Detecting Overlapping and Correlated Communities without Pure Nodes: Identifiability and Algorithm

Kejun Huang [1] **Xiao Fu** [2]

Abstract

Many machine learning problems come in the form of networks with relational data between entities, and one of the key unsupervised learning tasks is to detect communities in such a network. We adopt the mixed-membership stochastic blockmodel as the underlying probabilistic model, and give conditions under which the memberships of a subset of nodes can be uniquely identified. Our method starts by constructing a second-order graph moment, which can be shown to converge to a specific product of the true parameters as the size of the network increases. To correctly recover the true membership parameters, we formulate an optimization problem using insights from convex geometry. We show that if the true memberships satisfy a so-called sufficiently scattered condition, then solving the proposed problem correctly identifies the ground truth. We also propose an efficient algorithm for detecting communities, which is significantly faster than prior work and with better convergence properties. Experiments on synthetic and real data justify the validity of the proposed learning framework for network data.

1. Introduction

A lot of machine learning problems deal with pair-wise relational data. Examples include social networks and gene interactions. One of the key analytical questions is to detect latent communities from the ambient interactions in an unsupervised manner. Traditional methods usually deem this as a clustering problem on graphs, and classical algorithms like NormalizedCuts (Meila & Shi, 2001) and spectral clustering (Ng et al., 2002) have been successfully applied in practice. However, most clustering based learning frameworks lack the ability of representing more complicated network structures, for example by modeling interaction patterns *between* communities and *mixed-memberships* of the nodes. Several probabilistic latent variable models for networks have been proposed to enhance interpretability and expressibility of the model (Goldenberg et al., 2010).

1.1. Stochastic blockmodel and MMSB

The most famous probabilistic models for networks are perhaps the stochastic block model (Snijders & Nowicki, 1997; Nowicki & Snijders, 2001) and its mixed-membership variant (Airoldi et al., 2008). In these probabilistic models, an edge A_{ij} is present or absent follows a Bernoulli distribution with parameter P_{ij}, i.e.,

$$\Pr(A_{ij} = \{0, 1\}) = P_{ij}^{A_{ij}}(1 - P_{ij})^{1-A_{ij}}. \tag{1}$$

Furthermore, these models assume that the matrix $\boldsymbol{P}$, whose i, j-th entry is P_{ij}, admits the structure that

$$\boldsymbol{P} = \boldsymbol{M}^\top \boldsymbol{B} \boldsymbol{M}, \tag{2}$$

where $\boldsymbol{B}$ is $k \times k$, assuming there are k communities in the network, and $\boldsymbol{M}$ has k rows. The value $B_{pq} \in [0, 1]$ indicates the probability that a node in community p connects with a node in community q, for all $p, q = 1, ..., k$. The i-th column of $\boldsymbol{M}$, denoted as $\boldsymbol{m}_i$, represents the community membership of node i. In the stochastic blockmodel (SB), $\boldsymbol{m}_i$'s are restricted to be coordinate vectors, indicating that each node belongs to one and only one community; whereas in the mixed-membership stochastic blockmodel (MMSB), $\boldsymbol{m}_i$'s belong to the probability simplex Δ, defined as

$$\Delta = \{\boldsymbol{x} : \boldsymbol{x} \geq 0, \mathbf{1}^\top \boldsymbol{x} = 1\}, \tag{3}$$

allowing each node to hold mixed memberships across different communities. We will be focusing on the MMSB assumptions since it is more flexible to model complex membership structures in the network.

There are some interesting properties of the MMSB. The communities are allowed to be *overlapping* because of the inter-community interactions modeled by the nonzero off-diagonals in $\boldsymbol{B}$. Depending on whether we are dealing with

[1]Department of Computer and Information Science and Engineering, University of Florida, Gainesville, FL, USA [2]School of Electrical Engineering and Computer Science, Oregon State University, Corvallis, OR, USA. Correspondence to: Kejun Huang <kejun.huang@ufl.edu>, Xiao Fu <xiao.fu@oregonstate.edu>.

Proceedings of the 36th International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019.

an undirected or directed network, we can restrict the community interaction matrix $\boldsymbol{B}$ to be symmetric or asymmetric, respectively. In both cases the ground-truth network structure $\boldsymbol{M}^\top\boldsymbol{B}\boldsymbol{M}$ remains the same, so we can simply look at the community membership matrix $\boldsymbol{M}$ to determine communites. Another somewhat counter-intuitive property is that $\boldsymbol{B}$ does not have to be diagonally dominant, meaning that typical interaction patterns for a community may consists of more inter-community ones than intra-community ones. For example, the community of lawyers typically interact more with their clients than between peers. This type of interaction pattern cannot be captured by traditional graph-cut based methods, but can be easily modeled by MMSB by letting $\boldsymbol{B}$ taking larger off-diagonal entries than their corresponding diagonal terms.

On the other hand, parameter estimation for MMSB is challenging. First and foremost, the model may not be identifiable. The only restrictions on the model parameters are that the values in $\boldsymbol{B}$ are between 0 and 1, and the columns of $\boldsymbol{M}$ belong to the probability simplex Δ. Therefore, there may exist an invertible matrix $\boldsymbol{Q}$ such that $\boldsymbol{Q}\boldsymbol{M} \in \Delta$ and $0 \le \boldsymbol{Q}^{-\top}\boldsymbol{B}\boldsymbol{Q}^{-1} \le 1$ still hold, and clearly it does not affect their product.

Apart from a lack of identifiability, maximum-likelihood estimation for MMSB is also computationally hard. Due to the tri-linearity of the model parameterization $\boldsymbol{M}^\top\boldsymbol{B}\boldsymbol{M}$, optimizing the log-likelihood is a non-convex problem, which means obtaining a global optimum is in general hard to guarantee. Another difficulty is exploiting sparsity for large-scale networks. According to the Bernoulli generative model (1), the maximum-likelihood formulation of a network of n nodes involves the summation of $\mathcal{O}(n^2)$ terms, even if there are only $\mathcal{O}(n)$ or $\mathcal{O}(n \log n)$ edges. When Airoldi et al. (2008) first proposed the MMSB, they introduced a sparsity parameter $\rho \in [0, 1]$ to specify the proportion of non-edges that are generated from the Bernoulli model (1) while the rest are simply not observed. However, that does not mitigate the computational burden (mainly in terms of the $\mathcal{O}(n^2)$ memory requirement) even if the network is very sparse—in fact, introducing the additional parameter ρ is exactly equivalent to scaling down the $\boldsymbol{B}$ matrix by ρ, which does not affect the overall model except by lowering the probability of observing edges. Because of this, traditional methods like the expectation-maximization (Nowicki & Snijders, 2001), variational Bayes (Airoldi et al., 2008), and Gibbs sampling (Hanneke & Xing, 2007) are typically not scalable for large networks.

1.2. Matrix/tensor factorization methods

Recently, there has been a line of work that uses nonnegative matrix factorization for learning the parameters of MMSB on a large network, including Yang & Leskovec (2013); Zhang et al. (2014); Kaufmann et al. (2016); Jin et al. (2017); Panov et al. (2017); Mao et al. (2017; 2018). These methods all start by assuming that there is a way of estimating the $\boldsymbol{P}$ matrix defined in (2) underlying the ambient Bernoulli observations. Then the next step is to find a unique factorization $\boldsymbol{P} = \boldsymbol{M}^\top\boldsymbol{B}\boldsymbol{M}$ in order to extract the useful model parameters $\boldsymbol{M}$ and $\boldsymbol{B}$. Despite nuances in the assumptions, the key condition that gives rise to identifiability in these works is the existance of a *pure node* for every community. For community p, a pure node is defined as a node i such that $\boldsymbol{m}_i = \boldsymbol{e}_p$, meaning this node has no membership in all the communities except the p-th one.

The so-called *pure node* assumption is exactly the *separability* assumption first proposed by Donoho & Stodden (2004) to ensure identifiability of nonnegative matrix factorization (NMF). It has many anonyms in unsupervised learning applications where identifiability of the latent component is crucial, for example *anchor word* in topic modeling (Arora et al., 2012; 2013), *pure pixel* in hyperspectral unmixing (Nascimento & Dias, 2005; Ma et al., 2014), *local dominance* in blind source separation (Chan et al., 2008; Fu et al., 2015), and here the *pure node* assumption in community detection. A salient feature of separability is that it not only guarantees identifiability, but also leads to efficient algorithms to retrieve the solution, for example via greedy search algorithms that date back to Araújo et al. (2001). However, separability or *pure node* is a somewhat restrictive assumption in practice—for model robustness, it would be more ideal not to rely on such strong assumptions for the algorithm to work.

A question one may ask is whether it is possible to obtain an accurate estimate of $\boldsymbol{P}$. One may assume that we keep on observing Bernoulli samples generated according to the parameter $\boldsymbol{P}$. However, we should notice that the MMSB model $\boldsymbol{A} \sim \text{Bernoulli}(\boldsymbol{M}^\top\boldsymbol{B}\boldsymbol{M})$ only applies to the off-diagonal entries of the graph adjacency matrix $\boldsymbol{A}$. The diagonal entries of $\boldsymbol{A}$ are, by definition, equal to zero. In fact, a diagonal entry of $\boldsymbol{P}$, which equals to $\boldsymbol{m}_i^\top\boldsymbol{B}\boldsymbol{m}_i$, does not have any physical meaning, because the conditional independence assumption does not apply to a node with itself. This means even if we can observe multiple Bernoulli draws from $\boldsymbol{M}^\top\boldsymbol{B}\boldsymbol{M}$, it is impossible to directly estimate its diagonal entries. One can verify that the argument still holds for the graph Laplacian matrix constructed from $\boldsymbol{A}$.

Anandkumar et al. (2014a) proposed a more viable approach for applying factorization techniques based on the method of moments. Inspired by Bickel et al. (2011), they propose to construct a three-way tensor by dividing the large network into four disjoint sets of nodes, $\mathcal{S}_0$, $\mathcal{S}_1$, $\mathcal{S}_2$, and $\mathcal{S}_3$, and counting the number of 3-star subgraphs. Specifically, a 3-star is a subgraph in which a node in $\mathcal{S}_0$ connects to three other nodes, each belonging to $\mathcal{S}_1$, $\mathcal{S}_2$, and $\mathcal{S}_3$. Consider

four nodes $i_0 \in \mathcal{S}_0$, $i_1 \in \mathcal{S}_1$, $i_2 \in \mathcal{S}_2$, and $i_3 \in \mathcal{S}_3$, it is easy to see that such a 3-star subgraph exists with probability

$$(\boldsymbol{m}_{i_1}^\top \boldsymbol{B}\boldsymbol{m}_{i_0})(\boldsymbol{m}_{i_2}^\top \boldsymbol{B}\boldsymbol{m}_{i_0})(\boldsymbol{m}_{i_3}^\top \boldsymbol{B}\boldsymbol{m}_{i_0}).$$

Summing this up over all $i_0 \in \mathcal{S}_0$ and divide by the size of $\mathcal{S}_0$, this quantity goes asymptotically to

$$T_{i_1 i_2 i_3} = \sum_{j_1,j_2,j_3=1}^{k} G_{j_1 j_2 j_3} \widetilde{m}_{i_1 j_1} \widetilde{m}_{i_2 j_2} \widetilde{m}_{i_3 j_3}, \qquad (4)$$

where $\widetilde{\boldsymbol{m}}_i = \boldsymbol{B}^\top \boldsymbol{m}_i$ and $G_{j_1 j_2 j_3} = \mathrm{E}\left[m_{i_0 j_1} m_{i_0 j_2} m_{i_0 j_3}\right]$. Tensor $\boldsymbol{T}$ would admit a *canonical polyadic decomposition* (CPD) if $\boldsymbol{G}$ is a super-diagonal tensor, which is true if the simple stochastic blockmodel is considered. For MMSB, Anandkumar et al. (2014a) argued that such a CPD structure can still be constructed via a somewhat complicated "centering" procedure, which works if the $\boldsymbol{m}_{i_0}$'s are all generated from a Dirichlet distribution with a *known* intensity parameter α_0. Tensor CPD is known to be essentially unique under mild conditions (Sidiropoulos et al., 2017), and in the more restrictive case when all the latent factors have full column rank, there exist guaranteed algorithms like the higher-order power method to retrieve the solution efficiently (Anandkumar et al., 2014b).

1.3. This paper

In this paper, we adopt the graph moment approach by Anandkumar et al. (2014b), but instead count the number of 2-star subgraphs by dividing the network into only three sets $\mathcal{S}_0$, $\mathcal{S}_1$, and $\mathcal{S}_2$. One immediate advantage is that the sheer number of 2-stars in a network is significantly larger than the number of 3-stars, as illustrated in Figure 1. We will specifically set $\mathcal{S}_2$ to be the set of nodes that one is interested in detecting their communities, $\mathcal{S}_1$ to consist of only $k-1$ nodes (one less than the number of latent communities), and the rest all go to $\mathcal{S}_0$ to secure an accurate estimate of graph moment.

The main technique we use to uniquely recover the community membership structure of $\mathcal{S}_2$ is via a geometric intuition of finding the minimum volume enclosing simplex that covers the entire set of points obtained from the second-order moment. We will show that, for this method to be able to uniquely recover the community memberships, the only assumption we need is that mixed-membership coefficients of the nodes in $\mathcal{S}_2$ are *sufficiently scattered*, a geometric condition that will be explained in detail in the sequel. A notable difference between our result and all the aforementioned matrix factorization techniques is that we do not require the existence of pure nodes for any community. The sufficiently scattered condition includes separability, or pure nodes in the context of community detection, as a special case, but is much more general than the pure nodes assumption.

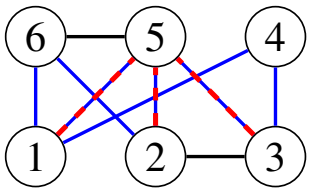


Figure 1. For this simple network of 6 nodes, if $\mathcal{S}_1 = \{1\}$, $\mathcal{S}_2 = \{2\}$, $\mathcal{S}_3 = \{3\}$, $\mathcal{S}_0 = \{4,5,6\}$, we have 3 samples to estimate the third-order moment of $\mathcal{S}_1, \mathcal{S}_2$, and $\mathcal{S}_3$ using 3-stars, out of which only one is nonzero $5 \to (1,2,3)$ shown in dashed red. If instead we let $\mathcal{S}_1 = \{1\}$, $\mathcal{S}_2 = \{2,3\}$, $\mathcal{S}_0 = \{4,5,6\}$ and estimate the second-order moment of $\mathcal{S}_1$ and $\mathcal{S}_2$, we have in total 6 samples of 2-stars, out of which 4 are nonzeros, highlighted in blue. The difference in sample sizes is much more significant for large networks in practice.

Compared to the tensor decomposition approach proposed by Anandkumar et al. (2014a), we note that our method does not require that the $\boldsymbol{m}_{i_0}$ vectors to be drawn from a known Dirichlet distribution. All we require is that the second moment $\mathrm{E}[\boldsymbol{m}_{i_0}\boldsymbol{m}_{i_0}^\top]$ exists, which is very easy to satisfy in practice. This means the components of $\boldsymbol{m}_{i_0}$ can be highly *correlated*, for example by following a logistic normal distribution (Blei & Lafferty, 2006), in which case it is not possible to modify the third-order moment to admit a low-rank CPD structure. Our method, on the other hand, is still able to recover the latent communities with identifiability guarantees, and there is no need to perform any sophisticated "centering" procedure.

Besides guaranteed identifiability under more relaxed conditions, we also provide a computationally efficient algorithm to detect the ground-truth communities. Even though the new learning framework induces a non-convex optimization problem that is in general NP-hard to solve, we show that our method ensures convergence to a stationary point. In the special case when the restrictive pure-node assumption does hold, our method is able to recover the true community memberships in one iteration.

2. A Simpler Graph Moment Construction

We start by describing how to construct the second-order moment using 2-stars to form the nonnegative matrix for us to learn the community memberships. The network is first divided into three disjoint sets of nodes $\mathcal{S}_0$, $\mathcal{S}_1$, and $\mathcal{S}_2$. The set $\mathcal{S}_2$ consists of n nodes that one is interested in finding their community memberships. The set $\mathcal{S}_1$ consists of $k-1$ nodes, which is one less than the number of communities (and assumed to be known, as in most other methods). The set $\mathcal{S}_0$ consists of all the other nodes to act as 2-star samples in order to construct a $(k-1) \times n$ matrix $\widehat{\boldsymbol{Y}}$.

Specifically, given the adjacency matrix $\boldsymbol{A}$ for the entire

network, with $(0,1)$ weights, the $i_1 i_2$-th entry of $\widehat{\boldsymbol{Y}}$ is defined as

$$\widehat{Y}_{i_1 i_2} = \frac{1}{|\mathcal{S}_0|} \sum_{i_0 \in \mathcal{S}_0} A_{i_0 i_1} A_{i_0 i_2}. \tag{5}$$

Under the generative model of MMSB, we know that $\mathrm{E}[A_{ij}] = \boldsymbol{m}_i^\top \boldsymbol{B} \boldsymbol{m}_j$ for $i \neq j$. Assuming each edge is *independently* sampled, we have $\mathrm{E}[A_{i_0 i_1} A_{i_0 i_2}] = \mathrm{E}[A_{i_0 i_1}]\,\mathrm{E}[A_{i_0 i_2}]$ for $i_1 \neq i_2$. As a result,

$$\begin{aligned}\mathrm{E}[\widehat{Y}_{i_1 i_2}] &= \frac{1}{|\mathcal{S}_0|} \sum_{i_0 \in \mathcal{S}_0} \boldsymbol{m}_{i_1}^\top \boldsymbol{B}^\top \boldsymbol{m}_{i_0} \boldsymbol{m}_{i_0}^\top \boldsymbol{B} \boldsymbol{m}_{i_2} \\ &= \boldsymbol{m}_{i_1}^\top \boldsymbol{B}^\top \left(\frac{1}{|\mathcal{S}_0|} \sum_{i_0 \in \mathcal{S}_0} \boldsymbol{m}_{i_0} \boldsymbol{m}_{i_0}^\top \right) \boldsymbol{B} \boldsymbol{m}_{i_2}.\end{aligned}$$

Let $\boldsymbol{\Sigma} = \mathrm{E}[\boldsymbol{m}_{i_0} \boldsymbol{m}_{i_0}^\top]$ and let $|\mathcal{S}_0|$ go to infinity, we have that

$$\widehat{\boldsymbol{Y}} \to \boldsymbol{M}_1^\top \boldsymbol{B}^\top \boldsymbol{\Sigma} \boldsymbol{B} \boldsymbol{M}_2\,,$$

where $\boldsymbol{M}_1$ is $k \times (k-1)$ indicating the membership of nodes in $\mathcal{S}_1$ in its columns, and similarly for the columns of $\boldsymbol{M}_2$. Define $\boldsymbol{\Xi} = \boldsymbol{M}_1^\top \boldsymbol{B}^\top \boldsymbol{\Sigma} \boldsymbol{B}$, and let $\boldsymbol{Y} = \mathrm{E}[\widehat{\boldsymbol{Y}}]$, then we result in

$$\boldsymbol{Y} = \boldsymbol{\Xi} \boldsymbol{M}_2\,. \tag{6}$$

A few comments are in order:

- Our construction of $\boldsymbol{Y}$ works for both undirected and directed networks. For directed networks, our construction asks specifically for 2-stars that goes from a node in $\mathcal{S}_0$ to nodes in $\mathcal{S}_1$ and $\mathcal{S}_2$.
- If the edges are weighted but with integer weights, one may interpret it as multiple draws from $\boldsymbol{M}^\top \boldsymbol{B} \boldsymbol{M}$, which simply means we have a lot more samples to estimate $\boldsymbol{Y}$.
- The idea of constructing this graph moment is inspired by Anandkumar et al. (2014a). The main difference is that we collect the number of 2-star subgraphs rather than 3-stars. This, again, naturally provides a lot more samples for us to estimate $\boldsymbol{Y}$.
- As we will see later, the estimate $\widehat{\boldsymbol{Y}}$ can directly be used *as is*, whereas Anandkumar et al. (2014a) require that nodes in $\mathcal{S}_0$ follow a Dirichlet distribution, implying that the $\boldsymbol{\Sigma}$ matrix has a "diagonal plus rank-one" structure. This is not required for our method to work.

3. Identifiability from Convex Geometry

As per our moment construction, we now have an estimate of $\boldsymbol{Y} = \boldsymbol{\Xi} \boldsymbol{M}_2$, and we want to find a unique representation of $\boldsymbol{M}_2$, which represents the mixed-membership of the nodes in $\mathcal{S}_2$. At first glance, it is a highly under-determined problem—how can we uniquely determine a $(k-1) \times k$ matrix $\boldsymbol{\Xi}$ and a $k \times n$ matrix $\boldsymbol{M}_2$ just from a $(k-1) \times n$ data matrix $\boldsymbol{Y}$? We do have the constraint that both $\boldsymbol{\Xi}$ and $\boldsymbol{M}_2$ can only take nonnegative values, but that is not enough to guarantee identifiability in general, especially when the latent dimension k is in fact larger than one of the ambient dimensions.

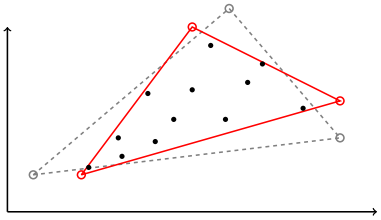


Figure 2. Geometrically, the columns of $\boldsymbol{Y}$ are points in the nonnegative orthant, and columns of $\boldsymbol{\Xi}$ define a simplex that contains all columns of $\boldsymbol{Y}$. However, such an enclosing simplex is not unique. There are in fact infintely many enclosing simplexes in the nonnegative orthant.

3.1. Geometric interpretation

The answer lies in the geometric interpretation of the model $\boldsymbol{Y} = \boldsymbol{\Xi} \boldsymbol{M}_2$. Consider the i_2-th column of $\boldsymbol{Y}$, denoted as $\boldsymbol{y}_{i_2}$, we have the relation that

$$\boldsymbol{y}_{i_2} = \boldsymbol{\Xi} \boldsymbol{m}_{i_2} = \sum_{j=1}^{k} \boldsymbol{\xi}_j m_{j i_2}.$$

According to the assumption in MMSB, $\boldsymbol{m}_{i_2} \in \Delta$ where Δ is the probability simplex defined in (3). This means that $\boldsymbol{y}_{i_2}$ is a *convex combination* of $\boldsymbol{\xi}_1, ..., \boldsymbol{\xi}_k$ in $\mathbb{R}^{k-1}$, for all $i_2 \in \mathcal{S}_2$. Geometrically, this means all the $\boldsymbol{y}_{i_2}$'s belong to the *convex hull* of $\boldsymbol{\xi}_1, ..., \boldsymbol{\xi}_k$, denoted as $\mathrm{conv}(\boldsymbol{\xi}_1, ..., \boldsymbol{\xi}_k)$. If the set of vectors $\boldsymbol{\xi}_1 - \boldsymbol{\xi}_k, ..., \boldsymbol{\xi}_{k-1} - \boldsymbol{\xi}_k$ are linearly independent, then $\mathrm{conv}(\boldsymbol{\xi}_1, ..., \boldsymbol{\xi}_k)$ is called a *simplex*. Details of these convex geometry concepts can be found in Boyd & Vandenberghe (2004).

An example is given in Figure 2, where black dots represent columns of $\boldsymbol{Y}$, and red dots represent columns of $\boldsymbol{\Xi}$. As we can see, columns of $\boldsymbol{Y}$ clearly lie inside the simplex defined by the columns of $\boldsymbol{\Xi}$. However, the enclosing simplex is not unique—the one depicted by gray dashed lines is another simplex containing all the columns of $\boldsymbol{Y}$, and there are infinitely many more. We do know that $\boldsymbol{\Xi}$ is nonnegative, which means $\mathrm{conv}(\boldsymbol{\xi}_1, ..., \boldsymbol{\xi}_k)$ should lie in the nonnegative orthant, but that does not help pin down the correct $\boldsymbol{\Xi}$.

However, among all possible enclosing simplexes, intuitively the most plausible solution would be the one with *minimum volume*. For the simplex $\mathrm{conv}(\boldsymbol{\xi}_1, ..., \boldsymbol{\xi}_k)$ in $\mathbb{R}^{k-1}$, its volume is equal to (Strang, 2006)

$$\mathrm{Vol}(\boldsymbol{\Xi}) = \frac{1}{(k-1)!} \left| \det \begin{bmatrix} \boldsymbol{\xi}_1 - \boldsymbol{\xi}_k & \cdots & \boldsymbol{\xi}_{k-1} - \boldsymbol{\xi}_k \end{bmatrix} \right|.$$

Using this intuition, we propose the following formulation

for recovering $\boldsymbol{\Xi}$ and $\boldsymbol{M}_2$

$$\begin{aligned} \underset{\boldsymbol{\Xi},\boldsymbol{M}_2}{\text{minimize}} \quad & \left|\det\left[\begin{array}{ccc} \boldsymbol{\xi}_1-\boldsymbol{\xi}_k & \cdots & \boldsymbol{\xi}_{k-1}-\boldsymbol{\xi}_k \end{array}\right]\right| \\ \text{subject to} \quad & \boldsymbol{Y}=\boldsymbol{\Xi}\boldsymbol{M}_2, \boldsymbol{M}_2\geq 0, \mathbf{1}^\top\boldsymbol{M}_2=\mathbf{1}. \end{aligned} \tag{7}$$

Problem (7) can be further simplified as follows. Define

$$\widetilde{\boldsymbol{Y}}=\begin{bmatrix}\boldsymbol{Y}\\ \mathbf{1}^\top\end{bmatrix}, \quad \widetilde{\boldsymbol{\Xi}}=\begin{bmatrix}\boldsymbol{\Xi}\\ \mathbf{1}^\top\end{bmatrix},$$

then the two equality constraints can be combined into one

$$\widetilde{\boldsymbol{Y}}=\widetilde{\boldsymbol{\Xi}}\boldsymbol{M}_2.$$

Furthermore, it is easy to verify that

$$\det\left[\begin{array}{ccc} \boldsymbol{\xi}_1-\boldsymbol{\xi}_k & \cdots & \boldsymbol{\xi}_{k-1}-\boldsymbol{\xi}_k \end{array}\right]=\det\widetilde{\boldsymbol{\Xi}},$$

using the Schur complement. Finally, we end up with the formulation

$$\begin{aligned} \underset{\widetilde{\boldsymbol{\Xi}},\boldsymbol{M}_2}{\text{minimize}} \quad & |\det\widetilde{\boldsymbol{\Xi}}| \\ \text{subject to} \quad & \widetilde{\boldsymbol{Y}}=\widetilde{\boldsymbol{\Xi}}\boldsymbol{M}_2, \boldsymbol{M}_2\geq 0, \boldsymbol{e}_k^\top\widetilde{\boldsymbol{\Xi}}=\mathbf{1}^\top. \end{aligned} \tag{8}$$

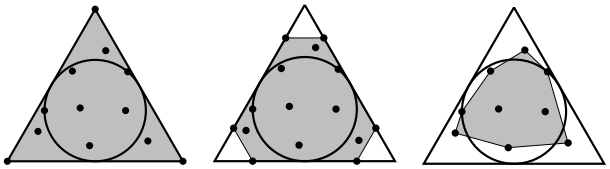
3.2. Identifiability

Problem (8) provides an intuitive identification criterion for recovering $\boldsymbol{\Xi}$ and $\boldsymbol{M}_2$, stemming from the geometric interpretation of the problem. In this subsection, we show that if the ground-truth membership matrix $\boldsymbol{M}_2$ satisfies a so-called *sufficiently scattered* condition, then optimally solving (8) guarantees unique recovery of $\boldsymbol{M}_2$ up to row permutations. Different permutations of the rows correspond to relabelling the communities, which is inconsequential in practice. From this point on, we denote the ground-truth matrices as $\boldsymbol{\Xi}^\natural$ and $\boldsymbol{M}_2^\natural$, and similarly for $\widetilde{\boldsymbol{\Xi}}^\natural$ since it is obtained by simply stacking $\boldsymbol{\Xi}^\natural$ with an all-one row vector $\mathbf{1}^\top$.

To study the identifiability of this model, we switch the space from $\mathbb{R}^{k-1}$ to $\mathbb{R}^k$, where the columns of $\boldsymbol{M}_2^\natural$ belong to. The columns of $\boldsymbol{M}_2^\natural$ have a one-to-one correpondance to the columns of $\boldsymbol{Y}$, and in turn the columns of $\boldsymbol{\Xi}^\natural$ correspond to the coordinate vectors $\boldsymbol{e}_1,...,\boldsymbol{e}_k$. If $\mathrm{conv}(\boldsymbol{\xi}_1,...,\boldsymbol{\xi}_k)$ is the smallest enclosing simplex for $\boldsymbol{Y}$, then so is Δ for $\boldsymbol{M}_2^\natural$, since $\Delta=\mathrm{conv}(\boldsymbol{e}_1,...,\boldsymbol{e}_k)$. Consequently, the columns of $\boldsymbol{M}_2^\natural$ should be "sufficiently scattered" in Δ, because otherwise we can further diminish the volume of the enclosing simplex. The formal definition of "sufficiently scattered" is given as follows.

Definition 1 (sufficiently scattered)**.** Let $\mathcal{D}$ be a "hyper-disc" on the hyperplane $\mathbf{1}^\top\boldsymbol{x}=1$ defined as

$$\mathcal{D}=\{\boldsymbol{x}\in\mathbb{R}^k : \|\boldsymbol{x}\|^2\leq\frac{1}{k-1}, \mathbf{1}^\top\boldsymbol{x}=1\}. \tag{9}$$



(a) Pure node (b) Sufficiently scattered (c) Not identifiable

Figure 3. A geometric illustration of the sufficiently scattered condition (middle), a special case that is separable / pure node (left), and a case that is not identifiable (right). The triangle denotes the probability simplex Δ, the circle denotes $\mathcal{D}$ defined in (9), and the shaded regions represent $\mathrm{conv}(\boldsymbol{M})$.

A matrix $\boldsymbol{M}$, with all its columns in Δ, is called **sufficiently scattered** if it satisfies that: (i) $\mathcal{D}\subseteq\mathrm{conv}(\boldsymbol{M})$, and (ii) $\mathrm{bd}\,\mathrm{conv}(\boldsymbol{M})\cap\mathrm{bd}\mathcal{D}=\{(1/k)(\mathbf{1}-\boldsymbol{e}_j) : j=1,...,k\}$, where bd denotes the boundary of a set.

The sufficiently scattered condition first appeared in (Huang et al., 2014) to establish uniqueness guarantees for the widely used nonnegative matrix factorization model. Fu et al. (2015) and Lin et al. (2015) simultaneously showed that under the same condition, the latent representation of $\boldsymbol{Y}=\boldsymbol{\Xi}\boldsymbol{M}$, where the columns of $\boldsymbol{M}$ are in Δ and $\boldsymbol{\Xi}$ is square and non-singular, can be uniquely identified by optimizing the volume criterion. The condition was first named "sufficiently scattered" by Huang et al. (2016) for yet another subtly different matrix tri-factorization model, which has applications in topic modeling (Huang et al., 2016) and hidden Markov model (HMM) identification (Huang et al., 2018). See Fu et al. (2019) for a recent survey on a family of related models with identifiability guarantees.

One may notice that in almost all of these aforementioned prior works, the condition is described using *conic hulls* and *dual cones*. In the supplementary material, we show that the way we present the condition is equivalent to the ones given before *when everything is on the hyperplane* $\mathbf{1}^\top\boldsymbol{x}=1$. We find the way we present the condition easier for the readers to understand, although the concepts of cones and dual cones are crucial in proving identifiability. In the context of our model, we have the following identifiability result.

Theorem 1. *(Fu et al., 2015; Lin et al., 2015) Suppose* $\boldsymbol{Y}=\boldsymbol{\Xi}^\natural\boldsymbol{M}_2^\natural$, *where* $\mathrm{rank}(\widetilde{\boldsymbol{\Xi}}^\natural)=k$ *and* $\boldsymbol{M}_2^\natural$ *is* **sufficiently scattered**. *Let* $(\boldsymbol{M}_\star,\boldsymbol{\Xi}_\star)$ *be an optimal solution for* (8)*, then there exists a permutation matrix* $\boldsymbol{\Pi}\in\mathbb{R}^{k\times k}$ *such that*

$$\boldsymbol{M}_2^\natural=\boldsymbol{\Pi}\boldsymbol{M}_\star, \qquad \widetilde{\boldsymbol{\Xi}}^\natural=\boldsymbol{\Xi}_\star\boldsymbol{\Pi}^\top.$$

Although this theorem has been independently proven by Fu et al. (2015) and Lin et al. (2015), we provide a complete proof in the supplementary material nonetheless. One of the reasons is because our formulated problem (8), while

mathematically equivalent, is not exactly the same as the ones written by Fu et al. (2015) and Lin et al. (2015). In the next section of algorithm design for solving (8), we will see that the way we write the problem leads to a more efficient algorithmic implementation with better convergence properties, which is not achieved by any prior work.

A geometric illustration of a matrix that satisfies the sufficiently scattered condition is shown in Figure 3b, where columns of the matrix are depicted as blue dots. As we can see, $\mathcal{D}$ is a subset of Δ, but touches the boundary of Δ at points $(1/k)(\mathbf{1}-\boldsymbol{e}_j)$, $j = 1, ..., k$. If a matrix $\boldsymbol{M}$ is sufficiently scattered, $\mathrm{conv}(\boldsymbol{M})$ contains $\mathcal{D}$ as a subset and, as a second requirement, $\mathcal{D}$ touches the boundary of $\mathrm{conv}(\boldsymbol{M})$ only at those points too.

One can also see from Figure 3a that the pure node assumption, considered in (Zhang et al., 2014; Kaufmann et al., 2016; Mao et al., 2017; Panov et al., 2017), is a very special case of sufficiently scattered. It requires that all the coordinate vectors are included in columns of $\boldsymbol{M}_2^\natural$, which makes $\mathrm{conv}(\boldsymbol{M}_2^\natural) = \Delta$, while our result shows that identifiability can be achieved even when this condition is violated. In fact, it has been empirically observed that a nonnegative sparse matrix satisfies the sufficiently scattered condition with very high probability (Huang et al., 2015; Fu et al., 2019).

4. Algorithm

In this section, we propose an algorithm for approximately solving Problem (8), which is non-convex and has been shown to be NP-hard (Packer, 2002). Nevertheless, we propose an algorithm after carefully reformulate Problem (8) and apply the general idea of successive upperbound minimization (BSUM) (Razaviyayn et al., 2013), which is guaranteed to converge to a stationary point. The proposed algorithm is called Community Detection via Minimum Volume Simplex Identification (CD-MVSI).

Inspired by the work of (Chan et al., 2009) and (Huang et al., 2016), we introduce a new variable that relates to $\widetilde{\boldsymbol{\Xi}}^{-1}$ to replace $\widetilde{\boldsymbol{\Xi}}$ in (8). However, before we do that, we note that in Problem (8), the first equality constraint decouples over the rows of $\widetilde{\boldsymbol{\Xi}}$, while the second equality constraint decouples over the columns of $\widetilde{\boldsymbol{\Xi}}$. This may lead to a difficulty if we adopt a cyclic column/row update scheme. Our tactic is to exploit the fact that the matrix product $\widetilde{\boldsymbol{Y}} = \widetilde{\boldsymbol{\Xi}}\boldsymbol{M}_2$ is not affected if we insert a diagonal matrix and its inverse in-between,

$$\widetilde{\boldsymbol{Y}} = \widetilde{\boldsymbol{\Xi}}\boldsymbol{M}_2 = \widetilde{\boldsymbol{\Xi}}\boldsymbol{D}\boldsymbol{D}^{-1}\boldsymbol{M}_2.$$

We choose $\boldsymbol{D}$ to be $\mathrm{Diag}(\boldsymbol{M}_2\mathbf{1})$, so that $\boldsymbol{D}^{-1}\boldsymbol{M}_2\mathbf{1} = \mathbf{1}$. Now we formally introduce $\boldsymbol{X} = (\widetilde{\boldsymbol{\Xi}}\boldsymbol{D})^{-1}$, and the constraints in (8) becomes

$$\boldsymbol{X}\widetilde{\boldsymbol{Y}} = \boldsymbol{D}^{-1}\boldsymbol{M}_2 \geq 0, \quad \boldsymbol{X}\widetilde{\boldsymbol{Y}}\mathbf{1} = \boldsymbol{D}^{-1}\boldsymbol{M}_2\mathbf{1} = \mathbf{1}.$$

Now we can eliminate variable $\boldsymbol{M}_2$. The loss function becomes $|\det \boldsymbol{D}\boldsymbol{X}|^{-1} = |\det \boldsymbol{D}|^{-1}|\det \boldsymbol{X}|^{-1}$, and we recognize that $|\det \boldsymbol{D}|^{-1}$, albeit unknown, is a positive scalar, which does not affect the optimization problem. Finally, we take the reciprocal square of the loss function, resulting in the following problem formulation

$$\begin{aligned} \underset{\boldsymbol{X}}{\text{maximize}} \quad & (\det \boldsymbol{X})^2 \\ \text{subject to} \quad & \boldsymbol{X}\widetilde{\boldsymbol{Y}} \geq 0, \quad \boldsymbol{X}\widetilde{\boldsymbol{Y}}\mathbf{1} = \mathbf{1}. \end{aligned} \tag{10}$$

After solving (10), we let $\boldsymbol{D} = \mathrm{Diag}(\boldsymbol{e}_k^\top \boldsymbol{X}_\star^{-1})$ recover a solution for $\boldsymbol{M}_2$ as

$$\boldsymbol{M}_{2\star} = \boldsymbol{D}\boldsymbol{X}_\star\widetilde{\boldsymbol{Y}}.$$

Problem (10) now has a convex constraint set that decouples over the rows of $\boldsymbol{X}$, although the objective function is still not concave. We propose to adopt the block successive upperbound minimization (BSUM) framework (Razaviyayn et al., 2013) and update the rows of $\boldsymbol{X}$ in a cyclic fashion. According to Laplace's formula, we know that $\det \boldsymbol{X}$ is a linear function with respect to the ℓ-th row of $\boldsymbol{X}$ using the co-factor expansion

$$\det \boldsymbol{X} = \sum_{m=1}^{k} (-1)^{\ell+m} x_{\ell m} \det \boldsymbol{X}_{\ell m},$$

where $\boldsymbol{X}_{\ell m}$ is obtained by deleting the ℓ-th row and m-th column of $\boldsymbol{X}$. For a particular $\boldsymbol{X}$, we define a vector $\boldsymbol{f} \in \mathbb{R}^k$ with the m-th entry equals to

$$f_m = \det \boldsymbol{X} \times (-1)^{\ell+m} \det \boldsymbol{X}_{\ell m}, \tag{11}$$

then we have $\boldsymbol{f}^\top \boldsymbol{z} \leq (\det \boldsymbol{Z})^2$ for all $\boldsymbol{z} \in \mathbb{R}^k$, where $\boldsymbol{Z}$ is obtained by replacing the ℓ-th row of $\boldsymbol{X}$ with $\boldsymbol{z}^\top$, and equality holds if $\boldsymbol{z} = \boldsymbol{x}_\ell$. Therefore, by successively solving

$$\begin{aligned} \underset{\boldsymbol{z}}{\text{maximize}} \quad & \boldsymbol{f}^\top \boldsymbol{z} \\ \text{subject to} \quad & \boldsymbol{z}^\top \widetilde{\boldsymbol{Y}} \geq 0, \quad \boldsymbol{z}^\top \widetilde{\boldsymbol{Y}}\mathbf{1} = 1, \end{aligned} \tag{12}$$

the cyclic column update scheme is guaranteed to monotonically increase the objective value of (10) if it is initialized with a feasible point. Since the constraints decouple over the rows, any random point, after one cycle of row updates, becomes feasible. Problem (12) is a linear program with only k variables and n inequality constraints, and there exist many reliable solvers to solve it efficiently. If the interior-point method is used, the worst case complexity is $\mathcal{O}(k^2 n^{1.5})$. Recall that n is the size of the subgroup of nodes that we are interested in detecting their communities, so it is not necessarily a large number even for a huge network.

The proposed CD-MVSI algorithm, starting from the moment generating step described in Section 2, is summarized

Algorithm 1 CD-MVSI: Community Detection via Minimum Volume Simplex Identification

Input: network adjacency matrix $\boldsymbol{A}$, number of communities k, index set $\mathcal{G}$ of a group of interested nodes

randomly pick a set of $k-1$ nodes $\mathcal{S}$
put the rest of nodes in $\mathcal{R}$
$\boldsymbol{Y} = \boldsymbol{A}(\mathcal{R},\mathcal{S})^\top \boldsymbol{A}(\mathcal{R},\mathcal{G})$
$\widetilde{\boldsymbol{Y}} = \begin{bmatrix} \boldsymbol{Y} \\ \boldsymbol{1}^\top \end{bmatrix}, \quad \boldsymbol{b} = \widetilde{\boldsymbol{Y}}\boldsymbol{1}$
randomly initialize $\boldsymbol{X} \in \mathbb{R}^{k\times k}$
repeat
  for $\ell = 1, ..., k$ **do**
    $\boldsymbol{f} \leftarrow \boldsymbol{X}^{-1}\boldsymbol{e}_\ell$
    $\boldsymbol{z} \leftarrow \arg\max_{\boldsymbol{z}} \boldsymbol{f}^\top \boldsymbol{z} \quad \text{s.t. } \boldsymbol{z}^\top \widetilde{\boldsymbol{Y}} \geq 0, \boldsymbol{z}^\top \boldsymbol{b} = 1$
    replace ℓ-th row of $\boldsymbol{X}$ with $\boldsymbol{z}^\top$
  end for
until convergence
$\boldsymbol{D} = \mathrm{Diag}(\boldsymbol{e}_k^\top \boldsymbol{X}^{-1})$
return $\boldsymbol{D}\boldsymbol{X}\widetilde{\boldsymbol{Y}}$

in Algorithm 1. The core algorithm shows some resemblance to the AnchorFree algorithm by Huang et al. (2016) for learning topic models, but we note the following differences:

1. Fundamentally, CD-MVSI and AnchorFree solve different problems. In AnchorFree, the equality constraint similar to the one in (10) is inherent, whereas ours comes from careful reformulation exploiting the algebraic structure of the problem.
2. CD-MVSI replaces the objective $|\det \boldsymbol{X}|$ with $(\det \boldsymbol{X})^2$, which is mathematically equivalent except that the objective is now smooth. The immediate benefit is that CD-MVSI only solves one LP per row update, whereas AnchorFree solves two, which means CD-MVSI is at least twice as fast. In the sequel we also show that it leads to provable convergence to a stationary point.
3. An interesting numerical trick we use in CD-MVSI is that the vector $\boldsymbol{f}$, defined in (11), is actually $(\det \boldsymbol{X})^2$ times the ℓ-th column of $\boldsymbol{X}^{-1}$ per Cramer's rule. It is well-known that directly calculating the determinant suffers from serious round-off errors. Our simulations show that computing $\boldsymbol{X}^{-1}$ significantly helps stablize the numerical performance.

We end this section with the following two convergence results, which is not known prior to our work. The proofs are relegated to the supplementary material.

Theorem 2. *Assume each LP sub-problem (12) has a unique solution, then CD-MVSI converges to a stationary point of Problem (10).*

Theorem 3. *Assume the ground truth $\boldsymbol{M}_2^\natural$ satisfies the separability condition, i.e., there exists a pure node for every community, and $\widetilde{\boldsymbol{\Xi}}^\natural$ is non-singular, then CD-MVSI recovers the true $\boldsymbol{M}_2^\natural$, up to row permutation, in one iteration.*

The caviar of Theorem 3 is that CD-MVSI also has computational guarantees when the more restrictive pure-node assumption is satisfied, similar to the work of (Zhang et al., 2014; Kaufmann et al., 2016; Mao et al., 2017; Panov et al., 2017). If the pure-node assumption does not hold, however, CD-MVSI can still correctly recover the underlying community structure under the more relaxed sufficiently scattered condition, which cannot be achieved by any other method.

5. Experiments

In this section, we provide some numerical experiments to showcase the effectiveness of CD-MVSI for learning community mixed-memberships from a large network. We apply a number of community detection methods to real-world co-authorship data sets obtained from Microsoft Academic Network and DBLP. In the supplementary material, we also validate the identifiability performance on synthetic data, which clearly shows that under more general scenarios, for example, in the absense of pure nodes or when the membership coefficients do not follow a Dirichlet distribution, CD-MVSI manages to perfectly identify the underlying community structure whereas other baseline methods are not able to.

Baseline methods. Mao et al. (2017) and Panov et al. (2017) demonstrated that their pure-node based methods, GeoNMF and SPOC, out-perform most of the other algorithms, including stochastic variational inference (Gopalan et al., 2012) designed for scaling up the original MMSB model. We therefore mainly compare with GeoNMF and SPOC in this section. We also construct the third-order graph moment proposed by Anandkumar et al. (2014a), and the subsequent canonical polyadic decomposition is executed using Tensorlab (Vervliet et al., 2016). All the experiments are conducted in MATLAB on an iMac Pro, and we use the built-in `linprog` function in MATLAB to solve each of the LP sub-problems.

Data sets. We consider the co-authorship data from Microsoft Academic Graph (MAG) and DBLP constructed by Mao et al. (2017). [1] Each network is provided with "ground-truth" community memberships of the nodes: In MAG, each paper is tagged with a "field of study" label (community), and the membership of a node (author) is the number of papers with certain tags normalized by the total number of papers. In DBLP, communities are defined by venues, and memberships are obtained by counting the number of papers published in specific venues. We refer the readers to Mao et al. (2017) for details.

[1] Downloaded from `http://www.cs.utexas.edu/~xmao/coauthorship`

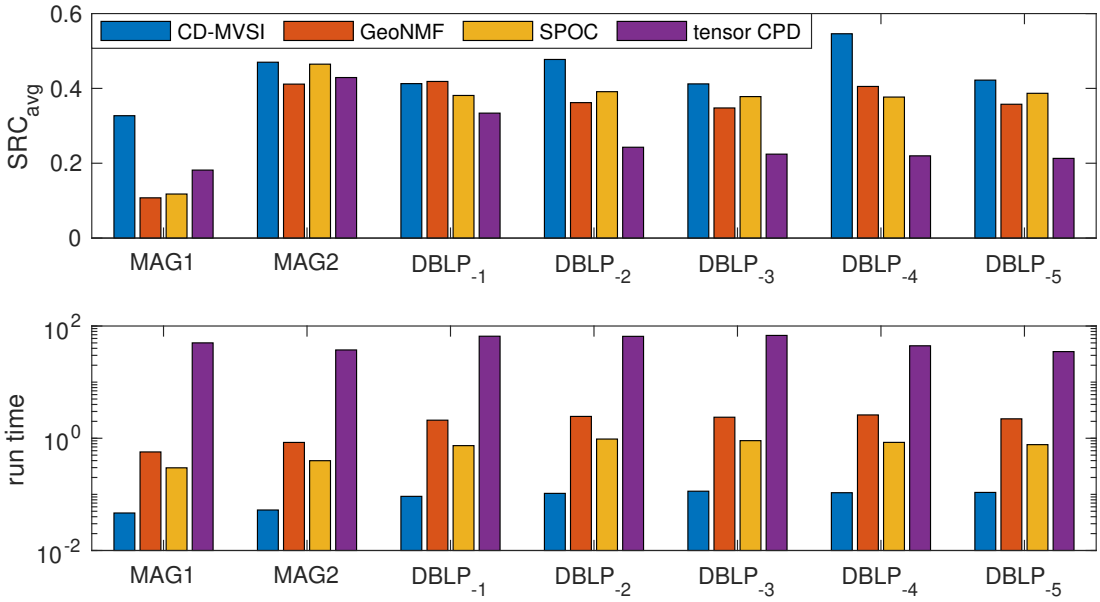


Figure 4. Averaged Spearman rank correlation coefficient (SRC_{avg}) and run time performance on co-authorship networks.

Mao et al. (2017) divided the networks into smaller ones in each experiment, so that each network contains only 3–6 communities. To make the experiment more challenging, we combine the smaller DBLP networks (with 10,000 to 30,000 nodes each) into bigger ones to work with. Specifically, DBLP_{-i} means we combine all the sub-networks except the i-th one. In this case each network contains 13–16 communities.

Evaluation metric. We evaluate the performance by calculating the averaged Spearman's rank correlation coefficient (SRC_{avg}) between the learned community memberships $\widehat{\boldsymbol{M}}$ and the ground-truth $\boldsymbol{M}^\natural$ provided by the data set, after fixing the permutation ambiguity of the communities using the Hungarian algorithm. This evaluation metric is the same as the one used in Mao et al. (2017) and Panov et al. (2017). The SRC_{avg} takes values between -1 and 1, and gives a larger number if the ranking of the elements in two vectors are similar, which fits well in our context. A larger SRC_{avg} implies better performance.

Performance. The performance in terms of SRC_{avg} and run time is shown in Figure 4. Because CD-MVSI and the tensor CPD method work with subsets of nodes, the experiment on each data set is the average of 10 random trials; each time a subset of 1000 nodes are randomly chosen as the interested group, on which the SRC_{avg} is calculated. For the tensor method, the Dirichlet 'concentration' parameter α_0 is set to be k; i.e., we assume $\boldsymbol{m}_i$ follows a uniform distribution in the probability simplex Δ.

As we can see, CD-MVSI consistently performs better than the other baseline methods in terms of SRC_{avg}, sometimes significantly better. Somewhat surprisingly, the run time performance is at least 10 times faster than GeoNMF and SPOC (notice the log-scale on the vertical axis). One main reason could be that GeoNMF and SPOC both require calculating the k-largest eigenvalues and their corresponding eigenvectors of the graph adjacency matrix, whereas CD-MVSI constructs a graph moment using basic matrix operations. The inferior performance of the tensor method is, in our opinion, partly due to the inaccurate estimate of the Dirichlet parameter α_0. Unfortunately, there is no good way of estimating that hyper-parameter, to the best of our knowledge.

6. Conclusion

In this paper we aimed to design a learning framework that is guaranteed to recover the underlying community memberships of the popular mixed-membership stochastic blockmodel (MMSB) for large networks. Our method started by constructing a second-order graph moment in order to overcome the binary nature of the data, by simply counting the number of 2-star sub-graphs. The resulting moment was shown to admit a very intuitive geometric interpretation, which led to our proposed problem formulation. We showed that if the membership matrix satisfies the sufficiently scattered condition, solving the proposed problem is guaranteed to recover the ground-truth. An efficient algorithm called CD-MVSI was then designed, which has robust convergence guarantees. Experiments on real-world co-authorship networks showcased the validity of our method.

Acknowledgments

This work is supported in part by the National Science Foundation (NSF) under grants ECCS 1608961, ECCS 1808159, and the Army Research Office (ARO) under grant W911NF-19-1-0247.

References

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

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

Anandkumar, A., Ge, R., Hsu, D., Kakade, S. M., and Telgarsky, M. Tensor decompositions for learning latent variable models. *Journal of Machine Learning Research*, 15(1):2773–2832, 2014b.

Araújo, M. C. U., Saldanha, T. C. B., Galvao, R. K. H., Yoneyama, T., Chame, H. C., and Visani, V. The successive projections algorithm for variable selection in spectroscopic multicomponent analysis. *Chemometrics and Intelligent Laboratory Systems*, 57(2):65–73, 2001.

Arora, S., Ge, R., and Moitra, A. Learning topic models—going beyond SVD. In *IEEE Symposium on Foundations of Computer Science (FOCS)*, pp. 1–10. IEEE, 2012.

Arora, S., Ge, R., Halpern, Y., Mimno, D., Moitra, A., Sontag, D., Wu, Y., and Zhu, M. A practical algorithm for topic modeling with provable guarantees. In *International Conference on Machine Learning*, pp. 280–288, 2013.

Bickel, P. J., Chen, A., and Levina, E. The method of moments and degree distributions for network models. *The Annals of Statistics*, 39(5):2280–2301, 2011.

Blei, D. and Lafferty, J. Correlated topic models. *Advances in Neural Information Processing Systems*, 18:147, 2006.

Boyd, S. and Vandenberghe, L. *Convex optimization*. Cambridge University Press, 2004.

Chan, T.-H., Ma, W.-K., Chi, C.-Y., and Wang, Y. A convex analysis framework for blind separation of non-negative sources. *IEEE Transactions on Signal Processing*, 56 (10):5120–5134, 2008.

Chan, T.-H., Chi, C.-Y., Huang, Y.-M., and Ma, W.-K. A convex analysis-based minimum-volume enclosing simplex algorithm for hyperspectral unmixing. *IEEE Transactions on Signal Processing*, 57(11):4418–4432, 2009.

Donoho, D. and Stodden, V. When does non-negative matrix factorization give a correct decomposition into parts? In *Advances in Neural Information Processing Systems*, pp. 1141–1148, 2004.

Fu, X., Ma, W.-K., Huang, K., and Sidiropoulos, N. D. Blind separation of quasi-stationary sources: Exploiting convex geometry in covariance domain. *IEEE Transactions on Signal Processing*, 63(9), 2015.

Fu, X., Huang, K., Sidiropoulos, N. D., and Ma, W.-K. Nonnegative matrix factorization for signal and data analytics: Identifiability, algorithms, and applications. *IEEE Signal Processing Magazine*, 36:59–80, 2019.

Goldenberg, A., Zheng, A. X., Fienberg, S. E., Airoldi, E. M., et al. A survey of statistical network models. *Foundations and Trends® in Machine Learning*, 2(2): 129–233, 2010.

Gopalan, P. K., Gerrish, S., Freedman, M., Blei, D. M., and Mimno, D. M. Scalable inference of overlapping communities. In *Advances in Neural Information Processing Systems*, pp. 2249–2257, 2012.

Hanneke, S. and Xing, E. P. Discrete temporal models of social networks. In *Statistical Network Analysis: Models, Issues, and New Directions*, pp. 115–125. Springer, 2007.

Huang, K., Sidiropoulos, N. D., and Swami, A. Nonnegative matrix factorization revisited: Uniqueness and algorithm for symmetric decomposition. *IEEE Transactions on Signal Processing*, 62(1):211–224, 2014.

Huang, K., Sidiropoulos, N. D., Papalexakis, E. E., Faloutsos, C., Talukdar, P. P., and Mitchell, T. M. Principled neuro-functional connectivity discovery. In *SIAM International Conference on Data Mining*, pp. 631–639. SIAM, 2015.

Huang, K., Fu, X., and Sidiropoulos, N. D. Anchor-free correlated topic modeling: Identifiability and algorithm. In *Advances in Neural Information Processing Systems*, pp. 1786–1794, 2016.

Huang, K., Fu, X., and Sidiropoulos, N. Learning hidden Markov models from pairwise co-occurrences with application to topic modeling. In *International Conference on Machine Learning*, pp. 2068–2077. PMLR, 2018.

Jin, J., Ke, Z. T., and Luo, S. Estimating network memberships by simplex vertex hunting. *arXiv preprint arXiv:1708.07852*, 2017.

Kaufmann, E., Bonald, T., and Lelarge, M. A spectral algorithm with additive clustering for the recovery of overlapping communities in networks. In *International Conference on Algorithmic Learning Theory*, pp. 355–370. Springer, 2016.

Lin, C.-H., Ma, W.-K., Li, W.-C., Chi, C.-Y., and Ambikapathi, A. Identifiability of the simplex volume minimization criterion for blind hyperspectral unmixing: The no-pure-pixel case. *IEEE Transactions on Geoscience and Remote Sensing*, 53(10):5530–5546, 2015.

Ma, W.-K., Bioucas-Dias, J. M., Chan, T.-H., Gillis, N., Gader, P., Plaza, A. J., Ambikapathi, A., and Chi, C.-Y. A signal processing perspective on hyperspectral unmixing: Insights from remote sensing. *IEEE Signal Processing Magazine*, 31(1):67–81, 2014.

Mao, X., Sarkar, P., and Chakrabarti, D. On mixed memberships and symmetric nonnegative matrix factorizations. In *International Conference on Machine Learning*, pp. 2324–2333, 2017.

Mao, X., Sarkar, P., and Chakrabarti, D. Overlapping clustering models, and one (class) svm to bind them all. In *Advances in Neural Information Processing Systems*, pp. 2126–2136, 2018.

Meila, M. and Shi, J. Learning segmentation by random walks. In *Advances in Neural Information Processing Systems*, pp. 873–879, 2001.

Nascimento, J. M. and Dias, J. M. Vertex component analysis: A fast algorithm to unmix hyperspectral data. *IEEE transactions on Geoscience and Remote Sensing*, 43(4): 898–910, 2005.

Ng, A. Y., Jordan, M. I., and Weiss, Y. On spectral clustering: Analysis and an algorithm. In *Advances in Neural Information Processing Systems*, pp. 849–856, 2002.

Nowicki, K. and Snijders, T. A. B. Estimation and prediction for stochastic blockstructures. *Journal of the American Statistical Association*, 96(455):1077–1087, 2001.

Packer, A. NP-hardness of largest contained and smallest containing simplices for V-and H-polytopes. *Discrete and Computational Geometry*, 28(3):349–377, 2002.

Panov, M., Slavnov, K., and Ushakov, R. Consistent estimation of mixed memberships with successive projections. In *International Workshop on Complex Networks and their Applications*, pp. 53–64. Springer, 2017.

Razaviyayn, M., Hong, M., and Luo, Z.-Q. A unified convergence analysis of block successive minimization methods for nonsmooth optimization. *SIAM Journal on Optimization*, 23(2):1126–1153, 2013.

Sidiropoulos, N. D., De Lathauwer, L., Fu, X., Huang, K., Papalexakis, E. E., and Faloutsos, C. Tensor decomposition for signal processing and machine learning. *IEEE Transactions on Signal Processing*, 65(13):3551–3582, 2017.

Snijders, T. A. and Nowicki, K. Estimation and prediction for stochastic blockmodels for graphs with latent block structure. *Journal of Classification*, 14(1):75–100, 1997.

Strang, G. *Linear Algebra and Its Applications*. Thompson, 2006.

Vervliet, N., Debals, O., Sorber, L., Van Barel, M., and De Lathauwer, L. Tensorlab 3.0, Mar. 2016. URL https://www.tensorlab.net. Available online.

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

Zhang, Y., Levina, E., and Zhu, J. Detecting overlapping communities in networks using spectral methods. *arXiv preprint arXiv:1412.3432*, 2014.