A TIGHT BOUND FOR TESTING PARTITION PROPERTIES

Abstract. A *partition property* of order k asks if a graph can be partitioned into k vertex sets of prescribed sizes so that the densities between any pair of sets falls within a prescribed range. This family of properties has been extensively studied in various areas of research ranging from theoretical computer science to statistical physics. Our main result is that every partition property of order k is testable with query complexity $\texttt{poly}(k/\varepsilon)$. We thus obtain an exponential improvement (in k) over the $(1/\varepsilon)^{O(k)}$ bound obtained by Goldreich, Goldwasser and Ron in their seminal FOCS 1996 paper. We further prove that our bound is tight in the sense that it cannot be made sub-polynomial in *either* k or ε.

Besides the intrinsic interest in obtaining a tight bound for the above well studied family of properties, our improved bound has several algorithmic implications, stemming from the fact that it remains polynomial even when testing partition properties of order $k = \texttt{poly}(1/\varepsilon)$.

1. Introduction

1.1. **Background and previous results.**

Property testers are fast randomized algorithms that can distinguish between objects satisfying some predetermined property P and those that are ε-far from satisfying P. In most cases being ε-far means that an ε-proportion of the object's representation needs to be changed in order to obtain a new object satisfying P. Hence, testing for P is a relaxed version of the classical decision problem which asks to decide whether an object satisfies P. While questions of this nature have been implicitly studied for many years in various areas of mathematics, the first explicit studies with a computational motivation where conducted by Blum, Luby and Rubinfeld [BLR93] and by Rubinfeld and Sudan [RS96]. In this paper we study properties of graphs in the so called *adjacency matrix model* (which is also sometimes referred to as the *dense graph model*). This is arguably one of the most well studied models in the area of property testing. For the sake of brevity, we refer the reader to [Gol17, BY22] for more background and references on property testing and graph property testing in particular.

We now introduce the precise definitions regarding testing graph properties in the adjacency matrix model. A graph property P is simply a family of graphs closed under isomorphism. A graph G on n vertices is ε-far from P if one should add/delete at least εn^2 edges to turn G into a graph satisfying P. If G is not ε-far from P then it is ε-close to P. A tester for P is a randomized algorithm that given $\varepsilon > 0$ distinguishes with high probability (say, $2/3$) between graphs satisfying P and those that are ε-far from P. We assume the algorithm can randomly sample a set S of vertices from $V(G)$ and then ask an oracle for $G[S]$, which is the graph induced by G on S. The query complexity of a tester is the size of the set S it samples[1]. If P has a tester whose query complexity depends only on ε (and is independent of n) then P is called *testable*.

Property testing in the adjacency matrix model (as describe above) was first introduced by Goldreich, Goldwasser and Ron [GGR98] in the seminal paper which lay the groundwork for the area of combinatorial property testing. The main result of [GGR98] was that a general family of so-called *partition properties* are all testable. Roughly speaking, a partition property asks if the vertex set of an input graph can be partitioned into a certain numbers of sets, of certain sizes, so that the number of edges between these sets fall within certain ranges. The precise definition reads as follows.

[1] It is more common to measure the query complexity of a tester using the number of edge queries it makes. By a theorem of Goldreich and Trevisan [GT03, Theorem 2], these two measures of query complexity are quadratically related. Since such gaps will not matter to us in this work, we opt to work with the measure we defined above.

Definition 1.1 (Partition Properties)**.** *A partition property of order k is given by a set of parameters $\Phi := \{\alpha_i^{\mathsf{LB}}, \alpha_i^{\mathsf{UB}}\}_{i=1}^k \cup \{d_{ij}^{\mathsf{LB}}, d_{ij}^{\mathsf{UB}}\}_{i\le j=1}^k$ in $[0,1]$. The graph property $\mathcal{P}_\Phi$ is defined as the set of all n-vertex graphs G having a partition $V(G) = \{V_1, \dots, V_k\}$ satisfying the following inequalities.*

$$\lfloor \alpha_i^{\mathsf{LB}} \cdot n \rfloor \le |V_i| \le \lceil \alpha_i^{\mathsf{UB}} \cdot n \rceil \qquad \forall i \in [k]; \tag{1}$$

$$\lfloor d_{ij}^{\mathsf{LB}} \cdot n^2 \rfloor \le e(V_i, V_j) \le \lceil d_{ij}^{\mathsf{UB}} \cdot n^2 \rceil \qquad \forall 1 \le i \le j \le k. \tag{2}$$

Remark 1.2. *We used floor/ceiling in (1) and (2) just to make sure that the property is non-empty for all n. Since these technicalities are immaterial for large n, we will drop these floor/ceiling signs from this point on.*

It is easy to see that many well studied graph properties such as k-colorability, Max-Cut, Max-Bisection and Max-Clique can all be described in the above framework.

We should also point that partition properties have been extensively studied (under different names) in various other areas such as machine learning, statistical physics and network analysis. Perhaps the most notable one is the so called *Stochastic Block Model* (sometimes called *Planted Partition Model*) in which the input is a random graph G generated according to some predetermined edge densities as in Definition 1.1, and the goal is to reconstruct the partition of $V(G)$ which witnesses this fact (observe that the fundamental *hidden clique problem* is just a special case). We refer the reader to the extensive surveys [Abb18, LW19] for more background and references regarding various aspects of this problem in various areas of research.

The main result of Goldreich, Goldwasser and Ron proved in [GGR98] was that every partition property Φ of order k, is testable with query complexity $(1/\varepsilon)^{O(k)}$. Following [GGR98], partition properties have been extensively studied in various papers. For example, Alon et al. [AFdlVKK03] studied testing of Constrained Satisfaction Problems (CSPs), which, in the setting of graphs[2], are essentially equivalent to partition properties with $k = 2$ and no constraints on the sizes of V_1, V_2. Their main result was an improved bound for $k = 2$ compared to the one given by [GGR98] (e.g. for testing Max-Cut). Their result was later improved by Rudelson and Vershynin [RV07]. More general CSPs were studied by Andersson and Engebretsen [AE02], Czumaj and Sohler [CS05] and Alon and Shapira [AS02], but they all correspond to very special types of partition properties, for example, only allowing $0/1$ constrains on the edge densities $d_{ij}^{\mathsf{LB}}, d_{ij}^{\mathsf{UB}}$ and not allowing any constraints on the sizes of $V_1, \dots, V_k$ (i.e. restricting to the case where all $\alpha_i^{\mathsf{LB}} = 0$ and all $\alpha_i^{\mathsf{UB}} = 1$).

1.2. **Our new results.**

Given the above discussion it is natural to ask if the $(1/\varepsilon)^{O(k)}$ bound of [GGR98] can be improved. Our main result in this paper gives the first improved bound for testing general partition properties, showing that the dependence on k can be improved from exponential to polynomial.

Theorem 1.3. *Every partition property of order k is testable with query complexity $\texttt{poly}(k/\varepsilon)$.*

Obviously the improved bound in Theorem 1.3 manifests itself for large k. For example, by Theorem 1.3, even partition properties of order $k = \texttt{poly}(1/\varepsilon)$ are testable with query complexity $\texttt{poly}(1/\varepsilon)$, which is an exponential improvement over the $2^{\texttt{poly}(1/\varepsilon)}$ bound supplied by [GGR98] for this setting. As we elaborate in Subsection 1.3, there are some important applications of Theorem 1.3 that call for testing partition properties of order k that depends on ε.

[2] The result of [AFdlVKK03] also applies to special cases of partition properties of r-uniform hypergraphs.

The above discussion naturally leads one to ask how tight is the dependence on k and ε in Theorem 1.3. While it is easy to see[3] that the dependence on ε cannot be made sub-polynomial, determining the dependence on k is more subtle. Although it might seem "obvious" that a tester of every (non-trivial) partition property of order k should have query complexity $\Omega(k)$, there are actually natural families of partition properties whose query complexity is only $\mathtt{poly}(\varepsilon^{-1}, \log k)$. Indeed, a recent result of Fiat and Ron [FR21] states that every partition property of order k in which all edge density parameters $d_{ij}^{\mathsf{LB}}, d_{ij}^{\mathsf{UB}}$ are 0/1 and where all $\alpha_i^{\mathsf{LB}} = 0$ and all $\alpha_i^{\mathsf{UB}} = 1$ (namely, where there is no restriction on the sizes of the sets V_i) is testable with query complexity $\mathtt{poly}(\varepsilon^{-1}, \log k)$. Having such an efficient bound for all partition properties would have been very useful for the study of graph estimation algorithms, as discussed in Subsection 1.3. Our next theorem rules out such an extension of the result of [FR21] to arbitrary partition properties, thus showing that in Theorem 1.3 the dependence on k should indeed be polynomial.

Theorem 1.4. *For every k there is a partition property $\mathcal{P}_\Phi$ of order k so that every 0.0001-tester of $\mathcal{P}_\Phi$ has query complexity $\Omega(\sqrt{k})$.*

While the above discussion revolved around algorithmic statements, we can deduce from the tools we develop in this paper a purely combinatorial/probabilitic statement which might be of independent interest. Here, and in what follows, an *equipartition* of a graph into sets $V_1, \dots, V_k$ is a partition where $||V_i| - |V_j|| \le 1$ for all i, j. We also use $d(A, B)$ to denote the *edge density* between two disjoint vertex sets A, B, that is, the number of edges connecting A, B divided by $|A| \cdot |B|$. What the next theorem says is that a small sample of vertices Q from a graph G has the following property with high probability: G has an equipartition with a prescribed set of edge densities *if and only if* $G[Q]$ has such an equipartition (up to a small error).

Theorem 1.5. *For every $\varepsilon > 0$, $\delta > 0$, and every integers $k > 0$, there is an integer $q = q_{1.5}(\varepsilon, k, \delta) = \mathtt{poly}(\varepsilon, k, \log \frac{1}{\delta}) > 0$ satisfying the following. Let Q be a set of q vertices taken uniformly at random from a graph G of size $n = |V(G)| \ge q$. Then, with probability at least $1 - \delta$, the following holds.*

(1) For every equipartition $\{Q_i\}_{i=1}^k$ of $G[Q]$, there is an equipartition $\{V_i\}_{i=1}^k$ of G satisfying

$$d(V_i, V_j) = d(Q_i, Q_j) \pm \varepsilon \tag{3}$$

for every $1 \le i \le j \le k$.

(2) For every equipartition $\{V_i\}_{i=1}^k$ of G, there is an equipartition $\{Q_i\}_{i=1}^k$ of $G[Q]$ satisfying

$$d(Q_i, Q_j) = d(V_i, V_j) \pm \varepsilon \tag{4}$$

for every $1 \le i \le j \le k$.

1.3. **Applications of Theorem 1.3.** Besides the intrinsic interest in obtaining a tight bound for testing the family of partition properties, our improved bound in Theorem 1.3 has some important algorithmic applications, stemming from the fact that our bound is polynomial in k. A property is *estimable* (or *tolerantly testable*) if for every $\varepsilon > 0$ there is an algorithm with constant query complexity (the constant may depend on ε) that approximates an object's distance to satisfying P within ε. This notion, which is at least as strong as testability, was introduced by Parnas, Ron and Rubinfeld [PRR06], and has been extensively studied in various settings. The most important question in this area asks about the relation between the hardness of testing a property and the hardness of estimating it. One of the central results in the area of graph property testing is the Fischer-Newman theorem [FN07] which states that every testable graph property P is also estimable.

[3] Indeed, consider the property of having no edges (i.e. being the empty graph), which is a partition property of order $k = 1$. Then one clearly needs a sample of size $\Omega(1/\varepsilon)$ in order to distinguish between the empty graph (which belongs to the property) and the complete bipartite graph with parts of sizes εn and $(1 - \varepsilon)n$ (which is ε-far from the property).

The proof in [FN07] relied on variants of Szemerédi's regularity lemma [Sze78, AFKS00] and thus supplied a very weak (tower-type) transformation from testing to estimating P. Several recent works [FR21, GKS23, HKL+20, HKL+21] studied the very natural problem of designing a more efficient transformation with a polynomial loss in the query complexity. All the above works relied on reducing the task of estimating G's distance to satisfying P, to the task of testing if G satisfies certain partition properties. Now, the key point is that in the above papers [FR21, GKS23, HKL+20, HKL+21], the partition properties are such that their order k depends on ε. For example, the partitions used in [GKS23, HKL+20, HKL+21] are of order $k = 2^{\mathtt{poly}(1/\varepsilon)}$, and so the results obtained in all these papers rely crucially on the $\mathtt{poly}(k/\varepsilon)$ bound given by Theorem 1.3 (using the $(1/\varepsilon)^{O(k)}$ bound of [GGR98] would have resulted in an exponential loss in the above papers). We expect that in addition to [GKS23, HKL+20, HKL+21], our main result will be instrumental in future studies related to testing and estimation of graph properties.

1.4. **Proof and paper overview.** The proof of Theorem 1.3 has two main steps. The first one, given by Lemma 3.2, deals entirely with collections of vertex sets, that is, it has nothing to do with graphs/edge-sets. Given a collection of vertex sets $S_1, \dots, S_t \subseteq V$, it is natural to look at all $t \times k$ intersection sizes between $S_1, \dots, S_t$ and some partition of V into k sets (relative to the size of V). What Lemma 3.2 states is that up to a small error, the possible intersection sizes one can obtain by partitioning V, is the same as those achievable by partitioning a small randomly selected set of vertices from V. The proof of this lemma relies on casting the problem as as a linear optimization problem and using the hyperplane separation theorem (i.e. Farkas's Lemma) to show that if V has no partition with certain parameters, then there is a *single* linear inequality witnessing this fact. One can then use a large deviation inequality to show that this linear inequality is also not satisfied by the sample, and therefore it also lacks a partition with the same parameters. The second main step in the proof, given by Lemma 3.6, states that in every graph, there is a small collection of sets $S_1, \dots, S_t$, so that for every A, B, knowing[4] the intersection sizes of A, B with $S_1, \dots, S_t$ determines, up to a small error, the number of edges between A, B. The proof of this lemma relies on the so called *Weak Regularity Lemma* of Frieze and Kannan [FK99].

We should stress that here we rely on the *Matrix Form* of the Frieze–Kannan lemma, which involves a decomposition of a *matrix* into $\mathtt{poly}(1/\varepsilon)$ many matrices. One can very easily prove variants of the lemmas we prove here from the *Graph Form* of the Frieze–Kannan lemma (see [FK96]), but this version of the lemma has bounds that are exponential in ε, so they fall short of giving Theorem 1.3. In a nutshell, the graph form of the Frieze–Kannan lemma follows from the matrix form by taking the Venn diagram of the sets in the matrix decomposition (see [FK99]). The reason why working with the graph partition is much easier is that the sets in the partition are disjoint. What Lemma 3.2 thus enables us is to overcome the difficulties that arise when working with the overlapping sets of the matrix decomposition. The proof of Theorem 1.3 is given in Section 3. In this section we also prove Theorem 1.5.

Finally, the proof of Theorem 1.4 is given in Section 4. The partition property that satisfies the assertion of Theorem 1.4 is that of being a blowup[5] of a graph K on k vertices. The graph K is chosen randomly thus making sure that it has certain pseudo-random properties that enable the deduction of Theorem 1.4 from Yao's min-max principle [Yao77]. The main (implicit) idea is to encode the task of testing graph isomorphism into the task of testing if a graph is a blowup of K.

[4] Goldreich asked in Subsection 8.3.2 of [Gol17] *"why are general partition properties easily testable ?"*. We think that our new proof of the result of [GGR98] supplies a very succinct answer. The main structural explanation is that in every graph G there are $t = \mathtt{poly}(1/\varepsilon)$ sets $S_1, \dots, S_t$ so that the densities between the sets of every partition $V_1, \dots, V_k$ of $V(G)$ are determined *solely* by the pairwise intersection sizes of $V_1, \dots, V_k$ and $S_1, \dots, S_t$ (this is Lemma 3.6 described above). Given this, one only needs Lemma 3.2 (described above) in order to turn this structural explanation into a $\mathtt{poly}(k/\varepsilon)$ sampling algorithm.

[5] A blowup of a k-vertex graph K is the n vertex graph obtained by replacing every vertex of K with a set of n/k vertices, and every edge with a complete bipartite graph.

2. Preliminaries

In this section we state three results that we will use in Sections 3 and 4. Given a graph $G = (V, E)$, whenever there is no risk of confusion, we will also use G to denote the adjacency matrix of G — i.e. the matrix $M \in \{0, 1\}^{V \times V}$, where $M(x, y) = 1$ if $xy \in E$ and $M(x, y) = 0$ otherwise.

Given two sets (not necessarily disjoint) $S, T \subseteq V$ and a symmetric matrix $M \in \mathbb{R}^{V \times V}$, we write $M(S, T) = \sum_{x \in S, y \in T} M(x, y)$. For any given (not necessarily disjoint) subsets $S, T \subseteq V$, we denote by $K_{S,T}$ the matrix $M \in \mathbb{R}^{V \times V}$ for which $M(x, y) = 1$ if $x \in S$ and $y \in T$ and $M(x, y) = 0$ otherwise. We also denote the matrix $K_{S,S}$ simply by K_S.

The (normalized) cut norm of a matrix $M \in \mathbb{R}^{V \times V}$ is defined as

$$\|M\|_\square = \frac{1}{|V|^2} \max_{S,T \subseteq V} |M(S, T)|.$$

We stress that M is allowed to have negative entries. Intuitively, a matrix of small cut norm is pseudo random. We also set $\|M\|_\infty$ as the maximum of the absolute values of the entries of M. The following is the matrix-form of the Frieze-Kannan Weak Regularity Lemma. It states that every matrix can be written as a sum of few simple (rank 1) matrices and a pseudo random matrix.

Lemma 2.1 (Matrix Decomposition [FK99, Theorem 1]). *For every $\varepsilon > 0$, there is an integer $T = T_{2.1}(\varepsilon) = \texttt{poly}(\varepsilon^{-1})$ for which the following holds. Given a matrix $M \in [0, 1]^{V \times V}$, there is an integer $t \leq T$, sets $X_r, Y_r \subseteq V$ and real numbers d_r (for all $r \in [t]$), such that*

$$M = \sum_{r=1}^{t} d_r \cdot K_{X_r,Y_r} + \Delta,$$

where $\Delta \in \mathbb{R}^{V \times V}$ satisfies $\|\Delta\|_\square \leq \varepsilon$, and $\sum_{r=1}^{t} d_r^2 \leq 1$, □

The next result states that if the cut norm of a matrix Δ is small, then so is the cut norm of the submatrix induced by a randomly selected small subset of the rows/columns of Δ.

Lemma 2.2 ([BCL+08, Theorem 2.10]). *For every $\gamma > 0$, there is $q = q_{2.2}(\gamma, \delta) = \texttt{poly}(\gamma^{-1}, \delta^{-1})$ for which the following holds. Let $\Delta \in \mathbb{R}^{V \times V}$ be a matrix satisfying $\|\Delta\|_\square \leq \frac{1}{2}\gamma$ and $\|\Delta\|_\infty = O(\gamma^{-1})$. If $Q \in \binom{V}{q}$ is a set of q vertices chosen uniformly at random from V, then*

$$\|\Delta|_{Q \times Q}\|_\square \leq \gamma,$$

with probability at least $1 - \delta$. □

The following concentration result is a consequence of Azuma's inequality for martingales.

Lemma 2.3 (McDiarmid's inequality [McD89, Lemma 1.2]). *Let $Z_1, \ldots, Z_k$ be independent random variables, where each Z_i takes values in some finite Ω_i. Suppose there is $C > 0$ and a function $f : \Omega_1 \times \cdots \times \Omega_k \to \mathbb{R}$ for which $|f(x) - f(y)| \leq C$ whenever $x = (x_1, \ldots, x_k)$ and $y = (y_1, \ldots, y_k)$ differ only in one coordinate. Then, for any $\lambda > 0$,*

$$\mathbb{P}(|f(Z) - \mathbb{E}(f(Z))| \geq \lambda) \leq 2 \exp\{-\lambda^2 / 2kC^2\}.$$ □

3. Proof of Theorem 1.3

Most of this section is devoted to proving the following key lemma, which relates partitions of a graph G and the partitions of a typical sample of G.

Lemma 3.1. *For every $\zeta > 0$, $\delta > 0$ and every integer $k > 0$, there is $q = q_{3.1}(\zeta, k, \delta) = \mathtt{poly}(\zeta, k, \log \frac{1}{\delta})$ satisfying the following. Let $Q \in_U \binom{V}{q}$ be a set of q vertices taken uniformly at random from a graph G of size $n = |V(G)| \geq q$. Then, with probability at least $1 - \delta$, for* every *partition $\{Q_i\}_{i=1}^k$ of $G[Q]$, there is a partition $\{V_i\}_{i=1}^k$ of G satisfying*

(1) $\frac{1}{n}|V_i| = \frac{1}{q}|Q_i| \pm \zeta$ *for every* $1 \leq i \leq k$;

(2) $\frac{1}{n^2}\,\mathrm{e}(V_i, V_j) = \frac{1}{q^2}\,\mathrm{e}(Q_i, Q_j) \pm \zeta$ *for every* $1 \leq i \leq j \leq k$.

Moreover, for every *partition $\{V_i\}_{i=1}^k$ of G, there is a partition $\{Q_i\}_{i=1}^k$ of $G[Q]$ satisfying the two constraints above.*

Let us first deduce Theorem 1.3 from the above Lemma 3.1.

Proof of Theorem 1.3. Let $\zeta = \frac{1}{4}\varepsilon/k^2$ and consider a tester T that, given an input graph $G = (V, E)$, with $|V| = n$, works as follows. The tester T samples a set $Q \in_U \binom{V}{q}$ of size $q = \max\{q_{3.1}(\zeta, k, \frac{1}{3}), 8k^2/\zeta\} = \mathtt{poly}(\varepsilon^{-1}, k)$ and accepts if and only if $G[Q]$ has a partition $\{Q_i\}_{i=1}^k$ satisfying the following inequalities.

$$\alpha_i^{\mathsf{LB}} - \zeta \leq \frac{1}{q}|Q_i| \leq \alpha_i^{\mathsf{UB}} + \zeta \qquad \forall 1 \leq i \leq k; \tag{5}$$

$$d_{ij}^{\mathsf{LB}} - \zeta \leq \frac{1}{n^2}\,\mathrm{e}(Q_i, Q_j) \leq d_{ij}^{\mathsf{UB}} + \zeta. \qquad \forall 1 \leq i \leq j \leq k. \tag{6}$$

Hereafter, we will assume the assertions of Lemma 3.1 holds, which happens with probability at least $\frac{2}{3}$.

First it is straightforward to see that if $G \in \mathcal{P}_\Phi$, then G is accepted. Indeed, it follows from the very definition of $\mathcal{P}_\Phi$ combined with Lemma 3.1, that $G[Q]$ must satisfy ineq. (5) and (6).

Next, assume G is ε-far from $\mathcal{P}_\Phi$. Suppose, by contradiction, that T accepts G. Then, by Lemma 3.1 and ineq. (5) and (6), G must have a partition $\{V_i'\}_{i=1}^k$ satisfying

$$\alpha_i^{\mathsf{LB}} - 2\zeta \leq \frac{1}{n}|V_i'| \leq \alpha_i^{\mathsf{UB}} + 2\zeta \qquad \forall i \in [k];$$

$$d_{ij}^{\mathsf{LB}} - 2\zeta \leq \frac{1}{n^2}\,\mathrm{e}(V_i', V_j') \leq d_{ij}^{\mathsf{UB}} + 2\zeta. \qquad \forall 1 \leq i \leq j \leq k.$$

We start by (arbitrarily) rearranging up to ζn vertices from each V_i' ($\forall i$) in order to get a partition $\{V_i\}_{i=1}^k$ satisfying ineq. (1). Since this rearranging can change each $e(V_i', V_j')$ by an additive factor of at most $(2\zeta n)n$, the partition $\{V_i\}_{i=1}^k$ now satisfies

$$d_{ij}^{\mathsf{LB}} - 4\zeta \leq \frac{1}{n^2}\,\mathrm{e}(V_i, V_j) \leq d_{ij}^{\mathsf{UB}} + 4\zeta. \qquad \forall 1 \leq i \leq j \leq k.$$

Hence, by adding/deleting up to $4\zeta n^2$ edges between each pair (V_i, V_j) ($\forall i, j$), we get a graph that also satisfies ineq. (2) and, therefore, is in $\mathcal{P}_\Phi$. Our choice of ζ then implies that G is not ε-far from $\mathcal{P}_\Phi$, which contradicts the hypothesis that G was ε-far from $\mathcal{P}_\Phi$. Therefore, T rejects graphs ε-far from $\mathcal{P}_\Phi$ (with probability at least $\frac{2}{3}$). □

We next show how to deduce Theorem 1.5 from Lemma 3.1.

Proof of Theorem 1.5. Set $\zeta = \min\{\varepsilon/(2k^2), \varepsilon/(8k)\}$ and $q = q_{3.1}(\zeta, k, \delta)$. Let $Q \in \binom{V}{q}$ be a set of q vertices chosen uniformly at random.

We will first prove item (1) from Theorem 1.5. By Lemma 3.1, with probability at least $1 - \delta$, for every *equipartition* $\{Q_i\}_{i=1}^k$ of $G[Q]$ there is a partition $\{W_i\}_{i=1}^k$ of G satisfying

(1) $\frac{1}{n}|W_i| = \frac{1}{q}|Q_i| \pm \varepsilon/8k = \frac{1}{k} \pm \varepsilon/8k.$

(2) $\frac{1}{n^2}\,\mathrm{e}(W_i, W_j) = \frac{1}{q^2}\,\mathrm{e}(Q_i, Q_j) \pm \varepsilon/2k^2$.

Let $\{V_i\}_{i=1}^k$ be an equipartition of G obtained from $\{W_i\}_{i=1}^k$ after arbitrarily redistributing up to $\frac{1}{8}\varepsilon(n/k)$ vertices from each class W_i. Since,

$$|\mathrm{e}(V_i, V_j) - \mathrm{e}(W_i, W_j)| \le \frac{\varepsilon n}{8k}\max\{|V_i|, |W_i|\} + \frac{\varepsilon n}{8k}\max\{|V_j|, |W_j|\} \le \frac{\varepsilon n}{4k}(\frac{n}{k} + \frac{\varepsilon n}{8k}) \le \frac{\varepsilon n^2}{2k^2},$$

we must have

$$d(V_i, V_j) = \frac{k^2\,\mathrm{e}(V_i, V_j)}{n^2} = \frac{k^2\,\mathrm{e}(W_i, W_j)}{n^2} \pm \frac{1}{2}\varepsilon = \frac{k^2\,\mathrm{e}(Q_i, Q_j)}{n^2} \pm k^2\left(\frac{1}{2}\varepsilon/k^2\right) \pm \frac{1}{2}\varepsilon = d(Q_i, Q_j) \pm \varepsilon,$$

as required by assertion (1).

The argument for assertion (2) of the lemma is symmetric. Indeed, by Lemma 3.1, for every *equipartition* $\{V_i\}_{i=1}^k$, there is an partition $\{Q_i\}_{i=1}^k$ of $G[Q]$ satisfying

(1) $\frac{1}{q}|W_i| = \frac{1}{n}|V_i| \pm \varepsilon/8k = \frac{1}{k} \pm \varepsilon/8k$.

(2) $\frac{1}{q^2}\,\mathrm{e}(W_i, W_j) = \frac{1}{n^2}\,\mathrm{e}(V_i, V_j) \pm \varepsilon/2k^2$.

By the first item above, one can get an equipartition $\{Q_i\}$ of $G[Q]$ by arbitrarily redistributing up to $\frac{1}{8}\varepsilon(q/k)$ vertices from each class W_i; similarly as before, we have that $|\mathrm{e}(Q_i, Q_j) - \mathrm{e}(W_i, W_j)| \le \frac{\varepsilon q^2}{2k^2}$. Hence,

$$d(Q_i, Q_j) = \frac{k^2\,\mathrm{e}(Q_i, Q_j)}{q^2} = \frac{k^2\,\mathrm{e}(W_i, W_j)}{n^2} \pm \frac{1}{2}\varepsilon = \frac{k^2\,\mathrm{e}(V_i, V_j)}{n^2} \pm k^2\left(\frac{1}{2}\varepsilon/k^2\right) \pm \frac{1}{2}\varepsilon = d(V_i, V_j) \pm \varepsilon,$$

as required by assertion (2). □

The remainder of this section is devoted to proving Lemma 3.1. Let V be a ground set of size n and let $\{S_r\}_{r=1}^t$ be a family of subsets of V. For any partition $\{U_i\}_{i=1}^k$ of a subset $U \subseteq V$, we refer to the following collection of intersection (relative) sizes $\{\alpha_r^{(i)}\}_{r\in[t]}^{i\in[k]}$, where $\alpha_r^{(i)} = |U_i \cap S_r|/|U|$, as the *intersection profile* of $\{U_i\}_{i=1}^k$ on $\{S_r\}_{r=1}^t$.

Our strategy to prove Lemma 3.1 consists of basically two parts. First, we show in Section 3.1 that given V and $\{S_r\}_{r=1}^t$ as above and a typical sample (of constant size) Q of V, there is a correspondence between the partitions of V and the partitions of Q, which essentially preserves the intersection profile on $\{S_r\}_{r=1}^t$. Then, we show that, given a graph $G = (V, E)$, there is a specific family $\{S_r\}_{r=1}^t$ on which the intersection profile is revealing, in the sense that, just from knowing the values of $\{\alpha_r^{(i)}\}_{r\in[t]}^{i\in[k]}$, one can provide a good estimate for (up to an additive error) for the number of edges $\mathrm{e}(A, B)$ between *any* two sets $A, B \subseteq V$.

3.1. **Intersection profile of a sample.** Our main goal in this subsection is to prove Lemma 3.2 below. We stress that this entire subsection deals solely with a vertex set V and a collection $S_1, \ldots, S_t$ of subsets of V, and it has nothing to do with edge sets of graphs.

Lemma 3.2. *For every $\eta > 0$ and any positive integers k, t and a, there is $q = q_{3.2}(\eta, k, t) \le 32\lceil a/\eta^2 + 2kt/(\eta^3)\rceil$ satisfying the following. Let $S_1, \ldots, S_t$ be subsets of vertices of set V and let Q be a subset of V chosen uniformly at random. Then, with probability at least $1 - e^{-a}$, for every partition $\{Q_i\}_{i=1}^k$ of Q, there is a partition $\{V_i\}_{i=1}^k$ of V such that*

$$\frac{1}{n}|V_i \cap S_r| = \frac{1}{q}|Q_i \cap S_r| \pm \eta \tag{7}$$

The following key lemma guarantees that if V has no partition with an intersection profile on $\{S_r\}_{r=1}^t$ close to some given values $\{\alpha_r^{(i)}\}$, then a typical sample $Q \subseteq V$ has no such partition as well.

Lemma 3.3. *Let t and k be positive integers and $\eta > 0$. Let $\{S_r\}_{r=1}^t$ be sets of vertices of a ground set V of size $|V| \geq 2tk/(\eta^2)$. Suppose there are numbers $\{\alpha_r^{(i)}\}_{i\in[k],r\in[t]}$ in $[0,1]$ for which* ***no*** *partition $\{V_i\}_{i=1}^k$ of V satisfies*

$$\forall r \in [t], \forall i \in [k]\colon \frac{|S_r \cap V_i|}{n} = \alpha_r^{(i)} \pm 3\eta. \tag{8}$$

Then, with probability at least $1 - e^{-q\eta^2/2}$, a set $Q \in \binom{V}{q}$ of q elements chosen uniformly at random has ***no*** *partition $\{Q_i\}_{i=1}^k$ satisfying*

$$\forall r \in [t], \forall i \in [k]\colon \frac{|S_r \cap Q_i|}{q} = \alpha_r^{(i)} \pm \eta \tag{9}$$

Proof. The general strategy to prove this lemma will be as follows: we first write the constraints given by Eq. (8) (regarding the intersection profile of $\{V_i\}_{i=1}^k$) as a system of linear inequalities, which, by hypothesis, has no solution. Hence, by a variant of Farka's lemma, we show there is a linear combination of such inequalities that witnesses the infeasibility of the system. We then show that a typical sample "inherits" a similar linear combination, which by its turn witnesses the infeasability of a system of inequalities related to (9).

Suppose V has no partition satisfying Eq. (8) and consider the following system $\mathcal{S}_V$ of linear inequalities (on nk variables $x_j^{(i)}$, $i \in [k]$, $j \in [n]$).

System of inequalities $\mathcal{S}_V$

$$\forall r \in [t], \forall i \in [k]\colon \qquad (\alpha_r^{(i)} - 2\eta)n \leq \sum_{j\in S_r} x_j^{(i)} \leq (\alpha_r^{(i)} + 2\eta)n \tag{$I_{r,i}$}$$

$$\forall j \in [n]\colon \qquad x_j^{(1)} \geq 0, \dots, x_j^{(k)} \geq 0, \sum_{i=1}^k x_j^{(i)} = 1 \tag{P_j}$$

Note: *we named the inequalities related with the* ***i****ntersection constraints as $(I_{r,i})$ and the ones related with $\{x_j^{(i)}\}$ specifying a* ***p****artition as (P_j).*

Claim 3.4. *The system $\mathcal{S}_V$ has no solution.*

Proof. Indeed, suppose there was a solution $\{x_j^{(i)}\}_{j\in[n]}^{i\in[k]}$ for $\mathcal{S}_V$. Then, consider a partition $\{V_i\}_{i=1}^k$ of V obtained by setting, independently for every vertex $j \in [n]$, the part V_i that contains j according to the distribution given by $\{x_j^{(i)}\}_{i\in[k]}$.

Let $r \in [t]$ and $i \in [k]$ be fixed. The cardinality $|S_r \cap V_i|$ can be written as the sum $\sum_{j\in S_r} \mathbb{1}_{j\in V_i}$. Since

$$\mathbb{E}(|S_r \cap V_i|) = \sum_{j\in S_r} \mathbb{E}(\mathbb{1}_{j\in V_i}) = \sum_{j\in S_r} x_j^{(i)} = n(\alpha_r^{(i)} \pm 2\eta),$$

and the indicator variables $\mathbb{1}_{j\in V_i}$ are all independent, we get by Lemma 2.3 that

$$\mathbb{P}(||S_r \cap V_i| - n\alpha_r^{(i)}| > 3\eta n) \leq 2\exp\left\{-\frac{\eta^2 n^2}{2|S_r|}\right\} \leq 2e^{-tk}.$$

Therefore, the probability that $\{V_i\}_{i=1}^k$ violates Eq. (8) for *some* $r \in [t]$ and $i \in [k]$ is at most $tk2e^{-tk} < 1$. Hence, we proved that with positive probability, the partition $\{V_i\}_{i=1}^k$ satisfies Eq. (8) for every $r \in [t]$ and $i \in [k]$, which contradicts the hypothesis of the lemma. This completes the proof of the claim. ■

We continue with the proof of the lemma. Let $Q \in_U \binom{V}{q}$ be a set of q vertices chosen uniformly at random over V. Consider the following system of inequalities on qk variables $\{y_j^{(i)}\}_{j\in Q}^{i\in[k]}$.

System of inequalities $\mathcal{S}_Q$

$$\forall r \in [t], \forall i \in [k]\colon \qquad (\alpha_r^{(i)} - \eta)q \le \sum_{j\in S_r} y_j^{(i)} \le (\alpha_r^{(i)} + \eta)q \qquad (I'_{r,i})$$

$$\forall j \in [q]\colon \qquad y_j^{(1)} \ge 0, \dots, y_j^{(k)} \ge 0, \sum_{i=1}^{k} y_j^{(i)} = 1 \qquad (P_j)$$

Notice that if Q has an equipartition $\{Q_i\}_{i=1}^k$ satisfying Eq. (9), then one can produce a solution to $\mathcal{S}_Q$ by setting $y_j^{(i)} = 1$ if $j \in Q_i$, and $y_j^{(i)} = 0$ otherwise. Hence, in order to prove the lemma, it suffices to upper bound the probability that $\mathcal{S}_Q$ has a solution by $e^{-q\eta^2/2}$.

Next, we will show that if $\mathcal{S}_V$ has no solution, then there is a linear combination of the constraints $\bigcup I_{r,i}$ for which there is no solution x satisfying $\bigcup P_j$. More formally, let $A \in \mathbb{R}^{2tk\times nk}$ and $b \in \mathbb{R}^{2tk}$ be such that the constraints $\bigcup I_{r,i}$ can all be writen as $Ax \le b$. In this context, we will denote by $L(r,i)$ and $U(r,i)$ the indices of $[2tk]$ associated, respectively, with the lower and the upper bound of $I_{r,i}$. Moreover we also set $P = \{x \in \mathbb{R}^{nk} : x \text{ satisfies } \bigcup_{j\in[n]} P_j\}$ as the set of all x satisfying every partition constraint.

Claim 3.5. *If $\mathcal{S}_V$ has no solution, then there is a vector $w \in \mathbb{R}^{2tk}$, with $w \ge 0$, such that if $x \in P$, then $(w^T A)x > w^T b$;*

Proof. In what follows, for every $\ell \in [2tk]$ we denote by e_ℓ the vector with entries ℓ equal to 1 and all other entries equal to 0. If $\mathcal{S}_V$ has no solution, then the sets $\{Ax : x \in P\} \subset \mathbb{R}^{2tk}$ and $\{b' : b' \le b\} \subset \mathbb{R}^{2tk}$ must have no intersection. Since the first set is compact, for it is the image of a compact set under a linear transformation, and the second set is closed there must be a hyperplane that *strictly* separates them (see e.g. [BT97, Hyperplane Separation Theorem]). In other words, there must be $w \in \mathbb{R}^{2tk}$ such that,

$$w^T(Ax) > w^T b \ge w^T b', \qquad \text{for every } x \in P \text{ and } b' \le b. \tag{10}$$

Since, for every $\ell \in [2tk]$, the second inequality above must hold for $b' := b - e_\ell \le b$, it follows that $w^T b \ge w^T b' = w^T b - u_\ell$, that is, $w_\ell \ge 0$. Thus, we proved that there is $w \ge 0$ for which $(w^T A)x > w^T b$ for every $x \in P$. This completes the proof of the claim. ■

We are now ready to complete the proof of the lemma. Since $\mathcal{S}_V$ has no solution, there is $w \in \mathbb{R}^{2tk}$ as in Claim 3.5. It is not difficult to see that one can assume, without loss of generality, that for every $(r,i) \in [t]\times[k]$ either $w_\ell = 0$ or $w_{\ell'} = 0$, where $\ell = L(r,i)$ and $\ell' = U(r,i)$. Thus, in order to make the notation easier, let $w_r^{(i)} := w_{\ell'}$, if $w_\ell = 0$; or $w_r^{(i)} = -w_\ell$ if $w_{\ell'} = 0$, where $\ell = L(r,i)$ and $\ell' = U(r,i)$. Let

$$\lambda_j^{(i)} := \sum_{r: j\in S_r} w_r^{(i)}$$

for every $i \in [k]$ and $j \in [n]$ and let

$$\lambda_0 := w^T b = \sum_{r=1}^{t}\sum_{i=1}^{k} (w_r^{(i)}\alpha_r^{(i)} + 2|w_r^{(i)}|\eta)n$$

It follows from Claim 3.5 that the inequality

$$(w^T A)x = \sum_{j=1}^{n}\sum_{i=1}^{k} \lambda_j^{(i)} x_j^{(i)} \le \lambda_0 = w^T b \quad \text{has no solution satisfying } x \in P.$$

But this happens if and only if

$$\sum_{j=1}^{n} \min_{i\in[k]} \lambda_j^{(i)} > \lambda_0, \tag{11}$$

since $\sum_{j=1}^{n}\sum_{i=1}^{k} \lambda_j^{(i)} x_j^{(i)}$ is minimized over $\{x_j^{(i)}\} \in P$ by simply setting $(\forall j \in [n])\ x_j^{(i)} = 1$ for some $i = \operatorname{argmin}_{i\in[k]} \lambda_j^{(i)}$ and $x_j^{(i)} = 0$ otherwise. We have thus arrived at a single equation (Eq. (11)) that witnesses $\mathcal{S}_V$ has no solution.

Now, suppose there is a solution $\{y_j^{(i)}\}$ to $\mathcal{S}_Q$. By taking a $w_r^{(i)}$-linear combination of the inequalities $I'_{r,i}$, we get

$$\begin{aligned}
\sum_{i\in Q}\sum_{i=1}^{k} \lambda_j^{(i)} y_j^{(i)} &\le \sum_{r=1}^{t}\sum_{i=1}^{k} (w_r^{(i)}\alpha_r^{(i)} + |w_r^{(i)}|\eta)q \\
&= \sum_{r=1}^{t}\sum_{i=1}^{k} (w_r^{(i)}\alpha_r^{(i)} + 2|w_r^{(i)}|\eta - |w_r^{(i)}|\eta)q \\
&= \lambda_0 \frac{q}{n} - \sum_{r=1}^{t}\sum_{i=1}^{k} |w_r^{(i)}|\eta q.
\end{aligned}$$

By putting $M = \sum_{r=1}^{t}\sum_{i=1}^{k} |w_r^{(i)}|$, it follows that

$$\sum_{i\in Q}\sum_{i=1}^{k} \lambda_j^{(i)} y_j^{(i)} \le \lambda_0 \frac{q}{n} - M\eta q,$$

which implies, similarly as before, by the (P_j) constraints on $y_j^{(i)}$, that Q must satisfy

$$\sum_{j\in Q} \min_{i\in[k]} \lambda_j^{(i)} \le \frac{q}{n}\lambda_0 - M\eta q. \tag{12}$$

Define $Y_j = \min_{i\in[k]} \lambda_j^{(i)}$ for every $j \in [n]$. Then the expectation of $\sum_{j\in Q} Y_j$ is

$$\mathbb{E}\Big(\sum_{j\in Q} Y_j\Big) = \mathbb{E}\Big(\sum_{j=1}^{n} Y_j \mathbb{1}_{j\in Q}\Big) = \frac{q}{n}\sum_{j=1}^{n} Y_j > \frac{q}{n}\lambda_0\ ,$$

where the last inequality follows from Eq. (11). Even though Q is chosen without replacement, it is well known that the hypergeometric distribution is at least as concentrated as the respective binomial distribution, we can use Lemma 2.3, (together with the fact that $Y_j \le M$ for every $j \in Q$) to obtain that

$$\begin{aligned}
\mathbb{P}\left(\sum_{j\in Q} \min_{i\in[k]} \lambda_j^{(i)} \le \frac{q}{n}\lambda_0 - M\eta q\right) &\le \mathbb{P}\left(\sum_{j=1}^{q} Y_j \le \mathbb{E}\Big(\sum_{j\in Q} Y_j\Big) - M\eta q\right) \\
&\le \exp\left\{\frac{-2(M\eta q)^2}{q(2M)^2}\right\} \\
&= e^{-q\eta^2/2}.
\end{aligned}$$

Hence, with probability at least $1 - e^{-q\eta^2/2}$ the sample Q is unable to satisfy Eq. (12). In that case, there cannot be a solution to $\mathcal{S}_Q$, as required. □

Proof of Lemma 3.2. Consider the following collection of intersection profiles of integer multiples of $\eta/4$.

$$\mathcal{N} = \{(\alpha_r^{(i)})_{i\in[k],r\in t} : \alpha_r^{(i)} = u_r^{(i)}\eta/4, \text{ for some integer } 0 \le u_r^{(i)} \le 4/\eta \ \}.$$

We have $|\mathcal{N}| \le (1+4/\eta)^{kt} \le e^{2kt/\eta}$ For each element of $\mathcal{N}$, apply Lemma 3.3 for the family $\{S_1,\dots,S_t\}$ with error parameter $\eta/4$. With probability at least

$$1 - |\mathcal{N}|e^{-q\eta^2/32} \ge 1 - e^{2kt/\eta}\cdot e^{-a-2kt/\eta} = 1 - e^{-a}$$

over the choice of Q, the assertion of Lemma 3.3 holds simultaneously for every element of $\mathcal{N}$.

Let $\{Q_i\}_{i=1}^k$ be any partition of Q. From the very definition of $\mathcal{N}$ there must be $(\alpha_r^{(i)}) \in \mathcal{N}$ for which

$$\forall i \in [k], \forall r \in [t] : \frac{1}{q}|Q_i \cap S_r| = \alpha_r^{(i)} \pm \eta/4.$$

By Lemma 3.3, there must be a partition of $\{V_i\}_{i=1}^k$ of V such that

$$\forall i \in [k], \forall r \in [t] : \frac{1}{n}|V_i \cap S_r| = \alpha_r^{(i)} \pm 3\eta/4 = \frac{1}{q}|Q_i \cap S_r| \pm \eta,$$

as desired. □

3.2. **A family on which the intersection profile is revealing.** The final piece we need for proving Lemma 3.1 is the following lemma. It roughly asserts that for every graph G, there is a family $\{S_r\}_{r=1}^t$ of subsets of $V(G)$ for which one can estimate the number of edges $e(A,B)$ between any two sets $A, B \subseteq V(G)$ by just computing the intersections $|A \cap S_r|$ and $|B \cap S_r|$. Moreover, it asserts that a typical sample of G must *also* satisfy this statement (with respect to the *same* sets $\{S_r\}_{r=1}^t$).

Lemma 3.6. *For every $\gamma > 0$ and $\delta > 0$, there are integers $T = T_{3.6}(\gamma) = \texttt{poly}(\gamma^{-1})$ and $q = q_{3.6}(\gamma,\delta) = \texttt{poly}(\frac{1}{\gamma},\frac{1}{\delta})$ satisfying the following. For every graph $G = (V,E)$, with $|V| \ge q$, there are sets $S_1,\dots,S_t \subset V(G)$ and real numbers $d_1,\dots,d_t$ (with $t \le T$) for which*

$$e(A,B) = \sum_{i=1}^{t} d_i \cdot |A\cap S_i||B \cap S_i| \pm \gamma n^2 \tag{13}$$

for every $A, B \subseteq V$. Moreover, if $Q \in \binom{V}{q}$ is a subset of q vertices chosen uniformly at random, then with probability at least $1-\delta$,

$$e(A,B) = \sum_{i=1}^{t} d_i \cdot |A\cap S_i||B \cap S_i| \pm \gamma q^2 \tag{14}$$

for every $A, B \subseteq Q$.

Proof. By a slight abuse of notation, we will also denote by $G \in \{0,1\}^{V\times V}$ be the adjacency matrix of the graph G. We will also denote the matrix $G|_{Q\times Q}$ simply by $G[Q]$.

Let $\zeta = \frac{1}{2}\gamma$. By Lemma 2.1, there is an integer $t' = O(\varepsilon^{-4})$, sets $X_1,\dots X_{t'}$ and $Y_1,\dots Y_{t'}$ and a matrix $\Delta' \in \mathbb{R}^{V\times V}$ for which

$$G = \sum_{r=1}^{t} d'_j \cdot K_{X_j,Y_j} + \Delta',$$

with $\|\Delta'\|_\square \le \zeta$ and $|d'_j| \le 1$ for every $1 \le j \le r$. Without loss of generality (by writing multiple times the same component of the decomposition), we can and will assume $|d'_j| \le 1$. Moreover, since for any set $X, Y \subseteq V$,

$$K_{X,Y} + K_{Y,X} = K_{X\cup Y, X\cup Y} - K_{X\setminus Y, X\setminus Y} - K_{Y\setminus X, Y\setminus X} + K_{X\cap Y, X\cap Y},$$

we can write

$$G = \frac{1}{2}(G + G^T) = \sum_{i=1}^{t} d_i K_{S_i} + \Delta, \tag{15}$$

by setting $t = 4t'$, $\Delta = \frac{1}{2}(\Delta' + \Delta'^T)$ and $d_{4j} = d_{4j+3} = \frac{1}{2}d'_j$, $d_{4j+1} = d_{4j+2} = \frac{1}{2}d'_j$, $S_{4j} = X_j \cup Y_j$, $S_{4j+1} = X_j \setminus Y_j$, $S_{4j+2} = Y_j \setminus X_j$, $S_{4j+3} = X_j \cap Y_j$ (for every $1 \le j \le r$). Note that $\|\Delta\|_\Box \le \|\Delta'\|_\Box \le \zeta$.

Equation (13) follows from Eq. (15) since, for every $A, B \subseteq V$,

$$e(A,B) = G(A,B) = \sum_{i=1}^{t} d_i K_{S_i}(A,B) + \Delta(A,B) = \sum_{i=1}^{t} d_i \cdot |A \cap S_i||B \cap S_i| \pm \gamma n^2.$$

Moreover, we set $q = q_{2.2}(\gamma, \delta)$ and let $Q \in \binom{V}{q}$ be a set of q vertices chosen uniformly at random from G. By Lemma 2.2, $\|\Delta[Q]\|_\Box \le \frac{1}{2}\gamma$, with probability at least δ. Hence, Eq. (14) follows easily from

$$G[Q] = \sum_{i=1}^{t} d_i \cdot K_{S_i \cap Q} + \Delta[Q] \qquad \Box$$

3.3. **Proof of Lemma 3.1.** We are ready to deduce Lemma 3.1 from Lemma 3.2 and Lemma 3.6. Put

$$\gamma = \frac{1}{3}\zeta, \quad T = T_{3.6}(\gamma), \quad \eta = \zeta/(3\sqrt{T}), \quad q = \max\{100kT \log\frac{1}{\delta}/\eta^3, q_{3.6}(\gamma, \frac{1}{2}\delta).\}$$

and let $Q \in \binom{V}{q}$ be chosen uniformly at random as in the lemma.

We start by proving the main direction of the lemma, in which we are given a partition $\{Q_i\}_{i=1}^k$ of Q and we need to produce a partition $\{V_i\}_{i=1}^k$ of V satisfying assertions (1) and (2).

Let $\{d_r\}_{r=1}^t$ be real numbers and $\{S_r\}_{r=1}^t$ be the sets of vertices of G satisfying Eq. (13) from Lemma 3.6.

With probability at least $1 - \frac{1}{2}\delta$, the set Q, taken uniformly at random, satisfies Eq. (14) from Lemma 3.6. With probability at least $1 - \frac{1}{2}\delta - \frac{1}{2}\delta = 1 - \delta$, Q also satisfies the assertion of Lemma 3.2 with respect to the following sets of vertices: $\{S_r\}_{i=1}^t \cup \{V(G)\}$. In particular, since $\{Q_i\}_{i=1}^k$ of Q, there exists a partition $\{V_i\}_{i=1}^k$ of G for which

$$\frac{|V_i|}{n} = \frac{|Q_i|}{q} \pm \eta = \frac{|Q_i|}{q} \pm \zeta, \tag{16}$$

for every $1 \le i \le k$ — which already sets assertion (1) from the lemma — and

$$\frac{|V_i \cap S_r|}{n} = \frac{|Q_i \cap S_r|}{q} \pm \eta, \tag{17}$$

for every $1 \le r \le t$ and $1 \le i \le k$.

For every $1 \le i, j \le k$, we have

$$\begin{aligned}\frac{1}{n^2}e(V_i, V_j) \pm \gamma &\overset{(13)}{=} \sum_{r=1}^{t} d_r \cdot \frac{|S_r \cap V_i|}{n} \cdot \frac{|S_r \cap V_j|}{n} \\ &\overset{(17)}{=} \sum_{r=1}^{t} d_r \cdot \frac{|S_r \cap Q_i| \pm \eta q}{q} \cdot \frac{|S_r \cap Q_j| \pm \eta q}{q}\end{aligned}$$

$$= \frac{1}{q^2}\sum_{r=1}^{t}\left(d_r \cdot |S_r \cap Q_i| \cdot |S_r \cap Q_j| \pm 3|d_r|\eta q^2\right)$$

$$\overset{(14)}{=} \frac{1}{q^2} e(Q_i,Q_j) \pm \left(\gamma + 3\eta\sum_{r=1}^{t}|d_r|\right)$$

Recall that Lemma 2.1 asserts that $\sum_{r=1}^{t} d_r^2 \leq 1$. Applying Cauchy-Schwarz, we get that

$$\frac{1}{n^2}\, e(V_i,V_j) = \frac{1}{q^2}\, e(Q_i,Q_j) \pm (2\gamma + 3\eta\sqrt{t}) = \frac{1}{q^2}\, e(Q_i,Q_j) \pm 3\gamma.$$

Hence, by our choice of γ, we get that $\frac{1}{n^2}e(V_i,V_j) = \frac{1}{q^2}e(Q_i,Q_j) \pm \zeta$, which concludes the proof of the main direction of assertion (2).

The "moreover" part of the lemma follows from a standard application of concentration results. Indeed, given a partition $\{V_i\}_{i=1}^{k}$ of G, one can simply set $Q_i = V_i \cap Q$ for every $i \in [k]$. Since, for every $i \in [k]$, $\mathbb{E}(|Q_i|/q) = |V_i|/n$ and the choice of each vertex of Q can change $|Q_i|/q$ by at most $1/q$, we get from Lemma 2.3 that

$$\mathbb{P}(|Q_i|/q - |V_i|/n| \geq \zeta) \leq 2\exp\{-(\zeta q)^2/2\}. \tag{18}$$

Moreover, for every $i,j \in [k]$, since $\mathbb{E}(e(Q_i,Q_j)/q^2) = \mathbb{E}(e(V_i,V_j)/n^2)$ and the choice of each vertex of Q can change $e(Q_i,Q_j)/q^2$ by at most $1/q$, we get from Lemma 2.3 that

$$\mathbb{P}(|\, e(Q_i,Q_j)/q^2 - e(V_i,V_j)|/n^2| \geq \zeta) \leq 2\exp\{-(\zeta q)^2/2\}. \tag{19}$$

Note that in both these applications of Lemma 2.3 we used the fact that the respective distribution are at least as concentrated as if Q was a multiset chosen with replacement. By equations (18) and (19) above, we can upper bound the probability that *any* of the assertions (1) and (2) are not satisfied by $4k^2\exp\{-(\zeta q)^2/2\} \leq \frac{2}{3}$, by the choice of q.

4. Proof of Theorem 1.4

Given a graph K on k vertices, we define $\Phi(K)$ to be the set of parameters as in Definition 1.1 satisfying $\alpha_i^{\mathsf{LB}} = \alpha_i^{\mathsf{UB}} = 1/k$ ($\forall i \in [k]$) and $d_{ij}^{\mathsf{LB}} = d_{ij}^{\mathsf{UB}} = 1$ for every $ij \in E(K)$ and $d_{ij}^{\mathsf{LB}} = d_{ij}^{\mathsf{UB}} = 0$ for every $ij \notin E(K)$. In particular, if $G \in \mathcal{P}_{\Phi(K)}$ (and $|V(G)|$ is divisible by k) then G is isomorphic to the n/k-blow-up of K.

The following definition is key to the proof of Theorem 1.4. We remind the reader that if $X \subseteq V(G)$ then we use $G[X]$ to denote the induced subgraph of G on X.

Definition 4.1 (Nice graphs)**.** *We say that a graph $K = (V,E)$ on k vertices is* nice *if for every subset X of $k/2$ vertices of V and every subset $Y \subseteq V \setminus X$ of $k/10$ vertices, the graph $K[Y]$ is $\frac{1}{800}$-far from being an induced subgraph of $K[X]$.*

Claim 4.2. *For every large enough k, there is a k-vertex nice graph.*

Proof. We claim that for large enough k, the uniform random graph $G(k,1/2)$ is nice with high probability. Indeed, suppose the vertex set of $G(k,1/2)$ is V and fix $X \subseteq V$ of size $k/2$, $Y \subseteq V \setminus X$ of size $k/10$ and an injective mapping $f : Y \mapsto X$ (f is a "supposed" isomorphism between Y and a subgraph of X). Observe that for every pair $u,v \in Y$, the probability that precisely one of the pairs $(u,v),(f(u),f(v))$ is an edge is $1/2$. Hence, the expected number of such pairs is $\frac{1}{2}\binom{|Y|}{2} \geq k^2/400$. Furthermore, since for distinct pairs the above events are independent, we have by Lemma 2.3 that the probability that the number of such pairs is smaller than (say) $k^2/800$ is $e^{-\Omega(k^2)}$. Finally, since the number of choices of X,Y,f is at most $2^k \cdot 2^k \cdot k^k = 2^{O(k\log k)}$ we infer by the union bound that with high probability $G(k,1/2)$ is nice. □

In what follows, given a graph H on h vertices. we denote by $H(n)$ the $\frac{n}{h}$-*blow-up* of H and we call the corresponding independent sets (of size $\frac{n}{h}$) the *clusters* of $H(n)$. Note that a graph G on n vertices is in $\mathcal{P}_{\Phi(K)}$ if and only if G is isomorphic to $K(n)$.

Claim 4.3. *The following holds for all large enough k. Let $K = (V, E)$ be a nice graph on k vertices and $K' = K[X]$ be the graph induced by a subset $X \subseteq V$ of size $|X| = k/2$. Then, $K'(n)$ is 0.0001-far from $\mathcal{P}_{\Phi(K)}$.*

Proof. Suppose wlog that $V = [k]$ and $X = [k/2]$. Let $\{U_j\}_{j=1}^{k/2}$ be the clusters of $K'(n)$, each of size $2n/k$. Suppose, by contradiction, that $K'(n)$ is 0.0001-close to $\mathcal{P}_{\Phi(K)}$. Then, there must be a partition $\{V_i\}_{i=1}^{k}$ of $K'(n)$ such that, by performing a set $\Delta \subseteq \binom{V}{2}$ of at most $0.0001n^2$ edge modifications, one gets a graph satisfying $d(V_i, V_j) = 1$ if $K(i, j) = 1$ and $d(V_i, V_j) = 0$ if $K(i, j) = 0$. For every $1 \leq j \leq k/2$ and every $k/2 + 1 \leq i \leq k$, let $c_{i,j} := |V_i \cap U_j|/|U_j| \leq 1$ and observe that for every $k/2 + 1 \leq i \leq k$ we have $c_{i,1} + \cdots + c_{i,k/2} = 1/2$.

For every $1 \leq j \leq k/2$, let y_j be a vertex chosen uniformly at random from the cluster U_j. Set $U := \{y_j\}_{j=1}^{k/2}$ and $Y := \{k/2 + 1 \leq i \leq k : |U \cap V_i| \geq 1\}$. For each fixed $k/2 + 1 \leq i \leq k$, the probability that none of the vertices $y_1, \ldots, y_{k/2}$ belongs to V_i is

$$(1 - c_{i,1}) \cdots (1 - c_{i,k/2}) \leq e^{-(c_{i,1}+\cdots+c_{i,k/2})} = e^{-0.5} < 3/4.$$

By linearity of expectation, this means that $\mathbb{E}(|Y|) \geq \frac{1}{8}k$, and since the outcome of each y_j can change $|Y|$ by at most 1, we have by Lemma 2.3, that the probability that $|Y| < k/10$ is at most 3/4 (for large k). Furthermore, since $\Delta < 0.0001n^2$, the expected number of pairs of Δ within $\binom{U}{2}$ is at most $0.0001\frac{k^2}{8} < k^2/1000$. Hence, by Lemma 2.3, the probability that this number is larger than $k^2/800$ is smaller than 1/4 — note that this implies that $K[Y]$ is $\frac{1}{800}$-close to a subgraph of $K[X]$ with probability at least 3/4. Indeed, this subgraph is the one spanned by the vertex set $X' \subseteq X$ defined as follows: for each $i \in Y$ put in X' precisely one of the $1 \leq j \leq k/2$ satisfying $y_j \in V_i$. We infer that, with probability at least 1/2, the set Y has size at least $k/10$ and $K[Y]$ is $\frac{1}{800}$-close to a subgraph of $K[X]$. Therefore, X and Y contradict the assumption that K is nice. □

Proof of Theorem 1.4: Let k be large enough so that Claim 4.3 holds and suppose K is a nice graph on k vertices. We claim that every 0.0001-tester for $\mathcal{P}_{\Phi(K)}$ has query complexity at least $\Omega(\sqrt{k})$. By Yao's min-max principal it is enough to show that, for every large enough n, there are two distributions D_1 and D_2 of n-vertex graphs, so that graphs in D_1 belong to $\mathcal{P}_{\Phi(K)}$ while those in D_2 are 0.0001-far from $\mathcal{P}_{\Phi(K)}$, and such that any deterministic algorithm with query complexity $o(\sqrt{k})$ has negligible probability of distinguishing between D_1 and D_2. The distribution D_1 contains a random permutation of the vertices of $K(n)$ (which belongs to $\mathcal{P}_{\Phi(K)}$). The distribution D_2 is defined as follows: we first pick a random subset $X \subseteq V(K)$ of size $k/2$ and then return a random permutation of the vertices of $K'(n)$, where $K' = K[X]$. By Claim 4.3, every graph in D_2 is 0.0001-far from $\mathcal{P}_{\Phi(K)}$. Finally, it is easy to see that any deterministic algorithm with query complexity $q = o(\sqrt{k})$ cannot distinguish between D_1 and D_2 with constant positive probability since in both cases what it sees is just a random subgraph of K on q vertices. Indeed, by taking $q = o(\sqrt{k})$ we guarantee that the algorithm does not pick more than one vertex from the same cluster of either $K'(n)$ or $K(n)$, that is, that in both cases it gets q *distinct* vertices of K. The fact that in D_1 it sees a random subset of K of size q follows from the fact that we randomly permute the vertices of $K(n)$. The fact that the same holds also for D_2 follows from the simple observation that a uniformly random subset of size q of a uniformly random subset of K of size $k/2$ is also a uniformly random subset of K of size q. □

References

[Abb18] Emmanuel Abbe. Community detection and stochastic block models: recent developments. *Journal of Machine Learning*, 18:1–86, 2018.

[AE02] Gunnar Andersson and Lars Engebretsen. Property testers for dense constraint satisfaction programs on finite domains. *Random Structures Algorithms*, 21(1):14–32, 2002.

[AFdlVKK03] Noga Alon, W. Fernandez de la Vega, Ravi Kannan, and Marek Karpinski. Random sampling and approximation of MAX-CSPs. *J. Comput. System Sci.*, 67(2):212–243, 2003. Special issue on STOC2002 (Montreal, QC).

[AFKS00] Noga Alon, Eldar Fischer, Michael Krivelevich, and Mario Szegedy. Efficient testing of large graphs. *Combinatorica*, 20(4):451–476, 2000.

[AS02] Noga Alon and Asaf Shapira. Testing satisfiability. In David Eppstein, editor, *Proceedings of the Thirteenth Annual ACM-SIAM Symposium on Discrete Algorithms, January 6-8, 2002, San Francisco, CA, USA*, pages 645–654. ACM/SIAM, 2002.

[BCL$^+$08] Christian Borgs, Jennifer T. Chayes, László Lovász, Vera T. Sós, and Katalin Vesztergombi. Convergent sequences of dense graphs. I. Subgraph frequencies, metric properties and testing. *Adv. Math.*, 219(6):1801–1851, 2008.

[BLR93] Manuel Blum, Michael Luby, and Ronitt Rubinfeld. Self-testing/correcting with applications to numerical problems. *J. Comput. Syst. Sci.*, 47(3):549–595, 1993.

[BT97] Dimitris Bertsimas and John N. Tsitsiklis. *Introduction to linear optimization*, volume 6 of *Athena scientific optimization and computation series*. Athena Scientific, 1997.

[BY22] Arnab Bhattacharyya and Yuichi Yoshida. *Property Testing - Problems and Techniques*. Springer, 2022.

[CS05] Artur Czumaj and Christian Sohler. Testing hypergraph colorability. *Theoret. Comput. Sci.*, 331(1):37–52, 2005.

[FK96] Alan Frieze and Ravi Kannan. The regularity lemma and approximation schemes for dense problems. In *37th Annual Symposium on Foundations of Computer Science (Burlington, VT, 1996)*, pages 12–20. IEEE Comput. Soc. Press, Los Alamitos, CA, 1996.

[FK99] Alan Frieze and Ravi Kannan. Quick approximation to matrices and applications. *Combinatorica*, 19(2):175–220, 1999.

[FN07] Eldar Fischer and Ilan Newman. Testing versus estimation of graph properties. *SIAM J. Comput.*, 37(2):482–501 (electronic), 2007.

[FR21] Nimrod Fiat and Dana Ron. On efficient distance approximation for graph properties. In *Proceedings of the 2021 ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 1618–1637. [Society for Industrial and Applied Mathematics (SIAM)], Philadelphia, PA, 2021.

[GGR98] Oded Goldreich, Shafi Goldwasser, and Dana Ron. Property testing and its connection to learning and approximation. *J. ACM*, 45(4):653–750, 1998.

[GKS23] Lior Gishboliner, Nick Kushnir, and Asaf Shapira. Testing versus estimation of graph properties, revisited. *Manuscript*, 2023.

[Gol17] Oded Goldreich. *Introduction to Property Testing*. Cambridge University Press, 2017.

[GT03] Oded Goldreich and Luca Trevisan. Three theorems regarding testing graph properties. *Random Structures Algorithms*, 23(1):23–57, 2003.

[HKL$^+$20] Carlos Hoppen, Yoshiharu Kohayakawa, Richard Lang, Hanno Lefmann, and Henrique Stagni. Estimating parameters associated with monotone properties. *Comb. Probab. Comput.*, 29(4):616–632, 2020.

[HKL$^+$21] Carlos Hoppen, Yoshiharu Kohayakawa, Richard Lang, Hanno Lefmann, and Henrique Stagni. On the query complexity of estimating the distance to hereditary graph properties. *SIAM J. Discret. Math.*, 35(2):1238–1251, 2021.

[LW19] Clement Lee and Darren J. Wilkinson. A review of stochastic block models and extensions of graph clustring. *Applied Network Science*, 4:122, 2019.

[McD89] Colin McDiarmid. On the method of bounded differences. In *Surveys in combinatorics, 1989 (Norwich, 1989)*, volume 141 of *London Math. Soc. Lecture Note Ser.*, pages 148–188. Cambridge Univ. Press, Cambridge, 1989.

[PRR06] Michal Parnas, Dana Ron, and Ronitt Rubinfeld. Tolerant property testing and distance approximation. *J. Comput. System Sci.*, 72(6):1012–1042, 2006.

[RS96] Ronitt Rubinfeld and Madhu Sudan. Robust characterizations of polynomials with applications to program testing. *SIAM J. Comput.*, 25(2):252–271, 1996.

[RV07] Mark Rudelson and Roman Vershynin. Sampling from large matrices: an approach through geometric functional analysis. *J. ACM*, 54(4):Art. 21, 19, 2007.

[Sze78] Endre Szemerédi. Regular partitions of graphs. In *Problèmes combinatoires et théorie des graphes (Colloq. Internat. CNRS, Univ. Orsay, Orsay, 1976)*, volume 260 of *Colloq. Internat. CNRS*, pages 399–401. CNRS, Paris, 1978.

[Yao77] Andrew Chi-Chih Yao. Probabilistic computations: Toward a unified measure of complexity (extended abstract). In *18th Annual Symposium on Foundations of Computer Science, Providence, Rhode Island, USA, 31 October - 1 November 1977*, pages 222–227. IEEE Computer Society, 1977.