10027, USA

Sham M. Kakade SKAKADE@MICROSOFT.COM
Microsoft Research
One Memorial Drive,
Cambridge MA 02142, USA

Editor: Nathan Srebro

Abstract

Community detection is the task of detecting hidden communities from observed interactions. Guaranteed community detection has so far been mostly limited to models with non-overlapping communities such as the stochastic block model. In this paper, we remove this restriction, and provide guaranteed community detection for a family of probabilistic network models with overlapping communities, termed as the mixed membership Dirichlet model, first introduced by Airoldi et al. (2008). This model allows for nodes to have fractional memberships in multiple communities and assumes that the community memberships are drawn from a Dirichlet distribution. Moreover, it contains the stochastic block model as a special case. We propose a unified approach to learning these models via a tensor spectral decomposition method. Our estimator is based on low-order moment tensor of the observed network, consisting of 3-star counts. Our learning method is fast and is based on simple linear algebraic operations, e.g., singular value decomposition and tensor power iterations. We provide guaranteed recovery of community memberships and model parameters and present a careful finite sample analysis of our learning method. As an important special case, our results match the best known scaling requirements for the (homogeneous) stochastic block model.

Keywords: community detection, spectral methods, tensor methods, moment-based estimation, mixed membership models

1. Introduction

Studying communities forms an integral part of social network analysis. A community generally refers to a group of individuals with shared interests (e.g., music, sports), or relationships (e.g., friends, co-workers). Community formation in social networks has been studied by many sociologists, (e.g., Moreno, 1934; Lazarsfeld et al., 1954; McPherson et al., 2001; Currarini et al., 2009), starting with the seminal work of Moreno (1934). They posit various factors such as *homophily*[1] among the individuals to be responsible for community formation. Various probabilistic and non-probabilistic network models attempt to explain community formation. In addition, they also attempt to quantify interactions and the extent of overlap between different communities, relative sizes among the communities, and various other network properties. Studying such community models are also of interest in other domains, e.g., in biological networks.

While there exists a vast literature on community models, learning these models is typically challenging, and various heuristics such as Markov Chain Monte Carlo (MCMC) or variational expectation maximization (EM) are employed in practice. Such heuristics tend to scale poorly for large networks. On the other hand, community models with guaranteed learning methods tend to be restrictive. A popular class of probabilistic models, termed as *stochastic blockmodels*, have been widely studied and enjoy strong theoretical learning guarantees, (e.g., White et al., 1976; Holland et al., 1983; Fienberg et al., 1985; Wang and Wong, 1987; Snijders and Nowicki, 1997; McSherry, 2001). On the other hand, they posit that an individual belongs to a single community, which does not hold in most real settings (Palla et al., 2005).

In this paper, we consider a class of mixed membership community models, originally introduced by Airoldi et al. (2008), and recently employed by Xing et al. (2010) and Gopalan et al. (2012). The model has been shown to be effective in many real-world settings, but so far, no learning approach exists with provable guarantees. In this paper, we provide a novel learning approach for learning these mixed membership models and prove that these methods succeed under a set of sufficient conditions.

The mixed membership community model of Airoldi et al. (2008) has a number of attractive properties. It retains many of the convenient properties of the stochastic block model. For instance, conditional independence of the edges is assumed, given the community memberships of the nodes in the network. At the same time, it allows for communities to overlap, and for every individual to be fractionally involved in different communities. It includes the stochastic block model as a special case (corresponding to zero overlap among the different communities). This enables us to compare our learning guarantees with existing works for stochastic block models and also study how the extent of overlap among different communities affects the learning performance.

1.1 Summary of Results

We now summarize the main contributions of this paper. We propose a novel approach for learning mixed membership community models of Airoldi et al. (2008). Our approach is a method of moments estimator and incorporates tensor spectral decomposition. We provide

1. The term *homophily* refers to the tendency that individuals belonging to the same community tend to connect more than individuals in different communities.

guarantees for our approach under a set of sufficient conditions. Finally, we compare our results to existing ones for the special case of the stochastic block model, where nodes belong to a single community.

1.1.1 Learning Mixed Membership Models

We present a tensor-based approach for learning the mixed membership stochastic block model (MMSB) proposed by Airoldi et al. (2008). In the MMSB model, the community membership vectors are drawn from the Dirichlet distribution, denoted by $\mathrm{Dir}(\alpha)$, where α is known the Dirichlet concentration vector. Employing the Dirichlet distribution results in sparse community memberships in certain regimes of α, which is realistic. The extent of overlap between different communities under the MMSB model is controlled (roughly) via a single scalar parameter, $\alpha_0 := \sum_i \alpha_i$, where $\alpha := [\alpha_i]$ is the Dirichlet concentration vector. When $\alpha_0 \to 0$, the mixed membership model degenerates to a stochastic block model and we have non-overlapping communities. When $\alpha_i < 1$ (and hence, $\alpha_0 < k$), the generated vectors tend to be sparse and we focus on this regime in this paper.

We propose a unified tensor-based learning method for the MMSB model and establish recovery guarantees under a set of sufficient conditions. These conditions are in in terms of the network size n, the number of communities k, extent of community overlaps (through α_0), and the average edge connectivity across various communities. Below, we present an overview of our guarantees for the special case of equal sized communities (each of size n/k) and homogeneous community connectivity: let p be the probability for any intra-community edge to occur, and q be the probability for any inter-community edge. Let Π be the community membership matrix, where $\Pi^{(i)}$ denotes the i^{th} row, which is the vector of membership weights of the nodes for the i^{th} community. Let P be the community connectivity matrix such that $P(i,i) = p$ and $P(i,j) = q$ for $i \neq j$.

Theorem 1 (Main Result) *For an MMSB model with network size n, number of communities k, connectivity parameters p, q and community overlap parameter α_0, when*[2]

$$n = \tilde{\Omega}(k^2(\alpha_0+1)^2), \qquad \frac{p-q}{\sqrt{p}} = \tilde{\Omega}\left(\frac{(\alpha_0+1)k}{n^{1/2}}\right), \tag{1}$$

our estimated community membership matrix $\hat{\Pi}$ and the edge connectivity matrix $\hat{P}$ satisfy with high probability (w.h.p.)

$$\frac{\varepsilon_{\pi,\ell_1}}{n} := \frac{1}{n}\max_{i\in[n]} \|\hat{\Pi}^i - \Pi^i\|_1 = \tilde{O}\left(\frac{(\alpha_0+1)^{3/2}\sqrt{p}}{(p-q)\sqrt{n}}\right) \tag{2}$$

$$\varepsilon_P := \max_{i,j\in[k]} |\hat{P}_{i,j} - P_{i,j}| = \tilde{O}\left(\frac{(\alpha_0+1)^{3/2}k\sqrt{p}}{\sqrt{n}}\right). \tag{3}$$

Further, our support estimates $\hat{S}$ satisfy w.h.p.,

$$\Pi(i,j) \geq \xi \Rightarrow \hat{S}(i,j) = 1 \quad \text{and} \quad \Pi(i,j) \leq \frac{\xi}{2} \Rightarrow \hat{S}(i,j) = 0, \quad \forall i \in [k], j \in [n], \tag{4}$$

where Π is the true community membership matrix and the threshold is chosen as $\xi = \Omega(\epsilon_P)$.

2. The notation $\tilde{\Omega}(\cdot), \tilde{O}(\cdot)$ denotes $\Omega(\cdot), O(\cdot)$ up to poly-log factors.

Remark: *Note that the scaling condition in* (1) *ensures that* $\frac{\varepsilon_{\pi,\ell_1}}{n}$ *in* (2) *is decaying, since we assume that* $\alpha_0 < k$ *(sparse regime). However, if we want the estimation error* ε_P *in* (3) *to decay, we require a slightly stronger condition with respect to* α_0 *that*

$$n = \tilde{\Omega}(k^2(\alpha_0+1)^3), \qquad \frac{p-q}{\sqrt{p}} = \tilde{\Omega}\left(\frac{(\alpha_0+1)^{1.5}k}{n^{1/2}}\right).$$

The complete details are in Section 4. We first provide some intuitions behind the sufficient conditions in (1). We require the network size n to be large enough compared to the number of communities k, and for the separation $p-q$ to be large enough, so that the learning method can distinguish the different communities. This is natural since a zero separation ($p=q$) implies that the communities are indistinguishable. Moreover, we see that the scaling requirements become more stringent as α_0 increases. This is intuitive since it is harder to learn communities with more overlap, and we quantify this scaling. For the Dirichlet distribution, it can be shown that the number of "significant" entries is roughly $O(\alpha_0)$ with high probability, and in many settings of practical interest, nodes may have significant memberships in only a few communities, and thus, α_0 is a constant (or growing slowly) in many instances.

In addition, we quantify the error bounds for estimating various parameters of the mixed membership model in (2) and (3). These errors decay under the sufficient conditions in (1). Lastly, we establish zero-error guarantees for support recovery in (4): our learning method correctly identifies (w.h.p) all the significant memberships of a node and also identifies the set of communities where a node does not have a strong presence, and we quantify the threshold ξ in Theorem 1. Further, we present the results for a general (non-homogeneous) MMSB model in Section 4.2.

1.1.2 Identifiability Result for the MMSB Model

A byproduct of our analysis yields novel identifiability results for the MMSB model based on low order graph moments. We establish that the MMSB model is identifiable, given access to third order moments in the form of counts of 3-star subgraphs, i.e., a star subgraph consisting of three leaves, for each triplet of leaves, when the community connectivity matrix P is full rank. Our learning approach involves decomposition of this third order tensor. Previous identifiability results required access to high order moments and were limited to the stochastic block model setting; see Section 1.3 for details.

1.1.3 Implications on Learning Stochastic Block Models

Our results have implications for learning stochastic block models, which is a special case of the MMSB model with $\alpha_0 \to 0$. In this case, the sufficient conditions in (1) reduce to

$$n = \tilde{\Omega}(k^2), \qquad \frac{p-q}{\sqrt{p}} = \tilde{\Omega}\left(\frac{k}{n^{1/2}}\right), \tag{5}$$

The scaling requirements in (5) match with the best known bounds[3] (up to poly-log factors) for learning uniform stochastic block models and were previously achieved by Chen et al. (2012) via convex optimization involving semi-definite programming (SDP). In contrast, we propose an iterative non-convex approach involving tensor power iterations and linear algebraic techniques, and obtain similar guarantees. For a detailed comparison of learning guarantees under various methods for learning (homogeneous) stochastic block models, see Chen et al. (2012).

Thus, we establish learning guarantees explicitly in terms of the extent of overlap among the different communities for general MMSB models. Many real-world networks involve sparse community memberships and the total number of communities is typically much larger than the extent of membership of a single individual, e.g., hobbies/interests of a person, university/company networks that a person belongs to, the set of transcription factors regulating a gene, and so on. Thus, we see that in this regime of practical interest, where $\alpha_0 = \Theta(1)$, the scaling requirements in (1) match those for the stochastic block model in (5) (up to polylog factors) without any degradation in learning performance. Thus, we establish that learning community models with sparse community memberships is akin to learning stochastic block models and we present a unified approach and analysis for learning these models.

To the best of our knowledge, this work is the first to establish polynomial time learning guarantees for probabilistic network models with overlapping communities and we provide a fast and an iterative learning approach through linear algebraic techniques and tensor power iterations. While the results of this paper are mostly limited to a theoretical analysis of the tensor method for learning overlapping communities, we note recent results which show that this method (with improvements and modifications) is very accurate in practice on real datasets from social networks, and is scalable to graphs with millions of nodes (Huang et al., 2013).

1.2 Overview of Techniques

We now describe the main techniques employed in our learning approach and in establishing the recovery guarantees.

1.2.1 Method of Moments and Subgraph Counts

We propose an efficient learning algorithm based on low order moments, viz., counts of small subgraphs. Specifically, we employ a third-order tensor which counts the number of 3-stars in the observed network. A 3-star is a star graph with three leaves (see Figure 1) and we count the occurrences of such 3-stars across different partitions. We establish that (an adjusted) 3-star count tensor has a simple relationship with the model parameters, when the network is drawn from a mixed membership model. We propose a multi-linear transformation using edge-count matrices (also termed as the process of whitening), which reduces the problem of learning mixed membership models to the *canonical polyadic (CP) decomposition* of an orthogonal symmetric tensor, for which tractable decomposition exists,

3. There are many methods which achieve the best known scaling for n in (5), but have worse scaling for the separation $p - q$. This includes variants of the spectral clustering method, (e.g., Chaudhuri et al., 2012). See Chen et al. (2012) for a detailed comparison.

as described below. Note that the decomposition of a general tensor into its rank-one components is referred to as its CP decomposition (Kolda and Bader, 2009) and is in general NP-hard (Hillar and Lim, 2013). However, the decomposition is tractable in the special case of an orthogonal symmetric tensor considered here.

1.2.2 Tensor Spectral Decomposition via Power Iterations

Our tensor decomposition method is based on the popular power iterations (see Anandkumar et al., 2012a, e.g.,). It is a simple iterative method to compute the stable eigen-pairs of a tensor. In this paper, we propose various modifications to the basic power method to strengthen the recovery guarantees under perturbations. For instance, we introduce adaptive deflation techniques (which involves subtracting out the eigen-pairs previously estimated). Moreover, we initialize the tensor power method with (whitened) neighborhood vectors from the observed network, as opposed to random initialization. In the regime, where the community overlaps are small, this leads to an improved performance. Additionally, we incorporate thresholding as a post-processing operation, which again, leads to improved guarantees for sparse community memberships, i.e., when the overlap among different communities is small. We theoretically establish that all these modifications lead to improvement in performance guarantees and we discuss comparisons with the basic power method in Section 4.4.

1.2.3 Sample Analysis

We establish that our learning approach correctly recovers the model parameters and the community memberships of all nodes under exact moments. We then carry out a careful analysis of the empirical graph moments, computed using the network observations. We establish tensor concentration bounds and also control the perturbation of the various quantities used by our learning algorithm via matrix Bernstein's inequality (Tropp, 2012, thm. 1.4) and other inequalities. We impose the scaling requirements in (1) for various concentration bounds to hold.

1.3 Related Work

There is extensive work on modeling communities and various algorithms and heuristics for discovering them. We mostly limit our focus to works with theoretical guarantees.

1.3.1 Method of Moments

The method of moments approach dates back to Pearson (1894) and has been applied for learning various community models. Here, the moments correspond to counts of various subgraphs in the network. They typically consist of aggregate quantities, e.g., number of star subgraphs, triangles etc. in the network. For instance, Bickel et al. (2011) analyze the moments of a stochastic block model and establish that the subgraph counts of certain structures, termed as "wheels" (a family of trees), are sufficient for identifiability under some natural non-degeneracy conditions. In contrast, we establish that moments up to third order (corresponding to edge and 3-star counts) are sufficient for identifiability of the stochastic block model, and also more generally, for the mixed membership Dirichlet model. We

employ subgraph count tensors, corresponding to the number of subgraphs (such as stars) over a set of labeled vertices, while the work of Bickel et al. (2011) considers only aggregate (i.e., scalar) counts. Considering tensor moments allows us to use simple subgraphs (edges and 3 stars) corresponding to low order moments, rather than more complicated graphs (e.g., wheels considered by Bickel et al. (2011)) with larger number of nodes, for learning the community model.

The method of moments is also relevant for the family of random graph models termed as *exponential random graph models* (Holland and Leinhardt, 1981; Frank and Strauss, 1986). Subgraph counts of fixed graphs such as stars and triangles serve as sufficient statistics for these models. However, parameter estimation given the subgraph counts is in general NP-hard, due to the normalization constant in the likelihood (the partition function) and the model suffers from degeneracy issues; see Rinaldo et al. (2009); Chatterjee and Diaconis (2011) for detailed discussion. In contrast, we establish in this paper that the mixed membership model is amenable to simple estimation methods through linear algebraic operations and tensor power iterations using subgraph counts of 3-stars.

1.3.2 Stochastic Block Models

Many algorithms provide learning guarantees for stochastic block models. For a detailed comparison of these methods, see the recent work by Chen et al. (2012). A popular method is based on spectral clustering (McSherry, 2001), where community memberships are inferred through projection onto the spectrum of the Laplacian matrix (or its variants). This method is fast and easy to implement (via singular value decomposition). There are many variants of this method, e.g., the work of Chaudhuri et al. (2012) employs normalized Laplacian matrix to handle degree heterogeneities. In contrast, the work of Chen et al. (2012) uses convex optimization techniques via semi-definite programming learning block models. For a detailed comparison of learning guarantees under various methods for learning stochastic block models, see Chen et al. (2012).

1.3.3 Non-probabilistic Approaches

The classical approach to community detection tries to directly exploit the properties of the graph to define communities, without assuming a probabilistic model. Girvan and Newman (2002) use betweenness to remove edges until only communities are left. However, Bickel and Chen (2009) show that these algorithms are (asymptotically) biased and that using modularity scores can lead to the discovery of an incorrect community structure, even for large graphs. Jalali et al. (2011) define community structure as the structure that satisfies the maximum number of edge constraints (whether two individuals like/dislike each other). However, these models assume that every individual belongs to a single community.

Recently, some non-probabilistic approaches have been introduced with overlapping community models by Arora et al. (2012) and Balcan et al. (2012). The analysis of Arora et al. (2012) is mostly limited to dense graphs (i.e., $\Theta(n^2)$ edges for a n node graph), while our analysis provides learning guarantees for much sparser graphs (as seen by the scaling requirements in (1)). Moreover, the running time of the method of Arora et al. (2012) is *quasipolynomial time* (i.e., $O(n^{\log n})$) for the general case, and is based on a combinatorial learning approach. In contrast, our learning approach is based on simple linear algebraic

techniques and the running time is a low-order polynomial (roughly it is $O(n^2k)$ for a n node network with k communities under a serial computation model and $O(n+k^3)$ under a parallel computation model). The work of Balcan et al. (2012) assumes endogenously formed communities, by constraining the fraction of edges within a community compared to the outside. They provide a polynomial time algorithm for finding all such "self-determined" communities and the running time is $n^{O(\log 1/\alpha)/\alpha}$, where α is the fraction of edges within a self-determined community, and this bound is improved to linear time when $\alpha > 1/2$. On the other hand, the running time of our algorithm is mostly independent of the parameters of the assumed model, (and is roughly $O(n^2k)$). Moreover, both these works are limited to homophilic models, where there are more edges within each community, than between any two different communities. However, our learning approach is not limited to this setting and also does not assume homogeneity in edge connectivity across different communities (but instead it makes probabilistic assumptions on community formation). In addition, we provide improved guarantees for homophilic models by considering additional post-processing steps in our algorithm. Recently, Abraham et al. (2012) provide an algorithm for approximating the parameters of an Euclidean log-linear model in polynomial time. However, there setting is considerably different than the one in this paper.

1.3.4 Inhomogeneous Random Graphs, Graph Limits and Weak Regularity Lemma

Inhomogeneous random graphs have been analyzed in a variety of settings (e.g., Bollobás et al., 2007; Lovász, 2009) and are generalizations of the stochastic block model. Here, the probability of an edge between any two nodes is characterized by a general function (rather than by a $k \times k$ matrix as in the stochastic block model with k blocks). Note that the mixed membership model considered in this work is a special instance of this general framework. These models arise as the limits of convergent (dense) graph sequences and for this reason, the functions are also termed as "graphons" or graph limits (Lovász, 2009). A deep result in this context is the regularity lemma and its variants. The weak regularity lemma proposed by Frieze and Kannan (1999), showed that any convergent dense graph can be approximated by a stochastic block model. Moreover, they propose an algorithm to learn such a block model based on the so-called d_2 distance. The d_2 distance between two nodes measures similarity with respect to their "two-hop" neighbors and the block model is obtained by thresholding the d_2 distances. However, the method is limited to learning block models and not overlapping communities.

1.3.5 Learning Latent Variable Models (Topic Models)

The community models considered in this paper are closely related to the probabilistic topic models (Blei, 2012), employed for text modeling and document categorization. Topic models posit the occurrence of words in a corpus of documents, through the presence of multiple latent topics in each document. Latent Dirichlet allocation (LDA) is perhaps the most popular topic model, where the topic mixtures are assumed to be drawn from the Dirichlet distribution. In each document, a topic mixture is drawn from the Dirichlet distribution, and the words are drawn in a conditional independent manner, given the topic mixture. The mixed membership community model considered in this paper can be interpreted as

a generalization of the LDA model, where a node in the community model can function both as a document and a word. For instance, in the directed community model, when the outgoing links of a node are considered, the node functions as a document, and its outgoing neighbors can be interpreted as the words occurring in that document. Similarly, when the incoming links of a node in the network are considered, the node can be interpreted as a word, and its incoming links, as documents containing that particular word. In particular, we establish that certain graph moments under the mixed membership model have similar structure as the observed word moments under the LDA model. This allows us to leverage the recent developments from Anandkumar et. al. (Anandkumar et al., 2012c,a,b) for learning topic models, based on the method of moments. These works establish guaranteed learning using second- and third-order observed moments through linear algebraic and tensor-based techniques. In particular, in this paper, we exploit the tensor power iteration method of Anandkumar et al. (2012b), and propose additional improvements to obtain stronger recovery guarantees. Moreover, the sample analysis is quite different (and more challenging) in the community setting, compared to topic models analyzed in Anandkumar et al. (2012c,a,b). We clearly spell out the similarities and differences between the community model and other latent variable models in Section 4.4.

1.3.6 Lower Bounds

The work of Feldman et al. (2012) provides lower bounds on the complexity of statistical algorithms, and shows that for cliques of size $O(n^{1/2-\delta})$, for any constant $\delta > 0$, at least $n^{\Omega(\log \log n)}$ queries are needed to find the cliques. There are works relating the hardness of finding hidden cliques and the use of higher order moment tensors for this purpose. Frieze and Kannan (2008) relate the problem of finding a hidden clique to finding the top eigenvector of the third order tensor, corresponding to the maximum spectral norm. Charles and Vempala (2009) extend the result to arbitrary r^{th}-order tensors and the cliques have to be size $\Omega(n^{1/r})$ to enable recovery from r^{th}-order moment tensors in a n node network. However, this problem (finding the top eigenvector of a tensor) is known to be NP-hard in general (Hillar and Lim, 2013). Thus, tensors are useful for finding smaller hidden cliques in network (albeit by solving a computationally hard problem). In contrast, we consider tractable tensor decomposition through reduction to orthogonal tensors (under the scaling requirements of (1)), and our learning method is a fast and an iterative approach based on tensor power iterations and linear algebraic operations. Mossel et al. (2012) provide lower bounds on the separation $p - q$, the edge connectivity between intra-community and inter-community, for identifiability of communities in stochastic block models in the sparse regime (when $p, q \sim n^{-1}$), when the number of communities is a constant $k = O(1)$. Our method achieves the lower bounds on separation of edge connectivity up to poly-log factors.

1.3.7 Likelihood-based Approaches to Learning MMSB

Another class of approaches for learning MMSB models are based on optimizing the observed likelihood. Traditional approaches such as Gibbs sampling or expectation maximization (EM) can be too expensive apply in practice for MMSB models. Variational approaches which optimize the so-called evidence lower bound (Hoffman et al., 2012; Gopalan et al., 2012), which is a lower bound on the marginal likelihood of the observed data (typically by

applying a mean-field approximation), are efficient for practical implementation. Stochastic versions of the variational approach provide even further gains in efficiency and are state-of-art practical learning methods for MMSB models (Gopalan et al., 2012). However, these methods lack theoretical guarantees; since they optimize a bound on the likelihood, they are not guaranteed to recover the underlying communities consistently. A recent work (Celisse et al., 2012) establishes consistency of maximum likelihood and variational estimators for stochastic block models, which are special cases of the MMSB model. However, it is not known if the results extend to general MMSB models. Moreover, the framework of Celisse et al. (2012) assumes a fixed number of communities and growing network size, and provide only asymptotic consistency guarantees. Thus, they do not allow for high-dimensional settings, where the parameters of the learning problem also grow as the observed dimensionality grows. In contrast, in this paper, we allow for the number of communities to grow, and provide precise constraints on the scaling bounds for consistent estimation under finite samples. It is an open problem to obtain such bounds for maximum likelihood and variational estimators. On the practical side, a recent work deploying the tensor approach proposed in this paper by Huang et al. (2013) shows that the tensor approach is more than an order of magnitude faster in recovering the communities than the variational approach, is scalable to networks with millions of nodes, and also has better accuracy in recovering the communities.

2. Community Models and Graph Moments

In the first part of section, we describe the mixed membership community model based on Dirichlet priors for the community draws by the individuals. Then in Section 2.2, we define and analyze the graph moments for these models.

2.1 Community Membership Models

We first introduce the special case of the popular stochastic block model, where each node belongs to a single community.

2.1.1 Notation

We consider networks with n nodes and let $[n] := \{1, 2, \ldots, n\}$. Let G be the $\{0, 1\}$ adjacency[4] matrix for the random network and let $G_{A,B}$ be the submatrix of G corresponding to rows $A \subseteq [n]$ and columns $B \subseteq [n]$. We consider models with k underlying (hidden) communities. For node i, let $\pi_i \in \mathbb{R}^k$ denote its *community membership vector*, i.e., the vector is supported on the communities to which the node belongs. In the special case of the popular stochastic block model described below, π_i is a basis coordinate vector, while the more general mixed membership model relaxes this assumption and a node can be in multiple communities with fractional memberships. Define $\Pi := [\pi_1|\pi_2|\cdots|\pi_n] \in \mathbb{R}^{k\times n}$. and let $\Pi_A := [\pi_i : i \in A] \in \mathbb{R}^{k\times|A|}$ denote the set of column vectors restricted to $A \subseteq [n]$. For a matrix M, let $(M)_i$ and $(M)^i$ denote its i^{th} column and row respectively. For a matrix M with singular value decomposition (SVD) $M = UDV^\top$, let $(M)_{k-svd} := U\tilde{D}V^\top$ denote the k-rank SVD of M, where $\tilde{D}$ is limited to top-k singular values of M. Let $M^\dagger$ denote

4. Our analysis can easily be extended to weighted adjacency matrices with bounded entries.

the MoorePenrose pseudo-inverse of M. Let $\mathbb{I}(\cdot)$ be the indicator function. Let $\mathrm{Diag}(v)$ denote a diagonal matrix with diagonal entries given by a vector v. We use the term high probability to mean with probability $1 - n^{-c}$ for any constant $c > 0$.

2.1.2 Stochastic Block Model (special case)

In this model, each individual is independently assigned to a single community, chosen at random: each node i chooses community j independently with probability $\widehat{\alpha}_j$, for $i \in [n], j \in [k]$, and we assign $\pi_i = e_j$ in this case, where $e_j \in \{0,1\}^k$ is the j^{th} coordinate basis vector. Given the community assignments Π, every directed[5] edge in the network is independently drawn: if node u is in community i and node v is in community j (and $u \neq v$), then the probability of having the edge (u, v) in the network is $P_{i,j}$. Here, $P \in [0,1]^{k \times k}$ and we refer to it as the *community connectivity matrix*. This implies that given the community membership vectors π_u and π_v, the probability of an edge from u to v is $\pi_u^\top P \pi_v$ (since when $\pi_u = e_i$ and $\pi_v = e_j$, we have $\pi_u^\top P \pi_v = P_{i,j}$.). The stochastic model has been extensively studied and can be learnt efficiently through various methods, e.g., spectral clustering (McSherry, 2001), convex optimization (Chen et al., 2012). and so on. Many of these methods rely on conditional independence assumptions of the edges in the block model for guaranteed learning.

2.1.3 Mixed Membership Model

We now consider the extension of the stochastic block model which allows for an individual to belong to multiple communities and yet preserves some of the convenient independence assumptions of the block model. In this model, the community membership vector π_u at node u is a probability vector, i.e., $\sum_{i \in [k]} \pi_u(i) = 1$, for all $u \in [n]$. Given the community membership vectors, the generation of the edges is identical to the block model: given vectors π_u and π_v, the probability of an edge from u to v is $\pi_u^\top P \pi_v$, and the edges are independently drawn. This formulation allows for the nodes to be in multiple communities, and at the same time, preserves the conditional independence of the edges, given the community memberships of the nodes.

2.1.4 Dirichlet Prior for Community Membership

The only aspect left to be specified for the mixed membership model is the distribution from which the community membership vectors Π are drawn. We consider the popular setting of Airoldi et al. (2008), where the community vectors $\{\pi_u\}$ are i.i.d. draws from the Dirichlet distribution, denoted by $\mathrm{Dir}(\alpha)$, with parameter vector $\alpha \in \mathbb{R}^k_{>0}$. The probability density function of the Dirichlet distribution is given by

$$\mathbb{P}[\pi] = \frac{\prod_{i=1}^k \Gamma(\alpha_i)}{\Gamma(\alpha_0)} \prod_{i=1}^k \pi_i^{\alpha_i - 1}, \quad \pi \sim \mathrm{Dir}(\alpha), \alpha_0 := \sum_i \alpha_i, \tag{6}$$

where $\Gamma(\cdot)$ is the Gamma function and the ratio of the Gamma function serves as the normalization constant.

5. We limit our discussion to directed networks in this paper, but note that the results also hold for undirected community models, where P is a symmetric matrix, and an edge (u, v) is formed with probability $\pi_u^\top P \pi_v = \pi_v^\top P \pi_u$.

The Dirichlet distribution is widely employed for specifying priors in Bayesian statistics, e.g., latent Dirichlet allocation (Blei et al., 2003). The Dirichlet distribution is the conjugate prior of the multinomial distribution which makes it attractive for Bayesian inference.

Let $\widehat{\alpha}$ denote the normalized parameter vector α/α_0, where $\alpha_0 := \sum_i \alpha_i$. In particular, note that $\widehat{\alpha}$ is a probability vector: $\sum_i \widehat{\alpha}_i = 1$. Intuitively, $\widehat{\alpha}$ denotes the relative expected sizes of the communities (since $\mathbb{E}[n^{-1} \sum_{u \in [n]} \pi_u[i]] = \widehat{\alpha}_i$). Let $\widehat{\alpha}_{\max}$ be the largest entry in $\widehat{\alpha}$, and $\widehat{\alpha}_{\min}$ be the smallest entry. Our learning guarantees will depend on these parameters.

The stochastic block model is a limiting case of the mixed membership model when the Dirichlet parameter is $\alpha = \alpha_0 \cdot \widehat{\alpha}$, where the probability vector $\widehat{\alpha}$ is held fixed and $\alpha_0 \to 0$. In the other extreme when $\alpha_0 \to \infty$, the Dirichlet distribution becomes peaked around a single point, for instance, if $\alpha_i \equiv c$ and $c \to \infty$, the Dirichlet distribution is peaked at $k^{-1} \cdot \vec{1}$, where $\vec{1}$ is the all-ones vector. Thus, the parameter α_0 serves as a measure of the average sparsity of the Dirichlet draws or equivalently, of how concentrated the Dirichlet measure is along the different coordinates. This in effect, controls the extent of overlap among different communities.

2.1.5 Sparse Regime of Dirichlet Distribution

When the Dirichlet parameter vector satisfies[6] $\alpha_i < 1$, for all $i \in [k]$, the Dirichlet distribution $\mathrm{Dir}(\alpha)$ generates "sparse" vectors with high probability;[7]; see Telgarsky (2012) (and in the extreme case of the block model where $\alpha_0 \to 0$, it generates 1-sparse vectors). Many real-world settings involve sparse community membership and the total number of communities is typically much larger than the extent of membership of a single individual, e.g., hobbies/interests of a person, university/company networks that a person belongs to, the set of transcription factors regulating a gene, and so on. Our learning guarantees are limited to the sparse regime of the Dirichlet model.

2.2 Graph Moments under Mixed Membership Models

Our approach for learning a mixed membership community model relies on the form of the graph moments[8] under the mixed membership model. We now describe the specific graph moments used by our learning algorithm (based on 3-star and edge counts) and provide explicit forms for the moments, assuming draws from a mixed membership model.

2.2.1 Notation

Recall that G denotes the adjacency matrix and that $G_{X,A}$ denotes the submatrix corresponding to edges going from X to A. Recall that $P \in [0,1]^{k \times k}$ denotes the community connectivity matrix. Define

$$F := \Pi^\top P^\top = [\pi_1 | \pi_2 | \cdots | \pi_n]^\top P^\top. \tag{7}$$

6. The assumption that the Dirichlet distribution be in the sparse regime is not strictly needed. Our results can be extended to general Dirichlet distributions, but with worse scaling requirements on the network size n for guaranteed learning.
7. Roughly the number of entries in π exceeding a threshold τ is at most $O(\alpha_0 \log(1/\tau))$ with high probability, when $\pi \sim \mathrm{Dir}(\alpha)$.
8. We interchangeably use the term first order moments for edge counts and third order moments for 3-star counts.

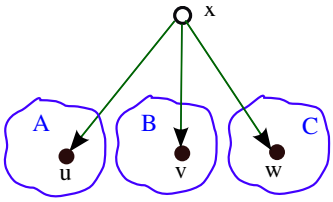


Figure 1: Our moment-based learning algorithm uses 3-star count tensor from set X to sets A, B, C (and the roles of the sets are interchanged to get various estimates). Specifically, T is a third order tensor, where $\mathrm{T}(u, v, w)$ is the normalized count of the 3-stars with u, v, w as leaves over all $x \in X$.

For a subset $A \subseteq [n]$ of individuals, let $F_A \in \mathbb{R}^{|A| \times k}$ denote the submatrix of F corresponding to nodes in A, i.e., $F_A := \Pi_A^\top P^\top$. We will subsequently show that F_A is linear map which takes any community vector π_i as input and outputs the corresponding neighborhood vector $G_{i,A}^\top$ in expectation.

Our learning algorithm uses moments up to the third-order, represented as a tensor. A third-order tensor T is a three-dimensional array whose (p, q, r)-th entry denoted by $T_{p,q,r}$. The symbol $\otimes$ denotes the standard Kronecker product: if u, v, w are three vectors, then

$$(u \otimes v \otimes w)_{p,q,r} := u_p \cdot v_q \cdot w_r. \tag{8}$$

A tensor of the form $u \otimes v \otimes w$ is referred to as a rank-one tensor. The decomposition of a general tensor into a sum of its rank-one components is referred to as *canonical polyadic (CP) decomposition* Kolda and Bader (2009). We will subsequently see that the graph moments can be expressed as a tensor and that the CP decomposition of the graph-moment tensor yields the model parameters and the community vectors under the mixed membership community model.

2.2.2 Graph Moments under Stochastic Block Model

We first analyze the graph moments in the special case of a stochastic block model (i.e., $\alpha_0 = \sum_i \alpha_i \to 0$ in the Dirichlet prior in (6)) and then extend it to general mixed membership model. We provide explicit expressions for the graph moments corresponding to edge counts and 3-star counts. We later establish in Section 3 that these moments are sufficient to learn the community memberships of the nodes and the model parameters of the block model.

2.2.3 3-star Counts

The primary quantity of interest is a third-order tensor which counts the number of 3-stars. A 3-star is a star graph with three leaves $\{a, b, c\}$ and we refer to the internal node x of the star as its "head", and denote the structure by $x \to \{a, b, c\}$ (see Figure 1). We partition the network into four[9] parts and consider 3-stars such that each node in the 3-star belongs to a different partition. This is necessary to obtain a simple form of the moments, based on the conditional independence assumptions of the block model, see Proposition 2. Specifically,

9. For sample complexity analysis, we require dividing the graph into more than four partitions to deal with statistical dependency issues, and we outline it in Section 3.

consider[10] a partition A, B, C, X of the network. We count the number of 3-stars from X to A, B, C and our quantity of interest is

$$\mathrm{T}_{X\to\{A,B,C\}} := \frac{1}{|X|}\sum_{i\in X}[G_{i,A}^\top \otimes G_{i,B}^\top \otimes G_{i,C}^\top], \tag{9}$$

where $\otimes$ is the Kronecker product, defined in (8) and $G_{i,A}$ is the row vector supported on the set of neighbors of i belonging to set A. $\mathrm{T} \in \mathbb{R}^{|A|\times|B|\times|C|}$ is a third order tensor, and an element of the tensor is given by

$$\mathrm{T}_{X\to\{A,B,C\}}(a,b,c) = \frac{1}{|X|}\sum_{x\in X} G(x,a)G(x,b)G(x,c), \quad \forall a\in A, b\in B, c\in C, \tag{10}$$

which is the normalized count of the number of 3-stars with leaves a, b, c such that its "head" is in set X.

We now relate the tensor $\mathrm{T}_{X\to\{A,B,C\}}$ to the parameters of the stochastic block model, viz., the community connectivity matrix P and the community probability vector $\widehat{\alpha}$, where $\widehat{\alpha}_i$ is the probability of choosing community i.

Proposition 2 (Moments in Stochastic Block Model) *Given partitions A, B, C, X, and $F := \Pi^\top P^\top$, where P is the community connectivity matrix and Π is the matrix of community membership vectors, we have*

$$\mathbb{E}[G_{X,A}^\top|\Pi_A,\Pi_X] = F_A\Pi_X, \tag{11}$$

$$\mathbb{E}[\mathrm{T}_{X\to\{A,B,C\}}\,|\Pi_A,\Pi_B,\Pi_C] = \sum_{i\in[k]}\widehat{\alpha}_i(F_A)_i\otimes(F_B)_i\otimes(F_C)_i, \tag{12}$$

where $\widehat{\alpha}_i$ is the probability for a node to select community i.

Remark: *In Equation* (11)*, we see that the edge generation occurs under a linear model, and more precisely, the matrix $F_A \in \mathbb{R}^{|A|\times k}$ is a linear map which takes a community vector $\pi_i \in \mathbb{R}^k$ to a neighborhood vector $G_{i,A}^\top \in \mathbb{R}^{|A|}$ in expectation.*

Remark: *(Identifiability under third order moments)) Note the form of the* 3*-star count tensor* T *in* (12)*. It provides a CP decomposition of* T *since each term in the summation, viz., $\widehat{\alpha}_i(F_A)_i\otimes(F_B)_i\otimes(F_C)_i$, is a rank one tensor. Thus, we can learn the matrices F_A, F_B, F_C and the vector $\widehat{\alpha}$ through CP decomposition of tensor* T*. Once these parameters are learnt, learning the communities is straight-forward under exact moments: by exploiting* (11)*, we find Π_X as*

$$\Pi_X = F_A^\dagger \cdot \mathbb{E}[G_{X,A}^\top|\Pi_A,\Pi_X].$$

Similarly, we can consider another tensor consisting of 3*-stars from A to X, B, C, and obtain matrices F_X, F_B and F_C through a CP decomposition, and so on. Once we obtain matrices F and Π for the entire set of nodes in this manner, we can obtain the community connectivity matrix P, since $F := \Pi^\top P^\top$. Thus, in principle, we are able to learn all the*

10. To establish our theoretical guarantees, we assume that the partitions A, B, C, X are randomly chosen and are of size $\Theta(n)$.

model parameters ($\widehat{\alpha}$ and P) and the community membership matrix Π under the stochastic block model, given exact moments. This establishes identifiability of the model given moments up to third order and forms a high-level approach for learning the communities. When only samples are available, we establish that the empirical versions are close to the exact moments considered above, and we modify the basic learning approach to obtain robust guarantees. See Section 3 for details.

Remark: *(Significance of conditional independence relationships) The main property exploited in proving the tensor form in* (12) *is the conditional-independence assumption under the stochastic block model: the realization of the edges in each 3-star, say in $x \to \{a, b, c\}$, is conditionally independent given the community membership vector π_x, when $x \neq a \neq b \neq c$. This is because the community membership vectors Π are assumed to be drawn independently at the different nodes and the edges are drawn independently given the community vectors.*

Considering 3-stars from X to A, B, C where X, A, B, C form a partition ensures that this conditional independence is satisfied for all the 3-stars in tensor T.

Proof: Recall that the probability of an edge from u to v given π_u, π_v is

$$\mathbb{E}[G_{u,v}|\pi_u, \pi_v] = \pi_u^\top P \pi_v = \pi_v^\top P^\top \pi_u = F_v \pi_u,$$

and $\mathbb{E}[G_{X,A}|\Pi_A, \Pi_X] = \Pi_X^\top P \Pi_A = \Pi_X^\top F_A^\top$ and thus (11) holds. For the tensor form, first consider an element of the tensor, with $a \in A, b \in B, c \in C$,

$$\mathbb{E}\left[\mathrm{T}_{X \to \{A,B,C\}}(a, b, c)|\pi_a, \pi_b, \pi_c, \pi_x\right] = \frac{1}{|X|} \sum_{x \in X} F_a \pi_x \cdot F_b \pi_x \cdot F_c \pi_x,$$

The equation follows from the conditional-independence assumption of the edges (assuming $a \neq b \neq c$). Now taking expectation over the nodes in X, we have

$$\begin{aligned}
\mathbb{E}\left[\mathrm{T}_{X \to \{A,B,C\}}(a, b, c)|\pi_a, \pi_b, \pi_c\right] &= \frac{1}{|X|} \sum_{x \in X} \mathbb{E}\left[F_a \pi_x \cdot F_b \pi_x \cdot F_c \pi_x | \pi_a, \pi_b, \pi_c\right] \\
&= \mathbb{E}\left[F_a \pi \cdot F_b \pi \cdot F_c \pi | \pi_a, \pi_b, \pi_c\right] \\
&= \sum_{j \in [k]} \widehat{\alpha}_j (F_a)_j \cdot (F_b)_j \cdot (F_c)_j,
\end{aligned}$$

where the last step follows from the fact that $\pi = e_j$ with probability $\widehat{\alpha}_j$ and the result holds when $x \neq a, b, c$. Recall that $(F_a)_j$ denotes the j^{th} column of F_a (since $F_a e_j = (F_a)_j$). Collecting all the elements of the tensor, we obtain the desired result. ■

2.2.4 Graph Moments under Mixed Membership Dirichlet Model

We now analyze the graph moments for the general mixed membership Dirichlet model. Instead of the raw moments (i.e., edge and 3-star counts), we consider modified moments to obtain similar expressions as in the case of the stochastic block model.

Let $\mu_{X \to A} \in \mathbb{R}^{|A|}$ denote a vector which gives the normalized count of edges from X to A:

$$\mu_{X \to A} := \frac{1}{|X|} \sum_{i \in X} [G_{i,A}^\top]. \tag{13}$$

We now define a modified adjacency matrix[11] $G_{X,A}^{\alpha_0}$ as

$$G_{X,A}^{\alpha_0} := \left(\sqrt{\alpha_0+1}G_{X,A} - (\sqrt{\alpha_0+1}-1)\vec{1}\mu_{X\to A}^\top\right). \tag{14}$$

In the special case of the stochastic block model ($\alpha_0 \to 0$), $G_{X,A}^{\alpha_0} = G_{X,A}$ is the submatrix of the adjacency matrix G. Similarly, we define modified third-order statistics,

$$\begin{aligned}\mathrm{T}_{X\to\{A,B,C\}}^{\alpha_0} &:= (\alpha_0+1)(\alpha_0+2)\,\mathrm{T}_{X\to\{A,B,C\}} + 2\,\alpha_0^2\,\mu_{X\to A}\otimes\mu_{X\to B}\otimes\mu_{X\to C} \\ &- \frac{\alpha_0(\alpha_0+1)}{|X|}\sum_{i\in X}\left[G_{i,A}^\top\otimes G_{i,B}^\top\otimes\mu_{X\to C} + G_{i,A}^\top\otimes\mu_{X\to B}\otimes G_{i,C}^\top + \mu_{X\to A}\otimes G_{i,B}^\top\otimes G_{i,C}^\top\right],\end{aligned} \tag{15}$$

and it reduces to (a scaled version of) the 3-star count $\mathrm{T}_{X\to\{A,B,C\}}$ defined in (9) for the stochastic block model ($\alpha_0 \to 0$). The modified adjacency matrix and the 3-star count tensor can be viewed as a form of "centering" of the raw moments which simplifies the expressions for the moments. The following relationships hold between the modified graph moments $G_{X,A}^{\alpha_0}$, T^{α_0} and the model parameters P and $\widehat{\alpha}$ of the mixed membership model.

Proposition 3 (Moments in Mixed Membership Model) *Given partitions A, B, C, X and $G_{X,A}^{\alpha_0}$ and T^{α_0}, as in (14) and (15), normalized Dirichlet concentration vector $\widehat{\alpha}$, and $F := \Pi^\top P^\top$, where P is the community connectivity matrix and Π is the matrix of community memberships, we have*

$$\mathbb{E}[(G_{X,A}^{\alpha_0})^\top|\Pi_A,\Pi_X] = F_A\,\mathrm{Diag}(\widehat{\alpha}^{1/2})\Psi_X, \tag{16}$$

$$\mathbb{E}[\mathrm{T}_{X\to\{A,B,C\}}^{\alpha_0}\,|\Pi_A,\Pi_B,\Pi_C] = \sum_{i=1}^k \widehat{\alpha}_i(F_A)_i\otimes(F_B)_i\otimes(F_C)_i, \tag{17}$$

where $(F_A)_i$ corresponds to i^{th} column of F_A and Ψ_X relates to the community membership matrix Π_X as

$$\Psi_X := \mathrm{Diag}(\widehat{\alpha}^{-1/2})\left(\sqrt{\alpha_0+1}\Pi_X - (\sqrt{\alpha_0+1}-1)\left(\frac{1}{|X|}\sum_{i\in X}\pi_i\right)\vec{1}^\top\right).$$

Moreover, we have that

$$|X|^{-1}\mathbb{E}_{\Pi_X}[\Psi_X\Psi_X^\top] = I. \tag{18}$$

Remark: *The 3-star count tensor T^{α_0} is carefully chosen so that the CP decomposition of the tensor directly yields the matrices F_A, F_B, F_C and $\widehat{\alpha}_i$, as in the case of the stochastic block model. Similarly, the modified adjacency matrix $(G_{X,A}^{\alpha_0})^\top$ is carefully chosen to eliminate second-order correlation in the Dirichlet distribution and we have that $|X|^{-1}\mathbb{E}_{\Pi_X}[\Psi\Psi^\top] = I$ is the identity matrix. These properties will be exploited by our learning algorithm in Section 3.*

11. To compute the modified moments G^{α_0}, and T^{α_0}, we need to know the value of the scalar $\alpha_0 := \sum_i \alpha_i$, which is the concentration parameter of the Dirichlet distribution and is a measure of the extent of overlap between the communities. We assume its knowledge here.

Remark: *Recall that α_0 quantifies the extent of overlap among the communities. The computation of the modified moment T^{α_0} requires the knowledge of α_0, which is assumed to be known. Since this is a scalar quantity, in practice, we can easily tune this parameter via cross validation.*

Proof: The proof is on lines of Proposition 2 for stochastic block models ($\alpha_0 \to 0$) but more involved due to the form of Dirichlet moments. Recall $\mathbb{E}[G_{i,A}^\top|\pi_i, \Pi_A] = F_A\pi_i$ for a mixed membership model, and $\mu_{X\to A} := \frac{1}{|X|}\sum_{i\in X} G_{i,A}^\top$, therefore $\mathbb{E}[\mu_{X\to A}|\Pi_A, \Pi_X] = F_A\left(\frac{1}{|X|}\sum_{i\in X}\pi_i\right)\vec{1}^\top$. Equation (16) follows directly. For Equation (18), we note the Dirichlet moment, $\mathbb{E}[\pi\pi^\top] = \frac{1}{\alpha_0+1}\operatorname{Diag}(\widehat{\alpha}) + \frac{\alpha_0}{\alpha_0+1}\widehat{\alpha}\widehat{\alpha}^\top$, when $\pi \sim \operatorname{Dir}(\alpha)$ and

$$\begin{aligned}
|X|^{-1}\mathbb{E}[\Psi_X\Psi_X^\top] &= \operatorname{Diag}(\widehat{\alpha}^{-1/2})\Big[(\alpha_0+1)\mathbb{E}[\pi\pi^\top] + (-2\sqrt{\alpha_0+1}(\sqrt{\alpha_0+1}-1) \\
&\quad +(\sqrt{\alpha_0+1}-1)^2)\mathbb{E}[\pi]\mathbb{E}[\pi]^\top\Big]\operatorname{Diag}(\widehat{\alpha}^{-1/2}) \\
&= \operatorname{Diag}(\widehat{\alpha}^{-1/2})\left(\operatorname{Diag}(\widehat{\alpha}) + \alpha_0\widehat{\alpha}\widehat{\alpha}^\top + (-\alpha_0)\widehat{\alpha}\widehat{\alpha}^\top\right)\operatorname{Diag}(\widehat{\alpha}^{-1/2}) \\
&= I.
\end{aligned}$$

On lines of the proof of Proposition 2 for the block model, the expectation in (17) involves multi-linear map of the expectation of the tensor products $\pi\otimes\pi\otimes\pi$ among other terms. Collecting these terms, we have that

$$\begin{aligned}
&(\alpha_0+1)(\alpha_0+2)\mathbb{E}[\pi\otimes\pi\otimes\pi] - (\alpha_0)(\alpha_0+1)(\mathbb{E}[\pi\otimes\pi\otimes\mathbb{E}[\pi]] \\
&+\mathbb{E}[\pi\otimes\mathbb{E}[\pi]\otimes\pi] + \mathbb{E}[\mathbb{E}[\pi]\otimes\pi\otimes\pi]) + 2\alpha_0^2\mathbb{E}[\pi]\otimes\mathbb{E}[\pi]\otimes\mathbb{E}[\pi]
\end{aligned}$$

is a diagonal tensor, in the sense that its (p,p,p)-th entry is $\widehat{\alpha}_p$, and its (p,q,r)-th entry is 0 when p,q,r are not all equal. With this, we have (17). ■

Note the nearly identical forms of the graph moments for the stochastic block model in (11), (12) and for the general mixed membership model in (16), (17). In other words, the modified moments $G_{X,A}^{\alpha_0}$ and T^{α_0} have similar relationships to underlying parameters as the raw moments in the case of the stochastic block model. This enables us to use a unified learning approach for the two models, outlined in the next section.

3. Algorithm for Learning Mixed Membership Models

The simple form of the graph moments derived in the previous section is now utilized to recover the community vectors Π and model parameters $P, \widehat{\alpha}$ of the mixed membership model. The method is based on the so-called tensor power method, used to obtain a tensor decomposition. We first outline the basic tensor decomposition method below and then demonstrate how the method can be adapted to learning using the graph moments at hand. We first analyze the simpler case when exact moments are available in Section 3.2 and then extend the method to handle empirical moments computed from the network observations in Section 3.3.

3.1 Overview of Tensor Decomposition through Power Iterations

In this section, we review the basic method for tensor decomposition based on power iterations for a special class of tensors, viz., symmetric orthogonal tensors. Subsequently, in Section 3.2 and 3.3, we modify this method to learn the mixed membership model from graph moments, described in the previous section. For details on the tensor power method, refer to Anandkumar et al. (2012a); Kolda and Mayo (2011).

Recall that a third-order tensor T is a three-dimensional array and we use $T_{p,q,r}$ to denote the (p,q,r)-th entry of the tensor T. The standard symbol $\otimes$ is used to denote the Kronecker product, and $(u \otimes v \otimes w)$ is a rank one tensor. The decomposition of a tensor into its rank one components is called the CP decomposition.

3.1.1 Multi-linear Maps

We can view a tensor $T \in \mathbb{R}^{d \times d \times d}$ as a multilinear map in the following sense: for a set of matrices $\{V_i \in \mathbb{R}^{d \times m_i} : i \in [3]\}$, the (i_1, i_2, i_3)-th entry in the three-way array representation of $T(V_1, V_2, V_3) \in \mathbb{R}^{m_1 \times m_2 \times m_3}$ is

$$[T(V_1, V_2, V_3)]_{i_1,i_2,i_3} := \sum_{j_1,j_2,j_3 \in [d]} T_{j_1,j_2,j_3} \, [V_1]_{j_1,i_1} \, [V_2]_{j_2,i_2} \, [V_3]_{j_3,i_3}.$$

The term multilinear map arises from the fact that the above map is linear in each of the coordinates, e.g., if we replace V_1 by $aV_1 + bW_1$ in the above equation, where W_1 is a matrix of appropriate dimensions, and a, b are any scalars, the output is a linear combination of the outputs under V_1 and W_1 respectively. We will use the above notion of multi-linear transforms to describe various tensor operations. For instance, $T(I, I, v)$ yields a matrix, $T(I, v, v)$, a vector, and $T(v, v, v)$, a scalar.

3.1.2 Symmetric Tensors and Orthogonal Decomposition

A special class of tensors are the symmetric tensors $T \in \mathbb{R}^{d \times d \times d}$ which are invariant to permutation of the array indices. Symmetric tensors have CP decomposition of the form

$$T = \sum_{i \in [r]} \lambda_i v_i \otimes v_i \otimes v_i = \sum_{i \in [r]} \lambda_i v_i^{\otimes 3}, \tag{19}$$

where r denotes the tensor CP rank and we use the notation $v_i^{\otimes 3} := v_i \otimes v_i \otimes v_i$. It is convenient to first analyze methods for decomposition of symmetric tensors and we then extend them to the general case of asymmetric tensors.

Further, a sub-class of symmetric tensors are those which possess a decomposition into orthogonal components, i.e., the vectors $v_i \in \mathbb{R}^d$ are orthogonal to one another in the above decomposition in (19) (without loss of generality, we assume that vectors $\{v_i\}$ are orthonormal in this case). An orthogonal decomposition implies that the tensor rank $r \leq d$ and there are tractable methods for recovering the rank-one components in this setting. We limit ourselves to this setting in this paper.

3.1.3 Tensor Eigen Analysis

For symmetric tensors T possessing an orthogonal decomposition of the form in (19), each pair (λ_i, v_i), for $i \in [r]$, can be interpreted as an eigen-pair for the tensor T, since

$$T(I, v_i, v_i) = \sum_{j \in [r]} \lambda_j \langle v_i, v_j \rangle^2 v_j = \lambda_i v_i, \quad \forall i \in [r],$$

due to the fact that $\langle v_i, v_j \rangle = \delta_{i,j}$. Thus, the vectors $\{v_i\}_{i \in [r]}$ can be interpreted as fixed points of the map

$$v \mapsto \frac{T(I, v, v)}{\|T(I, v, v)\|}, \tag{20}$$

where $\|\cdot\|$ denotes the spectral norm (and $\|T(I, v, v)\|$ is a vector norm), and is used to normalize the vector v in (20).

3.1.4 Basic Tensor Power Iteration Method

A straightforward approach to computing the orthogonal decomposition of a symmetric tensor is to iterate according to the fixed-point map in (20) with an arbitrary initialization vector. This is referred to as the tensor power iteration method. Additionally, it is known that the vectors $\{v_i\}_{i \in [r]}$ are the only *stable* fixed points of the map in (20). In other words, the set of initialization vectors which converge to vectors other than $\{v_i\}_{i \in [r]}$ are of measure zero. This ensures that we obtain the correct set of vectors through power iterations and that no spurious answers are obtained. See Anandkumar et al. (2012b, Thm. 4.1) for details. Moreover, after an approximately fixed point is obtained (after many power iterations), the estimated eigen-pair can be subtracted out (i.e., *deflated*) and subsequent vectors can be similarly obtained through power iterations. Thus, we can obtain all the stable eigen-pairs $\{\lambda_i, v_i\}_{i \in [r]}$ which are the components of the orthogonal tensor decomposition. The method needs to be suitably modified when the tensor T is perturbed (e.g., as in the case when empirical moments are used) and we discuss it in Section 3.3.

3.2 Learning Mixed Membership Models under Exact Moments

We first describe the learning approach when exact moments are available. In Section 3.3, we suitably modify the approach to handle perturbations, which are introduced when only empirical moments are available.

We now employ the tensor power method described above to obtain a CP decomposition of the graph moment tensor T^{α_0} in (15). We first describe a "symmetrization" procedure to convert the graph moment tensor T^{α_0} to a symmetric orthogonal tensor through a multi-linear transformation of T^{α_0}. We then employ the power method to obtain a symmetric orthogonal decomposition. Finally, the original CP decomposition is obtained by reversing the multi-linear transform of the symmetrization procedure. This yields a guaranteed method for obtaining the decomposition of graph moment tensor T^{α_0} under exact moments. We note that this symmetrization approach has been earlier employed in other contexts, e.g., for learning hidden Markov models (Anandkumar et al., 2012b, Sec. 3.3).

3.2.1 Reduction of the Graph-moment Tensor to Symmetric Orthogonal Form (Whitening)

Recall from Proposition 3 that the modified 3-star count tensor T^{α_0} has a CP decomposition as

$$\mathbb{E}[\mathrm{T}^{\alpha_0} | \Pi_A, \Pi_B, \Pi_C] = \sum_{i=1}^{k} \widehat{\alpha}_i (F_A)_i \otimes (F_B)_i \otimes (F_C)_i.$$

We now describe a symmetrization procedure to convert T^{α_0} to a symmetric orthogonal tensor through a multi-linear transformation using the modified adjacency matrix G^{α_0}, defined in (14). Consider the singular value decomposition (SVD) of the modified adjacency matrix G^{α_0} under exact moments:

$$|X|^{-1/2} \mathbb{E}[(G^{\alpha_0}_{X,A})^\top | \Pi] = U_A D_A V_A^\top.$$

Define $W_A := U_A D_A^{-1}$, and similarly define W_B and W_C using the corresponding matrices $G^{\alpha_0}_{X,B}$ and $G^{\alpha_0}_{X,C}$ respectively. Now define

$$R_{A,B} := \frac{1}{|X|} W_B^\top \mathbb{E}[(G^{\alpha_0}_{X,B})^\top | \Pi] \cdot \mathbb{E}[(G^{\alpha_0}_{X,A}) | \Pi] W_A, \quad \tilde{W}_B := W_B R_{A,B}, \tag{21}$$

and similarly define $\tilde{W}_C$. We establish that a multilinear transformation (as defined in (3.1.1)) of the graph-moment tensor T^{α_0} using matrices $W_A, \tilde{W}_B$, and $\tilde{W}_C$ results in a symmetric orthogonal form.

Lemma 4 (Orthogonal Symmetric Tensor) *Assume that the matrices F_A, F_B, F_C and Π_X have rank k, where k is the number of communities. We have an orthogonal symmetric tensor form for the modified 3-star count tensor T^{α_0} in (15) under a multilinear transformation using matrices $W_A, \tilde{W}_B$, and $\tilde{W}_C$:*

$$\mathbb{E}[\mathrm{T}^{\alpha_0}(W_A, \tilde{W}_B, \tilde{W}_C) | \Pi_A, \Pi_B, \Pi_C] = \sum_{i \in [k]} \lambda_i (\Phi)_i^{\otimes 3} \in \mathbb{R}^{k \times k \times k}, \tag{22}$$

where $\lambda_i := \widehat{\alpha}_i^{-0.5}$ and $\Phi \in \mathbb{R}^{k \times k}$ is an orthogonal matrix, given by

$$\Phi := W_A^\top F_A \, \mathrm{Diag}(\widehat{\alpha}^{0.5}). \tag{23}$$

Remark: *Note that the matrix W_A orthogonalizes F_A under exact moments, and is referred to as a* whitening matrix. *Similarly, the matrices $\tilde{W}_B = R_{A,B} W_B$ and $\tilde{W}_C = R_{A,C} W_C$ consist of whitening matrices W_B and W_C, and in addition, the matrices $R_{A,B}$ and $R_{A,C}$ serve to symmetrize the tensor. We can interpret $\{\lambda_i, (\Phi)_i\}_{i \in [k]}$ as the stable eigenpairs of the transformed tensor (henceforth, referred to as the* whitened and symmetrized tensor*).*

Remark: *The full rank assumption on matrix $F_A = \Pi_A^\top P^\top \in \mathbb{R}^{|A| \times k}$ implies that $|A| \geq k$, and similarly $|B|, |C|, |X| \geq k$. Moreover, we require the community connectivity matrix $P \in \mathbb{R}^{k \times k}$ to be of full rank*[12] *(which is a natural non-degeneracy condition). In this case,*

12. In the work of McSherry (2001), where spectral clustering for stochastic block models is analyzed, rank deficient P is allowed as long as the neighborhood vectors generated by any pair of communities are sufficiently different. On the other hand, our method requires P to be full rank. We argue that this is a mild restriction since we allow for mixed memberships while McSherry (2001) limit to the stochastic block model.

we can reduce the graph-moment tensor T^{α_0} *to a* k*-rank orthogonal symmetric tensor, which has a unique decomposition. This implies that the mixed membership model is identifiable using* 3*-star and edge count moments, when the network size* $n = |A|+|B|+|C|+|X| \geq 4k$*, matrix* P *is full rank and the community membership matrices* $\Pi_A, \Pi_B, \Pi_C, \Pi_X$ *each have rank* k*. On the other hand, when only empirical moments are available, roughly, we require the network size* $n = \Omega(k^2(\alpha_0+1)^2)$ *(where* $\alpha_0 := \sum_i \alpha_i$ *is related to the extent of overlap between the communities) to provide guaranteed learning of the community membership and model parameters. See Section 4 for a detailed sample analysis.*

Proof: Recall that the modified adjacency matrix G^{α_0} satisfies

$$\mathbb{E}[(G^{\alpha_0}_{X,A})^\top | \Pi_A, \Pi_X] = F_A \operatorname{Diag}(\widehat{\alpha}^{1/2}) \Psi_X,$$
$$\Psi_X := \operatorname{Diag}(\widehat{\alpha}^{-1/2}) \left(\sqrt{\alpha_0+1}\Pi_X - (\sqrt{\alpha_0+1}-1) \left(\frac{1}{|X|} \sum_{i \in X} \pi_i \right) \vec{1}^\top \right).$$

From the definition of Ψ_X above, we see that it has rank k when Π_X has rank k. Using the Sylvester's rank inequality, we have that the rank of $F_A \operatorname{Diag}(\widehat{\alpha}^{1/2})\Psi_X$ is at least $2k-k=k$. This implies that the whitening matrix W_A also has rank k. Notice that

$$|X|^{-1} W_A^\top \mathbb{E}[(G^{\alpha_0}_{X,A})^\top | \Pi] \cdot \mathbb{E}[(G^{\alpha_0}_{X,A}) | \Pi] W_A = D_A^{-1} U_A^\top U_A D_A^2 U_A^\top U_A D_A^{-1} = I \in \mathbb{R}^{k \times k},$$

or in other words, $|X|^{-1} M M^\top = I$, where $M := W_A^\top F_A \operatorname{Diag}(\widehat{\alpha}^{1/2}) \Psi_X$. We now have that

$$\begin{aligned} I = |X|^{-1} \mathbb{E}_{\Pi_X}\left[M M^\top \right] &= |X|^{-1} W_A^\top F_A \operatorname{Diag}(\widehat{\alpha}^{1/2}) \mathbb{E}[\Psi_X \Psi_X^\top] \operatorname{Diag}(\widehat{\alpha}^{1/2}) F_A^\top W_A \\ &= W_A^\top F_A \operatorname{Diag}(\widehat{\alpha}) F_A^\top W_A, \end{aligned}$$

since $|X|^{-1}\mathbb{E}_{\Pi_X}[\Psi_X \Psi_X^\top] = I$ from (18), and we use the fact that the sets A and X do not overlap. Thus, W_A whitens $F_A \operatorname{Diag}(\widehat{\alpha}^{1/2})$ under exact moments (up on taking expectation over Π_X) and the columns of $W_A^\top F_A \operatorname{Diag}(\widehat{\alpha}^{1/2})$ are orthonormal. Now note from the definition of $\tilde{W}_B$ that

$$\tilde{W}_B^\top \mathbb{E}[(G^{\alpha_0}_{X,B})^\top | \Pi] = W_A^\top \mathbb{E}[(G^{\alpha_0}_{X,A})^\top | \Pi],$$

since W_B satisfies

$$|X|^{-1} W_B^\top \mathbb{E}[(G^{\alpha_0}_{X,B})^\top | \Pi] \cdot \mathbb{E}[(G^{\alpha_0}_{X,B}) | \Pi] W_B = I,$$

and similar result holds for $\tilde{W}_C$. The final result in (22) follows by taking expectation of tensor T^{α_0} over Π_X. ■

3.2.2 Overview of the Learning Approach under Exact Moments

With the above result in place, we are now ready to describe the high-level approach for learning the mixed membership model under exact moments. First, symmetrize the graph-moment tensor T^{α_0} as described above and then apply the tensor power method described in the previous section. This enables us to obtain the vector of eigenvalues $\lambda := \widehat{\alpha}^{-1/2}$ and the matrix of eigenvectors $\Phi = W_A^\top F_A \operatorname{Diag}(\widehat{\alpha}^{0.5})$ using tensor power iterations. We can

then recover the community membership vectors of set A^c (i.e., nodes not in set A) under exact moments as

$$\Pi_{A^c} \leftarrow \mathrm{Diag}(\lambda)^{-1}\Phi^\top W_A^\top \mathbb{E}[G_{A^c,A}^\top|\Pi],$$

since $\mathbb{E}[G_{A^c,A}^\top|\Pi] = F_A\Pi_{A^c}$ (since A and A^c do not overlap) and $\mathrm{Diag}(\lambda)^{-1}\Phi^\top W_A^\top = \mathrm{Diag}(\widehat{\alpha})F_A^\top W_A W_A^\top$ under exact moments. In order to recover the community membership vectors of set A, viz., Π_A, we can use the edge-set $G_{A,B}$. Once all the community membership vectors Π are obtained, we can obtain the community connectivity matrix P, using the relationship: $\Pi^\top P\Pi = \mathbb{E}[G|\Pi]$ and noting that we assume Π to be of rank k. Thus, we are able to learn the community membership vectors Π and the model parameters $\widehat{\alpha}$ and P of the mixed membership model using edge counts and the 3-star count tensor. We now describe modifications to this approach to handle empirical moments.

3.3 Learning Algorithm under Empirical Moments

In the previous section, we explored a tensor-based approach for learning mixed membership model under exact moments. However, in practice, we only have samples (i.e., the observed network), and the method needs to be robust to perturbations when empirical moments are employed.

Algorithm 1 $\{\hat{\Pi}, \hat{P}, \widehat{\alpha}\} \leftarrow \mathrm{LearnMixedMembership}(G, k, \alpha_0, N, \tau)$

Input: Adjacency matrix $G \in \mathbb{R}^{n\times n}$, k is the number of communities, $\alpha_0 := \sum_i \alpha_i$, where α is the Dirichlet parameter vector, N is the number of iterations for the tensor power method, and τ is used for thresholding the estimated community membership vectors, specified in (29) in assumption A5. Let $A^c := [n] \setminus A$ denote the set of nodes not in A.

Output: Estimates of the community membership vectors $\Pi \in \mathbb{R}^{n\times k}$, community connectivity matrix $P \in [0,1]^{k\times k}$, and the normalized Dirichlet parameter vector $\widehat{\alpha}$.

Partition the vertex set $[n]$ into 5 parts X, Y, A, B, C.

Compute moments $G^{\alpha_0}_{X,A}$, $G^{\alpha_0}_{X,B}$, $G^{\alpha_0}_{X,C}$, $\mathrm{T}^{\alpha_0}_{Y\to\{A,B,C\}}$ using (14) and (15).

$\{\hat{\Pi}, \widehat{\alpha}\} \leftarrow \mathrm{LearnPartitionCommunity}(G^{\alpha_0}_{X,A}, G^{\alpha_0}_{X,B}, G^{\alpha_0}_{X,C}, \mathrm{T}^{\alpha_0}_{Y\to\{A,B,C\}}, G, N, \tau)$.

Define $\hat{Q}$ such that its i-th row is $\hat{Q}^i := (\alpha_0+1)\frac{\hat{\Pi}^i}{|\hat{\Pi}^i|_1} - \frac{\alpha_0}{n}\vec{1}^\top$. {We will establish that $\hat{Q} \approx (\Pi^\dagger)^\top$ under conditions A1-A5.}

Estimate $\hat{P} \leftarrow \hat{Q}G\hat{Q}^\top$. {Recall that $\mathbb{E}[G] = \Pi^\top P\Pi$ in our model.}

Return $\hat{\Pi}, \hat{P}, \widehat{\alpha}$

3.3.1 Pre-processing Step: Partitioning

In the previous section, we partitioned the nodes into four sets A, B, C, X for learning under exact moments. However, we require more partitions under empirical moments to avoid statistical dependency issues and obtain stronger reconstruction guarantees. We now divide the network into five non-overlapping sets A, B, C, X, Y. The set X is employed to compute whitening matrices $\hat{W}_A$, $\hat{W}_B$ and $\hat{W}_C$, described in detail subsequently, the set Y is employed to compute the 3-star count tensor T^{α_0} and sets A, B, C contain the leaves of

Procedure 1 $\{\hat{\Pi}, \widehat{\alpha}\} \leftarrow$ LearnPartitionCommunity($G^{\alpha_0}_{X,A}$, $G^{\alpha_0}_{X,B}$, $G^{\alpha_0}_{X,C}$, $\mathrm{T}^{\alpha_0}_{Y\to\{A,B,C\}}$, G, N, τ)

Input: Require modified adjacency submatrices $G^{\alpha_0}_{X,A}$, $G^{\alpha_0}_{X,B}$, $G^{\alpha_0}_{X,C}$, 3-star count tensor $\mathrm{T}^{\alpha_0}_{Y\to\{A,B,C\}}$, adjacency matrix G, number of iterations N for the tensor power method and threshold τ for thresholding estimated community membership vectors. Let $\mathrm{Thres}(A,\tau)$ denote the element-wise thresholding operation using threshold τ, i.e., $\mathrm{Thres}(A,\tau)_{i,j} = A_{i,j}$ if $A_{i,j} \geq \tau$ and 0 otherwise. Let e_i denote basis vector along coordinate i.

Output: Estimates of Π and $\widehat{\alpha}$.

Compute rank-k SVD: $(|X|^{-1/2}G^{\alpha_0}_{X,A})^{\top}_{k-svd} = U_A D_A V_A^{\top}$ and compute whitening matrices $\hat{W}_A := U_A D_A^{-1}$. Similarly, compute $\hat{W}_B, \hat{W}_C$ and $\hat{R}_{AB}, \hat{R}_{AC}$ using (24).

Compute whitened and symmetrized tensor $T \leftarrow \mathrm{T}^{\alpha_0}_{Y\to\{A,B,C\}}(\hat{W}_A, \hat{W}_B\hat{R}_{AB}, \hat{W}_C\hat{R}_{AC})$.

$\{\hat{\lambda}, \hat{\Phi}\} \leftarrow$TensorEigen$(T, \{\hat{W}_A^{\top}G^{\top}_{i,A}\}_{i\notin A}, N)$. $\{\hat{\Phi}$ is a $k\times k$ matrix with each columns being an estimated eigenvector and $\hat{\lambda}$ is the vector of estimated eigenvalues.$\}$

$\hat{\Pi}_{A^c} \leftarrow \mathrm{Thres}(\mathrm{Diag}(\hat{\lambda})^{-1}\hat{\Phi}^{\top}\hat{W}_A^{\top}G^{\top}_{A^c,A}\,,\ \tau)$ and $\hat{\alpha}_i \leftarrow \hat{\lambda}_i^{-2}$, for $i \in [k]$.

$\hat{\Pi}_A \leftarrow \mathrm{Thres}(\mathrm{Diag}(\hat{\lambda})^{-1}\hat{\Phi}^{\top}\hat{R}_{AB}^{\top}\hat{W}_B^{\top}G^{\top}_{A,B}\,,\ \tau)$.

Return $\hat{\Pi}$ and $\hat{\alpha}$.

the 3-stars under consideration. The roles of the sets can be interchanged to obtain the community membership vectors of all the sets.

3.3.2 Pre-processing Step: Whitening

The whitening procedure is along the same lines as described in the previous section, except that now empirical moments are used. Specifically, consider the k-rank singular value decomposition (SVD) of the modified adjacency matrix G^{α_0} defined in (14),

$$(|X|^{-1/2}G^{\alpha_0}_{X,A})^{\top}_{k-svd} = U_A D_A V_A^{\top}.$$

Define $\hat{W}_A := U_A D_A^{-1}$, and similarly define $\hat{W}_B$ and $\hat{W}_C$ using the corresponding matrices $G^{\alpha_0}_{X,B}$ and $G^{\alpha_0}_{X,C}$ respectively. Now define

$$\hat{R}_{A,B} := \frac{1}{|X|}\hat{W}_B^{\top}(G^{\alpha_0}_{X,B})^{\top}_{k-svd}\cdot(G^{\alpha_0}_{X,A})_{k-svd}\hat{W}_A, \tag{24}$$

and similarly define $\hat{R}_{AC}$. The whitened and symmetrized graph-moment tensor is now computed as

$$\mathrm{T}^{\alpha_0}_{Y\to\{A,B,C\}}(\hat{W}_A, \hat{W}_B\hat{R}_{AB}, \hat{W}_C\hat{R}_{AC}),$$

where T^{α_0} is given by (15) and the multi-linear transformation of a tensor is defined in (3.1.1).

3.3.3 Modifications to the Tensor Power Method

Recall that under exact moments, the stable eigen-pairs of a symmetric orthogonal tensor can be computed in a straightforward manner through the basic power iteration method in

(20), along with the deflation procedure. However, this is not sufficient to get good reconstruction guarantees under empirical moments. We now propose a robust tensor method, detailed in Procedure 2. The main modifications involve: *(i)* efficient initialization and *(ii)* adaptive deflation, which are detailed below. Employing these modifications allows us to tolerate a far greater perturbation of the third order moment tensor, than the basic tensor power procedure employed in Anandkumar et al. (2012b). See remarks following Theorem 11 in Appendix A for the precise comparison.

3.3.4 Modification 1: Efficient Initialization

Recall that the basic tensor power method incorporates generic initialization vectors and this procedure recovers all the stable eigenvectors correctly (except for initialization vectors over a set of measure zero). However, under empirical moments, we have a perturbed tensor, and here, it is advantageous to instead employ specific initialization vectors. For instance, to obtain one of the eigenvectors $(\Phi)_i$, it is advantageous to initialize with a vector in the neighborhood of $(\Phi)_i$. This not only reduces the number of power iterations required to converge (approximately), but more importantly, this makes the power method more robust to perturbations. See Theorem 11 in Appendix A.1 for a detailed analysis quantifying the relationship between initialization vectors, tensor perturbation and the resulting guarantees for recovery of the tensor eigenvectors.

For a mixed membership model in the sparse regime, recall that the community membership vectors Π are sparse (with high probability). Under this regime of the model, we note that the whitened neighborhood vectors contain good initializers for the power iterations. Specifically, in Procedure 2, we initialize with the whitened neighborhood vectors $\hat{W}_A^\top G_{i,A}^\top$, for $i \notin A$. The intuition behind this is as follows: for a suitable choice of parameters (such as the scaling of network size n with respect to the number of communities k), we expect neighborhood vectors $G_{i,A}^\top$ to concentrate around their mean values, viz., , $F_A \pi_i$. Since π_i is sparse (w.h.p) for the model regime under consideration, this implies that there exist vectors $\hat{W}_A^\top F_A \pi_i$, for $i \in A^c$, which concentrate (w.h.p) on only along a few eigen-directions of the whitened tensor, and hence, serve as an effective initializer.

3.3.5 Modification 2: Adaptive Deflation

Recall that in the basic power iteration procedure, we can obtain the eigen-pairs one after another through simple deflation: subtracting the estimates of the current eigen-pairs and running the power iterations again to obtain new eigenvectors. However, it turns out that we can establish better theoretical guarantees (in terms of greater robustness) when we adaptively deflate the tensor in each power iteration. In Procedure 2, among the estimated eigen-pairs, we only deflate those which "compete" with the current estimate of the power iteration. In other words, if the vector in the current iteration $\theta_t^{(\tau)}$ has a significant projection along the direction of an estimated eigen-pair ϕ_j, i.e.,

$$|\lambda_j \left\langle \theta_t^{(\tau)}, \phi_j \right\rangle| > \xi,$$

for some threshold ξ, then the eigen-pair is deflated; otherwise the eigenvector ϕ_j is not deflated. This allows us to carefully control the error build-up for each estimated eigenpair in our analysis. Intuitively, if an eigenvector does not have a good correlation with the

current estimate, then it does not interfere with the update of the current vector, while if the eigenvector has a good correlation, then it is pertinent that it be deflated so as to discourage convergence in the direction of the already estimated eigenvector. See Theorem 11 in Appendix A.1 for details.

Finally, we note that stabilization, as proposed by Kolda and Mayo (2011) for general tensor eigen-decomposition (as opposed to orthogonal decomposition in this paper), can be effective in improving convergence, especially on real data, and we defer its detailed analysis to future work.

Procedure 2 $\{\lambda, \Phi\} \leftarrow$TensorEigen$(T, \{v_i\}_{i\in[L]}, N)$

Input: Tensor $T \in \mathbb{R}^{k\times k\times k}$, L initialization vectors $\{v_i\}_{i\in L}$, number of iterations N.
Output: the estimated eigenvalue/eigenvector pairs $\{\lambda, \Phi\}$, where λ is the vector of eigenvalues and Φ is the matrix of eigenvectors.

```
for i = 1 to k do
  for τ = 1 to L do
    θ_0 ← v_τ.
    for t = 1 to N do
      T̃ ← T.
      for j = 1 to i − 1 (when i > 1) do
        if |λ_j ⟨θ_t^(τ), φ_j⟩| > ξ then
          T̃ ← T̃ − λ_j φ_j^⊗3.
        end if
      end for
      Compute power iteration update θ_t^(τ) := T̃(I, θ_{t−1}^(τ), θ_{t−1}^(τ)) / ‖T̃(I, θ_{t−1}^(τ), θ_{t−1}^(τ))‖
    end for
  end for
  Let τ* := arg max_{τ∈L} {T̃(θ_N^(τ), θ_N^(τ), θ_N^(τ))}.
  Do N power iteration updates starting from θ_N^(τ*) to obtain eigenvector estimate φ_i,
  and set λ_i := T̃(φ_i, φ_i, φ_i).
end for
return the estimated eigenvalue/eigenvectors (λ, Φ).
```

3.3.6 Reconstruction after Tensor Power Method

Recall that previously in Section 3.2, when exact moments are available, estimating the community membership vectors Π is straightforward, once we recover all the stable tensor eigen-pairs. However, in case of empirical moments, we can obtain better guarantees with the following modification: the estimated community membership vectors Π are further subject to thresholding so that the weak values are set to zero. Since we are limiting ourselves to the regime of the mixed membership model, where the community vectors Π are sparse (w.h.p), this modification strengthens our reconstruction guarantees. This thresholding step is incorporated in Algorithm 1.

Moreover, recall that under exact moments, estimating the community connectivity matrix P is straightforward, once we recover the community membership vectors since $P \leftarrow (\Pi^\top)^\dagger \mathbb{E}[G|\Pi]\Pi^\dagger$. However, when empirical moments are available, we are able to establish better reconstruction guarantees through a different method, outlined in Algorithm 1. We define $\hat{Q}$ such that its i-th row is

$$\hat{Q}^i := (\alpha_0 + 1)\frac{\hat{\Pi}^i}{|\hat{\Pi}^i|_1} - \frac{\alpha_0}{n}\vec{1}^\top,$$

based on estimates $\hat{\Pi}$, and the matrix $\hat{P}$ is obtained as $\hat{P} \leftarrow \hat{Q}G\hat{Q}^\top$. We subsequently establish that $\hat{Q}\hat{\Pi}^\top \approx I$, under a set of sufficient conditions outlined in the next section.

3.3.7 Improved Support Recovery Estimates in Homophilic Models

A sub-class of community model are those satisfying *homophily*. As discussed in Section 1, homophily or the tendency to form edges within the members of the same community has been posited as an important factor in community formation, especially in social settings. Many of the existing learning algorithms (e.g., Chen et al., 2012) require this assumption to provide guarantees in the stochastic block model setting. Moreover, our procedure described below can be easily modified to work in situations where the order of intra-connectivity and inter-connectivity among communities is reversed, i.e., in the community connectivity matrix $P \in [0,1]^{k\times k}$, $P(i,i) \equiv p < P(i,j) \equiv q$, for all $i \neq j$. For instance, in the k-coloring model (McSherry, 2001), $p = 0$ and $q > 0$.

We describe the post-processing method in Procedure 3 for models with community connectivity matrix P satisfying $P(i,i) \equiv p > P(i,j) \equiv q$ for all $i \neq j$. For such models, we can obtain improved estimates by averaging. Specifically, consider nodes in set C and edges going from C to nodes in B. First, consider the special case of the stochastic block model: for each node $c \in C$, compute the number of neighbors in B belonging to each community (as given by the estimate $\hat{\Pi}$ from Algorithm 1), and declare the community with the maximum number of such neighbors as the community of node c. Intuitively, this provides a better estimate for Π_C since we average over the edges in B. This method has been used before in the context of spectral clustering (McSherry, 2001).

The same idea can be extended to the general mixed membership (homophilic) models: declare communities to be significant if they exceed a certain threshold, as evaluated by the average number of edges to each community. The correctness of the procedure can be gleaned from the fact that if the true F matrix is input, it satisfies

$$F_{j,i} = q + \Pi_{i,j}(p - q), \quad \forall\, i \in [k], j \in [n],$$

and if the true P matrix is input, $H = p$ and $L = q$. Thus, under a suitable threshold ξ, the entries $F_{j,i}$ provide information on whether the corresponding community weight $\Pi_{i,j}$ is significant.

In the next section, we establish that in certain regime of parameters, this support recovery procedure can lead to zero-error support recovery of significant community memberships of the nodes and also rule out communities where a node does not have a strong presence.

Procedure 3 $\{\hat{S}\} \leftarrow \text{SupportRecoveryHomophilicModels}(G, k, \alpha_0, \xi, \hat{\Pi})$

Input: Adjacency matrix $G \in \mathbb{R}^{n \times n}$, k is the number of communities, $\alpha_0 := \sum_i \alpha_i$, where α is the Dirichlet parameter vector, ξ is the threshold for support recovery, corresponding to significant community memberships of an individual. Get estimate $\hat{\Pi}$ from Algorithm 1. Also asume the model is homophilic: $P(i,i) \equiv p > P(i,j) \equiv q$, for all $i \neq j$.

Output: $\hat{S} \in \{0,1\}^{n \times k}$ is the estimated support for significant community memberships (see Theorem 7 for guarantees).

Consider partitions A, B, C, X, Y as in Algorithm 1.

Define $\hat{Q}$ on lines of definition in Algorithm 1, using estimates $\hat{\Pi}$. Let the i-th row for set B be $\hat{Q}^i_B := (\alpha_0 + 1)\frac{\hat{\Pi}^i_B}{|\hat{\Pi}^i_B|_1} - \frac{\alpha_0}{n}\vec{1}^\top$. Similarly define $\hat{Q}^i_C$.

Estimate $\hat{F}_C \leftarrow G_{C,B}\hat{Q}_B^\top$, $\hat{P} \leftarrow \hat{Q}_C\hat{F}_C$.

if $\alpha_0 = 0$ (stochastic block model) **then**

  for $x \in C$ **do**

    Let $i^* \leftarrow \arg\max_{i \in [k]} \hat{F}_C(x,i)$ and $\hat{S}(i^*, x) \leftarrow 1$ and 0 o.w.

  end for

else

  Let H be the average of diagonals of $\hat{P}$, L be the average of off-diagonals of $\hat{P}$

  for $x \in C$, $i \in [k]$ **do**

    $\hat{S}(i,x) \leftarrow 1$ if $\hat{F}_C(x,i) \geq L + (H - L) \cdot \frac{3\xi}{4}$ and zero otherwise.{Identify large entries}

  end for

end if

Permute the roles of the sets A, B, C, X, Y to get results for remaining nodes.

3.3.8 Computational Complexity

We note that the computational complexity of the method, implemented naively, is $O(n^2k + k^{4.43}\widehat{\alpha}_{\min}^{-1})$ when $\alpha_0 > 1$ and $O(n^2k)$ when $\alpha_0 < 1$. This is because the time for computing whitening matrices is dominated by SVD of the top k singular vectors of $n \times n$ matrix, which takes $O(n^2k)$ time. We then compute the whitened tensor T which requires time $O(n^2k + k^3n) = O(n^2k)$, since for each $i \in Y$, we multiply $G_{i,A}, G_{i,B}, G_{i,C}$ with the corresponding whitening matrices, and this step takes $O(nk)$ time. We then average this $k \times k \times k$ tensor over different nodes $i \in Y$ to the result, which takes $O(k^3)$ time in each step.

For the tensor power method, the time required for a single iteration is $O(k^3)$. We need at most $\log n$ iterations per initial vector, and we need to consider $O(\widehat{\alpha}_{\min}^{-1}k^{0.43})$ initial vectors (this could be smaller when $\alpha_0 < 1$). Hence the total running time of tensor power method is $O(k^{4.43}\widehat{\alpha}_{\min}^{-1})$ (and when α_0 is small this can be improved to $O(k^4\widehat{\alpha}_{\min}^{-1})$ which is dominated by $O(n^2k)$.

In the process of estimating Π and P, the dominant operation is multiplying $k \times n$ matrix by $n \times n$ matrix, which takes $O(n^2k)$ time. For support recovery, the dominant operation is computing the "average degree", which again takes $O(n^2k)$ time. Thus, we have that the overall computational time is $O(n^2k + k^{4.43}\widehat{\alpha}_{\min}^{-1})$ when $\alpha_0 > 1$ and $O(n^2k)$ when $\alpha_0 < 1$.

Note that the above bound on complexity of our method nearly matches the bound for spectral clustering method (McSherry, 2001), since computing the k-rank SVD requires $O(n^2k)$ time. Another method for learning stochastic block models is based on convex optimization involving semi-definite programming (SDP) (Chen et al., 2012), and it provides the best scaling bounds (for both the network size n and the separation $p - q$ for edge connectivity) known so far. The specific convex problem can be solved via the method of *augmented Lagrange multipliers* (Lin et al., 2010), where each step consists of an SVD operation and q-linear convergence is established by Lin et al. (2010). This implies that the method has complexity $O(n^3)$, since it involves taking SVD of a general $n \times n$ matrix, rather than a k-rank SVD. Thus, our method has significant advantage in terms of computational complexity, when the number of communities is much smaller than the network size ($k \ll n$).

Further, a subsequent work provides a more sophisticated implementation of the proposed tensor method through parallelization and the use of stochastic gradient descent for tensor decomposition (Huang et al., 2013). Additionally, the k-rank SVD operations are approximated via randomized methods such as the Nystrom's method leading to more efficient implementations (Gittens and Mahoney, 2013). Huang et al. (2013) deploy the tensor approach for community detection and establish that it has a running time of $O(n + k^3)$ using nk cores under a parallel computation model (JáJá, 1992).

4. Sample Analysis for Proposed Learning Algorithm

In this section we analyze our algorithm when the moments are estimated from the sample. Unlike common sample complexity analysis, here we are only given one instance of the graph. We treat the edges in the graph as independent samples (conditioned on community membership), and the "sample complexity" will control how many communities we can learn in a graph with n vertices.

4.1 Homogeneous Mixed Membership Models

It is easier to first present the results for our proposed algorithm for the special case, where all the communities have the same expected size and the entries of the community connectivity matrix P are equal on diagonal and off-diagonal locations:

$$\widehat{\alpha}_i \equiv \frac{1}{k}, \qquad P(i,j) = p \cdot \mathbb{I}(i=j) + q \cdot \mathbb{I}(i \neq j), \quad p \geq q. \tag{25}$$

In other words, the probability of an edge according to P only depends on whether it is between two individuals of the same community or between different communities. The above setting is also well studied for stochastic block models ($\alpha_0 = 0$), allowing us to compare our results with existing ones. The results for general mixed membership models are deferred to Section 4.2.

[A1] Sparse regime of Dirichlet parameters: The community membership vectors are drawn from the Dirichlet distribution, $\mathrm{Dir}(\alpha)$, under the mixed membership model. We assume that $\alpha_i < 1$ for $i \in [k]$ (see Section 2.1 for an extended discussion on the sparse regime of the Dirichlet distribution) and that α_0 is known.

[A2] Condition on the network size: Given the concentration parameter of the Dirichlet distribution, $\alpha_0 := \sum_i \alpha_i$, we require that

$$n = \tilde{\Omega}(k^2(\alpha_0+1)^2), \tag{26}$$

and that the disjoint sets A, B, C, X, Y are chosen randomly and are of size $\Theta(n)$. Note that from assumption A1, $\alpha_i < 1$ which implies that $\alpha_0 < k$. Thus, in the worst-case, when $\alpha_0 = \Theta(k)$, we require[13] $n = \tilde{\Omega}(k^4)$, and in the best case, when $\alpha_0 = \Theta(1)$, we require $n = \tilde{\Omega}(k^2)$. The latter case includes the stochastic block model ($\alpha_0 = 0$), and thus, our results match the state-of-art bounds for learning stochastic block models.

[A3] Condition on edge connectivity: Recall that p is the probability of intra-community connectivity and q is the probability of inter-community connectivity. We require that

$$\frac{p-q}{\sqrt{p}} = \Omega\left(\frac{(\alpha_0+1)k}{n^{1/2}}\right) \tag{27}$$

The above condition is on the standardized separation between intra-community and inter-community connectivity (note that $\sqrt{p}$ is the standard deviation of a Bernoulli random variable). The above condition is required to control the perturbation in the whitened tensor (computed using observed network samples), thereby, providing guarantees on the estimated eigen-pairs through the tensor power method.

[A4] Condition on number of iterations of the power method: We assume that the number of iterations N of the tensor power method in Procedure 2 satisfies

$$N \geq C_2 \cdot \left(\log(k) + \log\log\left(\frac{p-q}{p}\right)\right), \tag{28}$$

for some constant C_2.

13. The notation $\tilde{\Omega}(\cdot), \tilde{O}(\cdot)$ denotes $\Omega(\cdot), O(\cdot)$ up to poly-log factors.

[A5] Choice of τ for thresholding community vector estimates: The threshold τ for obtaining estimates $\hat{\Pi}$ of community membership vectors in Algorithm 1 is chosen as

$$\tau = \begin{cases} \Theta\left(\frac{k\sqrt{\alpha_0}}{\sqrt{n}} \cdot \frac{\sqrt{p}}{p-q}\right), & \alpha_0 \neq 0, \quad (29) \\ 0.5, & \alpha_0 = 0, \quad (30) \end{cases}$$

For the stochastic block model ($\alpha_0 = 0$), since π_i is a basis vector, we can use a large threshold. For general models ($\alpha_0 \neq 0$), τ can be viewed as a regularization parameter and decays as $n^{-1/2}$ when other parameters are held fixed. We are now ready to state the error bounds on the estimates of community membership vectors Π and the block connectivity matrix P. $\hat{\Pi}$ and $\hat{P}$ are the estimates computed in Algorithm 1.

Recall that for a matrix M, $(M)^i$ and $(M)_i$ denote the i^{th} row and column respectively. We say that an event holds with high probability, if it occurs with probability $1 - n^{-c}$ for some constant $c > 0$.

Theorem 5 (Guarantees on Estimating P, Π) *Under assumptions A1-A5, we have with high probability*

$$\varepsilon_{\pi,\ell_1} := \max_{i\in[n]} \|\hat{\Pi}^i - \Pi^i\|_1 = \tilde{O}\left(\frac{(\alpha_0+1)^{3/2}\sqrt{np}}{(p-q)}\right) \quad (31)$$

$$\varepsilon_P := \max_{i,j\in[k]} |\hat{P}_{i,j} - P_{i,j}| = \tilde{O}\left(\frac{(\alpha_0+1)^{3/2}k\sqrt{p}}{\sqrt{n}}\right). \quad (32)$$

The proofs are given in the Appendix and a proof outline is provided in Section 4.3.

The main ingredient in establishing the above result is the tensor concentration bound and additionally, recovery guarantees under the tensor power method in Procedure 2. We now provide these results below.

Recall that $F_A := \Pi_A^\top P^\top$ and $\Phi = W_A^\top F_A \operatorname{Diag}(\widehat{\alpha}^{1/2})$ denotes the set of tensor eigenvectors under exact moments in (23), and $\hat{\Phi}$ is the set of estimated eigenvectors under empirical moments, obtained using Procedure 1. We establish the following guarantees.

Lemma 6 (Perturbation bound for estimated eigen-pairs) *Under the assumptions A1-A4, the recovered eigenvector-eigenvalue pairs $(\hat{\Phi}_i, \hat{\lambda}_i)$ from the tensor power method in Procedure 2 satisfies with high probability, for a permutation θ, such that*

$$\max_{i\in[k]} \|\hat{\Phi}_i - \Phi_{\theta(i)}\| \leq 8k^{-1/2}\varepsilon_T, \qquad \max_{i\in[k]} |\lambda_i - \widehat{\alpha}_{\theta(i)}^{-1/2}| \leq 5\varepsilon_T, \quad (33)$$

The tensor perturbation bound ε_T is given by

$$\varepsilon_T := \Big\| \mathrm{T}^{\alpha_0}_{Y\to\{A,B,C\}}(\hat{W}_A, \hat{W}_B\hat{R}_{AB}, \hat{W}_C\hat{R}_{AC}) - \mathbb{E}[\mathrm{T}^{\alpha_0}_{Y\to\{A,B,C\}}(W_A, W_BR_{AB}, W_CR_{AC})|\Pi_{A\cup B\cup C}] \Big\| \quad (34)$$

$$= \tilde{O}\left(\frac{(\alpha_0+1)k^{3/2}\sqrt{p}}{(p-q)\sqrt{n}}\right), \quad (35)$$

where $\|T\|$ for a tensor T refers to its spectral norm.

Remark: *(Stochastic Block Models ($\alpha_0 = 0$)) For stochastic block models, assumptions A2 and A3 reduce to*

$$n = \tilde{\Omega}(k^2), \qquad \zeta = \Theta\left(\frac{\sqrt{p}}{p-q}\right) = O\left(\frac{n^{1/2}}{k}\right). \tag{36}$$

This matches with the best known scaling (up to poly-log factors), and was previously achieved via convex optimization by Chen et al. (2012) for stochastic block models. However, our results in Theorem 5 do not provide zero error guarantees as in Chen et al. (2012). We strengthen our results to provide zero-error guarantees in Section 4.1.1 below and thus, match the scaling of Chen et al. (2012) for stochastic block models. Moreover, we also provide zero-error support recovery guarantees for recovering significant memberships of nodes in mixed membership models in Section 4.1.1.

Remark: *(Dependence on α_0) The guarantees degrade as α_0 increases, which is intuitive since the extent of community overlap increases. The requirement for scaling of n also grows as α_0 increases. Note that the guarantees on ε_π and ε_P can be improved by assuming a more stringent scaling of n with respect to α_0, rather than the one specified by A2.*

4.1.1 Zero-error Guarantees for Support Recovery

Recall that we proposed Procedure 3 as a post-processing step to provide improved support recovery estimates. We now provide guarantees for this method.

We now specify the threshold ξ for support recovery in Procedure 3.

[A6] Choice of ξ for support recovery: We assume that the threshold ξ in Procedure 3 satisfies

$$\xi = \Omega(\varepsilon_P),$$

where ε_P is specified in Theorem 5.

We now state the guarantees for support recovery.

Theorem 7 (Support recovery guarantees) *Assuming A1-A6 and (25) hold, the support recovery method in Procedure 3 has the following guarantees on the estimated support set $\hat{S}$: with high probability,*

$$\Pi(i,j) \geq \xi \Rightarrow \hat{S}(i,j) = 1 \quad \text{and} \quad \Pi(i,j) \leq \frac{\xi}{2} \Rightarrow \hat{S}(i,j) = 0, \quad \forall i \in [k], j \in [n], \tag{37}$$

where Π is the true community membership matrix.

Thus, the above result guarantees that the Procedure 3 correctly recovers all the "large" entries of Π and also correctly rules out all the "small" entries in Π. In other words, we can correctly infer all the significant memberships of each node and also rule out the set of communities where a node does not have a strong presence.

The only shortcoming of the above result is that there is a gap between the "large" and "small" values, and for an intermediate set of values (in $[\xi/2, \xi]$), we cannot guarantee correct inferences about the community memberships. Note this gap depends on ε_P, the error in estimating the P matrix. This is intuitive, since as the error ε_P decreases, we can infer the community memberships over a large range of values.

For the special case of stochastic block models (i.e., $\lim \alpha_0 \to 0$), we can improve the above result and give a zero error guarantee at all nodes (w.h.p). Note that we no longer require a threshold ξ in this case, and only infer one community for each node.

Corollary 8 (Zero error guarantee for block models) *Assuming A1-A5 and* (25) *hold, the support recovery method in Procedure 3 correctly identifies the community memberships for all nodes with high probability in case of stochastic block models* ($\alpha_0 \to 0$).

Thus, with the above result, we match the state-of-art results of Chen et al. (2012) for stochastic block models in terms of scaling requirements and recovery guarantees.

4.2 General (Non-homogeneous) Mixed Membership Models

In the previous sections, we provided learning guarantees for learning homogeneous mixed membership models. Here, we extend the results to learning general non-homogeneous mixed membership models under a sufficient set of conditions, involving scaling of various parameters such as network size n, number of communities k, concentration parameter α_0 of the Dirichlet distribution (which is a measure of overlap of the communities) and so on.
[B1] Sparse regime of Dirichlet parameters: The community membership vectors are drawn from the Dirichlet distribution, $\mathrm{Dir}(\alpha)$, under the mixed membership model. We assume that[14] $\alpha_i < 1$ for $i \in [k]$ $\alpha_i < 1$ (see Section 2.1 for an extended discussion on the sparse regime of the Dirichlet distribution).
[B2] Condition on the network size: Given the concentration parameter of the Dirichlet distribution, $\alpha_0 := \sum_i \alpha_i$, and $\widehat{\alpha}_{\min} := \alpha_{\min}/\alpha_0$, the expected size of the smallest community, define

$$\rho := \frac{\alpha_0 + 1}{\widehat{\alpha}_{\min}}. \tag{38}$$

We require that the network size scale as

$$n = \Omega\left(\rho^2 \log^2 k\right), \tag{39}$$

and that the sets A, B, C, X, Y are $\Theta(n)$. Note that from assumption B1, $\alpha_i < 1$ which implies that $\alpha_0 < k$. Thus, in the worst-case, when $\alpha_0 = \Theta(k)$, we require[15] $n = \tilde{\Omega}(k^4)$, assuming equal sizes: $\widehat{\alpha}_i = 1/k$, and in the best case, when $\alpha_0 = \Theta(1)$, we require $n = \tilde{\Omega}(k^2)$. The latter case includes the stochastic block model ($\alpha_0 = 0$), and thus, our results match the state-of-art bounds for learning stochastic block models. See Section 4.1 for an extended discussion.
[B3] Condition on relative community sizes and block connectivity matrix: Recall that $P \in [0,1]^{k \times k}$ denotes the block connectivity matrix. Define

$$\zeta := \left(\frac{\widehat{\alpha}_{\max}}{\widehat{\alpha}_{\min}}\right)^{1/2} \frac{\sqrt{(\max_i (P\widehat{\alpha})_i)}}{\sigma_{\min}(P)}, \tag{40}$$

14. The assumption B1 that the Dirichlet distribution be in the sparse regime is not strictly needed. Our results can be extended to general Dirichlet distributions, but with worse scaling requirements on n. The dependence of n is still polynomial in α_0, i.e., we require $n = \tilde{\Omega}((\alpha_0+1)^c \widehat{\alpha}_{\min}^{-2})$, where $c \geq 2$ is some constant.
15. The notation $\tilde{\Omega}(\cdot), \tilde{O}(\cdot)$ denotes $\Omega(\cdot), O(\cdot)$ up to log factors.

where $\sigma_{\min}(P)$ is the minimum singular value of P. We require that

$$\zeta = \begin{cases} O\left(\frac{n^{1/2}}{\rho}\right), & \alpha_0 < 1 \quad (41) \\ O\left(\frac{n^{1/2}}{\rho k \widehat{\alpha}_{\max}}\right) & \alpha_0 \geq 1. \quad (42) \end{cases}$$

Intuitively, the above condition requires the ratio of maximum and minimum expected community sizes to be not too large and for the matrix P to be well conditioned. The above condition is required to control the perturbation in the whitened tensor (computed using observed network samples), thereby, providing guarantees on the estimated eigen-pairs through the tensor power method. The above condition can be interpreted as a separation requirement between intra-community and inter-community connectivity in the special case considered in Section 4.1. Specifically, for the special case of homogeneous mixed membership model, we have

$$\sigma_{\min}(P) = \Theta(p-q), \quad \max_i (P\widehat{\alpha})_i = \frac{p}{k} + (k-1)\frac{q}{k} \leq p.$$

Thus, the assumptions A2 and A3 in Section 4.1 given by

$$n = \tilde{\Omega}(k^2(\alpha_0+1)^2), \qquad \zeta = \Theta\left(\frac{\sqrt{p}}{p-q}\right) = O\left(\frac{n^{1/2}}{(\alpha_0+1)k}\right)$$

are special cases of the assumptions B2 and B3 above.

[B4] Condition on number of iterations of the power method: We assume that the number of iterations N of the tensor power method in Procedure 2 satisfies

$$N \geq C_2 \cdot \left(\log(k) + \log\log\left(\frac{\sigma_{\min}(P)}{(\max_i (P\widehat{\alpha})_i)}\right)\right), \quad (43)$$

for some constant C_2.

[B5] Choice of τ for thresholding community vector estimates: The threshold τ for obtaining estimates $\hat{\Pi}$ of community membership vectors in Algorithm 1 is chosen as

$$\tau = \begin{cases} \Theta\left(\frac{\rho^{1/2} \cdot \zeta \cdot \widehat{\alpha}_{\max}^{1/2}}{n^{1/2} \cdot \widehat{\alpha}_{\min}}\right), & \alpha_0 \neq 0, \quad (44) \\ 0.5, & \alpha_0 = 0, \quad (45) \end{cases}$$

For the stochastic block model ($\alpha_0 = 0$), since π_i is a basis vector, we can use a large threshold. For general models ($\alpha_0 \neq 0$), τ can be viewed as a regularization parameter and decays as $n^{-1/2}$ when other parameters are held fixed. Moreover, when $n = \tilde{\Theta}(\rho^2)$, we have that $\tau \sim \rho^{-1/2}$ when other terms are held fixed. Recall that $\rho \propto (\alpha_0+1)$ when the expected community sizes $\widehat{\alpha}_i$ are held fixed. In this case, $\tau \sim \rho^{-1/2}$ allows for smaller values to be picked up after thresholding as α_0 is increased. This is intuitive since as α_0 increases, the community vectors π are more "spread out" across different communities and have smaller values.

We are now ready to state the error bounds on the estimates of community membership vectors Π and the block connectivity matrix P. $\hat{\Pi}$ and $\hat{P}$ are the estimates computed in Algorithm 1.

Recall that for a matrix M, $(M)^i$ and $(M)_i$ denote the i^{th} row and column respectively. We say that an event holds with high probability, if it occurs with probability $1 - n^{-c}$ for some constant $c > 0$.

Theorem 9 (Guarantees on estimating P, Π) *Under assumptions B1-B5, The estimates $\hat{P}$ and $\hat{\Pi}$ obtained from Algorithm 1 satisfy with high probability,*

$$\varepsilon_{\pi,\ell_1} := \max_{i\in[k]} |(\hat{\Pi})^i - (\Pi)^i|_1 = \tilde{O}\left(n^{1/2}\cdot\rho^{3/2}\cdot\zeta\cdot\widehat{\alpha}_{\max}\right) \tag{46}$$

$$\varepsilon_P := \max_{i,j\in[n]} |\hat{P}_{i,j} - P_{i,j}| = \tilde{O}\left(n^{-1/2}\cdot\rho^{5/2}\cdot\zeta\cdot\widehat{\alpha}_{\max}^{3/2}\cdot(P_{\max}-P_{\min})\right) \tag{47}$$

The proofs are in Appendix B and a proof outline is provided in Section 4.3.

The main ingredient in establishing the above result is the tensor concentration bound and additionally, recovery guarantees under the tensor power method in Procedure 2. We now provide these results below.

Recall that $F_A := \Pi_A^\top P^\top$ and $\Phi = W_A^\top F_A \operatorname{Diag}(\widehat{\alpha}^{1/2})$ denotes the set of tensor eigenvectors under exact moments in (23), and $\hat{\Phi}$ is the set of estimated eigenvectors under empirical moments, obtained using Procedure 1. We establish the following guarantees.

Lemma 10 (Perturbation bound for estimated eigen-pairs) *Under the assumptions B1-B4, the recovered eigenvector-eigenvalue pairs $(\hat{\Phi}_i, \hat{\lambda}_i)$ from the tensor power method in Procedure 2 satisfies with high probability, for a permutation θ, such that*

$$\max_{i\in[k]} \|\hat{\Phi}_i - \Phi_{\theta(i)}\| \le 8\widehat{\alpha}_{\max}^{1/2}\varepsilon_T, \qquad \max_i |\lambda_i - \widehat{\alpha}_{\theta(i)}^{-1/2}| \le 5\varepsilon_T, \tag{48}$$

The tensor perturbation bound ε_T is given by

$$\begin{aligned}\varepsilon_T &:= \Big\| \mathrm{T}^{\alpha_0}_{Y\to\{A,B,C\}}(\hat{W}_A, \hat{W}_B\hat{R}_{AB}, \hat{W}_C\hat{R}_{AC}) \\ &\qquad -\mathbb{E}[\mathrm{T}^{\alpha_0}_{Y\to\{A,B,C\}}(W_A, W_BR_{AB}, W_CR_{AC})|\Pi_{A\cup B\cup C}]\Big\| \qquad (49)\\ &= \tilde{O}\left(\frac{\rho}{\sqrt{n}}\cdot\frac{\zeta}{\widehat{\alpha}_{\max}^{1/2}}\right), \qquad (50)\end{aligned}$$

where $\|T\|$ for a tensor T refers to its spectral norm, ρ is defined in (38) *and ζ in* (40).

4.2.1 Application to Planted Clique Problem

The planted clique problem is a special case of the stochastic block model Condon and Karp (1999), and is arguably the simplest setting for the community problem. Here, a clique of size s is uniformly planted (or placed) in an Erdős-Rényi graph with edge probability 0.5. This can be viewed as a stochastic block model with $k = 2$ communities, where $\widehat{\alpha}_{\min} = s/n$ is the probability of a node being in a clique and $\widehat{\alpha}_{\max} = 1 - s/n$. The connectivity matrix

is $P = [1, q; q, q]$ with $q = 0.5$, since the probability of connectivity within the clique is 1 and the probability of connectivity for any other node pair is 0.5.

Since the planted clique setting has unequal sized communities, the general result in Section 9 is applicable, and we demonstrate how the assumptions $(B1)$-$(B5)$ simplify for the planted clique setting. We have that $\alpha_0 = 0$, since the communities are non-overlapping. For assumption $B2$, we have that

$$\rho = \frac{\alpha_0 + 1}{\widehat{\alpha}_{\min}} = \frac{n}{s}, \quad n = \tilde{\Omega}(\rho^2) \Rightarrow s = \tilde{\Omega}(\sqrt{n}). \tag{51}$$

For assumption $B3$, we have that $\sigma_{\min}(P) = \Theta(1)$ and that $\max_i (P\widehat{\alpha})_i \leq s/n + q \leq 2$, and thus the assumption $B3$ simplifies as

$$\zeta := \left(\frac{\widehat{\alpha}_{\max}}{\widehat{\alpha}_{\min}}\right)^{1/2} \frac{\sqrt{(\max_i (P\widehat{\alpha})_i)}}{\sigma_{\min}(P)} = \tilde{O}\left(\frac{\sqrt{n}}{\rho}\right) \Rightarrow s = \tilde{\Omega}\left(n^{2/3}\right). \tag{52}$$

The condition in (51) that $s = \tilde{\Omega}(n^{1/2})$ matches the computational lower bounds for recovering the clique (Feldman et al., 2012). Unfortunately, the condition in (52) that $s = \tilde{\Omega}\left(n^{2/3}\right)$ is worse. This is required for assumption $(B3)$ to hold, which is needed to ensure the success of the tensor power method. The whitening step is particularly sensitive to the condition number of the matrix to be whitened (i.e., matrices F_A, F_B, F_C in our case and the condition numbers for these matrices depend on the ratio of the community sizes), which results in a weaker guarantee. Thus, our method does not perform very well when the community sizes are drastically different. It remains an open question if our method can be improved in this setting. We conjecture that using "peeling" ideas similar to Ailon et al. (2013), where the communities are recovered one by one can improve our dependence on the ratio of community sizes.

4.3 Proof Outline

We now summarize the main techniques involved in proving Theorem 9. The details are in the Appendix. The main ingredient is the concentration of the adjacency matrix: since the edges are drawn independently conditioned on the community memberships, we establish that the adjacency matrix concentrates around its mean under the stated assumptions. See Appendix C.4 for details. With this in hand, we can then establish concentration of various quantities used by our learning algorithm.

Step 1: Whitening matrices. We first establish concentration bounds on the whitening matrices $\hat{W}_A, \hat{W}_B, \hat{W}_C$ computed using empirical moments, described in Section 3.3.1. With this in hand, we can approximately recover the span of matrix F_A since $\hat{W}_A^\top F \operatorname{Diag}(\widehat{\alpha}_i)^{1/2}$ is a rotation matrix. The main technique employed is the Matrix Bernstein's inequality (Tropp, 2012, thm. 1.4). See Appendix C.2 for details.

Step 2: Tensor concentration bounds. Recall that we use the whitening matrices to obtain a symmetric orthogonal tensor. We establish that the whitened and symmetrized tensor concentrates around its mean. (Note that the empirical third order tensor $T_{X \to A,B,C}$ tends to its expectation conditioned on Π_A, Π_B, Π_C when $|X| \to \infty$). This is done in several stages and we carefully control the tensor perturbation bounds. See Appendix C.1 for details.

Step 3: Tensor power method analysis. We analyze the performance of Procedure 2 under empirical moments. We employ the various improvements, detailed in Section 3.3.3 to establish guarantees on the recovered eigen-pairs. This includes coming up with a condition on the tensor perturbation bound, for the tensor power method to succeed. It also involves establishing that there exist good initializers for the power method among (whitened) neighborhood vectors. This allows us to obtain stronger guarantees for the tensor power method, compared to earlier analysis by Anandkumar et al. (2012b). This analysis is crucial for us to obtain state-of-art scaling bounds for guaranteed recovery (for the special case of stochastic block model). See Appendix A for details.

Step 4: Thresholding of estimated community vectors. In Step 3, we provide guarantees for recovery of each eigenvector in ℓ_2 norm. Direct application of this result only allows us to obtain ℓ_2 norm bounds for row-wise recovery of the community matrix Π. In order to strengthen the result to an ℓ_1 norm bound, we threshold the estimated Π vectors. Here, we exploit the sparsity in Dirichlet draws and carefully control the contribution of weak entries in the vector. Finally, we establish perturbation bounds on P through rather straightforward concentration bound arguments. See Appendix B.2 for details.

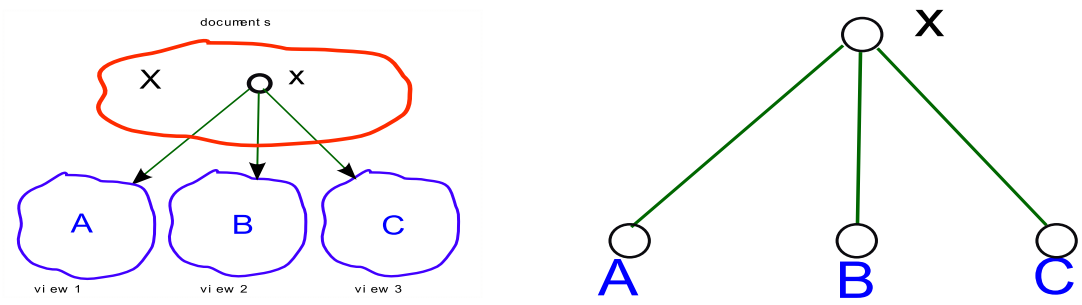
Step 5: Support recovery guarantees. To simplify the argument, consider the stochastic block model. Recall that Procedure 3 readjusts the community membership estimates based on degree averaging. For each vertex, if we count the average degree towards these "approximate communities", for the correct community the result is concentrated around value p and for the wrong community the result is around value q. Therefore, we can correctly identify the community memberships of all the nodes, when $p - q$ is sufficiently large, as specified by A3. The argument can be easily extended to general mixed membership models. See Appendix B.4 for details.

4.4 Comparison with Previous Results

We now compare the results of this paper to our previous work (Anandkumar et al., 2012b) on the use of tensor-based approaches for learning various latent variable models such as topic models, hidden Markov models (HMM) and Gaussian mixtures. At a high level, the tensor approach is exploited in a similar manner in all these models (including the community model in this paper), viz., that the conditional-independence relationships of the model result in a low rank tensor, constructed from low order moments under the given model. However, there are several important differences between the community model and the other latent variable models considered by Anandkumar et al. (2012b) and we list them below. We also precisely list the various algorithmic improvements proposed in this paper with respect to the tensor power method, and how they can be applicable to other latent variable models.

4.4.1 Topic Model vs. Community Model

Among the latent variable models studied by Anandkumar et al. (2012b), the topic model, viz., latent Dirichlet allocation (LDA), bears the closest resemblance to MMSB. In fact, the MMSB model was originally inspired by the LDA model. The analogy between the MMSB model and the LDA is direct under our framework and we describe it below.



(a) Community model as a topic model

(b) Graphical model representation

Figure 2: Casting the community model as a topic model, we obtain conditional independence of the three views.

Recall that for learning MMSBs, we consider a partition of the nodes $\{X, A, B, C\}$ and we consider the set of 3-stars from set X to A, B, C. We can construct an equivalent topic model as follows: the nodes in X form the "documents" and for each document $x \in X$, the neighborhood vectors $G_{xA}^\top, G_{xB}^\top, G_{xC}^\top$ form the three "words" or "views" for that document. In each document $x \in X$, the community vector π_x corresponds to the "topic vector" and the matrices F_A, F_B and F_C correspond to the topic-word matrices. Note that the three views $G_{xA}^\top, G_{xB}^\top, G_{xC}^\top$ are conditionally independent given the topic vector π_x. Thus, the community model can be cast as a topic model or a multi-view model. See Figure 2.

Although the community model can be viewed as a topic model, it has some important special properties which allows us to provide better guarantees. The topic-word matrices F_A, F_B, F_C are not arbitrary matrices. Recall that $F_A := \Pi_A^\top P^\top$ and similarly F_B, F_C are random matrices and we can provide strong concentration bounds for these matrices by appealing to random matrix theory. Moreover, each of the views in the community model has additional structure, viz., the vector $G_{x,A}^\top$ has independent Bernoulli entries conditioned on the community vector π_x, while in a general multi-view model, we only specify the conditional distribution of each view given the hidden topic vector. This further allows us to provide specialized concentration bounds for the community model. Importantly, we can recover the community memberships (or topic vectors) accurately while for a general multi-view model this cannot be guaranteed and we can only hope to recover the model parameters.

4.4.2 Improvements to Tensor Recovery Guarantees in This Paper

In this paper, we make modifications to the tensor power method of Anandkumar et al. (2012b) and obtain better guarantees for the community setting. Recall that the two modifications are adaptive deflation and initialization using whitened neighborhood vectors. The adaptive deflation leads to a weaker gap condition for an initialization vector to succeed in estimating a tensor eigenvector efficiently. Initialization using whitened neighborhood vectors allows us to tolerate more noise in the estimated 3-star tensor, thereby improving our sample complexity result. We make this improvement precise below.

If we directly apply the tensor power method of Anandkumar et al. (2012b), without considering the modifications, we require a stronger condition on the sample complexity and edge connectivity. For simplicity, consider the homogeneous setting of Section 4.1. The conditions $(A2)$ and $(A3)$ now need to be replaced with stronger conditions: **[A2']** **Sample complexity:** The number of samples satisfies

$$n = \tilde{\Omega}(k^4(\alpha_0 + 1)^2).$$

[A3'] **Edge connectivity:** The edge connectivity parameters p, q satisfy

$$\frac{p - q}{\sqrt{p}} = \Omega\left(\frac{(\alpha_0 + 1)k^2}{\sqrt{n}}\right).$$

Thus, we obtain significant improvements in recovery guarantees via algorithmic modifications and careful analysis of concentration bounds.

The guarantees derived in this paper are specific to the community setting, and we outlined previously the special properties of the community model when compared to a general multi-view model. However, when the documents of the topic model are sufficiently long, the word frequency vector within a document has good concentration, and our modified tensor method has better recovery guarantees in this setting as well. Thus, the improved tensor recovery guarantees derived in this paper are applicable in scenarios where we have access to better initialization vectors rather than simple random initialization.

5. Conclusion

In this paper, we presented a novel approach for learning overlapping communities based on a tensor decomposition approach. We established that our method is guaranteed to recover the underlying community memberships correctly, when the communities are drawn from a mixed membership stochastic block model (MMSB). Our method is also computationally efficient and requires simple linear algebraic operations and tensor iterations. Moreover, our method is tight for the special case of the stochastic block model (up to poly-log factors), both in terms of sample complexity and the separation between edge connectivity within a community and across different communities.

We now note a number of interesting open problems and extensions. While we obtained tight guarantees for MMSB models with uniform sized communities, our guarantees are weak when the community sizes are drastically different, such as in the planted clique setting where we do not match the computational lower bound (Feldman et al., 2012). The whitening step in the tensor decomposition method is particularly sensitive to the ratio of community sizes and it is interesting to see if modifications can be made to our algorithm to provide tight guarantees under unequal community sizes. While this paper mostly dealt with the theoretical analysis of the tensor method for community detection, we note recent experimental results where the tensor method is deployed on graphs with millions of nodes with very good accuracy and running times (Huang et al., 2013). In fact, the running times are more than an order of magnitude better than the state-of-art variational approach for learning MMSB models. The work of (Huang et al., 2013) makes an important modification to make the method scalable, viz., that the tensor decomposition is

carried out through stochastic updates in parallel unlike the serial batch updates considered here. Establishing theoretical guarantees for stochastic tensor decomposition is an important problem. Moreover, we have limited ourselves to the MMSB models, which assumes a linear model for edge formation, which is not applicable universally. For instance, exclusionary relationships, where two nodes cannot be connected because of their memberships in certain communities cannot be imposed in the MMSB model. Are there other classes of mixed membership models which do not suffer from this restriction, and yet are identifiable and are amenable for learning? Moreover, the Dirichlet distribution in the MMSB model imposes constraints on the memberships across different communities. Can we incorporate mixed memberships with arbitrary correlations? The answers to these questions will further push the boundaries of tractable learning of mixed membership communities models.

Acknowledgements

We thank the JMLR Action Editor Nathan Srebro and the anonymous reviewers for comments which significantly improved this manuscript. We thank Jure Leskovec for helpful discussions regarding various community models. Part of this work was done when AA, RG, and DH were at MSR New England. AA is supported in part by the Microsoft faculty fellowship, NSF Career award CCF-1254106, NSF Award CCF-1219234 and the ARO YIP Award W911NF-13-1-0084.

Appendix A. Tensor Power Method Analysis

In this section, we leverage on the perturbation analysis for tensor power method in Anandkumar et al. (2012b). As discussed in Section 3.3.3, we propose the following modifications to the tensor power method and obtain guarantees below for the modified method. The two main modifications are: (1) we modify the tensor deflation process in the robust power method in Procedure 2. Rather than a fixed deflation step after obtaining an estimate of the eigenvalue-eigenvector pair, in this paper, we deflate adaptively depending on the current estimate, and (2)rather than selecting random initialization vectors, as in Anandkumar et al. (2012b), we initialize with vectors obtained from adjacency matrix.

Below in Section A.1, we establish success of the modified tensor method under "good" initialization vectors, as defined below. This involves improved error bounds for the modified deflation procedure provided in Section A.2. In Section C.5, we subsequently establish that under the Dirichlet distribution (for small α_0), we obtain "good" initialization vectors.

A.1 Analysis under Good Initialization Vectors

We now show that when "good" initialization vectors are input to tensor power method in Procedure 2, we obtain good estimates of eigen-pairs under appropriate choice of number of iterations N and spectral norm ϵ of tensor perturbation.

Let $T = \sum_{i\in[k]} \lambda_i v_i$, where v_i are orthonormal vectors and $\lambda_1 \geq \lambda_2 \geq \ldots \lambda_k$. Let $\tilde{T} = T + E$ be the perturbed tensor with $\|E\| \leq \epsilon$. Recall that N denotes the number of iterations of the tensor power method.

We call an initialization vector u to be (γ, R_0)-good if there exists v_i such that $\langle u, v_i \rangle > R_0$ and

$$|\langle u, v_i \rangle| - \max_{j<i} |\langle u, v_j \rangle| > \gamma |\langle u, v_i \rangle|. \tag{53}$$

Choose $\gamma = 1/100$.

Theorem 11 *There exists universal constants $C_1, C_2 > 0$ such that the following holds.*

$$\epsilon \leq C_1 \cdot \lambda_{\min} R_0^2, \qquad N \geq C_2 \cdot \left(\log(k) + \log\log\left(\frac{\lambda_{\max}}{\epsilon} \right) \right), \tag{54}$$

Assume there is at least one good initialization vector corresponding to each v_i, $i \in [k]$. The parameter ξ for choosing deflation vectors in each iteration of the tensor power method in Procedure 2 is chosen as $\xi \geq 25\epsilon$. We obtain eigenvalue-eigenvector pairs $(\hat{\lambda}_1, \hat{v}_1), (\hat{\lambda}_2, \hat{v}_2), \ldots, (\hat{\lambda}_k, \hat{v}_k)$ such that there exists a permutation π on $[k]$ with

$$\|v_{\pi(j)} - \hat{v}_j\| \leq 8\epsilon/\lambda_{\pi(j)}, \qquad |\lambda_{\pi(j)} - \hat{\lambda}_j| \leq 5\epsilon, \quad \forall j \in [k],$$

and

$$\left\| T - \sum_{j=1}^{k} \hat{\lambda}_j \hat{v}_j^{\otimes 3} \right\| \leq 55\epsilon.$$

Remark: *(need for adaptive deflation) We now compare the above result with the result in (Anandkumar et al., 2012b, Thm. 5.1), where similar guarantees are obtained for a simpler version of the tensor power method without any adaptive deflation and using random initialization. The main difference is in our requirement of the gap γ in (53) for an initialization vector is weaker than the gap requirement in (Anandkumar et al., 2012b, Thm. 5.1). This is due to the use of adaptive deflation in this paper.*

Remark: *(need for non-random initialization) In this paper, we employ whitened neighborhood vectors generated under the MMSB model for initialization, while (Anandkumar et al., 2012b, Thm. 5.1) assumes a random initialization. Under random initialization, we obtain $R_0 \sim 1/\sqrt{k}$ (with poly(k) trials), while for initialization using whitened neighborhood vectors, we subsequently establish that $R_0 = \Omega(1)$ is a constant, when number of samples n is large enough. We also establish that the gap requirement in (53) is satisfied for the choice of $\gamma = 1/100$ above. See Lemma 25 for details. Thus, we can tolerate much larger perturbation ϵ of the third order moment tensor, when non-random initializations are employed.*

Proof: The proof is on lines of the proof of (Anandkumar et al., 2012b, Thm. 5.1) but here, we consider the modified deflation procedure, which improves the condition on ϵ in (54). We provide the full proof below for completeness.

We prove by induction on i, the number of eigenpairs estimated so far by Procedure 2. Assume that there exists a permutation π on $[k]$ such that the following assertions hold.

1. For all $j \leq i$, $\|v_{\pi(j)} - \hat{v}_j\| \leq 8\epsilon/\lambda_{\pi(j)}$ and $|\lambda_{\pi(j)} - \hat{\lambda}_j| \leq 12\epsilon$.
2. $D(u, i)$ is the set of deflated vectors given current estimate of the power method is $u \in S^{k-1}$:
$$D(u, i; \xi) := \{j : |\hat{\lambda}_i \hat{\theta}_i| \geq \xi\} \cap [i],$$
where $\hat{\theta}_i := \langle u, \hat{v}_i \rangle$.

3. The error tensor

$$\begin{aligned}\tilde{E}_{i+1,u} &:= \left(\hat{T} - \sum_{j \in D(u,i;\xi)} \hat{\lambda}_j \hat{v}_j^{\otimes 3}\right) - \sum_{j \notin D(u,i;\xi)} \lambda_{\pi(j)} v_{\pi(j)}^{\otimes 3} \\ &= E + \sum_{j \in D(u,i;\xi)} \left(\lambda_{\pi(j)} v_{\pi(j)}^{\otimes 3} - \hat{\lambda}_j \hat{v}_j^{\otimes 3}\right)\end{aligned}$$

satisfies

$$\|\tilde{E}_{i+1,u}(I,u,u)\| \leq 56\epsilon, \quad \forall u \in S^{k-1}; \tag{55}$$

$$\|\tilde{E}_{i+1,u}(I,u,u)\| \leq 2\epsilon, \quad \forall u \in S^{k-1} \text{ s.t. } \exists j \geq i+1 \centerdot (u^\top v_{\pi(j)})^2 \geq 1 - (168\epsilon/\lambda_{\pi(j)})^2. \tag{56}$$

We take $i = 0$ as the base case, so we can ignore the first assertion, and just observe that for $i = 0$, $D(u, 0; \xi) = \emptyset$ and thus

$$\tilde{E}_{1,u} = \hat{T} - \sum_{j=1}^{k} \lambda_i v_i^{\otimes 3} = E, \quad \forall\, u \in S^{k-1}.$$

We have $\|\tilde{E}_1\| = \|E\| = \epsilon$, and therefore the second assertion holds.

Now fix some $i \in [k]$, and assume as the inductive hypothesis. The power iterations now take a subset of $j \in [i]$ for deflation, depending on the current estimate. Set

$$C_1 := \min\left\{(56 \cdot 9 \cdot 102)^{-1}, (100 \cdot 168)^{-1}, \Delta' \text{ from Lemma 12 with } \Delta = 1/50\right\}. \tag{57}$$

For all good initialization vectors which are γ-separated relative to $\pi(j_{\max})$, we have (i) $|\theta_{j_{\max},0}^{(\tau)}| \geq R_0$, and (ii) that by (Anandkumar et al., 2012b, Lemma B.4) (using $\tilde{\epsilon}/p := 2\epsilon$, $\kappa := 1$, and $i^* := \pi(j_{\max})$, and providing C_2),

$$|\tilde{T}_i(\theta_N^{(\tau)}, \theta_N^{(\tau)}, \theta_N^{(\tau)}) - \lambda_{\pi(j_{\max})}| \leq 5\epsilon,$$

(notice by definition that $\gamma \geq 1/100$ implies $\gamma_0 \geq 1 - 1/(1+\gamma) \geq 1/101$, thus it follows from the bounds on the other quantities that $\tilde{\epsilon} = 2p\epsilon \leq 56C_1 \cdot \lambda_{\min} R_0^2 < \frac{\gamma_0}{2(1+8\kappa)} \cdot \tilde{\lambda}_{\min} \cdot \theta_{i^*,0}^2$ as necessary). Therefore $\theta_N := \theta_N^{(\tau^*)}$ must satisfy

$$\tilde{T}_i(\theta_N, \theta_N, \theta_N) = \max_{\tau \in [L]} \tilde{T}_i(\theta_N^{(\tau)}, \theta_N^{(\tau)}, \theta_N^{(\tau)}) \geq \max_{j \geq i} \lambda_{\pi(j)} - 5\epsilon = \lambda_{\pi(j_{\max})} - 5\epsilon.$$

On the other hand, by the triangle inequality,

$$\begin{aligned}\tilde{T}_i(\theta_N, \theta_N, \theta_N) &\leq \sum_{j \geq i} \lambda_{\pi(j)} \theta_{\pi(j),N}^3 + |\tilde{E}_i(\theta_N, \theta_N, \theta_N)| \\ &\leq \sum_{j \geq i} \lambda_{\pi(j)} |\theta_{\pi(j),N}| \theta_{\pi(j),N}^2 + 56\epsilon \\ &\leq \lambda_{\pi(j^*)} |\theta_{\pi(j^*),N}| + 56\epsilon,\end{aligned}$$

where $j^* := \arg\max_{j \geq i} \lambda_{\pi(j)} |\theta_{\pi(j),N}|$. Therefore

$$\lambda_{\pi(j^*)} |\theta_{\pi(j^*),N}| \geq \lambda_{\pi(j_{\max})} - 5\epsilon - 56\epsilon \geq \frac{4}{5} \lambda_{\pi(j_{\max})}.$$

Squaring both sides and using the fact that $\theta^2_{\pi(j^*),N} + \theta^2_{\pi(j),N} \leq 1$ for any $j \neq j^*$,

$$\begin{aligned}
\big(\lambda_{\pi(j^*)} \theta_{\pi(j^*),N}\big)^2 &\geq \frac{16}{25} \big(\lambda_{\pi(j_{\max})} \theta_{\pi(j^*),N}\big)^2 + \frac{16}{25} \big(\lambda_{\pi(j_{\max})} \theta_{\pi(j),N}\big)^2 \\
&\geq \frac{16}{25} \big(\lambda_{\pi(j^*)} \theta_{\pi(j^*),N}\big)^2 + \frac{16}{25} \big(\lambda_{\pi(j)} \theta_{\pi(j),N}\big)^2,
\end{aligned}$$

which in turn implies

$$\lambda_{\pi(j)} |\theta_{\pi(j),N}| \leq \frac{3}{4} \lambda_{\pi(j^*)} |\theta_{\pi(j^*),N}|, \quad j \neq j^*.$$

This means that θ_N is $(1/4)$-separated relative to $\pi(j^*)$. Also, observe that

$$|\theta_{\pi(j^*),N}| \geq \frac{4}{5} \cdot \frac{\lambda_{\pi(j_{\max})}}{\lambda_{\pi(j^*)}} \geq \frac{4}{5}, \quad \frac{\lambda_{\pi(j_{\max})}}{\lambda_{\pi(j^*)}} \leq \frac{5}{4}.$$

Therefore by (Anandkumar et al., 2012b, Lemma B.4) (using $\tilde{\epsilon}/p := 2\epsilon$, $\gamma := 1/4$, and $\kappa := 5/4$), executing another N power iterations starting from θ_N gives a vector $\hat{\theta}$ that satisfies

$$\|\hat{\theta} - v_{\pi(j^*)}\| \leq \frac{8\epsilon}{\lambda_{\pi(j^*)}}, \qquad |\hat{\lambda} - \lambda_{\pi(j^*)}| \leq 5\epsilon.$$

Since $\hat{v}_i = \hat{\theta}$ and $\hat{\lambda}_i = \hat{\lambda}$, the first assertion of the inductive hypothesis is satisfied, as we can modify the permutation π by swapping $\pi(i)$ and $\pi(j^*)$ without affecting the values of $\{\pi(j) : j \leq i-1\}$ (recall $j^* \geq i$).

We now argue that $\tilde{E}_{i+1,u}$ has the required properties to complete the inductive step. By Lemma 12 (using $\tilde{\epsilon} := 5\epsilon$, $\xi = 5\tilde{\epsilon} = 25\epsilon$ and $\Delta := 1/50$, the latter providing one upper bound on C_1 as per (57)), we have for any unit vector $u \in S^{k-1}$,

$$\left\| \left(\sum_{j \leq i} \Big(\lambda_{\pi(j)} v_{\pi(j)}^{\otimes 3} - \hat{\lambda}_j \hat{v}_j^{\otimes 3} \Big) \right) (I, u, u) \right\| \leq \left(1/50 + 100 \sum_{j=1}^{i} (u^\top v_{\pi(j)})^2 \right)^{1/2} 5\epsilon \leq 55\epsilon. \tag{58}$$

Therefore by the triangle inequality,

$$\|\tilde{E}_{i+1}(I, u, u)\| \leq \|E(I, u, u)\| + \left\| \left(\sum_{j \leq i} \Big(\lambda_{\pi(j)} v_{\pi(j)}^{\otimes 3} - \hat{\lambda}_j \hat{v}_j^{\otimes 3} \Big) \right) (I, u, u) \right\| \leq 56\epsilon.$$

Thus the bound (55) holds.

To prove that (56) holds, for any unit vector $u \in S^{k-1}$ such that there exists $j' \geq i+1$ with $(u^\top v_{\pi(j')})^2 \geq 1 - (168\epsilon/\lambda_{\pi(j')})^2$. We have (via the second bound on C_1 in (57) and the corresponding assumed bound $\epsilon \leq C_1 \cdot \lambda_{\min} R_0^2$)

$$100 \sum_{j=1}^{i} (u^\top v_{\pi(j)})^2 \leq 100 \Big(1 - (u^\top v_{\pi(j')})^2 \Big) \leq 100 \left(\frac{168\epsilon}{\lambda_{\pi(j')}} \right)^2 \leq \frac{1}{50},$$

and therefore

$$\left(1/50 + 100\sum_{j=1}^{i}(u^\top v_{\pi(j)})^2\right)^{1/2} 5\epsilon \leq (1/50+1/50)^{1/2}5\epsilon \leq \epsilon.$$

By the triangle inequality, we have $\|\tilde{E}_{i+1}(I,u,u)\| \leq 2\epsilon$. Therefore (56) holds, so the second assertion of the inductive hypothesis holds. We conclude that by the induction principle, there exists a permutation π such that two assertions hold for $i = k$. From the last induction step ($i = k$), it is also clear from (58) that $\|T - \sum_{j=1}^{k}\hat{\lambda}_j\hat{v}_j^{\otimes 3}\| \leq 55\epsilon$. This completes the proof of the theorem. ■

A.2 Deflation Analysis

Lemma 12 (Deflation analysis) *Let $\tilde{\epsilon} > 0$ and let $\{v_1, \ldots, v_k\}$ be an orthonormal basis for $\mathbb{R}^k$ and $\lambda_i \geq 0$ for $i \in [k]$. Let $\{\hat{v}_1, \ldots, \hat{v}_k\} \in \mathbb{R}^k$ be a set of unit vectors and $\hat{\lambda}_i \geq 0$. Define third order tensor $\mathcal{E}_i$ such that*

$$\mathcal{E}_i := \lambda_i v_i^{\otimes 3} - \hat{\lambda}_i \hat{v}_i^{\otimes 3}, \quad \forall\, i \in k.$$

For some $t \in [k]$ and a unit vector $u \in S^{k-1}$ such that $u = \sum_{i\in[k]}\theta_i v_i$ and $\hat{\theta}_i := \langle u, \hat{v}_i\rangle$, we have for $i \in [t]$,

$$\begin{aligned} |\hat{\lambda}_i\hat{\theta}_i| &\geq \xi \geq 5\tilde{\epsilon}, \\ |\hat{\lambda}_i - \lambda_i| &\leq \tilde{\epsilon}, \\ \|\hat{v}_i - v_i\| &\leq \min\{\sqrt{2},\ 2\tilde{\epsilon}/\lambda_i\}, \end{aligned}$$

then, the following holds

$$\left\|\sum_{i=1}^{t}\mathcal{E}_i(I,u,u)\right\|_2^2 \leq \left(4(5+11\tilde{\epsilon}/\lambda_{\min})^2 + 128(1+\tilde{\epsilon}/\lambda_{\min})^2(\tilde{\epsilon}/\lambda_{\min})^2\right)\tilde{\epsilon}^2\sum_{i=1}^{t}\theta_i^2 + 64(1+\tilde{\epsilon}/\lambda_{\min})^2\tilde{\epsilon}^2 + 2048(1+\tilde{\epsilon}/\lambda_{\min})^2\tilde{\epsilon}^2.$$

In particular, for any $\Delta \in (0,1)$, there exists a constant $\Delta' > 0$ (depending only on Δ) such that $\tilde{\epsilon} \leq \Delta'\lambda_{\min}$ implies

$$\left\|\sum_{i=1}^{t}\mathcal{E}_i(I,u,u)\right\|_2^2 \leq \left(\Delta + 100\sum_{i=1}^{t}\theta_i^2\right)\tilde{\epsilon}^2.$$

Proof: The proof is on lines of deflation analysis in (Anandkumar et al., 2012b, Lemma B.5), but we improve the bounds based on additional properties of vector u. From Anandkumar et al. (2012b), we have that for all $i \in [t]$, and any unit vector u,

$$\left\|\sum_{i=1}^{t}\mathcal{E}_i(I,u,u)\right\|_2^2 \leq \left(4(5+11\tilde{\epsilon}/\lambda_{\min})^2 + 128(1+\tilde{\epsilon}/\lambda_{\min})^2(\tilde{\epsilon}/\lambda_{\min})^2\right)\tilde{\epsilon}^2\sum_{i=1}^{t}\theta_i^2 + 64(1+\tilde{\epsilon}/\lambda_{\min})^2\tilde{\epsilon}^2\sum_{i=1}^{t}(\tilde{\epsilon}/\lambda_i)^2 + 2048(1+\tilde{\epsilon}/\lambda_{\min})^2\tilde{\epsilon}^2\left(\sum_{i=1}^{t}(\tilde{\epsilon}/\lambda_i)^3\right)^2. \tag{59}$$

Let $\hat{\lambda}_i = \lambda_i + \delta_i$ and $\hat{\theta}_i = \theta_i + \beta_i$. We have $\delta_i \leq \tilde{\epsilon}$ and $\beta_i \leq 2\tilde{\epsilon}/\lambda_i$, and that $|\hat{\lambda}_i \hat{\theta}_i| \geq \xi$.

$$\begin{aligned} ||\hat{\lambda}_i \hat{\theta}_i| - |\lambda_i \theta_i|| &\leq |\hat{\lambda}_i \hat{\theta}_i - \lambda_i \theta_i| \\ &\leq |(\lambda_i + \delta_i)(\theta_i + \beta_i) - \lambda_i \theta_i| \\ &\leq |\delta_i \theta_i + \lambda_i \beta_i + \delta_i \beta_i| \\ &\leq 4\tilde{\epsilon}. \end{aligned}$$

Thus, we have that $|\lambda_i \theta_i| \geq 5\tilde{\epsilon} - 4\tilde{\epsilon} = \tilde{\epsilon}$. Thus $\sum_{i=1}^{t} \tilde{\epsilon}^2/\lambda_i^2 \leq \sum_i \theta_i^2 \leq 1$. Substituting in (59), we have the result. ■

Appendix B. Proof of Theorem 9

We now prove the main results on error bounds claimed in Theorem 9 for the estimated community vectors $\hat{\Pi}$ and estimated block probability matrix $\hat{P}$ in Algorithm 1. Below, we first show that the tensor perturbation bounds claimed in Lemma 10 holds.

Let $\|T\|$ denote the spectral norm for a tensor T (or in special cases a matrix or a vector). Let $\|M\|_F$ denote the Frobenius norm. Let $|M_1|$ denote the operator ℓ_1 norm, i.e., the maximum ℓ_1 norm of its columns and $\|M\|_\infty$ denote the maximum ℓ_1 norm of its rows. Let $\kappa(M)$ denote the condition number, i.e., $\frac{\|M\|}{\sigma_{\min}(M)}$.

B.1 Proof of Lemma 10

From Theorem 11 in Appendix A, we see that the tensor power method returns eigenvalue-vector pair $(\hat{\lambda}_i, \hat{\Phi}_i)$ such that there exists a permutation θ with

$$\max_{i \in [k]} \|\hat{\Phi}_i - \Phi_{\theta(i)}\| \leq 8\widehat{\alpha}_{\max}^{1/2} \varepsilon_T, \tag{60}$$

and

$$\max_i |\lambda_i - \widehat{\alpha}_{\theta(i)}^{-1/2}| \leq 5\varepsilon_T, \tag{61}$$

when the perturbation of the tensor is small enough, according to

$$\varepsilon_T \leq C_1 \widehat{\alpha}_{\max}^{-1/2} r_0^2, \tag{62}$$

for some constant C_1, when initialized with a (γ, r_0) good vector.

With the above result, two aspects need to be established: (1) the whitened tensor perturbation ϵ_T is as claimed, (2) the condition in (62) is satisfied and (3) there exist good initialization vectors when whitened neighborhood vectors are employed. The tensor perturbation bound ϵ_T is established in Theorem 16 in Appendix C.1.

Lemma 25 establishes that when $\zeta = O(\sqrt{n} r_0^2/\rho)$, we have good initialization vectors with Recall $r_0^2 = \Omega(1/\widehat{\alpha}_{\max} k)$ when $\alpha_0 > 1$ and $r_0^2 = \Omega(1)$ for $\alpha_0 \leq 1$, and $\gamma = 1/100$ with probability $1 - 9\delta$ under Dirichlet distribution, when

$$n = \tilde{\Omega}\left(\alpha_{\min}^{-1} k^{0.43} \log(k/\delta)\right), \tag{63}$$

which is satisfied since we assume $\widehat{\alpha}_{\min}^{-2} < n$.

We now show that the condition in (62) is satisfied under the assumptions B1-B4. Since ϵ_T is given by

$$\varepsilon_T = \tilde{O}\left(\frac{\rho}{\sqrt{n}} \cdot \frac{\zeta}{\widehat{\alpha}_{\max}^{1/2}}\right),$$

the condition in (62) is equivalent to $\zeta = O(\sqrt{n} r_0^2/\rho)$. Therefore when $\zeta = O(\sqrt{n} r_0^2/\rho)$, the assumptions of Theorem 11 are satisfied.

B.2 Reconstruction of Π after Tensor Power Method

Let $(M)^i$ and $(M)_i$ denote the i^{th} row and i^{th} column in matrix M respectively. Let $Z \subseteq A^c$ denote any subset of nodes not in A, considered in Procedure LearnPartition Community. Define

$$\tilde{\Pi}_Z := \mathrm{Diag}(\lambda)^{-1} \Phi^\top \hat{W}_A^\top G_{Z,A}^\top. \tag{64}$$

Recall that the final estimate $\hat{\Pi}_Z$ is obtained by thresholding $\tilde{\Pi}_Z$ element-wise with threshold τ in Procedure 1. We first analyze perturbation of $\tilde{\Pi}_Z$.

Lemma 13 (Reconstruction Guarantees for $\tilde{\Pi}_Z$) *Assuming Lemma 10 holds and the tensor power method recovers eigenvectors and eigenvalues up to the guaranteed errors, we have with probability* $1 - 122\delta$,

$$\begin{aligned}\varepsilon_\pi := \max_{i \in Z} \|(\tilde{\Pi}_Z)^i - (\Pi_Z)^i\| &= O\left(\varepsilon_T \widehat{\alpha}_{\max}^{1/2} \left(\frac{\widehat{\alpha}_{\max}}{\widehat{\alpha}_{\min}}\right)^{1/2} \|\Pi_Z\|\right), \\ &= O\left(\rho \cdot \zeta \cdot \widehat{\alpha}_{\max}^{1/2} \left(\frac{\widehat{\alpha}_{\max}}{\widehat{\alpha}_{\min}}\right)^{1/2}\right)\end{aligned}$$

where ε_T *is given by* (72).

Proof: We have $(\tilde{\Pi}_Z)^i = \lambda_i^{-1}((\Phi)_i)^\top \hat{W}_A^\top G_{Z,A}^\top$. We will now use perturbation bounds for each of the terms to get the result.

The first term is

$$\begin{aligned}&\|\mathrm{Diag}(\lambda_i)^{-1} - \mathrm{Diag}(\widehat{\alpha}_i^{1/2})\| \cdot \|\mathrm{Diag}(\widehat{\alpha}^{1/2}) \tilde{F}_A^\top\| \cdot \|\tilde{F}_A\| \cdot \|\Pi_Z\| \\ &\leq 5\varepsilon_T \widehat{\alpha}_{\max} \widehat{\alpha}_{\min}^{-1/2} (1+\varepsilon_1)^2 \|\Pi_Z\|\end{aligned}$$

from the fact that $\|\mathrm{Diag}(\widehat{\alpha}^{1/2}) \tilde{F}_A^\top\| \leq 1 + \varepsilon_1$, where ε_1 is given by (87). The second term is

$$\begin{aligned}&\|\mathrm{Diag}(\widehat{\alpha}^{1/2})\| \cdot \|(\Phi)_i - \widehat{\alpha}_i^{1/2} (\tilde{F}_A)_i\| \cdot \|\tilde{F}_A\| \cdot \|\Pi_Z\| \\ &\leq 8\widehat{\alpha}_{\max} \varepsilon_T \widehat{\alpha}_{\min}^{-1/2} (1+\varepsilon_1) \|\Pi_Z\|\end{aligned}$$

The third term is

$$\begin{aligned}&\|\widehat{\alpha}_i^{1/2}\| \cdot \|(\hat{W}_A^\top - W_A^\top) F_A \Pi_Z\| \\ &\leq \widehat{\alpha}_{\max}^{1/2} \widehat{\alpha}_{\min}^{-1/2} \|\Pi_Z\| \epsilon_W &(65) \\ &\leq O\left(\left(\frac{\widehat{\alpha}_{\max}}{\widehat{\alpha}_{\min}}\right)^{1/2} \varepsilon_T \widehat{\alpha}_{\min}^{1/2} \|\Pi_Z\|\right), &(66)\end{aligned}$$

from Lemma 17 and finally, we have

$$\begin{aligned}
&\|\widehat{\alpha}_i^{1/2}\| \cdot \|W_A\| \cdot \|G_{Z,A}^\top - F_A \Pi_Z\| \\
&\leq O\left(\widehat{\alpha}_{\max}^{1/2} \frac{\sqrt{\alpha_0+1}}{\widehat{\alpha}_{\min}\sigma_{\min}(P)} \sqrt{(\max_i (P\widehat{\alpha})_i)(1+\varepsilon_2+\varepsilon_3)\log\frac{k}{\delta}}\right) \qquad (67)\\
&\leq O\left(\left(\frac{\widehat{\alpha}_{\max}}{\widehat{\alpha}_{\min}}\right)^{1/2} \varepsilon_T \sqrt{\alpha_0+1}(1+\varepsilon_2+\varepsilon_3)\sqrt{\frac{\log k}{\delta}}\right) \qquad (68)
\end{aligned}$$

from Lemma 22 and Lemma 23.

The third term in (66) dominates the last term in (68) since $(\alpha_0+1)\log k/\delta < n\widehat{\alpha}_{\min}$ (due to assumption B2 on scaling of n). ■

We now show that if we threshold the entries of $\tilde{\Pi}_Z$, the the resulting matrix $\hat{\Pi}_Z$ has rows close to those in Π_Z in ℓ_1 norm.

Lemma 14 (Guarantees after thresholding) *For $\hat{\Pi}_Z := \mathrm{Thres}(\tilde{\Pi}_Z, \tau)$, where τ is the threshold, we have with probability $1-2\delta$, that*

$$\begin{aligned}
\varepsilon_{\pi,\ell_1} := \max_{i\in[k]} |(\hat{\Pi}_Z)^i - (\Pi_Z)^i|_1 = O\Bigg(&\sqrt{n\eta}\,\varepsilon_\pi \sqrt{\log\frac{1}{2\tau}}\left(1-\sqrt{\frac{2\log(k/\delta)}{n\eta\log(1/2\tau)}}\right)\\
&+n\eta\tau + \sqrt{(n\eta+4\tau^2)\log\frac{k}{\delta}} + \frac{\varepsilon_\pi^2}{\tau}\Bigg),
\end{aligned}$$

where $\eta = \widehat{\alpha}_{\max}$ when $\alpha_0 < 1$ and $\eta = \alpha_{\max}$ when $\alpha_0 \in [1,k)$.

Remark: *The above guarantee on $\hat{\Pi}_Z$ is stronger than for $\tilde{\Pi}_Z$ in Lemma 13 since this is an ℓ_1 guarantee on the rows compared to ℓ_2 guarantee on rows for $\tilde{\Pi}_Z$.*
Remark: *When τ is chosen as*

$$\tau = \Theta(\frac{\varepsilon_\pi}{\sqrt{n\eta}}) = \Theta\left(\frac{\rho^{1/2}\cdot\zeta\cdot\widehat{\alpha}_{\max}^{1/2}}{n^{1/2}\cdot\widehat{\alpha}_{\min}}\right),$$

we have that

$$\begin{aligned}
\max_{i\in[k]} |(\hat{\Pi}_Z)^i - (\Pi_Z)^i|_1 &= \tilde{O}\left(\sqrt{n\eta}\cdot\varepsilon_\pi\right)\\
&= \tilde{O}\left(n^{1/2}\cdot\rho^{3/2}\cdot\zeta\cdot\widehat{\alpha}_{\max}\right)
\end{aligned}$$

Proof: Let $S_i := \{j : \hat{\Pi}_Z(i,j) > 2\tau\}$. For a vector v, let v_S denote the sub-vector by considering entries in set S. We now have

$$|(\hat{\Pi}_Z)^i - (\Pi_Z)^i|_1 \leq |(\hat{\Pi}_Z)^i_{S_i} - (\Pi_Z)^i_{S_i}|_1 + |(\Pi_Z)^i_{S_i^c}|_1 + |(\hat{\Pi}_Z)^i_{S_i^c}|_1$$

Case $\alpha_0 < 1$**:** From Lemma 26, we have $\mathbb{P}[\Pi(i,j) \geq 2\tau] \leq 8\widehat{\alpha}_i \log(1/2\tau)$. Since $\Pi(i,j)$ are independent for $j \in Z$, we have from multiplicative Chernoff bound (Kearns and Vazirani, 1994, Thm 9.2), that with probability $1-\delta$,

$$\max_{i\in[k]} |S_i| < 8n\widehat{\alpha}_{\max} \log\left(\frac{1}{2\tau}\right)\left(1 - \sqrt{\frac{2\log(k/\delta)}{n\widehat{\alpha}_i \log(1/2\tau)}}\right).$$

We have

$$|(\tilde{\Pi}_Z)^i_{S_i} - (\Pi_Z)^i_{S_i}|_1 \leq \varepsilon_\pi |S_i|^{1/2},$$

and the i^{th} rows of $\tilde{\Pi}_Z$ and $\hat{\Pi}_Z$ can differ on S_i, we have $|\tilde{\Pi}_Z(i,j) - \hat{\Pi}_Z(i,j)| \leq \tau$, for $j \in S_i$, and number of such terms is at most ε_π^2/τ^2. Thus,

$$|(\tilde{\Pi}_Z)^i_{S_i} - (\hat{\Pi}_Z)^i_{S_i}|_1 \leq \frac{\varepsilon_\pi^2}{\tau}.$$

For the other term, from Lemma 26, we have

$$\mathbb{E}[\Pi_Z(i,j) \cdot \delta(\Pi_Z(i,j) \leq 2\tau)] \leq \widehat{\alpha}_i(2\tau).$$

Applying Bernstein's bound we have with probability $1-\delta$

$$\max_{i\in[k]} \sum_{j\in Z} \Pi_Z(i,j) \cdot \delta(\Pi_Z(i,j) \leq 2\tau) \leq n\widehat{\alpha}_{\max}(2\tau) + \sqrt{2(n\widehat{\alpha}_{\max} + 4\tau^2)\log\frac{k}{\delta}}.$$

For $\hat{\Pi}^i_{S_i^c}$, we further divide S_i^c into T_i and U_i, where $T_i := \{j : \tau/2 < \Pi_Z(i,j) \leq 2\tau\}$ and $U_i := \{j : \Pi_Z(i,j) \leq \tau/2\}$.

In the set T_i, using similar argument we know $|(\Pi_Z)^i_{T_i} - (\tilde{\Pi}_Z)^i_{T_i}|_1 \leq O(\varepsilon_\pi\sqrt{n\widehat{\alpha}_{\max}\log 1/\tau})$, therefore

$$|\hat{\Pi}^i_{T_i}|_1 \leq |\tilde{\Pi}^i_{T_i}|_1 \leq |\Pi^i_{T_i} - \tilde{\Pi}^i_{T_i}|_1 + |\Pi^i_{S_i^c}|_1 \leq O(\varepsilon_\pi\sqrt{n\widehat{\alpha}_{\max}\log 1/\tau}).$$

Finally, for index $j \in U_i$, in order for $\hat{\Pi}_Z(i,j)$ be positive, it is required that $\tilde{\Pi}_Z(i,j) - \Pi_Z(i,j) \geq \tau/2$. In this case, we have

$$|(\hat{\Pi}_Z)^i_{U_i}|_1 \leq \frac{4}{\tau}\left\|(\tilde{\Pi}_Z)^i_{U_i} - \Pi^i_{U_i}\right\|^2 \leq \frac{4\varepsilon_\pi^2}{\tau}.$$

Case $\alpha_0 \in [1,k)$**:** From Lemma 26, we see that the results hold when we replace $\widehat{\alpha}_{\max}$ with $\alpha_{\max}$. ■

B.3 Reconstruction of P after Tensor Power Method

Finally we would like to use the community vectors Π and the adjacency matrix G to estimate the P matrix. Recall that in the generative model, we have $\mathbb{E}[G] = \Pi^\top P \Pi$. Thus, a straightforward estimate is to use $(\hat{\Pi}^\dagger)^\top G \hat{\Pi}^\dagger$. However, our guarantees on $\hat{\Pi}$ are not strong enough to control the error on $\hat{\Pi}^\dagger$ (since we only have row-wise ℓ_1 guarantees).

We propose an alternative estimator $\hat{Q}$ for $\hat{\Pi}^\dagger$ and use it to find $\hat{P}$ in Algorithm 1. Recall that the i-th row of $\hat{Q}$ is given by

$$\hat{Q}^i := (\alpha_0+1)\frac{\hat{\Pi}^i}{|\hat{\Pi}^i|_1} - \frac{\alpha_0}{n}\vec{1}^\top.$$

Define Q using exact communities, i.e.,

$$Q^i := (\alpha_0+1)\frac{\Pi^i}{|\Pi^i|_1} - \frac{\alpha_0}{n}\vec{1}^\top.$$

We show below that $\hat{Q}$ is close to $\Pi^\dagger$, and therefore, $\hat{P} := \hat{Q}^\top G \hat{Q}$ is close to P w.h.p.

Lemma 15 (Reconstruction of P) *With probability $1-5\delta$,*

$$\varepsilon_P := \max_{i,j\in[n]} |\hat{P}_{i,j} - P_{i,j}| \leq O\left(\frac{(\alpha_0+1)^{3/2}\varepsilon_\pi(P_{\max}-P_{\min})}{\sqrt{n}}\widehat{\alpha}_{\min}^{-1}\widehat{\alpha}_{\max}^{1/2}\log\frac{nk}{\delta}\right)$$

Remark: *If we define a new matrix Q' as $(Q')^i := \frac{\alpha_0+1}{n\widehat{\alpha}_i}\Pi^i - \frac{\alpha_0}{n}\vec{1}^\top$, then $\mathbb{E}_\Pi[Q'\Pi^\top] = I$. Below, we show that Q' is close to Q since $\mathbb{E}[|\Pi^i|_1] = n\widehat{\alpha}_i$ and thus the above result holds. We require Q to be normalized by $|\Pi^i|_1$ in order to ensure that the first term of Q has equal column norms, which will be used in our proofs subsequently.*

Proof: The proof goes in three steps:

$$P \approx Q\Pi^\top P\Pi Q^\top \approx QGQ^\top \approx \hat{Q}G\hat{Q}^\top.$$

Note that $\mathbb{E}_\Pi[\Pi Q^\top] = I$ and by Bernstein's bound, we can claim that $\Pi Q^\top$ is close to I and can show that the i-th row of $Q\Pi^\top$ satisfies

$$\Delta_i := |(Q\Pi^\top)^i - e_i^\top|_1 = O\left(k\sqrt{\log\left(\frac{nk}{\delta}\right)\frac{\widehat{\alpha}_{\max}}{\widehat{\alpha}_{\min}}}\frac{1}{\sqrt{n}}\right)$$

with probability $1-\delta$. Moreover,

$$\begin{aligned}|(\Pi^\top P\Pi Q^\top)_{i,j} - (\Pi^\top P)_{i,j}| &\leq |(\Pi^\top P)^i((Q)_j - e_j)| = |(\Pi^\top P)^i\Delta_j| \\ &\leq O\left(\frac{P_{\max}k\cdot\sqrt{\widehat{\alpha}_{\max}/\widehat{\alpha}_{\min}}}{\sqrt{n}}\sqrt{\log\frac{nk}{\delta}}\right).\end{aligned}$$

using the fact that $(\Pi^\top P)_{i,j} \leq P_{\max}$.

Now we claim that $\hat{Q}$ is close to Q and it can be shown that

$$|Q^i - \hat{Q}^i|_1 \leq O\left(\frac{\varepsilon_P}{P_{\max}-P_{\min}}\right) \tag{69}$$

Using (69), we have

$$\begin{aligned}|(\Pi^\top P\Pi Q^\top)_{i,j} - (\Pi^\top P\Pi\hat{Q}^\top)_{i,j}| &= |(\Pi^\top P\Pi)^i(Q^\top - \hat{Q}^\top)_j| \\ &= ((\Pi^\top P\Pi)^i - P_{\min}\vec{1}^\top)|(Q^\top - \hat{Q}^\top)_j|_1 \\ &\leq O((P_{\max}-P_{\min})|(Q^\top - \hat{Q}^\top)_j|_1) = O(\varepsilon_P).\end{aligned}$$

using the fact that $(Q^j - \hat{Q}^j)\vec{1} = 0$, due to the normalization.

Finally, $|(G\hat{Q}^\top)_{i,j}(\Pi^\top P \Pi \hat{Q}^\top)_{i,j}|$ are small by standard concentration bounds (and the differences are of lower order). Combining these $|\hat{P}_{i,j} - P_{i,j}| \leq O(\varepsilon_P)$.

■

B.4 Zero-error Support Recovery Guarantees

Recall that we proposed Procedure 3 to provide improved support recovery estimates in the special case of homophilic models (where there are more edges within a community than to any community outside). We limit our analysis to the special case of uniform sized communities ($\alpha_i = 1/k$) and matrix P such that $P(i,j) = p\mathbb{I}(i=j) + q\mathbb{I}(i \neq j)$ and $p \geq q$. In principle, the analysis can be extended to homophilic models with more general P matrix (with suitably chosen thresholds for support recovery).

We first consider analysis for the stochastic block model (i.e., $\alpha_0 \to 0$) and prove the guarantees claimed in Corollary 8.

Proof of Corollary 8: Recall the definition of $\tilde{\Pi}$ in (64) and $\hat{\Pi}$ is obtained by thresholding $\tilde{\Pi}$ with threshold τ. Since the threshold τ for stochastic block models is 0.5 (assumption B5), we have

$$|(\hat{\Pi})^i - (\Pi)^i|_1 = O(\varepsilon_\pi^2), \tag{70}$$

where ε_π is the row-wise ℓ_2 error for $\tilde{\Pi}$ in Lemma 13. This is because $\Pi(i,j) \in \{0,1\}$, and in order for our method to make a mistake, it takes $1/4$ in the ℓ_2^2 error.

In Procedure 3, for the stochastic block model ($\alpha_0 = 0$), for a node $x \in [n]$, we have

$$\hat{F}(x,i) = \sum_{y \in [n]} G_{x,y} \frac{\hat{\Pi}(i,y)}{|\hat{\Pi}^i|_1} \approx \sum_{y \in [n]} G_{x,y} \frac{\hat{\Pi}(i,y)}{|\Pi^i|_1} \approx \frac{k}{n} \sum_{y \in [n]} G_{x,y} \hat{\Pi}(i,y),$$

using (70) and the fact that the size of each community on average is n/k. In other words, for each vertex x, we compute the average number of edges from this vertex to all the estimated communities according to $\hat{\Pi}$, and set it to belong to the one with largest average degree. Note that the margin of error on average for each node to be assigned the correct community according to the above procedure is $(p-q)n/k$, since the size of each community is n/k and the average number of intra-community edges at a node is pn/k and edges to any different community at a node is qn/k. From (70), we have that the average number of errors made is $O((p-q)\varepsilon_\pi^2)$. Note that the degrees concentrate around their expectations according to Bernstein's bound and the fact that the edges used for averaging is independent from the edges used for estimating $\hat{\Pi}$. Thus, for our method to succeed in inferring the correct community at a node, we require,

$$O((p-q)\varepsilon_\pi^2) \leq (p-q)\frac{n}{k},$$

which implies

$$p - q \geq \tilde{\Omega}\left(\frac{\sqrt{p}k}{\sqrt{n}}\right).$$

■

We now prove the general result on support recovery.

Proof of Theorem 7: From Lemma 15,

$$|\hat{P}_{i,j} - P_{i,j}| \leq O(\varepsilon_P)$$

which implies bounds for the average of diagonals H and average of off-diagonals L:

$$|H - p| = O(\varepsilon_P), \quad |L - q| = O(\varepsilon_P).$$

On similar lines as the proof of Lemma 15 and from independence of edges used to define $\hat{F}$ from the edges used to estimate $\hat{\Pi}$, we also have

$$|\hat{F}(j,i) - F(j,i)| \leq O(\varepsilon_P).$$

Note that $F_{j,i} = q + \Pi_{i,j}(p-q)$. The threshold ξ satisfies $\xi = \Omega(\varepsilon_P)$, therefore, all the entries in F that are larger than $q + (p-q)\xi$, the corresponding entries in S are declared to be one, while none of the entries that are smaller than $q + (p-q)\xi/2$ are set to one in S. ■

Appendix C. Concentration Bounds

In this section we prove concentration bounds for the tensors and matrices appeared in the algorithm.

C.1 Main Result: Tensor Perturbation Bound

We now provide the main result that the third-order whitened tensor computed from samples concentrates. Recall that $\mathrm{T}^{\alpha_0}_{Y\to\{A,B,C\}}$ denotes the third order moment computed using edges from partition Y to partitions A, B, C in (15). $\hat{W}_A, \hat{W}_B\hat{R}_{AB}, \hat{W}_C\hat{R}_{AC}$ are the whitening matrices defined in (24). The corresponding whitening matrices $W_A, W_B R_{AB}, W_C R_{AC}$ for exact moment third order tensor $\mathbb{E}[\mathrm{T}^{\alpha_0}_{Y\to\{A,B,C\}}\,|\Pi]$ will be defined later. Recall that ρ is defined in (38) as $\rho := \frac{\alpha_0+1}{\widehat{\alpha}_{\min}}$. Given $\delta \in (0,1)$, throughout assume that

$$n = \Omega\left(\rho^2 \log^2 \frac{k}{\delta}\right), \tag{71}$$

as in Assumption $(B2)$.

Theorem 16 (Perturbation of whitened tensor) *When the partitions A, B, C, X, Y satisfy* (71)*, we have with probability $1 - 100\delta$,*

$$\begin{aligned}
\varepsilon_T &:= \left\| \mathrm{T}^{\alpha_0}_{Y\to\{A,B,C\}}(\hat{W}_A, \hat{W}_B\hat{R}_{AB}, \hat{W}_C\hat{R}_{AC}) - \mathbb{E}[\mathrm{T}^{\alpha_0}_{Y\to\{A,B,C\}}(W_A, \tilde{W}_B, \tilde{W}_C)|\Pi_A, \Pi_B, \Pi_C] \right\| \\
&= O\left(\frac{(\alpha_0+1)\sqrt{(\max_i (P\widehat{\alpha})_i)}}{n^{1/2}\widehat{\alpha}^{3/2}_{\min}\sigma_{\min}(P)} \cdot \left(1 + \left(\frac{\rho^2}{n}\log^2\frac{k}{\delta}\right)^{1/4}\right)\sqrt{\frac{\log k}{\delta}} \right) \\
&= \tilde{O}\left(\frac{\rho}{\sqrt{n}} \cdot \frac{\zeta}{\widehat{\alpha}^{1/2}_{\max}} \right).
\end{aligned} \tag{72}$$

C.1.1 Proof Overview

The proof of the above result follows. It consists mainly of the following steps: (1) Controlling the perturbations of the whitening matrices and (2) Establishing concentration of the third moment tensor (before whitening). Combining the two, we can then obtain perturbation of the whitened tensor. Perturbations for the whitening step is established in Appendix C.2. Auxiliary concentration bounds required for the whitening step, and for the claims below are in Appendix C.3 and C.4.

Proof of Theorem 16: In tensor T^{α_0} in (15), the first term is

$$(\alpha_0+1)(\alpha_0+2)\sum_{i\in Y}\left(G_{i,A}^\top \otimes G_{i,B}^\top \otimes G_{i,C}^\top\right).$$

We claim that this term dominates in the perturbation analysis since the mean vector perturbation is of lower order. We now consider perturbation of the whitened tensor

$$\Lambda_0 = \frac{1}{|Y|}\sum_{i\in Y}\left((\hat{W}_A^\top G_{i,A}^\top)\otimes(\hat{R}_{AB}^\top \hat{W}_B^\top G_{i,B}^\top)\otimes(\hat{R}_{AC}^\top \hat{W}_C^\top G_{i,C}^\top)\right).$$

We show that this tensor is close to the corresponding term in the expectation in three steps.

First we show it is close to

$$\Lambda_1 = \frac{1}{|Y|}\sum_{i\in Y}\left((\hat{W}_A^\top F_A\pi_i)\otimes(\hat{R}_{AB}^\top \hat{W}_B^\top F_B\pi_i)\otimes(\hat{R}_{AC}^\top \hat{W}_C^\top F_C\pi_i)\right).$$

Then this vector is close to the expectation over Π_Y.

$$\Lambda_2 = \mathbb{E}_{\pi\sim\mathrm{Dir}(\alpha)}\left((\hat{W}_A^\top F_A\pi)\otimes(\hat{R}_{AB}^\top \hat{W}_B^\top F_B\pi)\otimes(\hat{R}_{AC}^\top \hat{W}_C^\top F_C\pi)\right).$$

Finally we replace the estimated whitening matrix $\hat{W}_A$ with W_A, defined in (73), and note that W_A whitens the exact moments.

$$\Lambda_3 = \mathbb{E}_{\pi\sim\mathrm{Dir}(\alpha)}\left((W_A^\top F_A\pi)\otimes(\tilde{W}_B^\top F_B\pi)\otimes(\tilde{W}_C^\top F_C\pi)\right).$$

For $\Lambda_0-\Lambda_1$, the dominant term in the perturbation bound (assuming partitions A,B,C,X,Y are of size n) is (since for any rank 1 tensor, $\|u\otimes v\otimes w\| = \|u\|\cdot\|v\|\cdot\|w\|$),

$$O\left(\frac{1}{|Y|}\|\tilde{W}_B^\top F_B\|^2\left\|\sum_{i\in Y}\left(\hat{W}_A^\top G_{i,A}^\top - \hat{W}_A^\top F_A\pi_i\right)\right\|\right)$$
$$O\left(\frac{1}{|Y|}\widehat{\alpha}_{\min}^{-1}\cdot\frac{(\alpha_0+1)(\max_i(P\widehat{\alpha})_i)}{\widehat{\alpha}_{\min}\sigma_{\min}(P)}\cdot(1+\varepsilon_1+\varepsilon_2+\varepsilon_3)\sqrt{\log\frac{n}{\delta}}\right),$$

with probability $1-13\delta$ (Lemma 18). Since there are 7 terms in the third order tensor T^{α_0}, we have the bound with probability $1-91\delta$.

For $\Lambda_1 - \Lambda_2$, since $\hat{W}_A F_A \operatorname{Diag}(\widehat{\alpha})^{1/2}$ has spectral norm almost 1, by Lemma 20 the spectral norm of the perturbation is at most

$$\left\|\hat{W}_A F_A \operatorname{Diag}(\widehat{\alpha})^{1/2}\right\|^3 \left\|\frac{1}{|Y|}\sum_{i\in Y}(\operatorname{Diag}(\widehat{\alpha})^{-1/2}\pi_i)^{\otimes 3} - \mathbb{E}_{\pi\sim\mathrm{Dir}(\alpha)}(\operatorname{Diag}(\widehat{\alpha})^{-1/2}\pi_i)^{\otimes 3}\right\|$$
$$\leq O\left(\frac{1}{\widehat{\alpha}_{\min}\sqrt{n}}\cdot\sqrt{\log\frac{n}{\delta}}\right).$$

For the final term $\Lambda_2 - \Lambda_3$, the dominating term is

$$(\hat{W}_A - W_A)F_A \operatorname{Diag}(\widehat{\alpha})^{1/2}\|\Lambda_3\| \leq \varepsilon_{W_A}\|\Lambda_3\|$$
$$\leq O\left(\frac{(\alpha_0+1)\sqrt{\max_i(P\widehat{\alpha})_i}}{n^{1/2}\widehat{\alpha}_{\min}^{3/2}\sigma_{\min}(P)}(1+\varepsilon_1+\varepsilon_2+\varepsilon_3)\sqrt{\log\frac{n}{\delta}}\right).$$

Putting all these together, the third term $\|\Lambda_2 - \Lambda_3\|$ dominates. We know with probability at least $1 - 100\delta$, the perturbation in the tensor is at most

$$O\left(\frac{(\alpha_0+1)\sqrt{\max_i(P\widehat{\alpha})_i}}{n^{1/2}\widehat{\alpha}_{\min}^{3/2}\sigma_{\min}(P)}(1+\varepsilon_1+\varepsilon_2+\varepsilon_3)\sqrt{\log\frac{n}{\delta}}\right).$$

■

C.2 Whitening Matrix Perturbations

Consider rank-k SVD of $|X|^{-1/2}(G^{\alpha_0}_{X,A})^\top_{k-svd} = \hat{U}_A\hat{D}_A\hat{V}_A^\top$, and the whitening matrix is given by $\hat{W}_A := \hat{U}_A\hat{D}_A^{-1}$ and thus $|X|^{-1}\hat{W}_A^\top(G^{\alpha_0}_{X,A})^\top_{k-svd}(G^{\alpha_0}_{X,A})_{k-svd}\hat{W}_A = I$. Now consider the singular value decomposition of

$$|X|^{-1}\hat{W}_A^\top\mathbb{E}[(G^{\alpha_0}_{X,A})^\top|\Pi]\cdot\mathbb{E}[(G^{\alpha_0}_{X,A})|\Pi]\hat{W}_A = \Phi\tilde{D}\Phi^\top.$$

$\hat{W}_A$ does not whiten the exact moments in general. On the other hand, consider

$$W_A := \hat{W}_A\Phi_A\tilde{D}_A^{-1/2}\Phi_A^\top. \tag{73}$$

Observe that W_A whitens $|X|^{-1/2}\mathbb{E}[(G^{\alpha_0}_{X,A})|\Pi]$

$$|X|^{-1}W_A^\top\mathbb{E}[(G^{\alpha_0}_{X,A})^\top|\Pi]\mathbb{E}[(G^{\alpha_0}_{X,A})|\Pi]W_A = (\Phi_A\tilde{D}_A^{-1/2}\Phi_A^\top)^\top\Phi_A\tilde{D}_A\Phi_A^\top\Phi_A\tilde{D}_A^{-1/2}\Phi_A^\top = I$$

Now the ranges of W_A and $\hat{W}_A$ may differ and we control the perturbations below.

Also note that $\hat{R}_{A,B}$, $\hat{R}_{A,C}$ are given by

$$\hat{R}_{AB} := |X|^{-1}\hat{W}_B^\top(G^{\alpha_0}_{X,B})^\top_{k-svd}(G^{\alpha_0}_{X,A})_{k-svd}\hat{W}_A. \tag{74}$$
$$R_{AB} := |X|^{-1}W_B^\top\mathbb{E}[(G^{\alpha_0}_{X,B})^\top|\Pi]\cdot\mathbb{E}[G^{\alpha_0}_{X,A}|\Pi]\cdot W_A. \tag{75}$$

Recall ϵ_G is given by (80), and $\sigma_{\min}\left(\mathbb{E}[G^{\alpha_0}_{X,A}|\Pi]\right)$ is given in (23) and $|A| = |B| = |X| = n$.

Lemma 17 (Whitening matrix perturbations) *With probability* $1-\delta$,

$$\epsilon_{W_A} := \|\operatorname{Diag}(\widehat{\alpha})^{1/2} F_A^\top (\hat{W}_A - W_A)\| = O\left(\frac{(1-\varepsilon_1)^{-1/2}\epsilon_G}{\sigma_{\min}\left(\mathbb{E}[G^{\alpha_0}_{X,A}|\Pi]\right)}\right) \tag{76}$$

$$\epsilon_{\tilde{W}_B} := \|\operatorname{Diag}(\widehat{\alpha})^{1/2} F_B^\top (\hat{W}_B \hat{R}_{AB} - W_B R_{AB})\| = O\left(\frac{(1-\varepsilon_1)^{-1/2}\epsilon_G}{\sigma_{\min}\left(\mathbb{E}[G^{\alpha_0}_{X,B}|\Pi]\right)}\right) \tag{77}$$

Thus, with probability $1-6\delta$,

$$\epsilon_{W_A} = \epsilon_{\tilde{W}_B} = O\left(\frac{(\alpha_0+1)\sqrt{\max_i (P\widehat{\alpha})_i}}{n^{1/2}\widehat{\alpha}_{\min}\sigma_{\min}(P)} \cdot (1+\varepsilon_1+\varepsilon_2+\varepsilon_3)\right), \tag{78}$$

where ε_1, ε_2 *and* ε_3 *are given by* (86) *and* (87).

Remark: *Note that when partitions* X, A *satisfy* (71), $\varepsilon_1, \varepsilon_2, \varepsilon_3$ *are small. When* P *is well conditioned and* $\widehat{\alpha}_{\min} = \widehat{\alpha}_{\max} = 1/k$, *we have* $\epsilon_{W_A}, \epsilon_{\tilde{W}_B} = O(k/\sqrt{n})$.

Proof: Using the fact that $W_A = \hat{W}_A \Phi_A \tilde{D}_A^{-1/2} \Phi_A^\top$ or $\hat{W}_A = W_A \Phi_A \tilde{D}_A^{1/2} \Phi_A^\top$ we have that

$$\begin{aligned}
\|\operatorname{Diag}(\widehat{\alpha})^{1/2} F_A^\top (\hat{W}_A - W_A)\| &\le \|\operatorname{Diag}(\widehat{\alpha})^{1/2} F_A^\top W_A (I - \Phi_A \tilde{D}_A^{1/2} \Phi_A^\top)\| \\
&= \|\operatorname{Diag}(\widehat{\alpha})^{1/2} F_A^\top W_A (I - \tilde{D}_A^{1/2})\| \\
&\le \|\operatorname{Diag}(\widehat{\alpha})^{1/2} F_A^\top W_A (I - \tilde{D}_A^{1/2})(I + \tilde{D}_A^{1/2})\| \\
&\le \|\operatorname{Diag}(\widehat{\alpha})^{1/2} F_A^\top W_A\| \cdot \|I - \tilde{D}_A\|
\end{aligned}$$

using the fact that $\tilde{D}_A$ is a diagonal matrix.

Now note that W_A whitens $|X|^{-1/2}\mathbb{E}[G^{\alpha_0}_{X,A}|\Pi] = |X|^{-1/2} F_A \operatorname{Diag}(\alpha^{1/2}) \Psi_X$, where Ψ_X is defined in (85). Further it is shown in Lemma 23 that Ψ_X satisfies with probability $1-\delta$ that

$$\varepsilon_1 := \|I - |X|^{-1} \Psi_X \Psi_X^\top\| \le O\left(\sqrt{\frac{(\alpha_0+1)}{\widehat{\alpha}_{\min}|X|}} \cdot \log\frac{k}{\delta}\right)$$

Since $\varepsilon_1 \ll 1$ when X, A satisfy (71). We have that $|X|^{-1/2}\Psi_X$ has singular values around 1. Since W_A whitens $|X|^{-1/2}\mathbb{E}[G^{\alpha_0}_{X,A}|\Pi]$, we have

$$|X|^{-1} W_A^\top F_A \operatorname{Diag}(\alpha^{1/2}) \Psi_X \Psi_X^\top \operatorname{Diag}(\alpha^{1/2}) F_A^\top W_A = I.$$

Thus, with probability $1-\delta$,

$$\|\operatorname{Diag}(\widehat{\alpha})^{1/2} F_A^\top W_A\| = O((1-\varepsilon_1)^{-1/2}).$$

Let $\mathbb{E}[(G_{X,A}^{\alpha_0})|\Pi] = (G_{X,A}^{\alpha_0})_{k-svd} + \Delta$. We have

$$\begin{aligned}
\|I - \tilde{D}_A\| &= \|I - \Phi_A \tilde{D}_A \Phi_A^\top\| \\
&= \|I - |X|^{-1} \hat{W}_A^\top \mathbb{E}[(G_{X,A}^{\alpha_0})^\top|\Pi] \cdot \mathbb{E}[(G_{X,A}^{\alpha_0})|\Pi] \hat{W}_A\| \\
&= O\left(|X|^{-1} \|\hat{W}_A^\top \left(\Delta^\top (G_{X,A}^{\alpha_0})_{k-svd} + \Delta (G_{X,A}^{\alpha_0})_{k-svd}^\top\right) \hat{W}_A\|\right) \\
&= O\left(|X|^{-1/2} \|\hat{W}_A^\top \Delta^\top \hat{V}_A + \hat{V}_A^\top \Delta \hat{W}_A\|\right), \\
&= O\left(|X|^{-1/2} \|\hat{W}_A\| \|\Delta\|\right) \\
&= O\left(|X|^{-1/2} \|W_A\| \epsilon_G\right),
\end{aligned}$$

since $\|\Delta\| \leq \epsilon_G + \sigma_{k+1}(G_{X,A}^{\alpha_0}) \leq 2\epsilon_G$, using Weyl's theorem for singular value perturbation and the fact that $\epsilon_G \cdot \|W_A\| \ll 1$ and $\|W_A\| = |X|^{1/2}/\sigma_{\min}\left(\mathbb{E}[G_{X,A}^{\alpha_0}|\Pi]\right)$.

We now consider perturbation of $W_B R_{AB}$. By definition, we have that

$$\mathbb{E}[G_{X,B}^{\alpha_0}|\Pi] \cdot W_B R_{AB} = \mathbb{E}[G_{X,A}^{\alpha_0}|\Pi] \cdot W_A.$$

and

$$\|W_B R_{AB}\| = |X|^{1/2} \sigma_{\min}(\mathbb{E}[G_{X,B}^{\alpha_0}|\Pi])^{-1}.$$

Along the lines of previous derivation for ϵ_{W_A}, let

$$|X|^{-1} (\hat{W}_B \hat{R}_{AB})^\top \cdot \mathbb{E}[(G_{X,B}^{\alpha_0})^\top|\Pi] \cdot \mathbb{E}[G_{X,B}^{\alpha_0}|\Pi] \hat{W}_B \hat{R}_{AB} = \Phi_B \tilde{D}_B \Phi_B^\top.$$

Again using the fact that $|X|^{-1}\Psi_X \Psi_X^\top \approx I$, we have

$$\|\operatorname{Diag}(\widehat{\alpha})^{1/2} F_B^\top W_B R_{AB}\| \approx \|\operatorname{Diag}(\widehat{\alpha})^{1/2} F_A^\top W_A\|,$$

and the rest of the proof follows. ■

C.3 Auxiliary Concentration Bounds

Lemma 18 (Concentration of sum of whitened vectors) *Assuming all the partitions satisfy* (71)*, with probability* $1 - 7\delta$*,*

$$\begin{aligned}
\left\|\sum_{i \in Y} \left(\hat{W}_A^\top G_{i,A}^\top - \hat{W}_A^\top F_A \pi_i\right)\right\| &= O(\sqrt{|Y| \widehat{\alpha}_{\max}} \epsilon_{W_A}) \\
&= O\left(\frac{\sqrt{(\alpha_0+1)(\max_i (P\widehat{\alpha})_i)}}{\widehat{\alpha}_{\min} \sigma_{\min}(P)} \cdot (1 + \varepsilon_2 + \varepsilon_3) \sqrt{\log n/\delta}\right), \\
\left\|\sum_{i \in Y} \left((\hat{W}_B \hat{R}_{AB})^\top (G_{i,B}^\top - F_B \pi_i)\right)\right\| &= O\left(\frac{\sqrt{(\alpha_0+1)(\max_i (P\widehat{\alpha})_i)}}{\widehat{\alpha}_{\min} \sigma_{\min}(P)} \cdot (1 + \varepsilon_1 + \varepsilon_2 + \varepsilon_3) \sqrt{\log n/\delta}\right).
\end{aligned}$$

Remark: *Note that when P is well conditioned and $\widehat{\alpha}_{\min} = \widehat{\alpha}_{\max} = 1/k$, we have the above bounds as $O(k)$. Thus, when it is normalized with $1/|Y| = 1/n$, we have the bound as $O(k/n)$.*

Proof: Note that $\hat{W}_A$ is computed using partition X and $G_{i,A}$ is obtained from $i \in Y$. We have independence for edges across different partitions X and Y. Let $\Xi_i := \hat{W}_A^\top (G_{i,A}^\top - F_A \pi_i)$.Applying matrix Bernstein's inequality to each of the variables, we have

$$\begin{aligned}\|\Xi_i\| &\leq \|\hat{W}_A\| \cdot \|G_{i,A}^\top - F_A \pi_i\| \\ &\leq \|\hat{W}_A\| \sqrt{\|F_A\|_1},\end{aligned}$$

from Lemma 22. The variances are given by

$$\begin{aligned}\|\sum_{i\in Y} \mathbb{E}[\Xi_i \Xi_i^\top | \Pi]\| &\leq \sum_{i\in Y} \hat{W}_A^\top \operatorname{Diag}(F_A \pi_i) \hat{W}_A, \\ &\leq \|\hat{W}_A\|^2 \|F_Y\|_1 \\ &= O\left(\frac{|Y|}{|A|} \cdot \frac{(\alpha_0+1)(\max_i (P\widehat{\alpha})_i)}{\widehat{\alpha}_{\min}^2 \sigma_{\min}^2(P)} \cdot (1+\varepsilon_2+\varepsilon_3) \right),\end{aligned}$$

with probability $1-2\delta$ from (83) and (84), and $\varepsilon_2, \varepsilon_3$ are given by (87). Similarly, $\|\sum_{i\in Y} \mathbb{E}[\Xi_i^\top \Xi_i | \Pi]\| \leq \|\hat{W}_A\|^2 \|F_Y\|_1$. Thus, from matrix Bernstein's inequality, we have with probability $1-3\delta$

$$\begin{aligned}\|\sum_{i\in Y} \Xi_i\| &= O(\|\hat{W}_A\| \sqrt{\max(\|F_A\|_1, \|F_X\|_1)}). \\ &= O\left(\frac{\sqrt{(\alpha_0+1)(\max_i (P\widehat{\alpha})_i)}}{\widehat{\alpha}_{\min} \sigma_{\min}(P)} \cdot (1+\varepsilon_2+\varepsilon_3) \sqrt{\log n/\delta} \right)\end{aligned}$$

On similar lines, we have the result for B and C, and also use the independence assumption on edges in various partitions. ■

We now show that not only the sum of whitened vectors concentrates, but that each individual whitened vector $\hat{W}_A^\top G_{i,A}^\top$ concentrates, when A is large enough.

Lemma 19 (Concentration of a random whitened vector) *Conditioned on π_i, with probability at least $1/4$,*

$$\left\| \hat{W}_A^\top G_{i,A}^\top - W_A^\top F_A \pi_i \right\| \leq O(\varepsilon_{W_A} \widehat{\alpha}_{\min}^{-1/2}) = \tilde{O}\left(\frac{\sqrt{(\alpha_0+1)(\max_i (P\widehat{\alpha})_i)}}{n^{1/2} \widehat{\alpha}_{\min}^{3/2} \sigma_{\min}(P)} \right).$$

Remark: *The above result is not a high probability event since we employ Chebyshev's inequality to establish it. However, this is not an issue for us, since we will employ it to show that out of $\Theta(n)$ whitened vectors, there exists at least one good initialization vector corresponding to each eigen-direction, as required in Theorem 11 in Appendix A. See Lemma 25 for details.*

Proof We have

$$\left\| \hat{W}_A^\top G_{i,A}^\top - W_A^\top F_A \pi_i \right\| \leq \left\| (\hat{W}_A - W_A)^\top F_A \pi_i \right\| + \left\| \hat{W}_A^\top (G_{i,A}^\top - F_A \pi_i) \right\|.$$

The first term is satisfies satisfies with probability $1 - 3\delta$

$$\begin{aligned}\|(\hat{W}_A^\top - W_A^\top)F_A\pi_i\| &\leq \epsilon_{W_A}\widehat{\alpha}_{min}^{-1/2} \\ &= O\left(\frac{(\alpha_0+1)\widehat{\alpha}_{\max}^{1/2}\sqrt{(\max_i(P\widehat{\alpha})_i)}}{n^{1/2}\widehat{\alpha}_{\min}^{3/2}\sigma_{\min}(P)} \cdot (1+\varepsilon_1+\varepsilon_2+\varepsilon_3)\right)\end{aligned}$$

Now we bound the second term. Note that $G_{i,A}^\top$ is independent of $\hat{W}_A^\top$, since they are related to disjoint subset of edges. The whitened neighborhood vector can be viewed as a sum of vectors:

$$\hat{W}_A^\top G_{i,A}^\top = \sum_{j\in A} G_{i,j}(\hat{W}_A^\top)_j = \sum_{j\in A} G_{i,j}(\hat{D}_A\hat{U}_A^\top)_j = \hat{D}_A \sum_{j\in A} G_{i,j}(\hat{U}_A^\top)_j.$$

Conditioned on π_i and F_A, $G_{i,j}$ are Bernoulli variables with probability $(F_A\pi_i)_j$. The goal is to compute the variance of the sum, and then use Chebyshev's inequality noted in Proposition 32.

Note that the variance is given by

$$\|\mathbb{E}[(G_{i,A}^\top - F_A\pi_i)^\top \hat{W}_A\hat{W}_A^\top(G_{i,A}^\top - F_A\pi_i)]\| \leq \|\hat{W}_A\|^2 \sum_{j\in A}(F_A\pi_i)_j \left\|(\hat{U}_A^\top)_j\right\|^2.$$

We now bound the variance. By Wedin's theorem, we know the span of columns of $\hat{U}_A$ is $O(\epsilon_G/\sigma_{\min}(G_X^{\alpha_0}, A)) = O(\epsilon_{W_A})$ close to the span of columns of F_A. The span of columns of F_A is the same as the span of rows in Π_A. In particular, let $Proj_\Pi$ be the projection matrix of the span of rows in Π_A, we have

$$\left\|\hat{U}_A\hat{U}_A^\top - Proj_\Pi\right\| \leq O(\epsilon_{W_A}).$$

Using the spectral norm bound, we have the Frobenius norm

$$\left\|\hat{U}_A\hat{U}_A^\top - Proj_\Pi\right\|_F \leq O(\epsilon_{W_A}\sqrt{k})$$

since they are rank k matrices. This implies that

$$\sum_{j\in A}\left(\left\|(\hat{U}_A^\top)_j\right\| - \left\|Proj_\Pi^j\right\|\right)^2 = O(\epsilon_{W_A}^2 k).$$

Now

$$\|Proj_\Pi^j\| \leq \frac{\|\pi_j\|}{\sigma_{\min}(\Pi_A)} = O\left(\sqrt{\frac{(\alpha_0+1)}{n\widehat{\alpha}_{\min}}}\right),$$

from Lemma 23

Now we can bound the variance of the vectors $\sum_{j\in A} G_{i,j}(\hat{U}_A^\top)_j$, since the variance of $G_{i,j}$ is bounded by $(F_A\pi_i)_j$ (its probability), and the variance of the vectors is at most

$$\begin{aligned}\sum_{j\in A}(F_A\pi_i)_j\left\|(\hat{U}_A^\top)_j\right\|^2 &\le 2\sum_{j\in A}(F_A\pi_i)_j\left\|Proj_\Pi^j\right\|^2 + 2\sum_{j\in A}(F_A\pi_i)_j\left(\left\|(\hat{U}_A^\top)_j\right\| - \left\|Proj_\Pi^j\right\|\right)^2\\ &\le 2\sum_{j\in A}(F_A\pi_i)_j\max_{j\in A}\left(\left\|Proj_\Pi^j\right\|^2\right) + \max_{i,j}P_{i,j}\sum_{j\in A}\left(\left\|(\hat{U}_A^\top)_j\right\| - \left\|Proj_\Pi^j\right\|\right)^2\\ &\le O\left(\frac{|F_A|_1(\alpha_0+1)}{n\widehat{\alpha}_{\min}}\right)\end{aligned}$$

Now Chebyshev's inequality implies that with probability at least 1/4 (or any other constant),

$$\left\|\sum_{j\in A}(G_{i,j}-F_A\pi_i)(\hat{U}_A^\top)_j\right\|^2 \le O\left(\frac{|F_A|_1(\alpha_0+1)}{n\widehat{\alpha}_{\min}}\right).$$

And thus, we have

$$\hat{W}_A^\top(G_{i,A}-F_A\pi_i) \le \sqrt{\frac{|F_A|_1(\alpha_0+1)}{n\widehat{\alpha}_{\min}}}\cdot\left\|\hat{W}_A^\top\right\| \le O\left(\epsilon_{W_A}\widehat{\alpha}_{\min}^{-1/2}\right).$$

Combining the two terms, we have the result. ■

Finally, we establish the following perturbation bound between empirical and expected tensor under the Dirichlet distribution, which is used in the proof of Theorem 16.

Lemma 20 (Concentration of third moment tensor under Dirichlet distribution) *With probability* $1-\delta$, *for* $\pi_i \overset{iid}{\sim} \mathrm{Dir}(\alpha)$,

$$\begin{aligned}\left\|\frac{1}{|Y|}\sum_{i\in Y}(\mathrm{Diag}(\widehat{\alpha})^{-1/2}\pi_i)^{\otimes 3} - \mathbb{E}_{\pi\sim\mathrm{Dir}(\alpha)}(\mathrm{Diag}(\widehat{\alpha})^{-1/2}\pi)^{\otimes 3}\right\| &\le O\left(\cdot\frac{1}{\widehat{\alpha}_{\min}\sqrt{n}}\sqrt{\log\frac{n}{\delta}}\right)\\ &= \tilde{O}\left(\frac{1}{\widehat{\alpha}_{min}\sqrt{n}}\right)\end{aligned}$$

Proof The spectral norm of this tensor cannot be larger than the spectral norm of a $k\times k^2$ matrix that we obtain be "collapsing" the last two dimensions (by definitions of norms). Let $\phi_i := \mathrm{Diag}(\hat{\alpha})^{-1/2}\pi_i$ and the "collapsed" tensor is the matrix $\phi_i(\phi_i\otimes\phi_i)^\top$ (here we view $\phi_i\otimes\phi_i$ as a vector in $\mathbb{R}^{k^2}$). We apply Matrix Bernstein on the matrices $Z_i = \phi_i(\phi_i\otimes\phi_i)^\top$. Now

$$\left\|\sum_{i\in Y}\mathbb{E}[Z_iZ_i^\top]\right\| \le |Y|\max\|\phi\|^4\left\|\mathbb{E}[\phi\phi^\top]\right\| \le |Y|\widehat{\alpha}_{\min}^{-2}$$

since $\left\|\mathbb{E}[\phi\phi^\top]\right\| \le 2$. For the other variance term $\left\|\sum_{i\in Y}\mathbb{E}[Z_i^\top Z_i]\right\|$, we have

$$\left\|\sum_{i\in Y}\mathbb{E}[Z_i^\top Z_i]\right\| \le |Y|\widehat{\alpha}_{\min}\left\|\mathbb{E}[(\phi\otimes\phi)(\phi\otimes\phi)^\top]\right\|.$$

It remains to bound the norm of $\mathbb{E}[(\phi \otimes \phi)(\phi \otimes \phi)^\top]$. We have

$$\|\mathbb{E}[(\phi \otimes \phi)(\phi \otimes \phi)^\top]\| = \sup \left(\|\mathbb{E}[M^2]\|,\ s.t.\, M = \sum_{i,j} N_{i,j} \phi_i \phi_j^\top,\ \|N\|_F = 1 \right).$$

by definition. We now group the terms of $\mathbb{E}[M^2]$ and bound them separately.

$$\begin{aligned} M^2 = & \sum_i N_{i,i}^2 \phi_i \phi_i^\top \|\phi_i\|^2 + \sum_{i \neq j} N_{i,j}^2 \phi_i \phi_j^\top \langle \phi_i, \phi_j \rangle \\ & + \sum_{i \neq j \neq a} N_{i,i} N_{j,a} \phi_i \phi_a^\top \langle \phi_i, \phi_j \rangle + \sum_{i \neq j \neq a \neq b} N_{i,j} N_{a,b} \phi_i \phi_b^\top \langle \phi_j, \phi_a \rangle \end{aligned} \tag{79}$$

We bound the terms individually now.

$\|\phi(i)\|^4$ terms: By properties of Dirichlet distribution we know

$$\mathbb{E}[\|\phi(i)\|^4] = \Theta(\widehat{\alpha}_i^{-1}) \leq O(\widehat{\alpha}_{\min}^{-1}).$$

Thus, for the first term in (79), we have

$$\sup_{N:\|N\|_F=1} \| \sum_i \mathbb{E}[N_{i,i}^2 \phi_i \phi_i^\top \|\phi_i\|]\|^2\| = O(\widehat{\alpha}_{\min}^{-1}).$$

$\|\phi(i)\|^3 \cdot \|\phi(j)\|$ terms: We have

$$\|\mathbb{E}[\sum_{i,j} N_{i,i} N_{i,j} \phi(i)^3 \phi(j)]\| \leq \mathbb{E}[\|\phi_i\|^2 \cdot \|\phi_j\|] \leq O(\sqrt{\sum_{i,j} (N_{i,i}^2 \hat{\alpha}(j)) \sum_{i,j} N_{i,j}^2 \hat{\alpha}(i)^{-1}}) \leq O(\widehat{\alpha}_{\min}^{-1/2}).$$

$\|\phi(i)\|^2 \cdot \|\phi(j)\|^2$ terms: the total number of such terms is $O(k^2)$ and we have

$$\mathbb{E}[\|\phi(i)\|^2 \cdot \|\phi(j)\|^2] = \Theta(1),$$

and thus the Frobenius norm of these set of terms is smaller than $O(k)$

$\|\phi(i)\|^2 \cdot \|\phi(j)\| \cdot \|\phi(a)\|$ terms: there are $O(k^3)$ such terms, and we have

$$\|\mathbb{E}[\phi(i)\|^2 \cdot \|\phi(j)\| \cdot \|\phi(a)]\| = \Theta(\hat{\alpha}(i_2)^{1/2} \hat{\alpha}(i_3)^{1/2}).$$

The Frobenius norm of this part of matrix is bounded by

$$O\left(\sqrt{\sum_{i,j,a \in [k]} \hat{\alpha}(j) \hat{\alpha}(a)} \right) \leq O(\sqrt{k}) \sqrt{\sum_j \sum_a \widehat{\alpha}_j \widehat{\alpha}_a} \leq O(\sqrt{k}).$$

the rest: the sum is

$$\mathbb{E}[\sum_{i \neq j \neq a \neq b} N_{i,j} N_{a,b} \hat{\alpha}(i)^{1/2} \hat{\alpha}(j)^{1/2} \hat{\alpha}(a)^{1/2} \hat{\alpha}(b)^{1/2}].$$

It is easy to break the bounds into the product of two sums ($\sum_{i,j}$ and $\sum_{a,b}$) and then bound each one by Cauchy-Schwartz, the result is 1.

Hence the variance term in Matrix Bernstein's inequality can be bounded by $\sigma^2 \leq O(n\widehat{\alpha}_{\min}^{-2})$, each term has norm at most $\widehat{\alpha}_{\min}^{-3/2}$. When $\widehat{\alpha}_{\min}^{-2} < n$ we know the variance term dominates and the spectral norm of the difference is at most $O(\widehat{\alpha}_{\min}^{-1} n^{-1/2}\sqrt{\log n/\delta})$ with probability $1-\delta$.

■

C.4 Basic Results on Spectral Concentration of Adjacency Matrix

Let $n := \max(|A|, |X|)$.

Lemma 21 (Concentration of $G_{X,A}^{\alpha_0}$) *When $\pi_i \sim \mathrm{Dir}(\alpha)$, for $i \in V$, with probability $1-4\delta$,*

$$\epsilon_G := \|G_{X,A}^{\alpha_0} - \mathbb{E}[(G_{X,A}^{\alpha_0})^\top|\Pi]\| = O\left(\sqrt{(\alpha_0+1)n \cdot (\max_i (P\widehat{\alpha})_i)(1+\varepsilon_2)\log\frac{n}{\delta}}\right) \tag{80}$$

Proof: From definition of $G_{X,A}^{\alpha_0}$, we have

$$\epsilon_G \leq \sqrt{\alpha_0+1}\|G_{X,A} - \mathbb{E}[G_{X,A}|\Pi]\| + (\sqrt{\alpha_0+1}-1)\sqrt{|X|}\|\mu_{X,A} - \mathbb{E}[\mu_{X,A}|\Pi]\|.$$

We have concentration for $\mu_{X,A}$ and adjacency submatrix $G_{X,A}$ from Lemma 22. ■

We now provide concentration bounds for adjacency sub-matrix $G_{X,A}$ from partition X to A and the corresponding mean vector. Recall that $\mathbb{E}[\mu_{X\to A}|F_A, \pi_X] = F_A\pi_X$ and $\mathbb{E}[\mu_{X\to A}|F_A] = F_A\widehat{\alpha}$.

Lemma 22 (Concentration of adjacency submatrices) *When $\pi_i \overset{iid}{\sim} \mathrm{Dir}(\alpha)$ for $i \in V$, with probability $1-2\delta$,*

$$\|G_{X,A} - \mathbb{E}[G_{X,A}|\Pi]\| = O\left(\sqrt{n \cdot (\max(\max_i (P\widehat{\alpha})_i, \max_i (P^\top\widehat{\alpha})_i))(1+\varepsilon_2)\log\frac{n}{\delta}}\right). \tag{81}$$

$$\|\mu_A - \mathbb{E}[\mu_A|\Pi]\| = O\left(\frac{1}{|X|}\sqrt{n \cdot (\max(\max_i (P\widehat{\alpha})_i, \max_i (P^\top\widehat{\alpha})_i))(1+\varepsilon_2)\log\frac{n}{\delta}}\right), \tag{82}$$

where ε_2 is given by (87).

Proof: Recall $\mathbb{E}[G_{X,A}|\Pi] = F_A\Pi_X$ and $G_{A,X} = \mathrm{Ber}(F_A\Pi_X)$ where $\mathrm{Ber}(\cdot)$ denotes the Bernoulli random matrix with independent entries. Let

$$Z_i := (G_{i,A}^\top - F_A\pi_i)e_i^\top.$$

We have $G_{X,A}^\top - F_A\Pi_X = \sum_{i\in X} Z_i$. We apply matrix Bernstein's inequality.

We compute the variances $\sum_i \mathbb{E}[Z_iZ_i^\top|\Pi]$ and $\sum_i \mathbb{E}[Z_i^\top Z_i|\Pi]$. We have that $\sum_i \mathbb{E}[Z_iZ_i^\top|\Pi]$ only the diagonal terms are non-zero due to independence of Bernoulli variables, and

$$\mathbb{E}[Z_iZ_i^\top|\Pi] \leq \mathrm{Diag}(F_A\pi_i) \tag{83}$$

entry-wise. Thus,

$$\begin{aligned}\|\sum_{i\in X}\mathbb{E}[Z_iZ_i^\top|\Pi]\| &\le \max_{a\in A}\sum_{i\in X,b\in[k]}F_A(a,b)\pi_i(b)\\ &= \max_{a\in A}\sum_{i\in X,b\in[k]}F_A(a,b)\Pi_X(b,i)\\ &\le \max_{c\in[k]}\sum_{i\in X,b\in[k]}P(b,c)\Pi_X(b,i)\\ &= \|P^\top\Pi_X\|_\infty. \end{aligned} \tag{84}$$

Similarly $\|\sum_{i\in X}\mathbb{E}[Z_i^\top Z_i]\| = \sum_{i\in X}\mathrm{Diag}(\mathbb{E}[\|G_{i,A}^\top - F_A\pi_i\|^2]) \le \|P^\top\Pi_X\|_\infty$. On lines of Lemma 27, we have $\|P^\top\Pi_X\|_\infty = O(|X|\cdot(\max_i(P^\top\widehat{\alpha})_i))$ when $|X|$ satisfies (71).

We now bound $\|Z_i\|$. First note that the entries in $G_{i,A}$ are independent and we can use the vector Bernstein's inequality to bound $\|G_{i,A}-F_A\pi_i\|$. We have $\max_{j\in A}|G_{i,j}-(F_A\pi_i)_j| \le 2$ and $\sum_j\mathbb{E}[G_{i,j}-(F_A\pi_i)_j]^2 \le \sum_j(F_A\pi_i)_j \le \|F_A\|_1$. Thus with probability $1-\delta$, we have

$$\|G_{i,A}-F_A\pi_i\| \le (1+\sqrt{8\log(1/\delta)})\sqrt{\|F_A\|_1} + 8/3\log(1/\delta).$$

Thus, we have the bound that $\|\sum_i Z_i\| = O(\max(\sqrt{\|F_A\|_1}, \sqrt{\|P^\top\Pi_X\|_\infty}))$. The concentration of the mean term follows from this result. ■

We now provide spectral bounds on $\mathbb{E}[(G_{X,A}^{\alpha_0})^\top|\Pi]$. Define

$$\psi_i := \mathrm{Diag}(\hat{\alpha})^{-1/2}(\sqrt{\alpha_0+1}\pi_i - (\sqrt{\alpha_0+1}-1)\mu). \tag{85}$$

Let Ψ_X be the matrix with columns ψ_i, for $i\in X$. We have

$$\mathbb{E}[(G_{X,A}^{\alpha_0})^\top|\Pi] = F_A\,\mathrm{Diag}(\hat{\alpha})^{1/2}\Psi_X,$$

from definition of $\mathbb{E}[(G_{X,A}^{\alpha_0})^\top|\Pi]$.

Lemma 23 (Spectral bounds) *With probability* $1-\delta$,

$$\varepsilon_1 := \|I - |X|^{-1}\Psi_X\Psi_X^\top\| \le O\left(\sqrt{\frac{(\alpha_0+1)}{\widehat{\alpha}_{\min}|X|}}\cdot\log\frac{k}{\delta}\right) \tag{86}$$

With probability $1-2\delta$,

$$\begin{aligned}\|\mathbb{E}[(G_{X,A}^{\alpha_0})^\top|\Pi]\| &= O\left(\|P\|\widehat{\alpha}_{\max}\sqrt{|X||A|(1+\varepsilon_1+\varepsilon_2)}\right)\\ \sigma_{\min}\left(\mathbb{E}[(G_{X,A}^{\alpha_0})^\top|\Pi]\right) &= \Omega\left(\widehat{\alpha}_{\min}\sqrt{\frac{|A||X|}{\alpha_0+1}(1-\varepsilon_1-\varepsilon_3)}\cdot\sigma_{\min}(P)\cdot\right),\end{aligned}$$

where

$$\varepsilon_2 := O\left(\left(\frac{1}{|A|\widehat{\alpha}_{\max}^2}\log\frac{k}{\delta}\right)^{1/4}\right),\quad \varepsilon_3 := O\left(\left(\frac{(\alpha_0+1)^2}{|A|\widehat{\alpha}_{\min}^2}\log\frac{k}{\delta}\right)^{1/4}\right). \tag{87}$$

Remark: *When partitions X, A satisfy* (71)*, $\varepsilon_1, \varepsilon_2, \varepsilon_3$ are small.*
Proof: Note that ψ_i is a random vector with norm bounded by $O(\sqrt{(\alpha_0+1)/\widehat{\alpha}_{min}})$ from Lemma 27 and $\mathbb{E}[\psi_i \psi_i^\top] = I$. We now prove (86). using Matrix Bernstein Inequality. Each matrix $\psi_i\psi_i^\top/|X|$ has spectral norm at most $O((\alpha_0+1)/\widehat{\alpha}_{min}|X|)$. The variance σ^2 is bounded by

$$\left\| \frac{1}{|X|^2} \mathbb{E}[\sum_{i\in X} \|\psi_i\|^2 \psi_i\psi_i^\top] \right\| \leq \left\| \frac{1}{|X|^2} \max \|\psi_i\|^2 \mathbb{E}[\sum_{i\in X} \psi_i\psi_i^\top] \right\| \leq O((\alpha_0+1)/\widehat{\alpha}_{\min}|X|).$$

Since $O((\alpha_0+1)/\alpha_{\min}|X|) < 1$, the variance dominates in Matrix Bernstein's inequality.

Let $B := |X|^{-1}\Psi_X \Psi_X^\top$. We have with probability $1-\delta$,

$$\begin{aligned}\sigma_{\min}(\mathbb{E}[(G_{X,A}^{\alpha_0})^\top|\Pi]) &= \sqrt{|X|\sigma_{\min}(F_A \operatorname{Diag}(\hat{\alpha})^{1/2} B \operatorname{Diag}(\hat{\alpha})^{1/2} F_A^\top)}, \\ &= \Omega(\sqrt{\widehat{\alpha}_{\min}|X|(1-\epsilon_1)} \cdot \sigma_{\min}(F_A)).\end{aligned}$$

From Lemma 27, with probability $1-\delta$,

$$\sigma_{\min}(F_A) \geq \left(\sqrt{\frac{|A|\widehat{\alpha}_{\min}}{\alpha_0+1}} - O((|A|\log k/\delta)^{1/4}) \right) \cdot \sigma_{\min}(P).$$

Similarly other results follow. ■

C.5 Properties of Dirichlet Distribution

In this section, we list various properties of Dirichlet distribution.

C.5.1 Sparsity Inducing Property

We first note that the Dirichlet distribution $\mathrm{Dir}(\alpha)$ is sparse depending on values of α_i, which is shown in Telgarsky (2012).

Lemma 24 *Let reals $\tau \in (0,1]$, $\alpha_i > 0$, $\alpha_0 := \sum_i \alpha_i$ and integers $1 \leq s \leq k$ be given. Let $(X_i, \ldots, X_k) \sim \mathrm{Dir}(\alpha)$. Then*

$$\Pr\left[|\{i : X_i \geq \tau\}| \leq s\right] \geq 1 - \tau^{-\alpha_0} e^{-(s+1)/3} - e^{-4(s+1)/9},$$

when $s+1 < 3k$.

We now show that we obtain good initialization vectors under Dirichlet distribution.

Arrange the $\widehat{\alpha}_j$'s in ascending order, i.e., $\widehat{\alpha}_1 = \widehat{\alpha}_{\min} \leq \widehat{\alpha}_2 \ldots \leq \widehat{\alpha}_k = \widehat{\alpha}_{\max}$. Recall that columns vectors $\hat{W}_A^\top G_{i,A}^\top$, for $i \notin A$, are used as initialization vectors to the tensor power method. We say that $u_i := \frac{\hat{W}_A^\top G_{i,A}^\top}{\|\hat{W}_A^\top G_{i,A}^\top\|}$ is a (γ, R_0)-good initialization vector corresponding to $j \in [k]$ if

$$|\langle u_i, \Phi_j\rangle| \geq R_0, \quad |\langle u_i, \Phi_j\rangle| - \max_{m<j} |\langle u_i, \Phi_m\rangle| \geq \gamma\, |\langle u_i, \Phi_j\rangle|\,, \tag{88}$$

where $\Phi_j := \widehat{\alpha}_j^{1/2}(\tilde{F}_A)_j$, where $(\tilde{F}_A)_j$ is the j^{th} column of $\tilde{F}_A := W_A^\top F_A$. Note that the $\{\Phi_j\}$ are orthonormal and are the eigenvectors to be estimated by the tensor power method.

Lemma 25 (Good initialization vectors under Dirichlet distribution) *When* $\pi_i \overset{iid}{\sim} \mathrm{Dir}(\alpha)$, *and* $\alpha_j < 1$, *let*

$$\Delta := O\left(\frac{\zeta\rho}{\sqrt{n}r_0}\right). \tag{89}$$

For $j \in [k]$, *there is at least one* $(\gamma - \frac{2\Delta}{r_0-\Delta}, r_0 - \Delta)$*-good vector corresponding to each* Φ_j, *for* $j \in [k]$, *among* $\{u_i\}_{i\in[n]}$ *with probability* $1 - 9\delta$, *when*

$$n = \tilde{\Omega}\left(\alpha_{\min}^{-1} e^{r_0 \widehat{\alpha}_{\max}^{1/2}(\alpha_0 + c_1\sqrt{k\alpha_0})}(2k)^{r_0 c_2}\log(k/\delta)\right), \tag{90}$$

where $c_1 := (1 + \sqrt{8\log 4})$ *and* $c_2 := 4/3(\log 4)$, *when*

$$(1-\gamma)r_0\widehat{\alpha}_{\min}^{1/2}(\alpha_0 + (1+\sqrt{8\log 4})\sqrt{k\alpha_0} + 4/3(\log 4)\widehat{\alpha}_{\min}^{-1/2}\log 2k) > 1. \tag{91}$$

When $\alpha_0 < 1$, *the bound can be improved for* $r_0 \in (0.5, (\alpha_0+1)^{-1})$ *and* $1 - \gamma \geq \frac{1-r_0}{r_0}$ *as*

$$n > \frac{(1+\alpha_0)(1 - r_0\widehat{\alpha}_{\min})}{\widehat{\alpha}_{\min}(\alpha_{\min} + 1 - r_0(\alpha_0+1))}\log(k/\delta). \tag{92}$$

Remark: *(when* $\alpha_0 \geq 1$, $\alpha_0 = \Theta(1)$*) When* r_0 *is chosen as* $r_0 = \alpha_{\max}^{-1/2}(\sqrt{\alpha_0} + c_1\sqrt{k})^{-1}$, *the term* $e^{r_0\widehat{\alpha}_{\max}^{1/2}(\alpha_0 + c_1\sqrt{k\alpha_0})} = e$, *and we require*

$$n = \tilde{\Omega}\left(\alpha_{\min}^{-1} k^{0.43}\log(k/\delta)\right), \quad r_0 = \alpha_{\max}^{-1/2}(\sqrt{\alpha_0} + c_1\sqrt{k})^{-1}, \tag{93}$$

by substituting $c_2/c_1 = 0.43$. *Moreover,* (91) *is satisfied for the above choice of* r_0 *when* $\gamma = \Theta(1)$.

In this case we also need $\Delta < r_0/2$, *which implies*

$$\zeta = O\left(\frac{\sqrt{n}}{\rho k\widehat{\alpha}_{\max}}\right) \tag{94}$$

Remark: *(when* $\alpha_0 < 1$*) In this regime,* (92) *implies that we require* $n = \Omega(\widehat{\alpha}_{\min}^{-1})$. *Also,* r_0 *is a constant, we just need* $\zeta = O(\sqrt{n}/\rho)$.

Proof: Define $\tilde{u}_i := W_A^\top F_A\pi_i / \|W_A^\top F_A\pi_i\|$, when whitening matrix W_A and F_A corresponding to exact statistics are input.

We first observe that if $\tilde{u}_i$ is (γ, r_0) good, then u_i is $(\gamma - \frac{2\Delta}{r_0-\Delta}, r_0 - \Delta)$ good.

When $\tilde{u}_i$ is (γ, r_0) good, note that $W_A^\top F_A\pi_i \geq \widehat{\alpha}_{\max}^{-1/2} r_0$ since $\sigma_{\min}(W_A^\top F_A) = \widehat{\alpha}_{\max}^{-1/2}$ and $\|\pi_i\| \geq r_0$. Now with probability 1/4, conditioned on π_i, we have the event $\mathcal{B}(i)$,

$$\mathcal{B}(i) := \{\|u_i - \tilde{u}_i\| \leq \Delta\},$$

where Δ is given by

$$\Delta = \tilde{O}\left(\frac{\widehat{\alpha}_{\max}^{0.5}\sqrt{(\alpha_0+1)(\max_i(P\widehat{\alpha})_i)}}{r_0 n^{1/2}\widehat{\alpha}_{\min}^{1.5}\sigma_{\min}(P)}\right)$$

from Lemma 19. Thus, we have $\mathbb{P}[\mathcal{B}(i)|\pi_i] \geq 1/4$, i.e., $\mathcal{B}(i)$ occurs with probability $1/4$ for any realization of π_i.

If we perturb a (γ, r_0) good vector by Δ (while maintaining unit norm), then it is still $(\gamma - \frac{2\Delta}{r_0 - \Delta}, r_0 - \Delta)$ good.

We now show that the set $\{\tilde{u}_i\}$ contains good initialization vectors when n is large enough. Consider $Y_i \sim \Gamma(\alpha_i, 1)$, where $\Gamma(\cdot,\cdot)$ denotes the Gamma distribution and we have $Y/\sum_i Y_i \sim \mathrm{Dir}(\alpha)$. We first compute the probability that $\tilde{u}_i := W_A^\top F_A \pi_i / \|W_A^\top F_A \pi_i\|$ is a (r_0, γ)-good vector with respect to $j = 1$ (recall that $\widehat{\alpha}_1 = \widehat{\alpha}_{\min}$). The desired event is

$$\mathcal{A}_1 := (\widehat{\alpha}_1^{-1/2} Y_1 \geq r_0 \sqrt{\sum_j \widehat{\alpha}_j^{-1} Y_j^2}) \cap (\widehat{\alpha}_1^{-1/2} Y_1 \geq \frac{1}{1-\gamma} \max_{j>1} \widehat{\alpha}_j^{-1/2} Y_j) \tag{95}$$

We have

$$\begin{aligned}
\mathbb{P}[\mathcal{A}_1] &\geq \mathbb{P}\left[(\widehat{\alpha}_{\min}^{-1/2} Y_1 \geq r_0 \sqrt{\sum_j \widehat{\alpha}_j^{-1} Y_j^2}) \cap (Y_1 \geq \frac{1}{1-\gamma} \max_{j>1} Y_j) \right] \\
&\geq \mathbb{P}\left[(\widehat{\alpha}_{\min}^{-1/2} Y_1 > r_0 t) \bigcap (\sum_j \widehat{\alpha}_j^{-1} Y_j^2 \leq t^2) \bigcap_{j>1} (Y_1 \leq (1-\gamma) r_0 t \widehat{\alpha}_{\min}^{1/2}) \right], \quad \text{for some } t \\
&\geq \mathbb{P}\left[\widehat{\alpha}_{\min}^{-1/2} Y_1 > r_0 t \right] \mathbb{P}\left[\sum_j \widehat{\alpha}_j^{-1} Y_j^2 \leq t^2 \Big| \widehat{\alpha}_j^{-1/2} Y_j \leq (1-\gamma) r_0 t \widehat{\alpha}_{\min}^{1/2} \right] \\
&\quad \mathbb{P}\left[\max_{j>1} Y_j \leq (1-\gamma) r_0 t \widehat{\alpha}_{\min}^{1/2} \right] \\
&\geq \mathbb{P}\left[\widehat{\alpha}_{\min}^{-1/2} Y_1 > r_0 t \right] \mathbb{P}\left[\sum_j \widehat{\alpha}_j^{-1} Y_j^2 \leq t^2 \right] \mathbb{P}\left[\max_{j>1} Y_j \leq (1-\gamma) r_0 t \widehat{\alpha}_{\min}^{1/2} \right].
\end{aligned}$$

When $\alpha_j \leq 1$, we have

$$\mathbb{P}[\cup_j Y_j \geq \log 2k] \leq 0.5,$$

since $P(Y_j \geq t) \leq t^{\alpha_j - 1} e^{-t} \leq e^{-t}$ when $t > 1$ and $\alpha_j \leq 1$. Applying vector Bernstein's inequality, we have with probability $0.5 - e^{-m}$ that

$$\|\operatorname{Diag}(\widehat{\alpha}_j^{-1/2})(Y - \mathbb{E}(Y))\|_2 \leq (1 + \sqrt{8m})\sqrt{k\alpha_0} + 4/3 m \widehat{\alpha}_{\min}^{-1/2} \log 2k,$$

since $\mathbb{E}[\sum_j \widehat{\alpha}_j^{-1} \operatorname{Var}(Y_j)] = k\alpha_0$ since $\widehat{\alpha}_j = \alpha_j/\alpha_0$ and $\operatorname{Var}(Y_j) = \alpha_j$. Thus, we have

$$\|\operatorname{Diag}(\widehat{\alpha}_j^{-1/2}) Y\|_2 \leq \alpha_0 + (1 + \sqrt{8m})\sqrt{k\alpha_0} + 4/3 m \widehat{\alpha}_{\min}^{-1/2} \log 2k,$$

since $\|\operatorname{Diag}(\widehat{\alpha}_j^{-1/2})\mathbb{E}(Y)\|_2 = \sqrt{\sum_j \widehat{\alpha}_j^{-1} \alpha_j^2} = \alpha_0$. Choosing $m = \log 4$, we have with probability $1/4$ that

$$\begin{aligned}
\|\operatorname{Diag}(\widehat{\alpha}_j^{-1/2}) Y\|_2 \leq t &:= \alpha_0 + (1 + \sqrt{8 \log 4})\sqrt{k\alpha_0} + 4/3 (\log 4) \widehat{\alpha}_{\min}^{-1/2} \log 2k, &(96)\\
&= \alpha_0 + c_1 \sqrt{k\alpha_0} + c_2 \widehat{\alpha}_{\min}^{-1/2} \log 2k. &(97)
\end{aligned}$$

We now have

$$\mathbb{P}\left[\widehat{\alpha}_{\min}^{-1/2} Y_1 > r_0 t\right] \geq \frac{\alpha_{\min}}{4C}\left(r_0 t \widehat{\alpha}_{\min}^{1/2}\right)^{\alpha_{\min}-1} e^{-r_0 t \widehat{\alpha}_{\min}^{1/2}},$$

from Lemma 28.

Similarly,

$$\begin{aligned}\mathbb{P}\left[\max_{j\neq 1} Y_j \leq \widehat{\alpha}_{\min}^{1/2}(1-\gamma) r_0 t\right] &\geq 1 - \sum_j \left((1-\gamma) r_0 t \widehat{\alpha}_{\min}^{1/2}\right)^{\sum_j \alpha_j - 1} e^{-(1-\gamma) r_0 \widehat{\alpha}_{\min}^{1/2} t} \\ &\geq 1 - k e^{-(1-\gamma) r_0 \widehat{\alpha}_{\min}^{1/2} t},\end{aligned}$$

assuming that $(1-\gamma) r_0 \widehat{\alpha}_{\min}^{1/2} t > 1$.

Choosing t as in (96), we have the probability of the event in (95) is greater than

$$\begin{aligned}&\frac{\alpha_{\min}}{16C}\left(1 - \frac{e^{-(1-\gamma) r_0 \widehat{\alpha}_{\min}^{1/2}(\alpha_0 + c_1\sqrt{k\alpha_0})}}{2(2k)^{(1-\gamma) r_0 c_2 - 1}}\right)\frac{e^{-r_0 \widehat{\alpha}_{\min}^{1/2}(\alpha_0 + c_1\sqrt{k\alpha_0})}}{(2k)^{r_0 c_2}} \\ &\cdot \left(r_0 \widehat{\alpha}_{\min}^{1/2}(\alpha_0 + c_1\sqrt{k\alpha_0} + c_2 \widehat{\alpha}_{\min}^{-1/2}\log 2k)\right)^{\alpha_{\min}-1}.\end{aligned}$$

Similarly the (marginal) probability of events $\mathcal{A}_2$ can be bounded from below by replacing $\alpha_{\min}$ with α_2 and so on. Thus, we have

$$\mathbb{P}[\mathcal{A}_m] = \tilde{\Omega}\left(\alpha_{\min}\frac{e^{-r_0 \widehat{\alpha}_{\max}^{1/2}(\alpha_0 + c_1\sqrt{k\alpha_0})}}{(2k)^{r_0 c_2}}\right),$$

for all $m \in [k]$.

Thus, we have each of the events $\mathcal{A}_1(i)\cap\mathcal{B}(i), \mathcal{A}_2(i)\cap\mathcal{B}(i), \ldots, \mathcal{A}_k\cap\mathcal{B}(i)$ occur at least once in $i \in [n]$ i.i.d. tries with probability

$$\begin{aligned}&1 - \mathbb{P}\left[\bigcup_{j\in[k]}\left(\bigcap_{i\in[n]}(\mathcal{A}_j(i)\cap\mathcal{B}(i))^c\right)\right] \\ &\geq 1 - \sum_{j\in[k]}\mathbb{P}\left[\bigcap_{i\in[n]}(\mathcal{A}_j(i) - \mathcal{B}(i))^c\right] \\ &\geq 1 - \sum_{j\in[k]}\exp\left[-n\mathbb{P}(\mathcal{A}_j\cap\mathcal{B})\right], \\ &\geq 1 - k\exp\left[-n\tilde{\Omega}\left(\alpha_{\min}\frac{e^{-r_0 \widehat{\alpha}_{\max}^{1/2}(\alpha_0 + c_1\sqrt{k\alpha_0})}}{(2k)^{r_0 c_2}}\right)\right],\end{aligned}$$

where $\mathcal{A}_j(i)$ denotes the event that $\mathcal{A}_1$ occurs for i^{th} trial and we have that $\mathbb{P}[\mathcal{B}|\mathcal{A}_j] \geq 0.25$ since $\mathcal{B}$ occurs in any trial with probability 0.25 for any realization of π_i and the events $\mathcal{A}_j$ depend only on π_i. We use that $1 - x \leq e^{-x}$ when $x \in [0,1]$. Thus, for the event to occur with probability $1-\delta$, we require

$$n = \tilde{\Omega}\left(\alpha_{\min}^{-1} e^{r_0 \widehat{\alpha}_{\max}^{1/2}(\alpha_0 + c_1\sqrt{k\alpha_0})}(2k)^{r_0 c_2}\log(1/\delta)\right).$$

We can improve the above bound by directly working with the Dirichlet distribution. Let $\pi \sim \mathrm{Dir}(\alpha)$. The desired event corresponding to $j = 1$ is given by

$$\mathcal{A}_1 = \left(\frac{\widehat{\alpha}_1^{-1/2} \pi_1}{\| \mathrm{Diag}(\widehat{\alpha}_i^{-1/2}) \pi \|} \geq r_0 \right) \bigcap_{i>1} \left(\pi_1 \geq \frac{\pi_i}{1-\gamma} \right).$$

Thus, we have

$$\begin{aligned} \mathbb{P}[\mathcal{A}_1] &\geq \mathbb{P}\left[(\pi_1 \geq r_0) \bigcap_{i>1} (\pi_i \leq (1-\gamma) r_0) \right] \\ &\geq \mathbb{P}[\pi_1 \geq r_0] \mathbb{P}\left(\bigcap_{i>1} \pi_i \leq (1-\gamma) r_0 | \pi_1 \geq r_0 \right), \end{aligned}$$

since $\mathbb{P}\left(\bigcap_{i>1} \pi_i \leq (1-\gamma) r_0 | \pi_1 \geq r_0\right) \geq \mathbb{P}\left(\bigcap_{i>1} \pi_i \leq (1-\gamma) r_0\right)$. By properties of Dirichlet distribution, we know $\mathbb{E}[\pi_i] = \widehat{\alpha}_i$ and $\mathbb{E}[\pi_i^2] = \widehat{\alpha}_i \frac{\alpha_i + 1}{\alpha_0 + 1}$. Let $p := \Pr[\pi_1 \geq r_0]$. We have

$$\begin{aligned} \mathbb{E}[\pi_i^2] &= p \mathbb{E}[\pi_i^2 | \pi_i \geq r_0] + (1-p) \mathbb{E}[\pi_i^2 | \pi_i < r_0] \\ &\leq p + (1-p) r_0 \mathbb{E}[\pi_i | \pi_i < r_0] \\ &\leq p + (1-p) r_0 \mathbb{E}[\pi_i]. \end{aligned}$$

Thus, $p \geq \frac{\widehat{\alpha}_{\min}(\alpha_{\min} + 1 - r_0(\alpha_0+1))}{(\alpha_0+1)(1-r_0 \widehat{\alpha}_{\min})}$, which is useful when $r_0(\alpha_0 + 1) < 1$. Also when $\pi_1 \geq r_0$, we have that $\pi_i \leq 1 - r_0$ since $\pi_i \geq 0$ and $\sum_i \pi_i = 1$. Thus, choosing $1 - \gamma = \frac{1-r_0}{r_0}$, we have the other conditions for $\mathcal{A}_1$ are satisfied. Also, verify that we have $\gamma < 1$ when $r_0 > 0.5$ and this is feasible when $\alpha_0 < 1$. ■

We now prove a result that the entries of π_i, which are marginals of the Dirichlet distribution, are likely to be small in the sparse regime of the Dirichlet parameters. Recall that the marginal distribution of π_i is distributed as $B(\alpha_i, \alpha_0 - \alpha_i)$, where $B(a, b)$ is the beta distribution and

$$\mathbb{P}[Z = z] \propto z^{a-1}(1-z)^{b-1}, \quad Z \sim B(a, b).$$

Lemma 26 (Marginal Dirichlet distribution in sparse regime) *For $Z \sim B(a, b)$, the following results hold:*
Case $b \leq 1$**,** $C \in [0, 1/2]$**:**

$$\Pr[Z \geq C] \leq 8 \log(1/C) \cdot \frac{a}{a+b}, \tag{98}$$

$$\mathbb{E}[Z \cdot \delta(Z \leq C)] \leq C \cdot \mathbb{E}[Z] = C \cdot \frac{a}{a+b}. \tag{99}$$

Case $b \geq 1$**,** $C \leq (b+1)^{-1}$**:** *we have*

$$\Pr[Z \geq C] \leq a \log(1/C), \tag{100}$$

$$\mathbb{E}[Z \cdot \delta(Z \leq C)] \leq 6aC. \tag{101}$$

Remark: *The guarantee for $b \geq 1$ is worse and this agrees with the intuition that the Dirichlet vectors are more spread out (or less sparse) when $b = \alpha_0 - \alpha_i$ is large.*

Proof We have

$$\begin{aligned}\mathbb{E}[Z \cdot \delta(Z \leq C)] &= \int_0^C \frac{1}{B(a,b)} x^a (1-x)^{b-1} dx \\ &\leq \frac{(1-C)^{b-1}}{B(a,b)} \int_0^C x^a dx \\ &= \frac{(1-C)^{b-1} C^{a+1}}{(a+1)B(a,b)}.\end{aligned}$$

For $\mathbb{E}[Z \cdot \delta(Z \geq C)]$, we have,

$$\begin{aligned}\mathbb{E}[Z \cdot \delta(Z \geq C)] &= \int_C^1 \frac{1}{B(a,b)} x^a (1-x)^{b-1} dx \\ &\geq \frac{C^a}{B(a,b)} \int_C^1 (1-x)^{b-1} dx \\ &= \frac{(1-C)^b C^a}{bB(a,b)}.\end{aligned}$$

The ratio between these two is at least

$$\frac{\mathbb{E}[Z \cdot \delta(Z \geq C)]}{\mathbb{E}[Z \cdot \delta(Z \leq C)]} \geq \frac{(1-C)(a+1)}{bC} \geq \frac{1}{C}.$$

The last inequality holds when $a, b < 1$ and $C < 1/2$. The sum of the two is exactly $\mathbb{E}[Z]$, so when $C < 1/2$ we know $\mathbb{E}[Z \cdot \delta(Z \leq C)] < C \cdot \mathbb{E}[Z]$.

Next we bound the probability $\Pr[Z \geq C]$. Note that $\Pr[Z \geq 1/2] \leq 2\mathbb{E}[Z] = \frac{2a}{a+b}$ by Markov's inequality. Now we show $\Pr[Z \in [C, 1/2]]$ is not much larger than $\Pr[Z \geq 1/2]$ by bounding the integrals.

$$A = \int_{1/2}^1 x^{a-1}(1-x)^{b-1} dx \geq \int_{1/2}^1 (1-x)^{b-1} dx = (1/2)^b / b.$$

$$\begin{aligned}B = \int_C^{1/2} x^{a-1}(1-x)^{b-1} &\leq (1/2)^{b-1} \int_C^{1/2} x^{a-1} dx \\ &\leq (1/2)^{b-1} \frac{0.5^a - C^a}{a} \\ &\leq (1/2)^{b-1} \frac{1 - (1 - a \log 1/C)}{a} \\ &= (1/2)^{b-1} \log(1/C).\end{aligned}$$

The last inequality uses the fact that $e^x \geq 1 + x$ for all x. Now

$$\Pr[Z \geq C] = (1 + \frac{B}{A}) \Pr[Z \geq 1/2] \leq (1 + 2b \log(1/C)) \frac{2a}{a+b} \leq 8 \log(1/C) \cdot \frac{a}{a+b}$$

and we have the result.

When $b \geq 1$, we have an alternative bound. We use the fact that if $X \sim \Gamma(a,1)$ and $Y \sim \Gamma(b,1)$ then $Z \sim X/(X+Y)$. Since Y is distributed as $\Gamma(b,1)$, its PDF is $\frac{1}{\Gamma(b)}x^{b-1}e^{-x}$. This is proportional to the PDF of $\Gamma(1)$ (e^{-x}) multiplied by a increasing function x^{b-1}.

Therefore we know $\Pr[Y \geq t] \geq \Pr_{Y' \sim \Gamma(1)}[Y' \geq t] = e^{-t}$.

Now we use this bound to compute the probability that $Z \leq 1/R$ for all $R \geq 1$.

This is equivalent to

$$\begin{aligned}\Pr[\frac{X}{X+Y} \leq \frac{1}{R}] &= \int_0^\infty Pr[X=x]Pr[Y \geq (R-1)X]dx \\ &\geq \int_0^\infty \frac{1}{\Gamma(a)}x^{a-1}e^{-Rx}dx \\ &= R^{-a}\int_0^\infty \frac{1}{\Gamma(a)}y^{a-1}e^{-y}dy \\ &= R^{-a}.\end{aligned}$$

In particular, $\Pr[Z \leq C] \geq C^a$, which means $\Pr[Z \geq C] \leq 1 - C^a \leq a\log(1/C)$.

For $\mathbb{E}[Z\delta(Z<C)]$, the proof is similar as before:

$$P = \mathbb{E}[Z\delta(Z<C)] = \int_0^C \frac{1}{B(a,b)}x^a(1-x)^b dx \leq \frac{C^{a+1}}{B(a,b)(a+1)},$$

$$Q = \mathbb{E}[Z\delta(Z \geq C)] = \int_C^1 \frac{1}{B(a,b)}x^a(1-x)^b dx \geq \frac{C^a(1-C)^{b+1}}{B(a,b)(b+1)}.$$

Now $\mathbb{E}[Z\delta(Z \leq C)] \leq \frac{P}{Q}\mathbb{E}[Z] \leq 6aC$ when $C < 1/(b+1)$. ■

C.5.2 Norm Bounds

Lemma 27 (Norm Bounds under Dirichlet distribution) *For $\pi_i \overset{iid}{\sim} \mathrm{Dir}(\alpha)$ for $i \in A$, with probability $1-\delta$, we have*

$$\begin{aligned}\sigma_{\min}(\Pi_A) &\geq \sqrt{\frac{|A|\widehat{\alpha}_{\min}}{\alpha_0+1}} - O((|A|\log k/\delta)^{1/4}), \\ \|\Pi_A\| &\leq \sqrt{|A|\widehat{\alpha}_{\max}} + O((|A|\log k/\delta)^{1/4}), \\ \kappa(\Pi_A) &\leq \sqrt{\frac{(\alpha_0+1)\widehat{\alpha}_{\max}}{\widehat{\alpha}_{\min}}} + O((|A|\log k/\delta)^{1/4}).\end{aligned}$$

This implies that $\|F_A\| \leq \|P\|\sqrt{|A|\widehat{\alpha}_{\max}}$, $\kappa(F_A) \leq O(\kappa(P)\sqrt{(\alpha_0+1)\widehat{\alpha}_{\max}/\widehat{\alpha}_{min}})$. Moreover, with probability $1-\delta$

$$\|F_A\|_1 \leq |A| \cdot \max_i (P\widehat{\alpha})_i + O\left(\|P\|\sqrt{|A|\log\frac{|A|}{\delta}}\right). \tag{102}$$

Remark: *When* $|A| = \Omega\left(\log \frac{k}{\delta}\left(\frac{\alpha_0+1}{\widehat{\alpha}_{\min}}\right)^2\right)$, *we have* $\sigma_{\min}(\Pi_A) = \Omega(\sqrt{\frac{|A|\widehat{\alpha}_{\min}}{\alpha_0+1}})$ *with probability* $1-\delta$ *for any fixed* $\delta \in (0,1)$.
Proof: Consider $\Pi_A \Pi_A^\top = \sum_{i\in A} \pi_i \pi_i^\top$.

$$\begin{aligned}\frac{1}{|A|}\mathbb{E}[\Pi_A \Pi_A^\top] =& \mathbb{E}_{\pi\sim Dir(\alpha)}[\pi\pi^\top] \\ =& \frac{\alpha_0}{\alpha_0+1}\widehat{\alpha}\widehat{\alpha}^\top + \frac{1}{\alpha_0+1}\operatorname{Diag}(\widehat{\alpha}),\end{aligned}$$

from Proposition 29. The first term is positive semi-definite so the eigenvalues of the sum are at least the eigenvalues of the second component. Smallest eigenvalue of second component gives lower bound on $\sigma_{\min}(\mathbb{E}[\Pi_A \Pi_A^\top])$. The spectral norm of the first component is bounded by $\frac{\alpha_0}{\alpha_0+1}\|\hat{\alpha}\| \le \frac{\alpha_0}{\alpha_0+1}\widehat{\alpha}_{\max}$, the spectral norm of second component is $\frac{1}{\alpha_0+1}\alpha_{\max}$. Thus $\left\|\mathbb{E}[\Pi_A \Pi_A^\top]\right\| \le |A| \cdot \widehat{\alpha}_{\max}$.

Now applying Matrix Bernstein's inequality to $\frac{1}{|A|}\sum_i \left(\pi_i\pi_i^\top - \mathbb{E}[\pi\pi^\top]\right)$. We have that the variance is $O(1/|A|)$. Thus with probability $1-\delta$,

$$\left\|\frac{1}{|A|}\left(\Pi_A\Pi_A^\top - \mathbb{E}[\Pi_A\Pi_A^\top]\right)\right\| = O\left(\sqrt{\frac{\log(k/\delta)}{|A|}}\right).$$

For the result on F, we use the property that for any two matrices A, B, $\|AB\| \le \|A\|\,\|B\|$ and $\kappa(AB) \le \kappa(A)\kappa(B)$.

To show bound on $\|F_A\|_1$, note that each column of F_A satisfies $\mathbb{E}[(F_A)_i] = \langle \widehat{\alpha}, (P)_i\rangle\, 1^\top$, and thus $\|\mathbb{E}[F_A]\|_1 \le |A| \max_i (P\widehat{\alpha})_i$. Using Bernstein's inequality, for each column of F_A, we have, with probability $1-\delta$,

$$\left|\, \|(F_A)_i\|_1 - |A|\left\langle \widehat{\alpha}, (P)^i\right\rangle\right| = O\left(\|P\|\sqrt{|A|\log\frac{|A|}{\delta}}\right),$$

by applying Bernstein's inequality, since $|\left\langle \widehat{\alpha}, (P)^i\right\rangle| \le \|P\|$, and thus we have $\sum_{i\in A}\|\mathbb{E}[(P)^j\pi_i\pi_i^\top((P)^j)^\top]\|$, and $\sum_{i\in A}\|\mathbb{E}[\pi_i^\top((P)^j)^\top(P)^j\pi_i]\| \le |A|\cdot\|P\|$. ■

C.5.3 Properties of Gamma and Dirichlet Distributions

Recall Gamma distribution $\Gamma(\alpha,\beta)$ is a distribution on nonnegative real values with density function $\frac{\beta^\alpha}{\Gamma(\alpha)}x^{\alpha-1}e^{-\beta x}$.

Proposition 28 (Dirichlet and Gamma distributions) *The following facts are known for Dirichlet distribution and Gamma distribution.*

1. *Let* $Y_i \sim \Gamma(\alpha_i, 1)$ *be independent random variables, then the vector* $(Y_1, Y_2, ..., Y_k)/\sum_{i=1}^k Y_k$ *is distributed as* $Dir(\alpha)$.
2. *The* Γ *function satisfies Euler's reflection formula:* $\Gamma(1-z)\Gamma(z) \le \pi/\sin\pi_z$.
3. *The* $\Gamma(z) \ge 1$ *when* $0 < z < 1$.

4. *There exists a universal constant C such that $\Gamma(z) \leq C/z$ when $0 < z < 1$.*

5. *For $Y \sim \Gamma(\alpha, 1)$ and $t > 0$ and $\alpha \in (0, 1)$, we have*

$$\frac{\alpha}{4C} t^{\alpha-1} e^{-t} \leq \Pr[Y \geq t] \leq t^{\alpha-1} e^{-t}, \tag{103}$$

and for any $\eta, c > 1$, we have

$$\mathbb{P}[Y > \eta t | Y \geq t] \geq (c\eta)^{\alpha-1} e^{-(\eta-1)t}. \tag{104}$$

Proof: The bounds in (103) is derived using the fact that $1 \leq \Gamma(\alpha) \leq C/\alpha$ when $\alpha \in (0, 1)$ and

$$\int_t^\infty \frac{1}{\Gamma(\alpha_i)} x^{\alpha_i - 1} e^{-x} dx \leq \frac{1}{\Gamma(\alpha_i)} \int_t^\infty t^{\alpha_i - 1} e^{-x} dx \leq t^{\alpha_i - 1} e^{-t},$$

and

$$\int_t^\infty \frac{1}{\Gamma(\alpha_i)} x^{\alpha_i - 1} e^{-x} dx \geq \frac{1}{\Gamma(\alpha_i)} \int_t^{2t} x^{\alpha_i - 1} e^{-x} dx \geq \alpha_i / C \int_t^{2t} (2t)^{\alpha_i - 1} e^{-x} dx \geq \frac{\alpha_i}{4C} t^{\alpha_i - 1} e^{-t}.$$

■

Proposition 29 (Moments under Dirichlet distribution) *Suppose $v \sim Dir(\alpha)$, the moments of v satisfies the following formulas:*

$$\begin{aligned} \mathbb{E}[v_i] &= \frac{\alpha_i}{\alpha_0} \\ \mathbb{E}[v_i^2] &= \frac{\alpha_i(\alpha_i + 1)}{\alpha_0(\alpha_0 + 1)} \\ \mathbb{E}[v_i v_j] &= \frac{\alpha_i \alpha_j}{\alpha_0(\alpha_0 + 1)}, \quad i \neq j. \end{aligned}$$

More generally, if $a^{(t)} = \prod_{i=0}^{t-1}(a + i)$, then we have

$$\mathbb{E}[\prod_{i=1}^k v_i^{(a_i)}] = \frac{\prod_{i=1}^k \alpha_i^{(a_i)}}{\alpha_0^{(\sum_{i=1}^k a_i)}}.$$

C.6 Standard Results

C.6.1 Bernstein's Inequalities

One of the key tools we use is the standard matrix Bernstein inequality (Tropp, 2012, thm. 1.4).

Proposition 30 (Matrix Bernstein Inequality) *Suppose $Z = \sum_j W_j$ where*

1. *W_j are independent random matrices with dimension $d_1 \times d_2$,*
2. *$\mathbb{E}[W_j] = 0$ for all j,*

3. $\|W_j\| \leq R$ *almost surely.*

Let $d = d_1 + d_2$*, and* $\sigma^2 = \max\left\{\left\|\sum_j \mathbb{E}[W_j W_j^\top]\right\|, \left\|\sum_j \mathbb{E}[W_j^\top W_j]\right\|\right\}$*, then we have*

$$\Pr[\|Z\| \geq t] \leq d \cdot exp\left\{\frac{-t^2/2}{\sigma^2 + Rt/3}\right\}.$$

Proposition 31 (Vector Bernstein Inequality) *Let* $z = (z_1, z_2, ..., z_n) \in \mathbb{R}^n$ *be a random vector with independent entries,* $\mathbb{E}[z_i] = 0$*,* $\mathbb{E}[z_i^2] = \sigma_i^2$*, and* $\Pr[|z_i| \leq 1] = 1$*. Let* $A = [a_1|a_2|\cdots|a_n] \in \mathbb{R}^{m\times n}$ *be a matrix, then*

$$\Pr[\|Az\| \leq (1+\sqrt{8t})\sqrt{\sum_{i=1}^n \|a_i\|^2 \sigma_i^2} + (4/3)\max_{i\in[n]}\|a_i\| t] \geq 1 - e^{-t}.$$

C.6.2 Vector Chebyshev Inequality

We will require a vector version of the Chebyshev inequality Ferentios (1982).

Proposition 32 *Let* $z = (z_1, z_2, ..., z_n) \in \mathbb{R}^n$ *be a random vector with independent entries,* $\mathbb{E}[z_i] = \mu$*,* $\sigma := \|\operatorname{Diag}(\mathbb{E}[(z-\mu)^\top(z-\mu)])\|$*. Then we have that*

$$\mathbb{P}[\|z - \mu\| > t\sigma] \leq t^{-2}.$$

C.6.3 Wedin's Theorem

We make use of Wedin's theorem to control subspace perturbations.

Lemma 33 (Wedin's theorem; Theorem 4.4, p. 262 in Stewart and Sun (1990).) *Let* $A, E \in \mathbb{R}^{m\times n}$ *with* $m \geq n$ *be given. Let* A *have the singular value decomposition*

$$\begin{bmatrix} U_1^\top \\ U_2^\top \\ U_3^\top \end{bmatrix} A \begin{bmatrix} V_1 & V_2 \end{bmatrix} = \begin{bmatrix} \Sigma_1 & 0 \\ 0 & \Sigma_2 \\ 0 & 0 \end{bmatrix}.$$

Let $\tilde{A} := A + E$*, with analogous singular value decomposition* $(\tilde{U}_1, \tilde{U}_2, \tilde{U}_3, \tilde{\Sigma}_1, \tilde{\Sigma}_2, \tilde{V}_1 \tilde{V}_2)$*. Let* Φ *be the matrix of canonical angles between* $\operatorname{range}(U_1)$ *and* $\operatorname{range}(\tilde{U}_1)$*, and* Θ *be the matrix of canonical angles between* $\operatorname{range}(V_1)$ *and* $\operatorname{range}(\tilde{V}_1)$*. If there exists* $\delta, \alpha > 0$ *such that* $\min_i \sigma_i(\tilde{\Sigma}_1) \geq \alpha + \delta$ *and* $\max_i \sigma_i(\Sigma_2) \leq \alpha$*, then*

$$\max\{\|\sin\Phi\|_2, \|\sin\Theta\|_2\} \leq \frac{\|E\|_2}{\delta}.$$

References

I. Abraham, S. Chechik, D. Kempe, and A. Slivkins. Low-distortion inference of latent similarities from a multiplex social network. *arXiv preprint arXiv:1202.0922*, 2012.

N. Ailon, Y. Chen, and X. Huan. Breaking the small cluster barrier of graph clustering. *arXiv preprint arXiv:1302.4549*, 2013.

E. M. Airoldi, D. M. Blei, S. E. Fienberg, and E. P. Xing. Mixed membership stochastic blockmodels. *Journal of Machine Learning Research*, 9:1981–2014, June 2008.

A. Anandkumar, D. P. Foster, D. Hsu, S. M. Kakade, and Y. Liu. Two svds suffice: Spectral decompositions for probabilistic topic modeling and latent dirichlet allocation. *arXiv preprint arXiv:1202.0922*, 2012a.

A. Anandkumar, R. Ge, D. Hsu, S. M. Kakade, and M. Telgarsky. Tensor decompositions for latent variable models. *arXiv preprint arXiv:1210.7559*, 2012b.

A. Anandkumar, D. Hsu, and S. M. Kakade. A method of moments for mixture models and hidden markov models. In *Proc. of Conf. on Learning Theory*, June 2012c.

S. Arora, R. Ge, S. Sachdeva, and G. Schoenebeck. Finding overlapping communities in social networks: toward a rigorous approach. In *Proceedings of the 13th ACM Conference on Electronic Commerce*, 2012.

M. F. Balcan, C. Borgs, M. Braverman, J. T. Chayes, and S. Teng. I like her more than you: Self-determined communities. *arXiv preprint arXiv:1201.4899*, 2012.

P. J. Bickel and A. Chen. A nonparametric view of network models and Newman-Girvan and other modularities. *Proceedings of the National Academy of Sciences*, 106(50):21068–21073, 2009.

P. J. Bickel, A. Chen, and E. Levina. The method of moments and degree distributions for network models. *The Annals of Statistics*, 39(5):38–59, 2011.

D. M. Blei. Probabilistic topic models. *Communications of the ACM*, 55(4):77–84, 2012.

D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, March 2003.

B. Bollobás, S. Janson, and O. Riordan. The phase transition in inhomogeneous random graphs. *Random Structures & Algorithms*, 31(1):3–122, 2007.

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

B. S. Charles and S. S. Vempala. Random tensors and planted cliques. In *RANDOM*, 2009.

S. Chatterjee and P. Diaconis. Estimating and understanding exponential random graph models. *arXiv preprint arXiv:1102.2650*, 2011.

K. Chaudhuri, F. Chung, and A. Tsiatas. Spectral clustering of graphs with general degrees in the extended planted partition model. *Journal of Machine Learning Research*, pages 1–23, 2012.

Y. Chen, S. Sanghavi, and H. Xu. Clustering sparse graphs. In *Advances in Neural Information Processing*, 2012.

A. Condon and R. M. Karp. Algorithms for graph partitioning on the planted partition model. In *Randomization, Approximation, and Combinatorial Optimization. Algorithms and Techniques*, pages 221–232. Springer, 1999.

S. Currarini, M. O. Jackson, and P. Pin. An economic model of friendship: Homophily, minorities, and segregation. *Econometrica*, 77(4):1003–1045, 2009.

V. Feldman, E. Grigorescu, L. Reyzin, S. Vempala, and Y. Xiao. Statistical algorithms and a lower bound for planted clique. *Electronic Colloquium on Computational Complexity (ECCC)*, 19:64, 2012.

K. Ferentios. On Tcebycheff's type inequalities. *Trabajos de Estadística y de Investigación Operativa*, 33(1):125–132, 1982.

S. E. Fienberg, M. M. Meyer, and S. S. Wasserman. Statistical analysis of multiple sociometric relations. *Journal of the American Statistical Association*, 80(389):51–67, 1985.

O. Frank and D. Strauss. Markov graphs. *Journal of the American Statistical Association*, 81(395):832–842, 1986.

A. M. Frieze and R. Kannan. Quick approximation to matrices and applications. *Combinatorica*, 19(2):175–220, 1999.

A. M. Frieze and R. Kannan. A new approach to the planted clique problem. In *FSTTCS*, 2008.

M. Girvan and M. E. J. Newman. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, 99(12):7821–7826, 2002.

A. Gittens and M. W. Mahoney. Revisiting the Nystrom method for improved large-scale machine learning. *arXiv preprint arXiv:1303.1849*, 2013.

P. Gopalan, D. Mimno, S. Gerrish, M. Freedman, and D. Blei. Scalable inference of overlapping communities. In *Advances in Neural Information Processing Systems 25*, pages 2258–2266, 2012.

C. Hillar and L.-H. Lim. Most tensor problems are NP hard. *Journal of the ACM*, 60(6): 45:1–45:39, nov 2013.

M. Hoffman, D. M. Blei, C. Wang, and J. Paisley. Stochastic variational inference. *JMLR*, 14:1303–1347, 2012.

P. W. Holland and S. Leinhardt. An exponential family of probability distributions for directed graphs. *Journal of the American Statistical Association*, 76(373):33–50, 1981.

P. W. Holland, K. B. Laskey, and S. Leinhardt. Stochastic blockmodels: first steps. *Social Networks*, 5(2):109–137, 1983.

F. Huang, U. N. Niranjan, M. Hakeem, and A. Anandkumar. Fast detection of overlapping communities via online tensor methods. *arXiv preprint arXiv:1309.0787*, Sept. 2013.

J. JáJá. *An Introduction to Parallel Algorithms*. Addison Wesley Longman Publishing Co., Inc., 1992.

A. Jalali, Y. Chen, S. Sanghavi, and H. Xu. Clustering partially observed graphs via convex optimization. *arXiv preprint arXiv:1104.4803*, 2011.

M. J. Kearns and U. V. Vazirani. *An Introduction to Computational Learning Theory*. MIT Press., Cambridge, MA, 1994.

T. G. Kolda and B. W. Bader. Tensor decompositions and applications. *SIAM Review*, 51 (3):455, 2009.

T. G. Kolda and J. R. Mayo. Shifted power method for computing tensor eigenpairs. *SIAM Journal on Matrix Analysis and Applications*, 32(4):1095–1124, October 2011.

P. F. Lazarsfeld, R. K. Merton, et al. Friendship as a social process: A substantive and methodological analysis. *Freedom and Control in Modern Society*, 18(1):18–66, 1954.

Z. Lin, M. Chen, and Y. Ma. The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices. *arXiv preprint arXiv:1009.5055*, 2010.

L. Lovász. Very large graphs. *Current Developments in Mathematics*, 2008:67–128, 2009.

M. McPherson, L. Smith-Lovin, and J. M. Cook. Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, pages 415–444, 2001.

F. McSherry. Spectral partitioning of random graphs. In *FOCS*, 2001.

J. L. Moreno. *Who shall survive?: A new approach to the problem of human interrelations.* Nervous and Mental Disease Publishing Co, 1934.

E. Mossel, J. Neeman, and A. Sly. Stochastic block models and reconstruction. *arXiv preprint arXiv:1202.1499*, 2012.

G. Palla, I. Derényi, I. Farkas, and T. Vicsek. Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 435(7043):814–818, 2005.

K. Pearson. Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society, London, A.*, page 71, 1894.

A. Rinaldo, S. E. Fienberg, and Y. Zhou. On the geometry of discrete exponential families with application to exponential random graph models. *Electronic Journal of Statistics*, 3:446–484, 2009.

T. A. B. Snijders and K. Nowicki. Estimation and prediction for stochastic blockmodels for graphs with latent block structure. *Journal of Classification*, 14(1):75–100, 1997.

G. W. Stewart and J. Sun. *Matrix Perturbation Theory*, volume 175. Academic press New York, 1990.

M. Telgarsky. Dirichlet draws are sparse with high probability. *arXiv preprint arXiv:1301.4917*, 2012.

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

Y. J. Wang and G. Y. Wong. Stochastic blockmodels for directed graphs. *Journal of the American Statistical Association*, 82(397):8–19, 1987.

H. C. White, S. A. Boorman, and R. L. Breiger. Social structure from multiple networks. i. blockmodels of roles and positions. *American Journal of Sociology*, pages 730–780, 1976.

E. P. Xing, W. Fu, and L. Song. A state-space mixed membership blockmodel for dynamic network tomography. *The Annals of Applied Statistics*, 4(2):535–566, 2010.