Sharp Impossibility Results for Hyper-graph Global Testing

Anonymous Author(s)
Affiliation
Address
email

Abstract

In a broad Degree-Corrected Mixed-Membership (DCMM) setting, we test whether a hyper-graph has only one community or has multiple communities. Since both the null and alternative hypotheses have many unknown parameters, the challenge is, given an alternative, how to identify the null that is *hardest* to separate from the alternative. We approach this by proposing a *degree matching strategy* where we leverage a tensor-scaling technique. We present a result on standard minimax lower bound theory and a result on *Region of Possibility*, which is comparably more informative than the former. We also show these results are sharp and study a degree-based χ^2-test.

1 Introduction

While social network is a popular approach to modeling social relationships, some social relationships are better modelled by hypergraphs. For example, suppose we have a set of published papers in a scientific field (e.g., statistics), a co-authorship network (where each author is a node, and two nodes have an edge if and only if they have co-authored some papers) is not appropriate for analyzing the co-authorship topology. It is preferable to use a hypergraph, where each author is a node, and a subset of authors have a hyperedge if and only if they have co-authored a paper (e.g., [5, 10]). A hypergraph consists of a set of n nodes and a set of hyperedges, where a hyperedge consists of 2 or more nodes. We say a hypergraph is an m-uniform hypergraph if all hyperedges have exactly m nodes. A network can be viewed as a 2-uniform hypergraph. The coauthorship hypergraph above is a non-uniform hypergraph, which can be viewed as a union of m-uniform hypergraphs for different values of m.

We are interested in the problem of *global testing*, where we test whether the hypergraph has one community or multiple communities (communities in a hypergraph are clusters of nodes that have more hyperedges within than across). A good understanding of the problem is useful for measuring co-authorship and citation diversity [11] and for discovering non-obvious social groups and patterns [4]. It is also useful for understanding several related problems such as community detection and change-point detection in dynamic hypergraphs [13]. For simplicity, we first focus on 3-uniform hypergraphs and then extend the results to higher-order hypergraphs and non-uniform hypergraphs. Let $\mathcal{A}$ be the adjacency tensor of a uniform and symmetric 3-hypergraph with n nodes, where

$$\mathcal{A}_{i_1 i_2 i_3} = \begin{cases} 1, & \text{if } i_1, i_2, i_3 \text{ shares a hyperedge,} \\ 0, & \text{otherwise,} \end{cases} \qquad \text{for } 1 \le i_1, i_2, i_3 (\text{distinct}) \le n. \tag{1.1}$$

Since the tensor is symmetric, $\mathcal{A}(i_1, i_2, i_3) = \mathcal{A}(j_1, j_2, j_3)$ for two sets of indices $\{i_1, i_2, i_3\}$ and $\{j_1, j_2, j_3\}$ if one is a permutation of the other. We do not consider self hyperedges, hence, $\mathcal{A}_{i_1 i_2 i_3} = 0$ whenever i_1, i_2, i_3 are non-distinct. The notations below are frequently used in tensor analysis.

Definition 1.1. *(matricization, slicing, and slice aggregation). Let $\mathcal{A}$ be a 3-symmetric tensor of n-dimension. First, we call the $n \times n^2$ matrix A the matricization of $\mathcal{A}$, defined by $A_{i,j+n(k-1)} = \mathcal{A}_{ijk}$,*

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

$1 \leq i, j, k \leq n$. Second, for $1 \leq k \leq n$, we use $\mathcal{A}_{::k}$ to denote the $n \times n$ matrix whose row-i-and-column-j is $\mathcal{A}_{ijk}$, $1 \leq i, j \leq n$, and call it the k-th slice of $\mathcal{A}$. Last, let x be a $n \times 1$ vector, we use $(\mathcal{A}x)$ to denote the matrix $\sum_{k=1}^{n} x_k \mathcal{A}_{::k}$, which is an aggregation of the slices.

Real world hypergraphs have several noteworthy features. First, there may be severe degree heterogeneity (i.e., the degree of one node is many times higher than that of the other). Second, the overall sparsity levels may vary significantly from one hypergraph to another. Last, a node may have mixed-memberships across multiple communities To accommodate these features, we adopt the *degree-corrected mixed-membership* (tensor-DCMM) model. Suppose the tensor has K perceivable communities. Let $\mathcal{P}$ be a symmetric (3-uniform) tensor of K-dimension that models the community structure, let $\theta = (\theta_1, \theta_2, \ldots, \theta_n)'$ be positive parameters that model the degree heterogeneity of nodes, and $\pi_1, \pi_2, \ldots, \pi_n$ be K-dimensional membership vectors where $\pi_i(k) =$ the weight node i puts on community k, $1 \leq k \leq K$. We assume for all $1 \leq i_1, i_2, i_3 \leq n$ that are three distinct indicies, $\mathcal{A}_{i_1 i_2 i_3}$ are independent Bernoulli random variables where $\mathbb{P}(\mathcal{A}_{i_1 i_2 i_3} = 1) = \theta_{i_1}\theta_{i_2}\theta_{i_3} \sum_{k_1,k_2,k_3=1}^{K} \mathcal{P}_{k_1,k_2,k_3}\pi_{i_1}(k_1)\pi_{i_2}(k_2)\pi_{i_3}(k_3)$, where by Definition 1.1, the right hand side equals to $\theta_{i_1}\theta_{i_2}\theta_{i_3}\pi'_{i_1}(\mathcal{P}\pi_{i_3})\pi_{i_2}$. Introduce a non-stochastic 3-uniform tensor $\mathcal{Q}$ of n-dimension where $\mathcal{Q}_{i_1 i_2 i_3} = \theta_{i_1}\theta_{i_2}\theta_{i_3}\pi'_{i_1}(\mathcal{P}\pi_{i_3})\pi_{i_2}$ for $1 \leq i_1, i_2, i_3 \leq n$. Let $\mathrm{diag}(\mathcal{Q})$ be the tensor with the same size of $\mathcal{Q}$ where $(\mathrm{diag}(\mathcal{Q}))_{i_1 i_2 i_3} = \mathcal{Q}_{i_1 i_2 i_3}$ if i_1, i_2, i_3 are non-distinct, and $(\mathrm{diag}(\mathcal{Q}))_{i_1 i_2 i_3} = 0$ otherwise. It follows that

$$\mathbb{E}(\mathcal{A}) = \mathcal{Q} - \mathrm{diag}(\mathcal{Q}), \qquad \text{where} \quad \mathcal{Q}_{i_1 i_2 i_3} = \theta_{i_1}\theta_{i_2}\theta_{i_3}\pi'_{i_1}(\mathcal{P}\pi_{i_3})\pi_{i_2}. \tag{1.2}$$

For identifiability, let $P \in \mathbb{R}^{K,K^2}$ be the matricization of $\mathcal{P}$ (see Definition 1.1). We assume

$$\mathrm{rank}(P) = K, \quad \text{and} \quad \mathcal{P}_{iii} = 1 \text{ for all } 1 \leq i \leq K. \tag{1.3}$$

Definition 1.2. *We call model (1.1)-(1.3) the tensor-DCMM model. We call $\mathcal{Q}$ and $\mathcal{P}$ the Bernoulli probability tensor and community structure tensor for DCMM, respectively.*

In tensor-DCMM, if we require all π_i to be degenerate (i.e., one entry is 1, all other entries are 0), then tensor-DCMM reduces to the *Degree-Corrected Block Model (tensor-DCBM)* [13]. If we further require $\theta_1 = \ldots = \theta_n$ (but the second condition in (1.3) can be removed), then tensor-DCBM further reduces to the *Stochastic Block Model* (tensor-SBM) [8]. For simplicity, we may drop "tensor" in these terms if there is no confusion. The global testing problem above is then to test

$$H_0 : K = 1 \qquad \text{vs.} \qquad H_1 : K > 1. \tag{1.4}$$

Our primary interest is to find a sharp information lower bound for the global testing problem. To facilitate the lower bound study, we frequently adopt a *Random Mixed-Membership (RMM)* model. Introduce a subset $V_0 = \{x \in \mathbb{R}^K, x_k \geq 0, \sum_{k=1}^{K} x_k = 1\}$, and let F be a K-variate distribution with support contained in V_0. We assume

$$\pi_i \overset{iid}{\sim} F; \qquad (\text{let } h = \mathbb{E}_F[\pi_i]). \tag{1.5}$$

Moreover, let $V_0^* = \{e_1, e_2, \ldots, e_K\} \subset V_0$, where e_k is the k-th basis vector of $\mathbb{R}^K$. Similarly, RMM-DCMM reduces to RMM-DCBM if we require $\mathrm{supp}(F) = V_0^*$, and reduces to RMM-SBM if we further require $\theta_1 = \theta_2 = \ldots = \theta_n$ (but the second condition in (1.3) can be removed). Let $\theta = (\theta_1, \theta_2, \ldots, \theta_n)'$. We allow $(\theta, \mathcal{P}, h, F)$ to vary with n so to cover a wide variety of settings allowing severe degree heterogeneity, mixed-memberships, flexible sparsity levels, and weak signals.

Example 1. The following 2-parameter SBM has been used in some recent papers (e.g., [1, 18]), where we assume for two parameters α_n and ρ_n that may depend on n (but not on i, j, k), (i) no degree heterogeneity ($\theta_1 = \ldots = \theta_n = \alpha_n$), (ii) no mixed membership (all π_i are degenerate), and (iii) $\mathcal{P}_{ijk} = 1$ if $i = j = k$ and $\mathcal{P}_{ijk} = \rho_n$ otherwise. Also, Lin et al. [16] studied a 3-parameter SBM, which is same as above except for that in (iii), we assume for three parameters α_n and ρ_n, and τ_n, $\mathcal{P}_{ijk}$ equals to 1, ρ_n, or τ_n if i, j, k take 1, 2, or 3 distinct values, respectively. Compared to DCMM, these models are much narrower: they do not allow severe degree heterogeneity and mixed-memberships, and $\mathcal{P}$ only assumes 2 or 3 different values.

A sharp lower bound for tensor global testing may help us understand the fundamental limits of many problems (e.g., community detection, dynamic hypergraphs), but how to derive such a bound is a challenging problem. A reasonable approach is to borrow ideas from existing works on the simpler

setting of network global testing (e.g., [17, 2, 12]). However, such an approach faces challenges. First, tensor analysis is technically more challenging than matrix analysis (e.g., for tensors, even rank, eigenvalues, eigenvectors are not well defined) [15]. Second, compared to the network setting, the χ^2-divergence in the tensor setting involves more layers of expectation taken on the random membership as there are more nodes involved in a hyperedge and has several new terms not seen in the network setting, and thus requires much more delicate analysis. If we ignore such terms by imposing stronger conditions, we will end up with lower bounds that are not sharp. Third, it is always desirable to identify *a simple quantity* that characterizes the fundamental limits of the global testing problem. However, the more complex the model is, the harder it is to identify such a quantity.

Possibly due to such challenges, existing works on lower bounds for hypergraph models have been focused the special SBM as in Example 1. For example, Yuan et al. [18] derived the lower bounds for global testing with the 2-parameter SBM, focusing on the extreme sparse case, Ahn et al. [1] provided lower bound results for exactly recovering the communities (see Kim et al. [14] for a related setting) with a similar model. Lin et al. [16] and Chien et al. [6] used the 3-parameter SBM in Example 1 for study on the lower bounds for community detection. While these papers are very interesting, their lower bounds are characterized by only 2 or 3 parameters (i.e., α_n, ρ_n, τ_n) assumed in their models. For the much broader DCMM model considered here, we have many parameters $\theta, \mathcal{P}, h, F$ (θ is an n-vector and $\mathcal{P}$ is a tensor), and how to extend existing results on special SBM to the DCMM setting here is a challenging problem. In comparison, our results below reveal significant differences between SBM and DCMM.

The key problem is that, given an alternative hypothesis (where $K > 1$), how to identify the null hypothesis (where $K = 1$) that is *most difficult to distinguish* from the alternative? This is a challenging problem for both the null and alternative in DCMM with at least n unknown parameters. As our main contribution, we approach this by proposing a *degree matching strategy*, where for any given DCMM model with $K > 1$, we pair it with a DCMM model with $K = 1$ in a way so that for each node, the expected degree under the null matches with that under the alternative. This way, it is hard to separate the two hypotheses by a naive degree-based statistic. In fact, if the alternative satisfies $\|\theta\|_1\|\theta\|^2\mu_2^2 \to 0$ as $n \to \infty$, then the χ^2-divergence between the pair goes to 0, so two hypotheses are inseparable. Here, $\|\theta\|_q$ is the ℓ^q-norm of θ (we drop q for simplicity if $q = 2$), and μ_k is the k-th singular value of P, where $P \in \mathbb{R}^{K,K^2}$ is the matricization of the tensor $\mathcal{P}$. Comparing with existing works, our results are new and theoretically more satisfying, for we show that the lower bounds and Region of Possibility are prescribed by the simple condition of $\|\theta\|_1\|\theta\|^2\mu_2^2 \to 0$ as $n \to \infty$, regardless of the levels of sparsity, degree heterogeneity, mixed-memberships, and specific entry-values of $\mathcal{P}$.

Our results are as follows. First, we present the standard minimax theory. Define a class of RMM-DCMM models with a finite K, where $K > 1$, $\mu_2^2\|\theta\|^2\|\theta\|_1 \to 0$, and some mild regularity conditions hold. In this class, we can find a RMM-DCMM model (the alternative), and pair it with a DCMM model with $K = 1$ (the null), such that the χ^2-divergence between the pair converges to 0 as $n \to \infty$. Therefore, there *exists an alternative model* that is asymptotically inseparable from the null.

Second, we present the *Region of Impossibility*. The standard minimax theory is popular but it only claims that *there exists an alternative model* within a specified class that is inseparable from the null. It is desirable to show that *for any alternative in the class*, we can pair it with a null so that the χ^2-divergence of the pair goes to 0 as $n \to \infty$. We show that in the parameter space $(\theta, \mathcal{P}, h, F)$ of RMM-DCBM, there is a *Region of Impossibility* defined by $\mu_2^2\|\theta\|^2\|\theta\|_1 \to 0$ (and some mild regularity conditions). For *any alternative* in Region of Impossibility, we pair it with a null such that for each node, the expected degree of the alternative matches with that of the null (i.e., degree-matching strategy). To prove this, we leverage classic results in scaling uniform tensors (e.g., [3, 7]) . Moreover, we show that the χ^2-divergence between the pair goes 0 as $n \to \infty$. The discussion for the broader RMM-DCMM model is similar provided with a mild condition holds. Compared with existing results on minimax lower bounds, these results are more informative and theoretically more satisfying.

Third, we reveal three differences between DCBM and SBM (recall that in SBM, $\theta_1 = \theta_2 = \ldots = \theta_n$). First, for RMM-DCBM, degree matching aforementioned is always possible. However, for SBM, we may have such a degree matching only if the alternative is *symmetric* (precise definition TBD): such a degree matching is impossible when the alternative is *asymmetric*. Second, some naive degree-based statistical tests may be powerful for some SBM with asymmetric settings (to be precise).

DCBM is much broader, allowing degree heterogeneity, and we should not expect such tests be valid tools for the global testing problem.Third, for the two classes of SBM where we *include* and *exclude* the symmetric cases respectively, the optimal rates for testing errors are different. However, DCBM is much broader, and we no longer have such a phenomenon.

Fourth, we discuss the tightness of our lower bounds. Consider a DCBM with moderate degree heterogeneity (i.e., $\theta_{\max} \leq C\theta_{\min}$), we show that if $\mu_2^2\|\theta\|^2\|\theta\|_1/\log(n) \to \infty$, then there is a test such that the sum of Type I and Type II errors $\to 0$ as $n \to \infty$. Then, except for a $\log(n)$ factor here, our lower bounds are tight. We also discuss the more challenging problem of achievability: whether the lower bound is achievable by some fast algorithms. We mention some ideas on achievability, but since our primary focus is on lower bounds, we leave further discussion to the future.

Last, we extend the lower bounds and the tightness of lower bounds to higher-order hypergraphs and non-uniform hypergraphs. We found that XXXXX. We also discovered that XXXX.

In summary, for any given alternative, how to identify the null that is hardest to separate from the alternative is a challenging problem. We solve this by proposing a degree matching strategy. The degree matching strategy allows us to perform delicate analysis of the χ^2-divergence, and so to derive new and sharp results on minimax lower bounds and Region of Possibility for general DCMM models. Compared with existing results, our setting is much broader, and our results are theoretically more satisfying: we reveal that the minimax lower bounds and Region of Impossibility are governed by the simple quantity of $\|\theta\|_1\|\theta\|^2\mu_2^2$, regardless of degree heterogeneity, sparsity, mixed-memberships, and specific entry-values of $\mathcal{P}$. Our ideas are readily extendable to broader settings.

2 Lower Bounds for Uniform and Non-uniform Hypergraphs

For notation simplicity, we focus on 3-uniform hypergraphs in most of this section. The study of higher-order and non-uniform hypergraphs is deffered to Section 2.4. Consider a (3-uniform) DCMM model. Recall that $h = \mathbb{E}_F[\pi_i]$ and that $P \in \mathbb{R}^{K,K^2}$ is the matricization of the tensor $\mathcal{P}$. In this paper, we use $C > 0$ as a generic constant which may vary from occurrence to occurrence. We assume

$$\|P\| \leq C, \qquad \theta_{\max} \leq C, \qquad \max_{1\leq k\leq K}\{h_k\} \leq C \min_{1\leq k\leq K}\{h_k\}, \quad \text{and} \quad \|\theta\|^2\|\theta\|_1 \to \infty. \tag{2.6}$$

The first condition is mild, because the model identifiability already requires that all diagonal entries of $\mathcal{P}$ are 1. The second one is also mild (note that while the largest possible value of $\theta_{\max}$ is $O(1)$, $\theta_{\max}$ is allowed to tend to 0 relatively fast). The third one assumes that the community memberships are balanced, which is also mild. For the last condition, we will see soon that if $\|\theta\|^2\|\theta\|_1 \to 0$, then the signal is so weak that successful global testing is impossible; this condition is also mild.

Below, Section 2.1 presents the standard minimax theory and least favorable configuration; Section 2.2 presents the impossibility results for RMM-DCBM; Section 2.3 discusses several differences between SBM and DCBM; Section 2.4 extends the results to general hypergraphs. The tightness of lower bounds are addressed in Section 3.

2.1 Standard minimax lower bounds (RMM-DCMM)

We start with the least favorable configuration. The goal is to find a pair of models (a null model with $K = 1$, and an alternative model with $K > 1$) which are hard to distinguish from each other. We use the following degree-matching technique: we choose the pair in a way so that for each node, the expected degree under the null matches with that under the alternative, approximately. The idea is, if the degrees are not matching, then we may separate the two hypotheses by a simple degree-based statistic, so we should not expect the χ^2-divergence between the pair to be small.

In detail, consider a pair of models, a DCMM model with $K = 1$ and a RMM-DCMM model with $K > 1$, where for all $1 \leq i_1, i_2, i_3 \leq n$, the Bernoulli probability tensors $\mathcal{Q}$ and $\mathcal{Q}^*$ satisfy

$$\mathcal{Q}_{i_1i_2i_3} = \theta_{i_1}\theta_{i_2}\theta_{i_3}, \qquad \mathcal{Q}^*_{i_1i_2i_3} = \theta^*_{i_1}\theta^*_{i_2}\theta^*_{i_3} \cdot \pi'_{i_1}(\mathcal{P}\pi_{i_3})\pi_{i_2}, \tag{2.7}$$

where we recall that $(\mathcal{P}\pi_{i_3})$ is a $K \times K$ matrix (see Definition 1.1). We call the two models the *null and the alternative*, respectively. In (2.7), the community structure tensor $\mathcal{P}$ is as in (1.2), and π_i and $h = \mathbb{E}_F[\pi_i]$ are as in (1.5). Recall that $\theta = (\theta_1, \theta_2, \ldots, \theta_n)'$. Similarly, let $\theta^* = (\theta^*_1, \theta^*_2, \ldots, \theta^*_n)'$.

Fix $1 \le i_1 \le n$. By definitions and elementary statistics, the leading term of the expected degree of node i_1 under the null and alternative are

$$\theta_{i_1}\|\theta\|_1^2 \qquad \text{and} \qquad \theta^*_{i_1}\cdot\|\theta^*\|_1^2\cdot\pi'_{i_1}a, \tag{2.8}$$

respectively, where $a = (\mathcal{P}h)h \in \mathbb{R}^K$. For least favorable construction, we choose $(\mathcal{P}, h)$ in a way so that

$$a = (\mathcal{P}h)h = c_0^3\mathbf{1}_K, \qquad \text{for a scalar } c_0 > 0. \tag{2.9}$$

For broadness, we allow c_0 to depend on n. There are many $(\mathcal{P}, h)$ that satisfy (2.9). For example, in the 2-parameter SBM model as in Example 1 with $h = (1/K, \ldots, 1/K)'$, $a_k = (1/K^2)[1 + (K^2 - 1)\rho_n]$ for all $1 \le k \le K$, so (2.9) is satisfied with $c_0^3 = (1/K^2)[1 + (K^2 - 1)\rho_n]$. Moreover, we choose θ_i^* in the alternative model such that

$$\theta_i^* = (1/c_0)\theta_i, \qquad 1 \le i \le n. \tag{2.10}$$

Now, by (2.8)-(2.9), for all $1 \le i_1 \le n$, $\theta^*_{i_1}\|\theta^*\|_1^2\cdot\pi'_{i_1}a = c_0^3\theta^*_{i_1}\|\theta^*\|_1^2 = \theta_{i_1}\|\theta\|_1^2$. Therefore, for each node, the expected degree under the alternative matches that under the null, approximately, so it is hard to separate the null and the alternative by some naive degree-based statistics. Only when this degree-matching holds, we can hope two models are asymptotically indistinguishable. This is the key for our least favorable configuration. Recall that $P \in \mathbb{R}^{K,K^2}$ is the matricization of the community structure tensor $\mathcal{P}$ and μ_k is the k-th largest singular value of P. We have the following theorem.

Theorem 2.1 (Least favorable configuration)**.** *Fix $K > 1$ and consider a pair of models, a null and an alternative with K communities, given in (2.7), where (2.9)-(2.10) hold. Assume (2.6) holds and $\|\theta\|_1\|\theta\|^2\mu_2^2 = o(1)$. As $n \to \infty$, the χ^2-divergence between the pair tends to 0. Therefore, the two models are asymptotically indistinguishable: for any test, the sum of Type I and Type II errors is no smaller than $1 + o(1)$.*

Example 2. Consider a special case where in the alternative, $\mathcal{P}_{i_1i_2i_3} = 1$ if $i_1 = i_2 = i_3$ and $\mathcal{P}_{i_1i_2i_3} = \rho_n$ otherwise. This includes Example 1 as a special case, but we allow unequal θ_i's and mixed-memberships. In this case, $\mu_1 = [(K^3 - K^2)\rho_n^2 + (1 + (K-1)\rho_n)^2]^{1/2}$ and $\mu_2 = \ldots = \mu_K = |1 - \rho_n|$. The main condition of $\|\theta\|_1\|\theta\|^2\mu_2^2 \to 0$ is equivalent to that of $\|\theta\|_1\|\theta\|^2(1-\rho_n)^2 \to 0$.

The least favorable configuration gives rises to the standard minimax theorem. Fix $K \ge 1$ and consider a DCMM model (where π_i are non-random). Introduce a vector $g \in \mathbb{R}^K$ by $g_k = (1/\|\theta\|_1)\sum_{i=1}^n \theta_i\pi_i(k)$, $1 \le k \le K$. For a constant $0 < c_0 < 1$, and two positive sequences $\{\alpha_n\}_{n=1}^\infty$ and $\{\beta_n\}_{n=1}^\infty$, we define a class of DCMM models by

$$\mathcal{M}_n(K, c_0, \alpha_n, \beta_n) = \left\{ (\theta, \Pi, P): \begin{array}{l} \|P\| \le c_0,\ \theta_{\max} \le c_0,\ \max_{1\le k\le K} g_k \le c_0^{-1}\min_{1\le k\le K} g_k, \\ \|\theta\|_1\|\theta\|^2 \ge 1/\beta_n,\ \|\theta\|_1\|\theta\|^2\mu_2^2 \le \alpha_n \end{array} \right\}.$$

Here, we assume $\max_{1\le k\le K}\{g_k\} \le C\min_{1\le k\le K}\{g_k\}$, so the tensor is balanced. This is similar to the third condition in (2.6), except that π_i's are random there. For the null case, $K = P = \pi_i = 1$, and the above defines a class of θ, which we write for short by $\mathcal{M}_n(1, c_0, \alpha_n, \beta_n) = \mathcal{M}_n^*(\beta_n)$. The following theorem is proved in the supplement.

Theorem 2.2 (Minimax lower bound)**.** *Fix $K \ge 2$, a constant $c_0 > 0$, and any sequences $\{\alpha_n\}_{n=1}^\infty$ and $\{\beta_n\}_{n=1}^\infty$ where $\alpha_n = o(1)$ and $\beta_n = o(1)$. As $n \to \infty$, $\inf_\psi\{\sup_{\theta\in\mathcal{M}_n^*(\beta_n)}\mathbb{P}(\psi = 1) + \sup_{(\theta,\Pi,P)\in\mathcal{M}_n(K,c_0,\alpha_n,\beta_n)}\mathbb{P}(\psi = 0)\} = 1 - o(1)$, where the infimum is over all possible tests ψ.*

2.2 The region of impossibility for RMM-DCBM

The standard minimax theorem in Section 2.1 only says that in the class of all alternative models with $\|\theta\|^2\|\theta\|_1\mu_2^2 \to 0$, *there exists one* where we can pair it with a null so the pair are asymptotically inseparable. A much more satisfactory result is to show that, for *any alternative* in the same class, we can pair it with a null such that the pair are asymptotically inseparable. In this section, we prove this for the RMM-DCBM case (mixed-memberships are not allowed). The discussion for the RMM-DCMM (mixed-memberships allowed) is slightly different and is in the supplemental material.

Consider again a pair of models, a DCBM null model and an RMM-DCBM model with $K > 1$, where the Bernoulli probability tensors are $\mathcal{Q}$ and $\mathcal{Q}^*$, respectively. We assume for all $1 \le i_1, i_2, i_3 \le n$,

$$\mathcal{Q}_{i_1i_2i_3} = \theta_{i_1}\theta_{i_2}\theta_{i_3}, \qquad \mathcal{Q}^*_{i_1i_2i_3} = \theta^*_{i_1}\theta^*_{i_2}\theta^*_{i_3}\pi'_{i_1}(\mathcal{P}\pi_{i_3})\pi_{i_2}. \tag{2.11}$$

Here, the community structure tensor $\mathcal{P}$ is as in (1.2), π_i and $h = \mathbb{E}_F[\pi_i]$ are as in (1.5), and $\mathrm{supp}(F) = V_0^*$. Similarly, the goal is degree-matching: for any $(\theta, \mathcal{P}, h, F)$, we construct θ^* in a way so that for each node, the expected degrees under the null and the alternative match with each other approximately. Recall that in Section 2.2, in order to have a desired degree matching, it is crucial to pick an alternative model where the $(\mathcal{P}, h)$ satisfies $a \equiv (\mathcal{P}h)h = c_0^3 \mathbf{1}_K$ for some scalar $c_0 > 0$; once this holds, we have the desired degree matching by taking $\theta^* = (1/c_0)\theta$. Unfortunately, for general $(\mathcal{P}, h)$, we don't have $a \equiv (\mathcal{P}h)h \propto \mathbf{1}_K$, so we should not expect to have the desired degree matching by taking $\theta^* \propto \theta$. In short, for our purpose here, the approach in Section 2.1 no longer works, and we must find a new approach to constructing the model pair.

Our proposal is as follows. For any diagonal matrix $D = \mathrm{diag}(d_1, d_2, \ldots, d_K)$ with $d_k > 0$, $1 \leq k \leq K$, define a K-dimensional vector a^D by $a^D = (\mathcal{P}^D h)h$, where $\mathcal{P}^D$ is a 3-tensor in dimension K where $\mathcal{P}^D_{k_1k_2k_3} = d_{k_1}d_{k_2}d_{k_3}\mathcal{P}_{k_1k_2k_3}$, $1 \leq k_1, k_2, k_3 \leq K$. We now select a matrix D such that $a^D = \mathbf{1}_K$. The following lemma states that such a D always exists and is unique. It leverages classic result on tensor scaling (e.g., [3, 7]) and is proved in the supplement.

Lemma 2.1. *Fix $K > 1$ and let $\mathcal{P}$, h, D, and a^D be as above. Suppose $\min\{h_1, h_2, \ldots, h_K\} \geq C$. There exists an unique diagonal matrix $D = \mathrm{diag}(d_1, d_2, \ldots, d_K)$ such that $a^D = \mathbf{1}_K$.*

Now, given $(\theta, \mathcal{P}, h, F)$ and the two models in (2.11), let $D = \mathrm{diag}(d_1, d_2, \ldots, d_K)$ be as in Lemma 2.1 (if there are more than 1, pick any of them and fix it). Let $\mathcal{P}^*$ be $\mathcal{P}^D$ evaluated at the D selected above, and let $a^* = (\mathcal{P}^* h)h$. Moreover, in (2.11), choose θ_i^* in a way such that

$$\theta_i^* = d_k \theta_i, \quad \text{if node } i \text{ belongs to community } k. \tag{2.12}$$

Combining these with (2.11) and basic algebra, we have $\mathcal{Q}^*_{i_1i_2i_3} = \theta_{i_1}\theta_{i_2}\theta_{i_3}\pi'_{i_1}(\mathcal{P}^*\pi_{i_3})\pi_{i_2}$. Note that by Lemma 2.1, $a^* = \mathbf{1}_K$. Therefore, by similar calculations as in Section 2.1, for $1 \leq i_1 \leq n$, in the alternative model, the leading term of the expected degree of node i_1 is $\theta_{i_1}(\pi'_{i_1}a^*)\|\theta\|_1^2 = \theta_{i_1}\|\theta\|_1^2$. Combining this with (2.8), we have that for each node, the expected degrees under the null and the alternative are approximately matching. Therefore, it is hard to separate two models by a naive degree-based statistic. Recall that P is the matricization of $\mathcal{P}$ in the alternative, and μ_k is the k-th singular value of P. We have the the following theorem.

Theorem 2.3 (Impossibility for RMM-DCBM)**.** *Fix $K > 1$. For any given $(\theta, \mathcal{P}, h, F)$, consider a pair of models, a null and an alternative with K communities, as in (2.11), where θ_i^* are given by (2.12) with the matrix D as in Lemma 2.1. Suppose (2.6) holds and $\|\theta\|_1\|\theta\|^2\mu_2^2 = o(1)$. As $n \to \infty$, the χ^2-divergence between the pair tends to 0. Therefore, the two models are asymptotically indistinguishable: for any test, the sum of Type I and Type II errors is no smaller than $1 + o(1)$.*

In the parameter space $(\theta, \mathcal{P}, h, F)$ for DCBM, we call the region prescribed by $\|\theta\|_1\|\theta\|^2\mu_2^2 \to 0$ the *Region of Impossibility*: for any model in this region, we can pair it with a null so that the pair are asymptotically inseparable.

In the supplemental material, we give a similar theorem for RMM-DCMM, which allows for mixed-memberships. Consider a pair of models as in (2.11), where π_i's are drawn from a distribution F on the standard simplex V. We choose θ_i^* by $\theta_i^* = \theta_i / \|D_0^{-1}\pi_i\|_1$, where D_0 is a positive diagonal matrix jointly determined by $\mathcal{P}$ and $h = \mathbb{E}_F[\pi_i]$. When F only has point masses on the vertices of the simplex (i.e., RMM-DCMM reduces to RMM-DCBM), D_0 reduces to the D in Lemma 2.1 (which always exists and is unique), and θ_i^* is the same as in (2.12). However, for a general F, the existence of D_0 requires some additional regularity conditions. See the supplemental material for details.

2.3 Some differences between SBM and DCBM models

The SBM is a special case of DCBM where $\theta_1 = \ldots = \theta_n$, and is frequently used in recent works (e.g., [8, 1, 16]). We compare the global testing for SBM with that for DCBM. Recall that for an alternative RMM-DCBM model, we can always (by a tensor scaling technique) pair it with a null, such that for each node, the expected degrees under the null and the alternative match with each other. This is made possible because a null model in DCBM has n unknown parameters $\theta_1, \ldots, \theta_n$, giving us sufficient flexibility in choosing the null for degree matching. However, for the RMM-SBM, we can not always have such a degree matching, as RMM-SBM is much narrower than RMM-DCBM. Especially, a null model in SBM only has 1 unknown parameter, so we do not have sufficient flexibility in choosing the null for degree matching. A direct consequence is that some naive degree-based test may have non-trivial power for SBM.

Consider a pair of models, where the Bernoulli probability tensors $\mathcal{Q}$ and $\mathcal{Q}^*$ under two hypotheses are such that, for $1 \leq i_1, i_2, i_3 \leq n$,

$$\mathcal{Q}_{i_1 i_2 i_3} = \alpha_n, \qquad \mathcal{Q}^*_{i_2 i_2 i_3} = \pi'_{i_1}(\mathcal{P}\pi_{i_3})\pi_{i_3}. \tag{2.13}$$

Same as before, π_i are *iid* generated from a distribution F on $\{e_1, e_2, \ldots, e_K\}$ with $h = \mathbb{E}_F[\pi_i]$. Let $\hat{\alpha}_n = (n(n-1)(n-2))^{-1}\mathbf{1}'_n(\mathcal{A}\mathbf{1}_n)\mathbf{1}_n$, $\eta = (1/2)(\mathcal{A}\mathbf{1}_n)\mathbf{1}_n$, and $\bar{\eta}$ be the mean of $\eta_1, \eta_2, \ldots, \eta_K$. Consider the centered-$\chi^2$-statistic

$$\psi_n = \sum_{1 \leq i \leq n} \left[\frac{(\eta_i - \bar{\eta})^2}{\binom{n-1}{2}\hat{\alpha}_n(1-\hat{\alpha}_n)} - 1 \right] / \sqrt{2n}, \tag{2.14}$$

Lemma 2.2. *Consider the global testing problem (1.4) under the SBM model (2.13) for H_0 and H_1, respectively. Let $\widetilde{\alpha}_n = \mathbb{E}[\hat{\alpha}_n]$, $h = (1/n)\sum_{i=1}^n \pi_i$ and $\Sigma = \frac{1}{n}\sum_{i=1}^n (\pi_i - h)(\pi_i - h)'$. Let $\lambda_{K-1}(\Sigma)$ be the $(K-1)$-th largest eigenvalue (in magnitude) of Σ. Assume $\alpha_n \leq c_0$, $\max_{1 \leq i,j,k \leq K}\{\mathcal{P}_{ijk}\} \leq c_0$, $n^2\alpha_n \to \infty$, and $n^2\widetilde{\alpha}_n \to \infty$. Also, assume $\min\{h_1, h_2, \ldots, h_K\} \geq C$ and $\lambda_{K-1}(\Sigma) \geq C$. Write $\delta_n = \|\widetilde{\alpha}_n^{-1}(I_K - H_K)(\mathcal{P}h)h\|$, where $H_K = (1/K)\mathbf{1}_K\mathbf{1}'_K$ and I_K is the identity matrix of the same size. As $n \to \infty$, $\psi_n \to N(0,1)$ if H_0 holds, and $\psi_n \to \infty$ if H_1 holds and $n^{3/2}\widetilde{\alpha}_n^{1/2}\delta_n^2 \to \infty$*

The power of the χ^2-test hinges on δ_n. Note that $\delta_n = 0$ if and only if $(\mathcal{P}h)h \propto \mathbf{1}_K$. We call the cases of $(\mathcal{P}h)h \propto \mathbf{1}_K$ and $(\mathcal{P}h)h \not\propto \mathbf{1}_K$ the *symmetric* case and the *asymmetric* case, respectively. For symmetric SBM, we can implement the same degree matching strategy as before and prove a similar lower bound to that in Theorem 2.3. However, for asymmetric SBM, the χ^2-test may have non-trivial power, implying a shift of the lower bound. This is better seen from the example below.

Example 3. Consider an SBM global testing setting where we either have $K = 1$ (null) or $K = 2$ (alternative). When $K = 2$, we assume $h = (a, 1-a)'$ for some $0 < a < 1$, $\mathcal{P}_{ijk}$ is equal to ρ_0 if $i = j = k$ and ρ_1 otherwise. Suppose $n^2\rho_0 \to \infty$ and $\rho_1/\rho_0 \to 1$.

- Case 1: a can be any number in $(0,1)$. This case includes the symmetric SBM ($a = 1/2$). All tests are asymptotically powerless in the least favorable configuration if $\tau_n \equiv n^2\rho_0^{-1}(\rho_1 - \rho_0)^2 \to 0$ (this is a direct corollary of Theorem 2.3, noting that $D = I_K$ in (2.12)).
- Case 2: a is restricted to $(0, 1/2 - c_0] \cup [1/2 + c_0, 1)$. It excludes the symmetric SBM. By Lemma 2.2 and direct calculations, the χ^2-test has asymptotically full power if $\delta_n \equiv (n\rho_0)^{-1/2}(2a-1)^2\tau_n \to \infty$. In this range, it is possible that $\tau_n \to 0$. Hence, the detection boundary is different from that of Case 1.

In conclusion, we have three interesting discoveries for SBM. First, the degree matching strategy for constructing the least favorable configuration does not always work for RMM-SBM. Second, a simple degree-based χ^2-test may have power for asymmetric SBM, but it will not have power for symmetric SBM. Last, the detection boundary can be significantly different when the symmetric cases are *included* and *excluded*. Note that we have these discoveries because SBM is too specific; restricted to this narrower class, it is sometimes much easier to separate two hypotheses.

2.4 The Results for General Hypergraphs

We consider a general hypergraph, where the hyper-edges may contain different numbers of nodes. Let $M \geq 2$ be the maximum number of nodes in a hyper-edge. For each $2 \leq m \leq M$, let $\mathcal{A}^{(m)}$ be the adjacency tensor that stores only the order-m hyper-edges, i.e., $\mathcal{A}^{(m)}(i_1, i_2, \ldots, i_m)$ is equal to 1 if $\{i_1, i_2, \ldots, i_m\}$ is a hyper-edge and 0 otherwise. In this way, a non-uniform hypergraph is sliced into a collection of uniform hypergraphs. We denote it by $\mathcal{A}[M] \equiv \{\mathcal{A}^{(2)}, \mathcal{A}^{(3)}, \ldots, \mathcal{A}^{(M)}\}$. We model each $\mathcal{A}^{(m)}$ by a tensor-DCMM model. Let $\mathcal{P}^{(m)}$ be a symmetric order-m tensor in dimension K, and let $\theta^{(m)} = (\theta_1^{(m)}, \theta_2^{(m)}, \ldots, \theta_n^{(m)})'$ be a positive vector of degree parameters. Let $\pi_1, \pi_2, \ldots, \pi_n$ be the K-dimensional membership vectors.

Definition 2.1. *Given an order-m tensor $\mathcal{M}$ in dimension K and vectors $b_1, b_2, \ldots, b_m \in \mathbb{R}^K$, let $[\mathcal{M}; b_1, \cdots, b_K]$ be an order-m tensor in dimension K, whose entry in $(i_1, i_2, \ldots, i_m)$ is equal to $\sum_{1 \leq k_1, k_2, \ldots, k_m \leq K}[\mathcal{M}_{k_1 k_2 \ldots k_m} b_1(i_1) b_2(i_2) \cdots b_m(i_m)]$, for any $1 \leq i_1, \ldots, i_m \leq K$.*

We assume $\{\mathcal{A}^{(m)}(i_1, i_2, \ldots, i_m)\}_{1\le i_1<i_2<\ldots<i_m\le n}$ are independent Bernoulli variables, where the Bernoulli probabilities are specified by the tensor $\mathcal{Q}^{(m)}$, given by

$$\mathcal{Q}^{(m)}_{i_1 i_2 \ldots i_m} = \theta^{(m)}_{i_1} \ldots \theta^{(m)}_{i_m} \times [\mathcal{P}^{(m)}; \pi_{i_1}, \cdots, \pi_{i_m}], \qquad 1 \le i_1, i_2, \ldots, i_m \le n. \tag{2.15}$$

Similar to (1.2), we have $\mathbb{E}[\mathcal{A}^{(m)}] = \mathcal{Q}^{(m)} - \mathrm{diag}(\mathcal{Q}^{(m)})$. This extends the tensor-DCMM model to m-uniform hypergraphs for a general m.

Definition 2.2. *We say that $\mathcal{A}[M] = \{\mathcal{A}^{(2)}, \ldots, \mathcal{A}^{(M)}\}$ follows a (general) tensor-DCMM model if $\{\mathcal{A}^{(j)}\}_{2\le j\le M}$ are independent of each other and each $\mathcal{A}^{(j)}$ follows an m-uniform tensor-DCMM model as in (2.15), where $\pi_1, \pi_2, \ldots, \pi_n$ are shared by all $2 \le j \le M$.*

This definition includes the uniform tensor-DCMM as a special case, for which we set $\mathcal{P}^{(m)}$ to be a zero matrix for all but one value of m. The assumptions that $\mathcal{A}^{(j)}$'s are independent of each other and share the same membership vectors were also used in the literature; for example, see [9]. The model in [9] is a special case with $\theta^{(m)}_1 = \theta^{(m)}_2 = \cdots = \theta^{(m)}_i$ for each $2 \le m \le M$.

We present the *Region of Impossibility* for testing H_0: $K = 1$ versus H_1: $K > 1$. Similarly as in Section 2.2, we focus on the special case of tensor-DCBM models (i.e., each π_i is degenerate); the study of tensor-DCMM is similar. Fix $K > 1$. Consider a DCBM null model with probability tensors $\mathcal{Q}[M] = \{\mathcal{Q}^{(2)}, \ldots, \mathcal{Q}^{(M)}\}$ and a RMM-DCBM model with probability tensors $\mathcal{Q}^*[M] = \{\mathcal{Q}^{*(2)}, \ldots, \mathcal{Q}^{*(M)}\}$, where for every $2 \le m \le M$ and $1 \le i_1, i_2, \ldots, i_m \le n$,

$$\mathcal{Q}^{(m)}_{i_1, i_2, \ldots, i_m} = \theta^{(m)}_{i_1} \theta^{(m)}_{i_2} \cdots \theta^{(m)}_{i_m}, \tag{2.16}$$

$$\mathcal{Q}^{*(m)}_{i_1, i_2, \ldots, i_m} = \theta^{*(m)}_{i_1} \cdots \theta^{*(m)}_{i_m} \times [\mathcal{P}^{(m)}_{k_1 k_2 \ldots k_m}; \pi_{i_1}, \ldots, \pi_{i_m}], \qquad \pi_i \overset{iid}{\sim} F. \tag{2.17}$$

The support of F is in $V_0 = \{e_1, e_2, \ldots, e_K\}$. Let $h = \mathbb{E}[\pi_i]$ and suppose $\min\{h_1, \ldots, h_K\} \ge C$. In the supplemental material, we provide a lemma analogous to Lemma 2.1: For each $2 \le m \le M$, there exists a unique diagonal matrix $D^{(m)} = \mathrm{diag}(d^{(m)}_1, d^{(m)}_2, \ldots, d^{(m)}_K)$ such that $\sum_{1\le i_2, \ldots, i_m \le K} d^{(m)}_{i_1} \cdot \mathcal{P}^{(m)}_{i_1 \cdots i_m} \cdot (d^{(m)}_{i_2} h_{i_2}) \cdots (d^{(m)}_{i_m} h_{i_m}) = 1$, for every $1 \le i_1 \le K$. Its proof leverages the Sinkhorn-type results for higher-order tensors (e.g., [3, 7]). In (2.17), we construct $\theta^{*(m)} = (\theta^{*(m)}_1, \ldots, \theta^{*(m)}_n)$ by

$$\theta^{*(m)}_i = d^{(m)}_k \theta^{(m)}_i, \quad \text{if node } i \text{ belongs to community } k. \tag{2.18}$$

Let $P^{(m)} \in \mathbb{R}^{K \times K^{m-1}}$ be the matricization of $\mathcal{P}^{(m)}$ and let $\mu^{(m)}_2$ be the second singular value of $P^{(m)}$. The following theorem is proved in the supplemental material:

Theorem 2.4 (Impossibility for non-uniform RMM-DCBM)**.** *Fix $K > 1$ and $M \ge 2$. For any given (h, F) and $\{(\theta^{(m)}, \mathcal{P}^{(m)})\}_{2\le m\le M}$, consider a pair of models, a null as in (2.17) and an alternative with K communities as in (2.16), where $\{\theta^{*(m)}_i\}_{1\le i\le n, 2\le m\le M}$ are as in (2.18). Suppose $\|P^{(m)}\| \le C$, $\max_{1\le i\le n} \theta^{(m)}_i \le C$, and $\min_{1\le k\le K} h_k \ge C$. If $\max_{2\le m\le M}\{\|\theta^{(m)}\|_1^{m-2} \|\theta^{(m)}\|^2 (\mu^{(m)}_2)^2\} = o(1)$, then as $n \to \infty$, the χ^2-divergence between the pair tends to 0.*

3 Tightness of the lower bound

We propose a test for the global testing problem (1.4) and show that it attains the lower bounds in Theorems 2.3-2.4. For simplicity, we only discuss this for DCBM with moderate degree heterogeneity but the tightness holds in much broader settings. For each $2 \le m \le M$, we first compute a vector $\eta^{(m)}$ by an iterative algorithm. Fix m. Given a positive vector $u \in \mathbb{R}^n$, define $g(u) \in \mathbb{R}^n$ by

$$g_i(u) = \frac{\sum_{i_2, \ldots, i_m (\text{distinct})} \mathcal{A}^{(m)}_{i\, i_2 \ldots i_m} + \sum_{i_2, \ldots, i_m (\text{non-distinct})} u_i u_{i_2} \cdots u_{i_m}}{\big(\sum_{i_1, \ldots, i_m (\text{distinct})} \mathcal{A}^{(m)}_{i_1 i_2 \ldots i_m} + \sum_{i_1, \ldots, i_m (\text{non-distinct})} u_{i_1} \cdots u_{i_m}\big)^{(m-1)/m}}. \tag{3.19}$$

The algorithm initializes with $u^{(0)} = \mathbf{0}_n$ and computes $u^{(k)} = g(u^{(k-1)})$ for all $k \ge 1$. For $m = 3$, we run this iteration only once and output $u^{(1)}$ as $\eta^{(m)}$. For $m \ge 4$, we run this iteration for k_m times, where the minimum requirement of k_m is given in Theorem 3.1 below, and output $u^{(k_m)}$ as $\eta^{(m)}$. Next, we construct a statistic $Q^{(m)}_n$ for each $2 \le m \le M$. Let $\mathcal{A}^{*(m)}$ be a tensor with the same size of $\mathcal{A}^{(m)}$

where $\mathcal{A}^{*(m)}_{i_1\ldots i_m} = \mathcal{A}^{(m)}_{i_1\ldots i_m} - \eta^{(m)}_{i_1}\cdots\eta^{(m)}_{i_m}$, $1 \le i_1,\ldots,i_m \le n$. We say $S = (S_1, S_2, \ldots, S_{m+1})$ is a partition of $\{1, 2, \ldots, n\}$ if $S_1, S_2, \ldots, S_{m+1}$ are disjoint and $\{1, 2, \ldots, n\} = S_1 \cup S_2 \cup \cdots \cup S_{m+1}$. Let B be the set of all such partitions. Fixing $S = (S_1, S_2, \ldots, S_{m+1}) \in B$ and $1 \le k_1, \ldots, k_m \le m+1$, let $X_{S,k_1\ldots k_m} = \sum_{i_1\in S_{k_1},\ldots,i_m\in S_{k_m}(dist)} \mathcal{A}^{*(m)}_{i_1\ldots i_m}$. Introduce

$$Q^{(m)}_n = \max_{S=(S_1,\ldots,S_{m+1})\in B} \ \max_{1\le k_1,\ldots,k_m\le m+1} \{|X^{(m)}_{S,k_1\ldots k_m}|\}. \tag{3.20}$$

Let $V^{(m)}_n = \binom{n}{m}\hat{\alpha}_n(1-\hat{\alpha}_n)$, where $\hat{\alpha}_n = [(n-m)!/n!]\sum_{i_1,\ldots,i_m=1}^n \mathcal{A}^{(m)}_{i_1\ldots i_m}$. We propose the test statistic as $\phi_n = \max_{2\le m\le M}\phi^{(m)}_n$, where $\phi^{(m)}_n = Q^{(m)}_n/[n\log(n)V^{(m)}_n]^{1/2}$. The following theorem is proved in the supplement material.

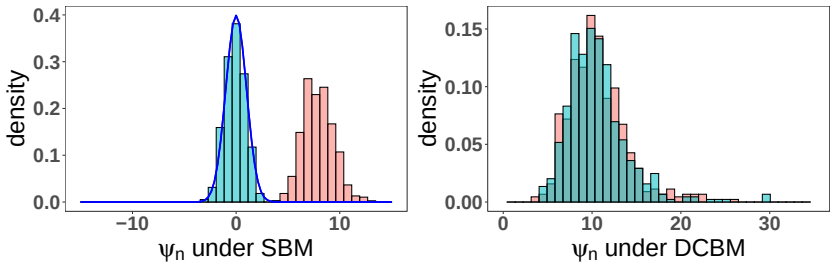
Theorem 3.1 (Tightness of lower bounds)**.** *Consider the general tensor-DCBM model with $M \ge 2$. Let $h = n^{1/2}\sum_{i=1}^n \pi_i$. Suppose, for each*

Suppose, for each $m = 2, \ldots, M$, assume that $\max_{1\le k\le K}\{g^{(m)}_k\} \le C\min_{1\le k\le K}\{g^{(m)}_k\}$, $\theta^{(m)}_{\max} \le C\theta^{(m)}_{\min}$, and $\theta^{(m)}_{\max} \le c_0$ for a constant $c_0 < 1$. As $n \to \infty$, if $\|\theta^{(m)}\|_1^{m-2}\|\theta^{(m)}\|^2 \to \infty$, we have $\mathbb{P}(\phi^{(m)}_n \le 1) = 1 + o(1)$ if H_0 holds, and $\mathbb{P}(\phi^{(m)}_n \le 1) = o(1)$ if H_1 holds and $\|\theta^{(m)}\|_1^{m-2}\|\theta^{(m)}\|^2(\mu^{(m)}_2)^2/\log(n) \to \infty$, $\|\theta^{(m)}\|_1^{m-2}\|\theta^{(m)}\|^2/(n^2\log(n)) = o(1)$.

Remark (Achievability by a fast algorithm) The above test is computationally expensive, so it is desirable to have some fast algorithms that achieve the lower bounds (i.e., to have asymptotically full power once $\|\theta\|_1\|\theta\|^2\mu_2^2 \to \infty$). [19] studied a special DCBM setting where the tensor $\mathcal{P}$ only has two different values, and showed that the test has asymptotically full power if the tensor is sufficiently sparse and a specific condition holds (it is however unclear whether this condition holds whenever $\|\theta\|_1\|\theta\|^2\mu_2^2 \to \infty$). However, in the broad DCMM setting considered here (where we allow mixed-membership, severe degree heterogeneity, flexible sparsity levels, and where we do not require $\mathcal{P}$ to take only 2 or 3 different values), it remains unclear how to develop a fast algorithm that achieve the lower bound. We leave it to future work.

4 Numerical study

Consider the χ^2-statistic ψ_n in (2.14) for SBM. By Lemma 2.2, in the null case, $\psi_n \approx N(0,1)$, and in the alternative case, the test may lack power in the symmetric case but have power in the asymmetric case. We now investigate this with simulations. Let $(n, K) = (500, 2)$, $\theta_1 = \ldots = \theta_n = n^{-1/2}$, and $\mathcal{P}$ be the tensor that takes values of 1 and $1/4$ on diagonal and off-diagonal entries, respectively. We consider a symmetric case where each communities have 250 nodes (so $h = (1/2, 1/2)'$) and an asymmetric case where two communities have 375 and 125 nodes, respectively (so $h = (3/4, 1/4)$). For each setting, we randomly generate the hypergraphs, apply the χ^2-test statistic in (2.14), and repeat for 500 times independently. The histogram of the χ^2 test statistics ψ_n are in the left panel of Figure 1 (green: symmetric; red: asymmetric case; blue: density of $N(0,1)$). The results suggests (a) in the symmetric case, the histogram of ψ_n fits well with $N(0,1)$ so the test lacks power, and (b) in the asymmetric case, the test is clearly seen to have power. This supports our theory.



Consider a pair of DCBM as in (2.11). Our theory says if they have matching expected degrees, then the χ^2-test won't be able to separate them. Fixing $(n, K) = (500, 2)$, consider a null setting where $\theta_i \overset{iid}{\sim} \text{Pareto}(0.5, 5)$, and an alternative setting where $(n, K, \mathcal{P}, h)$ are the same as in the

asymmetric case above, but $\theta_i^* = d_k\theta_i$ if node i belongs to community k, $1 \le k \le 2, 1 \le i \le n$. Here, $D = \mathrm{diag}(d_1, d_2) = \mathrm{diag}(1.10, 1.80)$ is the diagonal matrix obtained by applying the `solve` function in Matlab on the system of K equations given by Lemma 2.1. The two models are seen to have matching expected degrees. For each model, we randomly generate the hypergraphs and then apply the χ^2-test statistic ψ_n, and repeat 500 times independently. The results are in the right panel of Figure **??** (green: null; red: alternative). It is seen that both histograms fit well with $N(0, 1)$, so the χ^2-test has very little power in separating the two models. This supports our discovery.

References

[1] Kwangjun Ahn, Kangwook Lee, and Changho Suh. Community recovery in hypergraphs. *IEEE Transactions on Information Theory*, 65(10):6561–6579, 2019.

[2] Jess Banks, Cristopher Moore, Joe Neeman, and Praneeth Netrapalli. Information-theoretic thresholds for community detection in sparse networks. In *COLT*, pages 383–416, 2016.

[3] Ravindra Bapat. D1ad2 theorems for multidimensional matrices. *Linear Algebra and its Applications*, 48:437–442, 1982.

[4] Javier Béjar, Sergio Álvarez, Dario García, Ignasi Gómez, Luis Oliva, Arturo Tejeda, and Javier Vázquez-Salceda. Discovery of spatio-temporal patterns from location-based social networks. *J. Exp. Theor. Artif. Intell.*, 28(1-2):313–329, 2016.

[5] Austin R Benson, David F Gleich, and Jure Leskovec. Higher-order organization of complex networks. *Science*, 353(6295):163–166, 2016.

[6] I Chien, Chung-Yi Lin, and I-Hsiang Wang. Community detection in hypergraphs: Optimal statistical limit and efficient algorithms. In *International Conference on Artificial Intelligence and Statistics*, pages 871–879, 2018.

[7] Shmuel Friedland. Positive diagonal scaling of a nonnegative tensor to one with prescribed slice sums. *Linear algebra and its applications*, 434(7):1615–1619, 2011.

[8] Debarghya Ghoshdastidar and Ambedkar Dukkipati. Consistency of spectral partitioning of uniform hypergraphs under planted partition model. *Advances in Neural Information Processing Systems*, 27:397–405, 2014.

[9] Debarghya Ghoshdastidar and Ambedkar Dukkipati. Consistency of spectral hypergraph partitioning under planted partition model. *The Annals of Statistics*, 45(1):289–315, 2017.

[10] Jacopo Grilli, György Barabás, Matthew J Michalska-Smith, and Stefano Allesina. Higher-order interactions stabilize dynamics in competitive network models. *Nature*, 548(7666):210, 2017.

[11] Pengsheng Ji, Jiashun Jin, Zheng Tracy Ke, and Wanshan Li. The phase ii data for the networks of statisticians. *Manuscript*, 2019.

[12] Jiashun Jin, Zheng Tracy Ke, and Shengming Luo. Optimal adaptivity of signed-polygon statistics for network testing. *arXiv:1904.09532*, 2019.

[13] Zheng Tracy Ke, Feng Shi, and Dong Xia. Community detection for hypergraph networks via regularized tensor power iteration. *arXiv preprint arXiv:1909.06503*, 2019.

[14] Chiheon Kim, Afonso S Bandeira, and Michel X Goemans. Stochastic block model for hypergraphs: Statistical limits and a semidefinite programming approach. *arXiv:1807.02884*, 2018.

[15] Tamara G Kolda and Brett W Bader. Tensor decompositions and applications. *SIAM review*, 51(3):455–500, 2009.

[16] Chung-Yi Lin, I Eli Chien, and I-Hsiang Wang. On the fundamental statistical limit of community detection in random hypergraphs. In *2017 IEEE International Symposium on Information Theory (ISIT)*, pages 2178–2182. IEEE, 2017.

[17] Elchanan Mossel, Joe Neeman, and Allan Sly. Reconstruction and estimation in the planted partition model. *Probab. Theory Relat. Fields*, 162(3-4):431–461, 2015.

[18] Mingao Yuan, Ruiqi Liu, Yang Feng, and Zuofeng Shang. Testing community structures for hypergraphs. *arXiv preprint arXiv:1810.04617*, 2018.

[19] Mingao Yuan, Ruiqi Liu, Yang Feng, and Zuofeng Shang. Testing community structures for hypergraphs. *arXiv:1810.04617*, 2018.

A Appendix