The Annals of Statistics
2017, Vol. 45, No. 1, 289–315
DOI: 10.1214/16-AOS1453
© Institute of Mathematical Statistics, 2017

CONSISTENCY OF SPECTRAL HYPERGRAPH PARTITIONING UNDER PLANTED PARTITION MODEL

BY DEBARGHYA GHOSHDASTIDAR AND AMBEDKAR DUKKIPATI

Indian Institute of Science

Hypergraph partitioning lies at the heart of a number of problems in machine learning and network sciences. Many algorithms for hypergraph partitioning have been proposed that extend standard approaches for graph partitioning to the case of hypergraphs. However, theoretical aspects of such methods have seldom received attention in the literature as compared to the extensive studies on the guarantees of graph partitioning. For instance, consistency results of spectral graph partitioning under the stochastic block model are well known. In this paper, we present a planted partition model for sparse random nonuniform hypergraphs that generalizes the stochastic block model. We derive an error bound for a spectral hypergraph partitioning algorithm under this model using matrix concentration inequalities. To the best of our knowledge, this is the first consistency result related to partitioning nonuniform hypergraphs.

1. Introduction. A wide variety of complex real-world systems can be understood by analyzing the interactions among various entities or components of the system. This has made network analysis a subject of both theoretical and practical interest. A plethora of challenging problems related to social, biological, communication networks have intrigued researchers over the past decades, and has led to the development of some sophisticated techniques for network analysis. This is clearly witnessed in the problems related to network or graph partitioning, where the task is to find strongly connected groups of nodes with sparse connections across groups. The problem appears in several engineering applications such as circuit or program segmentation [Kernighan and Lin (1970)], community detection in social or biological networks [Guimera and Amaral (2005), Wasserman (1994)], data analysis and clustering [Ng, Jordan and Weiss (2002)] among others.

Graph partitioning and the stochastic block model. The problem of finding a balanced partition of a graph is known to be computationally hard. However, a number of approximate methods have been studied in the literature. These include spectral algorithms [Fiedler (1973), Krzakala et al. (2013), Ng, Jordan and Weiss (2002)], modularity and likelihood based methods [Bickel and Chen (2009),

Received May 2015; revised February 2016.
MSC2010 subject classifications. Primary 62H30, 62F12; secondary 05C65, 05C80.
Key words and phrases. Hypergraph, stochastic block model, spectral algorithm.

Choi, Wolfe and Airoldi (2012), Girvan and Newman (2002)], convex optimization [Amini and Levina (2014), Chen, Sanghavi and Xu (2014)], belief propagation [Decelle et al. (2011)] among others. The empirical success of such methods is not a mere coincidence, and theoretical guarantees for most of these methods have been extensively studied. In this respect, it is quite common to study partitioning algorithms under statistical models for random networks, such as the stochastic block model or planted partition model [Holland, Laskey and Leinhardt (1983)] or its variants. In this model, one considers a random graph on n nodes with a well defined k-way partition $\psi : \{1, \dots, n\} \to \{1, \dots, k\}$. The edges are randomly added with probabilities depending on the class labels of the participating nodes. Thus, the following interesting question arises.

QUESTION. Let ψ' be the partition obtained from an algorithm, then what is the number of mismatches between ψ and ψ'?

One typically asks for a high probability bound on the above error in terms of n. Such error bounds have been established for a variety of partitioning algorithms including aforementioned approaches. Chen, Sanghavi and Xu (2014) compare the theoretical guarantees for different approaches.

In the case of random graphs under stochastic block model, analysis of a spectral algorithm was first considered by McSherry (2001). However, the popular variant of spectral graph partitioning, commonly known as *spectral clustering*, was studied only in recent times [Lei and Rinaldo (2015), Rohe, Chatterjee and Yu (2011)]. It is now known that for a planted graph with $\Omega(\ln n)$ minimum node degree, spectral clustering achieves an $o(n)$ error rate. One commonly refers to this property as the *weak consistency* of spectral clustering. This not the best known error rate as exact recovery of the partitions are known to be possible using other approaches [Amini and Levina (2014)]. Recent results [Gao et al. (2015), Lei and Zhu (2014), Vu (2014)] show that an additional refinement process can improve the partitioning of spectral clustering to exactly recover the partitions, thereby achieving *strong consistency*. The condition on minimum node degree can be also relaxed by considering alternative spectral techniques [Krzakala et al. (2013), Le, Levina and Vershynin (2015)], and these algorithms can detect partitions in sparse random graphs that are close to the algorithmic barrier for community detection [Decelle et al. (2011)].

Hypergraph partitioning. In spite of the vast applicability of network modeling and analysis, there exists more complex scenarios, where pairwise interactions cannot accurately model the system of interest. A common example is folksonomy, where individuals annotate online resources, such as images or research papers. Such problems appear to have a tri-partite structure in form of "user–resource–annotation", and is naturally represented as a 3-uniform hypergraph [Ghoshal et al. (2009)], where each edge connects three nodes. Earlier works in data mining

[Gibson, Kleinberg and Raghavan (2000)] as well as in computer vision [Govindu (2005)] have also demonstrated the necessity of uniform hypergraphs. Moreover, in large scale circuit design [Karypis and Kumar (2000)] and molecular interaction networks [Michoel and Nachtergaele (2012)], one needs to consider group interactions that is appropriately modeled by a nonuniform hypergraph.

The current work focuses on hypergraph partitioning that appears in various applications such as circuit partitioning [Schweikert and Kernighan (1979)], categorical data clustering [Gibson, Kleinberg and Raghavan (2000)], geometric grouping [Govindu (2005)] and others. Various partitioning techniques are used in practice including move-based algorithms [Karypis and Kumar (2000), Schweikert and Kernighan (1979)], spectral algorithms [Rodríguez (2002), Zhou, Huang and Schölkopf (2007)], tensor based methods [Govindu (2005)], etc.

Hypergraph partitioning and related problems have been of theoretical interest for quite some time [Berge (1984)]. While early works on hypergraph partitioning studied various properties of hypergraph cuts [Bolla (1993), Chung (1993)], more recent results provide insights into the algebraic connectivity and chromatic numbers of hypergraphs [Cooper and Dutle (2012), Hu and Qi (2012)]. However, to date, little is known about the theoretical guarantees of hypergraph partitioning methods that are popular amongst practitioners. The primary reason for this lack of results, at least in a stochastic framework, is due to the absence of random models of nonuniform hypergraphs that can incorporate a planted structure. Planted structures in uniform hypergraphs have been studied in context of hypergraph coloring [Chen and Frieze (1996)], and in a more general setting in Ghoshdastidar and Dukkipati (2014), where almost sure error bounds are derived for partitioning planted uniform hypergraphs via tensor decomposition. But, extension of such models or analysis to nonuniform hypergraphs does not follow directly.

Contributions. The primary focus of this paper is to derive an error bound for a hypergraph partitioning algorithm that solves a spectral relaxation of the normalized hypergraph cut problem. This is achieved in the form of a two-fold contribution.

We present a model for generating random hypergraphs with a planted solution. Extensions of the Erdős–Rényi model to nonuniform hypergraphs have been studied in the literature [Darling and Norris (2005), Schmidt-Pruzan and Shamir (1985)], where, for each m, the probability of generating edges of size m is controlled by a parameter p_m. The recent work of Stasi et al. (2014) present a similar model, but with a specified degree sequence. Such models implicitly suggest that one can consider a nonuniform hypergraph as a collection of m-uniform hypergraphs for varying m. Thus, it is possible to construct planted models for nonuniform hypergraphs from a collection of uniform hypergraph models. Based on this idea, we present a planted hypergraph model that naturally extends the sparse stochastic block model for graphs [Lei and Rinaldo (2015)], and also encompasses previously studied models for uniform hypergraphs [Ghoshdastidar and Dukkipati (2014)].

We consider a popular spectral algorithm for hypergraph partitioning, and derive a bound on the number of nodes incorrectly assigned by the algorithm under the above model. We prove that for random planted hypergraphs with minimum node degree above a certain threshold, the algorithm is weakly consistent in general. However, the algorithm can also exactly recover the partitions from dense hypergraphs without any subsequent refinement procedure. Our analysis relies on an alternative characterization of the incidence matrix of the random hypergraph, and the use of matrix concentration inequalities [Chung and Radcliffe (2011), Tropp (2012)].

Typically, spectral partitioning algorithms involve a post-processing stage of distance based clustering. Though the k-means algorithm [Lloyd (1982)] or its approximate variants [Kumar, Sabharwal and Sen (2004), Ostrovsky et al. (2012)] are the popular choice in practice, such algorithms are not always guaranteed to provide good clustering. Gao et al. (2015) discusses the implication of this drawback on the consistency results for spectral clustering [Lei and Rinaldo (2015)]. On the other hand, we establish that under certain conditions, the approximate k-means algorithm of Ostrovsky et al. (2012) indeed provides a good clustering with very high probability.

Finally, we consider special cases of the planted model. We comment on the allowable model parameters, and illustrate their effect on the derived error bound. Numerical studies reveal the practical significance of spectral hypergraph partitioning as well as the applicability of our analysis.

Organization. We first describe the spectral hypergraph partitioning algorithm under consideration in Section 2. Section 3 describes the model for random hypergraphs with a planted partition. We provide the main consistency result in Section 4, followed by a series of examples of planted models studied in Section 5. Section 6 contains experimental results that validate our model and analysis, and Section 7 presents the concluding remarks. The proofs of the technical lemmas can be found in the supplement to this paper [Ghoshdastidar and Dukkipati (2016)].

Notations. Some of the notation that is often used in this paper are mentioned here. $\mathbb{1}\{\cdot\}$ is the indicator function and $\ln(\cdot)$ refers to the natural logarithm. $\mathsf{E}[\cdot]$ denotes expectation with respect to the distribution of the planted model. For a matrix A, we use $A_{i\cdot}$ to refer to the ith row of A and $A_{\cdot i}$ refers to its ith column. $\|\cdot\|_2$ denotes the Euclidean norm for vectors and the spectral norm for matrices, while $\|\cdot\|_F$ denotes the Frobenius norm. We sometimes compute standard matrix functions like $\mathrm{Trace}(\cdot)$ and $\det(\cdot)$. In addition, we also use asymptotic notation $O(\cdot), o(\cdot), \Omega(\cdot)$, etc., where we view these quantities as functions of the number of nodes n.

2. Spectral hypergraph partitioning. A hypergraph is defined as a tuple $(\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ is a set of objects and $\mathcal{E}$ is a collection of subsets of $\mathcal{V}$. Though early works in combinatorics viewed this structure purely as a set system, it was soon realized that one may view $\mathcal{V}$ as a set of nodes and every element of $\mathcal{E}$ as an edge (or connection) among a subset of nodes. As noted in Berge (1984), such a generalization of graphs helps to simplify several combinatorial results in the graph literature. A hypergraph is said to be r-uniform if every edge $e \in \mathcal{E}$ contains exactly r nodes.

In this paper, we assume that there are no edges of size 0 or 1 as they do not convey any information in a partitioning framework. We also assume that the hypergraph is undirected, that is, there is no ordering of nodes in any edge. Under this setting, the most simple representation of a hypergraph is in terms of its incidence matrix $H \in \{0, 1\}^{|\mathcal{V}| \times |\mathcal{E}|}$, where $H_{ve} = 1$ if the node v is contained in the edge e, and 0 otherwise. One can note that the degree of any node v can be written as $\deg(v) = \sum_{e \in \mathcal{E}} H_{ve}$, which is simply the sum of the vth row of H. Similarly, the cardinality of any edge e is $|e| = \sum_{v \in \mathcal{V}} H_{ve}$.

Several notions of hypergraph cut and hypergraph Laplacian have been proposed in the literature [Bolla (1993), Chung (1993), Rodríguez (2002)] that generalize the standard notions well studied in the graph literature. In this work, we consider the generalization studied in Zhou, Huang and Schölkopf (2007). Let $\mathcal{V}_1 \subset \mathcal{V}$, then $\mathrm{vol}(\mathcal{V}_1) = \sum_{v \in \mathcal{V}_1} \deg(v)$ is called the volume of $\mathcal{V}_1$, while the boundary of $\mathcal{V}_1$, defined as $\partial \mathcal{V}_1 = \{e \in \mathcal{E} : e \cap \mathcal{V}_1 \neq \phi, e \cap \mathcal{V}_1^c \neq \phi\}$, denotes the set of edges that are cut when the nodes are divided into $\mathcal{V}_1$ and $\mathcal{V}_1^c = \mathcal{V} \setminus \mathcal{V}_1$. The volume of $\partial \mathcal{V}_1$ is defined as

$$\mathrm{vol}(\partial \mathcal{V}_1) = \sum_{e \in \partial \mathcal{V}_1} \frac{|e \cap \mathcal{V}_1||e \cap \mathcal{V}_1^c|}{|e|}.$$

We consider the problem of partitioning the vertex set $\mathcal{V}$ into k disjoint sets, $\mathcal{V}_1, \ldots, \mathcal{V}_k$, that minimizes the normalized hypergraph cut

$$\text{NH-cut}(\mathcal{V}_1, \ldots, \mathcal{V}_k) = \sum_{j=1}^{k} \frac{\mathrm{vol}(\partial \mathcal{V}_j)}{\mathrm{vol}(\mathcal{V}_j)}. \tag{1}$$

One can observe that for graphs, the above definition (1) retrieves the standard notion of a normalized cut [von Luxburg (2007)]. Zhou, Huang and Schölkopf (2007) also define the notion of a normalized hypergraph Laplacian matrix $L \in \mathbb{R}^{|\mathcal{V}| \times |\mathcal{V}|}$ given by

$$L = I - D^{-1/2} H \Delta^{-1} H^T D^{-1/2}, \tag{2}$$

where the matrices $D \in \mathbb{R}^{|\mathcal{V}| \times |\mathcal{V}|}$, $\Delta \in \mathbb{R}^{|\mathcal{E}| \times |\mathcal{E}|}$ are diagonal with $D_{vv} = \deg(v)$ and $\Delta_{ee} = |e|$. A simple calculation shows that the problem of minimizing the quantity in (1) is equivalent to the problem

$$\underset{\mathcal{V}_1, \ldots, \mathcal{V}_k}{\text{minimize}} \ \mathrm{Trace}(\widehat{X}^T L \widehat{X}), \tag{3}$$

Algorithm 1 Spectral hypergraph partitioning algorithm

```
input Incidence matrix $H$ of the hypergraph.
1: Compute the hypergraph Laplacian $L$ as in (2).
2: Compute the leading eigenvector matrix $X \in \mathbb{R}^{|\mathcal{V}|\times k}$.
3: Normalize rows of $X$ to have unit norm. Call this matrix $\overline{X}$.
4: Run $k$-means on the rows of $\overline{X}$.
output Partition of $\mathcal{V}$ that corresponds to the clusters obtained from $k$-means.
```

where $\widehat{X} \in \mathbb{R}^{|\mathcal{V}|\times k}$ is such that $\widehat{X}_{vj} = \sqrt{\frac{\deg(v)}{\operatorname{vol}(\mathcal{V}_j)}}\mathbb{1}\{v \in \mathcal{V}_j\}$, and satisfies $\widehat{X}^T\widehat{X} = I$. Since the optimization in (3) is NP-hard, one relaxes the problem by minimizing over all $X \in \mathbb{R}^{|\mathcal{V}|\times k}$ with orthonormal columns. It is well known that the solution to this relaxed problem is the matrix of k leading orthonormal eigenvectors of L. Note that L is a positive semi-definite matrix with at least one eigenvalue equal to zero. The term "leading eigenvectors" refers to the eigenvectors that correspond to the k smallest eigenvalues of L.

The above discussion motivates a spectral k-way partitioning approach based on minimizing NH-cut. The method is listed in Algorithm 1. The form of Laplacian matrix in (2) also suggests that the problem of minimizing NH-cut may be alternatively expressed as the problem of partitioning a graph with weighted adjacency matrix

$$A = H\Delta^{-1}H^T. \tag{4}$$

Such a graph is related to the star expansion of the hypergraph [Agarwal, Branson and Belongie (2006)].

The intuition behind the k-means step in Algorithm 1 is as follows. If the solution of the spectral relaxation results in $X = \widehat{X}$, where $\widehat{X}$ is defined as in (3), then after row normalization, $\overline{X}$ corresponds to a binary matrix with exactly one nonzero term in each row. Hence, one obtains the partitions desired in (3) by performing k-means on the rows of $\overline{X}$. In this paper, we assume that the approximate k-means method of Ostrovsky et al. (2012) is used that provides a near optimal solution in a single iteration.

3. Planted partition in random hypergraphs.

In the rest of the paper, we study the error incurred by Algorithm 1. For this, we consider a model for generating random hypergraphs with a planted solution.

3.1. *The model.*

Let $\mathcal{V} = \{1, 2, \ldots, n\}$ be a set of nodes, and let $\psi : \{1, 2, \ldots, n\} \to \{1, 2, \ldots, k\}$ be a partition of the nodes into k classes. Here, ψ is the unknown planted partition that one needs to extract from a hypergraph generated on $\mathcal{V}$. For a node i, we denote its class by ψ_i. Let $M \geq 2$ be an integer, representing the range or the maximum edge cardinality in the hypergraph. In view of practical situations, we allow both k and M to vary with n though this dependence is not made explicit in the notation. One may set $M = n$ to allow occurrence of all possible edges, but in practice, one can assume that $M = O(\ln n)$. Also, for each n and for each $m = 2, \ldots, M$, let $\alpha_{m,n} \in [0, 1]$, and $B^{(m)} \in [0, 1]^{k \times k \times \cdots \times k}$ be a symmetric k-dimensional tensor of order m.

A random hypergraph on $\mathcal{V}$ is generated as follows. For each $m = 2, \ldots, M$, and for every set $\{i_1, i_2, \ldots, i_m\} \subset \mathcal{V}$, an edge is included independently with probability $\alpha_{m,n} B^{(m)}_{\psi_{i_1}\psi_{i_2}\cdots\psi_{i_m}}$. This process generates a random hypergraph of maximum edge cardinality M. The tensor $B^{(m)}$ contains the probabilities of forming m-way edges among the different classes if $\alpha_{m,n} = 1$. On the other hand, $\alpha_{m,n}$ allows for a sparsity scaling that does not depend on the partitions. In the case of sparse graphs, $\alpha_{2,n}$ regulates the edge density. However, in real-world nonuniform hypergraphs, one often finds than the density of 2 or 3-way edges is much more than edges of larger size (say, 10). To account for this generality, we allow $\alpha_{m,n}$ to vary both with m and n. For instance, if $\alpha_{2,n} = 1$ and $\alpha_{m,n} = \frac{1}{n^{m-1}}$ for all $m > 2$, then the generated hypergraph contains $O(n^2)$ number of 2-way edges, but only $O(n)$ number of m-way edges for every $m > 2$.

As a special case, note that for graphs, $M = 2$ for all n, and the model corresponds to the sparse stochastic block model, where an edge (u, v) is formed with probability $\alpha_{2,n} B^{(2)}_{\psi_u\psi_v}$. In other words, if $Z \in \{0, 1\}^{n \times k}$ denotes the assignment matrix, then the probability of edge (u, v) is same as the corresponding entry of $\alpha_{2,n} Z B^{(2)} Z^T$. For r-uniform uniform hypergraphs, one has $\alpha_{m,n} = 0$ for all $m \neq r$. Ghoshdastidar and Dukkipati (2014) considered a dense uniform hypergraph, that is, $\alpha_{r,n} = 1$, and edge probabilities specified by a rth-order k-dimensional tensor $B^{(r)}$. It was shown that the population adjacency tensor can be expressed in terms of $B^{(r)}$ and Z.

The intuition behind the above described model is that one may view a hypergraph of range M as a collection of uniform hypergraphs of orders $m = 2, \ldots, M$. In the random setting, each m-uniform hypergraph is specified in terms of $\alpha_{m,n}$ and $B^{(m)}$. The above model can be easily extended to directed hypergraphs, and also to the case of weighted hypergraphs.

3.2. *The random hypergraph Laplacian.* For the stochastic block model, a random instance of a graph is specified by its $n \times n$ adjacency matrix. However, for random hypergraphs, the size of the incidence matrix H is a random quantity as it depends on the number of generated edges. Since this poses difficulties in working with the form of hypergraph Laplacian in (2), we present an alternative representation. The Laplacian can be written as

$$L = I - \sum_{e \in \mathcal{E}} \frac{1}{|e|} D^{-1/2} a_e a_e^T D^{-1/2}, \tag{5}$$

where for $e \subset \mathcal{E}$, $a_e \in \{0, 1\}^n$ with $(a_e)_i = 1$, if node $i \in e$, and 0 otherwise.

Let $\beta_M = \sum_{m=2}^{M} \binom{n}{m}$. Note that β_M is the maximum number of edges the hypergraph can contain given the fact that its range is M. For convenience, we define a bijective map $\xi : \{1, 2, \ldots, \beta_M\} \to \{e \subset \mathcal{V} : 2 \leq |e| \leq M\}$, where each ξ_j refers to a subset of nodes, that is, a possible edge in the given hypergraph. Then the Laplacian can be expressed as

$$L = I - \sum_{j=1}^{\beta_M} \frac{\mathbb{1}\{\xi_j \in \mathcal{E}\}}{|\xi_j|} D^{-1/2} a_{\xi_j} a_{\xi_j}^T D^{-1/2}, \tag{6}$$

where the summation is over all possible edges of size at most M, but the missing edges do not contribute to the sum. Similarly, one can express the degree matrix D as

$$D_{ii} = \deg(i) = \sum_{e \in \mathcal{E}} (a_e)_i = \sum_{j=1}^{\beta_M} \mathbb{1}\{\xi_j \in \mathcal{E}\}(a_{\xi_j})_i. \tag{7}$$

The above representation corresponds to an ‘extended’ version of the incidence matrix as $\overline{H} \in \{0, 1\}^{n \times \beta_M}$, whose jth column is $\mathbb{1}\{\xi_j \in \mathcal{E}\} a_{\xi_j}$, that is, $\overline{H}$ contains the columns of H with additional zero columns inserted to account for missing edges. This holds for any hypergraph of range M defined on the set $\mathcal{V}$. We use this representation to keep the number of columns as a deterministic quantity. We now discuss how the described planted partition model for hypergraphs, with maximum edge size M, can be expressed in terms of the extended incidence matrix $\overline{H} \in \{0, 1\}^{n \times \beta_M}$. Let h_j, $j = 1, 2, \ldots, \beta_M$ be independent Bernoulli random variables that indicate the presence of the edge $\xi_j \subset \mathcal{V}$. By description of the model, if $\xi_j = \{i_1, i_2, \ldots, i_{m_j}\}$ for some $m_j \in \{2, \ldots, M\}$, then the random variable $h_j \sim \text{Bernoulli}(\alpha_{m_j,n} B^{(m_j)}_{\psi_{i_1}\psi_{i_2}\cdots\psi_{i_{m_j}}})$. The jth column of $\overline{H}$ is $h_j a_{\xi_j}$, and hence, the Laplacian matrix for the random hypergraph is

$$L = I - \sum_{j=1}^{\beta_M} \frac{h_j}{|\xi_j|} D^{-1/2} a_{\xi_j} a_{\xi_j}^T D^{-1/2} \qquad \text{where } D_{ii} = \sum_{j=1}^{\beta_M} h_j (a_{\xi_j})_i. \tag{8}$$

At this stage, we note that the above matrices depend on the number of nodes n. For ease of notation, we do not explicitly mention this dependence.

4. Consistency of spectral hypergraph partitioning. This section presents the main result of this paper that gives a bound on the error incurred by the spectral hypergraph partitioning algorithm described in Algorithm 1. Let $\psi' : \{1, \ldots, n\} \to \{1, \ldots, k\}$ denote the labels obtained from the algorithm. The partitioning error is given by the number of nodes incorrectly assigned by Algorithm 1, that is,

$$\mathsf{Err}(\psi, \psi') = \min_{\sigma} \sum_{i=1}^{n} \mathbb{1}\{\psi_i \neq \sigma(\psi'_i)\}, \tag{9}$$

where the minimum is taken over all permutation σ of labels. We show that if (i) the partitions are *identifiable*, and (ii) the hypergraph is not too *sparse*, then indeed $\mathsf{Err}(\psi, \psi')$ is bounded by a quantity that is at most sub-linear in n. Furthermore, the bound holds with probability $(1 - o(1))$. This immediately implies that Algorithm 1 is *weakly consistent*. However, we show later that for particular model parameters, Algorithm 1 can even recover the partitions exactly, that is, $\mathsf{Err}(\psi, \psi') = o(1)$.

4.1. *The main result.* The consistency result studied in this paper is quite similar, in spirit, to those studied in the case of stochastic block model for graphs. In such a case, one typically analyzes the population version of a spectral algorithm, and then uses the fact that the spectral properties of the Laplacian eventually concentrates around those of the population Laplacian.

From this point of view, we consider the population version of the hypergraph Laplacian (8) defined as

$$\mathcal{L} = I - \sum_{j=1}^{\beta_M} \frac{\mathsf{E}[h_j]}{|\xi_j|} \mathcal{D}^{-1/2} a_{\xi_j} a_{\xi_j}^T \mathcal{D}^{-1/2}, \tag{10}$$

where $\mathcal{D}$ is the expected degree matrix, that is, $\mathcal{D}_{ii} = \sum_{j=1}^{\beta_M} \mathsf{E}[h_j](a_{\xi_j})_i$. We also define the quantity $d = \min_{i \in \{1,\dots,n\}} \mathcal{D}_{ii}$. Without loss of generality, we may also assume that for a given n, the community sizes are $n_1 \geq n_2 \geq \cdots \geq n_k$.

Before stating the main result, it is useful to elaborate on the aforementioned conditions under which the derived error bound holds. A lower bound on the sparsity of the hypergraph is a standard requirement to ensure that the concentration of the spectral properties eventually hold, and has been often used in the graph literature [Le, Levina and Vershynin (2015), Lei and Rinaldo (2015)]. In our setting, this can be stated in terms of the sparsity factors $\alpha_{m,n}$, or more simply, in terms of the minimum expected degree d, that grows with n but at a rate controlled by the sparsity factors.

A more critical condition is the identifiability of the partitions. Note that the definition of the hypergraph Laplacian essentially implies that the hypergraph is reduced to a graph with self-loops. Hence, the performance of Algorithm 1 crucially depends on the identifiability of the partitions from $\mathcal{L}$, or rather from this reduced graph with the population adjacency matrix

$$\mathcal{A} = \mathsf{E}[A] = \sum_{j=1}^{\beta_M} \frac{\mathsf{E}[h_j]}{|\xi_j|} a_{\xi_j} a_{\xi_j}^T.$$

The following result provides a characterization of $\mathcal{L}$ and $\mathcal{A}$, which in turn helps to quantify the condition for identifiability of the partitions from $\mathcal{L}$.

LEMMA 4.1. *Let $Z \in \{0,1\}^{n\times k}$ denote the assignment matrix corresponding to the partition ψ. Then the population hypergraph Laplacian is given by*

$$\mathcal{L} = I - \mathcal{D}^{-1/2}\mathcal{A}\mathcal{D}^{-1/2}, \tag{11}$$

where $\mathcal{A}$ can be expressed as

$$\mathcal{A} = ZGZ^T - J. \tag{12}$$

Here, $J \in \mathbb{R}^{n\times n}$ is diagonal with $J_{ii} = J_{jj}$ whenever $\psi_i = \psi_j$, and $G \in \mathbb{R}^{k\times k}$.

Furthermore, $\mathcal{L}$ contains k eigenvalues for which the corresponding orthonormal eigenvectors are the columns of the matrix $\mathcal{X} = Z(Z^TZ)^{-1/2}U$, where $U \in \mathbb{R}^{k\times k}$ is orthonormal.

The representation in (12) shows that $\mathcal{A}$ is essentially of rank k, except for the diagonal entries. Owing to the first term in (12), one does expect $\mathcal{L}$ to have k eigenvectors whose entries are constant in each community. As discussed later, a close inspection of $\mathcal{X}$ reveals that indeed the columns of $\mathcal{X}$ satisfy this property. Thus, if the spectral stage of Algorithm 1 can extract $\mathcal{X}$, then zero error can be achieved from the k-means step.

In general, $\mathcal{X}$ need not correspond to leading eigenvectors $\mathcal{L}$ (as computed in Algorithm 1). This is true even for certain types of graphs, for instance k-colorable graphs [Alon and Kahale (1997)]. This effect is more pronounced in nonuniform hypergraphs due to the presence of a large number of model parameters. To account for this factor, we define the following quantity:

$$\delta = \left(\lambda_{\min}(G) \min_{1\le i\le n} \frac{n_{\psi_i}}{\mathcal{D}_{ii}}\right) - \max_{1\le i,j\le n}\left|\frac{J_{ii}}{\mathcal{D}_{ii}} - \frac{J_{jj}}{\mathcal{D}_{jj}}\right|, \tag{13}$$

where n_{ψ_i} is the size of the community in which node i belongs. We show that if $\delta > 0$, then the columns of $\mathcal{X}$ are the k leading eigenvectors of $\mathcal{L}$. Here, $\lambda_{\min}(G)$ refers to the smallest eigenvalue of G. Thus, we can state the consistency result for Algorithm 1 as below.

THEOREM 4.2. *Consider a random hypergraph on n nodes generated according to the planted partition model described in Section 3. Assume that n is sufficiently large, and the size of the k partitions are $n_1 \ge n_2 \ge \cdots \ge n_k$. Let d be the minimum expected degree, and δ be the quantity defined in (13).*

There exists an absolute constant $C > 0$, such that, if $\delta > 0$ and

$$d > C\frac{kn_1(\ln n)^2}{\delta^2 n_k} \tag{14}$$

then with probability at least $1 - O((\ln n)^{-1/4})$,

$$\mathsf{Err}(\psi, \psi') = O\left(\frac{kn_1 \ln n}{\delta^2 d}\right). \tag{15}$$

Note here that the quantities δ, d and k can vary with n. On substituting the condition on d into (15), one can see that $\mathsf{Err}(\psi, \psi') = o(n)$ with probability $(1 - o(1))$. Hence, Algorithm 1 is weakly consistent if the conditions of the theorem are satisfied. However, we show later that in certain dense hypergraphs, the bound in (15) may eventually decay to zero. Thus, Algorithm 1 is guaranteed to exactly recover the communities in such cases.

In Section 5, we consider particular instances of the planted model, and illustrate the dependance of the above result on the model parameters. For instance, (14) implies that the result holds if the sparsity factor $(\alpha_{m,n})$ is above a certain threshold (see Corollaries 5.1 and 5.2). Even when (14) holds, higher error is incurred for a sparse hypergraph (small d) or when the number of communities k is large.

One may note that $\delta > 0$ is the condition for identifiability of the partitions, and is essential for success of the algorithm. Typically, one does find that $\delta \downarrow 0$ as $n \to \infty$. To this end, the condition (14) implies that δ cannot decay rapidly as $\delta^2 d$ needs to maintain a minimum growth rate. We also note that δ quantifies identifiability of the partitions and $\mathsf{Err}(\psi, \psi')$ varies as $\frac{1}{\delta^2}$. Hence, if the model parameters are such that δ is small, for instance if the probability of inter-community edges is very close to that of within community edges, then $\mathsf{Err}(\psi, \psi')$ is larger.

Before presenting the proof of Theorem 4.2, we comment on the assumption of sufficiently large n. Note that the sole purpose of this assumption is to ensure the success of the k-means algorithm. Later, in the proof, we establish that if n is large enough, the condition (14) ensures that the approximate k-means method of Ostrovsky et al. (2012) provides a near optimal solution, which is worse by only a constant factor. Earlier works on spectral graph partitioning [Lei and Rinaldo (2015), Rohe, Chatterjee and Yu (2011)] assumed the existence of such a near optimal solution with probability 1. To demonstrate the effect of such an assumption, we state the following result, which is a modification of Theorem 4.2 under the above assumption.

COROLLARY 4.3. *Consider a random hypergraph on n nodes generated according to the planted partition model, and let the other quantities be as defined in Theorem 4.2. Assume that for a constant $\gamma > 1$, there is a γ-approximate*[1] *k-means algorithm that succeeds with probability* 1.

There exists an absolute constant $C > 0$, such that, if $\delta > 0$ and

$$d > C\frac{\ln n}{\delta^2} \tag{16}$$

then with probability at least $1 - \frac{4}{n^2}$,

$$\mathsf{Err}(\psi, \psi') = O\left(\frac{kn_1 \ln n}{\delta^2 d}\right). \tag{17}$$

[1] Informally, a γ-approximate k-means methods returns a solution for which the objective of the k-means problem is at most γ times the global minimum, where $\gamma > 1$. A formal definition is postponed to (22) and subsequent discussions.

The result reveals that if a good k-means algorithm is available, then the success probability of Algorithm 1 increases, and the result is also applicable for more sparse hypergraph since the condition (16) is weaker than (14). However, based on the existing results in the k-means literature, one should consider the following remark.

REMARK. If the data satisfies certain *clusterability* criterion,[2] then the efficient variants of k-means [Kumar, Sabharwal and Sen (2004), Ostrovsky et al. (2012)] provide a γ-approximate solution with a constant probability $\rho < 1$. Both γ and ρ depend on various factors including k, clusterability criterion, etc.

In view of the above remark, Corollary 4.3 is too optimistic. Recently, Gao et al. (2015) pointed that if one uses the method of Kumar, Sabharwal and Sen (2004), then γ grows with k. In addition, one should also note that the success probability of this method is $\rho = c^k$ for an absolute constant $c \in (0, 1)$. Hence, a spectral partitioning algorithm using this method cannot succeed with probability $(1 - o(1))$. Instead, we use the method of Ostrovsky et al. (2012) to achieve a higher success rate as stated in Theorem 4.2. The only additional assumption is that of sufficiently large n. We note that this requirement, along with condition (14), can be relaxed if one only aims for a constant success probability. This is shown in the following modification of Theorem 4.2, where we assume that the k-means algorithm of Ostrovsky et al. (2012) is used.

COROLLARY 4.4. *Consider a random hypergraph on n nodes generated according to the planted partition model, and let the other quantities be as defined in Theorem 4.2.*

There exist absolute constants $C > 0$ and $\varepsilon \in (0, 0.015)$, such that, if $\delta > 0$ and

$$d > \frac{C}{\varepsilon^2}\frac{kn_1 \ln n}{\delta^2 n_k} \tag{18}$$

then with probability at least $1 - O(\sqrt{\varepsilon})$,

$$\mathsf{Err}(\psi, \psi') = O\left(\frac{kn_1 \ln n}{\delta^2 d}\right). \tag{19}$$

4.2. *Proof of Theorem 4.2.* We now present an outline of the proof of Theorem 4.2 using a series of lemmas. The proofs for these lemmas are given in the supplementary material [Ghoshdastidar and Dukkipati (2016)]. The result is obtained by proving the following facts:

[2] Various clusterability criteria have been studied in the literature. In this work, we consider the notion of ε-separability proposed by Ostrovsky et al. (2012).

1. If Algorithm 1 is performed on the population Laplacian $\mathcal{L}$, then under the condition of $\delta > 0$, the obtained partitions are correct.
2. The deviation of L from $\mathcal{L}$ is bounded above, and the bound holds with probability at least $(1 - \frac{4}{n^2})$.
3. As a consequence of above facts, the standard matrix perturbation bounds [Stewart and Sun (1990)] imply that the eigenvalues and the corresponding eigenspaces of L concentrate about those of $\mathcal{L}$.
4. If (14) holds, then k-means stage of Algorithm 1 succeeds in obtaining a near optimal solution with probability at least $1 - O((\ln n)^{-1/4})$.
5. The partitioning error can be expressed in terms of the above bounds, which leads to (15).

Corollaries 4.3 and 4.4 can be proved in similar manner. This is discussed in the supplementary material [Ghoshdastidar and Dukkipati (2016)]. We now prove the above facts. The following result extends Lemma 4.1.

LEMMA 4.5. *If $\delta > 0$, then the k leading orthonormal eigenvectors of $\mathcal{L}$ correspond to the columns of the matrix $\mathcal{X} = Z(Z^T Z)^{-1/2}U$.*

In the above result, $Z^T Z$ is a diagonal matrix with entries being the sizes of the k partitions. Hence, both $Z^T Z$ and U are of the rank k. Due to this, one can observe that the matrix $\mathcal{X}$ contains exactly k distinct rows, each corresponding to a particular partition, that is, if $A_{i\cdot}$ denotes ith row of a matrix A, then for any two nodes $i, j \in \mathcal{V}$,

$$\mathcal{X}_{i\cdot} = \mathcal{X}_{j\cdot} \quad \Longleftrightarrow \quad Z_{i\cdot} = Z_{j\cdot} \quad \Longleftrightarrow \quad \psi_i = \psi_j.$$

Moreover, since U is orthonormal, the distinct rows of $\mathcal{X}$ are orthogonal. Hence, after row normalization, the distinct rows correspond to k orthonormal vectors in $\mathbb{R}^k$, which can be easily clustered by k-means algorithm to obtain the true communities. Technically, δ is a lower bound on the eigen-gap between the kth and $(k+1)$th smallest eigenvalues of $\mathcal{L}$. Since, it is difficult to obtain a simple characterization of the eigen-gap, we resort to the use of δ as defined in (13).

Next, we bound the deviation of a random instance of L from the population Laplacian $\mathcal{L}$. This bound relies on the use of matrix Bernstein inequality [Chung and Radcliffe (2011), Tropp (2012)]. We note that for graphs, sharp deviation bounds have been used [Lei and Rinaldo (2015)], but such techniques cannot be directly extended to the case of hypergraphs.

LEMMA 4.6. *If $d > 9 \ln n$, then with probability at least $(1 - \frac{4}{n^2})$,*

$$\|L - \mathcal{L}\|_2 \le 12\sqrt{\frac{\ln n}{d}}. \tag{20}$$

We now use the principle subspace perturbation result due to Lei and Rinaldo (2015) to comment on the deviation of the leading eigenvectors of L from those of $\mathcal{L}$. A modified version of their result is proved that incorporates the row normalization of the eigenvector matrix. Let X be the matrix of the k leading eigenvectors of L, and $\overline{X}$ be its row normalized version. We have the following result. Note that since $\delta < 1$, the condition of Lemma 4.6 is subsumed by the condition stated below.

LEMMA 4.7. *If $\delta > 0$ and $d > \frac{576 \ln n}{\delta^2}$, then there is an orthonormal matrix $Q \in \mathbb{R}^{k\times k}$ such that*

$$\|\overline{X} - ZQ\|_F \leq \frac{24}{\delta}\sqrt{\frac{2kn_1 \ln n}{d}} \tag{21}$$

with probability at least $(1 - \frac{4}{n^2})$.

We now derive a bound on the error incurred by the k-means step in the algorithm. Formally, k-means minimizes $\|\overline{X} - S\|_F$ over all $S \in \mathcal{M}_{n\times k}(k)$, where $\mathcal{M}_{n\times k}(r)$ is the set of all $n \times k$ matrices with at most r distinct rows. In practice, the rows of S correspond to the centers of obtained clusters. Achieving a global optimum for this problem is NP-hard. However, there are algorithms [Kumar, Sabharwal and Sen (2004), Ostrovsky et al. (2012)] that can provide a solution S^* from the above class of matrices such that

$$\|\overline{X} - S^*\|_F \leq \gamma \min_S \|\overline{X} - S\|_F \tag{22}$$

for some $\gamma > 1$. The factor γ depends on the algorithm under consideration. For instance, γ grows with k in the case of Kumar, Sabharwal and Sen (2004). On the other hand, Ostrovsky et al. (2012) showed that a constant factor approximation is possible if the data (rows of $\overline{X}$ in our case) is *well separated.*

To be precise, define $\eta_r(\overline{X})$ to be the minimum of the objective function when r clusters are found, that is,

$$\eta_r(\overline{X}) = \min_{S\in\mathcal{M}_{n\times k}(r)} \|\overline{X} - S\|_F. \tag{23}$$

The rows of $\overline{X}$ is said to be ε-separated if $\eta_k(\overline{X}) \leq \varepsilon\eta_{k-1}(\overline{X})$. Theorem 4.15 in Ostrovsky et al. (2012) claims that if this condition holds for small enough ε, then the solution $S^* \in \mathcal{M}_{n\times k}(k)$ obtained from their approximate k-means algorithm satisfies (22) with probability $(1 - O(\sqrt{\varepsilon}))$, where γ is given as $\gamma = \sqrt{\frac{1-\varepsilon^2}{1-37\varepsilon^2}}$.

The following result shows that in our case, the rows of $\overline{X}$ are indeed well separated.

LEMMA 4.8. *If the condition in* (14) *holds, then the rows of $\overline{X}$ are ε-separated with $\varepsilon = (\ln n)^{-1/2}$.*

As a consequence of Lemma 4.8, it follows that if n is sufficiently large, then the result of Ostrovsky et al. (2012) holds. Moreover, one can also observe that for large n, we have $\gamma = O(1)$.

Finally, one needs to combine the above results in order to prove Theorem 4.2. For this, define the set $\mathcal{V}_{\mathrm{err}} \subset \mathcal{V}$ as

$$\mathcal{V}_{\mathrm{err}} = \left\{ i \in \mathcal{V} : \| S^*_{i\cdot} - Z_{i\cdot} Q \|_2 \geq \frac{1}{\sqrt{2}} \right\}. \tag{24}$$

Rohe, Chatterjee and Yu (2011) used a similar definition for the number of incorrectly assigned nodes, and discussed the intuition behind this definition. In the following result, we formally prove that the nodes that are not in $\mathcal{V}_{\mathrm{err}}$ are correctly assigned. We also provide an upper bound on the size of $\mathcal{V}_{\mathrm{err}}$.

LEMMA 4.9. *Let $i, j \notin \mathcal{V}_{\mathrm{err}}$ and $S^*_{i\cdot} = S^*_{j\cdot}$, then $\psi_i = \psi_j$. As a consequence,* $\mathsf{Err}(\psi, \psi') \leq |\mathcal{V}_{\mathrm{err}}|$. *In addition,*

$$|\mathcal{V}_{\mathrm{err}}| \leq 4(1+\gamma^2) \| \overline{X} - ZQ \|_F^2.$$

Theorem 4.2 follows by combining the above bound with (21).

5. Consistency for special cases. We now study the implications of Theorem 4.2 for partitioning particular models of uniform and nonuniform hypergraphs. We also discuss the conditions for identifiability in special cases.

5.1. *Balanced partitions in uniform hypergraph.* Let the n nodes be divided into k groups such that each group contains $\frac{n}{k}$ nodes. We now consider a random r-uniform hypergraph on the nodes generated as follows. Let $p, q \in [0, 1]$ be constants with $(p+q) \leq 1$, and $\alpha_{r,n} \in (0, 1]$ be the sparsity factor dependent on n. For any r nodes from the same group, there is an edge among them with probability $\alpha_{r,n}(p+q)$. If all the r nodes do not belong to same group, then there is an edge with probability $\alpha_{r,n} q$.

In terms of the model in Section 3, one can see that $M = r$, and for all $m < r$, $\alpha_{m,n} = 0$. The rth order k-dimensional tensor $B^{(r)}$ is given by

$$B^{(r)}_{j_1 j_2 \cdots j_r} = \begin{cases} p+q & \text{if } j_1 = j_2 = \cdots = j_r, \\ q & \text{otherwise.} \end{cases}$$

One can see that for $r = 2$, this model corresponds to the sparse stochastic block model considered in Lei and Rinaldo (2015) with balanced community sizes, and if $\alpha_{2,n} = 1$, one has the standard four parameter stochastic block model [Rohe, Chatterjee and Yu (2011)]. The following corollary to Theorem 4.2 shows the consistency of Algorithm 1.

COROLLARY 5.1. *In the above model*,

$$\delta = \frac{p\alpha_{r,n} n}{rkd} \binom{\frac{n}{k} - 2}{r - 2}, \tag{25}$$

and hence, *the partitions are identifiable for all* $p > 0$. *Moreover*, *if*

$$\alpha_{r,n} \geq C \frac{k^{2r-1} n (\ln n)^2}{\binom{n}{r}} \tag{26}$$

for some absolute constant $C > 0$, *then the conditions in Theorem* 4.2 *are satisfied*, *and hence*, *we have*

$$\mathsf{Err}(\psi, \psi') = O\left(\frac{k^{2r-2} n^2 \ln n}{p^2 \alpha_{r,n} \binom{n}{r}}\right) = o(n) \tag{27}$$

with probability $(1 - o(1))$.

The lower bound on $\alpha_{r,n}$ mentioned in Corollary 5.1 needs some discussion. One can verify that in the above model, the expected number of edges lie in the range $[q\alpha_{r,n}\binom{n}{r}, (p+q)\alpha_{r,n}\binom{n}{r}]$, that is, it is about $\alpha_{r,n}\binom{n}{r}$ up to a constant scaling. The lower bound on $\alpha_{r,n}$ specifies that the number of edges must be at least $\Omega(k^{2r-1} n (\ln n)^2)$. This also indicates that for a larger r, more edges are required to ensure the error bound of Corollary 5.1. Since $\alpha_{r,n} \leq 1$, one can see that the result is applicable for $k = O(n^{0.5-\varepsilon})$ for all $\varepsilon > \frac{1}{2(2r-1)}$. Even consistency results for graph partitioning require similar condition [Choi, Wolfe and Airoldi (2012), Rohe, Chatterjee and Yu (2011)].

A closer look at the condition (26) shows that if k is constant or increases slowly, $k = O(\ln n)$, then a sufficient condition for weak consistency of Algorithm 1 is $\alpha_{r,n} \geq C_r \frac{(\ln n)^{2r+1}}{n^{r-1}}$, where the constant C_r depends only on r. In case of graph partitioning, this level of sparsity is needed when one relies on matrix Bernstein inequality. However, recent results [Lei and Rinaldo (2015)] reduced the lower bound by using sharp concentration bounds for the binary adjacency matrix. Corollary 5.1 also indicates that if k increases at a higher rate, for example, $k = n^a$, then consistency can be guaranteed only when the hypergraph is more dense.

On the other extreme are uniform hypergraphs encountered in computer vision Ghoshdastidar and Dukkipati (2014, 2015) that are usually dense, that is, $\alpha_{r,n} = 1$. In this case, if $k = O(\ln n)$ then $\mathsf{Err}(\psi, \psi') = O(\frac{(\ln n)^{2r-1}}{n^{r-2}})$. Thus, the error decreases at a faster rate for r-uniform hypergraphs with larger r. In fact, for $r \geq 3$, above bound indicates that $\mathsf{Err}(\psi, \psi') = o(1)$, that is, Algorithm 1 guarantees exact recovery of the partitions for large n.

Lastly, we discuss the effect of δ and the parameters p, q in this setting. Note that the case $q = 0$ is not interesting as there are no edges among different groups, and hence, the partition can be identified by a simple breadth-first search. On the other hand, $p = 0$ generates a random uniform hypergraph with all identical edges.

Hence, the partitions cannot be identified in this case. This can also be seen from (25), where $\delta = 0$. In general, p denotes the gap between the probability of edge occurrence among nodes from same community and the probability with which nodes from different communities form an edge. Since δ is linear in p, one can observe from Theorem 4.2 that $\mathsf{Err}(\psi, \psi')$ varies as $\frac{1}{p^2}$ with p. However, note that the model assumes that p does not vary with n, and may be treated as a constant in the asymptotic case.

5.2. *Balanced partitions in nonuniform hypergraph.* We now consider the case of nonuniform hypergraph of range M, where M may vary with n. As in Section 5.1, assume that n nodes are equally split into k groups. Also let $p, q \in (0, 1)$ such that $(p+q) \leq 1$, and for $m = 2, \ldots, M$, let $B^{(m)}$ be the mth-order symmetric k-dimensional tensor with

$$B^{(m)}_{j_1 j_2 \ldots j_m} = \begin{cases} p+q & \text{if } j_1 = j_2 = \cdots = j_m, \\ q & \text{otherwise.} \end{cases}$$

Setting $\alpha_{m,n} \in (0, 1]$ as the sparsity factors, we obtain a model, where the edges appear independently, and for each m, an edge on m nodes from the same group appears with probability $\alpha_{m,n}(p+q)$. For any set of m nodes from different groups, there is an edge among them with probability $\alpha_{m,n} q$.

Since, the nonunifom hypergraph is a superposition of the m-uniform hypergraphs for $m = 2, \ldots, M$, one can easily derive a consistency result in the nonuniform case by applying Corollary 5.1 for each of the uniform components. However, observe that the number of edges of size m is $\Theta(\alpha_{m,n}\binom{n}{m})$, and hence, the requirement $\alpha_{m,n}\binom{n}{m} \geq C_m k^{2m-1} n (\ln n)^2$ for each m implies that the number of m-size edges should increase with m. This contradicts the natural intuition in existing random models [Darling and Norris (2005)], where the hypergraph contains less edges of higher cardinality. The same phenomenon is also observed in practice (see Section 6.1). The following consistency result takes this fact into account.

COROLLARY 5.2. *The partitions in the above model are identifiable for all* $p > 0$. *In addition, let* $(\theta_m)_{m=2}^{\infty}$ *be a nonnegative sequence independent of* n, *and assume that for any* $n \in \mathbb{N}$ *and* $m = 2, \ldots, M$, *the sparsity factor*

$$\alpha_{m,n} = \frac{\theta_m n^a (\ln n)^b}{\binom{n}{m}}$$

for some $a \geq 1$ *and* $b \geq 2$. *There exists an absolute constant* C, *such that, if*

$$\sum_{m=r}^{M} m\theta_m \leq C\left(\frac{n^{a-1}(\ln n)^{b-2}}{k^{2r-1}}\right) \tag{28}$$

for $r = \min\{m : \theta_m > 0\}$, *then* $\mathsf{Err}(\psi, \psi') = o(n)$ *with probability* $(1 - o(1))$.

In the above result, r denotes the smallest size of an edge in the hypergraph. In practice $(\theta_m)_{m=2}^{\infty}$ is a decreasing sequence, and hence, the number of m-size edges also decreases with m. In particular, if $\theta_2 > 0$, $\sum_{m=2}^{\infty} m\theta_m < \infty$, and $k = O(n^{(a-1)/3}(\ln n)^{(b-2)/3})$, then Algorithm 1 is weakly consistent. Thus, if the hypergraph is sparse, that is, $a = 1$, consistency is guaranteed only for logarithmic growth in k, whereas larger number of partitions can be consistently detected only in dense hypergraphs. Observe that the problem gets harder if $r > 2$.

5.3. *Identifiability of the partitions.* In the previous two sections, we considered problems where the partitions are identifiable from $\mathcal{L}$. This need not hold for arbitrary model parameters. We now briefly discuss few cases, which show that the partitions are typically identifiable under reasonable choice of model parameters.

EXAMPLE 1. Consider a 3-uniform hypergraph on n nodes. For simplicity, assume there are $k \geq 3$ partitions of equal size. We define $B^{(3)}$ as follows:

$$B^{(3)}_{j_1 j_2 j_3} = \begin{cases} p_1 & \text{if } j_1 = j_2 = j_3, \\ p_2 & \text{if exactly two of them are identical,} \\ p_3 & \text{if } j_1, j_2, j_3 \text{ are all different} \end{cases}$$

for some constants $p_1, p_2, p_3 \in [0, 1]$. Observe that the above situation is the most general case provided that the partitions are statistically identical. In this setting, it is easy to see that the following statement holds.

LEMMA 5.3. *Assume that n is a multiple of k. Then $\delta > 0$ if and only if*

$$(p_2 - p_3) + \frac{1}{k}(p_1 - 3p_2 + 2p_3) - \frac{2}{n}(p_1 - p_2) > 0. \tag{29}$$

In particular, $\delta > 0$ when $p_1 > p_2 > p_3$, or at most one inequality is replaced by equality.

Note that the setting of Section 5.1 follows when $p_1 > p_2 = p_3$, while the case $p_1 = p_2 = p_3$ corresponds to a random hypergraph with all edges following the same law. Obviously, the partitions are not identifiable in the latter case. More generally, the order of probabilities $p_1 > p_2 > p_3$ is intuitive as it implies that an edge has a larger probability of occurrence if it has more nodes from the same community. One may compare this observation with the case of graphs, where partitioning based on the leading eigenvectors of Laplacian works only when edges within each community occur more frequently than edges across communities. The opposite scenario, found in colorable graphs, requires one to consider eigenvectors corresponding to the other end of the spectrum [Alon and Kahale (1997)]. Moreover, if $p_2 > p_3$ and k grows with n, one can observe that δ mostly depends on the gap $(p_2 - p_3)$, and hence, the error $\mathsf{Err}(\psi, \psi')$ is proportional to $\frac{1}{(p_2-p_3)^2}$.

EXAMPLE 2. We now modify the above model by allowing edges of size 2 to be present. In particular, assume $\alpha_{2,n} = 1$ and $B^{(2)} = I$, which means all pairwise edges within each community are present, and no two nodes from different communities form a pairwise edge. In addition, let $\alpha_{3,n} \in [0, 1]$ be arbitrary. Then one can observe the following.

LEMMA 5.4. *Assume that n is a multiple of k. Then $\delta > 0$ if and only if*

$$\frac{1}{2} + \frac{n\alpha_{3,n}}{3}\left((p_2 - p_3) + \frac{(p_1 - 3p_2 + 2p_3)}{k} - \frac{2(p_1 - p_2)}{n}\right) > 0. \tag{30}$$

It is easy to see that if $\alpha_{3,n} = 0$, then the hypergraph is a graph with k disconnected components, and hence, the partitions are identifiable. However, even when $\alpha_{3,n} = o(\frac{1}{n})$, the pairwise edges eventually dominate and the partitions can be identified for arbitrary values of p_1, p_2, p_3. On the other hand, if $\alpha_{3,n}$ grows faster than $\frac{1}{n}$ (for instance, $\alpha_{3,n} = 1$), then the situation is eventually similar to that of Lemma 5.3. The critical case is $\alpha_{3,n} = \Theta(\frac{1}{n})$, where the expected number of 2-way and 3-way edges are of similar order. In this case, (30) suggests that the partitions can be identified ($\delta > 0$) even when $p_2 < p_3$ provided p_3 is sufficiently small.

EXAMPLE 3. In the above cases, we restricted ourselves to communities of equal size. The arguments also hold for $\frac{n_1}{n_k} = O(1)$. However, if $n_k \ll n_1$ or the probability of edges vary across different communities, then the second term in (13) can lead to $\delta \leq 0$, or equivalently, may affect the identifiability of the partitions. To study this effect, we consider the following model for r-uniform hypergraphs.

Let $\alpha_{r,n} = 1$, and there are $k = 2$ partitions of size s and $(n - s)$. We assume $s = o(n)$, and define $B^{(r)} \in \mathbb{R}^{2\times 2\times\cdots\times 2}$ as

$$B^{(r)}_{j_1 j_2 \cdots j_r} = \begin{cases} 1 & \text{if } j_1 = j_2 = \cdots = j_r = 1, \\ \dfrac{1}{2} & \text{otherwise.} \end{cases}$$

For $r = 2$, the model is same as that of a s clique planted in a Erdős–Rényi graph. This model presents a high disparity in both community sizes and degree distributions. We make the following comment on the identifiability of the partitions under this model.

LEMMA 5.5. *For a given $r \geq 2$, there exists a finite constant s_r such that $\delta > 0$ for the above model for all $s \geq s_r$.*

Thus, when s grows with n, the partitions can be eventually identified from $\mathcal{L}$. The proof of the above result shows that both the terms in (13) decay with n, but the ratio of the first term to the second grows as $\Omega(s)$. We believe that a similar observation can be made in more general situations, where this growth rate depends on the size of the smallest community.

In view of the above lemma, it is interesting to know whether Algorithm 1 is able to detect small cliques in uniform hypergraphs. This is indeed true, but due to the generality of the approach, as listed in this paper, the minimal growth rate for s needed to accurately find the partitions from L is not optimal. More precisely, it is worse by a logarithmic factor in the case of graphs. However, Lemma 5.5 shows that one can use spectral techniques similar to [Alon, Krivelevich and Sudakov (1998)] for finding planted cliques in hypergraphs.

6. Experimental results. In this section, we empirically demonstrate that the conditions in Corollaries 5.1 and 5.2 are reasonable. For this, we consider a number of hypergraphs that have been studied in practical problems [Alpert (1998), Ghoshal et al. (2009)]. We also study the performance of Algorithm 1 in some benchmark clustering problems.

6.1. *Sparsity of real-world hypergraphs.* The consistency results in this paper are applicable only under certain restrictions on the hypergraph to be partitioned. To be precise, Corollaries 5.1 and 5.2 hold when the sparsity of the hypergraph is above a certain threshold. We study the practicability of the conditions in the case of real-world hypergraphs. We consider two types of applications—*folksonomy*, where the underlying model is a 3-uniform hypergraph, and *circuit design*, which involves nonuniform hypergraph partitioning.

To study the nature of hypergraphs in folksonomy, we consider 11 networks from KONECT, HetRec'2011 and MovieLens.[3] Each network is a tri-partite 3-uniform hypergraph containing three types of nodes: user, resource and annotation. Each edge is an entry in the database that occurs when an user describes a certain resource by a particular tag or rating. The number of nodes vary between 2630 to 9.8×10^5. Assuming that $k = O(1)$, the sufficient condition in Corollary 5.1 requires that the number of edges in a 3-uniform hypergraph grows as $\Omega(n(\ln n)^2)$. In Figure 1, we compare the number of edges $|\mathcal{E}|$ with $n(\ln n)^2$ for above networks. We observe that in few cases (last four in Figure 1), these quantities are similar, whereas for the remaining networks, $|\mathcal{E}|$ is smaller by a nearly constant factor.

The next study is related to nonuniform hypergraphs that are encountered in circuit partitioning. We consider 18 circuits from the ISPD98 circuit benchmark suite [Alpert (1998)]. From a hypergraph view, the components of the circuit are the nodes of the hypergraph, while the multi-way connections among them are the edges. These networks are also sparse as the number of nodes vary from 1.27×10^4

[3] The HetRec'2011 and MovieLens datasets are maintained by the GroupLens research group, and are available at: http://grouplens.org/.

KONECT refers to the Koblenz network collection: http://konect.uni-koblenz.de/.

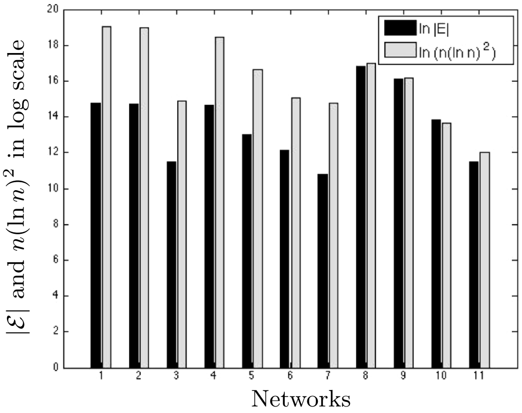

FIG. 1. *Bar plot for* $|\mathcal{E}|$ *and* $n(\ln n)^2$ *in logarithmic scale for* 11 *folksonomy networks.*

to 2.1×10^5, while the number of edges range between 1.4×10^4 to 2×10^5. Moreover, these networks contain relatively large number of edges of sizes 2 or 3, and the number of edges of size m gradually decreases with m. We assume $a = 1$, $b = 2$, and ignoring constant factors, we estimate θ_m as $\theta_m = \frac{|\mathcal{E}_m|}{n(\ln n)^2}$, where $\mathcal{E}_m$ is the set of edges of size m in the network. Figure 2 shows a plot of this quantity as a function of m for different networks. We find that the estimate of θ_m is bounded by exponentially decaying functions, and hence, one can argue that $\sum_m m\theta_m < \infty$.

6.2. *Experiments on benchmark problems.* Partitioning the networks discussed in the previous section is an interesting problem. However, for such networks, the underlying partitions are not known, and hence, for these networks,

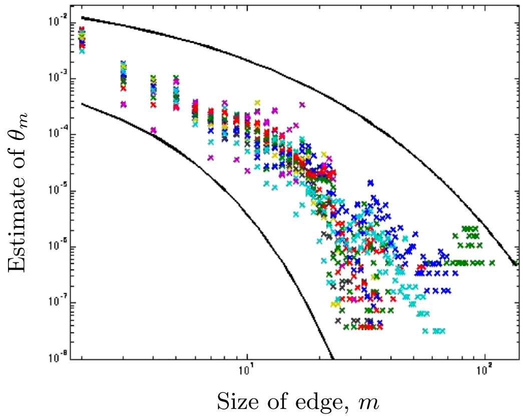

FIG. 2. *Scatter plot for estimated* $\theta_m = \frac{|\mathcal{E}_m|}{n(\ln n)^2}$ *versus m for the* 18 *circuits. Plot for each circuit is shown in a different color. The bounding curves correspond to the functions* $0.05\exp(-m^{0.5})$ *from above and* $0.002\exp(-m^{0.8})$ *from below.*

TABLE 1

Fraction of nodes incorrectly assigned by different clustering algorithms. The results for ROCK, COOLCAT and LIMBO are taken from Andritsos *et al.* (2004)

Database	**ROCK**	**COOLCAT**	**LIMBO**	**hMETIS**	**Clique**	**Algorithm 1**
Voting	0.16	0.15	0.13	0.24	0.12	0.12
Mushroom	0.43	0.27	0.11	0.48	0.11	0.11

the performance of Algorithm 1 cannot be measured in terms of the number of incorrectly assigned nodes. So, we consider benchmark problems for categorical data clustering, where the true partitions are known. Here, one needs to group instances of a database, each described by a number of categorical attributes. Two such benchmark databases include the *1984 US Congressional Voting Records* and the *Mushroom Database* available at the UCI repository [Lichman (2013)]. The first set contains votes of 435 Congress men on 16 issues. The task is to group the Congress men into Democrats and Republicans based on whether they voted for or against the issue, or abstained their votes. The mushroom database contains information about 22 features of 8124 varieties of mushrooms. Based on the categorical features, one needs to separate the edible varieties from the poisonous ones. Thus, both databases have two well-defined partitions.

One may consider the instances of the database as the nodes of the hypergraph. For each possible value of each attribute, an edge is considered among all instances that take the particular value of the attribute. This generates a sparse nonuniform hypergraph that can be partitioned to obtain the clusters. Table 1 compares the performance of Algorithm 1 with some popular categorical clustering algorithms. We also study the performance when the hypergraph partitioning is done using a multilevel approach (hMETIS) [Karypis and Kumar (2000)] or eigen-decomposition of the normalized Laplacian obtained from clique expansion [Rodríguez (2002)]. The error is measured as $\frac{1}{n}\mathsf{Err}(\psi, \psi')$. Table 1 shows that Algorithm 1 performs quite well compared to other methods.

7. Conclusion. The primary focus of this work was to study the consistency of hypergraph partitioning in the presence of a planted structure in the hypergraph. This is achieved by considering a model for random hypergraphs that extends the stochastic block model in a natural way. The algorithm studied in this work is quite simple, where one essentially reduces a given hypergraph to a graph with weighted adjacency matrix given by (4), and then performs spectral clustering on this graph. Our analysis mainly relies on a matrix concentration inequality that was previously used to derive concentration bounds for the Laplacian matrix of sparse random graphs [Chung and Radcliffe (2011)]. We also establish that the k-means step indeed achieves a constant factor approximation with probability $(1 - o(1))$. This question had remained unanswered for a long time in the spectral clustering literature.

Note on the optimality of our result. Theorem 4.2 is quite similar, in spirit, to the existing results in the block model literature, for instance Lei and Rinaldo (2015), where it is shown that spectral clustering is weakly consistent when the expected degree of any node is $\Omega(\ln n)$, or equivalently the edge density $\alpha_{2,n} = \Omega(\frac{\ln n}{n})$. One can easily see from Corollary 5.1 that our result is not optimal, at least in the case of graphs. The primary factor contributing to this difference is a sharp concentration result [Friedman, Kahn and Szemeredi (1989)] that holds for the sparse binary adjacency matrix. While such a sharp bound may not hold for a weighted adjacency matrix, we do wonder whether one can consider the following approach.

QUESTION. Viewing a hypergraph as a collection of m-uniform hypergraphs, one can represent the adjacencies as a collection of m-way binary tensors for varying m. Does a generalization of Friedman, Kahn and Szemeredi (1989) hold for sparse binary tensors? If so, then what is its implication on the allowable sparsity for community detection in hypergraphs?

One can show that for dense tensors, the operator norm (equivalently, largest eigenvalue for matrices) does concentrate similar to the matrix case [Ghoshdastidar and Dukkipati (2015)], but the sparse case has not been studied yet.

Considering the most sparse regime for community detection [Decelle et al. (2011)], it is now known that spectral techniques based on eigenvectors of suitably defined matrices work even for graphs with density $\alpha_{2,n} = \Theta(\frac{1}{n})$ [Krzakala et al. (2013), Le, Levina and Vershynin (2015)]. To this end, the following problem is quite interesting.

QUESTION. What is an appropriate extension of the regularized adjacency matrix [Le, Levina and Vershynin (2015)] and the nonbacktracking matrix [Krzakala et al. (2013)] in the case of hypergraphs? More generally, what is the algorithmic barrier for community detection in hypergraphs?

Phase transitions in uniform hypergraphs have been studied in the literature [Achlioptas and Coja-Oghlan (2008), Panagiotou and Coja-Oghlan (2012)], and thresholds for 2-colorability and Boolean satisfiability are known up to constant factors. However, the case of nonuniform hypergraphs still remains unexplored.

Extensions of our results. One can observe that both the model and the analysis can be further extended to more general situations. For instance, one often encounters weighted hypergraphs in practical applications [Ghoshdastidar and Dukkipati (2014)], where every edge e has a weight $w(e)$ associated with it. In our random model, we assumed $w(e)$ to be a Bernoulli random variable. A direct extension to the weighted hypergraphs is obtained by allowing $w(e)$ to take real values. To this end, we note that our results are only based on the first two moments of $w(e)$.

Hence, if we restrict $w(e) \in [0, 1]$ and assume that first moment is same as that of the Bernoulli variables in our model, then Theorem 4.2 holds even in this setting.

In the case of planted graphs, the stochastic block model has been extended to account for factors such as degree heterogenity or overlapping communities [Lei and Rinaldo (2015), Zhang, Levina and Zhu (2014)]. Similar modifications of the hypergraph model is an interesting extension. However, it also seems possible that some information, such as community overlap, may be lost when the edge information is 'compressed' into the hypergraph Laplacian. Hence, one may have to consider spectral properties of the incidence matrix H or other alternatives for the Laplacian in (2).

While we restricted our discussions to a popular hypergraph partitioning approach, the results can be extended to variants of Algorithm 1. For instance, one may use the eigenvectors of the weighted adjacency matrix A instead of the Laplacian L. Minor modifications to Theorem 4.2 can guarantee the consistency of such an approach. Moreover, Theorem 4.2 is based on a theoretical result of approximate k-means [Ostrovsky et al. (2012)]. As mentioned in Section 4, we need to assume n to be sufficiently large in order to ensure that the k-means step provides a near optimal solution. Alternatively, one could also use the greedy clustering algorithm of Gao et al. (2015) that may alleviate the condition on n.

It also is known that one can iteratively refine the solution of a spectral algorithm to exactly recover the partitions [Lei and Zhu (2014), Vu (2014)]. Such an approach usually constructs an embedding of the nodes based on the adjacency matrix. We believe that similar results will hold for hypergraphs if one constructs the embedding using the weighted adjacency matrix A defined in (4).

We noted in Lemma 5.5 that Algorithm 1 can be used to find a planted clique in a random hypergraph. For a more optimal result, one could possibly extend the approach of Alon, Krivelevich and Sudakov (1998) to the case of hypergraphs. To this end, it is interesting to note that the hypergraph clique problem is often encountered in computer vision applications [Ghoshdastidar and Dukkipati (2014)]. Thus, several variations of the hypergraph partitioning problem surface in engineering applications.

This work explores into the theoretical analysis of hypergraph partitioning, and provides the first step for expanding the extensive studies on planted graphs to the case of hypergraphs.

Acknowledgements. The authors thank the reviewers for pointing out key references and for helpful suggestions that led to inclusion of the result for k-means algorithm, and discussions on identifiability of the partitions.

SUPPLEMENTARY MATERIAL

Supplement to "Consistency of spectral hypergraph partitioning under planted partition model" (DOI: 10.1214/16-AOS1453SUPP; .pdf). The supplementary material contains detailed proofs of all the lemmas and corollaries stated in Sections 4 and 5.

REFERENCES

ACHLIOPTAS, D. and COJA-OGHLAN, A. (2008). Algorithmic barriers from phase transitions. In *Proceedings of* 49*th Annual Symposium on Foundations of Computer Science*.

AGARWAL, S., BRANSON, K. and BELONGIE, S. (2006). Higher order learning with graphs. In *Proceedings of the International Conference on Machine Learning* (*Pittsburgh*, *Pennsylvania*, 2006) 17–24. ACM, New York.

ALON, N. and KAHALE, N. (1997). A spectral technique for coloring random 3-colorable graphs. *SIAM J. Comput*. **26** 1733–1748. MR1484153

ALON, N., KRIVELEVICH, M. and SUDAKOV, B. (1998). Finding a large hidden clique in a random graph. In *Proceedings of the Ninth Annual ACM–SIAM Symposium on Discrete Algorithms* (*San Francisco*, *CA*, 1998) 594–598. ACM, New York. MR1642973

ALPERT, C. J. (1998). The ISPD98 circuit benchmark suite. In *ISPD'*98: *Proceedings of the* 1998 *International Symposium on Physical Design* 80–85. ACM, New York.

AMINI, A. A. and LEVINA, E. (2014). On semi-definite relaxations for the block model. Available at arXiv:1406.5647.

ANDRITSOS, P., TSAPARAS, P., MILLER, R. J. and SEVCIK, K. C. (2004). LIMBO: Scalable clustering of categorical data. In *International Conference on Extending Database Technology* (*Heraklion*, *Crete*, *Greece*, 2004) 123–146. Springer, Berlin, Heidelberg.

BERGE, C. (1984). *Hypergraphs*: *Combinatorics of Finite Sets*. Elsevier, Amsterdam.

BICKEL, P. J. and CHEN, A. (2009). A nonparametric view of network models and Newman–Girvan and other modularities. *Proc*. *Natl*. *Acad*. *Sci*. *USA* **106** 21068–21073.

BOLLA, M. (1993). Spectra, Euclidean representations and clusterings of hypergraphs. *Discrete Math*. **117** 19–39. MR1226129

CHEN, H. and FRIEZE, A. (1996). Coloring bipartite hypergraphs. In *Integer Programming and Combinatorial Optimization* (*Vancouver*, *BC*, 1996). *Lecture Notes in Computer Science* **1084** 345–358. Springer, Berlin. MR1441812

CHEN, Y., SANGHAVI, S. and XU, H. (2014). Improved graph clustering. *IEEE Trans*. *Inform*. *Theory* **60** 6440–6455. MR3265033

CHOI, D. S., WOLFE, P. J. and AIROLDI, E. M. (2012). Stochastic blockmodels with a growing number of classes. *Biometrika* **99** 273–284. MR2931253

CHUNG, F. R. K. (1993). The Laplacian of a hypergraph. In *Expanding Graphs* (*Princeton*, *NJ*, 1992). *DIMACS Ser*. *Discrete Math*. *Theoret*. *Comput*. *Sci*. **10** 21–36. Amer. Math. Soc., Providence, RI. MR1235565

CHUNG, F. and RADCLIFFE, M. (2011). On the spectra of general random graphs. *Electron*. *J*. *Combin*. **18** Paper 215. MR2853072

COOPER, J. and DUTLE, A. (2012). Spectra of uniform hypergraphs. *Linear Algebra Appl*. **436** 3268–3292. MR2900714

DARLING, R. W. R. and NORRIS, J. R. (2005). Structure of large random hypergraphs. *Ann*. *Appl*. *Probab*. **15** 125–152. MR2115039

DECELLE, A., KRZAKALA, F., MOORE, C. and ZDEBOROVÁ, L. (2011). Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications. *Phys*. *Rev*. *E* **84** 066106.

FIEDLER, M. (1973). Algebraic connectivity of graphs. *Czechoslovak Math*. *J*. **23** (98) 298–305. MR0318007

FRIEDMAN, J., KAHN, J. and SZEMEREDI, E. (1989). On the second eigenvalue of random regular graphs. In *Proceedings of the Twenty-First Annual ACM Symposium on Theory of Computing*.

GAO, C., MA, Z., ZHANG, A. Y. and ZHOU, H. H. (2015). Achieving optimal misclassification proportion in stochastic block model. Available at arXiv:1505.03772.

GHOSHAL, G., ZLATIĆ, V., CALDARELLI, G. and NEWMAN, M. E. J. (2009). Random hypergraphs and their applications. *Phys*. *Rev*. *E* **79** 066118. MR2551286

GHOSHDASTIDAR, D. and DUKKIPATI, A. (2014). Consistency of spectral partitioning of uniform hypergraphs under planted partition model. In *Advances in Neural Information Processing Systems* (*Montréal*, *Canada*, 2014) 397–405. Curran Associates, Inc., Red Hook, NY.

GHOSHDASTIDAR, D. and DUKKIPATI, A. (2015). A provable generalized tensor spectral method for uniform hypergraph partitioning. In *Proceedings of the International Conference on Machine Learning* (*Lille*, *France*, 2015). 400–409.

GHOSHDASTIDAR, D. and DUKKIPATI, A. (2016). Supplement to "Consistency of spectral hypergraph partitioning under planted partition model." DOI:10.1214/16-AOS1453SUPP.

GIBSON, D., KLEINBERG, J. and RAGHAVAN, P. (2000). Clustering categorical data: An approach based on dynamical systems. *VLDB J.* **8** 222–236.

GIRVAN, M. and NEWMAN, M. E. J. (2002). Community structure in social and biological networks. *Proc. Natl. Acad. Sci. USA* **99** 7821–7826 (electronic). MR1908073

GOVINDU, V. M. (2005). A tensor decomposition for geometric grouping and segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition* (*San Diego*, *CA*, *USA*, 2005) 1150–1157. IEEE Computer Society, Washington, DC.

GUIMERA, R. and AMARAL, L. A. N. (2005). Functional cartography of complex metabolic networks. *Nature* **433** 895–900.

HOLLAND, P. W., LASKEY, K. B. and LEINHARDT, S. (1983). Stochastic blockmodels: First steps. *Soc. Netw.* **5** 109–137. MR0718088

HU, S. and QI, L. (2012). Algebraic connectivity of an even uniform hypergraph. *J. Comb. Optim.* **24** 564–579. MR2974300

KARYPIS, G. and KUMAR, V. (2000). Multilevel k-way hypergraph partitioning. *VLSI Des.* **11** 285–300.

KERNIGHAN, B. W. and LIN, S. (1970). An efficient heuristic procedure for partitioning graphs. *Bell Syst. Tech. J.* **49** 291–307.

KRZAKALA, F., MOORE, C., MOSSEL, E., NEEMAN, J., SLY, A., ZDEBOROVÁ, L. and ZHANG, P. (2013). Spectral redemption in clustering sparse networks. *Proc. Natl. Acad. Sci. USA* **110** 20935–20940. MR3174850

KUMAR, A., SABHARWAL, Y. and SEN, S. (2004). A simple linear time $(1 + \varepsilon)$-approximation algorithm for geometric k-means clustering in any dimensions. In *Proceedings of the Annual Symposium on Foundations of Computer Science* (*Rome*, *Italy*, 2004) 454–462. IEEE Computer Society, Washington, DC.

LE, C. M., LEVINA, E. and VERSHYNIN, R. (2015). Sparse random graphs: Regularization and concentration of the Laplacian. Available at arXiv:1502.03049.

LEI, J. and RINALDO, A. (2015). Consistency of spectral clustering in stochastic block models. *Ann. Statist.* **43** 215–237. MR3285605

LEI, J. and ZHU, L. (2014). A generic sample splitting approach for refined community recovery in stochastic block models. Available at arXiv:1411.1469.

LICHMAN, M. (2013). UCI machine learning repository. Available at http://archive.ics.uci.edu/ml.

LLOYD, S. P. (1982). Least squares quantization in PCM. *IEEE Trans. Inform. Theory* **28** 129–137. MR0651807

MCSHERRY, F. (2001). Spectral partitioning of random graphs. In 42*nd IEEE Symposium on Foundations of Computer Science* (*Las Vegas*, *NV*, 2001) 529–537. IEEE Computer Soc., Los Alamitos, CA. MR1948742

MICHOEL, T. and NACHTERGAELE, B. (2012). Alignment and integration of complex networks by hypergraph-based spectral clustering. *Phys. Rev. E* **86** 056111.

NG, A., JORDAN, M. and WEISS, Y. (2002). On spectral clustering: Analysis and an algorithm. In *Advances in Neural Information Processing Systems* (*Vancouver*, *British Columbia*, *Canada*) 849–856. MIT Press, Cambridge, MA.

OSTROVSKY, R., RABANI, Y., SCHULMAN, L. J. and SWAMY, C. (2012). The effectiveness of Lloyd-type methods for the k-means problem. *J. ACM* **59** Art. 28. MR3008400

PANAGIOTOU, K. and COJA-OGHLAN, A. (2012). Catching the k-NAESAT threshold. In *ACM Symposium on Theory of Computing*.

RODRÍGUEZ, J. A. (2002). On the Laplacian eigenvalues and metric parameters of hypergraphs. *Linear Multilinear Algebra* **50** 1–14. MR1890984

ROHE, K., CHATTERJEE, S. and YU, B. (2011). Spectral clustering and the high-dimensional stochastic blockmodel. *Ann. Statist.* **39** 1878–1915. MR2893856

SCHMIDT-PRUZAN, J. and SHAMIR, E. (1985). Component structure in the evolution of random hypergraphs. *Combinatorica* **5** 81–94. MR0803242

SCHWEIKERT, G. and KERNIGHAN, B. W. (1979). A proper model for the partitioning of electrical circuits. In *Proceedings of the* 9*th Design Automation Workshop* 57–62. ACM, New York, NY.

STASI, D., SADEGHI, K., RINALDO, A., PETROVIC, S. and FIENBERG, S. (2014). β models for random hypergraphs with a given degree sequence. In *Proceedings of COMPSTAT* 2014—21*st International Conference on Computational Statistics* 593–600. Internat. Statist. Inst., The Hague. MR3372442

STEWART, G. W. and SUN, J. G. (1990). *Matrix Perturbation Theory. Computer Science and Scientific Computing*. Academic Press, Boston, MA. MR1061154

TROPP, J. A. (2012). User-friendly tail bounds for sums of random matrices. *Found. Comput. Math.* **12** 389–434. MR2946459

VON LUXBURG, U. (2007). A tutorial on spectral clustering. *Stat. Comput.* **17** 395–416. MR2409803

VU, V. (2014). A simple SVD algorithm for finding hidden partitions. Available at arXiv:1404.3918.

WASSERMAN, S. (1994). *Social Network Analysis*: *Methods and Applications*. Cambridge Univ. Press, Cambridge.

ZHANG, Y., LEVINA, E. and ZHU, J. (2014). Detecting overlapping communities in networks with spectral methods. Available at arXiv:1412.3432v2.

ZHOU, D., HUANG, J. and SCHÖLKOPF, B. (2007). Learning with hypergraphs: Clustering, classification, and embedding. In *Advances in Neural Information Processing Systems* (*Vancouver, British Columbia, Canada*) 1601–1608. MIT Press, Cambridge, MA.

DEPARTMENT OF COMPUTER SCIENCE & AUTOMATION
INDIAN INSTITUTE OF SCIENCE
BANGALORE, 560012
KARNATAKA
INDIA
E-MAIL: debarghya.g@csa.iisc.ernet.in
ad@csa.iisc.ernet.in