Journal of Machine Learning Research 22 (2021) 1-35 Submitted 11/19; Revised 2/21; Published 7/21

Nonparametric Modeling of Higher-Order Interactions via Hypergraphons

Krishnakumar Balasubramanian KBALA@UCDAVIS.EDU
Department of Statistics
University of California
Davis, CA 95616 USA

Editor: Edo Airoldi

Abstract

We study statistical and algorithmic aspects of using hypergraphons, that are limits of large hypergraphs, for modeling higher-order interactions. Although hypergraphons are extremely powerful from a modeling perspective, we consider a restricted class of Simple Lipschitz Hypergraphons (SLH), that are amenable to practically efficient estimation. We also provide rates of convergence for our estimator that are optimal for the class of SLH. Simulation results are provided to corroborate the theory.

Keywords: Hypergraphons, Rate of Convergence, Nonparametric statistics, Tensor methods.

1. Introduction

Let $V = \{1, \ldots, n\}$ be a set of n items that could represent, for example, people in a social network, genes in a biological network or researchers in an academic network. Models of interaction among the n items could be conveniently represented in the form of a graph or a hypergraph, $G(V, E)$, where the items form the nodes of the graph and the hyperedge set E represents the interactions among the items. Network datasets that capture such complex interactions between a set of objects are becoming increasingly prevalent in several scientific fields. Developing realistic generative models for such networks is a challenging problem that has been an active subject of research across diverse fields spanning from statistics, physics, computer science; see Kolaczyk (2009); Goldenberg et al. (2010); Battiston et al. (2020) for comprehensive overview.

A majority of the existing work has focussed on the case of modeling *pairwise* interactions. In this situation, the edge set models interactions between two nodes at a time, by means of presence or absence of a link. In several modern applications, such pairwise interactions do not completely characterize the complex interactions existing among the items. Often times it is more meaningful to consider higher-order interactions (Bonacich et al., 2004; Benson et al., 2016). For example, Agarwal et al. (2005) provided convincing empirical evidence showing that going beyond pairwise interactions helps in computer vision applications; a hypergraph based approach for graph matching was provided in Duchenne et al. (2011). Yet another application is the study of complex networks, where modeling intricate interdependencies between multiple networks are typically represented via hypergraphs (Ghoshal et al., 2009; Zlatić et al., 2009; Michoel and Nachtergaele, 2012; Kivelä

et al., 2014) so as to capture the higher-order interactions among the different networks. We refer the interested reader to Battiston et al. (2020), for a detailed survey of several existing models of higher-order interactions and their applications to various scientific fields.

As a simple yet concrete example for the limitations of the just using pairwise interactions, consider the following co-author citation toy network. Consider a simple set of five authors $\{A, B, C, D, E\}$ and assume that the set of co-author relationship among the authors is as follows: $(A, B), (B, E), (A, E)(C, D, E)$. That is the five authors wrote five papers in total with the above set of authors for each paper. If we represent this network via a graph, with edges representing if two authors have collaborated or not, we get a graph with edge set $\{(A, B), (B, E), (A, E), (C, E), (C, D), (D, E)\}$. We now see that the crucial coauthor information is lost by such representation - one could wrongly interpret that authors A, B, E have co-authored a paper together or one could fail to conclude that authors C, D, E co-authored a paper together. This highlights the limitation of modeling such co-author data set via simple graphs. Clearly if we were to model this data as a hypergraph, we do not run into such issues. A more comprehensive co-author citation network was recently considered in Ji and Jin (2016) based on papers published in the Statistics community. The need for modeling higher-order interaction for that dataset was in particular also suggested in the discussion by Karwa and Petrovia (2016), following the publication of Ji and Jin (2016).

Motivated by such problem with pairwise interaction networks or graph-based networks, recently Ghoshdastidar and Dukkipati (2017a) considered a model for generating hypergraphs. Specifically, they considered a generalization of a stochastic block models (we refer the reader to Abbe (2017) for a detailed overview) developed for the case of graphs to the hypergraph setting. Furthermore Florescu and Perkins (2016) considered a special case of hypergraphs with a bipartite structure and suggested interesting algorithmic conjectures. Algorithmic results were also provided in Ghoshdastidar and Dukkipati (2017b); Ahn et al. (2018); Ke et al. (2019), for community detection in hypergraphs based on tensor methods. While the above works invariably work in the realm of stochastic block modeling, there has been other approaches to model hypergraph data as well. A β-model and a Latent Class Analysis (LCA) based model (focussing on clustering) was proposed in Stasi et al. (2014) and Ng and Murphy (2018) respectively. Furthermore, Turnbull et al. (2019) recently proposed a latent space model for random hypergraphs based on concepts from computational topology. We will revisit the above models in Section 2.3 for a brief discussion on their connection to the approach that we propose in this work. Furthermore, a random geometric model for random hypergraph was also proposed in Lunagómez et al. (2017) focussing on graphical modeling. A significant drawback of the above approaches for modeling higher-order interactions is that they are predominantly parametric models. It is well known that nonparametric models offer increased modeling flexibility at the expense of requiring larger sample sizes in standard regression and density estimation problems. Thus providing such nonparametric models for networks is extremely appealing, particularly in the context of modeling large graphs, as the nodes in the graphs corresponds to samples in the context of traditional regression models.

In this work, we leverage the theory of large hypergraph limits and propose to use hypergraphons (Gowers, 2007; Elek and Szegedy, 2012; Lovász, 2012; Zhao, 2015), as a nonparametric model to capture m-uniform higher-order interactions in networks. Mod-

eling such m-uniform interactions arises, for example, in the context of protein network alignment problem used to represent interactions across different organisms. In this context, we are given two graphs G_1 and G_2 whose vertices are connected by a bipartite graph B. Based on this, a 4-uniform alignment hypergraph is formed with an hyperedge connecting the nodes i, j, k and l if and only if the notes i and j are connected in the graph G_1 and nodes k and l are connected in the graph G_2. We refer to the interested to reader to Michoel and Nachtergaele (2012) for more details. As we will see in the rest of the paper, in the limit of large nodes, the difference between the expressibility of parametric (for example, block hypergraph models) and nonparametric models (the proposed hypergraphons) for modeling m-uniform higher-order interactions is significant. This is different from the case of modeling pairwise interactions via stochastic block models and graphons for graph-valued networks. Roughly speaking while graphs are represented by adjacency matrices (second-order tensors) and the corresponding graphons are two-dimensional functions, this is not true for hypergraphs and hypergraphons. An m-uniform hypergraph could be represented by m-th order tensor, whereas the corresponding hypergraphon is represented by a $2^m - 2$ dimensional function. This expressive power comes at an increased statistical and computational price.

In order to facilitate efficient estimation and computation, we restrict ourself to a class of Simple Lipschitz Hypergraphons (SLH). Furthermore, Lemma 10 and Theorem 2 in Kallenberg (1999) provide guarantees for approximating general hypergraphons with simple hypergraphons, which provides further motivation for understanding this class of hypergraphons statistically and computationally. For this class of hypergraphons, we propose an estimator along with its rates of convergence. The proposed estimator is based on approximating this class of hypergraphons with a parametric stochastic hypergraph block models with appropriately selected number of blocks. Indeed such an approach yields rate-optimal estimators for the case of graphons (Gao et al., 2015; Klopp et al., 2017). Unfortunately, from a computational perspective the estimator is non-convex and hence NP-hard to compute in the worst case. We provide an algorithm which is based on a well-motivated heuristic, and show thorough simulations that it works well in practice. We also mention here that the approximation result provided in Kallenberg (1999) is existential, not entirely constructive and provided in a weaker metric. It is an extremely interesting open problem to provide a constructive proof of such a result in a stronger metric, so that one could find the optimal simple hypergraphon approximation for a given hypergraphon. Such a result would constructively quantify the efficiency of approximation from a practical perspective.

Finally, the modern theory of large hypergraph limits has been established through the analytic regularity approach, for example, in Gowers (2007); Zhao (2015). However, a probabilistic approach based on node-exchangeability has also been examined, for example, in Hoover (1979); Aldous (1981); Kallenberg (1999) to study limits of large hypergraphs. Recently, a closely related concept of edge-exchangeability has been proposed and leveraged by Crane and Dempsey (2018); Campbell et al. (2018); Dempsey et al. (2019) to propose models of higher-order interactions. The above works do not rigorously analyze the statistical estimation procedure associated with the model. It is interesting to explore further relationships and dissimilarities between the proposed hypergraphon model and edge-exchangeable models as future work. Furthermore, it is extremely interesting to quan-

tify the degree of expressibility offered by the two approaches, and compare the statistical and computational complexity of estimation in the different models.

1.1 Notations

We denote by $[n] = \{1, \ldots, n\}$, the set of integers from 1 to n and $\binom{n}{k}$ to denote the number of ways of selecting k objects out of n. Furthermore, we use $\mathfrak{S}_n$ to be the set of all permutations of the set $[n]$. We denote a vector v in d-dimensional Euclidean space by small case letters $a \in \mathbb{R}^d$. Similarly we denote matrices by upper case letters $A \in \mathbb{R}^{n \times n}$. A bold upper case letter $\mathbf{A} \in \mathbb{R}_m^{n \times n \times \ldots \times n}$ corresponds to an m^{th} order tensor. The $(j_1, \ldots, j_m) - th$ entry of the tensor is denoted as $\mathbf{A}_{j_1, j_2, \ldots, j_m}$. For a tensor $\mathbf{A}$ we use $\|\mathbf{A}\|_F$ and $\|\mathbf{A}\|_\infty$ to denote the standard Frobenius norm and the max-norm (maximum value of its entries) respectively.

Furthermore, whenever there is no confusion, to avoid notation overload (mostly in the proofs) we also use the following notation to index the entries of the tensor. Let $j = (j_1, \ldots, j_m)$ with each $j_i \in [n]$ so that $j \in [n]^m$. Whenever it indexes a tensor $\mathbf{A}$ as $\mathbf{A}_j$, it denotes the $(j_1, \ldots, j_m) - th$ entry of the tensor. Hence we have $\mathbf{A}_j = \mathbf{A}_{j_1, \ldots, j_m}$. Furthermore, we use $\sigma(j)$ to denote the set of all permutations of a given $j \in [n]^m$. Here $\sigma \in \mathfrak{S}_m$. We also denote multiple summations like $\sum_{j_1=1}^{n} \cdots \sum_{j_m=1}^{n}$ as $\sum_{j \in [n]^m}$. For the sake of brevity, in the rest of the paper $z(j)$ represents $(z(j_1), z(j_2), \ldots, z(j_m))$, when $j \in [n]^m$, and similarly for $z^{-1}(a)$ when $a \in [k]^m$. We also need the following definition of collapsing a tensor to a matrix.

Definition 1 (Tensorcollapse) *For a tensor $\mathbf{A} \in \mathbb{R}_m^{n \times n \times \ldots n}$, we define the operation of collapsing a tensor to a matrix, $\mathcal{M}(\cdot, \cdot, \cdot) : \mathbb{R}_m^{n \times n \times \ldots n} \times [m] \times [m] \mapsto \mathbb{R}_2^{d \times d}$ as*

$$\mathcal{M}(\mathbf{A}, 1, 2) = A_{j_1, j_2} = \sum_{j_3, j_4, \ldots, j_m = 1}^{d} \mathbf{A}_{j_1, j_2, j_3, j_4, \ldots, j_m}.$$

2. Hypergraph Block Models and Hypergraphons

In this section, we introduce the hypergraphon model that we use for modeling higher order interactions. Before we do so, we introduce some basic definitions.

Definition 2 (Hypergraph) *A hypergraph $G = (V, E)$ consists of a set of vertices denoted by a set and labeled as $V = [n]$ and a set of hyperedges $e \in E$ where each hyperedge e consists of a subsets of vertices from V. A hypergraph is said to be m-uniform if every edge consists of exactly m vertices. Note that when $m = 2$ this corresponds to a graph.*

In this paper, we restrict ourselves to hypergraphs for which hyperedges contain only unique vertices. Furthermore, we consider the case of m-uniform hypergraphs, motivated by the applications mentioned in Section 1. As mentioned previously, a model for pairwise or higher-order interaction is represented by assuming particular structure on the adjacency matrix or tensor, denoted by $\mathbf{A} \in \{0, 1\}^{n \times n \times \cdots \times n}$. Each entry in this tensor represents the presence or absence of an hyperedge.

2.1 Stochastic Block Models

The stochastic block model in the graph setting was proposed by Holland et al. (1983) and has been analyzed extensively in the statistics (Rohe et al., 2011; Lei and Rinaldo, 2015; Choi et al., 2012; Chatterjee, 2015; Abbe et al., 2017; Le et al., 2017; Gao et al., 2018), computer science (Abbe and Sandon, 2015), statistical physics (Fortunato, 2010; Krzakala et al., 2013) communities, thereby establishing a wide variety of results illustration several interesting phenomenon. Due to the large amount of literature developed on the topic of stochastic block models in the recent years, we refer the reader to Abbe (2017) for an in-depth survey. As mentioned before, all the above works consider the case of pairwise interaction models. We now introduce the stochastic block model. Let $z : [n] \mapsto [k]$ denote a mapping of the set of n vertices to one of the k communities which is unknown. Let $\mathcal{Z}_{n,k}$ denote the set of all possible mappings $z : [n] \mapsto [k]$. Note that the cardinality of the set $\mathcal{Z}$ is k^n. Let $\mathbf{Q} \in [0,1]^{k\times k}$ be a symmetric matrix (i.e., a 2^{nd} order tensor), representing the probabilities of connections between and within each of the k communities. The $\mathbf{A}_{j_1 j_2}$-th entry of the adjacency matrix is modeled as a Bernoulli random variable with mean parameter $\mathbf{\Theta}_{j_1 j_2}$. In the stochastic block model, the matrix $\mathbf{\Theta}$ has a block structure and takes values from the matrix $\mathbf{Q}$ in a particular fashion that we describe now. Specifically, the matrix $\mathbf{\Theta}$ is assumed to come from the following set $\mathcal{T}_z$ of matrices:

$$\mathcal{T}_z = \{\mathbf{\Theta} : \exists \mathbf{Q} \in [0,1]^{k\times k} \;\&\; z \in \mathcal{Z}_{n,k} \text{ S.T. } \mathbf{\Theta}_{j_1,j_2} = \mathbf{Q}_{a,b} = \mathbf{Q}_{b,a} \text{ for } (j_1,j_2) \in z^{-1}(a) \times z^{-1}(b), \text{ and } \mathbf{\Theta}_{j_1,j_1} = 0 \text{ otherwise}\}.$$

Recently several authors have considered the straight-forward extension of the stochastic block model from the graph setting to the hypergraph setting. See for example Ghoshdastidar and Dukkipati (2017a); Ke et al. (2019); Ahn et al. (2018); Pal and Zhu (2019) for statistical and computational results in this setting. Furthermore, fundamental limits, in both estimation and hypothesis testing context was studied in Angelini et al.; Chien et al. (2018); Kim et al. (2018); Ahn et al. (2019). Below, we describe the stochastic hypergraph block models, of which the standard stochastic block models are a special case. Let $\mathbf{Q} \in [0,1]^{k\times k\ldots\times k}$ be an m^{th} order tensor that is assumed to by symmetric, i.e., $\mathbf{Q}_j = \mathbf{Q}_{\sigma(j)}$ for $\sigma \in \mathfrak{S}_m$. This tensor has the probabilities of connections between and within each of the k communities. The $\mathbf{A}_j$-th entry of the tensor is modeled as a Bernoulli random variable with mean parameter $\mathbf{\Theta}_j$. In the stochastic block model, the tensor $\mathbf{\Theta}$ has a block structure and takes values from the tensor $\mathbf{Q}$ in a particular fashion that we describe now. In this work we do not consider hyperedges which have at least one repeated vertex. Define $G \subset [n]^m$ to be the set of hyperedges with no repeated vertices. That is, $j \notin G$ iff $\exists\ u, v$ s.t $j_u = j_v$. Then the tensor $\mathbf{\Theta}$ is assumed to come from the following set $\mathcal{T}_z$ of order m tensors:

$$\mathcal{T}_z = \{\mathbf{\Theta} : \exists \mathbf{Q} \in [0,1]_m^{k\times\cdots\times k} \;\&\; z \in \mathcal{Z}_{n,k} \text{ S.T. } \mathbf{\Theta}_j = \mathbf{Q}_{z(j)} \text{ for } j \in G, \text{ and } \mathbf{\Theta}_j = 0 \text{ otherwise}\} \tag{1}$$

Note that when $m = 2$, the above definition corresponds to the case of standard stochastic block models defined for the case of graphs (Holland et al., 1983) and described above.

2.2 Hypergraphons

The stochastic block model, as mentioned before serves as an intriguing model for understanding several fundamental principles behind statistical analysis of networks. Motivated by the theory of large limits (Lovász, 2012), recently nonparametric models and estimators for such models, was proposed and analyzed in Bickel and Chen (2009); Airoldi et al. (2013); Wolfe and Olhede (2013); Yang et al. (2014); Chan and Airoldi (2014); Borgs et al. (2015a,b); Gao et al. (2015); Klopp et al. (2017); Zhang et al. (2017); Pensky (2019), for modeling pairwise interactions. The graphon models are in-homogenous random graph models that contain several specific models proposed in the literature, including the stochastic block model and latent space model (Hoff et al., 2002), as special cases.

Definition 3 (Graphon and Sampling) *Graphons are symmetric measurable functions* $f : [0,1]^2 \mapsto [0,1]$. *Given a graphon, the process of sampling a random graph proceeds as follows:*

1. *A sequence of real numbers* $X_1, \ldots, X_n$ *is generated independently and uniformly on* $[0,1]$.
2. *Let* $\theta_{i,j} = f(X_i, X_j)$, *where* f *is a graphon.*
3. *Let* $A_{i,j}$ *be the entry of identity adjacency matrix. It is generated as* $A_{i,j} = \text{BER}(\theta_{i,j})$.

Note that topology of the network generated as above is invariant to the permuting the labels. Hence, we have a graph generated as above with a graphon $f(u,v)$ *and another graphon* $f(\vartheta(u), \vartheta(v))$ *for a measure preserving bijection* $\vartheta : [0,1] \mapsto [0,1]$ *invoke the same random graph model.*

We refer the reader to Lovász (2012) for a complete and rigorous characterization of graphons. From the statistical literature, a model of particular relevance to us is that of sparse graphons as considered in Klopp et al. (2017). Under that model, step 2 in the above sampling procedure is modified as $\theta_{i,j} = \rho_n f(X_i, X_j)$, where f is a graphon. Here $\rho_n > 0$ and it tends to zero as $\rho_n \to 0$ when $n \to \infty$. This provides a control on the number of edges generated in the random graph model. Specifically, if $\rho_n = 1$ and fixed, then the expected number of edges is proportional to n^2 which corresponds to a dense graph. But for a value of $\rho_n \to 0$, as $n \to \infty$, we have the expected number of edge to be of the oder $O(\rho_n n^2)$. Hence by controlling the rate of decay of ρ_n, one can obtain sparse graphs that mimic real-world networks. In particular, this modification is much needed when we later work with the hypergraphon setting as the number of higher order interactions are typically sparse as the order increases. We would also like to mention that there are other approaches for sampling sparse exchangeable graph; we refer the intersted reader to Caron and Fox (2017); Caron and Rousseau (2017); Veitch and Roy (2019); Borgs et al. (2019).

The study of the limit of hypergraphs and regularity and left-convergence results was initiated in Gowers (2007). Several authors extended the results with most relevant ones being Elek and Szegedy (2012) and Zhao (2015). Below we first provide the definition of the hypergraphon. We introduce some new notations before we proceed. For a set $[m]$, denote by $[\![m]\!]$, the collection of all nonempty proper subsets of $[m]$ or equivalently the collection of nonempty subset of $[m]$ of size at most $m-1$.

Definition 4 (m-uniform Hypergraphon and Sampling) *A m-uniform symmetric hypergraphon is a $2^m - 2$ dimensional measurable function $f : [0,1]^{[\![m]\!]} \mapsto [0,1]$ whose coordinates are indexed by proper and non-empty proper subset of $[m]$. It is also assumed to be symmetric in the following sense: it remains invariant under any permutation of the coordinates induced by any permutation of $[m]$. Similar to the graphon model, we have the hypergraph generated as:*

1. *A vector $X \in [0,1]^{[\![n]\!]}$ is generated uniformly randomly.*
2. *For a $j \in [n]^m$, let $\mathbf{\Theta}_j = f(X_j)$, where f is graphon satisfying the conditions above and X_j denotes co-ordinates of X indexed by nonempty proper subsets of the set j.*
3. *Let $\mathbf{A}_j$ be the entry of identity adjacency tensor, where j contains only unique vertices. It is generated as $\mathbf{A}_j = \text{BER}(\mathbf{\Theta}_j)$. All other entries are set to zero.*

Note that the definition of symmetry in the above model is slightly different from the more standard notion of symmetry. It is best illustrated through an example. Consider $m = 3$. In this case the hypergraphon is a 6-dimensional function indexed by the sets $\big\{\{1\},\{2\},\{3\},\{1,2\},\{2,3\},\{1,3\}\big\}$. In this case, the function is invariant for any permutation $\sigma \in \mathfrak{S}_3$ operating on $[3]$. That is, it is invariant for any permutation of the first three coordinates and the last three co-ordinates. Finally, similar to the case of sparse graphons we replace the step 2 in the above definition with $\mathbf{\Theta}_j = \rho_n f(X)$, where f is a hypergraphon. Here $\rho_n > 0$ and it tends to zero as $\rho_n \to 0$ when $n \to \infty$ and hence provides a control on the number of hyperedges generated. Specifically the expected number of hyperedges generated under this model is of the order $O(\rho_n n^m)$.

2.3 The Class of Smooth Lipschitz Hypergraphons

The difference between the graphons and hypergraphons is clear from the above Definition 3 and 4. While graphons, which corresponds to the order-2 adjacency tensors (or the standard adjacency matrices) are two dimensional functions, hypergraphons which corresponds to order-m adjacency tensors are 2^m-2 dimensional functions and not m-dimensional functions as one might expect it to be. The need for these additional dimensions could be understood by considering the two cases of generating 3-uniform hypergraph: in the first case each hyperedge (triangle) is sampled independently with same probability (1/2), and in the second case first a standard Erdős-Renyi graphs on n vertices is generated and then the 3-uniform hypergraph is formed based on the triangles in the graph. The limit of the former case is the constant (1/2) hypergraphon function. Where as, the limit of the later is different from the constant hypergraphon function and requires the additional coordinates to be represented. See Zhao (2015) or the example in Figure 5 for the form of the limit in the later case.

The class of general hypergraphons suffers from the problem of efficient estimation as the number of nodes requires must scale at a rate doubly exponentially in terms of the parameter m for it to be consistent. Indeed, firstly the dimensionality of the graphon grows exponentially in terms of the cardinality of the hyperedges and secondly, the number of samples (in this case the nodes) required to achieve consistent estimation of an d dimensional function already grows exponentially in d (note here $d = 2^m$), as is well-known in

nonparametric literature (Klopp et al., 2017; Giné and Nickl, 2015). Furthermore, without smoothness assumptions, it might become more complicated to estimate such functions. In this section, we define the class of Simple Lipschitz Hypergraphon (SLH) which we will be main object we concentrate on in this paper.

Definition 5 *A m-uniform Simple Lipschitz Hypergraphon,* $\mathrm{SLH}(m, L)$*, is a measurable function $f : [0,1]^m \mapsto [0,1]$ whose coordinates are indexed by size-1 subsets of $[m]$. It is assumed to be symmetric in the following sense: it remains invariant under any permutation of the coordinates induced by any permutation of $[m]$. Furthermore, it is assumed to be L-Lipschitz. A consequence of the lipschitz assumption is that for any $(x_1, \ldots, x_m), (y_1, \ldots y_m) \in [0,1]^m$, we have*

$$|f(x_1, \ldots, x_m) - f(y_1, \ldots y_m)| \leq L \max_i |x_i - y_i|. \tag{2}$$

Note that the class of Lipschitz functions that we consider are a subset of α-holder smooth functions (with $\alpha = 1$), which are standard function classes considered in the nonparametric regression literature (see for e.g., (Klopp et al., 2017; Giné and Nickl, 2015). Given the above class of functions, we follow the usual sampling procedure to generate a random hypergraph. The class of *simple hypergraphons* was considered first in Kallenberg (1999), from the view point of modeling exchangeable arrays. The problem of estimation in graphons and hypergraphons was also considered Kallenberg (1999) that provided asymptotic statements for the estimators. Furthermore, an interesting asymptotic result was proved in the same work on approximating general hypergraphons with simple hypergraphons, that provided precise statements on the quality of approximation. Motivated by this work, we restrict ourself to the case of simple hypergraphons as a tradeoff between modeling flexibility and efficient estimability.

Once we restrict ourself to the class of SLH described above, one could use stochastic hypergraph models, with an appropriately growing number of blocks, to estimate hypergraphons efficiently. Indeed, such an approach is motivated by the corresponding rate-optimal estimators proposed for the class of smooth graphons (Gao et al., 2015; Klopp et al., 2017). While such estimators are suitable for estimation in simple hypergraphons, it is not suited for the case of general hypergraphons. This also highlights the limitations of block models in the modeling higher order interactions – whereas general hypergraphons have much higher modeling capacity, stochastic block models with growing number of classes are suitable to capture only the case of SLH.

We end this section by comparing the hypergraphon approach to several other approaches existing in the literature for modeling hypergraphs. Recall that Stasi et al. (2014) proposed a model for random hypergraphs with a given degree sequence, extending the work of Chatterjee et al. (2011) who proposed a similar model for graphs. Such models are called as β-models in the literature. In Chatterjee et al. (2011), it was shown that in the case of graphs, the limit of such models could be characterized as a specific graphon function. Following a similar argument in Chatterjee et al. (2011), it is possible to show that the β-model for hypergraphs proposed in Stasi et al. (2014) also has a limiting hypergraphon construction, and characterize it precisely. It is interesting to examine and characterize the limiting hypergraphon corresponding to the approaches proposed by Ng and Murphy (2018)

and Turnbull et al. (2019). This, however is beyond the scope of current work. A detailed study of this problem will be done in a future work.

3. Estimator and Main Results

We now introduce our estimator and establish rates of convergence for estimating the probability tensor, under the hypergraphon model. Our proof involves first establishing the result when the probability tensor is generated by a block model and then showing that as we increase the number of blocks any function in the class SLH could be well approximated by the block model. Our proof technique extends the results of Klopp et al. (2017) that provided a similar result for the case of graphons. In order do so, we extend their results with appropriate modifications to the case of simple hypergraphons. We also discuss the consequences of our result – specifically the scaling with respect to the problem parameters like smoothness and order of the hypergraphon.

As mentioned above, our estimator is based on a stochastic block model approximation to the class of SLH. Let $\bar{\mathbf{\Theta}}$ denote the true probability tensor generated by a fixed hypergrahon f_0 from the class $SLH(m, L)$. Our estimator $\hat{\mathbf{\Theta}}$, is defined as the following two-steps. In the first step, we solve a least-squares optimization problem. In the next step, we use the estimated assignment function and the probabilities to provide a final estimator of $\hat{\mathbf{\Theta}}$. They are summarized as follows:

$$\begin{aligned} &\textbf{Step 1: } \left(\hat{z}, \hat{\mathbf{Q}}\right) = \underset{z \in \mathcal{Z}_{n,k,n_0};\ Q \in [0,1]_m^{k \times \cdots \times k}}{\operatorname{argmin}} L(z, \mathbf{Q}), \\ &\textbf{Step 2: } \quad \hat{\mathbf{\Theta}}_j = \quad \hat{\mathbf{Q}}_{\hat{z}(j)}. \end{aligned} \tag{3}$$

Here, the loss function, $L(z, \mathbf{Q})$ is given by

$$L(z, \mathbf{Q}) = \sum_{a \in [k]^m} \sum_{j \in z^{-1}(a) \cap H} \left(\mathbf{A}_j - \mathbf{Q}_a\right)^2$$

where $\mathcal{Z}_{n,k,n_0}$ is a slight modification of the set $\mathcal{Z}_{n,k}$ defined in Section 2.1 – it is defined as the set of all mappings from $[n]$ to $[k]$ such that $\min_{a \in [k]} |z^{-1}(a)| \geq n_0$. This modification is made to consider stochastic hypergraph models with *balanced blocks*, while approximating the hypergraphon. For the above estimator, we have the following theorem that quantifies the rate of convergence of estimating the probability tensor $\bar{\mathbf{\Theta}}$.

Theorem 1 (Main Result) *Let $f_0 \in \mathrm{SLH}(m, L)$, where $L > 0$ and $0 < \rho_n \leq 1$. Let $k = \lceil (\rho_n n^m)^{\frac{1}{m+2}} \rceil$ and $n_0 \geq n/k$ and suppose $\rho_n \geq (\log n)^2 n^{-\frac{2}{m+1}}$. Then there is a constant C depending only on α and L such that the least squares estimate $\hat{\mathbf{\Theta}}$ constructed with this choice of n_0 satisfies*

$$\mathbb{E}_{\mathbf{A},X}\left[\frac{1}{n^m}\|\hat{\mathbf{\Theta}} - \bar{\mathbf{\Theta}}\|_F^2\right] \leq C\left\{\rho_n^{\frac{2m+2}{m+2}} n^{-\frac{2m}{m+2}} + \rho_n\left(\frac{\log n}{n^{m-1}}\right)\right\}.$$

Proof [Proof of Theorem 1] As mentioned before, our strategy is to approximate the hypergraphon from the class $SLH(m, L)$ by a hypergraph stochastic block model with appropriately defined number of communities k. Given such a construction, we then capture the

estimation error of parameter estimation in the hypergraph stochastic block model. Let $\mathbf{\Theta}_*$ be the best k-class hypergraph stochastic block model approximation of $\bar{\mathbf{\Theta}}$ in Frobenius norm from the set,

$$\mathcal{T}_{z,n_0} = \{\mathbf{\Theta} : \exists \mathbf{Q} \in [0,1]_m^{k\times\cdots\times k} \ \& \ z \in \mathcal{Z}_{n,k,n_0} \text{ S.T. } \mathbf{\Theta}_j = Q_{z(j)} \text{ for } j \in G, \text{ and } \mathbf{\Theta}_j = 0 \text{ otherwise}\}.$$

Note that the set $\mathcal{T}_{z,n_0}$ is a slight modification of the set $\mathcal{T}_z$ defined in Equation 1 to enforce the balanced partition constraint via choosing n_0. Based on this, we first have the following lemma, that decomposes the overall error into an estimation error term and an approximation error term.

Lemma 1 (Estimation Error) *In the hypergraph stochastic block model setting, we can find an estimator $\hat{\mathbf{\Theta}}$ such that there exists an absolute positive constant $C_1 > 0$ and positive constants $C_2, C_3 > 0$ depending on m such that for $n_0 \geq 2$,*

$$\mathbb{E}_{\mathbf{A},X}\left[\frac{1}{n^m}\|\hat{\mathbf{\Theta}} - \bar{\mathbf{\Theta}}\|_F^2\right] \leq \frac{C_1}{n^m}\mathbb{E}_X\left[\|\bar{\mathbf{\Theta}} - \mathbf{\Theta}_*\|_F^2\right] + C_2\|\bar{\mathbf{\Theta}}\|_\infty\left(\frac{\log k}{n^{m-1}} + \frac{k^m}{n^m}\right) + \frac{C_3 \log n/n_0}{n_0}\left(\frac{\log k}{n^{m-1}} + \frac{k^m}{n^m}\right)$$

Note that our assumption $\rho_n \geq (\log n)^2 n^{-\frac{2}{m+1}}$, after some algebra gives us $\rho_n \geq \frac{\log n/n_0}{n_0}$. Under this scaling, note that third term in the expectation bound of Lemma 1 could be absorbed in the second term and hence we get

$$\mathbb{E}_{\mathbf{A},X}\left[\frac{1}{n^m}\|\hat{\mathbf{\Theta}} - \bar{\mathbf{\Theta}}\|_F^2\right] \leq \frac{C}{n^m}\mathbb{E}_X\left[\|\bar{\mathbf{\Theta}} - \mathbf{\Theta}_*\|_F^2\right] + C\rho_n\left(\frac{\log k}{n^{m-1}} + \frac{k^m}{n^m}\right)$$

Here the term $\frac{1}{n^m}\mathbb{E}_X\left[\|\bar{\mathbf{\Theta}} - \mathbf{\Theta}_*\|_F^2\right]$, corresponds to the approximation error part. We now have the following lemma that bounds the approximation error.

Lemma 2 (Approximation Error) *Consider the hypergraphon model with $f_0 \in \mathrm{SLH}(m, L)$ where $L > 0$. Let $n_0 \geq 2$ and $k = \lfloor n/n_0 \rfloor$. Then,*

$$\mathbb{E}_X\left[\frac{1}{n^m}\|\bar{\mathbf{\Theta}} - \mathbf{\Theta}_*\|_F^2\right] \leq CM^2\rho_n^2\left(\frac{1}{k^2}\right)$$

By Lemma 2, our bound on approximation error, we get the following bound for the overall error:

$$\mathbb{E}_{\mathbf{A},X}\left[\frac{1}{n^m}\|\hat{\mathbf{\Theta}} - \bar{\mathbf{\Theta}}\|_F^2\right] \leq C\left\{\frac{\rho_n^2}{k^2} + \rho_n\left(\frac{\log k}{n^{m-1}} + \frac{k^m}{n^m}\right)\right\}.$$

We now choose $k = \lceil (\rho_n n^m)^{\frac{1}{m+2}} \rceil$ with the goal of balancing the first and third terms in the above expression. This choice of k gives us

$$\mathbb{E}_{\mathbf{A},X}\left[\frac{1}{n^m}\|\hat{\mathbf{\Theta}} - \bar{\mathbf{\Theta}}\|_F^2\right] \leq C\left\{\rho_n^{\frac{2m+2}{m+2}} n^{-\frac{2m}{m+2}} + \rho_n\left(\frac{\log n}{n^{m-1}}\right)\right\},$$

which completes the proof of Theorem 1. The proofs of Lemma 1 and 2 are more involved and are provided in the Appendix A. ■

Remark 3 *Note that for our result, we require the sparsity parameter $\rho_n > n^{-2/(m+1)}$. This corresponds to moderately sparse regime. Extending the result to the case of highly sparse regime (i.e., $\rho_n > n^{-(m-1)}$) is an interesting problem. It appears that the current proof technique is incapable of handling the highly sparse regime and a completely different approximation of the hypergraphon (for example, a multi-level approximation) with a parametric model might be required. This highlights another important distinction between the graphons and hypergraphons. We leave this interesting problem as future work.*

Remark 4 *Our result above could be extended to the case of α-Holder simple hypergraphons in a straight forward manner. In that case, there are two regimes in which our estimator behaves differently. For simplicity, consider the case when $\rho_n = 1$. In this case, we have*

$$\mathbb{E}_{\mathbf{A},X}\left[\frac{1}{n^m}\|\hat{\mathbf{\Theta}} - \bar{\mathbf{\Theta}}\|_F^2\right] = \begin{cases} n^{-\frac{2\alpha' m}{m+2\alpha}} & \text{when } 0 \le \alpha < 1 \\ n^{-\frac{2m}{m+2}} & \text{when } \alpha \ge 1. \end{cases}$$

In particular, for no value of α, the second term $(\frac{\log n}{n^{m-1}})$ becomes the dominant term. Furthermore, when $m = 2$, which corresponds to the case of graphons, for $\alpha \ge 1$, the first term and the second term are equivalent up to log factors and we recover the results of Klopp et al. (2017). We conjecture that our rates are essentially minimax optimal for the SLH when $\rho = 1$. In order to see that note that the problem of estimating a hypergraphon from the class SLH corresponds to the case of estimating a m-dimensional Lipschitz function. The minimax rate of non-parametric regression in this case essentially coincides with the rates we obtained (see for e.g., Giné and Nickl (2015) for a comprehensive overview of nonparametric models in the context of regression and density estimation). It is interesting to examine the minimax optimality of our result for the general case of ρ_n. A proof of minimax rates for general ρ has eluded us thus far.

4. Algorithm

Recall that our estimator $\hat{\mathbf{\Theta}}$, involves solving the following least-squares optimization problem

$$\hat{z}, \hat{\mathbf{Q}} = \operatorname*{argmin}_{z,Q} L(z, \mathbf{Q}), \quad \text{where} \quad L(z, \mathbf{Q}) = \sum_{j \in [n]^m} \left(\mathbf{A}_j - \mathbf{Q}_{z(j)}\right)^2.$$

An immediate algorithm to optimize the above objective function is to perform coordinate descent on the above objective function. This approach has the following drawback. While for a fixed z optimizing for $\mathbf{Q}$ is a standard least-squared problem, for a fixed $\mathbf{Q}$ optimizing for z evaluating all possible assignment functions $z \in \mathcal{Z}_{n,k}$, which would take exponential time and would thus be computationally infeasible even for moderate values of n, k. One could further break optimizing for z into optimizing for the individual coordinates $z(j)$. But we found through preliminary experiments such an approach runs into several issues like non-convergence of the iterates and worse performance (when it converges). In order to obtain a computable estimator and demonstrate our results empirically, we propose to use the procedure described in Algorithm 1.

Intuition: We now describe the main intuition behind the algorithm. Firstly note that when $a \in [k]$, η_a is the number of vertices which are assigned community a, and when

Algorithm 1 Alternating Minimization for hypergraphon

Input: $\mathbf{A}$, k.

repeat

Let E_{ia} be the number of hyperedges containing both i and a member of a according to the current $\hat{z}$.

$$E_{ia} = \sum_{j_2 \in \hat{z}^{(-1)}(a)} \sum_{j_3, \cdots, j_m \in [n]} \mathbf{A}_{ij_2 \cdots j_m} \tag{4}$$

Update $\hat{z}$ as

$$\hat{z}(i) = \operatorname*{argmax}_a \frac{1}{\varkappa_a} E_{ia}$$

where $\varkappa_a$ is defined in Equation 5.

until Convergence

Compute $\hat{\mathbf{Q}}_a = \frac{1}{\eta_a} \sum_{j \in \hat{z}^{(-1)}(a)} A_j$

Output: $\hat{\mathbf{Q}}_a$ and $\hat{z}$.

$a \in [k]^m$, η_a is the number of hyperedges whose community assignments match a node-wise. In other words, $j \in [n]^m$ matches $a \in [k]^m$ if there exists a permutation $\sigma \in \mathfrak{S}_k$, such that for all $i \in [n]$, $a_i = \sigma(z(j_i))$. With this observation, for a fixed z, we have $\mathbf{Q}_a = \frac{1}{\eta_\alpha} \sum_{j \in z^{-1}(a)} \mathbf{A}_j$. Furthermore for any minimizer $\hat{\mathbf{Q}}$ and $\hat{z}$ of $L(\mathbf{Q}, z)$ we have

$$\hat{\mathbf{Q}}_a = \frac{1}{\eta_\alpha} \sum_{j \in \hat{z}^{-1}(a)} \mathbf{A}_j.$$

Hence we concentrate first on estimate $\hat{z}$ and then use the above relation to obtain an estimate for $\hat{\mathbf{Q}}$ and subsequently $\hat{\mathbf{\Theta}}$.

The procedure in Algorithm 1 is motivated by the modified version of k-means type algorithm analyzed in Lu and Zhou (2016) for community detection in standard stochastic block model. Their algorithm repeatedly updates $z(i)$ according to the *fraction* of nodes in each community a that node i connects to. This can be viewed as the fraction of possible edges into a which i is a part of. But this intuition does not extend straight-forwardly to the case of hypergraphs. In order to extend their approach to the case of hypergraphs, we first propose to collapse the m-th order adjacency tensor $\mathbf{A}$ to the matrix $\mathcal{M}(\mathbf{A})$ (recall the definition from section 1.1). Now we define E_{ia}, the number of hyperedges containing for i and a member of a, according to the Equation 4. Note that is not exactly equal to the number of hyperedges containing for i and a member of a, as we are overcounting hyperedges which contain more than one element assigned community a. Even with this discrepancy, our algorithm performs well empirically as we demonstrate in section 5. This matrix is then used to estimate $z(i)$. Our update for $z(i)$ is based on the fraction of possible hyperedges that node i is a part of. The total number of possible hyperedges containing

both i and a member of community a is given by

$$\varkappa_a = \binom{\eta_a}{1}\binom{n-\eta_a}{m-2} + 2!\binom{\eta_a}{2}\binom{n-\eta_a}{m-3} + \ldots + (m-1)!\binom{\eta_a}{m-1}\binom{n-\eta_a}{0}, \quad (5)$$

where we count a hyperedge j times if it contains j elements from a, to match the over-counting of E_{ia}.

Initialization: Like most alternating methods, our algorithm is susceptible to local minima. To overcome this issue, we use two methods. First, we try many random initializations, running the algorithm to convergence, and then selecting the best result according to our empirical loss function $L(z, Q)$. Our second method is to initialize z using a spectral classifier, as given by Ghoshdastidar and Dukkipati (2017b). Their method is again based on the performing spectral clustering on the $\mathcal{M}(\mathbf{A})$ matrix, as this operation preserves the factors approximately. We report the results using both initialization methods in section 5 and discuss the similarities and differences between them.

We also note that the tensor collapse operation also bears similarities with the Leurgan's algorithm for decomposing tensors (Leurgans et al., 1993). The algorithm, proposed specifically for the case of three-way tensors, involves taking weighted sums of the slices. Note that in our case, there is no need for taking a weighted sum. Finally, the theoretical analysis of this algorithm is more involved than the standard graph based stochastic block model case. We plan to report the theoretical results in the context of community detection in hypergraphs in the near future.

5. Simulation Results: Well-specified Case

We now provide simulation results depicting the performance of our algorithm for the case when the random graphs are generated from the following two simple hypergraphons:

$$\text{Case 1: } f(u,v,w) = uvw$$
$$\text{Case 2: } f(u,v,w) = \frac{1}{1+e^{\{-(c_1u^2+c_2v^2+c_3w^2)\}}}$$

These experiments corresponds to well-specified modeling situation – we assumed the true graphon is a simple hypergraphon and we use the proposed algorithm to estimate the probability matrix. In Section 6, we also consider the misspecified case – here we test how well the algorithm performs for the case when the true hypergraphon is not a simple hypergraphon. Our error metric is the normalized L_2 reconstruction error (i.e., $\|\hat{\mathbf{\Theta}} - \bar{\mathbf{\Theta}}\|_F^2 / \|\bar{\mathbf{\Theta}}\|_F^2$). We use both the random and spectral initialization described in previous section. Our results for the well-specified case are reported in Figures 1 and 2. For each case, we performed the simulation for 50 independent trials and the average values are reported. The bars in the figures represent the corresponding standard error. The value of ρ_n was set to 1 and 0.7 in Figure 1 and 2 respectively, to correspond to the **dense** edges setting and **moderately sparse** edges situation. The value of k was fixed at $0.6 \cdot n^3$ and $0.5 \cdot n^3$ in Figure 1 and 2 respectively, based on the insight provided by the theorem. We experimented with both random initialization and spectral initialization as described in the previous section. Below are our observations from the experiments.

- The main difference between the two hypergraphons is that the one in case 1 could be thought of as a rank-1 function. Since our algorithm is based on the TENSORCOLLAPSE operation, the performance depends on how well such an operation captures the spectrum of the true function. It is easy to see that in the case of rank-1 function, the operation completely preserves the spectrum. Hence we expect our algorithm to perform well in case 1 than in case 2. Indeed this is reflect in Figures 1 and 2 for both the dense and sparse cases.

- The difference between the two initialization schemes is nearly negligible in the dense setting but the spectral initialization performs comparatively better in the sparse setting. We believe with a more sophisticated regularized spectral initialization one could achieve better performance. Precisely characterizing this in a theoretical framework is left a future work.

- As expected, as the number of nodes increases the expected reconstruction error decreases confirming the theoretical result presented in Theorem 1. Note that the value of k is set at a fixed value in the experiment.

5.1 Effect of number of blocks

Recall that our algorithm takes in as input k, the number of communities in the hypergraph stochastic block model that is used to approximate the hypergraphon. In Figure 3 the results of increasing the number of communities is depicted. The simulation setup is same as in Section 5 for the case of dense graphs; that is, $\rho_n = 1$. We considered $n = 20$ so that we could clearly observe the bias-variance tradeoff. For each case, we performed the simulation for 50 independent trials and the average values are reported. The bars in the figures represent the corresponding standard error. We use hypergraphs samples from both case 1 and 2 hypergraphons and increase k from $0.1 \cdot n^3$ to $0.9 \cdot n^3$ in steps of 0.1. The graphs in Figure 3 demonstrate a bias-variance tradeoff in terms of the effect of k on estimating the probability matrix. For lower values of k, the corresponding block model does not provide a good approximation for the underlying simple hypergraphons while for higher values of k, the number of nodes in each community is too less to provide any information for estimating the probabilities of edge formulation within that community. There is a sweet-spot in the middle which achieves the optimal value of k that balance the above bias and variance. This insight was used in the experiments presented in the previous section to fix the value of k.

5.2 Runtime comparison

Modeling higher-order interactions certainly comes at a computational cost compared to that of pairwise interactions. In this section, we compare the wall-clock times of random-initialization method for the above experiments. We consider random-initialization method as we observe that it performs as good as the one with spectral initialization. The runtimes of the spectral-initialization method would take into account the time for obtaining the initialization, in addition.

All the simulation settings are as described in Section 5. Figure 4 (left) shows the wall-clock time of our proposed algorithm as a function of number of nodes for a fixed value of

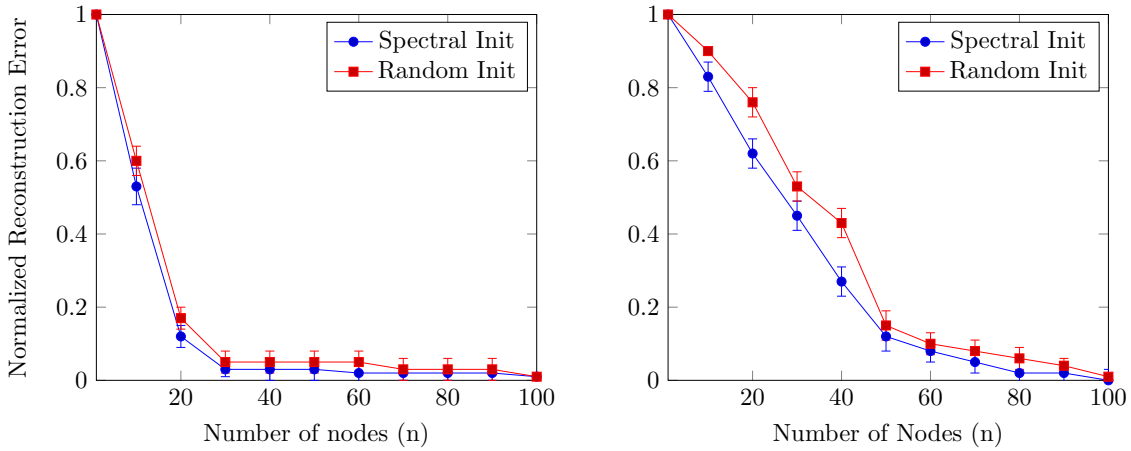

Figure 1: The graphs correspond to normalized average estimation error, for estimating the probability matrix, as a function of the number of nodes. The left hand side corresponds to when the hypergraph was sampled from hypergraphon in case 1 and the right hand side corresponds to the case 2. The value of ρ_n was set to 1 and hence this corresponds to the **dense** edges situation. Each point in the graphs above corresponds to a average over 50 independent trials and the bars represent the corresponding standard error. The value of k was fixed at $0.6 \cdot n^3$ (based on the insight provided by the theorem).

k, corresponding to the hypergraphon in case 1 and case 2 in both the dense and sparse settings. Similarly, Figure 4 (right) shows the wall-clock time of our proposed algorithm as a function of k for a fixed value of n for the dense setting. We notice that for a fixed k, as n is increased, the wall-clock time increases near-linearly first and then starts to increase rapidly. For a fixed n, as k increases the wall-clock time increases near-linearly. It is intriguing to come up with faster implementations of the algorithm, or even better, faster algorithms for estimating SLH hypergraphons.

6. Simulation Results: Misspecified Case

In this section we motivate the use of simple hypergraphons by demonstrating that they can be good approximations to general hypergraphons. We consider the most elementary hypergraphons for this experiment – piecewise constant hypergraphons. First we specify the model through a multistage construction, and then formalize the model by constructing its associated hypergraphon. We keep $m = 3$ fixed for this experiment.

Let n be the number of vertices, and k be the number of communities. We have four additional parameters, (p_1, q_1, p_2, q_2). Intuitively speaking, depending on vertex communities, we will construct edges between pairs of vertices, which can be viewed as latent variables, and then construct hyperedges between triples of vertices depending on those latent edges, but not directly on the communities of the constituent nodes. First we assign each vertex to one of the k groups uniformly at random. Then between each pair of vertices, we put an edge with probability p_1 if the vertices belong to the same community, and with probability q_1 otherwise. Next, between each triple of vertices, we put a hyperedge with probability p_2

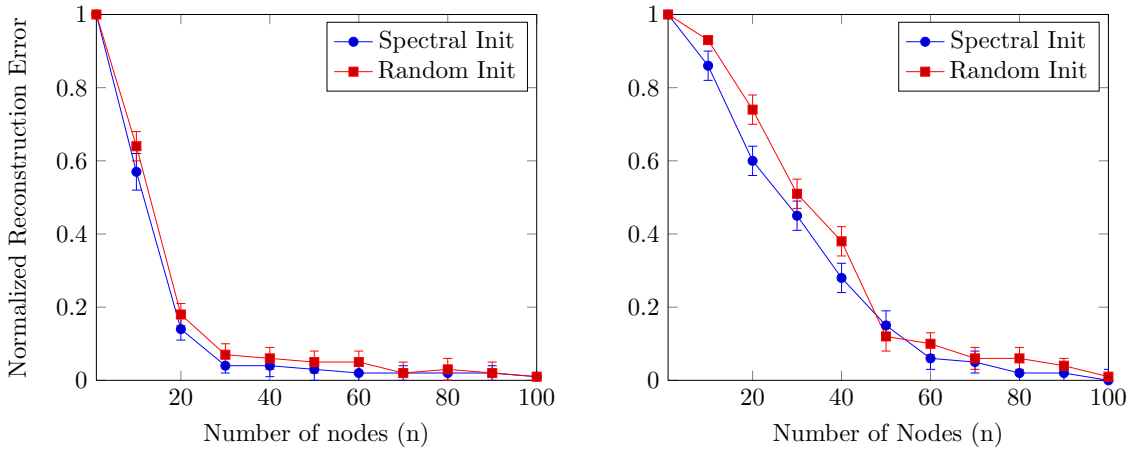

Figure 2: The graphs correspond to normalized average estimation error, for estimating the probability matrix, as a function of the number of nodes. The left hand side corresponds to when the hypergraph was sampled from hypergraphon in case 1 and the right hand side corresponds to the case 2. The value of ρ_n was set to 0.7 and hence this corresponds to the **sparse** edges situation. Each point in the graphs above corresponds to a average over 50 independent trials and the bars represent the corresponding standard error. The value of k was fixed at $0.5 \cdot n^3$ (based on the insight provided by the theorem).

if all three latent edges are present, and with probability q_2 otherwise. See Zhao (2015) for a hypergraphon construction of a similar flavor.

We now construct a (six-dimensional) hypergraphon which matches the distribution of the above model. We break the function into ten cases which correspond to the more intuitive procedural construction we just considered. We begin by defining the sets $A_1, A_2, \ldots, A_5$ and $B_1, B_2, \ldots, B_5$. To aid in our exposition, let I_x be the i such that $x \in \left[\frac{i-1}{k}, \frac{i}{k}\right)$. $I_x = I_y$ will mean that the vertices corresponding to x and y are in the same community. The set A_1 corresponds to the event where all three vertices are in the same community. A_2 will correspond to the event where the first two vertices are in the same community, A_3 to where the first and third are in the same community, and A_4 when the second and third are. A_5 corresponds to the event where all of the vertices are in the same community. For every assignment of communities to vertices, we have different probabilities of all three edges being present. When all of the edges are present, we have a probability p_2 of the hyperedge being in the graph, a probability of q_2 otherwise. B_i will correspond to the event where all three edges are present in case A_i. Finally, we define our hypergraphon f as displayed in Figure 5.

6.1 Results

We considered two cases for mis-specified models: case of $p_1 = 0.6$ and $q_1 = 0.4$ and the case of $p_1 = 0.8$ and $q_1 = 0.4$. The number of nodes n was set to be 100. We found that the the choice of $k = 0.5 \cdot n^3$ gave the best result. We emphasize here that this choice was found through trial and error as there is no supporting theory in the mis-specified case. Our results for estimation are plotted in Figures 5. We find that in certain regimes of parameters, our

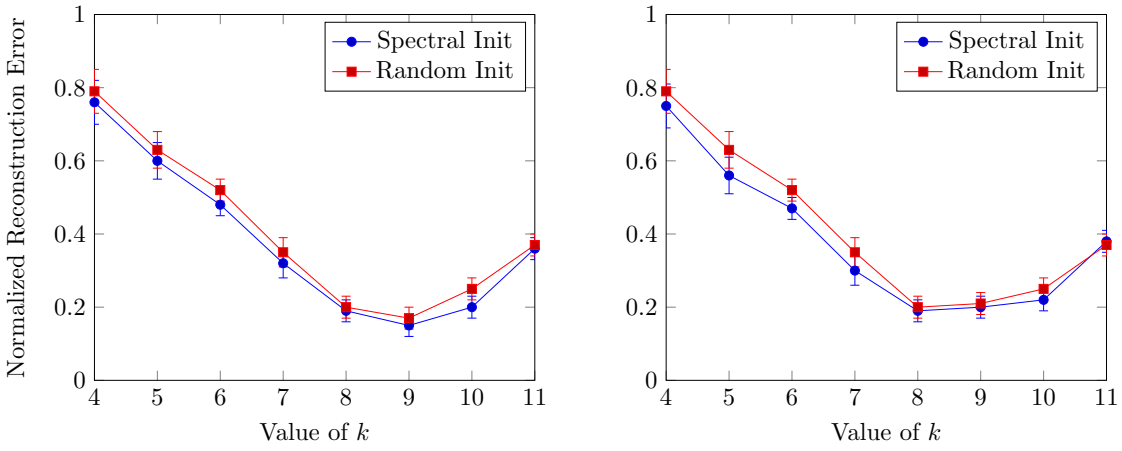

Figure 3: The graphs correspond to normalized average estimation error, for estimating the probability matrix, as a function of increasing the number of blocks. The left hand side corresponds to when the hypergraph was sampled from hypergraphon in case 1 and the right hand side corresponds to the case 2. The value of ρ_n was set to 1 and hence this corresponds to the dense edges situation. Each point in the graphs above corresponds to a average over 50 independent trials and the bars represent the corresponding standard error. The x-axis is to be interpreted after scaling it with n^3.

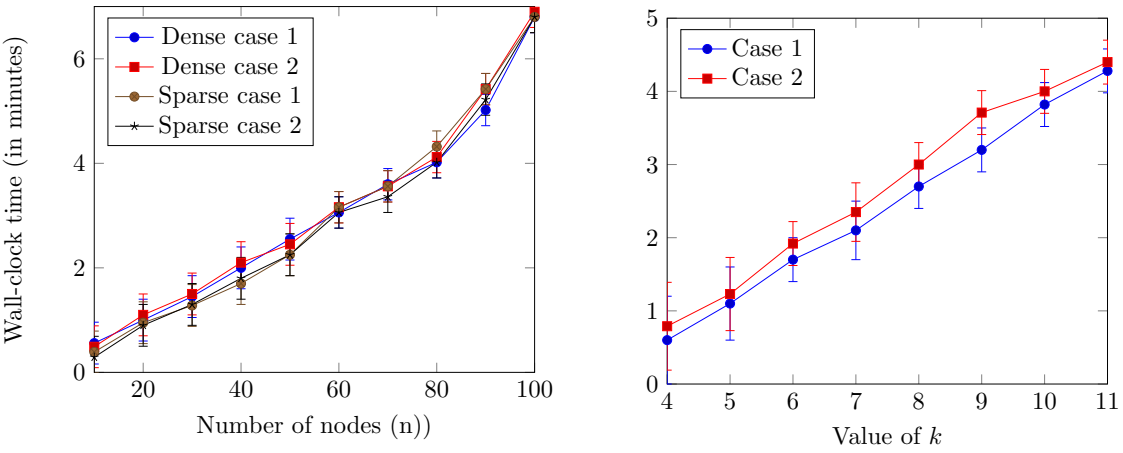

Figure 4: Wall-clock times of the experiments described in Section 5. The simulation settings are the same as before. Here, we only consider random-initialization based experiments. Case 1 and 2 corresponds to the two different hypergraphons considered and dense and sparse refers to the values of ρ_n set in the experiments.

estimation algorithm performs well. Firstly, when p_2 is close to q_2, the graphon resembles a constant hypergraphon and our estimation algorithm performs perfectly. To see this, notice that when $p_2 = q_2$, the resulting hypergraphon is in fact the constant hypergraphon with parameter $p_2 = q_2$. This suggests that our estimation algorithm works when the more complex hypergraphon is structurally close to a simple hypergraphon. We notice that as

$$
\begin{aligned}
&f(x_1,x_2,x_3,x_{12},x_{13},x_{23}) \\
&\quad = \begin{cases}
p_2 & \text{for } (x_1,x_2,x_3) \in A_1, (x_{12},x_{13},x_{23}) \in B_1 \\
q_2 & \text{for } (x_1,x_2,x_3) \in A_1, (x_{12},x_{13},x_{23}) \in B_1^C \\
p_2 & \text{for } (x_1,x_2,x_3) \in A_2, (x_{12},x_{13},x_{23}) \in B_2 \\
q_2 & \text{for } (x_1,x_2,x_3) \in A_2, (x_{12},x_{13},x_{23}) \in B_2^C \\
p_2 & \text{for } (x_1,x_2,x_3) \in A_3, (x_{12},x_{13},x_{23}) \in B_3 \\
q_2 & \text{for } (x_1,x_2,x_3) \in A_3, (x_{12},x_{13},x_{23}) \in B_3^C \\
p_2 & \text{for } (x_1,x_2,x_3) \in A_4, (x_{12},x_{13},x_{23}) \in B_4 \\
q_2 & \text{for } (x_1,x_2,x_3) \in A_4, (x_{12},x_{13},x_{23}) \in B_4^C \\
p_2 & \text{for } (x_1,x_2,x_3) \in A_5, (x_{12},x_{13},x_{23}) \in B_5 \\
q_2 & \text{for } (x_1,x_2,x_3) \in A_5, (x_{12},x_{13},x_{23}) \in B_5^C
\end{cases}
\end{aligned}
$$

$$
\begin{aligned}
A_1 &= \{(x_1,x_2,x_3) \in [0,1]^3 \text{ s.t. } I_{x_1} = I_{x_2} = I_{x_3}\} \\
A_2 &= \{(x_1,x_2,x_3) \in [0,1]^3 \text{ s.t. } I_{x_1} = I_{x_2} \neq I_{x_3}\} \\
A_3 &= \{(x_1,x_2,x_3) \in [0,1]^3 \text{ s.t. } I_{x_1} = I_{x_3} \neq I_{x_2}\} \\
A_4 &= \{(x_1,x_2,x_3) \in [0,1]^3 \text{ s.t. } I_{x_2} = I_{x_3} \neq I_{x_1}\} \\
A_5 &= \{(x_1,x_2,x_3) \in [0,1]^3 \text{ s.t. } I_{x_1} \neq I_{x_2}, I_{x_1} \neq I_{x_3}, I_{x_2} \neq I_{x_3}\} \\
B_1 &= \{(x_{12},x_{13},x_{23}) \in [0,p_1) \times [0,p_1) \times [0,p_1)\} \\
B_2 &= \{(x_{12},x_{13},x_{23}) \in [0,p_1) \times [0,q_1) \times [0,q_1)\} \\
B_3 &= \{(x_{12},x_{13},x_{23}) \in [0,q_1) \times [0,p_1) \times [0,q_1)\} \\
B_4 &= \{(x_{12},x_{13},x_{23}) \in [0,q_1) \times [0,q_1) \times [0,p_1)\} \\
B_5 &= \{(x_{12},x_{13},x_{23}) \in [0,q_1) \times [0,q_1) \times [0,q_1)\}
\end{aligned}
$$

Figure 5: The full hypergraphon used for testing model-misspecification error.

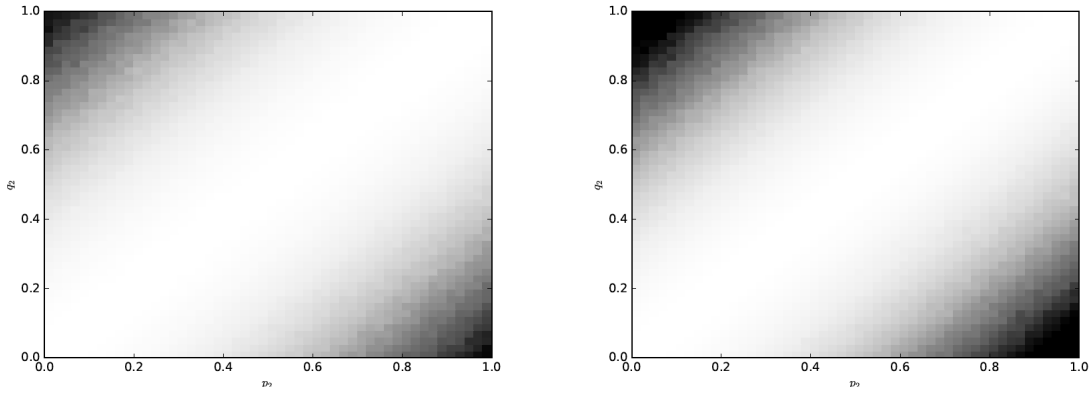

Figure 6: Left figure corresponds to the case of $p_1 = 0.6$ and $q_1 = 0.4$ respectively. Right figure corresponds to the case of $p_1 = 0.8$ and $q_1 = 0.4$. We note that the structure is similar, but as suggested by other figure, when p_1 and q_1 are big, the error is large. Whiter regions correspond to small normalize reconstruction error (around 0.2) and darker regions correspond to large estimation errors (around 0.8).

p_1, q_1 get bigger, so does the estimation error. This is because when p_1 and q_1 are both small, the resulting hypergraphon resembles the constant hypergraphon with parameter q_2, since edges are rare. For larger p_1, q_1, the resulting hypergraphon is too rich for our estimation procedure to handle, and we observe larger estimation error.

Finally, we observed that estimation succeeds when community detection fails, a common phenomenon since estimation is generally an easier problem than community detection. However, we also observe that community detection often succeeds when estimation fails, when $|p_2 - q_2|$ is large for example. This is telling because our estimation algorithm first does community detection, and estimates Q from the estimated communities. We can conclude that this is in fact model misspecification error, since estimation fails even when community detection succeeds.

7. Applications to Hyperedge Prediction

In this section, we demonstrate the applicability of the proposed hypergraphon estimator in Sections 3 and 4 to the problem of hyperedge prediction (also called hyperlink prdiction) in m-uniform hypergraphs. Here, we assume that we are not given the entire adjacency tensor $\mathbf{A}$. Instead, we only observe a fraction of entries; the set of observed entries is denoted as Ω. Based on the observed entries, the task is to predict the presence or absence of the missing hyperedges. In order to do so, we modify the estimator in (3) as $\hat{\mathbf{\Theta}} = \text{argmin}_{\Theta \in T_z} \{\|\Theta\|_F^2 - \frac{2n^m}{|\Omega|} \sum_{j \in \Omega} \mathbf{A}_j \mathbf{\Theta}_j\}$, in order to account for missing entries in the adjacency tensor. A similar method was also suggested by Gao et al. (2015) for the graph setting. Based on the estimated $\hat{\mathbf{\Theta}}$, we predict the missing hyperedge as present if the corresponding entry in the tensor $\hat{\mathbf{\Theta}}$ is greater than 0.5 and absent otherwise.

For our experiments, we use the the `GPS` dataset (Zheng et al., 2010) and the `MovieLens` dataset (Harper and Konstan, 2015). The `GPS` dataset consists of 146 users, at 70 locations, performing 5 different types of activities. This dataset consists of 1436 hyperedges in total

	GPS			MovieLens		
	70%	80%	90%	70%	80%	90%
CMM	0.64 ± 0.03	0.70 ± 0.05	0.79 ± 0.05	0.76 ± 0.06	0.82 ± 0.05	0.88 ± 0.06
C3MM	0.69 ± 0.04	0.76 ± 0.03	0.83 ± 0.05	0.81 ± 0.06	0.87 ± 0.06	0.90 ± 0.04
Hypergraphon	0.73 ± 0.03	0.76 ± 0.05	0.82 ± 0.02	0.85 ± 0.06	0.89 ± 0.04	0.90 ± 0.06

Table 1: Area under ROC curve for hyperedge prediction on `GPS` and `MovieLens` datasets.

that were 1s. For the `MovieLens` data, we used a smaller version with 500 users, 500 movies and 100 tags, for computational simplicity. This dataset consists of 8,349 hyperedges that were 1s. For the sake of experiments, we randomly pick 70%, 80% and 90% and treat them as observed data. The problem is to predict the missing hyperedges. The value of k for the hypergaphon method was set as the one that obtains the best prediction performance, by trial-and-error method. We remark that developing principled approaches (e.g., cross-validation) for hyperparameter selection with network data is an interesting problem (which has generated great interest in the community recently), which, however is beyond the scope of this work. We compare the performance of our method against two factorization-based methods termed as CMM (Zhang et al., 2018) and C3MM (Sharma et al., 2020). The above methods were chosen as they are recent works for hyperedge prediction with superior or comparable performance to several baselines. We performed the experiment for 50 trials and calculated the average Area under ROC curve (AUC) as a measure of performance. From the results in Table 1, we see that the hypergraphon based approach is comparable to the CMM and C3MM method with demonstrating superiority when the percentage of missing hyperedges in the given hypergraph is large.

8. Discussion

In this paper, we initiated the study of hypergraphon for nonparametric modeling of hypergraphs. We provided rates of convergence in expectation for estimating a class of Smooth Lipschitz Hypergraphons (SLH) and provided practical algorithms for implementing the estimators. There are several directions for future work, some of which we highlight below.

Estimating the graphon function: Note that in this paper, we consider estimating the probability of formation of hyperedges in the L_2 norm. A more challenging problem is to estimate the hypergraphon function itself from the given adjacency tensor. In order to do so, one needs to define appropriate norms on the space of hypergraphons. For the case of graphons, the cut-norm, defined as

$$\|f\|_{\square} = \sup_{S,T \subseteq [0,1]} \left| \int_{S \times T} f(x,y)\, dx\, dy \right|,$$

where S and T are measurable subset of $[0,1]$ serves as a good metric; see Klopp and Verzelen (2019). But in the case of hypergraphons the problem is more delicate. A straight forward generalization of the above cut norms (defined below for simple 3-uniform hypergraphons)

$$\|f\|_{\square} = \sup_{S,T,U \subseteq [0,1]} \left| \int_{S \times T \times U} f(x,y,z)\, dx\, dy\, dz \right|,$$

provides only a weaker metric on the space of simple hypergraphons. Care must be taken first to define first a metric that is meaningful; see Zhao (2015). We leave this problem of providing cut-norm convergence results for estimating hypergraphons as future work.

Computational Theory: The algorithm presented in section 4 works well as shown via the experimental results, surprisingly also with random initialization in spite of being a non-convex optimization problem. Recent advances in non-convex optimization, also highlights a similar phenomenon holds for other models; see, for example, Chen et al. (2019) for the case of phase retrieval model. It is interesting to leverage such results in the context of hypergraphon models. Furthermore, for the case of graphons, Zhang et al. (2017) proposed a neighborhood smoothing approach in particular for estimating the probability matrix. It is not clear if such an approach is immediately extendable for hypergraphons. But it is interesting to explore if a similar approach could be leveraged for hypergraphons.

2-layer interaction model: A direct step-function (or SBM) based approximation of the full hypergraphon is provided in Equation 13 in Zhao (2015). It would be interesting to explore to use a layer wise approach based on the Q-step approximation to estimate the full hypergraphon function in the future.

Hierarchical clustering: Based on the idea of graphons, recently a hierarchical clustering framework based was proposed in Eldridge et al. (2016). Also model-free interpretations of the standard clustering algorithm was provided in Diao et al. (2016). It would extremely interesting to explore similar extension to the case of hypergraphons.

Appendix A. Proofs for Section 3

In this section, we provide the proofs for Lemma 1 and Lemma 2 from Section 3. Throughout the proof we assume that C is a absolute constant that changes from step to step. Furthermore, in this section, we denote $\mathbb{E}_{\mathbf{A},X}$ and $\mathbb{E}_X$, as just $\mathbb{E}$ for simplicity, as it will be clear from the context. We also require some additional notations: Recall that $G \subset [n]^m$ denotes the set of hyperedges with no repeated vertices. That is, $j \notin G$ iff $\exists\ u, v$ s.t $j_u = j_v$. We define $H \in [n]^m$ to be the set of increasing hyperedges. That is, $j \in H$ iff $j_1 < j_2 < \ldots < j_m$. Then, note that $|H| = \binom{n}{m}$ and $|G| = m!|H|$.

Proof [Proof of Lemma 1] First, we consider the loss function

$$L(\mathbf{Q}, z) = \frac{1}{2} \sum_{a \in [k]^m} \sum_{j \in z^{-1}(a) \cap H} (\mathbf{A}_j - \mathbf{Q}_a)^2$$

Let $\hat{\mathbf{Q}}$, $\hat{z}$ be minimizers of $L(\mathbf{Q}, z)$ and let $\hat{\mathbf{\Theta}}$ be such that $\hat{\mathbf{\Theta}}_j = \hat{\mathbf{Q}}_{\hat{z}(j)}$. Note that this is a least squares estimator. Let $\mathbf{\Theta}_*$ be the best k-class block model approximation of $\bar{\mathbf{\Theta}}$ in Frobenius norm. In the case where $\bar{\mathbf{\Theta}}$ is truly a k-class block model, $\mathbf{\Theta}_* = \bar{\mathbf{\Theta}}$. Since $\hat{\mathbf{\Theta}}$ is a

least squares estimator, we have the following set of inequalities:

$$\begin{aligned}
&\|\hat{\mathbf{\Theta}} - \mathbf{A}\|_F^2 \leq \|\mathbf{\Theta}_* - \mathbf{A}\|_F^2 \\
\Leftrightarrow &\|\hat{\mathbf{\Theta}} - \bar{\mathbf{\Theta}}\|_F^2 + 2\langle \hat{\mathbf{\Theta}} - \bar{\mathbf{\Theta}}, \bar{\mathbf{\Theta}} - \mathbf{A}\rangle + \|\bar{\mathbf{\Theta}} - \mathbf{A}\|_F^2 \leq \|\mathbf{\Theta}_* - \bar{\mathbf{\Theta}}\|_F^2 + 2\langle \mathbf{\Theta}_* - \bar{\mathbf{\Theta}}, \bar{\mathbf{\Theta}} - \mathbf{A}\rangle + \|\bar{\mathbf{\Theta}} - \mathbf{A}\|_F^2 \\
\Leftrightarrow &\|\hat{\mathbf{\Theta}} - \bar{\mathbf{\Theta}}\|_F^2 \leq \|\mathbf{\Theta}_* - \bar{\mathbf{\Theta}}\|_F^2 + 2\langle \hat{\mathbf{\Theta}} - \bar{\mathbf{\Theta}}, \mathbf{A} - \bar{\mathbf{\Theta}}\rangle + 2\langle \bar{\mathbf{\Theta}} - \mathbf{\Theta}_*, \mathbf{A} - \bar{\mathbf{\Theta}}\rangle \\
\Leftrightarrow &\|\hat{\mathbf{\Theta}} - \bar{\mathbf{\Theta}}\|_F^2 \leq \|\mathbf{\Theta}_* - \bar{\mathbf{\Theta}}\|_F^2 + 2\langle \hat{\mathbf{\Theta}} - \bar{\mathbf{\Theta}}, \mathbf{E}\rangle + 2\langle \bar{\mathbf{\Theta}} - \mathbf{\Theta}_*, \mathbf{E}\rangle,
\end{aligned}$$

where $\mathbf{E} = \mathbf{A} - \bar{\mathbf{\Theta}}$ is the noise tensor. Now we are interesting in bounding the expectation of the left hand side. Since $\mathbf{\Theta}_*$ and $\bar{\mathbf{\Theta}}$ are deterministic and $\mathbb{E}[\mathbf{E}] = 0$, the final summand has zero mean, so it suffices to bound the expectation of $\langle \hat{\mathbf{\Theta}} - \bar{\mathbf{\Theta}}, \mathbf{E}\rangle$. For any $z \in \mathcal{Z}_{n.k}$, let $\tilde{\mathbf{\Theta}}_z$ be the best Frobenius norm approximation of $\bar{\mathbf{\Theta}}$ in the collection of rank m tensors

$$\mathcal{T}_z = \{\mathbf{\Theta} : \exists \mathbf{Q} \in \mathbb{R}_{\text{sym}}^{k\times\cdots\times k} \text{ such that } \mathbf{\Theta}_j = Q_{z(j)} \text{ for } j \in G, \text{ and } \mathbf{\Theta}_j = 0 \text{ otherwise}\}$$

Note that we can interpret $\mathcal{T}_z$ as the collection of $\mathbf{\Theta}$'s which correspond to a k-class block model with classes given by z. More concretely, $\tilde{\mathbf{\Theta}}_z$ is constructed by taking blockwise averages of $\bar{\mathbf{\Theta}}$ according to z, the same way $\hat{\mathbf{\Theta}}_z$ is constructed by taking blockwise averages of $\mathbf{A}$ with respect to z. With this interpretation, we have the decomposition

$$\langle \hat{\mathbf{\Theta}} - \bar{\mathbf{\Theta}}, \mathbf{E}\rangle = \langle \tilde{\mathbf{\Theta}}_z - \bar{\mathbf{\Theta}}, \mathbf{E}\rangle + \langle \hat{\mathbf{\Theta}} - \tilde{\mathbf{\Theta}}_z, \mathbf{E}\rangle.$$

Here the first summand, $\langle \tilde{\mathbf{\Theta}}_z - \bar{\mathbf{\Theta}}, \mathbf{E}\rangle$ corresponds to the error incurred from misclustering, and $\langle \hat{\mathbf{\Theta}} - \tilde{\mathbf{\Theta}}_z, \mathbf{E}\rangle$ corresponds to error from Bernoulli noise. We bound each of the two terms separately next.

Control of $\langle \tilde{\mathbf{\Theta}}_z - \bar{\mathbf{\Theta}}, \mathbf{E}\rangle$: We first express $\langle \tilde{\mathbf{\Theta}}_{\hat{z}} - \bar{\mathbf{\Theta}}, \mathbf{E}\rangle$ as a sum of independent random variables. To do this we only sum over j which are in increasing order, and multiply by the appropriate factor $m!$. Hence we have

$$\langle \tilde{\mathbf{\Theta}}_{\hat{z}} - \bar{\mathbf{\Theta}}, \mathbf{E}\rangle = \sum_{j\in[n]^m} (\tilde{\mathbf{\Theta}}_{\hat{z}} - \bar{\mathbf{\Theta}})_j \mathbf{E}_j = \sum_{j\in H} m!(\tilde{\mathbf{\Theta}}_{\hat{z}} - \bar{\mathbf{\Theta}})_j \mathbf{E}_j.$$

Note $|H| = \binom{n}{m}$, and that the summands are only positive in G. We use the above decomposition, and Bernstein's inequality (see Theorem 2 below), for a fixed, deterministic z to get following tail bounds:

$$\begin{aligned}
&\mathbb{P}\left(\langle \tilde{\mathbf{\Theta}}_z - \bar{\mathbf{\Theta}}, \mathbf{E}\rangle \geq \sqrt{2t\sum_{j\in H} \text{Var}(m!(\tilde{\mathbf{\Theta}}_z - \bar{\mathbf{\Theta}})_j \mathbf{E}_j)} + \frac{2m!\|\tilde{\mathbf{\Theta}}_z - \bar{\mathbf{\Theta}}\|_\infty}{3}\right) \leq e^{-t} \\
\Leftrightarrow &\mathbb{P}\left(\langle \tilde{\mathbf{\Theta}}_z - \bar{\mathbf{\Theta}}, \mathbf{E}\rangle \geq \sqrt{2t(m!)^2\sum_{j\in H} (\tilde{\mathbf{\Theta}}_z - \bar{\mathbf{\Theta}})_j^2 \text{Var}(\mathbf{E}_j)} + \frac{2m!\|\tilde{\mathbf{\Theta}}_z - \bar{\mathbf{\Theta}}\|_\infty}{3}\right) \leq e^{-t}.
\end{aligned}$$

Now note that the variance in the above expression is bound as

$$\text{Var}(\mathbf{E}_j) = (\bar{\mathbf{\Theta}})_j(1 - (\bar{\mathbf{\Theta}})_j) \leq (\bar{\mathbf{\Theta}})_j \leq \|\bar{\mathbf{\Theta}}\|_\infty.$$

Hence substituting this expression, we have

$$\begin{aligned}
&\mathbb{P}\left(\langle\tilde{\mathbf{\Theta}}_z-\bar{\mathbf{\Theta}},\mathbf{E}\rangle\geq\sqrt{2t(m!)^2\sum_{j\in H}(\tilde{\mathbf{\Theta}}_z-\bar{\mathbf{\Theta}})_j^2\|\bar{\mathbf{\Theta}}\|_\infty}+\frac{2m!\|\tilde{\mathbf{\Theta}}_z-\bar{\mathbf{\Theta}}\|_\infty}{3}\right)\leq e^{-t}\\
\Leftrightarrow&\mathbb{P}\left(\langle\tilde{\mathbf{\Theta}}_z-\bar{\mathbf{\Theta}},\mathbf{E}\rangle\geq\sqrt{2t\|\bar{\mathbf{\Theta}}\|_\infty(m!)^2\sum_{j\in H}(\tilde{\mathbf{\Theta}}_z-\bar{\mathbf{\Theta}})_j^2}+\frac{2m!\|\tilde{\mathbf{\Theta}}_z-\bar{\mathbf{\Theta}}\|_\infty}{3}\right)\leq e^{-t}\\
\Leftrightarrow&\mathbb{P}\left(\langle\tilde{\mathbf{\Theta}}_z-\bar{\mathbf{\Theta}},\mathbf{E}\rangle\geq\sqrt{2t\|\bar{\mathbf{\Theta}}\|_\infty m!\|\tilde{\mathbf{\Theta}}_z-\bar{\mathbf{\Theta}}\|_F^2}+\frac{2m!\|\tilde{\mathbf{\Theta}}_z-\bar{\mathbf{\Theta}}\|_\infty}{3}\right)\leq e^{-t}\\
\Leftrightarrow&\mathbb{P}\left(\langle\tilde{\mathbf{\Theta}}_z-\bar{\mathbf{\Theta}},\mathbf{E}\rangle\geq\|\tilde{\mathbf{\Theta}}_z-\bar{\mathbf{\Theta}}\|_F\sqrt{2m!\|\bar{\mathbf{\Theta}}\|_\infty t}+\frac{2m!\|\tilde{\mathbf{\Theta}}_z-\bar{\mathbf{\Theta}}\|_\infty}{3}\right)\leq e^{-t}.
\end{aligned}$$

Now we use a union bound to get a uniform in z version of the above expression tail bound. To do so, let

$$A_z=\left\{\langle\tilde{\mathbf{\Theta}}_z-\bar{\mathbf{\Theta}},\mathbf{E}\rangle\geq\|\tilde{\mathbf{\Theta}}_z-\bar{\mathbf{\Theta}}\|_F\sqrt{2m!\|\bar{\mathbf{\Theta}}\|_\infty t}+\frac{2m!\|\tilde{\mathbf{\Theta}}_z-\bar{\mathbf{\Theta}}\|_\infty}{3}\right\},$$

and notice that $A_{\hat{z}}\subset\bigcup_{z\in\mathcal{Z}_{n,k}}A_z$. We proceed with a union bound to get the following bound.

$$\mathbb{P}(A_{\hat{z}})\leq\mathbb{P}\left(\bigcup_{z\in\mathcal{Z}_{n,k}}A_z\right)\leq\sum_{z\in\mathcal{Z}_{n,k}}\mathbb{P}(A_z)\leq k^n e^{-t}=e^{-(t-n\log k)}.$$

Substituting $\tilde{t}=t-n\log k$ and then relabeling $\hat{t}$ by t yields the following:

$$\begin{aligned}
&\mathbb{P}\left(\langle\tilde{\mathbf{\Theta}}_{\hat{z}}-\bar{\mathbf{\Theta}},\mathbf{E}\rangle\geq\|\tilde{\mathbf{\Theta}}_{\hat{z}}-\bar{\mathbf{\Theta}}\|_F\sqrt{2m!\|\bar{\mathbf{\Theta}}\|_\infty t}+\frac{2m!\|\tilde{\mathbf{\Theta}}_{\hat{z}}-\bar{\mathbf{\Theta}}\|_\infty}{3}\right)\leq e^{-(t-n\log k)}\\
\Leftrightarrow&\mathbb{P}\left(\langle\tilde{\mathbf{\Theta}}_{\hat{z}}-\bar{\mathbf{\Theta}},\mathbf{E}\rangle\geq\|\tilde{\mathbf{\Theta}}_{\hat{z}}-\bar{\mathbf{\Theta}}\|_F\sqrt{2m!\|\bar{\mathbf{\Theta}}\|_\infty(\tilde{t}+n\log k)}+\frac{2m!\|\tilde{\mathbf{\Theta}}_{\hat{z}}-\bar{\mathbf{\Theta}}\|_\infty}{3}\right)\leq e^{-\tilde{t}}\\
\Leftrightarrow&\mathbb{P}\left(\langle\tilde{\mathbf{\Theta}}_{\hat{z}}-\bar{\mathbf{\Theta}},\mathbf{E}\rangle\geq\|\tilde{\mathbf{\Theta}}_{\hat{z}}-\bar{\mathbf{\Theta}}\|_F\sqrt{2m!\|\bar{\mathbf{\Theta}}\|_\infty(t+n\log k)}+\frac{2m!\|\tilde{\mathbf{\Theta}}_{\hat{z}}-\bar{\mathbf{\Theta}}\|_\infty}{3}\right)\leq e^{-t}.
\end{aligned}$$

Since $\tilde{\mathbf{\Theta}}_{\hat{z}}$ is an averaging of $\mathbf{\Theta}$ over blocks, we have $\|\tilde{\mathbf{\Theta}}_{\hat{z}}-\bar{\mathbf{\Theta}}\|_\infty$. Furthermore, using the inequality $2uv\leq u^2+v^2$, we have

$$\|\tilde{\mathbf{\Theta}}_{\hat{z}}-\bar{\mathbf{\Theta}}\|_F\sqrt{2m!\|\bar{\mathbf{\Theta}}\|_\infty(t+n\log k)}\leq\frac{1}{8}\|\tilde{\mathbf{\Theta}}_{\hat{z}}-\bar{\mathbf{\Theta}}\|_F^2+4m!\|\bar{\mathbf{\Theta}}\|_\infty(t+n\log k).$$

Combining these two observations yields

$$\begin{aligned}
&\mathbb{P}\left(\langle\tilde{\mathbf{\Theta}}_{\hat{z}}-\bar{\mathbf{\Theta}},\mathbf{E}\rangle\geq\frac{1}{8}\|\tilde{\mathbf{\Theta}}_{\hat{z}}-\bar{\mathbf{\Theta}}\|_F^2+4m!\|\bar{\mathbf{\Theta}}\|_\infty(t+n\log k)+\frac{2m!\|\bar{\mathbf{\Theta}}\|_\infty}{3}\right)\leq e^{-t}\\
\Leftrightarrow&\mathbb{P}\left(\langle\tilde{\mathbf{\Theta}}_{\hat{z}}-\bar{\mathbf{\Theta}},\mathbf{E}\rangle-\frac{1}{8}\|\tilde{\mathbf{\Theta}}_{\hat{z}}-\bar{\mathbf{\Theta}}\|_F^2\geq4m!\|\bar{\mathbf{\Theta}}\|_\infty(t+n\log k)+\frac{2m!\|\bar{\mathbf{\Theta}}\|_\infty}{3}\right)\leq e^{-t}.
\end{aligned}$$

Finally letting $u = 4m!\|\bar{\mathbf{\Theta}}\|_\infty(t + n\log k) + \frac{2m!\|\bar{\mathbf{\Theta}}\|_\infty}{3}$, we obtain

$$\mathbb{P}\left(\langle\tilde{\mathbf{\Theta}}_{\hat{z}} - \bar{\mathbf{\Theta}}, \mathbf{E}\rangle - \frac{1}{8}\|\tilde{\mathbf{\Theta}}_{\hat{z}} - \bar{\mathbf{\Theta}}\|_F^2 \geq u\right) \leq e^{n\log k + 1/6 - \frac{3u}{4m!\|\bar{\mathbf{\Theta}}\|_\infty}}.$$

Now we proceed to obtain the following expectation bound from the above probability bound. Integrating both sides with respect to u gives us

$$\begin{aligned}
\mathbb{E}[\langle\tilde{\mathbf{\Theta}}_{\hat{z}} - \bar{\mathbf{\Theta}}, \mathbf{E}\rangle - \frac{1}{8}\|\tilde{\mathbf{\Theta}}_{\hat{z}} - \bar{\mathbf{\Theta}}\|_F^2] &\leq \int_0^{u^*} \mathbb{P}\left(\langle\tilde{\mathbf{\Theta}}_{\hat{z}} - \bar{\mathbf{\Theta}}, \mathbf{E}\rangle - \frac{1}{8}\|\tilde{\mathbf{\Theta}}_{\hat{z}} - \bar{\mathbf{\Theta}}\|_F^2 \geq u\right) du \\
&\quad + \int_{u^*}^{\infty} \mathbb{P}\left(\langle\tilde{\mathbf{\Theta}}_{\hat{z}} - \bar{\mathbf{\Theta}}, \mathbf{E}\rangle - \frac{1}{8}\|\tilde{\mathbf{\Theta}}_{\hat{z}} - \bar{\mathbf{\Theta}}\|_F^2 \geq u\right) du \\
&\leq u^* + \int_{u^*}^{\infty} e^{n\log k + 1/6 - \frac{3u}{4m!\|\bar{\mathbf{\Theta}}\|_\infty}} du \\
&= u^* + \frac{m!\|\bar{\mathbf{\Theta}}\|_\infty}{2} e^{n\log k + 1/6 - \frac{3u}{4m!\|\bar{\mathbf{\Theta}}\|_\infty}}\Big|_\infty^{u^*} \\
&= u^* + \frac{m!\|\bar{\mathbf{\Theta}}\|_\infty}{2} e^{n\log k + 1/6 - \frac{3u^*}{4m!\|\bar{\mathbf{\Theta}}\|_\infty}}
\end{aligned}$$

Choosing $u^* = Cm!\|\bar{\mathbf{\Theta}}\|_\infty n\log k$ yields the following:

$$\mathbb{E}[\langle\tilde{\mathbf{\Theta}}_{\hat{z}} - \bar{\mathbf{\Theta}}, \mathbf{E}\rangle - \frac{1}{8}\|\tilde{\mathbf{\Theta}}_{\hat{z}} - \bar{\mathbf{\Theta}}\|_F^2] \leq Cm!\|\bar{\mathbf{\Theta}}\|_\infty n\log k \tag{6}$$

which completes the control of $\langle\tilde{\mathbf{\Theta}}_z - \bar{\mathbf{\Theta}}, \mathbf{E}\rangle$. We now proceed to control the second term.

Control of $\langle\hat{\mathbf{\Theta}}_z - \tilde{\mathbf{\Theta}}_z, \mathbf{E}\rangle$: Controlling this term is more involved than the first term. We give a brief overview of the steps need to get the required bound. We first construct a 1/4 net to ensure closeness in Frobenius norm, and then modify the net slightly to also ensure closeness in infinity norm. After that we follow the standard Bernstein pipeline as in the control of the first term. Then we handle a more involved bound on $\|\tilde{\mathbf{A}}_{\hat{z}} - \tilde{\mathbf{\Theta}}_{\hat{z}}\|_\infty$. We elaborate on the steps below.

For any $z \in \mathcal{Z}_{n,k}$, define $\mathcal{T}_{z,1} = \{\mathbf{\Theta} \in \mathcal{T}_z : \|\bar{\mathbf{\Theta}}\|_F \leq 1\}$ to be the set of all tensors whose block structure is determined by z, and whose Frobenius norm is bounded by 1. Denote by $\tilde{\mathbf{A}}_z$ the best Frobenius norm approximation of $\mathbf{A}$ in $\mathcal{T}_z$, i.e. the blockwise average of $\mathbf{A}$ according to z. Then $\tilde{\mathbf{E}}_z = \tilde{\mathbf{A}}_z - \tilde{\mathbf{\Theta}}_z$ is also the projection and blockwise average of $\mathbf{E}$ onto/of $\mathcal{T}_z$. By properties of projections, $\mathbf{\Theta} = \frac{\tilde{\mathbf{E}}_z}{\|\tilde{\mathbf{E}}_z\|_F}$ maximizes $\langle\mathbf{\Theta}, \mathbf{E}\rangle$ over $\mathcal{T}_{z,1}$. Let C_z be a minimal 1/4 net of $\mathcal{T}_{z,1}$ in Frobenius norm. Now to each $\mathbf{V} \in C_z$ associate

$$\tilde{\mathbf{V}} = \operatorname*{argmin}_{\mathbf{\Theta} \in \mathcal{T}_{z,1} \cap B(\mathbf{V}, 1/4)} \|\mathbf{\Theta}\|_\infty.$$

Define $\tilde{C}_z = \{\tilde{\mathbf{V}} : \mathbf{V} \in C_z\}$. A standard bound on covering numbers implies $\log|\tilde{C}_z| \leq Ck^m$. Now using Bernstein inequality (see Theorem 2) and a union bound over $z \in \mathcal{Z}_{n,k}$ and $\mathbf{\Theta} \in C_z$, we have,

$$\mathbb{P}\left(\langle\mathbf{\Theta}, \mathbf{E}\rangle \geq \sqrt{2m!\|\bar{\mathbf{\Theta}}\|_\infty(t + n\log k + k^m)} + \frac{2m!\|\mathbf{\Theta}\|_\infty}{3}(t + n\log k + k^m)\right) \leq e^{-t},$$

where we used the fact that $\|\mathbf{\Theta}\|_F \leq 1$ on $\tilde{C}_z$. By definition of $\tilde{C}_z$, there is a $\mathbf{\Theta} \in \tilde{C}_z$ such that

$$\left\| \mathbf{\Theta} - \frac{\tilde{\mathbf{E}}_{\hat{z}}}{\|\tilde{\mathbf{E}}_{\hat{z}}\|_F} \right\|_F \leq \frac{1}{2} \qquad \text{and} \qquad \|\mathbf{\Theta}\|_\infty \leq \frac{\|\tilde{\mathbf{E}}_{\hat{z}}\|_\infty}{\|\tilde{\mathbf{E}}_{\hat{z}}\|_F}.$$

Furthermore, for this $\mathbf{\Theta}$, we have $2\left(\frac{\tilde{\mathbf{E}}_{\hat{z}}}{\|\tilde{\mathbf{E}}_{\hat{z}}\|_F} - \mathbf{\Theta}\right)$ belongs to $\mathcal{T}_{z,1}$. Thus, we have

$$\left\langle 2\left(\frac{\tilde{\mathbf{E}}_{\hat{z}}}{\|\tilde{\mathbf{E}}_{\hat{z}}\|_F} - \mathbf{\Theta}\right), \mathbf{E} \right\rangle \leq \left\langle \frac{\tilde{\mathbf{E}}_{\hat{z}}}{\|\tilde{\mathbf{E}}_{\hat{z}}\|_F}, \mathbf{E} \right\rangle,$$

or equivalently,

$$\left\langle \frac{\tilde{\mathbf{E}}_{\hat{z}}}{\|\tilde{\mathbf{E}}_{\hat{z}}\|_F}, \mathbf{E} \right\rangle \leq 2\langle \mathbf{\Theta}, \mathbf{E} \rangle.$$

This gives us

$$\langle \tilde{\mathbf{E}}_{\hat{z}}, \mathbf{E} \rangle \leq 2\|\tilde{\mathbf{E}}_{\hat{z}}\|_F \sqrt{2m!\|\bar{\mathbf{\Theta}}\|_\infty (t + n\log k + k^m)} + \frac{4m!}{3}\|\tilde{\mathbf{E}}_{\hat{z}}\|_\infty (t + n\log k + k^m)$$

with probability at least $1 - e^{-t}$. Now we need a bound on $\|\tilde{\mathbf{E}}_{\hat{z}}\|_\infty$. Every entry of $\tilde{\mathbf{E}}_{\hat{z}}$ is an average over some block of $\mathbf{E}$, so we take a union bound over all blocks of all legal sizes. First we write what an entry of $\tilde{\mathbf{E}}_{\hat{z}}$ looks like:

$$\left[\tilde{\mathbf{E}}_{\hat{z}}\right]_j = \frac{\sum_{i \in G : \hat{z}(i) = \hat{z}(j)} \mathbf{E}_i}{|\{i \in G : \hat{z}(i) = \hat{z}(j)\}|}.$$

Note that we then have the following to be true

$$\|\tilde{\mathbf{E}}_{\hat{z}}\|_\infty \leq \sup_{s \in S} X_s \quad \text{and} \quad X_s = \sup_{V \subset [n]^m : |V_i| = s_i} \frac{\sum_{j \in G \cap V} \mathbf{E}_j}{|G \cap V|},$$

where $S = \{n_0, \ldots, n\}^m$.

First we bound $\frac{1}{|G \cap V|}\sum \mathbf{E}_j$ using Bernstein's inequality (see Theorem 2), and the inequality $2ab \leq a^2 + b^2$. We need to be careful that we only consider independent summands. We denote by $|\#j|$ the number of permutations of j which are contained in V, and use the trivial bound $|\#j| \leq m!$. Based on the above observations, we have the following set of

inequalities:

$$\begin{aligned}
&\mathbb{P}\left(\sum_{j\in G\cap V}\mathbf{E}_j \geq \sqrt{2t\sum_{j\in H\cap V}\mathrm{Var}(|\#j|X_j)}+\frac{2t}{3}\right)\leq e^{-t}\\
\Rightarrow&\mathbb{P}\left(\sum_{j\in G\cap V}\mathbf{E}_j \geq \sqrt{2tm!\sum_{j\in H\cap V}|\#j|\mathrm{Var}(X_j)}+\frac{2t}{3}\right)\leq e^{-t}\\
\Rightarrow&\mathbb{P}\left(\sum_{j\in G\cap V}\mathbf{E}_j \geq \sqrt{2tm!\sum_{j\in H\cap V}|\#j|\|\bar{\mathbf{\Theta}}\|_\infty}+\frac{2t}{3}\right)\leq e^{-t}\\
\Leftrightarrow&\mathbb{P}\left(\sum_{j\in G\cap V}\mathbf{E}_j \geq \sqrt{2tm!|G\cap V|\|\bar{\mathbf{\Theta}}\|_\infty}+\frac{2t}{3}\right)\leq e^{-t}\\
\Leftrightarrow&\mathbb{P}\left(\frac{1}{|G\cap V|}\sum_{j\in G\cap V}\mathbf{E}_j \geq \sqrt{\frac{2tm!}{|G\cap V|}\|\bar{\mathbf{\Theta}}\|_\infty}+\frac{2t}{3|G\cap V|}\right)\leq e^{-t}\\
\Rightarrow&\mathbb{P}\left(\frac{1}{|G\cap V|}\sum_{j\in G\cap V}\mathbf{E}_j \geq \|\bar{\mathbf{\Theta}}\|_\infty+\frac{tm!}{2|G\cap V|}+\frac{2t}{3|G\cap V|}\right)\leq e^{-t}\\
\Rightarrow&\mathbb{P}\left(\frac{1}{|G\cap V|}\sum_{j\in G\cap V}\mathbf{E}_j \geq \|\bar{\mathbf{\Theta}}\|_\infty+\frac{tm!}{|G\cap V|}\right)\leq e^{-t}
\end{aligned}$$

By Stirling's formula, we have $|\{V_i : |V_i| = s_i\}| \leq \binom{n}{s_i} \leq \left(\frac{en}{s_i}\right)^{s_i}$. Therefore, to bound X_s, we use a union bound over possible V_i, a set which has cardinality at most $\prod_{i=1}^m \left(\frac{en}{s_i}\right)^{s_i}$, which gives us the following:

$$\mathbb{P}\left(X_s \geq \|\bar{\mathbf{\Theta}}\|_\infty + m!\frac{t+\sum_{i=1}^m s_i\log\frac{en}{s_i}}{|G\cap V|}\right)\leq e^{-t}$$

Now note that $|G\cap V| \geq (\max_i s_i)(n_0-1)(n_0-2)\cdots(n_0-m+1) \geq Cn_0^{m-1}(\max_i s_i)$. This gives us

$$\begin{aligned}
&\mathbb{P}\left(X_s \geq \|\bar{\mathbf{\Theta}}\|_\infty + m!C\frac{t+\sum_{i=1}^m s_i\log\frac{en}{s_i}}{n_0^{m-1}(\max_i s_i)}\right)\leq e^{-t}\\
\Rightarrow&\mathbb{P}\left(X_s \geq \|\bar{\mathbf{\Theta}}\|_\infty + m!C\frac{t/n_0+\sum_{i=1}^m \log\frac{en}{s_i}}{n_0^{m-1}}\right)\leq e^{-t}\\
\Rightarrow&\mathbb{P}\left(X_s \geq \|\bar{\mathbf{\Theta}}\|_\infty + m!C\frac{t/n_0+m\log\frac{n}{n_0}}{n_0^{m-1}}\right)\leq e^{-t}.
\end{aligned}$$

Further taking a union bound over $s_i \in \{n_0, \dots, n\}$ produces a $\frac{m}{n_0} \log n$ term, which is dominated by the $m \log \frac{n}{n_0}$, and we get

$$\mathbb{P}\left(\|\tilde{\mathbf{E}}_{\hat{z}}\|_\infty \le \|\bar{\mathbf{\Theta}}\|_\infty + m! C \frac{t + m \log \frac{n}{n_0}}{n_0^{m-1}} \right) \ge 1 - e^{-t}.$$

Recall that

$$\langle \tilde{\mathbf{E}}_{\hat{z}}, \mathbf{E} \rangle \le 2 \|\tilde{\mathbf{E}}_{\hat{z}}\|_F \sqrt{2m! \|\bar{\mathbf{\Theta}}\|_\infty (t + n \log k + k^m)} + \frac{4m!}{3} \|\tilde{\mathbf{E}}_{\hat{z}}\|_\infty (t + n \log k + k^m)$$

with probability at least $1 - e^{-t}$. Combining this with our bound on $\|\tilde{\mathbf{E}}_{\hat{z}}\|_\infty$ and using AM-GM gives us

$$\langle \tilde{\mathbf{E}}_{\hat{z}}, \mathbf{E} \rangle \le \frac{\|\tilde{\mathbf{E}}_{\hat{z}}\|_F^2}{16} + (m!)^2 C \left(\|\bar{\mathbf{\Theta}}\|_\infty (t + n \log k + k^m) + \frac{t + m \log \frac{n}{n_0}}{n_0^{m-1}} (t + n \log k + k^m) \right)$$

with probability at least $1 - 3e^{-t}$.

Now we proceed to obtain a bound in expectation from the above probability bound. In order to do so, note that by Lemma 5, we have

$$\begin{aligned}
\mathbb{E}\left[\langle \tilde{\mathbf{E}}_{\hat{z}}, \mathbf{E} \rangle - \frac{\|\tilde{\mathbf{E}}_{\hat{z}}\|_F^2}{16} \right] \le & (m!)^2 C \left(\|\bar{\mathbf{\Theta}}\|_\infty (n \log k + k^m) + \frac{m \log \frac{n}{n_0}}{n_0^{m-1}} (n \log k + k^m) \right) \\
& + (m!)^2 C \left(\|\bar{\mathbf{\Theta}}\|_\infty + \frac{m \log \frac{n}{n_0}}{n_0^{m-1}} + \frac{n \log k + k^m}{n_0^{m-1}} \right) \\
\le & (m!)^2 C \left(\|\bar{\mathbf{\Theta}}\|_\infty (n \log k + k^m) + \frac{m \log \frac{n}{n_0}}{n_0^{m-1}} (n \log k + k^m) \right)
\end{aligned}$$

Recall that we have

$$\mathbb{E}\left[\langle \tilde{\mathbf{\Theta}}_{\hat{z}} - \bar{\mathbf{\Theta}}, \mathbf{E} \rangle - \frac{\|\tilde{\mathbf{\Theta}}_{\hat{z}} - \bar{\mathbf{\Theta}}\|_F^2}{8} \right] \le C m! \|\bar{\mathbf{\Theta}}\|_\infty n \log k$$

and by definition, $\|\tilde{\mathbf{\Theta}}_{\hat{z}} - \bar{\mathbf{\Theta}}\|_F \le \|\hat{\mathbf{\Theta}}_{\hat{z}} - \bar{\mathbf{\Theta}}\|_F$. Combining the above observation, we have

$$\begin{aligned}
\|\tilde{\mathbf{E}}_{\hat{z}}\|_F &= \|\hat{\mathbf{\Theta}}_{\hat{z}} - \tilde{\mathbf{\Theta}}_{\hat{z}}\|_F \\
&\le \|\hat{\mathbf{\Theta}}_{\hat{z}} - \bar{\mathbf{\Theta}}\|_F + \|\tilde{\mathbf{\Theta}}_{\hat{z}} - \bar{\mathbf{\Theta}}\|_F \\
&\le 2 \|\hat{\mathbf{\Theta}}_{\hat{z}} - \bar{\mathbf{\Theta}}\|_F.
\end{aligned}$$

This gives us the required bound in expectation.

$$\mathbb{E}\left[\langle \hat{\mathbf{\Theta}} - \bar{\mathbf{\Theta}}, \mathbf{E} \rangle \right] \le \frac{3}{8} \mathbb{E} \|\hat{\mathbf{\Theta}} - \bar{\mathbf{\Theta}}\|_F^2 + (m!)^2 C \left(\|\bar{\mathbf{\Theta}}\|_\infty (n \log k + k^m) + \frac{m \log \frac{n}{n_0}}{n_0^{m-1}} (n \log k + k^m) \right) \tag{7}$$

The statement of the proposition then follows immediately based on combing the above Equations 6 and 7.

■

Proof [Proof of Lemma 2] For the sake of clarity, in this proof we will write $\mathbf{\Theta}$ and f in place of $\bar{\mathbf{\Theta}}$ and f_0 respectively. Recall that in the hypergraphon setting, $\mathbf{\Theta}$ is define by $\mathbf{\Theta}_j = \rho_n f(X_j)$, and conditional on $\mathbf{\Theta}$, $\mathbf{\Theta}_*$ is deterministic. Instead of working directly with $\mathbf{\Theta}_*$, we define $\ddot{\mathbf{\Theta}}$ to be a blockwise average of $\mathbf{\Theta}$ on a balanced partition z^*, where the first $k-1$ blocks have n_0 elements, and the kth block contains $n-(k-1)n_0$. More precisely, define z^* by

$$z^{-1}(a) = \{i \in [n] : X_i = X_{(\ell)} \text{ for some } \ell \in [(a-1)n_0 + 1, an_0]\}$$

when $a \in \{1, 2, \ldots, k-1\}$, and for $a = k$ we have

$$z^{-1}(k) = \{i \in [n] : X_i = X_{(\ell)} \text{ for some } \ell \in [(k-1)n_0 + 1, n]\}$$

where $X_{(j)}$ denotes the jth order statistic of the vector X. Let $n = n_0 k + r$, where $r \in \{0, 1, \ldots, n_0 - 1\}$. Note that by construction, the first $k-1$ classes contain exactly n_0 elements, and the last one contains $n_0 + r$. For $a \in [k]^m$, we define η_a^* to be the number of (unordered) hyperedges in the collection of classes a. Let c_i be the number of times class i appears in the vector a. Then

$$\eta_a^* = \binom{n_0 + r}{c_k} \prod_{i \in [k-1]} \binom{n_0}{c_i}$$

Now we define the following blockwise average

$$\mathbf{Q}_a^* = \frac{1}{\eta_a^*} \sum_{j \in (z^*)^{-1}(a) \cap H} \mathbf{\Theta}_j = \frac{1}{\eta_a^*} \sum_{j \in (z^*)^{-1}(a) \cap H} \rho_n W(X_j)$$

Finally, we define $\ddot{\mathbf{\Theta}}$ to be given by $\ddot{\mathbf{\Theta}}_j = Q^*_{z(j)}$ for $j \in G$, and zero otherwise. Then we have,

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{n^m}\|\mathbf{\Theta} - \ddot{\mathbf{\Theta}}\|_F^2\right] &= \frac{1}{n^m} \sum_{a \in [k]^m} \mathbb{E} \sum_{j \in (z^*)^{-1}(a) \cap G} (\mathbf{\Theta}_j - \mathbf{Q}_a^*)^2 \\
&= \frac{1}{n^m} \sum_{a \in [k]^m} \mathbb{E} \sum_{j \in (z^*)^{-1}(a) \cap G} \left(\rho_n f(X_j) - \frac{1}{\eta_a^*} \sum_{u \in (z^*)^{-1}(a) \cap H} \rho_n f(X_u)\right)^2 \\
&= \frac{\rho_n^2}{n^m} \sum_{a \in [k]^m} \mathbb{E} \sum_{j \in (z^*)^{-1}(a) \cap G} \left(\frac{1}{\eta_a^*} \sum_{u \in (z^*)^{-1}(a) \cap H} (f(X_j) - f(X_u))\right)^2.
\end{aligned}$$

Note that we can interpret the inner sum as an expectation and use Jensen's inequality to obtain

$$\left(\frac{1}{\eta_a^*} \sum_{u \in (z^*)^{-1}(a) \cap H} (f(X_j) - f(X_u))\right)^2 \leq \frac{1}{\eta_a^*} \sum_{u \in (z^*)^{-1}(a) \cap H} (f(X_j) - f(X_u))^2.$$

This then implies

$$\mathbb{E}\left[\frac{1}{n^m}\|\mathbf{\Theta}-\ddot{\mathbf{\Theta}}\|_F^2\right] \leq \frac{\rho_n^2}{n^m}\sum_{a\in[k]^m}\sum_{j\in(z^*)^{-1}(a)\cap G}\frac{1}{\eta_a^*}\sum_{u\in(z^*)^{-1}(a)\cap H}\mathbb{E}[(f(X_j)-f(X_u))^2]$$

By leveraging the Lipschitz smoothness assumption in Equation 2, the inequality $(\sum_i a_i)^2 \leq 2\sum_i a_i^2$, and Jensen's inequality respectively, we then have the following:

$$\begin{aligned}\mathbb{E}[(f(X_j)-f(X_u))^2] &\leq M^2 \max_i \mathbb{E}\left[(|X_{j_1}-X_{u_1}|,\ldots,|X_{j_m}-X_{u_m}|)^2\right]\\ &\leq 2M^2\max_i\left(\mathbb{E}\left[|X_{j_1}-X_{u_1}|^2\right],\ldots,\mathbb{E}\left[|X_{j_m}-X_{u_m}|^2\right]\right)\\ &\leq 2M^2\max_i\left(\mathbb{E}\left[|X_{j_1}-X_{u_1}|^2\right]+\ldots+\mathbb{E}\left[|X_{j_m}-X_{u_m}|^2\right]\right),\end{aligned}$$

Let σ_x be such that $X_x = X_{(\sigma_x)}$. For example, σ applied to the maximal element of X would be 1. Since in the above expression, j_i and u_i belong to the same cluster for every i, by construction, we have $|\sigma_{j_i}-\sigma_{u_i}| \leq 2n_0$ for every i. By lemma 6 and that fact that $k = \lfloor n/n_0 \rfloor$,

$$\mathbb{E}\left[|X_{j_i}-X_{u_i}|^2\right] \leq \left(\frac{2n_0}{n}\right)^2 \leq 4\left(\frac{1}{k}\right)^2$$

which implies

$$\mathbb{E}[(f(X_j)-f(X_u))^2] \leq 8M^2\left(\frac{1}{k}\right)^2$$

Putting it all together, we obtain

$$\mathbb{E}\left[\frac{1}{n^m}\|\mathbf{\Theta}-\ddot{\mathbf{\Theta}}\|_F^2\right] \leq 8M^2\rho_n^2\left(\frac{1}{k^2}\right)$$

which completes the proof.

■

Appendix B. Auxiliary Results

In this section, we list auxiliary results that are used in the proofs of the main results.

Lemma 5 *Let X be a random variable such that its tail-decay satisfy: $\mathbb{P}(X > u^2+Bu+C) \leq e^{-u}$ for all u. Then $\mathbb{E}[X] \leq C+2+B$.*

Proof By assumption $\mathbb{P}(X > u^2+Bu+C) \leq e^{-u}$. We have

$$\mathbb{E}[X] \leq t^* + \int_{t^*}^{\infty}\mathbb{P}(X>t)dt$$

We choose $t^* = C$ so that on $[t^*,\infty)$, $t = u^2+Bu+C$ has at least one nonnegative solution and is necessarily increasing. Let $t = u^2+Bu+C$.

$$\begin{aligned}\mathbb{E}[X] &\leq C + \int_C^\infty \mathbb{P}(X > t)dt \\ &= C + \int_C^\infty \mathbb{P}(X > u^2 + Bu + C)dt \\ &\leq C + \int_0^\infty e^{-u}(2u + B)du \\ &= C + 2 + B.\end{aligned}$$

■

We also need the following result from Klopp et al. (2017).

Lemma 6 *Let $Z_1, \ldots, Z_n$ be i.i.d uniformly distributed random variables on $[0, 1]$. Let $Z_{(i)}$ be the i-th order statistic of the above set of random variables. Then for any $n_0 \leq n$ and $0 \leq s < n_0$, we have*

$$\mathbb{E}\left(Z_{(i)} - Z_{(i+s)}\right)^2 = \frac{s(s+1)}{(n+1)(n+2)} \leq \left(\frac{n_0}{n}\right)^2.$$

The proof of the above lemma is straightforward and could be found in Klopp et al. (2017). Furthermore, since there are several different versions of Bernstein's inequality in the literature, we also state the version that we used in our proofs, below.

Theorem 2 (Bernstein's Inequality) *Let $X_1, X_2, \ldots, X_N$ be zero-mean independent random variables such that $|X_i| \leq M$ almost surely. Then for all $t \geq 0$*

$$\mathbb{P}\left(\sum_{i=1}^N X_i \geq \sqrt{2t\sum_{i=1}^N E[X_i^2]} + \frac{2M}{3}t\right) \leq e^{-t}.$$

Acknowledgements

We would like to thank the editor, Edo Airoldi, and the anonymous reviewers for their insightful comments that helped greatly improve the paper. We also acknowledge support for this project from the National Science Foundation (via NSF grant DMS-2053918).

References

Emmanuel Abbe. Community detection and stochastic block models: recent developments. *The Journal of Machine Learning Research*, 18(1):6446–6531, 2017.

Emmanuel Abbe and Colin Sandon. Community detection in general stochastic block models: Fundamental limits and efficient algorithms for recovery. In *Foundations of Computer Science (FOCS), 2015 IEEE 56th Annual Symposium on*, pages 670–688, 2015.

Emmanuel Abbe, Jianqing Fan, Kaizheng Wang, and Yiqiao Zhong. Entrywise eigenvector analysis of random matrices with low expected rank. *arXiv preprint arXiv:1709.09565*, 2017.

S. Agarwal, J. Lim, L. Zelnik-Manor, P. Perona, D. Kriegman, and S. Belongie. Beyond pairwise clustering. *IEEE Conference on Computer Vision and Pattern Recognition*, 2005.

Kwangjun Ahn, Kangwook Lee, and Changho Suh. Hypergraph spectral clustering in the weighted stochastic block model. *IEEE Journal of Selected Topics in Signal Processing*, 12(5):959–974, 2018.

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

Edo M Airoldi, Thiago B Costa, and Stanley H Chan. Stochastic blockmodel approximation of a graphon: Theory and consistent estimation. In *Advances in Neural Information Processing Systems*, pages 692–700, 2013.

David J Aldous. Representations for partially exchangeable arrays of random variables. *Journal of Multivariate Analysis*, 11(4):581–598, 1981.

Maria Chiara Angelini, Francesco Caltagirone, Florent Krzakala, and Lenka Zdeborová. Spectral detection on sparse hypergraphs. In *2015 53rd Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 66–73. IEEE.

Federico Battiston, Giulia Cencetti, Iacopo Iacopini, Vito Latora, Maxime Lucas, Alice Patania, Jean-Gabriel Young, and Giovanni Petri. Networks beyond pairwise interactions: Structure and dynamics. *Physics Reports*, 2020.

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

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

P. Bonacich, A. C. Holdren, and M. Johnston. Hyper-edges and multidimensional centrality. *Social networks*, 26(3):189–203, 2004.

C. Borgs, J. T. Chayes, H. Cohn, and S. Ganguly. Consistent nonparametric estimation for heavy-tailed sparse graphs. *arXiv preprint arXiv:1508.06675*, 2015a.

Christian Borgs, Jennifer Chayes, and Adam Smith. Private graphon estimation for sparse graphs. In *Advances in Neural Information Processing Systems*, pages 1369–1377, 2015b.

Christian Borgs, Jennifer T Chayes, Henry Cohn, and Victor Veitch. Sampling perspectives on sparse exchangeable graphs. *The Annals of Probability*, 47(5):2754–2800, 2019.

Trevor Campbell, Diana Cai, and Tamara Broderick. Exchangeable trait allocations. *Electronic Journal of Statistics*, 12(2):2290–2322, 2018.

François Caron and Emily B Fox. Sparse graphs using exchangeable random measures. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(5):1295–1366, 2017.

François Caron and Judith Rousseau. On sparsity and power-law properties of graphs based on exchangeable point processes. *arXiv preprint arXiv:1708.03120*, 2017.

Stanley Chan and Edoardo Airoldi. A consistent histogram estimator for exchangeable graph models. In *International Conference on Machine Learning*, pages 208–216, 2014.

Sourav Chatterjee. Matrix estimation by universal singular value thresholding. *The Annals of Statistics*, 43(1):177–214, 2015.

Sourav Chatterjee, Persi Diaconis, and Allan Sly. Random graphs with a given degree sequence. *Annals of Applied Probability*, 21(4):1400–1435, 2011.

Yuxin Chen, Yuejie Chi, Jianqing Fan, and Cong Ma. Gradient descent with random initialization: fast global convergence for nonconvex phase retrieval. *Mathematical Programming*, 176(1):5–37, Jul 2019.

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.

David S Choi, Patrick J Wolfe, and Edoardo M Airoldi. Stochastic blockmodels with a growing number of classes. *Biometrika*, 99(2):273–284, 2012.

Harry Crane and Walter Dempsey. Edge exchangeable models for interaction networks. *Journal of the American Statistical Association*, 113(523):1311–1326, 2018.

Walter Dempsey, Brandon Oselio, and Alfred Hero. Hierarchical network models for structured exchangeable interaction processes. *arXiv preprint arXiv:1901.09982*, 2019.

Peter Diao, Dominique Guillot, Apoorva Khare, and Bala Rajaratnam. Model-free consistency of graph partitioning. *arXiv preprint arXiv:1608.03860*, 2016.

O. Duchenne, F. Bach, I. Kweon, and J. Ponce. A tensor-based algorithm for high-order graph matching. *IEEE transactions on pattern analysis and machine intelligence*, 33(12): 2383–2395, 2011.

Justin Eldridge, Mikhail Belkin, and Yusu Wang. Graphons, mergeons, and so on! In *Advances in Neural Information Processing Systems*, pages 2307–2315, 2016.

Gábor Elek and Balázs Szegedy. A measure-theoretic approach to the theory of dense hypergraphs. *Advances in Mathematics*, 231(3-4):1731–1772, 2012.

L. Florescu and W. Perkins. Spectral thresholds in the bipartite stochastic block model. *Conference on Learning Theory*, 2016.

Santo Fortunato. Community detection in graphs. *Physics reports*, 486(3):75–174, 2010.

Chao Gao, Yu Lu, and Harrison H Zhou. Rate-optimal graphon estimation. *The Annals of Statistics*, 43(6):2624–2652, 2015.

Chao Gao, Zongming Ma, Anderson Y Zhang, and Harrison H Zhou. Community detection in degree-corrected block models. *The Annals of Statistics*, 46(5):2153–2185, 2018.

Gourab Ghoshal, Vinko Zlatić, Guido Caldarelli, and MEJ Newman. Random hypergraphs and their applications. *Physical Review E*, 79(6):066–118, 2009.

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

Debarghya Ghoshdastidar and Ambedkar Dukkipati. Uniform hypergraph partitioning: Provable tensor methods and sampling techniques. *The Journal of Machine Learning Research*, 18(1):1638–1678, 2017b.

Evarist Giné and Richard Nickl. *Mathematical foundations of infinite-dimensional statistical models*, volume 40. Cambridge University Press, 2015.

Anna Goldenberg, Alice X Zheng, Stephen E Fienberg, and Edoardo M Airoldi. A survey of statistical network models. *Foundations and Trends in Machine Learning*, 2(2):129–233, 2010.

W Timothy Gowers. Hypergraph regularity and the multidimensional szemerédi theorem. *Annals of Mathematics*, pages 897–946, 2007.

F Maxwell Harper and Joseph A Konstan. The movielens datasets: History and context. *ACM transactions on interactive intelligent systems (TIIS)*, 5(4):1–19, 2015.

Peter D Hoff, Adrian E Raftery, and Mark S Handcock. Latent space approaches to social network analysis. *Journal of the american Statistical association*, 97(460):1090–1098, 2002.

Paul W Holland, Kathryn Blackmond Laskey, and Samuel Leinhardt. Stochastic blockmodels: First steps. *Social networks*, 1983.

Douglas N Hoover. Relations on probability spaces and arrays of random variables. *Preprint, Institute for Advanced Study, Princeton, NJ*, 2, 1979.

Pengsheng Ji and Jiashun Jin. Coauthorship and citation networks for statisticians. *The Annals of Applied Statistics*, 10(4):1779–1812, 2016.

Olav Kallenberg. Multivariate sampling and the estimation problem for exchangeable arrays. *Journal of Theoretical Probability*, 1999.

Vishesh Karwa and Sonja Petrovia. Discussion of coauthorship and citation networks for statisticians. *The Annals of Applied Statistics*, 2016.

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

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

Mikko Kivelä, Alex Arenas, Marc Barthelemy, James P Gleeson, Yamir Moreno, and Mason A Porter. Multilayer networks. *Journal of complex networks*, 2(3):203–271, 2014.

Olga Klopp and Nicolas Verzelen. Optimal graphon estimation in cut distance. *Probability Theory and Related Fields*, pages 1–58, 2019.

Olga Klopp, Alexandre B Tsybakov, and Nicolas Verzelen. Oracle inequalities for network models and sparse graphon estimation. *The Annals of Statistics*, 45(1):316–354, 2017.

Eric D Kolaczyk. Statistical Analysis of Network Data: Methods and Models (Springer Series in Statistics). 2009.

Florent Krzakala, Cristopher Moore, Elchanan Mossel, Joe Neeman, Allan Sly, Lenka Zdeborová, and Pan Zhang. Spectral redemption in clustering sparse networks. *Proceedings of the National Academy of Sciences*, 110(52):20935–20940, 2013.

Can M. Le, Elizaveta Levina, and Roman Vershynin. Concentration and regularization of random graphs. *Random Structures & Algorithms*, 51(3):538–561, 2017.

Jing Lei and Alessandro Rinaldo. Consistency of spectral clustering in stochastic block models. *The Annals of Statistics*, 43(1):215–237, 2015.

SE Leurgans, RT Ross, and RB Abel. A decomposition for three-way arrays. *SIAM Journal on Matrix Analysis and Applications*, 14(4):1064–1083, 1993.

László Lovász. *Large networks and graph limits*, volume 60. American Mathematical Soc., 2012.

Yu Lu and Harrison H Zhou. Statistical and Computational Guarantees of Lloyd's Algorithm and its Variants. *arXiv preprint arXiv:1612.02099*, 2016.

Simón Lunagómez, Sayan Mukherjee, Robert L Wolpert, and Edoardo M Airoldi. Geometric representations of random hypergraphs. *Journal of the American Statistical Association*, 112(517):363–383, 2017.

Tom Michoel and Bruno Nachtergaele. Alignment and integration of complex networks by hypergraph-based spectral clustering. *Physical Review E*, 86(5):056111, 2012.

Tin Lok James Ng and Thomas Brendan Murphy. Model-based clustering for random hypergraphs. *arXiv preprint arXiv:1808.05185*, 2018.

Soumik Pal and Yizhe Zhu. Community detection in the sparse hypergraph stochastic block model. *arXiv preprint arXiv:1904.05981*, 2019.

Marianna Pensky. Dynamic network models and graphon estimation. *The Annals of Statistics*, 47(4):2378–2403, 2019.

Karl Rohe, Sourav Chatterjee, and Bin Yu. Spectral clustering and the high-dimensional stochastic blockmodel. *The Annals of Statistics*, pages 1878–1915, 2011.

Govind Sharma, Prasanna Patil, and M Narasimha Murty. C3MM: Clique-Closure based Hyperlink Prediction. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence (IJCAI-20)*, 2020.

Despina Stasi, Kayvan Sadeghi, Alessandro Rinaldo, Sonja Petrovic, and Stephen Fienberg. β models for random hypergraphs with a given degree sequence. In *2014 21st International Conference on Computational Statistics*, page 593, 2014.

Kathryn Turnbull, Simón Lunagómez, Christopher Nemeth, and Edoardo Airoldi. Latent space representations of Hypergraphs. *arXiv preprint arXiv:1909.00472*, 2019.

Victor Veitch and Daniel M Roy. Sampling and estimation for (sparse) exchangeable graphs. *The Annals of Statistics*, 47(6):3274–3299, 2019.

Patrick J Wolfe and Sofia C Olhede. Nonparametric graphon estimation. *arXiv preprint arXiv:1309.5936*, 2013.

Justin Yang, Christina Han, and Edoardo Airoldi. Nonparametric estimation and testing of exchangeable graph models. In *Artificial Intelligence and Statistics*, pages 1060–1067, 2014.

Muhan Zhang, Zhicheng Cui, Shali Jiang, and Yixin Chen. Beyond link prediction: Predicting hyperlinks in adjacency space. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Yuan Zhang, Elizaveta Levina, and Ji Zhu. Estimating network edge probabilities by neighbourhood smoothing. *Biometrika*, 104(4):771–783, 2017.

Y. Zhao. Hypergraph limits: A regularity approach. *Random Structures & Algorithms*, 47 (2):205–226, 2015.

Vincent Zheng, Bin Cao, Yu Zheng, Xing Xie, and Qiang Yang. Collaborative filtering meets mobile recommendation: A user-centered approach. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 24, 2010.

Vinko Zlatić, Gourab Ghoshal, and Guido Caldarelli. Hypergraph topological quantities for tagged social networks. *Physical Review E*, 80(3):036118, 2009.