Product Measure Approximation of Symmetric Graph Properties

Dimitris Achlioptas *
Department of Computer Science
University of California, Santa Cruz
optas@cs.ucsc.edu

Paris Siminelakis †
Department of Electrical Engineering
Stanford University
psimin@stanford.edu

Abstract

Random structures often present a trade-off between realism and tractability, the latter predominantly enabled by independence. In pioneering random graph theory, Erdős and Rényi originally studied the set of all graphs with a given number of edges, seeking to identify its typical properties under the uniform measure, i.e., the $G(n, m)$ model of random graphs. However, this approach quickly presents major challenges, most notably the lack of independence, leading to the approximation of the $G(n, m)$ model by the $G(n, p)$ model, wherein edges appear independently with probability p. Let $\mathcal{G}_n$ be the set of all graphs on n vertices. In this work we pursue the following question: *What are general sufficient conditions for the uniform measure on a set of graphs $S \subseteq \mathcal{G}_n$ to be approximable by a product measure?*

*Research supported by a European Research Council (ERC) Starting Grant (StG-210743) and an Alfred P. Sloan Fellowship.
†Supported in part by an Onassis Foundation Scholarship.

1 Introduction

Let $\mathcal{G}_n$ be the set of all graphs on n vertices. Given a graph property (set) $S \subseteq \mathcal{G}_n$ and another property S', questions in Graph Theory can be stated as: *"Does S imply S' or, equivalently, is $S \subseteq S'$?"* For instance, if S is *planar* and $S^{'}$ is 4-*colourable*, the celebrated Four Color Theorem [2] asserts that the answer is positive. Requiring that S implies S' is a very strong statement that can fail for a single counterexample. However, it may well be the case that the vast majority of elements in S do have property $S^{'}$. Thus, it is often fruitful to consider relaxations of our original question. The two most common relaxations amount to either asking that each element of S is *close* to some element of S', leading to Property Testing [28, 29], or allowing the removal of a certain fraction of elements from S, leading to the study of random graphs [5, 17]. That is, the question is rephrased as "What is the probability that a uniformly random element of S is in S'?"

Erdős and Rényi pioneered random graph theory by considering such questions when S is the simplest graph property, namely containing a fixed number m of edges, i.e., $S = S_m := \{G \in \mathcal{G}_n : |E(G)| = m\}$. The resulting random graph model, known as $G(n, m)$, amounts to selecting a uniformly random m-subset of edges. Nevertheless, even for this simplest S, the probabilistic presence or absence of some other property $S^{'}$ is hard to analyze, since the uniform measure on S requires making one massive random choice at once rather than many independent choices.

Erdős and Rényi went a step further and developed tools to make the study of such questions for $G(n, m)$ tractable by introducing the $G(n, p)$ model of random graphs, wherein each edge is included independently with probability p. In $G(n, p)$ intricate probabilistic analysis is possible by simulating the creation process of the random graph. And although $G(n, m)$ and $G(n, p)$ induce different distributions on $\mathcal{G}_n$ they can be related so that results transfer. The reason is that since all edges in $G(n, p)$ are independent and equally likely, $G(n, p)$ can be seen as first selecting a random number of edges $m \sim \mathrm{Bin}(\binom{n}{2}, p)$ and then a uniformly random element of S_m. Thus, if one sets $p = p(m) = m/\binom{n}{2}$, the random graph $G \sim G(n, m)$ and the two random graphs $G^{\pm} \sim G(n, (1 \pm o(1))p)$ can be coupled[1] so that with high probability

$$G^{-} \subseteq G \subseteq G^{+} \ . \tag{1}$$

The significance of this relationship between $G(n, m)$ and $G(n, p)$, i.e., between what we *wish* to study and what we *can* study, manifests dramatically in the study of *monotone* graph properties. We can prove results about $G(n, m)$ concerning a monotone increasing property $S^{'}$ by showing that $G(n, p^{+})$ does not have the property (negative side) or that $G(n, p^{-})$ does have the property (positive side). This connection has been thoroughly exploited to establish threshold functions for a host of monotone graph properties such as Connectivity, Hamiltonicity, and Subgraph Existence, making it the workhorse of random graph theory.

In this work we seek to extend the above approximation of the uniform measure by a product measure to more general sets. We are motivated in this by the belief that the main point of any model is prediction. That is, that the reason to study random graphs is to understand what graph properties are *typically implied* by other graph properties. As a result, we hold that rather than piling on features in an effort to match observed properties of real networks we should remain agnostic about everything not explicitly modeled, i.e., consider the *uniform* measure on graphs with the assumed property S. Notably, this viewpoint does not eschew the presence of independences among edges (and more general events) in real networks. On the contrary, it asserts that besides enhancing intellectual honesty the discovery of independences has great utility in and of itself. Our main results are: a generalization of (1) to convex symmetric sets of graphs (Theorem 1), and a corresponding concentration inequality (Theorem 2).

[1] Fix an arbitrary enumeration of all $N = \binom{n}{2}$ edges and assign to each edge e, independently, a uniform random variable $U_e \sim U[0, 1]$. Let $U_{(m)}$ denote the m-th smallest number among $U_1, \ldots, U_N$. Then to generate G we take $E(G) = \{e : U_e \leq U_{(m)}\}$, while to generate $G(n, (1 \pm \epsilon)p)$ we take $E(G^{\pm}) = \{e : U_e \leq (1 \pm \epsilon)p\}$. Thus, (1) reduces to $(1 - \epsilon)p \leq U_(m) \leq (1 + \epsilon)p$.

2 Definitions and Results

The mathematical analysis of random structures in general, and random graphs in particular, is immensely enabled by independence. A product measure on the set of all undirected simple graphs $\mathcal{G}_n$ with n vertices is specified succinctly by a symmetric matrix $\mathbf{Q} \in [0,1]^{n \times n}$ of probabilities where $Q_{ii} = 0$ for $i \in [n]$. By analogy to $G(n,p)$, we denote by $G(n,\mathbf{Q})$ the measure in which each possible edge $\{i,j\}$ is included independently with probability Q_{ij}. Our main result is a sufficient condition for the uniform measure over a set S, denoted as $U(S)$, to be approximable by a product measure in the following sense.

Sandwichability. *The uniform measure over a set S is said to be (ϵ,δ)-sandwichable if there exists a $n \times n$ symmetric matrix $\mathbf{Q}$ such that the two distributions $G^{\pm} \sim G(n,(1\pm\epsilon)\mathbf{Q})$, and $G \sim U(S)$ can be coupled so that $G^- \subseteq G \subseteq G^+$ with probability at least $1-\delta$.*

Informally, the two conditions required for our theorem to hold are:

Partition Symmetry. We consider sets S that are symmetric with respect to *some* partition of the edges. That is, the characteristic function of S should depend only on *how many* edges are taken from each part but not *which* edges. So, for example, in the Erdős-Rényi $G(n,m)$ model the partition trivially consists of one part, i.e., all edges are exchangeable. More interesting cases are partitions induced by vertex partitions (Stochastic Block Models) and, most interestingly, partitions induced by geometry (lattices). We discuss the motivation for edge-partition symmetry at length in Section 3.

Convexity. Partition symmetry reduces the study of the uniform measure on S to the study of the induced measure on the set of feasible vectors, expressing the number of edges taken from each part. Since this set can be of arbitrary complexity, some regularity is necessary. We consider symmetric properties S such that the set of feasible vectors is *convex*. While, strictly speaking, convexity is not necessary for our proof method to work, it provides a clean conceptual framework that still allows very general properties to be expressed. These include properties expressible as Linear Programs over the number of edges from each part, and even non-linear constraints, expressing the presence or absence of percolation. Most importantly, since convex sets are closed under intersection, convex properties can be piled on (sets intersected) while maintaining approximability by a product distribution.

For each convex symmetric set S, our theorem identifies a unique product measure $\mathbf{Q}^*(S)$ as the solution of a constrained entropy-maximization problem. The precise quantitative version of the theorem requires the introduction of parameters capturing the geometry of the convex domain along with aspects of the edge-partition. As we will see these parameters are inevitable and our results sharp, up to constants.

Below we give a series of definitions concluding with a formal statement of our main result. We start with some notation. We will use lower case boldface letters to denote vectors and uppercase boldface letters to denote matrices. Further, we fix an arbitrary enumeration of the $N = \binom{n}{2}$ edges and sometimes represent the set of all graphs on n vertices as $H_N = \{0,1\}^N$. We will refer to an element of $x \in H_N$ interchangeably as a graph and a string. Given a partition $\mathcal{P} = (P_1,\ldots,P_k)$ of $[N]$, we define $\Pi_N(\mathcal{P})$ to be the set of all permutations acting only within blocks of the partition.

Edge Block Symmetry. *Fix a partition $\mathcal{P}$ of $[N]$. A set $S \subseteq H_N$ is called $\mathcal{P}$-symmetric if it is invariant under the action of $\Pi_N(\mathcal{P})$. Equivalently, if $\mathbb{I}_S(x)$ is the indicator function of set S, then $\mathbb{I}_S(x) = \mathbb{I}_S(\pi(x))$ for all $x \in H_N$ and $\pi \in \Pi_N(\mathcal{P})$.*

The number of blocks $k = |\mathcal{P}|$ gives a rough indication of the amount of symmetry present. For example, when $k = 1$ we have maximum symmetry and all edges are equivalent. In a stochastic block

model (SBM) with ℓ classes, $k = \binom{\ell}{2}$. For a d-dimensional lattice, partitioning the potential edges by distance results in roughly $k = n^{1/d}$ parts, whereas finally if $k = N$ there is no symmetry whatsoever. Our results accommodate partitions with as many as $O(n^{1-\epsilon})$ parts. This is way more than enough for most situations. For example, as we saw, in lattices there are $O(n^{1/d})$ distances, while if we have n generic points such that the nearest pair of points have distance 1 while the farthest have distance D, fixing any $\delta > 0$ and binning together all edges of length $[(1+\delta)^i, (1+\delta)^{i+1})$ for $i \geq 0$, yields only $O(\delta^{-1} \log D)$ classes.

Edge Profile. *Given a partition $\mathcal{P} = (P_1, \ldots, P_k)$ of H_N and a graph $x \in H_N$, for $i \in [k]$ let $m_i(x)$ be the number of edges of x from P_i. The* edge profile *of x is $\mathbf{m}(x) := (m_1(x), \ldots, m_k(x))$.*

We will denote the image of a $\mathcal{P}$-symmetric set S under $\mathbf{m}$ as $\mathbf{m}(S) \subseteq \mathbb{R}^k$. The edge-profile becomes particularly relevant to the study of $\mathcal{P}$-symmetric sets due to the following basic fact (proven in Appendix A).

Proposition 1. *Any function $f : H_N \to \mathbb{R}$ invariant under $\Pi_N(\mathcal{P})$, depends only on the edge-profile $\mathbf{m}(\mathbf{x})$.*

In particular, since membership in S depends solely on a graph's edge-profile, it follows that a uniformly random element of S can be selected as follows: first (i) generate an edge profile $\mathbf{v} = (v_1, \ldots, v_k) \in \mathbb{R}^k$ from the distribution on $\mathbf{m}(S)$ induced by $U(S)$, and then (ii) for each $i \in [k]$ independently select a uniformly random subset of v_i edges P_i. (Formally, this is Proposition 3 in Appendix A.) Thus, conditional on the edge-profile, not only is the distribution of edges known, but it factorizes in a product of $G(n,m)$ distributions. Thus, the complexity of the uniform measure on S manifests *entirely* in the induced distribution on $\mathbf{m}(S) \in \mathbb{N}^k$, whose structure we discuss next.

Definition 1. *Let $p_i = |P_i|$ denote the number of edges in part i of partition $\mathcal{P}$.*

Edge Profile Entropy. *Given an edge profile $\mathbf{v} \in \mathbf{m}(S)$ define the entropy of $\mathbf{v}$ as* $\textsc{Ent}(\mathbf{v}) = \sum_{i=1}^{k} \log \binom{p_i}{v_i}$.

Using the edge-profile entropy we can express the induced distribution on $\mathbf{m}(S)$ as $\mathbb{P}(\mathbf{v}) = \frac{1}{|S|} e^{\textsc{Ent}(\mathbf{v})}$. The crux of our argument is now this: the only genuine obstacle to S being approximable by a product measure is degeneracy, i.e., the existence of multiple, well-separated edge-profiles that maximize $\textsc{Ent}(\mathbf{v})$. The reason we refer to this situation as degeneracy is that it typically encodes a hidden symmetry of S with respect to $\mathcal{P}$. For example, imagine $\mathcal{P} = (P_1, P_2)$ with $p_1 = p_2$ and that S contains all graphs that either have $p_1/2$ edges from P_1 and $p_2/3$ edges from P_2, or vice versa. Then, in a uniformly random $G \in S$ the presence in G of edge $e \in P_i$ boosts the probability of all other edges in P_i. Note that since the set $\mathbf{m}(S)$ is discrete, quantifying what it means for the maximizer of $\textsc{Ent}$ to be "sufficiently unique" is non-trivial. For example, what happens if, strictly speaking, there is a unique maximizer of $\textsc{Ent}(\mathbf{v})$ but sufficiently many near-maximizers to potentially receive, in aggregate, a majority of the measure?

As our main goal is to show that approximation by a product measure holds for sets far more general than the set S_m containing all graphs with m edges, i.e., $G(n,m)$, we focus in this extended abstract on a setting that strikes a nice balance between conceptual clarity and generality.

Convexity. *Let $\mathrm{Conv}(A)$ denote the convex hull of a set A. Say that a $\mathcal{P}$-symmetric set $S \subseteq \mathcal{G}_N$ is convex iff the convex hull of $\mathbf{m}(S)$ contains no new integer points, i.e., if $\mathrm{Conv}(\mathbf{m}(S)) \cap \mathbb{N}^k = \mathbf{m}(S)$.*

Let $H_{\mathcal{P}}(\mathbf{v})$ be the approximation to $\textsc{Ent}(\mathbf{v})$ that results by replacing each binomial term with its binary entropy approximation via the first term in Stirling's approximation (see (13) in Appendix B).

Entropic Optimizer. *Let $\mathbf{m}^* = \mathbf{m}^*(S) \in \mathbb{R}^k$ be the unique solution to* $\max_{\mathbf{v} \in \mathrm{Conv}(\mathbf{m}(S))} H_{\mathcal{P}}(\mathbf{v})$.

Defining the optimization problem on the convex hull of $\mathbf{m}(S)$ is crucial. It will enable us to study the set S only through the properties of the maximizer $\mathbf{m}^*$. Naturally, if a $\mathcal{P}$-symmetric set S has entropic optimizer $\mathbf{m}^* = (m_1^*, \ldots, m_k^*)$, the natural candidate product measure for each $i \in [k]$ assigns probability m_i^*/p_i to all edges in part P_i. The challenge is to relate this measure to the uniform measure on S by proving concentration of the induced measure on $\mathbf{m}(S)$ around a point near $\mathbf{m}^*$. For that we need the vector $\mathbf{m}^*$ to be "close" to[2] a vector in $\mathbf{m}(S)$ and to be able to quantify the decrease in entropy "away" from $\mathbf{m}^*$. To quantify this second notion we need the following parameters, expressing the geometry of convex sets.

Definition 2. *Given a partition $\mathcal{P}$ and a $\mathcal{P}$-symmetric convex set S, we define*

$$\text{Thickness:} \qquad \mu = \mu(S) = \min_{i \in [k]} \min\{m_i^*, p_i - m_i^*\} \tag{2}$$

$$\text{Condition number:} \qquad \lambda = \lambda(S) = \frac{5k \log n}{\mu(S)} \tag{3}$$

$$\text{Resolution:} \qquad r = r(S) = \frac{\lambda + \sqrt{\lambda^2 + 4\lambda}}{2} > \lambda \tag{4}$$

The most important parameter is the *thickness* $\mu(S)$. It quantifies the L_∞ distance of the optimizer $\mathbf{m}^*(S)$ from the natural boundary $\{0, p_1\} \times \ldots \times \{0, p_k\}$ where the entropy of a class becomes zero. Thus, this parameter determines the *rate of coordinate-wise concentration* around the optimum.

The *condition number* $\lambda(S)$, on the other hand, quantifies the robustness of S. To provide intuition, in order for the product measure approximation to be accurate for every class of edges (part of $\mathcal{P}$), fluctuations in the number of edges of order $\sqrt{m_i^*}$ need to be "absorbed" in the mean m_i^*. For this to happen with polynomially high probability for a single part, we must have $m_i^* = \Omega(\log(n))$ by standard Chernoff bounds. We absorb the dependencies between parts by taking a union bound, thus multiplying by the number of parts, yielding the numerator in (3). Our results give strong probability bounds when $\lambda(S) \ll 1$, i.e., when in a typical graph in S the number of edges from each part is $\Omega(k \log n)$ edges away from triviality, a condition we expect to hold in all natural applications. We can now state our main result.

Theorem 1 (Main result)**.** *Let $\mathcal{P}$ be any edge-partition and let S be any $\mathcal{P}$-symmetric convex set. For every $\epsilon > \sqrt{12\lambda(S)}$, the uniform measure over S is (ϵ, δ)-sandwichable, for $\delta = 2\exp\left[-\mu(S)\left(\frac{\epsilon^2}{12} - \lambda(S)\right)\right]$.*

As a sanity check we see that up to the constants in the Θ-notation, Theorem 1 recovers the sandwichability of $G(n, m)$ as sharply as the Chernoff bound.

Example 1. *$G(n, m)$ is (ϵ, δ)-sandwichable where $\epsilon = \Theta\left(C\sqrt{\frac{\log m}{m}}\right)$ and $\delta = n^{-\Theta(C)}$, for any $C > 0$.*

Theorem 1 follows by analyzing the natural coupling between the uniform measure on S and the product measure corresponding to the entropic optimizer $\mathbf{m}^*$. Our main technical contribution is Theorem 2 below, that expresses a concentration inequality for the edge-profiles of convex symmetric sets S. The *resolution*, $r(S)$, defined in (4) above, reflects the smallest *concentration width* that can be provedt by our theorem. When $\lambda(S) \ll 1$, as required for the theorem to be meaningfully applied, it scales optimally as $\sqrt{\lambda(S)}$.

Theorem 2. *Let $\mathcal{P}$ be any edge-partition, let S be any $\mathcal{P}$-symmetric convex set and let $\mathbf{m}^*$ be the entropic optimizer of S. If G is a uniformly random element of S, then for all $\epsilon > r(S)$:*

$$\mathbb{P}_S\left(|\mathbf{m}(G) - \mathbf{m}^*| \le \epsilon \tilde{\mathbf{m}}^*\right) \ge 1 - \exp\left[-\mu(S)\left(\frac{\epsilon^2}{1+\epsilon} - \lambda(S)\right)\right] \quad , \tag{5}$$

where $\mathbf{x} \le \mathbf{y}$ means that $x_i \le y_i$ for all $i \in [k]$, and $\tilde{m}_i = \min\{m_i^*, p_i - m_i^*\}$.

[2]Indeed, this is the only use we make of convexity in the proof presented here.

The intuition behind the concentration result is that as *thickness* increases two phenomena occur: (i) vectors close to $\mathbf{m}^*$ capture a larger fraction of the measure, and (ii) the decay in entropy away from $\mathbf{m}^*$ becomes steeper. These joint forces compete against the probability mass captured by vectors "away" from the optimum. The point were they prevail corresponds to $\lambda(S) \ll 1$ or, equivalently, $\mu(S) \gg 5k\log(n)$. We note that except for the constant 5, this bound **can not** be improved without further assumptions on S. For instance, the probability bounds we give scale as $n^{-\Omega(k\epsilon^2)}$, assuming $\lambda(S) \ll 1$, but the actual deviation probability could be significantly smaller.

Finally, we derive results allowing the approximation of the probability of any "local" event in $U(S)$. Due to lack of space we state and prove these results in Appendix D.

3 Motivation

As mentioned in the introduction our goal is to study the uniform measure over sets of graphs. Since just the description of a set, either explicit by enumerating elements or implicit through its Fourier representation, can be of exponential size, to make progress the first step is to identify a "language" in which to express sets of graphs that is expressive enough to be interesting and restricted enough to be tractable.

3.1 Why an Edge Partition?

Arguably the most natural way to introduce structure on a set S is to impose symmetry. This is formally expressed as the *invariance* of its characteristic function $\mathbb{I}_S$ under the action of a group of transformations T, i.e., requiring that $\forall t \in T : \mathbb{I}_S(t(x)) = \mathbb{I}_S(x), \forall x$. We consider functions $\mathbb{I}_S$ for which there exists a partition $\mathcal{P}$ of the $\binom{n}{2}$ edges such that the function is invariant under any permutation of the edges (indices) within a part. Therefore, if graphs G, G' contain the same number of edges from each part, then $\mathbb{I}_S(G) = \mathbb{I}_S(G')$. We argue that this notion of symmetry is both natural and well-motivated. Considering a subgroup of the symmetric group is natural due to the Cayley Theorem [3], asserting that every group is isomorphic to a subgroup of the permutation group. Indeed, the symmetry we require is already embedded in many existing graph models (typically in a far more restrictive form than ours).

3.1.1 Edge-Partition Symmetry in Existing Graph Models

Erdős-Rényi. When $\mathcal{P}$ has a single part, all edges are equivalent and $\mathbb{I}_S(G)$ depends only on the *number* of edges in G. When $\mathbb{I}_S(G)$ is the indicator for $|E(G)| = m$, the uniform measure on S is $G(n, m)$. We note that every probability distribution where edges are exchangeable obeys this symmetry, e.g., $G(n, p)$. At the other extreme, if $\mathcal{P}$ consistis of $\binom{n}{2}$ atomic parts, there is no symmetry, recovering full generality.

Stochastic Kronecker Graphs. This model [20, 22, 15] takes the k-fold Kronecker product of a 2×2 symmetric matrix $P_{[1]}$ of probabilities to form an $n \times n$ matrix $P_{[k]}$, where $n = 2^k$. Edges are included independently with probability $P_{[k]}(i, j) = \prod^k P_{[1]}(i_\ell, j_\ell)$, which can be seen to depend on the binary representation of the endpoints. The resulting edge-partition, by probability, has up to $3^k \approx n^{1.58}$ parts.

Stochastic Block Models. These theoretically important models [24, 27, 23] presuppose the existence of a **vertex** partition $\mathcal{V} = (V_1, \ldots, V_\ell)$ and the existence of an $\ell \times \ell$ matrix of probabilities P. A random digraph is produced by including every arc (i, j) independently with probability $p_{\bar{i},\bar{j}}$, where $\bar{i}, \bar{j}$ are the blocks containing i, j, respectively. (To get undirected graphs arc direction is ignored.) Thus, $\mathcal{V}$ induces an edge partition $\mathcal{P}_\mathcal{V} \stackrel{\triangle}{=} \{V_i \times V_j\}_{i,j \le \ell}$, where edges are in the same part iff their endpoints are in the same ordered pair of blocks of $\mathcal{V}$.

Graph Limits. The fact that our symmetry requirement encompasses Stochastic Block Models is particularly pertinent in light of the theory of Graph Limits [21, 7]. According to that theory, given a sequence of graphs $G_1, G_2, \ldots$ obeying some regularity conditions, one can extract a limiting object that captures many properties of the sequence as a measurable integrable[3] function $W : [0,1]^2 \to [0,1]$, called a *graphon*, see [8, 9]. Inherent in the construction of the limiting object is an intermediate approximation of the sequence of graphs by a sequence of Stochastic Block Models, obtained by invoking the (weak) Szemerédi Regurality Lemma [14, 7]. The conclusions we can draw for our purposes is that any property that is encoded in the limiting object, typically subgraph densities, is expressible within our framework of symmetry. This is due to the fact that each element of the sequence of Stochastic Block Models used to identify the graphon satisfies our notion of symmetry.

3.1.2 Enabling the Expression of Geometry

A common feature of all examples in the previous section is that the edge-partition *factorized* along vertices, i.e., the part of every edge was determined by the parts of its endpoints according to some vertex partition. A strong driving force behind the development of further random graph models has been the incorporation of geometry. Typically this is done by embedding the vertices in some (low-dimensional) metric space and assigning probabilities to edges as a function of distance. This can be seen as a further reduction in symmetry, as the edge-partition does not factorize over vertices.

Our result also does not presume any kind of factorization of the edge-partition and this is perhaps its *most significant* feature, as it allows to establish product measure approximations even in highly non-uniform settings. In fact, we do not make any coherence requirement of the edge-partition. As a result, our notion of symmetry is neither confined to mean-field type structure, as in models with a vertex-factorization, nor does it impose a geometry on the vertices as done by geometric random graph models.

3.2 Why should S be Convex?

To focus on the approximate modes of the edge-profile distribution, for $\epsilon = \epsilon(n) > 0$, let

$$M_\epsilon(S) = \{\mathbf{v} \in \mathbf{m}(S) : \ \textsc{Ent}(\mathbf{v}) \geq \sup_{\mathbf{z} \in \mathbf{m}(S)} \textsc{Ent}(\mathbf{z}) - \epsilon\} \ . \tag{6}$$

Ideally, as $\epsilon \to 0$ the set $M_\epsilon(S)$ shrinks to a point, implying concentration around a single typical profile $\mathbf{m}^*$. As discussed, expecting $M_\epsilon(S)$ to shrink to a point is most natural under very general considerations. We chose to state our result under the assumption that S is convex for two reasons. The first is that since entropy is a concave function, convexity implies unimodality automatically. Thus, even though convexity is not necessary for concentration, it serves as a nice framework. Secondly, since convex sets are closed under intersection, we gain another automatism, i.e., convex properties can be piled on (sets intersected) while maintaining approximability by a product distribution. This opens the prospect of analyzing increasing more complex properties (sets) S.

A natural concern at this point is whether there are interesting convex graph properties, i.e., properties that exhibit permutation symmetry over a partition of the edges and whose set of edge profiles is convex. In the next section 4 we give two examples giving a sense of the range of applicability of our results.

[3] In general L^p-integrable for some $p \geq 1$.

4 Two Examples

4.1 Probabilistic Network Design via Linear Programs

Navigability. As an example of the intricate properties that can be treated within our framework, we discuss network navigability [18]. Kleinberg, in [18, 19] gave sufficient conditions for greedy routing to discover paths of poly-logarithmic length between any two vertices. One of the most general settings where navigability is possible is that of set-systems, a mathematical abstraction of the relevant geometric properties of grids, regular-trees and graphs of bounded doubling dimension. The essence of navigability lies in the requirement that for any vertex in the graph, the probability of having an edge to a vertex at distance in the range $[2^{i-1}, 2^i)$ is approximately uniform for all $i \in [\log n]$. In our language, we can partition the edges according to their distance scale so that part P_i would include all possible edges between vertices at scale i.

If we now associate a cost with each edge, possibly related to its length, and a total budget B, Theorem 1 implies that the edges within each distance scale will be independent. As a result, it can be shown [1], that navigability emerges without any coordination between the vertices (such as all using the same probability distribution), or independence assumption regarding network formation.

Linear Programs. More generally, rather than a single cost value for all the edges in a part and a single "budget" we can have a system of linear constraints whose variables are the components of the edge-profile $\mathbf{m}$. The uniform measure over the set satisfying the linear constraints can be written as:

$$S = \{\, A \cdot \mathbf{m} \leq \mathbf{b} \} \ , \tag{7}$$

for some appropriate matrix $A = [A_1 \dots A_k] \in \mathbb{R}^{\ell \times k}$ and vector $\mathbf{b} \in \mathbb{R}^{\ell}$. Such, a set could express capacity constraints, latency constraints, explicit upper and lower bounds on the acceptable number of edges from a class, etc. Besides this generality of expression, we see that the entropy-optimization problem now is not simply tractable but has a closed form analytic solution in terms of the dual variables $\lambda \in \mathbb{R}^{\ell}_{+}$:

$$m_i^*(S) = \frac{p_i}{1 + \exp(A_i^T \lambda)} \ . \tag{8}$$

Using this characterization of the edge probabilities (8), it is possible to show [1] that for a class of geometries generalizing Kleiberg's set-systems [19], navigability can be achieved using in total $\Theta(n \log(n))$ edges by imposing a single linear constraint: that of bounded total cost $S_B = \{\mathbf{v} : \sum_{i=1}^{k} c_i v_i \leq B\}$ where the cost of each edge corresponds to the *cost of indexing* $c_i := \log(p_i/n)$, i.e., the number of bits needed to identify a particular vertex at each distance scale i.

4.2 Percolation Avoidance

In the previous examples, we explored partition that did not factorize over vertices. Here, we see that when more structure is assumed about the partition stronger statements can be made.

Consider a social network with n vertices consisting of ℓ communities having sizes $\rho_i n$, where $\rho_i > 0$ for $i \in [\ell]$. Imagine that out of the ℓ classes, we require some specific class s to act as the "connector", i.e., that the graph induced by the remaining classes should have no component of size greater than ϵn for some arbitrarily small $\epsilon > 0$. Let $S = S_\epsilon$ be the set of all such graphs.

The vertex partition induces a natural edge partition by assigning edges between different pairs of communities to different parts. For a given graph $x \in H_N$, let $\mathbf{M} \in \mathbb{N}^{\ell \times \ell}$ be the matrix encoding the number of edges that x has in each part P_{ij}. The property of not having a giant component without the vertices of

a specific cluster is not symmetric and nor convex with respect to the partition. Nevertheless, utilizing a theorem of Bollobas, Janson, Riordan [6] we can provide a precise sufficient asymptotic condition for the property of interest. Given the edge-profile $\mathbf{M}$ define the $(\ell-1)\times(\ell-1)$ matrix:

$$T(\mathbf{M})_{ij} := \frac{m_{ij}}{n^2\rho_i}, \forall i,j \in [\ell]\setminus\{s\}$$

that encapsulates the dynamics of a multi-type branching process and is relevant to the study of the giant component in our setting.

Theorem 3 (Informal)**.** *Let* $\|\cdot\|_2$ *denote the* operator norm *(maximum singular value). If* $\|T(\mathbf{M})\|_2 \leq 1$ *no giant component exists, while if* $\|T(\mathbf{M})\|_2 > 1$ *a giant component exists.*

Thus, in our framework, the property $S_\epsilon = \{$ no giant component without vertices from $s\}$, not only can be accurately approximated under the specific partition $\mathcal{P}$ by $\hat{S} = \{\mathbf{M} : \ \|T(\mathbf{M})\|_2 \leq 1\}$ but $\hat{S}$ happens to also be a convex function of $\mathbf{M}$.

5 Technical Overview

In this section, we present an overview of the technical work involved in the proof of Theorems 1 and 2. Most of the work lies in the concentration result, Theorem 2.

Concentration. The general idea is to identify a high-probability subset $\mathcal{L} \subseteq \mathbf{m}(S)$ by integrating the probability measure around the entropy-maximizing profile $\mathbf{m}^*$. Since ultimately our goal is to couple the uniform measure with a product measure, we need to establish concentration for every part of the edge-partition, i.e., in every coordinate. There are two main obstacles to overcome: (i) the size of $|S|$, i.e., the partition function, is unknown, (ii) integrating the measure outside $\mathcal{L}$ while concurrently quantifying the decrease in entropy as a function of the L_∞ distance from the maximizer $\mathbf{m}^*$. Our strategy to resolve the above obstacles is:

Partition function. We approximate (lower bound) the log-partition function $\log|S|$ by the contribution to it of the entropic optimal edge-profile $\mathbf{m}^*$, thus upper-bounding the probability of every $\mathbf{v} \in \mathbf{m}(S)$ as

$$\log \mathbb{P}_S(\mathbf{v}) = \textsc{Ent}(\mathbf{v}) - \log(|S|) \leq \textsc{Ent}(\mathbf{v}) - \textsc{Ent}(\mathbf{m}^*) \ . \tag{9}$$

This is the crucial step that opens up the opportunity of relating the probability of a vector $\mathbf{v}$ to the distance $\|\mathbf{v}-\mathbf{m}^*\|_2$ through analytic properties of entropy. The importance of this step stems from the fact that all information about S resides in $\mathbf{m}^*$ due to our choice of solving the optimization problem on the convex hull. We show next that this can be a rather loose upper bound.

Proposition 2. *There exists a partition* $\mathcal{P}$ *with* k *parts and a convex* $\mathcal{P}$*-symmetric set* S *such that* $\log(|S|) - \textsc{Ent}(\mathbf{m}^*) = \Omega(k\log(n))$.

Unfortunately, to obtain a better lower bound on $\log(|S|)$ we would need to integrate over the set S, which can not be done without making assumptions about its geometry. As we are aiming for a general theorem we cannot really improve this step and the potential loss of up to $O(k\log(n))$ is present.

Distance bounds: To bound the rate at which entropy decays from below as a function of the component-wise distance from the maximizer $\mathbf{m}^*$, we first approximate $\textsc{Ent}(\mathbf{v})$ by the corresponding binary entropy

to get a smooth function. Exploiting the separability, concavity and differentiability of entropy we obtain component-wise distance bounds using a second-order Taylor approximation. At this step we also lose a cumulative factor of order $3k\log(n)$ stemming from Stirling approximations and the subtle point that the maximizer $\mathbf{m}^*$ might not be an integer point. The constant 3 can be improved, but in light of Proposition 2 this would be pointless and complicate the proof unnecessarily.

Union bound: Finally, we integrate the obtained bounds outside the set of interest by showing that even if all "bad" vectors where placed right at the boundary of the set, where the lower bound on the decay of entropy is smallest, the total probability mass would be exponentially small. The loss incurred at this step is of order $2k\log(n)$, since there are at most n^{2k} bad vectors.

Coupling. To prove Theorem 1 using our concentration result, we argue as follows. Conditional on the edge-profile, we can couple the generation of edges in different parts independently by a similar process as in the $G(n,m)$ to $G(n,p)$ case. Then, using a union bound we can bound the probability that all couplings succeed given an appropriate $\mathbf{v}$. Finally, using the concentration theorem we show that sampling an appropriate edge-profile happens with high probability.

Summing up, the conditions in our theorems are optimal up to small constants and cannot be improved without further assumptions. In Appendix D we also derive an upper bound on a *local version* (involving few edges) of the Total Variation distance between the uniform measure and its product measure approximation. This makes our results relevant in computing expectations and moments of local graph functions, such as subgraph counts.

6 Discussion and Related Work

6.1 Related Models of Random Graphs

While the shortcomings of the classical Erdős-Rényi models have long been recognized, proposing more realistic models is not an easy task. The difficulty lies in achieving a balance between realism and analytical tractability. By now there are thousands of papers proposing different ways to generate graphs with desirable properties [16] and the vast majority of them only provide heuristic arguments to back up their claims. For a gentle introduction the reader is referred to the book of Newman [25] and for a more mathematical treatment to the books of Chung and Lu [12] and of Durrett [13].

The most general tractable model to date are Inhomogeneous Random Graphs, introduced by Bollobás, Janson, and Riordan [6]. The generality and utility of the model stems from allowing two levels of modeling, i.e., both on vertices and on edges. Specifically, each of the n vertices is assigned a feature vector x_i in some separable metric space $\mathcal{S}$ and one takes a joint probability distribution μ on $\mathcal{S}^n$, such that the corresponding point process on $\mathcal{S}$ is well-defined and has a limit. Then, conditional on the vector $\mathbf{x}_n$, edges are formed *independently*, with probability $p_{ij} = \min\{\kappa(x_i,x_j)/n, 1\}$, where κ is a kernel on $\mathcal{S}\times\mathcal{S}$. The authors provide very general results about many graph properties, most notably regrading the existence or not of a giant component, but also about the asymptotic degree distribution and diameter. Inhomogeneous random graphs form a general theoretical framework within which many different models can be unified. At the same time, though, it offers no insight on how to set either of the two key global components μ and κ in order to focus the measure on graphs with a desired set of properties.

The work most closely related to ours is the much heralded idea of Exponential Random Graphs [30, 26]. In this model, one seeks to define a probability distribution on the set $\mathcal{G}_n$ of all graphs with n vertices such that the *expectation* of m functions $H_k(G)$ satisfy some *affine constraints*. Out of all valid distributions,

the max-entropy distribution is selected, having the form $\mathbb{P}_{\boldsymbol{\beta}}(G) = \frac{1}{Z(\boldsymbol{\beta})} \exp(\boldsymbol{\beta} \cdot \mathbf{H}(G))$, where $\boldsymbol{\beta}$ is a parameter vector acting as an inverse temperature and $\mathbf{H}(G) = (H_1(G), \ldots, H_m(G))$ is an (energy) function measuring the deviation from the different prescribed expectations. While this is a very general way of incorporating desired features *without making independence assumptions*, there are a number of issues. A first one, at least conceptually, is that constraints are imposed only in expectation. That means that there is no hope of having the distribution's support confined on a particular set, except of course in the "zero temperature" limit $|\beta_i| \to \infty$. Moreover, in many natural sub-cases, it was proven that even sampling computationally, can take exponential time [4]. Lastly, in most cases were the model is interesting, it becomes analytically intractable as probabilities of events can be calculated only through simulation. All in all, although this approach is conceptually appealing, it has remained largely unsatisfactory as a general modeling tool, as in the absence of structural assumptions the measure rapidly becomes impenetrable.

6.2 Large Deviations

The work of Erdős-Rényi relating $G(n, m)$ and $G(n, p)$ can be seen as fixing a symmetric scalar function $T(x) = 1^T x$ on $H_N = \{0, 1\}^N$, where $N = \binom{n}{2}$, and showing that the subset $\{x : T(x) \in \{m\}\}$ can be approximated by a product measure. We generalized this to a vector function, $\mathbf{m}(x)$, the edge-profile, and a convex set $C \subseteq \mathbb{N}^k$, i.e., to $S = \{x : \mathbf{m}(x) \in C\}$. We saw that the product measure approximability is possible due to the concentration of the random variable $\mathbf{m}(x)$ around a single entropic optimizer $\mathbf{m}^*(S)$. Below we discuss our work under different viewpoints and provide connections with other areas.

Large Deviations in Random Graphs. The uniform measure on the set of all graphs $\mathcal{G}_n$ is equivalent to $G(n, 1/2)$. Consequently, for any set of graphs S, a uniformly random element G of S can be generated by conditioning $G \sim G(n, 1/2)$ to lie in S. Our concentration theorem can be seen as a large deviation principle for $P_{1/2}(\mathbf{m}(G) \in A_\epsilon \,|S)$, where $\mathbb{P}_{1/2}(\cdot)$ refers to the distribution of $G(n, 1/2)$ and A_ϵ to the complement of the concentration set. The study of large deviations of $\mathbb{P}_{1/2}(\cdot)$ has a long history [17, 5] and has recently received renewed interest since the work of Chatterjee and Varadhan [11].

Large Deviations in the Hypercube. We have stated our results in terms of sets (properties) of graphs, having as the primary goal to to demonstrate that what we call the "agnostic" approach to the study of random graphs, i.e., studying the uniform measure over subsets of graphs, can be technically feasible via product measure approximation. In actuality, though, there is no inherent reason to think of N in $H_N = \{0, 1\}^N$ as enumerating the set of $N = \binom{n}{2}$ possible edges on n vertices. Our results hold for any subset S of the hypercube whose characteristic function (i) is $\mathcal{P}$-symmetric with respect to a fixed partition $\mathcal{P}$ of the N coordinates, and (ii) is such that the set of weight-profiles $\mathbf{m}(S)$ is convex (or, more generally, supports a unique entropic maximizer). This brings about the possibility of our results finding application in the more general context of *Boolean Analysis*, especially in relation to monotone properties. The work most similar to ours in this direction is the recent work of Chatterjee and Dembo [10] on *Non-linear Large Deviations*.

References

[1] Dimitris Achlioptas and Paris Siminelakis. Navigability is a robust property. *On-line Manuscript*, 2014.

[2] K. Appel and W. Haken. Every planar map is four colorable. part i: Discharging. *Illinois Journal of Mathematics*, 21(3):429–490, 09 1977.

[3] M. Artin. *Algebra*. Featured Titles for Abstract Algebra Series. Pearson Prentice Hall, 2011.

[4] Shankar Bhamidi, Guy Bresler, and Allan Sly. Mixing time of exponential random graphs. In *Foundations of Computer Science, 2008. FOCS'08. IEEE 49th Annual IEEE Symposium on*, pages 803–812. IEEE, 2008.

[5] Béla Bollobás. *Random graphs*, volume 73 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, Cambridge, second edition, 2001.

[6] Béla Bollobás, Svante Janson, and Oliver Riordan. The phase transition in inhomogeneous random graphs. *Random Structures & Algorithms*, 31(1):3–122, 2007.

[7] C. Borgs, J. T. Chayes, H. Cohn, and Y. Zhao. An L^p theory of sparse graph convergence I: limits, sparse random graph models, and power law distributions. *ArXiv e-prints*, January 2014.

[8] C. Borgs, J.T. Chayes, L. Lovász, V.T. Sós, and K. Vesztergombi. Convergent sequences of dense graphs I: Subgraph frequencies, metric properties and testing. *Advances in Mathematics*, 219(6):1801 – 1851, 2008.

[9] Christian Borgs, Jennifer T Chayes, László Lovász, Vera T Sós, and Katalin Vesztergombi. Convergent sequences of dense graphs ii. multiway cuts and statistical physics. *Ann. of Math.(2)*, 176(1):151–219, 2012.

[10] Sourav Chatterjee and Amir Dembo. Nonlinear large deviations. *arXiv:1401.3495*, 2014.

[11] Sourav Chatterjee and S.R.S. Varadhan. The large deviation principle for the Erdős-Rényi random graph. *European Journal of Combinatorics*, 32(7):1000 – 1017, 2011. Homomorphisms and Limits.

[12] Fan RK Chung and Linyuan Lu. *Complex graphs and networks*, volume 107. American mathematical society Providence, 2006.

[13] Richard Durrett. *Random graph dynamics*, volume 20. Cambridge university press, 2007.

[14] Alan Frieze and Ravi Kannan. Quick approximation to matrices and applications. *Combinatorica*, 19(2):175–220, 1999.

[15] David F Gleich and Art B Owen. Moment-based estimation of stochastic kronecker graph parameters. *Internet Mathematics*, 8(3):232–256, 2012.

[16] Anna Goldenberg, Alice X. Zheng, Stephen E. Fienberg, and Edoardo M. Airoldi. A survey of statistical network models. *Found. Trends Mach. Learn.*, 2(2):129–233, February 2010.

[17] Svante Janson, Tomasz Łuczak, and Andrzej Rucinski. *Random graphs*. Wiley-Interscience Series in Discrete Mathematics and Optimization. Wiley-Interscience, New York, 2000.

[18] Jon M. Kleinberg. Navigation in a small world. *Nature*, 406(6798):845, August 2000.

[19] Jon M. Kleinberg. Small-world phenomena and the dynamics of information. In Thomas G. Dietterich, Suzanna Becker, and Zoubin Ghahramani, editors, *NIPS*, pages 431–438. MIT Press, 2001.

[20] Jure Leskovec, Deepayan Chakrabarti, Jon Kleinberg, Christos Faloutsos, and Zoubin Ghahramani. Kronecker graphs: An approach to modeling networks. *The Journal of Machine Learning Research*, 11:985–1042, 2010.

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

[22] Mohammad Mahdian and Ying Xu. Stochastic kronecker graphs. *Random Structures & Algorithms*, 38(4):453–466, 2011.

[23] Laurent Massoulié. Community detection thresholds and the weak ramanujan property. *CoRR*, abs/1311.3085, 2013.

[24] Frank McSherry. Spectral partitioning of random graphs. In *Foundations of Computer Science, 2001. Proceedings. 42nd IEEE Symposium on*, pages 529–537. IEEE, 2001.

[25] Mark Newman. *Networks: an introduction*. Oxford University Press, 2010.

[26] Garry Robins, Pip Pattison, Yuval Kalish, and Dean Lusher. An introduction to exponential random graph p^* models for social networks. *Social networks*, 29(2):173–191, 2007.

[27] Karl Rohe, Sourav Chatterjee, Bin Yu, et al. Spectral clustering and the high-dimensional stochastic blockmodel. *The Annals of Statistics*, 39(4):1878–1915, 2011.

[28] Dana Ron. Property testing. In *Handbook of Randomized Computing, Vol. II*, pages 597–649. Kluwer Academic Publishers, 2000.

[29] Ronitt Rubinfeld and Asaf Shapira. Sublinear time algorithms. In *In Proceedings of International Congress of Mathematicians 2006*, pages 1095–1110, 2006.

[30] Tom AB Snijders. Markov chain monte carlo estimation of exponential random graph models. *Journal of Social Structure*, 3(2):1–40, 2002.

A Basic Proofs

Proposition 3. *Consider for all $i \in [k]$ disjoint sets of edges $I_i, O_i \subset P_i$ and define the events $A_i = \{G \in \mathcal{G}_n : I_i \subset E(G)$ and $O_i \cap E(G) = \emptyset\}$. Conditional on the edge profile of G being $\mathbf{v}$, the events are independent, i.e. it holds that: $\mathbb{P}_S(A_1 \cap \ldots \cap A_k|\mathbf{v}) = \prod_{i=1}^k \mathbb{P}_S(A_i|v_i)$.*

Proof of Proposition 3. Since $G \sim \mathcal{P}(S)$ the distribution of G is by definition uniform on $\mathcal{S}$. This also means that it is uniform on the subset of graphs having edge profile $\mathbf{m} \in \mathbb{N}^k$(conditioning). But then:

$$\mathbb{P}_S(A_1 \cap \ldots \cap A_k|\mathbf{v}) = \frac{\mathbb{P}_S(A_1 \cap \ldots \cap A_k \cap \mathbf{m}(G) = \mathbf{v})}{\mathbb{P}_S(\mathbf{m}(G) = \mathbf{v})} = \frac{|A_1 \cap \ldots \cap A_k \cap \mathbf{m}(G) = \mathbf{v}|}{|\mathbf{m}(G) = \mathbf{v}|}\mathbb{I}_{\mathbf{m}(S)}(\mathbf{v})$$

where the first equality follows from Bayes rule and the second due to uniformity and the fact that our symmetry assumption implies that membership in S depends only on the edge-profile $\mathbf{m}$. Recall that each set $A_i = \{G \in \mathcal{G}_n : I_i \subset E(G)$ and $O_i \cap E(G) = \emptyset\}$ imposes the requirement that the edges in I_i are included in G and that the edges in O_i are not included in G. Having conditioned on $\sqsubseteq$, we know that exactly v_i edges from P_i are included in G and that we can satisfy the requirements for edges in P_i by selecting any subset of $v_i - |I|_i$ edges out of $P_i \setminus (I_i \cup_i)$. For convenience set $|P_i| = p_i$, $|I|_i = n_i$, $|O_i \cup I_i| = r_i$, and let C_ℓ^n denote the number of k-combinations out of an n element set(binomial coefficient). The number of valid subsets of P_i is then given by $C_{v_i - n_i}^{p_i - r_i}$. As the constraints imposed are separable, we have:

$$\frac{|A_1 \cap \ldots \cap A_k \cap \mathbf{m}(G) = \mathbf{v}|}{|\mathbf{m}(G) = \mathbf{v}|} = \frac{\prod_{i=1}^k C_{v_i - n_i}^{p_i - r_i}}{|\mathbf{m}(G) = \mathbf{v}|} = \prod_{i=1}^k \frac{|A_i \cap \mathbf{m}(G) = \mathbf{v}|}{|\mathbf{m}(G) = \mathbf{v}|}$$

which gives the required identity by exploiting again uniformity of the probability measure. □

Proof of Proposition 1. Fix an $\mathbf{x} \in H_n$ and consider the set $O(\mathbf{x}) \stackrel{\triangle}{=} \{\mathbf{y} \in H_n : \exists \pi \in \Pi_n(\mathcal{P})$ such that $y = \pi(\mathbf{x})\}$ and call it the *orbit* of $\mathbf{x}$ under $\Pi_n(\mathcal{P})$(note that by group property orbits form a partition of H_n). The assumption of symmetry, implies that f is constant for all $\mathbf{y} \in O(\mathbf{x})$:

$$f(\mathbf{y}_1) = f(\mathbf{y}_2) = f(\mathbf{x}),\ \forall \mathbf{y}_1, \mathbf{y}_2 \in O(\mathbf{x})$$

By definition of $\Pi_n(\mathcal{P})$, for any $\mathbf{x} \in H_n$ there is a permutation $\pi_x \in \Pi_n(\mathcal{P})$, such that i) $\pi_x(\mathbf{x}) = (\pi_{x,1}(x_{P_1}), \ldots, \pi_{x,k}(x_{P_k})) \in O(\mathbf{x})$, ii) for all $i \in [k]$, $\pi_{x,i}(x_{P_i})$ is a bit-string where all 1's appear consequently starting from the first position. Let as identify with each orbit $O \subset H_n$ such a distinct element $\mathbf{x}_o$. As the function of f is constant along each orbit, its value depends only through $\mathbf{x}_o$, which in turn depends only on the number of 1's(edges) in each part, encoded in the edge profile $\mathbf{m} = (m_1, \ldots, m_k)$. □

Proof of Proposition 2. Consider $\mathcal{P}$ any balanced partition consisting of k-parts and let $S = \mathcal{G}_n$ to be the space of all graphs. Then $Z(S) = |S| = 2^{\binom{n}{2}}$ and $\mathbf{m}^*(S)$ is the all $\binom{n}{2}/2k$ vector (all blocks half full). Using Stirling's approximation of the factorial we have that:

$$\begin{aligned} \log(|S|) - \textsc{Ent}(\mathbf{m}^*) &\geq \binom{n}{2}\log 2 - 2k \log \binom{\binom{n}{2}/k}{\frac{1}{2}\binom{n}{2}/k} && (10) \\ &\geq k \log n - \frac{k}{2}\log k && (11) \end{aligned}$$

For $k = o(n^2)$ the last expression is of order $\Omega(k \log(n))$.

□

B Proof of Theorem 2

In this section we prove Theorem 2. For the purposes of the proof we are going to employ, instead of $\mathbf{m}(x)$, a different parametrization in terms of the *edge-profile* $\mathbf{a}(x) = (a_1(x), \ldots, a_k(x)) \in [0,1]^k$ where $a_i(x) = m_i(x)/p_i$. This will be convenient both in calculations as well as conceptually as $a_i(x)$ represents the effective edge density of a part P_i in the partition. We start by approximating the entropy of an edge-profile via the $\mathcal{P}$-entropy.

Definition 3. *Given a partition $\mathcal{P}$, the $\mathcal{P}$-entropy is define for every $\mathbf{a} \in [0,1]^k$ as*

$$H_{\mathcal{P}}(\mathbf{a}) = -\sum_{i=1}^{k} p_i \left[a_i \log a_i + (1-a_i)\log(1-a_i)\right] \tag{12}$$

The $\mathcal{P}$-entropy is simply the entropy of the product measure defined over edges through $\mathbf{a}$. We slightly abuse the notation and also define the $\mathcal{P}$-entropy in terms of the edge-profile:

$$H_{\mathcal{P}}(\mathbf{v}) = -\sum_{i=1}^{k} \left[v_i \log\left(\frac{v_i}{p_i}\right) + (p_i - v_i)\log\left(\frac{p_i - v_i}{p_i}\right)\right] \tag{13}$$

Let $\mathcal{M}_{\mathcal{P}} := \{0, \ldots, p_1\} \times \ldots \times \{0, \ldots, p_k\}$ be the space of all possible vectors $\mathbf{m}$. In what follows we sometimes supress the dependence of the quantities in $\mathbf{m}$ or $\mathbf{a}$ to ease the notation.

Lemma 1. *Let $\mathbf{m} \in \mathcal{M}_{\mathcal{P}}$ be an edge-profile and $\mathbf{a} \in [0,1]^k$ be the corresponding probability profile, then:*

$$\textsc{Ent}(\mathbf{m}) = \mathcal{H}_{\mathcal{P}}(\mathbf{a}) - \gamma(n)$$

where $0 \le \gamma(n) \le k \log n$ is a term that approaches zero as m_i and $p_i - m_i$ tend to infinity.

Proof. We begin by providing the first order Stirling approximation for a single term of the form $\log \binom{p_i}{m_i}$. Specifically, since $m_i = p_i a_i$ and by using $\log n! = n \log n - n + \frac{1}{2}\log n + \theta_n$, where $\theta_n \in (0,1]$, we get:

$$\begin{aligned} \log \binom{p_i}{m_i} &= \log(p_i!) - \log(m_i!) - \log((p_i - m_i)!) \\ &= -p_i \left[a_i \log a_i + (1-a_i)\log(1-a_i)\right] - \delta_n(a_i, p_i) \ , \end{aligned}$$

where $0 \le \delta_n(a_i, p_i) \le \log n$. Summing the derived expression for all $i \in [k]$ gives:

$$\begin{aligned} \textsc{Ent}(\mathbf{m}) &= -\sum_{i=1}^{k} p_i \left[a_i \log a_i + (1-a_i)\log(1-a_i)\right] - \sum_{i=1}^{k} \delta_n(a_i, p_i) \\ &= H_{\mathcal{P}}(\mathbf{a}) - \gamma(n) \ , \end{aligned}$$

where $0 \le \gamma(n) \le k \log n$. □

Next, using the Taylor remainder theorem and the connection with $\mathcal{P}$-entropy, we obtain geometric estimates on the decay of entropy around $\mathbf{m}^*$.

Theorem 4 (Taylor Remainder Theorem)**.** *Assume that f and all its partial derivatives are differentiable at every point of an open set $S \subseteq \mathbb{R}^K$. If $\mathbf{a}, \mathbf{b} \in S$ are such that the line segment $L(\mathbf{a}, \mathbf{b}) \subseteq S$, then there exists a point $\mathbf{z} \in L(\mathbf{a}, \mathbf{b})$ such that:*

$$f(\mathbf{b}) - f(\mathbf{a}) = \nabla f(\mathbf{a})^T(\mathbf{b} - \mathbf{a}) + \frac{1}{2}(\mathbf{b} - \mathbf{a})^T \nabla^2 f(\mathbf{z})(\mathbf{b} - \mathbf{a}) \ . \tag{14}$$

Lemma 2 (L_2 distance bounds)**.** *If $\mathbf{m}^*$ is the unique maximizer and $\mathbf{w} \in \mathrm{Conv}(\mathbf{m}(S))$, then*

$$\textsc{Ent}(\mathbf{w}) - \textsc{Ent}(\mathbf{m}^*) \leq -\sum_{i=1}^{k} \frac{(w_i - m_i^*)^2}{\max\{\tilde{m}_i^*, \tilde{w}_i\}} + 3k \log n \tag{15}$$

where $\mathbf{m}_i^* = \min\{m_i^*, p_i - m_i^*\}$(respectively $\tilde{w}_i$), denotes the thickness *of a part $i \in [k]$.*

Proof. Invoking Lemma 1, we rewrite the difference in entropy as a difference in $\mathcal{P}$-entropy, where $\mathbf{a}^*$ is the probability profile of the maximizer and $\mathbf{b}$ of $\mathbf{w}$:

$$\textsc{Ent}(\mathbf{w}) - \textsc{Ent}(\mathbf{m}^*) \leq H_{\mathcal{P}}(\mathbf{b}) - H_{\mathcal{P}}(\mathbf{a}^*) + 3k \log n$$

Here, we have additionally dealt with the subtle integrality issue, namely that $\mathbf{m}^*$ might not belong to $\mathbf{m}(S)$. Rounding the vector to the nearest integral point produces a cumulative error of at most $2k \log(n)$ in the entropy that adds to the $k \log(n)$ error coming from Stirling's approximation. Both errors can be reduced using higher order Stirling approximations but we avoid doing so since an error of order $k \log(n)$ is unavoidable due to the approximation of the partition function.

Convexity of the domain $\mathrm{Conv}(\mathbf{m}(S))$ and differentiability of the $\mathcal{P}$-entropy provide the necessary conditions to use the Taylor Remainder Theorem. Let $\mathbf{z}$ be a point in the linear segment $L(\mathbf{a}^*, \mathbf{b})$. We proceed with writing the expressions for partial derivatives of $H_{\mathcal{P}}$.

$$\partial_i H_{\mathcal{P}}(\mathbf{a}^*) = -p_i \log\left(\frac{a_i^*}{1 - a_i^*}\right) \tag{16}$$

$$\partial_{ii}^2 H_{\mathcal{P}}(\mathbf{z}) = -p_i \left(\frac{1}{1 - z_i} + \frac{1}{z_i}\right) \ , \tag{17}$$

while $\partial_{ij}^2 f = 0$ for $i \neq j$ due to separability of the function $H_{\mathcal{P}}$. The Taylor Remainder forumla, now reads:

$$H_{\mathcal{P}}(\mathbf{b}) - H_{\mathcal{P}}(\mathbf{a}^*) = \nabla H_{\mathcal{P}}(\mathbf{a}^*) \cdot (\mathbf{b} - \mathbf{a}^*) - \sum_{i=1}^{K} p_i (b_i - a_i^*)^2 \left(\frac{1}{1 - z_i} + \frac{1}{z_i}\right) \tag{18}$$

Since, $\mathbf{a}^*$ is the unique solution to the sc MaxEnt problem and the domain is convex, the first term in the above formula is always bounded above by zero. Otherwise, there would be a direction $\mathbf{u}$ and a small enough parameter $\epsilon > 0$ such that $\mathbf{a}^* + \epsilon \mathbf{u}$ has greater entropy, a contradiction. To bound the second sum from above, let $\tilde{z}_i = \min\{z_i, 1 - zi\}$(expressing the fact that binary entropy is symmetric around $1/2$) and use the trivial bound $\tilde{z}_i \leq \max\{\tilde{a}_i^*, \tilde{b}_i\}$. Thus,

$$H_{\mathcal{P}}(\mathbf{b}) - H_{\mathcal{P}}(\mathbf{a}^*) \leq -\sum_{i=1}^{K} p_i (b_i - a_i^*)^2 \frac{1}{\tilde{z}_i} \leq -\sum_{i=1}^{K} p_i \frac{(b_i - a_i^*)^2}{\max\{\tilde{a}_i^*, \tilde{b}_i\}} \ . \tag{19}$$

Dividing and multiplying by p_i, and writing $\tilde{w}_i = p_i \tilde{b}_i$, $\tilde{m}_i^* = p_i \tilde{a}_i^*$, gives:

$$H_{\mathcal{P}}(\mathbf{b}) - H_{\mathcal{P}}(\mathbf{a}^*) \leq -\sum_{i=1}^{K} \frac{(w_i - m_i^*)^2}{\max\{\tilde{m}_i^*, \tilde{w}_i\}} \ . \tag{20}$$

where $\mathbf{w}$ and $\mathbf{m}^*$ are the original edge profiles. We note that for most cases we have that $\tilde{z}_i = z_i$, i.e. a block is half-empty. □

In preparation of performing the "union bound", we prove that:

Proposition 4. *The number of distinct edge-profiles $|\mathbf{m}(S)|$ is bounded by $|\mathcal{M}_{\mathcal{P}}| \leq e^{2k \log n}$.*

Proof. Assuming that no constraint is placed upon $\mathbf{m}$ by S, then $\mathbf{m}(S) = \mathcal{M}_{\mathcal{P}}$. This number is equal to the product of $[p_i + 1] \leq n^2$ as there are at most $\binom{n}{2}$ edges within a block. Multiplying the last bound we get the statement. □

Before proceeding with the proof of the concentration theorem, we repeat the definitions of the crucial parameters mentioned in the introduction.

Definition 4. *Given a partition $\mathcal{P}$ and a $\mathcal{P}$-symmetric set S, define:*

$$\mu(S) = \min_{i \in [k]} \{m_i^* \wedge (p_i - m_i^*)\} \tag{21}$$

$$\lambda(S) = \frac{5k \log n}{\mu(S)} \tag{22}$$

$$r(S) = \frac{\lambda + \sqrt{\lambda^2 + 4\lambda}}{2} \tag{23}$$

the Thickness*,* condition number *and* resolution *of the convex set S and $x \wedge y := \min\{x, y\}$ denotes the* min *operation.*

Proof of Theorem 2. Our goal is to use the developed machinery to control the probability of deviations from the optimum at scale $\epsilon > r(S)$. Define the set $\mathcal{L}_\epsilon(\mathbf{m}^*) \triangleq \{x \in S : |\mathbf{m}(x) - \mathbf{m}^*| \leq \epsilon \tilde{\mathbf{m}}^*\}$. We are going to show that $\mathbb{P}_S(\mathcal{L}_\epsilon^c(\mathbf{m}^*)) \to 0$ "exponentially" fast and thus provide localization of the edge profile within a scale ϵ for each coordinate. To that end, we write:

$$\mathbb{P}_S(\mathcal{L}_\epsilon^c(\mathbf{m}^*)) = \sum_{\mathbf{w} \in \mathcal{L}_\epsilon^c(\mathbf{m}^*)} \mathbb{P}_S(\mathbf{w}) \leq \sum_{\mathbf{w} \in \mathcal{L}_\epsilon^c(\mathbf{m}^*)} \exp\left[\textsc{Ent}(\mathbf{m}) - \textsc{Ent}(\mathbf{m}^*)\right] \tag{24}$$

where we have used (9), the approximation of the log-partition function by the entropy of the optimum edge-profile, and added the contribution of points outside of $\mathbf{m}(S)$. At this point, we are going to leverage the lower bound for the decay of entropy away from $\mathbf{m}^*$. This is done by first performing a union bound, i.e. considering that all points in $\mathcal{L}_\epsilon^c(\mathbf{m}^*)$ are placed on the least favorable such point $\mathbf{w}^*$. Since, we are requiring coordinate-wise concentration, such point would differ from the optimal vector only in one-coordinate, and in particular should be the one that minimizes our lower bound. Any such vector $\mathbf{w} \in \mathcal{L}_\epsilon(\mathbf{m}^*)$, would have at least one coordinate $i \in [k]$ such that $|w_i - m_i^*| = \epsilon m_i^*$. By Lemma 15, we get

$$\textsc{Ent}(\mathbf{w}) - \textsc{Ent}(\mathbf{m}^*) \leq -\frac{\epsilon^2 (m_i^*)^2}{(1+\epsilon) m_i^*} + 3k \log n = -\frac{\epsilon^2}{(1+\epsilon)} m_i^* + 3k \log n \tag{25}$$

using the facts that $\max\{\tilde{m}_i, \tilde{w}_i\} \leq \tilde{m}_i + \tilde{w}_i \leq (1+\epsilon)m_i^*$. Now, by definition the thickness $\mu(S) \leq \tilde{m}_i^*$ for all $i \in [k]$, and so a a vector $\mathbf{w}^*$ that minimizes the bound is such that $\textsc{Ent}(\mathbf{w}^*) - \textsc{Ent}(\mathbf{m}^*) \leq -\frac{\epsilon^2}{(1+\epsilon)}\mu(G) + 3k\log n$. We perform the union bound by using $|\mathcal{L}_\epsilon^c(\mathbf{m}^*)| \leq |\mathcal{M}_\mathcal{P}| \leq \exp(2k\log n)$ from Proposition 4:

$$\begin{aligned}
\mathbb{P}_S(\mathcal{L}_\epsilon^c(\mathbf{m}^*)) &\leq |\mathcal{L}_\epsilon^c(\mathbf{m}^*)| \cdot \mathbb{P}_S(\mathbf{w}^*) && (26)\\
&\leq \exp\left[-\frac{\epsilon^2}{(1+\epsilon)}\mu(G) + 5k\log n\right] && (27)\\
&\leq \exp\left[-\mu(G)\left(\frac{\epsilon^2}{1+\epsilon} - \frac{5k\log n}{\mu(G)}\right)\right] && (28)
\end{aligned}$$

Finally, identifying $\lambda(S)$ in the expression provides the statement. We note here that the resolution $r(S)$ is defined exactly so that the expression in the exponent is negative. The condition $\lambda(S) \ll 1$ is a requirement that makes concentration possible in a small scale, i.e $\epsilon \ll 1$. □

B.1 Comments on the Proof

Tightness. The crucial steps in the proof are, firstly, the approximation of the log-partition function and, secondly, the L_2 distance bounds on the decay of entropy away from the optimum. Both steps are essentially optimal under general assumptions, as is shown in Proposition 2. Our proof can only improved by using higher order Stirling approximations and a more complicated integration process (incremental union bounds over L_∞-shells) instead of the simple union bound, to reduce the error from $5k\log n$ down to possibly the minimum of $2k\log(n)$. Since, the above considerations would complicate the proof significantly and the gain is a small improvement in the constant we deem this unnecessary.

Relaxing Conclusions. Our theorem seeks to provide concentration simultaneously for all parts. That motivates the definition of thickness parameter $\mu(S)$ as the minimum distance from the natural boundary of the number of edges that any part i has at the optimum $\mathbf{m}^*$. Quantifying everything in terms of $\mu(S)$ is a very conservative requirement. For instance, if we define the set S to have no edges in a particular part of the partition, then $\mu(S)$ is 0 and our conclusions become vacuous. Our proofs in reality generalize, to the case where we confine our attention only to a subset $I \subseteq [k]$ of blocks in the partition. In particular, if one defines I^* as the set of parts whose individual *thickness* parameter $\tilde{m}_i = \min\{m_i, p_i - m_i\}$ is greater than $5k\log(n)$, both theorems hold for the subset of edges $\cup_{i\in I^*} P_i$. In essence that means that for every class that is "well-conditioned", we can provide concentration of the number of edges and approximate monotone properties of only those parts by coupling them with product measures.

Relaxing Convexity. Besides partition symmetry that comprises our main premise and starting point, the second main assumption made about the structure of S is *convexity*. In the proof convexity is used only in Lemma 2 to argue that: (i) the maximizer $\mathbf{m}^*$ will be close to some vector in $\mathbf{m}(S)$, and (ii) that the first order term in the Taylor approximation of the entropy is always negative. However, since the optimization problem was defined on the convex hull of $\mathbf{m}(S)$, in point (ii) above we are only using convexity of $\mathrm{Conv}(\mathbf{m}(S))$ and not of the set S. Thus, the essential requirement on $\mathcal{P}$-symmetric sets is *approximate unimodality*.

Definition 5. *A $\mathcal{P}$-symmetric set S is called Δ-unimodal if the solution $\mathbf{m}^*$ to the entropy optimization problem defined in Section 2, satisfies:*

$$d_1(\mathbf{m}^*, S) := \min_{\mathbf{v}\in\mathbf{m}(S)} \|\mathbf{m}^* - \mathbf{v}\|_1 \leq \Delta \tag{29}$$

Convexity essentially implies that the set S is k-unimodal as we need to round each of the k coordinates of the solution to the optimization problem to the nearest integer. Under this assumption, all our results apply by only changing the condition number of the set to $\lambda(S) = \frac{(2\Delta+3k)\log(n)}{\mu(S)}$. In this extended abstract, we opted to present our results by using the familiar notion of convexity to convey intuition on our results and postpone the presentation in full generality for the full version of the paper.

C Proof of Theorem 1

In this section, we leverage the concentration theorem to prove that convex $\mathcal{P}$-symmetric sets are (ϵ, δ)-sandwichable. Before presenting the proof of the theorem we state a preliminary lemma.

C.1 Basic Coupling Lemma

Consider a set of random variables $X_1, \ldots, X_k$ with laws $\mu_1, \ldots, \mu_k$. A coupling between a set of random variables is a (joint) probability distribution μ, such that $\mathbb{P}_\mu(X_i) = \mathbb{P}_{\mu_i}(X_i)$ for $i \in [k]$, i.e. the marginals of the random variables are right. Let $A = \{1, \ldots, N\}$ be a finite set with N elements. Further let X denote a uniform subset of m elements of A, denoted as $X \sim \mathrm{Samp}(m, A)$, and Z a subset of A where each element of A is included with the same probability p, denote as $Z \sim \mathrm{Flip}(p, A)$.

Lemma 3. *Given a set A with N elements and a number m, define $p^{\pm}(m) = \frac{m}{(1\mp\delta)N}$. Consider the random variables $X \sim \mathrm{Samp}(m, A)$ and $Z^{\pm} \sim \mathrm{Flip}(p^{\pm}, A)$, then there exists a coupling μ such that:*

$$\mathbb{P}_\mu\left(Z^- \subseteq X \subseteq Z^+\right) \geq 1 - 2\exp\left(\frac{\delta^2}{3(1+\delta)m}\right). \tag{30}$$

Proof. Let μ be the joint distribution of $U_1, \ldots, U_N$ i.i.d uniform in $[0,1]$ random variables, and $U_{(m)}$ to denote the m-th smallest such random variable. Define $X(U) = \{i \in A : U_i \leq U_{(m)}\}$ and $Z^{\pm}(U) = \{i \in A : U_i \leq p^{\pm}\}$ to be random subsets of A. By construction it is easy to see that $X(U)$ and $Z^{\pm}(U)$ have the right marginals. By construction of the sets, it is easy to see that the following equivalence holds:

$$Z^- \subseteq X \subseteq Z^+ \Leftrightarrow |Z^-| \leq |X| \leq |Z^+|$$

To analyze the second event define the "bad" events:

$$\begin{aligned} B_- &= \{u \in [0,1]^N : \sum_{i=1}^N \mathbb{I}(u_i \leq p_-) > m\} \\ B_+ &= \{u \in [0,1]^N : \sum_{i=1}^N \mathbb{I}(u_i \leq p_+) < m\} \end{aligned}$$

Each event can be stated as the probability that the sum $X_\pm$ of n i.i.d Bernoulli $p_\pm$ random variables exceeds (smaller then) the expectation $np_\pm$. By employing standard Chernoff bounds, we get:

$$\begin{aligned} \mathbb{P}_\mu(B_-) &= \mathbb{P}_U(X_- > m) = \mathbb{P}_\mu(X > (1+\delta)np_-) \leq \exp(-\frac{\delta^2}{3(1+\delta)}m) \\ \mathbb{P}_U(B_+) &= \mathbb{P}_U(X_+ < m) = \mathbb{P}_U(X < (1-\delta)np_+) \leq \exp(-\frac{\delta^2}{2(1-\delta)}m) \end{aligned}$$

The proof is concluded through the use of union bound:

$$\mathbb{P}_\mu(B_- \cup B_+) \leq \mathbb{P}_\mu(B_-) + \mathbb{P}_U(B_+) \leq 2\exp\left(-\frac{\delta^2}{3(1+\delta)}m\right)$$

This concludes the lemma. □

Using, this simple lemma and Theorem 2, we prove the *sandwich theorem*.

C.2 Main proof.

Recall, that our aim is to prove that the uniform measure over the set S is (ϵ, δ)-sandwichable by some product measure $G(n, \mathbf{Q})$.

Proof of Theorem 1. Given a $\mathcal{P}$-symmetric convex set S, consider $\mathbf{m}^*(S)$ the optimal edge-profile and define the $n \times n$ matrix $\mathbf{Q}^*(S)$ as: $Q_{u,v} = \frac{m_i^*}{p_i}, \forall \{u, v\} \in P_i$ and $i \in [k]$. Further, define $\mathbf{q}_i := \frac{m_i^*}{p_i}, \forall \in [k]$ to be used later. In order to prove the required statement, we need to construct a coupling between the random variables $G \sim U(S)$, $G^\pm \sim G(n, (1\pm)\mathbf{Q}^*)$. By separating edges according to the partition, we express the edge set of the graphs as $E(G) = E_1 \cup \ldots \cup E_k$ and $E(G^\pm) = E_1^\pm \cup \ldots \cup E_k^\pm$.

Let μ denote the joint probability distribution of $N+1$ i.i.d. uniform random variables $U_1, \ldots, U_{N+1}$ on $[0, 1]$. As in the coupling lemma, we are going to use these random variables to jointly generate the random edge-sets of G^-, G, G^+. Using U_{N+1}, we can first generate the edge profile $\mathbf{v} \in \mathbf{m}(S)$ from its corresponding distribution. Then, conditional on the edge profile $\mathbf{v} \in \mathbb{N}^k$, the probability distribution of G factorizes in $G(n, m)$-like distributions for each block (Appendix D). Lastly, we associate with each edge e a unique random variable U_e and construct a coupling for edges in each block separately.

In our notation, $E_i \sim \mathrm{Samp}(v_i, P_i)$ and $E_i^\pm \sim \mathrm{Flip}(q_i^\pm, P_i)$. Using Lemma3, we construct a coupling for each $i \in [k]$ between the random variables E_i, E_i^+, E_i^- and bound the probability that the event $E_i^- \subseteq E_i \subseteq E_i^+$ does not hold. Using the union bound over the k parts, we then obtain an estimate of the probability that the property holds across blocks, always conditional on the edge-profile $\mathbf{v}$. The final step involves getting rid of the conditioning by invoking the concentration theorem.

Concretely, define B_i the event that the i-th block does not satisfy the property $E_i^- \subseteq E_i \subseteq E_i^+$ and $\mathcal{L}_\epsilon(\mathbf{m}^*)$ the set appearing in Theorem 2. We have that $\mathbb{P}_\mu(G^- \subseteq G \subseteq G^+) = 1 - \mathbb{P}_\mu(\cup B_i)$. Conditioning on the edge profile gives:

$$\begin{aligned}
\mathbb{P}_\mu(\cup B_i) &\leq \mathbb{P}_\mu(\mathcal{L}_\epsilon^c(\mathbf{m}^*)) + \sum_{\mathbf{v} \in L_\epsilon(\mathbf{m}^*)} \mathbb{P}_\mu(\cup B_i|\mathbf{v})\mathbb{P}_\mu(\mathbf{v}) \\
&\leq \mathbb{P}_\mu(\mathcal{L}_\epsilon^c(\mathbf{m}^*)) + \max_{\mathbf{v} \in L_\epsilon(\mathbf{m}^*)} \mathbb{P}_\mu(\cup B_i|\mathbf{v}) \\
&\leq \mathbb{P}_\mu(\mathcal{L}_\epsilon^c(\mathbf{m}^*)) + \max_{\mathbf{v} \in L_\epsilon(\mathbf{m}^*)} \left[\sum_{i=1}^k \mathbb{P}_\mu(B_i|\mathbf{v})\right]
\end{aligned}$$

The first inequality holds by conditioning on the edge profile and bounding the probability of the bad events from above by 1 for all "bad" profiles (outside of the concentration set). The second inequality, is derived by upper bounding the probability of the bad event by the most favorable such edge-profile and the last inequality follows from an application of the union bound. Applying Theorem 2, we get a bound on the first term and then invoking Lemma 3 we get a bound for each of the term in the sum:

$$\mathbb{P}_\mu(\cup B_i) \leq \exp\left[-\mu(S)\left(\frac{\epsilon^2}{1+\epsilon} - \lambda(S)\right)\right] + 2\max_{\mathbf{v} \in L_\epsilon(\mathbf{m}^*)} \left[\sum_{i=1}^k \exp\left(-\frac{\epsilon^2}{3(1+\epsilon)}v_i\right)\right]$$

Hence, we see that the upper bound is monotone in v_i for all $i \in [k]$. Additionally, we know that for all $\mathbf{v} \in L_\epsilon(\mathbf{m}^*)$ it holds that $\mathbf{v} \geq (1-\epsilon)\mathbf{m}^*$. Further, by definition we have $\mathbf{m}^* \geq \mu(S)$. The bound now becomes:

$$\begin{aligned} \mathbb{P}_\mu(\cup B_i) &\leq \exp\left[-\mu(S)\left(\frac{\epsilon^2}{1+\epsilon} - \lambda(S)\right)\right] + 2k\exp\left[-\frac{\epsilon^2(1-\epsilon)}{3(1+\epsilon)}\mu(S)\right] \quad (31) \\ &\leq \exp\left[-\mu(S)\left(\frac{\epsilon^2}{1+\epsilon} - \lambda(S)\right)\right] + \exp\left[-\mu(S)\left(\frac{\epsilon^2(1-\epsilon)}{3(1+\epsilon)} - \frac{\log(2k)}{\mu(S)}\right)\right] \quad (32) \end{aligned}$$

Finally, using $\epsilon < 1/2$ and $\log(2k)/\mu(S) \leq \lambda(S)$ we arrive at the required conclusion. □

D Approximate Independence for Local Events

The concentration and sandwich theorems dramatically enable the study of monotone properties under the uniform measure over convex-symmetric sets. Going beyond monotone properties, we would like to enable the study of more local (involving a few edges) events. For example, we would like to be able to make statements about moments of subgraph counts[17] (triangle, cliques, cycles) and other local-graph functions.

D.1 Overview

We will denote the probability under the uniform measure over S as $\mathbb{P}_S(\cdot)$, and the probability under the product measure, defined through the maximizer $\mathbf{m}^*(S)$, as $\mathbb{P}_{\mathbf{q}^*}(\cdot)$, where $\mathbf{q}^*$ denotes the corresponding probability-profile. Our goal is to provide a local version of the Total Variation Distance $\sup_A |\mathbb{P}_S(A) - \mathbb{P}_{\mathbf{q}^*}(A)|$ by restricting the sets A under consideration to be *local*. In fact, we are aiming for something far stronger than that, namely to prove that for local events the probability assigned by the uniform measure and its product measure approximation are *essentially the same*, and any differences are of lower order.

Consider any disjoint sets of edges $I_i, O_i \subseteq P_i$ and let $A_i = \{G \in \mathcal{G}_n : I_i \subseteq E(G) \text{ and } O_i \cap E(G) = \emptyset\}$ for $i \in [k]$, i.e., the event that involves the inclusion of the edges in I_i and the exclusion of the edges in O_i. To simplify the analysis, we assume that all the probabilities $q_i^* = m_i^*/p_i$ are bounded above by $1/2$. We now proceed with the definition of *local events*.

Definition 6 (Simple)**.** *Fix $\epsilon > r(S)$, a simple event A_i is* ϵ-local *if $|I_i \cup O_i| \leq \epsilon \cdot m_i^*$.*

Definition 7 (Composite)**.** *Furthermore, a composite event $A = \cap_{i=1}^k A_i$ is* ϵ-local *if all events A_i are ϵ-local and additionally: $\mathbb{P}_{\mathbf{q}^*}(A) \geq \exp\left[-\mu(S)\left(\frac{\epsilon^2}{1+\epsilon} - \lambda(S)\right)\right] \cdot \epsilon$.*

Our definition of local events involves the number of edges appearing as well as the probability of the events. The first requirement is very natural if one thinks of the $G(n,m)$ and $G(n,p)$ models, where the probabilities assigned to the events start to differ for events that involve the inclusion of $\sqrt{m}$ edges (order of fluctuations in $G(n,p)$). The second requirement stems from the fact that we do not have any hope to analyze events that happen with much lower probability than we have control over the space of edge-profiles through Theorem 2. At this point we have motivated both of our conditions on events, and we state our theorem. We purposefully have instantiated many of the constants in order to give more explicit as opposed to more general bounds.

Theorem 5. *Given a $\mathcal{P}$-symmetric convex set S with $\mu(S) \gg (5k\log n)^2$, define $\hat{\epsilon} \triangleq \sqrt{c^*\lambda(S)}$ where $c^* \approx 2$ is explicitly defined. For an $\hat{\epsilon}$-local event A involving at most $m \leq \frac{\sqrt[4]{\mu(S)}}{\sqrt{c^*5k\log n}}$ edges, it holds:*

$$\left|\frac{\mathbb{P}_S(A)}{\mathbb{P}_{\mathbf{q}^*}(A)} - 1\right| \leq \frac{8}{\sqrt[4]{\mu(S)}} \tag{33}$$

In reality, one can obtain different versions of our theorem depending on what she is willing to trade-off, accuracy in approximating the probabilities or bound on the total number of edges involved. Our choices where made so as to make the statement of the theorem as compact as possible.

To get a sense of the conclusion of the theorem lets assume that $\mu(S) = (5k\log n)^2\log(n)$. This is equivalent as saying that each block of the partition includes at least $\mu(S)$ edges at the optimum. Then, if we also assume that $k = \sqrt[3]{n}$, we get that $\mu(S) = 25n^{2/3}\log^3 n$ and $\lambda = 1/(5k^{1/3}\log^2 n) = O(n^{-1/3})$. The total number of edges in this case is bounded below by $k\mu(S) = 25n\log(n)^3$. In the above scenario, our theorem allows for $\Omega\left(n^{-1/6}\right)$-local events involving at most $m = O(\sqrt[4]{\log n})$ edges in total. Furthermore, for these constants the conclusion of the theorem states that the exact probability of the uniform measure assigned to the event, is approximated by the probability of the product measure within a ratio of $(1 \pm \delta)$, where $\delta = O(n^{-1/6})$.

At this point we note that even for this small number of edges $O(\sqrt[4]{\log n})$, there are interesting events that can be calculated. As an example, moments of the random variable Z the total number of triangles can be calculated up to order $\log\log n$. To be more precise our theorem can be applied for the calculation of expectations of *local functions*.

Definition 8. *A function $f : \mathcal{G}_n \to \mathbb{R}$ is called ϵ-local for a $\mathcal{P}$-symmetric set S if $\forall x \in \mathbb{R}$ the event $B_x \triangleq \{G \in \mathcal{G}_n : \ f(G) \leq x\}$ is ϵ-local.*

Corollary 1. *Under the conditions of Theorem 5, for any ϵ-local function $f : \mathcal{G}_n \to \mathbb{R}$, it holds that:*

$$|\mathbb{E}_S[f(G)] - \mathbb{E}_{\mathbf{q}^*}[f(G)]| \leq \frac{8\gamma}{\sqrt[4]{\mu(S)}}\mathbb{E}_{\mathbf{q}^*}[|f(G)|] \tag{34}$$

Proof. Observe that since the function is defined on the discrete space $\mathcal{G}_n$ it takes at most $2^{\binom{n}{2}}$ values. Let $I^+ = \{x \in \mathbb{R} : \exists G \in \mathcal{G}_n, f(G) = x > 0\}$ and I^-be (respectively) the positive and negative values that f can take. We simply multiply (33) by $\mathbb{P}_{\mathbf{q}^*}(A)$, and integrate it over the two sets. □

D.2 Proof of Theorem 5

Let $|I_i| = s_i$ and $|O_i| = n_i - s_i$ be the cardinalities of the sets appearing in A_i. We start with a lemma that provides estimates about the conditional probability distribution within each block:

Lemma 4. *For a $\mathcal{P}$-symmetric convex set S with $\lambda(S) < 1$, let $\epsilon > r(S)$ and consider $\mathbf{v} \in \mathcal{L}_\epsilon(\mathbf{m}^*)$(in the concentration region). For a ϵ-local event A_i, we have:*

$$(1-2\epsilon)^{n_i} \leq \frac{\mathbb{P}_S(A_i|\mathbf{v})}{P_{\mathbf{q}^*}(A_i)} \leq (1-\epsilon)^{-n_i}(1+\epsilon)^{n_i-s_i} \tag{35}$$

where $(q_i^*)^{s_i}(1-q_i^*)^{n_i-s_i} = P_{\mathbf{q}^*}(A_i)$ is the probability of the event under the product measure.

Proof. We start of by righting an exact expression for $\mathbb{P}_S(A_i|\mathbf{v})$:

$$\mathbb{P}_S(A_i|\mathbf{v}) = \frac{\binom{p_i-n_i}{v_i-s_i}}{\binom{p_i}{v_i}} = \frac{(v_i)_{s_i}(p_i-v_i)_{n_i-s_i}}{(p_i)_{n_i}} \tag{36}$$

where $(x)_n$ denotes the descending factorial, for which we get the following easy bounds $(x-n)^n \leq x(x-1)\dots(x-n+1) \leq x^n$. First, we obtain a lower bound on $\mathbb{P}_S(A_i|\mathbf{v})$:

$$\begin{aligned}
\mathbb{P}_S(A_i|\mathbf{v}) &\geq \frac{(v_i-s_i)^{s_i}(p_i-v_i-n_i+s_i)^{n_i-s_i}}{p_i^{n_i}} && (37)\\
&\geq \left[(1-2\epsilon)\frac{m_i^*}{p_i}\right]^{s_i}\left[1-(1+2\epsilon)\frac{m_i^*}{p_i}\right]^{n_i-s_i} && (38)\\
&\geq (1-2\epsilon)^{s_i}(q_i^*)^{s_i}(1-q_i^*)^{n_i-s_i}\left[1-\frac{q_i^*}{1-q_i^*}2\epsilon\right]^{n_i-s_i} && (39)\\
&\geq \mathbb{P}_{\mathbf{q}^*}(A_i)(1-2\epsilon)^{s_i}\left[1-2\epsilon\right]^{n_i-s_i} && (40)
\end{aligned}$$

The second inequality was derived using $\mathbf{v} \in \mathcal{L}_\epsilon(\mathbf{m}^*)$ and $n_i \leq \epsilon m_i^*$, while the third by recalling that $q_i^* = m_i^*/p_i$. To get the last inequality, we use the assumption $q_i^* < 1/2$. Proceeding in the same manner we get the upper bound: $\mathbb{P}_S(A_i|\mathbf{v}) \leq \mathbb{P}_{\mathbf{q}^*}(A_i)(1-\epsilon)^{-n_i}(1+\epsilon)^{n_i-s_i}$ □

The idea now is to combine those partial conditional estimates for all classes and use the concentration theorem to get rid of conditioning.

Proof of Theorem 5. To prove the lemma we use again the usual strategy of conditioning. In particular, we obtain initial upper and lower bounds by utilizing Theorem 2 and conditioning. Let $Q_S(\epsilon)$ be the "error" of the concentration theorem:

$$(1-Q_S(\epsilon)) \min_{\mathbf{v}\in\mathcal{L}_\epsilon(\mathbf{m}^*)} \mathbb{P}_S(A|\mathbf{v}) \leq \mathbb{P}_S(A) \leq \mathbb{P}_S\left(\mathcal{L}_\epsilon^c(\mathbf{m}^*)\right) + \max_{\mathbf{v}\in\mathcal{L}_\epsilon(\mathbf{m}^*)} \mathbb{P}_S(A|\mathbf{v})$$

By dividing with $\mathbb{P}_{\mathbf{q}^*}(A)$, we get:

$$(1-Q_S(\epsilon)) \min_{\mathbf{v}\in\mathcal{L}_\epsilon(\mathbf{m}^*)} \frac{\mathbb{P}_S(A|\mathbf{v})}{\mathbb{P}_{\mathbf{q}^*}(A)} \leq \frac{\mathbb{P}_S(A)}{\mathbb{P}_{\mathbf{q}^*}(A)} \leq \frac{\mathbb{P}_S\left(\mathcal{L}_\epsilon^c(\mathbf{m}^*)\right)}{\mathbb{P}_{\mathbf{q}^*}(A)} + \max_{\mathbf{v}\in\mathcal{L}_s(\mathbf{m}^*)} \frac{\mathbb{P}_S(A|\mathbf{v})}{\mathbb{P}_{\mathbf{q}^*}(A)}$$

Due to Proposition 3 and Lemma 4, we have the following bounds for the ratios involving conditioning on the edge-profile:

$$(1-2\epsilon)^m \leq \prod_{i=1}^k \frac{\mathbb{P}_S(A_i|\mathbf{v})}{\mathbb{P}_{\mathbf{q}^*}(A_i)} \leq (1-\epsilon)^{-m}(1+\epsilon)^m \tag{41}$$

where $m = \sum n_i$ is the total number of edges. Moreover, by definition of the ϵ-local event, we get that $\frac{\mathbb{P}_S\left(\bar{\mathcal{L}}_\epsilon(\mathbf{m}^*)\right)}{\mathbb{P}_{\mathbf{q}^*}(A)} \leq \epsilon$. Now, setting $\hat{\epsilon} = \sqrt{c^*\lambda(S)}$, where $c^* = 2(1+\lambda+\sqrt{\lambda^2+2\lambda}) \approx 2$, such that $\hat{\epsilon}^2/(1+\hat{\epsilon}) - \lambda = \lambda$ and recalling that $\mu(S)\lambda = 5k\log n$, gives:

$$\begin{aligned}
\frac{\mathbb{P}_S(A)}{\mathbb{P}_{\mathbf{q}^*}(A)} &\geq \left(1-\frac{1}{n^{5k}}\right)(1-\hat{\epsilon})^m && (42)\\
\frac{\mathbb{P}_S(A)}{\mathbb{P}_{\mathbf{q}^*}(A)} &\leq \hat{\epsilon} + (1-\hat{\epsilon})^{-m}(1+\hat{\epsilon})^m && (43)
\end{aligned}$$

For the lower bound we use the inequality $(1-x)^y \geq 1-xy,\ \forall x<1$, and obtain:

$$\begin{aligned}\frac{\mathbb{P}_S(A)}{\mathbb{P}_{\mathbf{q}^*}(A)} &\geq \left(1-\frac{1}{n^{5k}}\right)(1-\hat{\epsilon}m)\left[1-2\hat{\epsilon}m\right] &(44)\\ &\geq 1-8\hat{\epsilon}m &(45)\end{aligned}$$

The last line follows by expanding the product and bounding from below each term in the sum by the smallest such term. To get the upper bound we proceed similarly. We use the inequality $(1-x)^y \geq 1-xy,\ \forall x<1$ to simplify the fraction in the denominator. Further, for $\hat{\epsilon}m \to 0$ we have that $(1+\epsilon)^m \leq 1+2\hat{\epsilon}m$ and $\frac{1}{1-\hat{\epsilon}m} \leq 1+2\hat{\epsilon}m$ for adequately large n. Consequently:

$$\begin{aligned}\frac{\mathbb{P}_S(A)}{\mathbb{P}_{\mathbf{q}^*}(A)} &\leq \hat{\epsilon}+(1+2m\hat{\epsilon})(1+2m\hat{\epsilon}) &(46)\\ &\leq 1+6\hat{\epsilon} &(47)\end{aligned}$$

Putting everything together, we obtain that $\left|\frac{\mathbb{P}_S(A)}{\mathbb{P}_{\mathbf{q}^*}(A)}-1\right| \leq 8\hat{\epsilon}m$. Finally, by assumption $\hat{\epsilon}m \leq \frac{1}{\sqrt[4]{\mu(S)}}$ and, hence, the theorem is concluded. □