Learning Mixed Membership Community Models in Social Tagging Networks through Tensor Methods

Anima Anandkumar* Hanie Sedghi†

January 22, 2016

Abstract

Community detection in graphs has been extensively studied both in theory and in applications. However, detecting communities in hypergraphs is more challenging. In this paper, we propose a tensor decomposition approach for guaranteed learning of communities in a special class of hypergraphs modeling social tagging systems or *folksonomies*. A folksonomy is a tripartite 3-uniform hypergraph consisting of (user, tag, resource) hyperedges. We posit a probabilistic *mixed membership* community model, and prove that the tensor method consistently learns the communities under efficient sample complexity and separation requirements.

Keywords: Community models, social tagging systems/folksonomies, mixed membership models, tensor decomposition methods.

1 Introduction

Folksonomies or social tagging systems (Chakraborty et al., 2012) have been hugely popular in recent years. These are tripartite networks consisting of users, resources and tags. The resources can vary according to the system. For instance, in Delicious, the URLs are the resources, in Flickr, they are the images, in LastFm, they are the music files, in MovieLens, they are the reviews, and so on. The collaborative annotation of these resources by users with descriptive keywords, enables faster search and retrieval (Chakraborty and Ghosh, 2013).

The role of community detection in folksonomies cannot be overstated. Online social tagging systems are growing rapidly and it is important to group the nodes (i.e. users, resources and tags) for scalable operations in a number of applications such as personalized search (Xu et al., 2008), resource and friend recommendations (Konstas et al., 2009), and so on. Moreover, learning communities can provide an understanding of community formation behavior of humans, and the role of communities in human interaction and collaboration in online systems.

Folksonomies are special instances of hypergraphs. A folksonomy is a tripartite 3-uniform hypergraph consisting of hyperedges between users, resources and tags. Scalable community detection in hypergraphs is in general challenging, and most previous works are limited to pure membership models, where a node belongs to at most one group. This is highly unrealistic since users have multiple interests, and the tags

*University of California, Irvine, Email: a.anandkumar@uci.edu

†University of Southern California, Email: hsedghi@usc.edu

and resources have multiple contexts or topics. A few works which do consider overlapping communities in folksonomies are heuristic without any guarantees and do not incorporate any statistical modeling (see Section 1.2 for details).

In this paper, we propose a novel probabilistic approach for modeling folksonomies, and propose a guaranteed approach for detecting overlapping communities in them. A naive model for folksnomies would result in a large number of model parameters, and make learning intractable. Here we present a more scalable approach where realistic conditional independence constraints are imposed, leading to scalable modeling and tractable learning.

Our model is a hypergraph extension of the popular *mixed membership stochastic blockmodel* (MMSB), introduced by Airoldi et. al (Airoldi et al., 2008). We impose additional conditional independence constraints, which are natural for social tagging systems. We term our model as *mixed membership stochastic folksonomy* (MMSF). When hypergraphs are generated from such a class of MMSFs, we show that the hyper-edges can be much *more informative* about the underlying communities, than in the graph setting. Intuitively, this is because the hyper-edges represent multiple *views* of the hidden communities. In this paper, we show that these properties can be exploited for learning via spectral approaches.

1.1 Summary of Results

We develop a practically relevant mixed membership hypergraph model and propose novel methods to learn them with guarantees. We posit a probabilistic model for generation of hyper-edges $\{r, u, t\}$ between resources r, users u and tags t. We impose natural conditional independence assumptions that conditioned on the community memberships of individual nodes, the hyperedge generations are independent. In addition, we assume that the users select tags for a given resource, based on the context in which the resource is accessed. For instance, consider the resource as a paper that falls both in theoretical and applied machine learning, as shown in Figure 1. If a user accesses the resource under the context of theory, he/she uses tags that are indicative of theory. Note that we allow the users and tags to be in multiple communities; however, the actual realization of an hyper-edge depends only on the context in which the resource was accessed. Depending on what kind of user is tagging the paper, the likelihood of choosing various tags such as application, latent variable model etc changes. The conditional independence assumption states that once a user accesses the paper in certain context (e.g. looking for applications), the probability of using tags in a category (e.g. applications, experiments) only depends on that context. There are many other such examples. For example, a movie can be a drama about a political figure. A person who is mostly into politics will watch this movie in the context of politics and use political tags (for example name of the person, specific political events that where illustrated in the movie), while a person who is more into drama genre will use drama to tag the movie.

While community models on general hypergraphs is NP hard, our setting is geared towards the setting of folksonomies with users, resources and tags, and the assumptions we make naturally hold in this setting. Importantly, we allow for general distributions for mixed community memberships. The earlier work by Anandkumar et al. (2014a) on MMSB models on graphs is limited to the Dirichlet distribution. Note that the Dirichlet assumption for community memberships can be limiting and cannot model general correlations in memberships. Without the Dirichlet assumption, the earlier techniques, when applied directly, would yield tensors in the Tucker form, which do not possess a unique decomposition and thus, the communities cannot be learnt from the tensor forms. In addition, our moment forms are different since it is the hypergraph setting and conditional independence assumptions are different. Thus, earlier work on MMSB cannot be directly applied here.

In addition, we impose weak assumptions on the distribution of the community memberships. This is

required since the memberships are in general not identifiable when they are mixed. While the original MMSB model (Airoldi et al., 2008) assumes that the communities are drawn from a Dirichlet distribution, here, we do not require such a strong parametric assumption. Here, we impose a weak assumption that a certain fraction of resource nodes are "pure" and belong to a single community. This is reasonable to expect in practice. We establish that the communities are identifiable under these natural assumptions, and can be learnt efficiently using spectral approaches.

Here, we propose a novel algorithm to detect pure nodes belonging to a single community. The presence of pure nodes is natural to expect in practice and does not require the Dirichlet assumption. Our method consists of two main routines. First, we design a simple rank test to identify pure resource nodes. The algorithm involves first projecting hyperedges to subspace of top-k eigenvectors. It then involves performing rank test on the matricization of connectivity vectors of each resource node, where rows correspond to users and columns correspond to tags. We can then exploit these detected pure nodes to form tensors that can be decomposed efficiently to yield the communities for all the nodes (and not just the pure nodes). We prove that our proposed method correctly recovers the parameters of the MMSF model when exact moments are input. This two stage algorithm is expected to have much wider applicability than the MMSB model which is limited to the Dirichlet distribution. For this general model, we show a tight sample complexity that $n > k^3$ can recover the communities.

For the first step, we construct a matrix for each resource node, consisting of its edges to users and tags. We show that this matrix is rank-1 in expectation (over the hyperedges) for a pure resource node. This property enables us to identify such pure nodes. We then construct a 3-star count tensor using these estimated pure resource nodes. We count the pure resource nodes, which are common to triplets of (user,tag) tuples to form the tensor. We show that in expectation this tensor has a CP decomposition form, and requiring this decomposition yields the community memberships after some simple post-processing steps.

We then carefully analyze the perturbation bounds under empirical moments, and show that the communities can be accurately recovered under some natural assumptions. The perturbation analysis for this step is novel since it requires analyzing the effect of standard spectral perturbations on matricization and the subsequent rank test. We use subexponential Hanson Wright inequalities to obtain tight guarantees for this step. These assumptions determine how the number of nodes n is related to the number of communities k, and a lower bound on the separation $p - q$, where p denotes the connectivity within the same community, while q denotes the connectivity across different communities. Such requirements have been imposed before in the graph setting, for stochastic block models (Yudong et al., 2012) and mixed membership models (Anandkumar et al., 2014a). Here, we show that for MMSF, the requirement is stronger, since intuitively, we require concentration on a hypergraph instead of a graph. We employ sub-exponential forms of Hanson Wright's inequality to get tight bounds in the sparse regime, where the connectivity probabilities p, q are small. Thus, we obtain efficient guarantees for recovering mixed membership communities from social tagging networks.

We establish that for the success of rank test, if $p \simeq q$, we need the network size to scale as $n = \tilde{\Omega}\left(k^3\right)$ (when the correlation matrix of community membership distribution is well-conditioned). For the case where $q < p/k$, we require $n = \tilde{\Omega}\left(k^2\right)$. This is intuitive as the role of q is to make the different community components non-orthogonal for the rank test, i.e., q acts as noise. Therefore, a smaller q results in better guarantees. For the success of tensor decomposition method, we require $n = \tilde{\Omega}\left(k^3\right)$, when p, q are constants, in the well-conditioned setting. Note that in comparison, for learning mixed membership stochastic block model graphs, we require $n = \tilde{\Omega}\left(k^2\right)$, from Anandkumar et al. (2014a), which is lower sample complexity. This is because we need to learn more number of parameters in the hypergraph setting. Moreover, for sparse graphs, the parameters p, q decay with n, and we also handle this setting, and provide the precise bounds in Section 4.

1.2 Related Work

There is an extensive body of work for community detection in graphs. Popular methods with guarantees include spectral clustering (McSherry, 2001) and convex optimization (Yudong et al., 2012). For a detailed survey, see (Anandkumar et al., 2014a). However, these methods cannot handle mixed membership models, where a node can belong to more than one community.

Our algorithm is based on the tensor decomposition approach of (Anandkumar et al., 2013) for pairwise MMSB model in graphs. The method has been implemented for many real-world datasets and has shown significant improvement in running times and accuracy over the state of art stochastic variational techniques (Huang et al., 2013). The tensor consists of 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. The MMSB model assumes a Dirichlet distribution for community memberships, and in this case, a modified 3-star count tensor is used. It is shown that this tensor has a *CP*-decomposition form, and the components of the decomposition can be used to learn the parameters of the MMSB model. However, this method cannot be extended easily to general distributions, beyond the Dirichlet assumption, since for general distributions, the 3-star count tensor only has a Tucker decomposition form, and not a CP form. In general, the model parameters are not identifiable from a Tucker form. Thus, in graphs, mixed membership models cannot be easily learnt when general distributions (beyond the Dirichlet distribution) for mixed memberships are assumed. In this paper, we show that in the hypergraph setting, more general distributions of community memberships can be learnt, when certain conditional independence relationships are assumed for hyper-edge generation.

Another limitation of the MMSB model is that due to the Dirichlet assumption, only normalized community memberships can be incorporated. However, in this case, the mixed nodes (i.e. those belonging to more than one community) are less densely connected than the pure nodes, as pointed out by (Yang and Leskovec, 2013). In contrast, in our paper, we can handle un-normalized community memberships vectors (in a weighted graph), since we do not make the Dirichlet assumption, and thus, this limitation is not present. However, for simplicity, we present the results in the normalized setting.

Scalable community detection in hypergraphs is in general challenging and most previous works are limited to pure membership models, where a node belongs to at most one group (Brinkmeier et al., 2007; Lin et al., 2009; Murata, 2010; Neubauer and Obermayer, 2009; Vazquez, 2009). Clustering in multipartite hypergraphs can be seen as extensions of the *co-clustering* of matrices, where rows and columns are simultaneously clustered. In (Jegelka et al., 2009), extensions of co-clustering to the tensor setting is considered. However, this setting can only handle pure communities, where a node belongs to at most one community. A few works which do consider mixed communities in hypergraphs are heuristic without any guarantees, and do not incorporate any statistical modeling (Wang et al., 2010; Chakraborty et al., 2012; Papadopoulos et al., 2010). They mostly use modularity based scores without providing any guarantees. In this paper, we present the first guaranteed method for learning communities in mixed membership hypergraphs.

2 Mixed Membership Model for Folksonomies

Setup: We consider folksonomies modeled as tripartite 3-uniform hypergraphs over three sets of nodes, viz., set of users U, set of tags T and set of resources R. An hyperedge $\{u, t, r\}$ occurs when user u tags resource r with tag t. For convenience, we will consider a *matricized* version of the $\{0, 1\}$ hyper-adjacency tensor, denoted by $\widehat{G} \in \{0, 1\}^{|U|\cdot|T|\times|R|}$, which indicates the presence of hyper-edges. The reason behind considering matricization along the resource *mode* will soon become clear. We use the notation $\widehat{G}(\{u, t\}, r)$ to denote the entry corresponding to the hyper-edge $\{u, t, r\}$, and $\widehat{G}(\{U, T\}, r)$ to denote the column vector

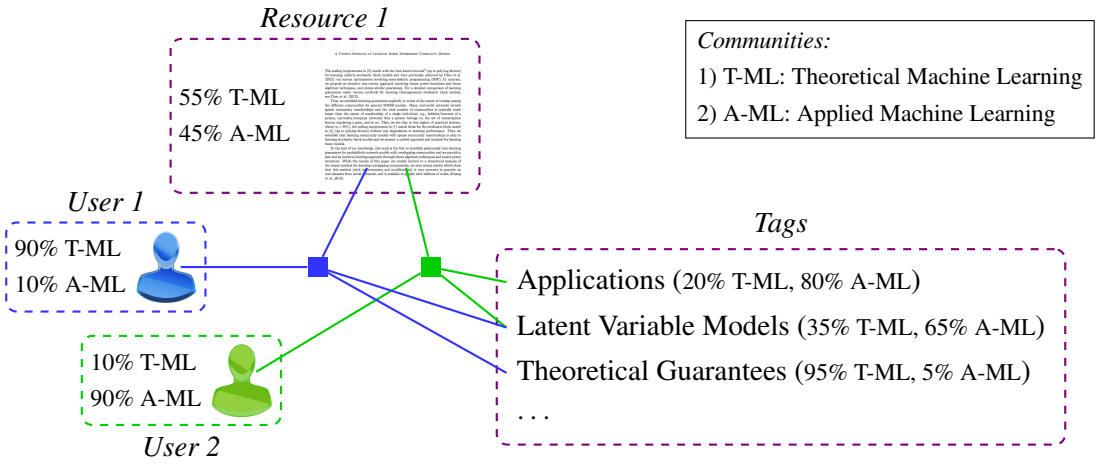

Figure 1: Overview of MMSF model for an example of machine learning articles (resources) tagged by users. One article (resource) and the corresponding tags by two users are shown. Two communites of Theoretical machine learning and Applied machine learning are assumed. The mixed community membership of resources, users and tags are also shown.

corresponding to the set of hyper-edges $\{U, T, r\}$.

We consider models with k underlying (hidden) communities and let $[k] := \{1, 2, \ldots, k\}$. 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. Let $\Pi_U := [\pi_i : i \in U] \in \mathbb{R}^{k \times |U|}$ denote the set of column vectors denoting the community memberships of users in U, and similarly define Π_T and Π_R. Let $\Pi := [\pi_i : i \in U \cup T \cup R]$.

We now provide a statistical model to explain the presence of hyper-edges $\{u, t, r\}$ among users, tags and resources through the community memberships. We consider a mixed memberships model, where there are multiple communities for users, tags and resources. Intuitively, users belonging to certain groups (i.e. interested in certain topics) will tend to select resources mainly comprised of those topics. The tags employed by the users are dependent on the *contextual* category of the resource selected by the user. This intuition is formalized under our proposed statistical model below.

Let $z_{u \to \{t,r\}} \in \mathbb{R}^k$ be a coordinate basis vector which denotes the community membership of user u when posting tag t and resource r, and similarly let $z_{r \to \{u,t\}}$, $z_{t \to \{u,r\}}$ denote the memberships of resource r and tag t when participating in the hyperedge $\{u, t, r\}$.

Let $P \in \mathbb{R}^{k \times k}$ be the community connectivity matrix, where $P_{i,j}$ denotes the probability that a user in community i selects a resource in community j. Similarly, let $\tilde{P} \in \mathbb{R}^{k \times k}$ denote a matrix such that each entry $\tilde{P}_{i,j}$ denotes the probability that a tag in community i is associated with resource in community j.

The proposed mixed membership stochastic folksonomy (MMSF) is as follows:

- For each node in $i \in U \cup T \cup R$, draw its community membership vector $\pi_i \in \mathbb{R}^k$, i.i.d. from some distribution f_π.
- For each triplet $\{u, t, r\}$, draw coordinate basis vectors $z_{u \to \{t,r\}} \sim \text{Multinomial}(\pi_u)$, $z_{t \to \{u,r\}} \sim \text{Multinomial}(\pi_t)$ and $z_{r \to \{u,t\}} \sim \text{Multinomial}(\pi_r)$ in a conditionally independent manner, given Π.

- Draw random variables

$$\widehat{B}_{r\to u;t} \sim \text{ Bernoulli}(z_{u\to\{t,r\}}^{\top} P z_{r\to\{u,t\}})$$
$$\widehat{B}_{r\to t;u} \sim \text{ Bernoulli}(z_{t\to\{u,r\}}^{\top} \tilde{P} z_{r\to\{u,t\}}). \quad (1)$$

The presence of hyper-edge $G(\{u,t\},r)$ is given by the product

$$\widehat{G}(\{u,t\},r) = \widehat{B}_{r\to u;t} \cdot \widehat{B}_{r\to t;u}. \quad (2)$$

The use of variables $z_{u\to\{t,r\}}$, $z_{t\to\{u,r\}}$ and $z_{r\to\{u,t\}}$ allows for *context-dependent* selection of group memberships as in the MMSB model. Given a resource and its context, a user may choose to access the resource, and probability of using a tag on a resource depends on context of the tag and the resource. Given the context of user, tag and the resource, these two events are independent. In order to have a hyper-edge, we need both events to happen and this explains Eqn. (2).

Ours is a resource centric model, where a resource can be regarded as comprising of many *topics* or communities. Which tags get associated with the resource is dependent on the context of the resource $z_{r\to\{u,t\}}$ and the tag $z_{t\to\{u,r\}}$ and similarly, which user selects a resource is dependent on the context of the user $z_{u\to\{t,r\}}$ and the resource $z_{r\to\{u,t\}}$. The hyper-edges are drawn according to (2) and thus, matricization along the resource mode is convenient for analysis. Our model is resource centric and not user centric. The intuition is that the tags associated with a resource are dependent on the context that the resource is being accessed and the likelihood of the user accessing a resource is dependent on his/her current group and the context of the resource. Figure 1 provides an instance of a hypergraph where the resource is a paper and communities consist of theoretical and applied machine learning.

Unlike the pairwise MMSB model (Airoldi et al., 2008), where the edges are conditionally independent given the community memberships, in the proposed MMSF model, the edges $\widehat{B}_{r\to t;u}$ and $\widehat{B}_{r\to u;t}$ contained in the hyperedge $\{u,t,r\}$ are *not* conditionally independent given the community memberships, since they are selected based on the common context $z_{r\to\{u,t\}}$ of the resource r. Thus, the MMSF model is capturing dependencies beyond the pairwise MMSB model. At the same time, the MMSF model has conditionally independent hyperedges given the community memberships, which leads to tractable learning.

We do *not* take the approach of modeling hyperedges directly, i.e., through a community connectivity tensor in $\tilde{P} \in \mathbb{R}^{k\times k\times k}$, where $\tilde{P}_{a,b,c}$ would give the probability that a user in community a would have an hyperedge with resource b and tag c. This would lead to k^3 unknown parameters, while our model has only k^2 unknown parameters. Moreover, if the user at a certain point is interested in some topic (i.e. draws $z_{u\to\{t,r\}}$ in some community), then he looks for resources and tags having significant membership in that topic (modeled through draws of $z_{t\to\{u,r\}}$ and $z_{r\to\{u,t\}}$) and this will generate the hyper-edge $u \to \{t,r\}$.

We assume that the community vectors are drawn i.i.d. from a general unknown distribution: for $i \in [n]$, $\pi_i \overset{i.i.d.}{\sim} f_\pi(\cdot)$, supported on the $(k-1)$-dimensional simplex Δ^{k-1}

$$\Delta^{k-1} := \{\pi \in \mathbb{R}^k, \pi(i) \in [0,1], \sum_i \pi(i) = 1\}.$$

The performance of our learning algorithms will depend on the distribution of π. In particular, we assume that with probability ρ, a realization of π is a coordinate basis vector, and thus, about ρ fraction of the nodes in the network are *pure*, i.e. they belong mostly to a single community. In this paper, we investigate how the tractability of learning the communities depends on ρ.

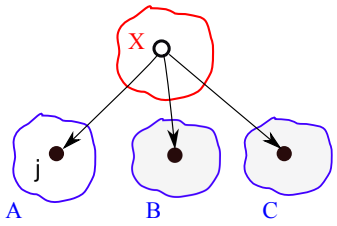

Figure 2: Our moment-based learning algorithm uses 3-star count tensor from set X to sets A, B, C.

3 Proposed Method

Notation: For a matrix M, where $M = UDV^\top$ is the SVD of M, let k-svd$(M) := U\tilde{D}V^\top$ denote the k-rank SVD of M, where $\tilde{D}$ is limited to top-k singular values of M. A matrix $A \in \mathbb{R}^{p\times q}$ is stacked as a vector $a \in \mathbb{R}^{pq}$ by the vec$(\cdot)$ operator,

$$a = \text{vec}(A) \Leftrightarrow a\big((i_1-1)q + i_2)\big) = A(i_1, i_2).$$

The reverse matricization operation is denoted by mat$(\cdot)$, i.e. above $A = \text{mat}(a)$. Let $A * B$ denote the Hadamard or entry-wise product. Let k-svd(M) of a matrix M denote its restriction to top-k singular values, i.e. if $M = U\Lambda V^\top$, k-svd$(M) = U_k\Lambda_k V_k^\top$, which denote the restriction of the subspaces and the singular values to the top-k ones.

In this paper, we consider the problem of learning the community vectors π_i, for $i \in [n]$, given a realization of the (matricized) hyper-adjacency matrix $G \in \mathbb{R}^{|R|\times|U|\cdot|T|}$. We will employ a clustering-based approach on the hyper-adjacency matrix, but employ a different clustering criterion than the usual distance based clustering. our method is shown in Algorithm 1.

Our method relies on finding *pure* resource nodes and using them to find communities for the resource, tag and users. A pure resource node is a node that is mainly corresponding to one hidden community. Therefore, finding that node paves the way for finding resource communities. In addition, since this is a resource-centric model, looking at the subset of hyper graph with pure resources, all tags and all users, suffices to find the communities for users and tags as well. Since we assume knowledge of community connectivity matrices, we can learn community memberships for mixed resource nodes as well. We now provide the details of our proposed method.

Projection matrix: We partition the resource set R into two parts X and Y to avoid dependency issues between the projection matrix and the projected vectors, and this is standard for analysis of spectral clustering. Now let k-svd$(\widehat{G}(\{U,T\},Y)) = M_k\Lambda_k V_k^\top$ and we employ $\widehat{\text{Proj}} := M_k M_k^\top$ as the projection matrix. We project the vectors $\widehat{G}(\{U,T\},x)$ for $x \in X$ using this projection matrix.

Rank test on projected vectors: In the usual spectral clustering method, once we have projected vectors $\widehat{\text{Proj}}\,\widehat{G}(\{U,T\},x) \in \mathbb{R}^{|U|\cdot|T|}$, any distance based clustering can be employed to classify the vectors into different (pure) communities. However, when mixed membership nodes are present, this method fails. We propose an alternative method which considers a rank test on the (matricized form of) the projected vectors. Specifically consider the matricized form $\text{mat}(\widehat{\text{Proj}}\,\widehat{G}(\{U,T\},x) \in \mathbb{R}^{|U|\times|T|}$ and check whether

$$\sigma_1(\text{mat}(\widehat{\text{Proj}}\,\widehat{G}(\{U,T\},x))) > \tau_1 \quad \text{and} \quad \sigma_2(\text{mat}(\widehat{\text{Proj}}\,\widehat{G}(\{U,T\},x))) < \tau_2$$

and if so, declare the node $x \in X$ as a *pure* node. Interchange roles of X and Y and similarly find pure nodes in Y.

Learning using estimated pure nodes: Once the pure nodes in resource set R are found, we can employ the tensor decomposition method, proposed in (Anandkumar et al., 2014a), for learning the mixed membership communities of all the nodes. The pure nodes are employed to obtain averaged 3-star subgraph counts. Partition $\{U, T\}$ into three sets A, B, C as shown in Figure 2. The 3-star subgraph count is defined as

$$\widehat{\mathcal{T}}_{\tilde{R}\to A,B,C} := \frac{1}{|\tilde{R}|}\sum_{r\in\tilde{R}} \widehat{G}(r,A)^\top \otimes \widehat{G}(r,B)^\top \otimes \widehat{G}(r,C)^\top, \tag{3}$$

where $\tilde{R}$ denotes the set of pure resource nodes. The method is explained in Appendix B.

Reconstruction after power method: Since we do not have access to the exact moments we need to do additional processing: the estimated community membership vectors are then subject to thresholding so that the weak values are set to zero. This modification makes our reconstruction strong as we are considering sparse community memberships. Also note that assuming knowledge of community connectivity matrices, we can learn community memberships for mixed resource nodes as well. This is shown in Algorithm 3 in the Appendix.

Algorithm 1 $\{\widehat{\Pi}\} \leftarrow$ LearnMixedMembership$(\widehat{G}, k, \tau_1, \tau_2)$

Input: Hyper-adjacency matrix $\widehat{G} \in \mathbb{R}^{|U|\cdot|T|\times|R|}$, k is the number of communities, and τ_1, τ_2 are thresholds for rank test.

Output: Estimates of the community membership vectors Π.

1: Partition the resource set R randomly into two parts X, Y.
2: $\tilde{R} =$Pure Resource Nodes Detection(X, Y, U, T).
3: $\widehat{\Pi} \leftarrow$ TensorDecomp$(\widehat{G}(\{U, T\}, \cdot), \tilde{R})$
4: Return $\widehat{\Pi}$.

Procedure 2 Pure Resource Nodes Detection

Input: X, Y, U, T.

1: Construct Projection matrix $\widehat{\mathrm{Proj}} = M_k M_k^\top$, where k-svd$(\widehat{G}(\{U, T\}, Y)) = M_k \Lambda_k V_k^\top$.
2: Set of pure nodes $\tilde{R} \leftarrow \emptyset$.
3: **for** $x \in X$ **do**
4: **if** $\sigma_1(\mathrm{mat}(\widehat{\mathrm{Proj}}\, \widehat{G}(\{U, T\}, x))) \geq \tau_1$ and $\sigma_2(\mathrm{mat}(\widehat{\mathrm{Proj}}\, \widehat{G}(\{U, T\}, x))) < \tau_2$ **then**
5: $\tilde{R} \leftarrow \tilde{R} \cup \{x\}$. {Note $\mathrm{mat}(\widehat{\mathrm{Proj}}\, \widehat{G}(\{U, T\}, x)) \in \mathbb{R}^{|U|\times|T|}$ is matricization}
6: **end if**
7: **end for**
8: Interchange roles of X and Y and find pure nodes in Y.
9: Return $\tilde{R}$.

4 Analysis of the Learning Algorithm

Notation: Let $\tilde{O}(\cdot)$ denote $O(\cdot)$ up to poly-log factors. We use the term high probability to mean with probability $1 - n^{-c}$ for any constant $c > 0$.

4.1 Assumptions

For simplicity, we assume that the community memberships of resources, tags and users are drawn from the same distribution. Further, we consider equal expected community sizes, i.e. $\mathbb{E}[\pi] = 1/k \cdot 1^\top$. Additionally, we assume that the community connectivity matrices P, $\tilde{P}$ are homogeneous[1] and equal

$$P = \tilde{P} = (p - q)I + q11^\top, \tag{4}$$

$p, q \in \mathbb{R}$. These simplifications are merely for convenience, and can be easily removed.

Requirement for success of rank test: We require that[2]

$$n = \tilde{\Omega}\left(\sigma_k(\mathbb{E}[\pi\pi^\top])^{-3} \cdot \kappa(\mathbb{E}[\pi\pi^\top])^{-2} \cdot \left(\frac{(p-q)/k + q}{(p-q)/\sqrt{k} + q}\right)^2\right), \tag{5}$$

where $\kappa(\cdot)$ denotes the condition number and $\sigma_k(\cdot)$ denotes the k^{th} largest singular value.

We assume that $\max_{i\in[k]} \pi_x(i) = 1 - \epsilon, \epsilon = O(1)$ and hence there exists no node such that their π is between 1 and $\pi_{\max}$.

Requirement for success of tensor decomposition: Recall that the tensor method uses only pure resource nodes. Let ρ be the fraction of such pure resource nodes. Let $w_i := \mathbb{P}[\pi_r(i) = 1 | r \in \tilde{R}]$. For simplicity, we assume that $w_i \equiv 1/k$. Again, this can be easily extended.

We require the separation in edge connectivity $p - q$ to satisfy

$$\frac{(p-q)^2}{p} = \tilde{\Omega}\left(\frac{\sqrt{k}}{\sqrt{n\rho} \cdot \sigma_k(\mathbb{E}[\pi\pi^\top])}\right). \tag{6}$$

Intuitively this implies that there should be enough separation between connectivity within a community and connectivity across communities.

Dependence on p, q: Note that for the rank test, (5), in the well-conditioned setting we have $\sigma_k(\mathbb{E}[\pi\pi^\top]) = O(1/k)$. Then if $p \simeq q$, we need $n = \tilde{\Omega}\left(k^3\right)$. For the case where $q < p/k$, we will require $n = \tilde{\Omega}\left(k^2\right)$. This is intuitive as the role of q is to make the components non-orthogonal, i.e., q acts as noise. Therefore, smaller q results in better guarantees. 55 For the tensor decomposition method to be successful,i.e., Equation (6), in the well-conditioned setting, if we have $n = \tilde{\Omega}\left(k^3\right)$, this means p, q are constants. Alternatively, for sparse graphs, we want p, q to decay. According to the constraints, we need a larger n. This is intuitive as in case of sparse graphs our observations convey less information about unknown community memberships. Therefore, we need more samples.

Note that Anandkumar et al. (2014a) require $n = O(k^2)$ while we need $n = O(k^3)$. The reason is that we are estimating a hypergraph (they estimate a graph) and we are estimating more parameters in this model. Therefore, we need more samples.

[1] Our results can be easily extended to the case when P and $\tilde{P}$ are full rank.

[2] $\tilde{\Omega}, \tilde{O}$ represent Ω, O up to poly-log factors.

4.2 Guarantees

We now establish main results on recovery at the end of our algorithm. We first show that under the assumptions in the previous section, we obtain an ℓ_2 guarantee for recovery of the membership weights of source nodes in each community. We should note that this result can be extended to recovery of membership for tag and user nodes as well. In this case, there will be additional perturbation terms.

Let $\tilde{\Pi}$ be the reconstruction of communities (of resources, users and tags) using the tensor method in Algorithm 3 in the Appendix, but before thresholding. For a matrix M, let $(M)^i$ denote the i^{th} row. Recall that $(\Pi)^i$ denotes the memberships of all the nodes in the i^{th} community, since $\Pi \in \mathbb{R}^{(|R|+|U|+|T|)\times k}$. We have the following result:

Theorem 1 (Reconstruction of communities (before thresholding)) *We have w.h.p.*

$$\epsilon_\pi := \max_{i\in[k]} \|(\tilde{\Pi})^i - (\Pi)^i\| = \tilde{O}\left(\frac{\sqrt{k}\cdot p\cdot \kappa(\mathbb{E}[\pi\pi^\top])}{\sqrt{\rho}(p-q)^2}\right). \tag{7}$$

Remark: Note that the ℓ_2 norm above is taken over all the nodes of the network and we expect this to be $O(\sqrt{n})$ if error at each node is $O(1)$. Assuming $\mathbb{E}[\pi\pi^\top]$ is well conditioned and when $\rho, p, q = \Omega(1)$, we get a better guarantee that $\epsilon_\pi = O(\sqrt{k})$.

Now we further show that when the distribution of π is "mostly" sparse, i.e. each node's membership vector does not have too many large entries, we can improve the above ℓ_2 guarantees into ℓ_1 guarantees via thresholding.

Specifically, assuming that the distribution of π satisfies

$$\mathbb{P}[\pi(i) \geq \tau] \leq \frac{C}{k}\log(1/\tau), \quad \forall i \in [k]$$

for $\tau = O(\epsilon_\pi \cdot \frac{k}{n})$, we have the following result. This is equivalent to the case that the tail τ is exponentially small in k, i.e., sparsity.

Remark: Dirichlet distribution satisfies this assumption when $\sum_i \alpha_i < 1$, where α_i represent the Dirichlet concentration parameters.

Theorem 2 (ℓ_1 guarantee for reconstruction after thresholding) *We have*

$$\|\hat{\Pi}^i - \Pi^i\|_1 = \tilde{O}\left(\epsilon_\pi \cdot \sqrt{\frac{n}{k}}\right) = \tilde{O}\left(\frac{\sqrt{n}\cdot p\cdot \kappa(\mathbb{E}[\pi\pi^\top])}{\sqrt{\rho}(p-q)^2}\right), \tag{8}$$

where $\hat{\Pi}^i$ is the result of thresholding with $\tau = O(\epsilon_\pi \cdot \frac{k}{n})$.

Remark: Note that the ℓ_1 norm above is taken over all the nodes of the network and we expect this to be $O(n)$ if error at each node is $O(1)$. Assuming $\mathbb{E}[\pi\pi^\top]$ is well-conditioned and when $\rho, p, q = \Omega(1)$, we get a better guarantee of $O(\sqrt{n})$. Hence, we obtain good error guarantees in both cases on ℓ_1 and ℓ_2 norms.

For proof of the Theorems, see Appendix C.

5 Overview of Proof

5.1 Analysis of Graph Moments under MMSF

5.1.1 Overview of Kronecker and Khatri-Rao products:

We require the notions of Kronecker $A \otimes B$ and Khatri-Rao products $A \odot B$ between two matrices A and B. First we define the *Kronecker product* $A \otimes B$ between matrices $A \in \mathbb{R}^{n_1 \times k_1}$ and $B \in \mathbb{R}^{n_2 \times k_2}$. Its $(\mathbf{i}, \mathbf{j})^{\text{th}}$ entry is given by

$$(A \otimes B)_{\mathbf{i},\mathbf{j}} := A_{i_1,j_1} B_{i_2,j_2}, \quad \mathbf{i} = \{i_1, i_2\} \in [n_1] \times [n_2], \mathbf{j} = \{j_1, j_2\} \in [k_1] \times [k_2].$$

Thus, for two vectors a and b, we have

$$(a \otimes b)_{\mathbf{i}} := a_{i_1} b_{i_2}, \quad \mathbf{i} = \{i_1, i_2\} \in [n_1] \times [n_2].$$

For the Khatri-Rao product $A \odot B$ between matrices $A \in \mathbb{R}^{n_1 \times k}$ and $B \in \mathbb{R}^{n_2 \times k}$, we have its $(\mathbf{i}, j)^{\text{th}}$ as

$$A \odot B(\mathbf{i}, j) := A_{i_1,j} B_{i_2,j}, \quad \mathbf{i} = \{i_1, i_2\} \in [n_1] \times [n_2], j \in [k].$$

In other words, we have

$$A \odot B := [a_1 \otimes b_1 \quad a_2 \otimes b_2 \quad \dots a_k \otimes b_k],$$

where a_i, b_i are the i^{th} columns of A and B. Note the difference between the Kronecker and the Khatri-Rao products. While the Kronecker product expands both the number of rows and columns, the Khatri-Rao product preserves the original number of columns. We will also use another simple fact that

$$(A \otimes B)(C \otimes D) = AC \otimes BD. \tag{9}$$

5.1.2 Result on Correctness of the Algorithm

Recall that $P \in [0,1]^{k \times k}$ denotes the connectivity matrix between communities of users and resources and $\tilde{P} \in [0,1]^{k \times k}$ denotes the corresponding connectivity matrix between communities of resources and tags. Define

$$F := \Pi_U^\top P, \quad \tilde{F} := \Pi_T^\top \tilde{P}. \tag{10}$$

Let $F_u = \pi_u^\top P$ be the row vector corresponding to user u and similarly $\tilde{F}_t$ corresponds to tag t. Similarly, let $F_A = \Pi_A^\top P$ be the sub-matrix of F.

We now provide a simple result on the average hyper-edge connectivity and the form of the 3-star counts, given the community memberships.

Proposition 1 (Form of Graph Moments) *Under the MMSF model proposed in Section 2, we have that the generated hyper-graph $\widehat{G} \in \{0,1\}^{|U| \cdot |T| \times |R|}$ satisfies*

$$G := \mathbb{E}[\widehat{G} | \Pi] = (F \odot \tilde{F}) \Pi_R, \tag{11}$$

where $\odot$ denotes the Khatri-Rao product. Moreover, for a given resource $r \in R$, the column vector $\widehat{G}(\{U, T\}, r)$ has conditionally independent entries given the community membership vector π_r. If $\tilde{R} \subset R$ is the set of (exactly) pure nodes, then the 3-star count defined in (3) satisfies

$$\mathcal{T}_{\tilde{R} \to A,B,C} := \mathbb{E}[\widehat{\mathcal{T}}_{\tilde{R} \to A,B,C} | \Pi] = \sum_{i \in [k]} w_i (H_A \otimes H_B \otimes H_C), \tag{12}$$

where w_i is

$$w_i := \mathbb{P}[\pi_r(i) = 1 | r \in \tilde{R}],$$

and $H_A := F_{U(A)} \odot \tilde{F}_{T(A)}$, and similarly, H_B and H_C.

The above results follow from modeling assumptions in Section 2, and in particular, the conditional independence relationships among the different variables. For details, see Appendix A.

In (11), note that if a column of $G(\{U,T\},X)$ corresponds to a pure node $x \in X$, then the matrix has rank of one, since π_x corresponds to a coordinate basis vector. On the other hand, for the case where columns correspond to mixed nodes, the matrix has rank bigger than one. Thus, the rank criterion succeeds in identifying the pure nodes in X under exact moments.

Lemma 3 (Correctness of the method under exact moments) *Assume $F \odot \tilde{F}$ has full column rank, and Π_Y has full row rank, where $Y \subset R$ is used for constructing the projection matrix, then the proposed method LearnMixedMembership in Algorithm 1 correctly learns the community membership matrix Π.*

Proof: Using the form of the moments in Proposition 1, we have that if $r \in \tilde{R}$ is a pure node, then $G(\{U,T\},r) = (F \odot \tilde{F})\pi_r$ is rank one since it selects only one column of $F \odot \tilde{F}$. Thus, the rank test in Algorithm 1 succeeds in recovering the pure nodes. The correctness of tensor method follows from (Anandkumar et al., 2014a). □

Since we only have sampled graph $\widehat{G}$ and not the exact moments, we need to carry out perturbation analysis, which is outlined below.

5.2 Perturbation Analysis

Recall that $\widehat{\text{Proj}} = M_k M_k^\top$ is the projection matrix corresponding to $\text{k-svd}(\widehat{G}(\{U,T\},Y)) = M_k \Lambda_k V_k^\top$. In the similar manner Proj is the projection matrix corresponding to $\text{k-svd}(G(\{U,T\},Y))$. Define the perturbation between empirical and exact moments upon projection as

$$m_x := \| \widehat{\text{Proj}}\, \widehat{G}(\{U,T\},x) - G(\{U,T\},x)\|, \quad \forall\, x \in X, \quad \epsilon_{\text{Rank}} := \max_x \|m_x\|. \tag{13}$$

The above perturbation can be divided into two parts

$$\|m_x\| \leq \| \widehat{\text{Proj}}(\widehat{G}(\{U,T\},x) - G(\{U,T\},x))\| + \|(\widehat{\text{Proj}} - \text{Proj})G(\{U,T\},x)\|.$$

The first term is commonly referred to as *distance perturbation* and the second term is the *subspace perturbation*. We establish these perturbation bounds below.

We begin our perturbation analysis by bounding m_x as defined in Eqn. (13).

Lemma 4 (Distance perturbation) *Under the assumptions of Section 4.1, with probability $1-\delta$, we have for all $x \in X$,*

$$\| \widehat{\text{Proj}}(\widehat{G}(\{U,T\},x) - G(\{U,T\},x))\| \leq \sqrt{k}p \left(1 + \frac{C'}{\sqrt{k}} (\log(n/\delta))^4\right)^{1/2},$$

for some constant $C' > 0$.

See Appendix C.1 and Appendix C.2 for details. Notice that the subspace perturbation dominates.

Lemma 5 (Subspace perturbation) *We have the subspace perturbation as*

$$\|(\widehat{\mathrm{Proj}} - \mathrm{Proj})G(\{U,T\},x)\| \leq 2\sigma_k^{-1}(\Pi_Y)\sqrt{\|F \odot \tilde{F}\|_1}.$$

Under the assumptions of Section 4.1, w.h.p. this reduces as

$$\|(\widehat{\mathrm{Proj}} - \mathrm{Proj})G(\{U,T\},x)\| \leq O\left(\frac{\sqrt{n}}{\sqrt{\sigma_k(\mathbb{E}[\pi\pi^\top])}} \cdot \left(\frac{p-q}{k} + q\right)\right).$$

See Appendix C.2.

5.3 Analysis of Rank Test

Recall that from the perturbation analysis, we have bound ϵ_{Rank} on the error vector m_x, defined in (13). We assume there exist no node such that $\max_{i\in[k]} \pi_x(i)$ is between the threshold given in (14) and 1. We have the following result on the rank test.

Lemma 6 (Conditions for Success of Rank Test) *When the thresholds in Algorithm 1 are chosen*

$$0 < \tau_1 < \min_i \|(F_U)_i\| \cdot \|(\tilde{F}_T)_i\| - \epsilon_{\mathrm{Rank}}, \quad \tau_2 > \epsilon_{\mathrm{Rank}},$$

then all the pure nodes pass the rank test. Moreover, any node $x \in X$ passing the rank test satisfies

$$\max_{i\in[k]} \pi_x(i) \geq \frac{\tau_1 - \tau_2 - 2\,\epsilon_{\mathrm{Rank}}}{\max_i \|(F_U)_i\| \cdot \|(\tilde{F}_T)_i\|}. \tag{14}$$

Proof: See Appendix C.3. □

The above result states that we can correctly detect pure nodes using the rank test. The conditions stem from the fact that we require the top eigen-value to pass the test and the second top eigen-value to not pass the test. For a pure node, $\sigma_1(\mathrm{mat}(\widehat{\mathrm{Proj}}\, G(\{U_1,T_1\},x)))$ is $\min_i \|(F_{U_1})_i\| \cdot \|(\tilde{F}_{T_1})_i\|$. To account for empirical error, we consider ϵ_{Rank}. In addition, the second-top eigen-value can be as small as 0. We also note the error in empirical estimation. This result allows us to control the perturbation in the 3-star tensor constructed using the nodes which passed the rank test.

6 Conclusion

In this paper, we propose a novel probabilistic approach for modeling folksonomies, and propose a guaranteed approach for detecting overlapping communities in them. We present a more scalable approach where realistic conditional independence constraints are imposed. These constraints are natural for social tagging systems, and they lead to scalable modeling and tractable learning. While the original MMSB model assumes that the communities are drawn from a Dirichlet distribution, here, we do not require such a strong parametric assumption. Note that the Dirichlet assumption for community memberships can be limiting and cannot model general correlations in memberships. Here, we impose a weak assumption that a certain fraction of resource nodes are "pure" and belong to a single community. This is reasonable to expect in practice. We establish that the communities are identifiable under these natural assumptions, and can be learnt efficiently using spectral approaches. Considering future directions, we note that social tagging assumes a specific structure. Therefore, it is of interest to extend this model to more general hypergraphs.

Acknowledgment

A. Anandkumar is supported in part by Microsoft Faculty Fellowship, NSF Career award CCF-1254106, ONR Award N00014-14-1-0665, ARO YIP Award W911NF-13-1-0084 and AFOSR YIP FA9550-15-1-0221. H. Sedghi is supported by NSF Career award FG15890.

The authors thank Majid Janzamin for detailed discussion on rank test analysis. The authors thank Rong Ge and Yash Deshpande for extensive initial discussions during the visit of AA to Microsoft Research New England in Summer 2013 regarding the pairwise mixed membership models without the Dirichlet assumption. The authors also acknowledge detailed discussions with Kamalika Chaudhuri regarding analysis of spectral clustering.

Appendix

A Moments under MMSF model and Algorithm Correctness

Proof of Proposition 1: We have

$$\begin{aligned}\mathbb{E}[\widehat{G}(\{u,t\},r)|\pi_r,\pi_t,\pi_u] &\overset{(a)}{=} \mathbb{E}[\mathbb{E}[\widehat{G}(\{u,t\},r)|z_{r\to\{u,t\}},\pi_t,\pi_r,\pi_u]] \\ &\overset{(b)}{=} \mathbb{E}[\mathbb{E}[\widehat{B}_{r\to u;t}\cdot\widehat{B}_{r\to t;u}|z_{r\to\{u,t\}},\pi_t,\pi_u]|\pi_r] \\ &\overset{(c)}{=} \mathbb{E}[F_u z_{r\to\{u,t\}}\cdot\tilde{F}_t z_{r\to\{u,t\}}|\pi_r], \end{aligned} \tag{15}$$

where (a) and (b) are from the assumption (2) that

$$\widehat{G}(\{u,t\},r) = \widehat{B}_{r\to u;t}\cdot\widehat{B}_{r\to t;u},$$

where $\widehat{B}_{r\to u;t}$ and $\widehat{B}_{r\to t;u}$ are Bernoulli draws, which only depend on the contextual variables $z_{r\to\{u,t\}}$, $z_{u\to\{r,t\}}$ and $z_{t\to\{u,r\}}$, and therefore $\widehat{G}(\{u,t\},r) - z_{r\to\{u,t\}} - \pi_r$ form a Markov chain. This also establishes that $\widehat{G}(\{u,t\},r)$ and $\widehat{G}(\{u',t'\},r)$ are conditionally independent given the community membership vector π_r, for $u \neq u'$ and $t \neq t'$.

For (c), we have that

$$\begin{aligned}\mathbb{E}[\widehat{B}_{r\to u;t}|z_{r\to\{u,t\}},\pi_u] &= \mathbb{E}[\mathbb{E}[\widehat{B}_{r\to u;t}|z_{r\to\{u,t\}},z_{u\to\{r,t\}}]|\pi_u] \\ &= \mathbb{E}[z_{u\to\{t,r\}}^\top P z_{r\to\{u,t\}}|z_{r\to\{u,t\}},\pi_u] \\ &= \pi_u^\top P z_{r\to\{u,t\}} \\ &= F_u z_{r\to\{u,t\}}\end{aligned}$$

from (1) and the fact that

$$\mathbb{E}[z_{u\to\{t,r\}}|\pi_u] = \pi_u.$$

Thus, we have

$$\begin{aligned}\mathbb{E}[\widehat{G}(\{U,T\},r)|\pi_r,\Pi_T,\Pi_U] &\overset{(a)}{=} \mathbb{E}[Fz_{r\to\{u,t\}}\otimes\tilde{F}z_{r\to\{u,t\}}|\pi_r]\\ &\overset{(b)}{=} \mathbb{E}[(F\otimes\tilde{F})(z_{r\to\{u,t\}}\otimes z_{r\to\{u,t\}})|\pi_r]\\ &\overset{(c)}{=} \sum_{i\in[k]}\pi_r(i)(F\otimes\tilde{F})(e_i\otimes e_i)\\ &\overset{(d)}{=} (F\odot\tilde{F})\pi_r,\end{aligned}$$

where (a) follows from (15) and (b) follows from the fact (9). (c) follows from the fact that $z_{r\to\{u,t\}}$ takes value e_i with probability $\pi_r(i)$, where $e_i\in\mathbb{R}^k$ is the basis vector in the i^{th} coordinate. (d) follows from the definition of Khatri-Rao product.

The form of the 3-star moment is from the lines of (Anandkumar et al., 2014a, Prop 2.1), and relies on the assumption that $\tilde{R}$ consists of pure nodes.

□

B Learning using Tensor Decomposition

We now recap the tensor decomposition approach proposed in (Anandkumar et al., 2014a) here. This is shown in Algorithm 3 with modifications specific to our framework.

We partition U,T into three sets for the different tasks explained in the Algorithm 3. Also note that with knowledge of community connectivity matrices, we can learn community memberships for mixed resource nodes as well.

Procedure 3 $(\widehat{\Pi})\leftarrow$ TensorDecomp$(\widehat{G},\tilde{R})$

Let $P\in\mathbb{R}^{k\times k}$ be the community connectivity matrix from user communities to resource communities and similarly $\tilde{P}$ is connectivity from tag communities to resource communities. $\tilde{R}$ are estimated pure resource nodes. Partition $\{U,T\}$ into $\{U_i,T_i\}$ for $i=1,2,3$.

Compute whitened and symmetrized tensor $\mathcal{T}\leftarrow\hat{G}_{\tilde{R}\to\{A,B,C\}}(\widehat{W}_A,\widehat{W}_B\widehat{S}_{AB},\widehat{W}_C\widehat{S}_{AC})$, where A,B,C form a partition of $\{U_2,T_2\}$. Use $\{U_3,T_3\}$ for computing the whitening matrices.

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

$\widehat{\Pi}_R\leftarrow$ Thres$(\text{Diag}(\widehat{\lambda})^{-1}\widehat{\Phi}^\top\widehat{W}_A^\top\widehat{G}_{R,A}^\top,\ \tau)$.

return $(\widehat{\Pi})$.

C Perturbation Analysis: Proof of Theorems 1, 2

Notation: For a vector v, let $\|v\|$ denote its 2-norm. Let $\text{Diag}(v)$ denote a diagonal matrix with diagonal entries given by a vector v. For a matrix M, let $(M)_i$ and $(M)^i$ denote its i^{th} column and row respectively. Let $\|M\|_1$ denote column absolute sum and $\|M\|_\infty$ denote row absolute sum of M. Let $M^\dagger$ denote the MoorePenrose pseudo-inverse of M.

Procedure 4 $\{\lambda, \Phi\} \leftarrow$TensorEigen$(T,\ \{v_i\}_{i\in[L]}, N)$ (Anandkumar et al., 2014a)

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 $\tau = 1$ to L **do**
    $\theta_0 \leftarrow v_\tau$.
    for $t = 1$ to N **do**
      $\tilde{T} \leftarrow T$.
      for $j = 1$ to $i-1$ (when $i > 1$) **do**
        if $|\lambda_j \langle \theta_t^{(\tau)}, \phi_j\rangle| > \xi$ **then**
          $\tilde{T} \leftarrow \tilde{T} - \lambda_j \phi_j^{\otimes 3}$.
        end if
      end for
      Compute power iteration update $\theta_t^{(\tau)} := \frac{\tilde{T}(I, \theta_{t-1}^{(\tau)}, \theta_{t-1}^{(\tau)})}{\|\tilde{T}(I, \theta_{t-1}^{(\tau)}, \theta_{t-1}^{(\tau)})\|}$
    end for
  end for
  Let $\tau^* := \arg\max_{\tau\in L}\{\tilde{T}(\theta_N^{(\tau)}, \theta_N^{(\tau)}, \theta_N^{(\tau)})\}$.
  Do N power iteration updates starting from $\theta_N^{(\tau^*)}$ to obtain eigenvector estimate ϕ_i, and set $\lambda_i := \tilde{T}(\phi_i, \phi_i, \phi_i)$.
end for
return the estimated eigenvalue/eigenvectors (λ, Φ).

C.1 Distance Concentration: Proof of Lemma 4

The proof is along the lines of (McSherry, 2001, Theorem 13) but we apply Hanson-Wright bound in Proposition 5 to get a better perturbation guarantee without the need for constructing the so-called combinatorial projection, as in (McSherry, 2001).

We have $h_x := \widehat{G}(x; \{U, T\}) - G(x; \{U, T\})$ and let $\sigma^2 = \max_i \mathbb{E}[h_x(i)^2 | \pi_x]$. Note the simple fact

$$\|\widehat{\mathrm{Proj}}\, h_x\|^2 = h_x^\top \widehat{\mathrm{Proj}}^2\, h_x = h_x^\top \widehat{\mathrm{Proj}}\, h_x,$$

since $\widehat{\mathrm{Proj}}$ is a projection matrix. From Proposition 1, we have that the entries of h_x are conditionally independent given π_x. Thus, the Hanson-Wright inequality in Proposition 5 is applicable, and we have with probability $1-\delta$, for all $x \in X$,

$$h_x^\top \widehat{\mathrm{Proj}}\, h_x \le \mathbb{E}[h_x^\top \widehat{\mathrm{Proj}}\, h_x | \pi_x] + C'\sigma^2 \|\widehat{\mathrm{Proj}}\|_{\mathbb{F}} (\log(n/\delta))^4 \tag{16}$$

Now $\|\widehat{\mathrm{Proj}}\|_{\mathbb{F}} \le \sqrt{k}\|\widehat{\mathrm{Proj}}\| = \sqrt{k}$. The expectation is

$$\mathbb{E}[h_x^\top \widehat{\mathrm{Proj}}\, h_x | \pi_x] \le \mathrm{tr}(\widehat{\mathrm{Proj}})\sigma^2 = k\sigma^2,$$

using the property that $\widehat{\mathrm{Proj}}$ is idempotent. Thus, we have from (16), with probability $1-\delta$, for all $x \in X$,

$$h_x^\top \widehat{\mathrm{Proj}}\, h_x \le k\sigma^2 + C'\sqrt{k}\sigma^2 (\log(n/\delta))^4,$$

and we see that the mean term dominates and the bound is $\tilde{O}(k\sigma^2)$.

Draw random variables

$$\widehat{B}_{r\to u;t} \sim \text{ Bernoulli}(z_{u\to\{t,r\}}^\top P z_{r\to\{u,t\}})$$
$$\widehat{B}_{r\to t;u} \sim \text{ Bernoulli}(z_{t\to\{u,r\}}^\top \tilde{P} z_{r\to\{u,t\}}).$$

The presence of hyper-edge $G(\{u,t\},r)$ is given by the product

$$\widehat{G}(\{u,t\},r) = \widehat{B}_{r\to u;t} \cdot \widehat{B}_{r\to t;u}.$$

The variance is on lines of proof of Lemma 10 and we repeat it here.

$$\begin{aligned}
\max_i \mathbb{E}[h_x(i)^2|\pi_x] &= \max_{u\in U, v\in V} \mathbb{E}[\hat{B}_{x\to u;t}\hat{B}_{x\to t;u} - ((F\odot\tilde{F})\pi_x)_{ut}]^2 \\
&\le \max_{u\in U, v\in V} ((F\odot\tilde{F})\pi_x)_{ut}, \\
&\le \max_{u\in U, v\in V} \sum_{j\in[k]} F(u,j)\tilde{F}(t,j)\pi_x(j) \\
&\le \max_{i\in[k]} \sum_{j\in[k]} P(i,j)\tilde{P}(i,j)\pi_x(j) \\
&\le P_{\max}^2
\end{aligned}$$

C.2 Proof of Lemma 5

From Davis-Kahan in Proposition 6, we have

$$\|(\widehat{\text{Proj}} - I)G(\{U,T\},Y)\| \le 2\|\widehat{G}(\{U,T\},Y) - G(\{U,T\},Y)\|.$$

and thus

$$\|(\widehat{\text{Proj}} - I)G(\{U,T\},x)\| \le 2\|\widehat{G}(\{U,T\},Y) - G(\{U,T\},Y)\| \cdot \|G(\{U,T\},Y)^\dagger \cdot G(\{U,T\},x)\|$$

Now,

$$G(\{U,T\},Y)^\dagger = \Big((F\odot\tilde{F})\Pi_Y\Big)^\dagger = \Pi_Y^\dagger (F\odot\tilde{F})^\dagger,$$

since the assumption is that $F\odot\tilde{F}$ has full column rank and Π_Y has full row rank. Thus, we have

$$G(\{U,T\},Y)^\dagger \cdot G(\{U,T\},x) = \Pi_Y^\dagger (F\odot\tilde{F})^\dagger (F\odot\tilde{F})\pi_x = \Pi_Y^\dagger \cdot \pi_x,$$

since $(F\odot\tilde{F})^\dagger(F\odot\tilde{F}) = I$ due to full column rank, when $|U|$ and $|T|$ are sufficiently large, due to concentration result from Lemma 11. Note that under assumption A3, the variance terms in Lemma 11 are decaying and we have that $F\odot\tilde{F}$ has full column rank w.h.p. From Lemma 10, we have the result.

C.3 Analysis of Rank Test: Lemma 6

Consider the test under expected moments $G := \mathbb{E}[\widehat{G}|\Pi]$. For every node $x \in X$ (R is randomly partitioned into X, Y), which passes the rank test in Algorithm 1, by definition,

$$\|\operatorname{mat}(\hat{G}(\{U,T\},x))\| > \tau_1, \quad \text{and} \quad \sigma_2(\operatorname{mat}(\hat{G}(\{U,T\},x))) < \tau_2.$$

We use the following approximation.

$$\|F_i\| \simeq \sqrt{(p-q)^2\|\Pi^i\|^2 + nq^2 + 2(p-q)q\|\Pi^i\|_1}$$

Recall the form of G from Proposition 1

$$\mathrm{mat}(G(\{U,T\},x)) = F_U \,\mathrm{Diag}(\pi_x)\tilde{F}_T^\top.$$

First we consider the case, $p \simeq q$. Following lines of Anandkumar et al. (2014b), we have that

$$|\sigma_1 - \pi_{\max} n(\frac{p-q}{k}+q)^2| \leq \|E\| + \epsilon_{\mathrm{Rank}}$$

where

$$\|E\| \leq \sqrt{k}\pi_{2,\max} n(\frac{p-q}{k}+q)^2 \frac{\|\mathbb{E}[\pi\pi^\top]\|q^2}{\left(\frac{p-q}{k}+q\right)^3}\left\{p\sqrt{\mathbb{E}[\pi\pi^\top]} + \frac{\|\mathbb{E}[\pi\pi^\top]\|q^2}{\left(\frac{p-q}{k}+q\right)}\right\}.$$

Hence, we have that

$$\sigma_2 \geq \pi_{2,\max} n(\frac{p-q}{k}+q)^2 - \|E\| - \epsilon_{\mathrm{Rank}} - (1/\tilde{\mu})\,\epsilon_{\mathrm{Rank}} - \pi_{3,\max} np^2 \|\mathbb{E}[\pi\pi^\top]\|,$$

where we assume $\pi_{\max} \geq (1+\mu)\pi_{2,\max}$ and $\tilde{\mu} := \frac{1+\mu-\mu_R-\mu_E}{1+\mu}$, $\mu_R := \frac{\|F\|}{\|F_i\|}$, $\mu_E := \frac{\|E\|}{\pi_{2,\max} n(\frac{p-q}{k}+q)^2}$.

We note that ϵ_{Rank} dominates $\|E\|$ and the last term. Therefore,

$$\tau_2 - \epsilon_{\mathrm{Rank}} \geq \sigma_2(\mathrm{mat}(G(\{U,T\},x))) \geq \pi_{2,\max} n(\frac{p-q}{k}+q)^2 - (1+1/\tilde{\mu})\,\epsilon_{\mathrm{Rank}},$$

and

$$\begin{aligned}
\tau_1 + \epsilon_{\mathrm{Rank}} &\leq \|F_U \,\mathrm{Diag}(\pi_x)\tilde{F}_T^\top\| \\
&\leq \pi_{\max} \max_i \|(F_{U_1})_i\| \cdot \|(\tilde{F}_T)_i\| + \pi_{2,\max} n(\frac{p-q}{k}+q)^2 \\
&\leq \pi_{\max} \max_i \|(F_U)_i\| \cdot \|(\tilde{F}_T)_i\| + \tau_2 + 1/\tilde{\mu}\,\epsilon_{\mathrm{Rank}}\,.
\end{aligned}$$

Combining we have that any vector which passes the rank test satisfies

$$\pi_{\max} \geq \frac{\tau_1 - \tau_2 + (1-1/\tilde{\mu})\,\epsilon_{\mathrm{Rank}}}{\max_i \|(F_U)_i\| \cdot \|(\tilde{F}_T)_i\|}.$$

Now, for the case where $q < p/k$, the bound on $\|E\|$ is almost 0, $\mu_R \simeq 1$ and $\mu_E = 0$. Hence Eqn. (C.3) always holds. This is intuitive as the role of q is to make the components non-orthogonal, i.e., q acts as noise. Therefore, smaller q results in better guarantees.

With $|U| = |T| = \Theta(n)$, and using the concentration bounds in Lemma 11, we have that with probability $1-\delta$,

$$\|(F_U)_i\| \cdot \|(\tilde{F}_T)_i\| = O\left(\sqrt{|U|\cdot|T|}\|\mathbb{E}[\pi\pi^\top]\| \cdot (p - q + \sqrt{k}q)\right)$$

assuming homegenous setting.

For ϵ_{Rank}, the subspace perturbation dominates. From Lemma 11, we have

$$\|F \odot \tilde{F}\|_1 = O\left(n^2 \left(\frac{p-q}{k} + q\right)^2\right).$$

Thus, we have the subspace perturbation from Lemma 5 as

$$\epsilon_{\text{Rank}} = O\left(\frac{\sqrt{n}}{\sqrt{\sigma_k(\mathbb{E}[\pi\pi^\top])}} \cdot \left(\frac{p-q}{k} + q\right)\right).$$

Substituting for the condition that $\tau_1 = \Omega(\epsilon_{\text{Rank}})$, we obtain assumption (5). Thus, the rank test succeeds in this setting.

C.4 Perturbation Analysis for the Tensor Method

This is along the lines of analysis in (Anandkumar et al., 2014a). However, notice here due to hypergraph setting, we need to redo the individual perturbations. Recall that $w_i := \mathbb{P}[i = \arg\max_j \pi(j) | \pi \text{ is pure}]$ and $\rho = \mathbb{P}[\pi \text{ is pure}]$. The size of recovered set of pure nodes $\tilde{R} = \Theta(n\rho)$, assuming $n\rho > 1$.

We provide the perturbation of the whitened tensor. Let $\Phi := W_A^\top H_A \operatorname{Diag}(\eta)^{1/2}$ be the eigenvectors of the whitened tensor under exact moments and $\lambda := \operatorname{Diag}(\eta)^{-1/2}$ be the eigenvalues. $S, \widehat{S}$ respectively denote the exact and empirical symmetrization matrix for different cases based on their subscript.

Lemma 7 (Perturbation of whitened tensor) *We have w.h.p.*

$$\begin{aligned}\epsilon_{\mathcal{T}} :=& \left\| \widehat{\mathcal{T}}_{\tilde{R}\to\{A,B,C\}}(\hat{W}_A, \hat{W}_B \widehat{S}_{AB}, \hat{W}_C \widehat{S}_{AC}) - \sum_{i\in[k]} \lambda_i \Phi^{\otimes 3} \right\| \\ =& O\left(\frac{p}{\sqrt{n\rho} w_{\min} \cdot (p-q)^2 \cdot \sigma_k(\mathbb{E}[\pi\pi^\top])}\right) \end{aligned} \tag{17}$$

Proof: Let $\mathcal{T} := \mathbb{E}[\widehat{\mathcal{T}} | \Pi_{A,B,C}]$.

$$\begin{aligned}\epsilon_1 &:= \left\| \widehat{\mathcal{T}}(\hat{W}_A, \hat{W}_B \widehat{S}_{AB}, \hat{W}_C \widehat{S}_{AC}) - \mathcal{T}(\hat{W}_A, \hat{W}_B \widehat{S}_{AB}, \hat{W}_C \widehat{S}_{AC}) \right\| \\ \epsilon_2 &:= \left\| \mathcal{T}(\hat{W}_A, \hat{W}_B \widehat{S}_{AB}, \hat{W}_C \widehat{S}_{AC}) - \mathcal{T}(W_A, W_B S_{AB}, W_C S_{AC}) \right\| \end{aligned}$$

For ϵ_1, the dominant term in the perturbation bound is

$$\begin{aligned} & O\left(\frac{1}{|\tilde{R}|} \|\tilde{W}_B^\top H_B\|^2 \left\| \sum_{i\in Y} \left(\widehat{W}_A^\top (\widehat{G}_{A,i} - H_A \pi_i)\right) \right\| \right) \\ &= O\left(\frac{1}{w_{\min}} \frac{1}{|\tilde{R}|} \left\| \sum_{i\in Y} \left(\widehat{W}_A^\top (\widehat{G}_{A,i} - H_A \pi_i)\right) \right\| \right) \end{aligned}$$

The second term is

$$\epsilon_2 \le \frac{\epsilon_W}{\sqrt{w_{\min}}},$$

since due to whitening property.

Now imposing the requirement that

$$\epsilon_i < \Theta\left(\lambda_{\min} r^2\right),$$

from Theorem 11 (Anandkumar et al., 2014a), $\lambda_{\min} = 1/\sqrt{w_{\max}}$, and we have $r = \Theta(1)$ by initialization using whitened neighborhood vectors (from lemma 25 (Anandkumar et al., 2014a)). ϵ_1 is not the dominant error, on lines of (Anandkumar et al., 2014a). Now for ϵ_2, we require

$$\epsilon_W \leq \sqrt{\frac{w_{\min}}{w_{\max}}} \leq 1,$$

and using Lemma 8, we have

$$\frac{(p-q)^2}{p} \geq \frac{\sqrt{w_{\max}}}{w_{\min}} \cdot \frac{1}{\sqrt{n\rho} \cdot \sigma_k(\mathbb{E}[\pi\pi^\top])}.$$

□

Lemma 8 (Whitening Perturbation) *We have the perturbation of the whitening matrix $\hat{W}_A$ as w.h.p.*

$$\epsilon_W := \|\operatorname{Diag}(\vec{w})^{1/2} H_A^\top (\hat{W}_A - W_A)\| = O\left(\frac{p}{\sqrt{n\rho w_{\min}} \cdot (p-q)^2 \cdot \sigma_k(\mathbb{E}[\pi\pi^\top])}\right).$$

Proof: From (Anandkumar et al., 2014a, Lemma 17), the whitening perturbation under the tensor method is given by

$$\epsilon_W := \|\operatorname{Diag}(w)^{1/2} H_A^\top (\hat{W}_A - W_A)\| = O\left(\frac{\epsilon_G}{\sigma_{\min}(G_{\tilde{R},A})}\right).$$

Using the bounds from Section C.5, we have

$$\epsilon_G := \|\widehat{G}(\{U,T\},\tilde{R}) - G(\{U,T\},\tilde{R})\| = O(\sqrt{\|F \odot \tilde{F}\|_1}) = O\left(n\left(\frac{p-q}{k} + q\right)\right),$$

and

$$\begin{aligned}\sigma_{\min}(G_{\tilde{R},A}) &= \Omega\left(\sqrt{|\tilde{R}| w_{\min}} \cdot \sigma_{\min}(H_A)\right) \\ &= \Omega\left(\sqrt{n \cdot \rho w_{\min}} \cdot \sigma_{\min}(H_A)\right).\end{aligned}$$

From Lemma 12, we have

$$\sigma_{\min}(H_A) = \sigma_{\min}(F_A \odot \tilde{F}_A) = \Omega\left(n(p-q)^2 \min_{i,j\neq i}\left(\mathbb{E}[\pi_i^2] - \mathbb{E}[\pi_i\pi_j]\right)\right).$$

Finally note that $\sigma_k(\mathbb{E}[\pi\pi^\top]) = \Theta\left(\min_{i,j\neq i}\left(\mathbb{E}[\pi_i^2] - \mathbb{E}[\pi_i\pi_j]\right)\right)$. Substituting we have the result. □

Let $\tilde{\Pi}_Z$ be the reconstruction after the tensor method (before thresholding) on resource subset $Z \subset R - \tilde{R}$ (we do not incorporate $\tilde{R}$ to avoid dependency issues), i.e.

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

Lemma 9 (Reconstruction of communities (before thresholding)) *We have w.h.p.*

$$\epsilon_\pi := \max_{i\in Z} \|(\tilde{\Pi}_Z)^i - (\Pi_Z)^i\| = \frac{\epsilon_{\mathcal{T}}}{\sqrt{k}} \|\Pi_Z\| = O\left(\frac{\epsilon_{\mathcal{T}}}{\sqrt{k}} \cdot \sqrt{n} \|\mathbb{E}[\pi\pi^\top]\|\right). \tag{18}$$

Proof: This is on lines of (Anandkumar et al., 2014a, Lemma 13). □

C.5 Concentration of Graph Moments

Lemma 10 (Concentration of hyper-edges) *With probability $1-\delta$, given community membership vectors Π,*

$$\epsilon_G := \|\widehat{G}(\{U,T\},Y) - G(\{U,T\},Y)\| = O(\max(\sqrt{\|F \odot \tilde{F}\|_1}, \sqrt{\|(P * \tilde{P})\Pi_Y\|_\infty}))$$

Remark: When number of nodes n is large enough, the first term, viz., $\sqrt{\|F \odot \tilde{F}\|_1}$ dominates.

Proof: The proof is on the lines of (Anandkumar et al., 2013, Lemma 22) but adapted to the setting of hyper-adjacency rather than adjacency matrices. Let $m_y := \widehat{G}(\{U,T\},y) - G(\{U,T\},y)$ and $M_y := m_y e_y^\top$ and thus

$$\widehat{G}(\{U,T\},Y) - G(\{U,T\},Y) = \sum_y M_y,$$

Note that the random matrices M_y are conditionally independent for $y \in Y$ since m_y are conditionally independent given π_y, and in each vector m_y, the entries are independent as well. We apply matrix Bernstein's inequality. We have $\mathbb{E}[M_y|\Pi] = 0$. We compute the variances $\sum_{y\in Y} \mathbb{E}[M_y M_y^\top|\Pi]$ and $\sum_y \mathbb{E}[M_y^\top M_y|\Pi]$. We have that $\sum_y \mathbb{E}[M_y M_y^\top|\Pi]$ only the diagonal terms are non-zero due to independence, and

$$\mathbb{E}[M_y M_y^\top|\Pi] \leq \mathrm{Diag}((F \odot \tilde{F})\pi_y) \tag{19}$$

entry-wise, assuming Bernoulli random variables. Thus,

$$\begin{aligned}
\|\sum_{y\in Y} \mathbb{E}[M_y M_y^\top|\Pi]\| &\leq \max_{u\in U, t\in T} \sum_{y\in Y, j\in[k]} F(u,j)\tilde{F}(t,j)\pi_y(j) \\
&= \max_{u\in U, t\in T} \sum_{y\in Y, j\in[k]} F(u,j)\tilde{F}(t,j)\Pi_Y(j,y) \\
&\leq \cdot \max_{i\in[k]} \sum_{y\in Y, j\in[k]} P(i,j)\tilde{P}(i,j)\Pi_Y(j,y) \\
&= \|(P * \tilde{P})\Pi_Y\|_\infty,
\end{aligned} \tag{20}$$

where $*$ indicates Hadamard or entry-wise product. Similarly $\sum_{y\in Y} \mathbb{E}[M_y^\top M_y] = \sum_{y\in Y} \mathrm{Diag}(\mathbb{E}[m_y^\top m_y]) \leq \|(P * \tilde{P})\Pi_Y\|_\infty$. From Lemma 11, we have a bound $\|(P * \tilde{P})\Pi_Y\|_\infty$.

We now bound $\|M_y\| = \|m_y\|$ through vector Bernstein's inequality. We have for Bernoulli $\widehat{G}$,

$$\max_{u\in U, t\in T} |\widehat{G}(\{u,t\},y) - G(\{u,t\},y)| \leq 2$$

and

$$\sum_{u\in U, t\in T} \mathbb{E}[\widehat{G}(\{u,t\},y) - G(\{u,t\},y)]^2 \leq \sum_{u\in U, t\in T} ((F \odot \tilde{F})\pi_y)_{ut} \leq \|F \odot \tilde{F}\|_1.$$

Thus with probability $1-\delta$, we have

$$\|M_y\| \leq (1 + \sqrt{8\log(1/\delta)})\sqrt{\|F \odot \tilde{F}\|_1 \cdot} + 8/3\log(1/\delta).$$

Thus, we have the bound that $\|\sum_y M_y\| = O(\max(\sqrt{\|F \odot \tilde{F}\|_1}, \sqrt{\|(P * \tilde{P})\Pi_Y\|_\infty}))$. □

For a given $\delta \in (0,1)$, we assume that the sets U, T and $Y \subset R$ are large enough to satisfy

$$\sqrt{|U| \cdot |T|} \geq \frac{8}{3} \log \frac{|U| \cdot |T|}{\delta}$$
$$\sqrt{|Y|} \geq \frac{8}{3} \log \frac{|Y|}{\delta}.$$

Lemma 11 (Concentration bounds) *With probability* $1 - \delta$*,*

$$\begin{aligned}
\|F \odot \tilde{F}\|_1 &\leq |U| \cdot |T| \max_i (P \cdot \mathbb{E}[\pi])_i \max_i (\tilde{P} \cdot \mathbb{E}[\pi])_i + \sqrt{\frac{8}{3}|U| \cdot |T| \cdot P_{\max}^4 \cdot \log \frac{|U| \cdot |T|}{\delta}}, \\
|\|(F \odot \tilde{F})_i\| &\leq \max_i \|\Pi_U\| \cdot \|\Pi_T\| \cdot \|P_i\| \cdot \|\tilde{P}_i\| \\
&= O\left(\sqrt{|U| \cdot |T|} \|\mathbb{E}[\pi\pi^\top]\| \cdot (p - q + \sqrt{k}q)\right),
\end{aligned}$$

for the homogeneous setting. Similarly for subset $Y \subset R$*, we have*

$$\|\Pi_Y \Pi_Y^\top\| \leq |Y| \cdot \|\mathbb{E}[\pi\pi^\top]\| + \sqrt{\frac{8}{3}|Y| \cdot \|\mathbb{E}[\pi\pi^\top]\|^2 \cdot \log \frac{|Y|}{\delta}}$$
$$\sigma_k(\Pi_Y \Pi_Y^\top) \geq |Y| \cdot \sigma_k(\mathbb{E}[\pi\pi^\top]) - \sqrt{\frac{8}{3}|Y| \cdot \|\mathbb{E}[\pi\pi^\top]\|^2 \cdot \log \frac{|Y|}{\delta}}$$
$$\|(P * \tilde{P})^\top \Pi_Y\|_\infty \leq |Y| \max_i (\mathbb{E}[\pi]^\top \cdot (P * \tilde{P}))_i + \sqrt{\frac{8}{3}|Y| \cdot P_{\max}^4 \cdot \log \frac{|Y|}{\delta}}$$

Remark: Note that $\sigma(P) = \Theta(p - q)$ and $\|P\| = \Theta(p + q)$ for homogeneous P. Under Assumption A3, the variance terms are small and the above quantities are close to their expectation.

Proof: To bound on $\|F \odot \tilde{F}\|_1$, we note that $\|\mathbb{E}[F \odot \tilde{F}]\|_1 \leq |U| \cdot |T| \max_i (P^\top \cdot \mathbb{E}[\pi])_i (\tilde{P}^\top \cdot \mathbb{E}[\pi])_i$. Using Bernstein's inequality, for each column of $F \odot \tilde{F}$, we have, with probability $1 - \delta$,

$$\left| \|(F \odot \tilde{F})_i\|_1 - |U| \cdot |T| \langle \mathbb{E}[\pi], (P)_i \rangle \langle \mathbb{E}[\pi], (\tilde{P})_i \rangle \right| \leq \sqrt{\frac{8}{3}|U| \cdot |T| \cdot P_{\max}^4 \cdot \log \frac{|U| \cdot |T|}{\delta}},$$

by applying Bernstein's inequality, since $\langle \pi, (P)_i \rangle \langle \pi, (\tilde{P})_i \rangle \leq \max_i (P^\top \pi)_i (\tilde{P}^\top \pi)_i \leq P_{\max}^2$, and

$$\max \left(\sum_{u \in U, t \in T} \|\mathbb{E}[(P)_i^\top \pi_u \pi_u^\top (P)_i] \cdot \mathbb{E}[(\tilde{P})_i^\top \pi_t \pi_t^\top (\tilde{P})_i]\|, \sum_{u \in U, t \in T} \|\mathbb{E}[\pi_u^\top (P)_i (P)_i^\top \pi_u] \cdot \mathbb{E}[\pi_t^\top (\tilde{P})_i (\tilde{P})_i^\top \pi_t]\| \right)$$
$$\leq |U| \cdot |T| \cdot P_{\max}^4.$$

The other results follow similarly. □

The lowest singular value for the Khatri-Rao product is a bit more involved and we provide the bound below.

Lemma 12 (Spectral Bound for KR-product)

$$\sigma_k^2(F \odot \tilde{F}) \geq |U| \cdot |T| \sigma_k(\Gamma * \Gamma) - \sqrt{\frac{8}{3}|U| \cdot |T| \cdot \|P\|^2 \cdot \|\tilde{P}\|^2 \cdot \|\mathbb{E}[\pi\pi^\top]\|^2 \cdot \log \frac{|U| \cdot |T|}{\delta}},$$

where $\Gamma := P^\top \mathbb{E}[\pi\pi^\top] \tilde{P}$ *and* $*$ *denotes Hadamard product.*

Proof: The result in the Lemma follows directly from the concentration result. For the homogeneous setting, we have for a matrix Γ,

$$\sigma_k(\Gamma * \Gamma) = \Theta\left(\min_i \Gamma(i,i)^2 - \max_{i \neq j} \Gamma(i,j)^2\right).$$

Substituting we have the result. □

Remark: For the homogeneous setting, with $P = \tilde{P}$ having p on the diagonal and q on the off-diagonal, we have

$$\begin{aligned}\Gamma &= \left[(p-q)I + q11^\top\right] \mathbb{E}[\pi\pi^\top] \left[(p-q)I + q11^\top\right] \\ &= (p-q)^2 \mathbb{E}[\pi\pi^\top] + 2(p-q)qv1^\top + q^2 \|\mathbb{E}[\pi\pi^\top]\|_{sum} 11^\top,\end{aligned}$$

where v is a vector where $v_i = \|\mathbb{E}[\pi\pi^\top]^{(i)}\|_1$, where $M^{(i)}$ denotes the i^{th} row of M. Thus, we have the following bound

$$\begin{aligned}\sigma_k(\Gamma * \Gamma) &= \left(\min_{i,j \neq i} \left(\Gamma(i,i)^2 - \Gamma(i,j)^2\right)\right) \\ &= \Theta\left((p-q)^4 \min_{i,j \neq i} \left(\mathbb{E}(\pi_i^2) - \mathbb{E}[\pi_i \pi_j]\right)^2\right),\end{aligned}$$

assuming that $\mathbb{E}[\pi_i^2] - \mathbb{E}[\pi_i\pi_j] = \Theta(\mathbb{E}[\pi_i^2])$ for all $i \neq j$, and the other terms which are dropped are positive. Thus, we have w.h.p.

$$\sigma_k(F \odot \tilde{F}) = \Omega\left(n(p-q)^2 \min_{i,j \neq i} \left(\mathbb{E}[\pi_i^2] - \mathbb{E}[\pi_i\pi_j]\right)\right) \tag{21}$$

D Standard Matrix Concentration and Perturbation Bounds

D.1 Bernstein's Inequalities

One of the key tools we use is the standard matrix Bernstein inequality (Tropp, 2012, thm. 6.1, 6.2).

Proposition 2 (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 $\sigma^2 = \max\left\{\| \sum_j \mathbb{E}[W_j W_j^\top], \| \sum_j \mathbb{E}[W_j^\top W_j]\right\}$, then we have

$$\begin{aligned}\Pr[\| Z \geq t] \leq & d \cdot exp\left\{\frac{-t^2/2}{\sigma^2 + Rt/3}\right\} \\ \leq & d \cdot exp\left\{\frac{-3t^2}{8\sigma^2}\right\}, \quad t \leq \sigma^2/R, \\ \leq & d \cdot exp\left\{\frac{-3t}{8R}\right\}, \quad t \geq \sigma^2/R\end{aligned}$$

Proposition 3 (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*

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

D.2 Hanson-Wright Inequalities

We require the Hanson-Wright inequality (Rudelson and Vershynin, 2013).

Proposition 4 (Hanson-Wright Inequality: sub-Gaussian bound) *Let $z = (z_1, z_2, ..., z_n) \in \mathbb{R}^n$ be a random vector with independent entries, $\mathbb{E}[z_i] = 0$ and $\Pr[|z_i| \leq 1] = 1$ and let $M \in \mathbb{R}^{n\times n}$ be any matrix. There exists a constant $c > 0$ s.t.*

$$\Pr\left[|z^\top Mz - \mathbb{E}(z^\top Mz)| > t\right] \leq 2\exp\left[-c\min\left(\frac{t^2}{\|M\|_{\mathbb{F}}^2}, \frac{t}{\|M\|}\right)\right]$$

Unfortunately the sub-Gaussian bound is not strong enough when z has small variance σ^2. In this case, we get the perturbation as $\tilde{O}(\|M\|_{\mathbb{F}})$ instead of $\tilde{O}(\sigma\|M\|_{\mathbb{F}})$, which is desired. This is because for a bounded random variable, the sub-Gaussian parameter only depends on the bound and not on the variance.

We will consider an extension of the Hanson-Wright inequality to sub-exponential random variables (Erdős et al., 2012; Vu and Wang, 2013) and employ the sub-exponential formulation for bounded random variables. We first define sub-exponential random variable (Vershynin, 2010, Definition 5.13).

Definition 1 (Sub-exponential Random Variable) *A zero-mean random variable X is said to be sub-exponential if there exists a parameter K such that $\mathbb{E}[e^{X/K}] \leq e$.*

Remark: There are other equivalent notions for sub-exponential random variables (Vershynin, 2010, Definition 5.13), but this will be the convenient one for proving sub-exponential bound for Bernoulli random variables. It is easy to see that the centered Bernoulli random variables are sub-exponential for some constant K.

We will employ the following version of Hanson-Wright's inequality for sub-exponential random variables (Erdős et al., 2012, Lemma B.2).

Proposition 5 (Hanson-Wright Inequality: sub-exponential bound) *Let $z = (z_1, z_2, ..., z_n) \in \mathbb{R}^n$ be a random vector with independent entries, $\mathbb{E}[z_i] = 0$, $\mathbb{E}[z_i^2] \leq \sigma^2$ and z_i are sub-exponential and let $M \in \mathbb{R}^{n\times n}$ be any matrix. There exists constants $c, C > 0$ s.t.*

$$\Pr\left[|z^\top Mz - \mathbb{E}(z^\top Mz)| > t\sigma^2\|M\|_{\mathbb{F}}\right] \leq C\exp\left[-ct^{1/4}\right].$$

Remark: The result in the form above appears in (Vu and Wang, 2013, (13)) and we set $\alpha = 1$ in (Vu and Wang, 2013, (13)). The parameter C above differs from the sub-exponential parameter K by only a constant factor.

Comparing sub-exponential formulation in Proposition 5 with sub-Gaussian formulation in Proposition 4, we see that in the former, the deviation is $\tilde{O}(\|M\|_{\mathbb{F}}\sigma)$, while in the latter it is only $\tilde{O}(\|M\|_{\mathbb{F}})$.

Thus, for centered Bernoulli random variables and we can employ Proposition 5, and we will use it for distance concentration bounds.

D.3 Davis-Kahan Inequality

We also use the standard Davis and Kahan bound for subspace perturbation.

Proposition 6 (Davis and Kahan) *For a matrix $\hat{A}$, let $\widehat{\mathrm{Proj}}$ be the projection matrix on to its top-k left singular vectors. For any rank-k matrix A, we have*

$$\|(\widehat{\mathrm{Proj}} - I)A\| \leq 2\|\hat{A} - A\|$$

Proof: This is directly from (McSherry, 2001, Lemma 12). By writing $A = \hat{A} - (\hat{A} - A)$, we have

$$\|(\widehat{\mathrm{Proj}} - I)A\| \leq \|(\widehat{\mathrm{Proj}} - I)\hat{A}\| + \|(\widehat{\mathrm{Proj}} - I)(\hat{A} - A)\|,$$

and each of the terms is less than $\|\hat{A} - A\|$. For the first term, it is because $\widehat{\mathrm{Proj}}\,\hat{A}$ is the best rank-k approximation of $\hat{A}$ and since A is also rank k, the residual $\|(\widehat{\mathrm{Proj}} - I)\hat{A}\| \leq \|\hat{A} - A\|$. For the second term, $\|(\widehat{\mathrm{Proj}} - I)(\hat{A} - A)\| \leq \|\hat{A} - A\|$ since $(\widehat{\mathrm{Proj}} - I)$ cannot increase norm. □

References

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

A. Anandkumar, R. Ge, D. Hsu, and S. M. Kakade. A Tensor Spectral Approach to Learning Mixed Membership Community Models. In *Conference on Learning Theory (COLT)*, June 2013.

Animashree Anandkumar, Rong Ge, Daniel Hsu, and Sham M Kakade. A tensor approach to learning mixed membership community models. *The Journal of Machine Learning Research*, 15(1):2239–2312, 2014a.

Animashree Anandkumar, Rong Ge, and Majid Janzamin. Guaranteed non-orthogonal tensor decomposition via alternating rank-1 updates. *arXiv preprint arXiv:1402.5180*, 2014b.

Michael Brinkmeier, Jeremias Werner, and Sven Recknagel. Communities in graphs and hypergraphs. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 869–872. ACM, 2007.

Abhijnan Chakraborty and Saptarshi Ghosh. Clustering hypergraphs for discovery of overlapping communities in folksonomies. In *Dynamics On and Of Complex Networks, Volume 2*, pages 201–220. Springer, 2013.

Abhijnan Chakraborty, Saptarshi Ghosh, and Niloy Ganguly. Detecting overlapping communities in folksonomies. In *Proceedings of the 23rd ACM conference on Hypertext and social media*, pages 213–218. ACM, 2012.

László Erdős, Horng-Tzer Yau, and Jun Yin. Bulk universality for generalized wigner matrices. *Probability Theory and Related Fields*, 154(1-2):341–407, 2012.

F. Huang, U.N. Niranjan, M. Hakeem, and A. Anandkumar. Fast Detection of Overlapping Communities via Online Tensor Methods. *ArXiv 1309.0787*, Sept. 2013.

Stefanie Jegelka, Suvrit Sra, and Arindam Banerjee. Approximation algorithms for tensor clustering. In *Algorithmic learning theory*, pages 368–383. Springer, 2009.

Ioannis Konstas, Vassilios Stathopoulos, and Joemon M Jose. On social networks and collaborative recommendation. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 195–202. ACM, 2009.

Yu-Ru Lin, Jimeng Sun, Paul Castro, Ravi Konuru, Hari Sundaram, and Aisling Kelliher. Metafac: community discovery via relational hypergraph factorization. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 527–536. ACM, 2009.

F. McSherry. Spectral partitioning of random graphs. In *FOCS*, 2001.

Tsuyoshi Murata. Detecting communities from tripartite networks. In *Proceedings of the 19th international conference on World wide web*, pages 1159–1160. ACM, 2010.

Nicolas Neubauer and Klaus Obermayer. Towards community detection in k-partite k-uniform hypergraphs. In *Proceedings of the NIPS 2009 Workshop on Analyzing Networks and Learning with Graphs*, pages 1–9, 2009.

Symeon Papadopoulos, Yiannis Kompatsiaris, and Athena Vakali. A graph-based clustering scheme for identifying related tags in folksonomies. In *Data Warehousing and Knowledge Discovery*, pages 65–76. Springer, 2010.

Mark Rudelson and Roman Vershynin. Hanson-wright inequality and sub-gaussian concentration. *arXiv preprint arXiv:1306.2872*, 2013.

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

Alexei Vazquez. Finding hypergraph communities: a bayesian approach and variational solution. *Journal of Statistical Mechanics: Theory and Experiment*, 2009(07):P07006, 2009.

Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

Van Vu and Ke Wang. Random weighted projections, random quadratic forms and random eigenvectors. *arXiv preprint arXiv:1306.3099*, 2013.

Xufei Wang, Lei Tang, Huiji Gao, and Huan Liu. Discovering overlapping groups in social media. In *Data Mining (ICDM), 2010 IEEE 10th International Conference on*, pages 569–578. IEEE, 2010.

Shengliang Xu, Shenghua Bao, Ben Fei, Zhong Su, and Yong Yu. Exploring folksonomy for personalized search. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 155–162. ACM, 2008.

Jaewon Yang and Jure Leskovec. Overlapping community detection at scale: A nonnegative matrix factorization approach. In *Proceedings of the sixth ACM international conference on Web search and data mining*, pages 587–596. ACM, 2013.

Chen Yudong, Sujay Sanghavi, and Huan Xu. Clustering sparse graphs. In *Advances in Neural Information Processing Systems 25*, 2012.