Testing Versus Estimation of Graph Properties, Revisited

Lior Gishboliner
ETH Zürich, Switzerland

Nick Kushnir
School of Mathematics, Tel Aviv University, Israel

Asaf Shapira
School of Mathematics, Tel Aviv University, Israel

Abstract

A graph G on n vertices is ε-far from property $\mathcal{P}$ if one should add/delete at least εn^2 edges to turn G into a graph satisfying $\mathcal{P}$. A *distance estimator* for $\mathcal{P}$ is an algorithm that given G and $\alpha, \varepsilon > 0$ distinguishes between the case that G is $(\alpha - \varepsilon)$-close to $\mathcal{P}$ and the case that G is α-far from $\mathcal{P}$. If $\mathcal{P}$ has a distance estimator whose query complexity depends only on ε, then $\mathcal{P}$ is said to be *estimable*.

Every estimable property is clearly also testable, since testing corresponds to estimating with $\alpha = \varepsilon$. A central result in the area of property testing is the Fischer–Newman theorem, stating that an inverse statement also holds, that is, that every testable property is in fact estimable. The proof of Fischer and Newmann was highly ineffective, since it incurred a tower-type loss when transforming a testing algorithm for $\mathcal{P}$ into a distance estimator. This raised the natural problem, studied recently by Fiat–Ron and by Hoppen–Kohayakawa–Lang–Lefmann–Stagni, whether one can find a transformation with a polynomial loss. We obtain the following results.

- We show that if $\mathcal{P}$ is hereditary, then one can turn a tester for $\mathcal{P}$ into a distance estimator with an exponential loss. This is an exponential improvement over the result of Hoppen et. al., who obtained a transformation with a double exponential loss.
- We show that for every $\mathcal{P}$, one can turn a testing algorithm for $\mathcal{P}$ into a distance estimator with a double exponential loss. This improves over the transformation of Fischer–Newman that incurred a tower-type loss.

Our main conceptual contribution in this work is that we manage to turn the approach of Fischer–Newman, which was inherently ineffective, into an efficient one. On the technical level, our main contribution is in establishing certain properties of Frieze–Kannan Weak Regular partitions that are of independent interest.

2012 ACM Subject Classification Mathematics of computing → Approximation algorithms

Keywords and phrases Testing, estimation, weak regularity, randomized algorithms, graph theory, Frieze-Kannan Regularity

Digital Object Identifier 10.4230/LIPIcs.APPROX/RANDOM.2023.46

Category RANDOM

Related Version *Full Version*: https://arxiv.org/abs/2305.05487

Funding *Lior Gishboliner*: Supported by SNSF grant 200021_196965.
Asaf Shapira: Supported in part by ERC Consolidator Grant 863438 and NSF-BSF Grant 20196.

1 Introduction

1.1 Background on graph property testing

Property testers are fast randomized algorithms that can distinguish between objects satisfying some predetermined property $\mathcal{P}$ and those that are ε-far from satisfying $\mathcal{P}$. In most cases, ε-far means that an ε-proportion of the object's representation needs to be changed in order

to obtain a new object satisfying $\mathcal{P}$. Hence, testing for $\mathcal{P}$ is a relaxed version of the classical decision problem which asks to decide whether an object satisfies $\mathcal{P}$. 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. The reader is referred to [20] for more background and references on property testing.

We now introduce the model of testing graph properties in the adjacency matrix model. A graph property $\mathcal{P}$ is a family of graphs closed under isomorphism. A graph G on n vertices is ε-far from $\mathcal{P}$ if one should add/delete at least εn^2 edges to turn G into a graph satisfying $\mathcal{P}$. If G is not ε-far from $\mathcal{P}$ then it is ε-close to $\mathcal{P}$. A *tester* for $\mathcal{P}$ is a randomized algorithm that given $\varepsilon > 0$ distinguishes with high probability (say, 2/3) between graphs satisfying $\mathcal{P}$ and those that are ε-far from $\mathcal{P}$. We assume the algorithm can query for each $1 \leq i, j \leq n$ whether the input G contains the edge (i, j). The *edge query complexity*, denoted $Q(\varepsilon)$, of a tester is the number of edge queries it performs. If $\mathcal{P}$ has a tester whose edge query complexity depends only on ε (and is independent of n) then $\mathcal{P}$ is called *testable*. In what follows we will mainly work with *vertex query complexity* which is the smallest $q = q(\varepsilon)$ so that we can ε-test $\mathcal{P}$ by inspecting a subgraph of the input graph G, induced by a set of q randomly selected vertices. By a theorem of Goldreich and Trevisan [22] we know that $q(\varepsilon) \leq 2Q(\varepsilon) \leq q^2(\varepsilon)$. In most (but not all) discussions below we will not care much about these quadratic factors. In such cases we might use the term *query complexity* without mentioning if this is vertex or edge query complexity.

Property testing in the adjacency matrix model was first introduced by Goldreich, Goldwasser and Ron [21], who proved that every *partition property* (e.g. k-colorability and MAX-CUT) is testable. There are several general results guaranteeing that a graph property is testable [3, 10]. A result of this nature was obtained by Alon and Shapira [5] who proved that every hereditary[1] graph property is testable. Their proof applied Szemerédi's regularity lemma [35] (see also [33]), which is one of the most useful tools when studying properties of dense graphs. Using this tool comes with a hefty price, since the bounds one obtains when using the regularity lemma are of tower-type[2].

One of the central open (meta) problems related to testing graph properties is when can one turn an ineffective (e.g. one with tower-type bounds) result into an efficient one, preferably with polynomial bounds. While this is a quantitative question, what lies beneath it is in fact the following qualitative problem; when can we prove a testability result while avoiding Szemerédi's regularity lemma, either by giving a direct combinatorial argument or by using a weaker variant of the regularity lemma (e.g. the Frieze–Kannan regularity lemma [18] which we discuss below). For example, Rödl and Duke [31] used the regularity lemma in order to (implicitly) prove that k-colorability is testable. The tower-type bounds obtained in [31] were improved to polynomial in [21] using a direct argument which avoided the use of the regularity lemma. A specific central open problem, due to Alon and Fox [4], concerns hereditary properties, and asks which hereditary properties are testable with query complexity poly$(1/\varepsilon)$. A systematic investigation of this problem was carried out in [19].

1.2 Distance estimation

In the dense graph model we say that a graph's distance from $\mathcal{P}$ is α, if α is the smallest real so that G is α-close to $\mathcal{P}$. In other words, this is the minimum number of edges one should add/delete in order to obtain a graph satisfying $\mathcal{P}$, normalised by n^2. We denote

[1] A graph property is hereditary if it is closed under vertex removal. Some examples are being 3-colorable, being triangle-free and being induced H-free, for some fixed H.

[2] The tower function tower(x) is a tower of exponents of height x.

this quantity by $\text{dist}_{\mathcal{P}}(G)$. A *distance estimator* for $\mathcal{P}$ is a randomized algorithm that given $\alpha, \varepsilon > 0$ distinguishes with high probability (say, 2/3) between graphs that are $(\alpha - \varepsilon)$-close to $\mathcal{P}$ and those that are α-far from $\mathcal{P}$. If for every α, ε there is a distance estimator for $\mathcal{P}$ whose query complexity depends only on ε, then $\mathcal{P}$ is said to be *estimable*. Note that testing $\mathcal{P}$ is equivalent to distance estimation with $\alpha = \varepsilon$, hence this notion is at least as strong as testability.

Distance estimation was first studied in [30] and has since been studied in various other settings such as distributions [7], strings [6], sparse graphs [11, 13, 28], boolean functions [1, 9], error correcting codes [23, 26] and image processing [8]. It is known that in certain settings, there are testable properties which are not estimable [15]. One of the central and most unexpected results in the area of graph property testing is the Fischer–Newman theorem [16], which states that in the setting of graphs, every testable property is also estimable. As with several of the main results in this area, the proof in [16] relied on Szemerédi's regularity lemma [35] and thus resulted in a tower-type loss when transforming a tester for $\mathcal{P}$ into a distance estimator for $\mathcal{P}$. Returning to the discussion in the last paragraph of the previous subsection, it is natural to ask if one can improve the transformation of [16] and turn a tester for $\mathcal{P}$ into a distance estimator with a polynomial loss.

1.3 New results concerning hereditary graph properties

As we mentioned in the previous subsection, the family of hereditary graph properties has been extensively studied within the setting of graph property testing. The fact that every hereditary property is testable follows from the following statement, where we use $ind(F, G)$ to denote the probability that a random mapping $\varphi : V(F) \to V(G)$ is an injective induced homomorphism. [3]

▶ **Lemma 1** (Induced Removal Lemma, [5]). *For every $\varepsilon > 0$ and every hereditary $\mathcal{P}$, there exists $M = M_1(\varepsilon, \mathcal{P})$, $\delta = \delta_1(\varepsilon, \mathcal{P}) > 0$ and $n_0 = n_1(\varepsilon, \mathcal{P})$ such that if a graph G on $n \geq n_0$ vertices is ε-far from $\mathcal{P}$ then there is a graph $F \notin \mathcal{P}$ with $|V(F)| \leq M$ such that $ind(F, G) \geq \delta$.*

The first version of the above lemma was obtained by Alon, Fischer, Krivelevich and Szegedy [2] who proved it when $\mathcal{P}$ can be characterized using a finite number of forbidden induced subgraphs. The lemma was proved in full generality by Alon and Shapira [5]. Alternative proofs were later obtained by Lovász and Szegedy [27], Conlon and Fox [12] and Borgs et al. [10]. It was also extended to the setting of hypergraphs by Rödl and Schacht [32].

Note that it follows immediately from Lemma 1 that every hereditary property is testable with vertex query complexity

$$q(\varepsilon) = \max\{n_0, M/\delta\} \,. \tag{1}$$

Indeed, the algorithm samples a set X of q vertices, queries about all pairs within X, and then accepts if and only if the graph on X satisfies $\mathcal{P}$. If G satisfies $\mathcal{P}$ then the algorithm clearly answers correctly (with probability 1). If G is ε-far from $\mathcal{P}$, then by Lemma 1 a random M-tuple of vertices spans an induced copy of a graph $F \notin \mathcal{P}$ with probability at least δ. Hence, a sample of size M/δ contains an induced copy of F with probability at least 2/3, thus guaranteeing that the sample of vertices does not satisfy $\mathcal{P}$ (since $\mathcal{P}$ is hereditary). Recall that [22] proved that if $\mathcal{P}$ is testable, then it is testable using an algorithm as above. Hence, the bounds in Lemma 1 more or less determine the query complexity of testing a

[3] A mapping $\varphi : V(F) \to V(G)$ is an induced homomorphism if $uv \in E(F)$ if and only if $\varphi(u)\varphi(v) \in E(G)$.

hereditary $\mathcal{P}$. This raises the following natural problem, introduced by Hoppen et al. [25, 24] and by Fiat and Ron [14], asking if it is possible to estimate every hereditary $\mathcal{P}$ with (roughly) the same query complexity with which it can be tested as in (1).

▶ **Problem 2.** *Determine if every hereditary graph property $\mathcal{P}$ is estimable with query complexity*

$$n_0 \cdot M/\delta \ ,$$

where $M = M_1(\varepsilon', \mathcal{P})$, $\delta = \delta_1(\varepsilon', \mathcal{P})$, $n_0 = n_1(\varepsilon', \mathcal{P})$ are given by Lemma 1 with $\varepsilon' = \mathrm{poly}(\varepsilon)$.

▶ Remark 3. There are hereditary graph properties (e.g. triangle-freeness) for which the best known bounds for M and δ in Lemma 1 are of tower-type. One can argue that in such cases there is little difference between the tower(M/δ) bounds given by [16] and those suggested by Problem 2. However, we should emphasize that for many of these properties (e.g. triangle-freeness) the tower-type bounds are not known to be tight (indeed, the best known lower bounds are just slightly super polynomial). Perhaps more importantly, there are numerous hereditary graph properties for which it is known that both M and δ in Lemma 1 are polynomial in ε (e.g. k-colorability, being an interval graph or being a line graph; see the detailed discussion in [19]). For all these properties, Problem 2 suggests a poly$(1/\varepsilon)$ bound, versus the tower$(1/\varepsilon)$ bound given by [16].

Problem 2 was studied by Hoppen et al. [25, 24]. Their main result was that every hereditary $\mathcal{P}$ is estimable with query complexity $2^{\mathrm{poly}((1/\delta)^{M^2}, \log n_0)}$. Our first main result is the following exponential improvement of this result, making a significant step towards resolving Problem 2.

▶ **Theorem 4.** *Every hereditary $\mathcal{P}$ is estimable with query complexity*

$$2^{\mathrm{poly}(M/\delta, \log n_0)} \ ,$$

where $M = M_1(\varepsilon/2, \mathcal{P})$, $\delta = \delta_1(\varepsilon/2, \mathcal{P})$ and $n_0 = n_1(\varepsilon/2, \mathcal{P})$ are the parameters of Lemma 1.

▶ Remark 5. In all known cases, the best bounds in Lemma 1 are such that $\log n_0 \ll 1/\delta$, hence the upper bound of [25] is $2^{(1/\delta)^{O(M^2)}}$ while the one in Theorem 4 is $2^{\mathrm{poly}(M/\delta)}$.

In almost all cases, results concerning testing of dense graphs rely on combinatorial statements which imply trivial algorithms. For example, the algorithm for testing a hereditary property $\mathcal{P}$ is trivial once we have Lemma 1 at our disposal. In sharp contrast, many estimation results involve sampling a set of vertices and then carrying out a highly non-trivial computation over this sample. This is certainly the case in the present paper, see the proofs of Lemmas 14 and 15. However, thanks to a well known sampling trick [21], one can transfer any estimation result into a combinatorial statement. For example, this trick gives the following corollary of Theorem 4.

▶ **Corollary 6.** *Set $q = 2^{\mathrm{poly}(M/\delta, \log n_0)}$ as in Theorem 4. Then*

$$\Pr_X \left[|\mathrm{dist}_{\mathcal{P}}(G[X]) - \mathrm{dist}_{\mathcal{P}}(G)| \le \varepsilon \right] \ge 2/3 \ ,$$

where the probability is over randomly selected subsets X of q vertices from G, and $G[X]$ is the graph induced by G on X.

It is interesting to note that with Corollary 6 at hand, we can now go back and reprove Theorem 4 using the "trivial/natural" algorithm which samples a set of q vertices X, computes $\mathrm{dist}_{\mathcal{P}}(G[X])$, and then states that G is $(\alpha-\varepsilon)$-close to $\mathcal{P}$ if $\mathrm{dist}_{\mathcal{P}}(G[X]) \leq \alpha - \varepsilon/2$ and is otherwise α-far from $\mathcal{P}$.

Our proof of Theorem 4 actually gives the bound $2^{\mathrm{poly}(M/\varepsilon\delta, \log n_0)}$. One can speculate that $\mathrm{poly}(M/\varepsilon\delta) = \mathrm{poly}(M/\delta)$ since in all known cases δ is at best polynomial in ε, and in many cases much smaller. In order to formally be able to remove the dependence on ε from our bound, we prove the following proposition, where $\mathcal{P}$ is *trivial* if either $\mathcal{P}$ contains all graphs or if it contains finitely many graphs. The proof of this proposition relies on a subtle application of Ramsey's theorem.

▶ **Proposition 7.** *The following holds for every non-trivial hereditary property $\mathcal{P}$. If $q(\varepsilon)$ denotes the vertex query complexity of $\mathcal{P}$ then for every small enough ε, we have*

$$M/\delta \geq q(\varepsilon) \geq \Omega(1/\varepsilon)\ , \tag{2}$$

where $M = M_1(\varepsilon, \mathcal{P})$ and $\delta = \delta_1(\varepsilon, \mathcal{P})$ are the constants of Lemma 1.

The left inequality above follows from (1). Observe that the lower bound on $q(\varepsilon)$ is best possible since it is tight when $\mathcal{P}$ is the property of having no edges (in which case $q(\varepsilon) = O(1/\varepsilon)$). The proof of the proposition will appear in the journal version of the paper.

It is of course natural to study Problem 2 also for specific hereditary properties. A natural problem of this type is whether every hereditary $\mathcal{P}$ that is testable with query complexity $\mathrm{poly}(1/\varepsilon)$ is also estimable with query complexity $\mathrm{poly}(1/\varepsilon)$. Such an investigation was initiated recently by Fiat and Ron [14] who proved such a statement for many natural hereditary properties such as Chordality and not containing an induced path on 4 vertices.

1.4 New results concerning general graph properties

Given the discussion above, the following problem seems natural.

▶ **Problem 8.** *Determine if every property $\mathcal{P}$ that is testable with vertex query complexity $q(\varepsilon)$, is estimable with query complexity $q(\varepsilon')$ for some $\varepsilon' = \mathrm{poly}(\varepsilon)$.*

Prior to this work, the only result concerning general graph properties $\mathcal{P}$ was the transformation of Fischer and Newman [16] which turns a testing algorithm for a graph property $\mathcal{P}$ with query complexity $q(\varepsilon)$ into a distance estimator with query complexity $\mathrm{tower}(q(\varepsilon/2))$. Using the tools we develop in order to obtain Theorem 4, we also obtain the following improved bound.

▶ **Theorem 9.** *If $\mathcal{P}$ is testable with query complexity $q(\varepsilon)$ then it is estimable with query complexity $2^{\mathrm{poly}(1/\varepsilon)\cdot 2^{q(\varepsilon/2)}}$.*

We would like to argue at this point that since any "natural" property satisfies $q(\varepsilon) \geq \log(1/\varepsilon)$ the above bound can be written as $\exp(\exp(\mathrm{poly}(q(\varepsilon/2))))$. In order to formally make such a claim, we prove the following variant of Proposition 7, in which $\mathcal{P}$ is *unnatural* if there is ε_0 so that the following holds for every $0 < \varepsilon < \varepsilon_0$ and $n \geq n_0(\varepsilon)$: either every n-vertex graphs is ε-close to $\mathcal{P}$, or every n-vertex graph does not belong to $\mathcal{P}$. If $\mathcal{P}$ is not unnatural then it is (naturally) *natural*.

▶ **Proposition 10.** *Let $\mathcal{P}$ be a natural property and let $q(\varepsilon)$ be its vertex query complexity, and $Q(\varepsilon)$ be its edge query complexity. Then*

$$Q(\varepsilon) = \Omega(1/\varepsilon)\ . \tag{3}$$

In particular, $q(\varepsilon) = \Omega(\sqrt{1/\varepsilon})$.

The "in particular" part above follows directly from the Goldreich–Trevisan [22] theorem mentioned earlier. Observe that the general lower bound given in (3) is best possible since it is tight when $\mathcal{P}$ is the property of having no edges, where $Q(\varepsilon) = O(1/\varepsilon)$. The proof of the proposition will appear in the journal version of the paper.

1.5 Main technical contributions and comparison to previous approaches

Summary of previous approaches

The main reason why Szemerédi's regularity lemma is so useful when studying testing/estimation problems is that an ε-regular partition of a graph G determines (approximately) the values of $ind(F, G)$ for all small F. Hence, on a very high level, the way one can estimate a graph's distance to a hereditary property $\mathcal{P}$ is to take a *single* ε-regular partition of G (one such exists by the regularity lemma) and then try to modify this partition using the smallest possible number of edge modifications, so that the new partition "predicts" that there are no induced copies of graphs $F \notin \mathcal{P}$ in the new graph G'. A key "continuity" feature one has to use at this stage is that if G has a regular partition with certain edge densities between the clusters of the partition, and one would like to modify G so that in the new graph G' one has a regular partition where the edge densities between the clusters will change on average by γ, then one can achieve this by modifying $(\gamma + o(1))n^2$ edges of G. Fischer and Newman [16] critically relied on the fact that regular partitions in the sense of Szemerédi have this continuity property. The approach of [16] was ineffective since although a regular partition has constant size (i.e., depending only on ε), this constant has tower-type dependence on ε. We should point that one of the key novel ideas of [16] was a method for obtaining the densities of a single Szemerédi partition of the input G.

The way Hoppen et al. [25, 24] managed to improve upon [16] (for hereditary $\mathcal{P}$) was by first observing that in order to estimate $ind(F, G)$ for all small F, one does not need the full power of Szemerédi's regularity lemma. Instead, one can use the weak regularity lemma of Frieze and Kannan [17] which involves constants that are only exponential in ε. The main reason why their proof gave a doubly exponential bound is that Frieze–Kannan regular partitions do not (seem to) have the same continuity feature we mentioned in the previous paragraph with respect to Szemerédi partitions. To overcome this, Hoppen et al. [25, 24] introduced a sophisticated method that somehow combines working with Frieze–Kannan regular partitions in some parts of the proof, together with vertex partitions that have no regularity[4] features at all (these are sometimes called GGR partitions, after [21]) in other parts of the proof.

Our main technical contribution

Our main technical contribution in this paper establishes that Frieze-Kannan weak regular partitions "almost" satisfy the same continuity feature we mentioned above with respect to Szemerédi partitions. What we show is that one can indeed efficiently modify a Frieze–Kannan partition if one starts with a partition with guarantees slightly stronger than those of Frieze–Kannan, and one is content with ending with a usual Frieze–Kannan partition.

[4] Working with partitions that have no regularity requirements has the advantage that they trivially have the continuity property. Indeed, if we want to change the edge density between two sets A, B by γ we just add/remove $\gamma|A||B|$ edges. Needless to say that working with such partitions has various disadvantages resulting from their lack of regularity features.

See Lemma 28 for the precise statement, whose proof relies on a randomized-rounding-type argument. With the above continuity feature at hand, we can now go back to the Fischer–Newman approach and turn it into an effective one, by taking full advantage of the Frieze–Kannan lemma. One additional hurdle we need to overcome in order to make sure we only incur an exponential loss in our proof, is a method for finding a Frieze–Kannan partition of a graph using a constant number of queries. Here we introduce a variant of the method of Fischer–Newman tailored for Frieze–Kannan partitions, see Lemma 14. The main tools we develop for proving Theorem 4 turn out to be also applicable for proving Theorem 9. The reason why in Theorem 9 we have a double exponential loss is that it is not enough to estimate $ind(F,G)$ for a single F (as in Theorem 4 thanks to Lemma 1) but we instead need to control $ind(F,G)$ for all graphs F of order $q(\varepsilon)$. We expect Lemmas 14 and 28 to be applicable in future studies related to efficient testing and estimation of graph properties.

Paper overview

In Section 2 we introduce the two main lemmas in the paper, and show how they imply Theorem 4. These lemmas are proved in Sections 3 and 4. In Section 5 we prove Theorem 9. We prove Proposition 7 at the end of Section 2 and Proposition 10 at the end of Section 5. We use $a = \text{poly}(x)$ to denote the fact that a is bounded from above (or below, when $0 < x < 1$) by x^d for some fixed d, which is independent of n or ε. Also, when we say that "for every $a = \text{poly}(x)$ there is $b = \text{poly}(x)$" we mean that for every d there is d' so that if $a \le x^d$ then there is a $b \le x^{d'}$.

2 The Key Lemmas and Proof of Theorem 4

Our goal in this section is to state Lemmas 14 and 15 and then use them to derive Theorem 4. We prove these lemmas in Sections 3 and 4. At the end of this section we also prove Proposition 7.

To state Lemmas 14 and 15 we need some definitions. We first recall that given a graph $G = (V,E)$, an equipartition $A = \{V_1,\dots,V_k\}$ of $V(G)$ is a partition satisfying $||V_i| - |V_j|| \le 1$. Given a graph G and subsets $X, Y \subseteq V(G)$, we use $e(X,Y)$ to denote the number of edges between X and Y, and $d(X,Y) = e(X,Y)/|X||Y|$ to denote the *density* between them.

▶ **Definition 11** (Signature). *For an equipartition $A = \{V_1,\dots,V_t\}$ of $V(G)$, a (γ,ε)-signature of A is a sequence of reals $S = (\eta_{i,j})_{1\le i<j\le t}$, such that $|d(V_i,V_j) - \eta_{i,j}| \le \gamma$ for all but at most $\varepsilon\binom{t}{2}$ of the pairs $i<j$. A (γ,γ)-signature is referred to as γ-signature.*

▶ **Definition 12** (Index of a partition). *For an equipartition A of a graph $V(G)$ into t sets, we define the* index *of A to be*

$$ind(A) = \frac{1}{t^2}\sum_{1\le i<j\le t} d^2(V_i,V_j)\ .$$

▶ **Definition 13** (Final partition). *For a function $f:\mathbb{N}\to\mathbb{N}$ and $\gamma>0$, we say that an equipartition A of G consisting of t sets is (f,γ)-final if there exists no equipartition B of $V(G)$ with at least t and up to $f(t)$ sets for which $ind(B) \ge ind(A) + \gamma$.*

The above notion of a final partition is useful since (as we show later) every graph has such a partition and furthermore, we can design an algorithm for finding a signature of one such partition of an input G. The first key lemma leading to the proof of Theorem 4 does exactly that.

▶ **Lemma 14.** *For every $k, \zeta > 0$, and every $\gamma = \text{poly}(\zeta)$ and $f_\zeta(x) = x \cdot 2^{\text{poly}(1/\zeta)}$, there are $q = q_{14}(\zeta, k)$, $N = N_{14}(\zeta, k)$ and $T = T_{14}(\zeta, k)$ so that*

$$q, N, T \leq \text{poly}(k) \cdot 2^{\text{poly}(1/\zeta)}$$

and such that the following holds. If G is a graph on at least N vertices then there is an algorithm making at most q queries to G, computing with probability at least $\frac{2}{3}$ a γ-signature of an (f_ζ, γ)-final partition of G into at least k and at most T sets.

We prove the above lemma is Section 3. The following is the second key lemma, which we prove in Section 4. In its statement we use the notion $ind(F, G)$ which we defined before the statement of Lemma 1. What it roughly states, is that having a signature of G (with good parameters) is enough for estimating G's distance to satisfying $\mathcal{P}$.

▶ **Lemma 15.** *For every $h, \varepsilon, \delta > 0$, there are $\gamma = \gamma_{15}(h, \varepsilon, \delta)$, $s = s_{15}(h, \varepsilon, \delta)$ and $f_{15}^{(h,\varepsilon,\delta)} : \mathbb{N} \to \mathbb{N}$ so that*

$$\gamma = \text{poly}(\varepsilon\delta/h), \quad s = \text{poly}(h/\varepsilon\delta), \quad f_{15}(x) = x \cdot 2^{\text{poly}(h/\varepsilon\delta)}$$

and the following holds. For every family $\mathcal{H}$ of graphs, each on at most h vertices, there exists a deterministic algorithm, that receives as an input a γ-signature S of an (f_{15}, γ)-final partition A into $t \geq s$ sets of a graph G with $n \geq N_{15}(h, \varepsilon, \delta, t) = \text{poly}(t) \cdot 2^{\text{poly}(h/\varepsilon\delta)}$ vertices, and distinguishes given any α between the following two cases:

(i) G is $(\alpha - \varepsilon)$ close to some graph G' for which $ind(H, G') = 0$ for every $H \in \mathcal{H}$.
(ii) G is α-far from every G' for which $ind(H, G') < \delta$ for every $H \in \mathcal{H}$.

Proof (of Theorem 4). Suppose $\mathcal{P}$ is a hereditary graph property, and let $\alpha, \varepsilon > 0$. Lemma 1 with inputs $\varepsilon/2$ and $\mathcal{P}$ asserts that there are

$$h = M_1(\varepsilon/2), \quad \delta = \delta_1(\varepsilon/2), \quad n_0 = n_1(\varepsilon/2) \,,$$

so that if a graph G on at least n_0 vertices is $\varepsilon/2$-far from $\mathcal{P}$, then $ind(H, G) \geq \delta$ for some $H \notin \mathcal{P}$ with $|V(H)| \leq h$. We need to describe an algorithm making $2^{\text{poly}(h/\delta, \log n_0)}$ queries to G and distinguishes with probability at least $2/3$ between the case that G is $(\alpha - \varepsilon)$-close to $\mathcal{P}$ and the case that G is α-far from $\mathcal{P}$. Set

$$\gamma = \gamma_{15}(h, \varepsilon/2, \delta), \quad s = s_{15}(h, \varepsilon/2, \delta), \quad f = f_{15}^{(h,\varepsilon/2,\delta)} \,.$$

Finally, set $\zeta = \delta\varepsilon/h$ and observe that

$$\gamma = \gamma_{15}(h, \varepsilon/2, \delta) = \text{poly}(\varepsilon\delta/2h) = \text{poly}(\zeta) \,,$$

that

$$f(x) = f_{15}^{(h,\varepsilon/2,\delta)}(x) = x \cdot 2^{\text{poly}(2h/\varepsilon\delta)} = x \cdot 2^{\text{poly}(1/\zeta)} \,,$$

that

$$s = s_{15}(h, \varepsilon/2, \delta) = \text{poly}(2h/\varepsilon\delta) = \text{poly}(1/\zeta) \,.$$

Also, note that by Proposition 7 we have $\text{poly}(1/\zeta) = \text{poly}(h/\delta)$. Let q, N, T be the parameters given by Lemma 14 when applied with $k = s$, and ζ, γ, f defined above. (note that γ and f satisfy the assumptions of the lemma). Lemma 14 then guarantees that $q, N, T \leq 2^{\text{poly}(1/\zeta)} \leq 2^{\text{poly}(h/\delta)}$.

If G has less than N vertices then we can just ask about all the edges of G and answer correctly with probability 1. The number of queries is then at most $N^2 \leq 2^{\text{poly}(h/\delta)}$ as needed. If G has more than N vertices then we can use the algorithm of Lemma 14 with the parameters k, ζ, γ, f defined above. The algorithm makes at most $q \leq 2^{\text{poly}(h/\delta)}$ queries and with probability at least $2/3$ returns a γ-signature S of an equipartition of G into $s \leq t \leq T$ sets that is (f, γ)-final. Let

$$N' = N_{15}(h, \varepsilon/2, \delta, T) = \text{poly}(T) \cdot 2^{\text{poly}(h/\varepsilon\delta)} = 2^{\text{poly}(h/\delta)} .$$

Again, if G has less than $N_1 = \max\{N', n_0\}$ vertices then we can just ask about all the edges of G and answer correctly with probability 1. The number of queries is then at most $(N_1)^2 \leq 2^{\text{poly}(h/\delta, \log n_0)}$ as needed.

Suppose then that G has at least $\max\{N, N_1\}$ vertices. Let $\mathcal{H}$ be the family of graph on at most h vertices which do not satisfy $\mathcal{P}$. Then we can now run the algorithm of Lemma 15 on the signature S, with respect to $\mathcal{H}$, with $\alpha' = \alpha - \varepsilon/2$ and with $\varepsilon/2$ instead of ε (note that we chose the parameters with $\varepsilon/2$). If the algorithm says that case (i) holds (namely that G is $(\alpha' - \varepsilon/2)$-close to some G' with $ind(H, G') = 0$ for every $H \in \mathcal{H}$) then we declare that G is $(\alpha - \varepsilon)$-close to $\mathcal{P}$, and if the algorithm says that case (ii) holds (namely that G is α'-far from every G' with $ind(H, G') < \delta$ for every $H \in \mathcal{H}$) then we declare that G is α-far from $\mathcal{P}$.

Let us prove the correctness of the above algorithm. If G is $(\alpha - \varepsilon)$-close to $\mathcal{P}$ then it is $(\alpha - \varepsilon)$-close to a graph G' satisfying $ind(H, G') = 0$ for every $H \in \mathcal{H}$. Since $\alpha - \varepsilon = \alpha' - \varepsilon/2$ the algorithm will say that case (i) holds, hence the algorithm answers correctly in this case. Suppose now that G is α-far from $\mathcal{P}$. Then any G' that is α'-close to G must be $\varepsilon/2$-far from $\mathcal{P}$. Hence, by Lemma 1 in any such G' we have $ind(H, G') \geq \delta$ for at least one $H \in \mathcal{H}$. We conclude that G is α'-far from every G' satisfying $ind(H, G') < \delta$ for every $H \in \mathcal{H}$. Hence, the algorithm of Lemma 15 will say that case (ii) holds , so our algorithm will answer correctly in this case as well. ◀

3 Proof of Lemma 14

The proof is similar to one in [16]. What they have shown is that for every f, γ, one can find an (f, γ)-final partition with a constant, albeit huge tower-type, query complexity. What we do here is show that for restricted types of f, one can get a much better bound. To do this we also need to rely on a recent result of [34].

3.1 Preliminary lemmas

In this subsection we describe some preliminary lemmas that will be used in the next subsection in which we prove Lemma 14. We will need the following Chernoff-type large deviation inequality.

▶ **Lemma 16.** *Suppose $X_1, \ldots, X_m$ are m independent Boolean random variables, so that for every $1 \leq i \leq m$ we have $\Pr[X_i = 1] = p_i$. Let $E = \sum_{i=1}^m p_i$. Then, $\Pr[|\sum_{i=1}^m X_i - E| \geq \theta m] \leq 2e^{-2\theta^2 m}$.*

▶ **Definition 17** (Partition Properties)**.** *A partition property is a triple $\pi = (s, \ell, u)$ where s is an integer (the size of the partition property), ℓ is a vector of $\binom{s}{2}$ reals $0 \leq \alpha_{i,j} \leq 1$ for each $1 \leq i < j \leq s$, and u is a vector of $\binom{s}{2}$ reals $0 \leq \beta_{i,j} \leq 1$ for each $1 \leq i < j \leq s$. We say that a graph G satisfies π if there is an equipartition $\{V_1, \ldots, V_s\}$ of $V(G)$, such that $\alpha_{ij} \leq d(V_i, V_j) \leq \beta_{ij}$ for every $1 \leq i < j \leq s$.*

Given s and μ we use $\pi(s,\mu)$ to denote the family of partition properties π of size s in which every $\alpha_{i,j}$ and $\beta_{i,j}$ is an integer multiple of μ (so $\pi(s,\mu)$ contains $\{0,\mu,2\mu,\ldots,1\}^{2\binom{s}{2}}$ partition properties). Finally, define $\Pi(t,\mu)=\bigcup_{s\leq t}\pi(s,\mu)$.

Note that each π as above is one of the partition properties studied in [21], where it was shown that they are μ-testable with query complexity $(1/\mu)^{\mathrm{poly}(s)}$. This was improved recently to $\mathrm{poly}(s/\mu)$ in [34]. The next lemma states that with (roughly) the same query complexity we can in fact *simultaneously* test all properties in $\Pi(t,\mu)$.

The proof of the next lemma will appear in the journal version of the paper.

▶ **Lemma 18.** *For every t and $\mu>0$ there is $q=q_{18}(t,\mu)=\mathrm{poly}(t/\mu)$ satisfying the following. There is a randomized algorithm, that given a graph G, makes q queries to G and with probability at least $2/3$, for every $\pi\in\Pi(t,\mu)$, distinguishes between the case that G satisfies π and the case that G is μ-far from π.*

Proof (of Lemma 14): Given k,ζ,γ and f_ζ as in the statement of the lemma, we define $T_0=k$ and for $i\geq 1$ define $T_i=f_\zeta(T_{i-1})$. Now set the following parameters.

$$N=N_{14}(k,\zeta)=T_{2/\gamma}=k\cdot 2^{\mathrm{poly}(1/\zeta)},\quad T=T_{14}(k,\zeta)=T_{2/\gamma}=k\cdot 2^{\mathrm{poly}(1/\zeta)}\ ,$$

and

$$t=f_\zeta(T)=k\cdot 2^{\mathrm{poly}(1/\zeta)},\quad \mu=\frac{\gamma}{48(f_\zeta(T))^2}=\frac{1}{\mathrm{poly}(k)\cdot 2^{\mathrm{poly}(1/\zeta)}}\ .$$

We now describe the algorithm for finding a signature S satisfying the requirement of the lemma. For what follows let $\pi'(s,\mu)$ be the partition properties in which $\beta_{i,j}=\alpha_{i,j}+\mu$ for every $1\leq i<j\leq s$. Also for each $\pi\in\pi'(s,\mu)$ define the index of π to be $ind(\pi)=\frac{1}{t^2}\sum_{1\leq i<j\leq t}\alpha_{ij}^2$. In the Step-1 we run the algorithm of Lemma 18 with the parameters t,μ defined above. This is the only randomized part of the algorithm. In the Step-2 of the algorithm we do the following.

- **(i)** For each $k\leq s\leq t$ set $M(s)=\max_\pi ind(\pi)$ where the maximum is taken over all $\pi\in\pi'(s,\mu)$ which the algorithm of Step-1 accepted.
- **(ii)** Let $s^\star$ be the smallest number in $\{k,\ldots,T\}$ such that $M(s')\leq M(s^\star)+\frac{3}{4}\gamma$ for every $s'\in\{s^\star+1,\ldots,f_\zeta(s^\star)\}$. If there exists such an $s^\star$, output the signature $S^\star$ that achieves the maximum over $s^\star$. Otherwise, the algorithm fails.

Note that the query complexity of the algorithm is $q=q_{18}(t,\mu)=\mathrm{poly}(t/\mu)=\mathrm{poly}(k)\cdot 2^{\mathrm{poly}(1/\zeta)}$, as needed. Also, Lemma 18 guarantees that Step-1 of the above described algorithm succeeds with probability at least $2/3$. It thus remains to show that assuming this event holds, Step-2 of the algorithm will return an (f_ζ,γ)-final partition. First of all note that if it succeeds then it returns a partition of size at least k and at most T, as required.

The proof that if Step-1 succeeded, then Step-2 returns an (f_ζ,γ)-final partition is identical to the proof of Claim 5.5 in [16], so we give a sketch of the proof. First, the reader might be wondering why every graph necessarily has an (f_ζ,γ)-final partition as in the statement of the lemma. Let us actually explain why every G has an $(f_\zeta,\gamma/2)$-final partition, while using the definitions we introduced above. Start from an arbitrary equipartition A_0 of G into $T_0=k$ sets, and let $ind_0=ind(A_0)$ denote the index of A_0 as in Definition 12. If A_0 is $(f_\zeta,\gamma/2)$-final then we are done. If not, then there must be another partition A_1 of G with at least T_0 and at most $f(T_0)=T_1$ parts, with index $ind(A_1)\geq ind(A_0)+\gamma/2$. Since $0\leq ind(A)\leq 1$ for every equipartition, we see that this process will eventually end up with

a partition A of size $k \leq s \leq T$ so that all partitions of G into at least s and at most $f(s)$ parts have index less than $ind(A) + \gamma/2$. But this means that A is $(f_\zeta, \gamma/2)$-final. Note that we thus get that G has a $(f_\zeta, \gamma/2)$-final partition A of size $s \leq T$.

Let us now explain how to turn the above existential proof into a proof of correctness of the algorithm describe earlier. Let $M_G(s)$ denote the largest index of an equipartition of G of size s. First we claim that for every $k \leq s \leq t$,

$$M(s) - \gamma/8 \leq M_G(s) \leq M(s) + \gamma/8. \tag{4}$$

For the second inequality in (4), let A be an equipartition with s parts such that $M_G(s) = ind(A)$. Let $\pi \in \pi'(s, \mu)$ be the partition property obtained from A by rounding down the densities to the closest integer multiple of μ. Then we have $|ind(A) - ind(\pi)| \leq 3\mu \leq \gamma/8$. Hence, $M(s) \geq ind(\pi) \geq ind(A) - \gamma/8 = M_G(s) - \gamma/8$.

For the first inequality in (4), let $\pi \in \pi'(s, \mu)$ be a partition property which the algorithm accepted and such that $M(s) = ind(\pi)$. Then G must be μ-close to π (as otherwise π should have been rejected). Let G' be a graph μ-close to G that satisfies π, and let A be the vertex partition of G' witnessing that G' satisfies π. Note that when turning G into G', for each pair of parts of A, we change the density between this pair by at most μs^2. Hence, in G, the partition property π is a $2\mu s^2$-signature of A (here and in what follows, we view π as a signature). So $|ind(A) - ind(\pi)| \leq 6\mu s^2 \leq \gamma/8$, using our choice of μ. Now, $M_G(s) \geq ind(A) \geq ind(\pi) - \gamma/8 = M(s) - \gamma/8$. This proves (4).

It follows from the existential proof above that there is $k \leq s^\star \leq T$ and an equipartition A of G into $s^\star$ parts which is $(f_\zeta, \gamma/2)$-final. We can assume that $M_G(s^\star) = ind(A)$, because the equipartition satisfying this must also be final. We have $M_G(s') \leq M_G(s^\star) + \gamma/2$ for every $s^\star \leq s' \leq f_\zeta(s^\star)$. By (4), this implies that $M(s') \leq M(s^\star) + 3\gamma/4$ for every $s^\star \leq s' \leq f_\zeta(s^\star)$. So the algorithm will return a partition.

Note that the algorithm does not necessarily return the same signature/partition-property as above π that is μ-close to the above partition A. The reason for the algorithm to choose a different partition is that there might be another partition of size s with a larger index (which is of course also (f_ζ, γ)-final) or there might be an $s^* < s$ with the same properties, or there might be other partitions with the same index. However, one can invert the reasoning in the previous paragraph and show that if a π is returned then it must be the γ-signature of an (f_ζ, γ)-final partition. ◀

4 Proof of Lemma 15

4.1 Preliminary lemmas

In this subsection we describe some preliminary lemmas that will be used in the next subsection in which we prove Lemma 15. We start with introducing the Frieze–Kannan regularity lemma [17, 18]. We first state their notion of γ-regularity.

▶ **Definition 19** (Frieze–Kannan Regularity [18]). *Let $G = (V, E)$ be a graph and $A = \{V_1 \dots, V_k\}$ be an equipartition of $V(G)$. For a subset $X \subseteq V$ and $1 \leq i \leq k$ denote $X_i = X \cap V_i$. We say that A is γ-Frieze–Kannan-regular if:*

$$d^A_\square(G) := \max_{S,T \subseteq V} \frac{1}{n^2} \left| \sum_{i,j \in [k]^2} \Big(d(S_i, T_j) - d_{ij} \Big) |S_i||T_j| \right| < \gamma \tag{5}$$

Roughly speaking, a partition A is γ-Frieze–Kannan-regular, or γ-FK-regular for short, if we can estimate the number of edges between large sets S, T from the intersection sizes $S \cap V_i$ and $T \cap V_i$. We will also need the following slightly stronger notion of weak regularity that was introduced in [29].

▶ **Definition 20** (Frieze–Kannan Regularity* [29])**.** *In the setting of Definition 19, we say that A is γ-Frieze–Kannan Regular* if:*

$$d_\square^{\star A}(G) := \max_{S,T\subseteq V} \frac{1}{n^2} \sum_{i,j\in[k]^2} \left| d(S_i, T_j) - d_{ij} \right| |S_i||T_j| < \gamma \tag{6}$$

The translation between these two notions will be crucial in Lemma 28 below. Suppose $A = \{V_1, \ldots, V_k\}$ is an equipartition of $V(G)$. Then an equipartition $B = \{W_1, ..., W_\ell\}$ of $V(G)$ is said to *refine* A if each $W_i \in B$ is contained in some $V_j \in A$. The following lemma is proved in [29] using a simple variant of the original proof of Frieze and Kannan [18].

▶ **Lemma 21** (Frieze–Kannan Weak Regularity Lemma [18, 29])**.** *For every k_0 and $\gamma > 0$ there is $T = T_{21}(k_0, \gamma) = k_0 \cdot 2^{\mathrm{poly}(1/\gamma)}$ so that the following holds for every graph G on at least T vertices. If A is an equipartition of $V(G)$ into at most k_0 sets, then there is a refinement B of A into at most T sets such that $d_\square^{\star B}(G) < \gamma$.*

Let us now extend the definition of $d_\square$ to distance between pairs of weighted graph, where a weighted graph R is a complete graph, so that every edge (i, j) is assigned a weight $0 \le R(i, j) \le 1$.

If R, R' are two weighted graphs on n vertices then we define

$$d_1(R, R') = \frac{1}{n^2} \sum_{i<j} |R(i, j) - R'(i, j)| \,, \tag{7}$$

and

$$d_\square(R, R') = \max_{\alpha,\beta} \frac{1}{n^2} \left| \sum_{i<j} \alpha(i)\beta(j)(R(i, j) - R'(i, j)) \right| \,, \tag{8}$$

where the maximum is taken over all functions $\alpha, \beta : [n] \to [0, 1]$.

▶ **Definition 22** ($ind(F, R)$)**.** *Let R be a weighted graph on $[k]$ and let φ be an injective function $\varphi : V(F) \to [k]$. We set*

$$ind_\varphi(F, R) = \prod_{i<j\in E(F)} R(\varphi(i), \varphi(j)) \prod_{i<j\notin E(F)} (1 - R(\varphi(i), \varphi(j)))$$

In the case of φ not being injective, we define

$$ind_\varphi(F, R) = 0$$

Denoting by Φ the set of functions from $V(F)$ to $[k]$, we define

$$ind(F, R) = \frac{1}{|\Phi|} \sum_{\varphi\in\Phi} ind_\varphi(F, R) \,. \tag{9}$$

Note that we can think of a signature $S = (\eta_{i,j})_{1\le i<j\le t}$ as a weighted graph on t vertices. This means that for a pair of signatures S, S' we can define $d_1(S, S')$ and $d_\square(S, S')$ as in (7) and (8) respectively, and we can also define $ind(F, S)$ as in (9). We will need the following lemmas from [25]. The proof of the next two lemmas will appear in the journal version of the paper.

▶ **Lemma 23.** *Suppose R, R' are two weighted graphs on n vertices, and H is a graph on h vertices. Then for any $\gamma \geq d_{\square}(R, R')$ and $n \geq \frac{2}{\gamma}$, we have $|ind(H, R) - ind(H, R')| \leq 2h^2 \cdot \gamma$*

Given a graph G on n vertices, and an equipartition $A = \{V_1, \ldots, V_k\}$, we define the graph G_A on $V(G)$ to be the weighted graph with weights $G_A(u, v) = d(V_i, V_j)$ for every $u \in V_i$ and $v \in V_j$. Let S_A be the 0-signature of A, that is, the weighted graph on k vertices with $S(i, j) = d(V_i, V_j)$. Observe that if k divides n (so all sets of A are of equal size) then $ind(H, G_A)$ is almost the same as $ind(H, S_A)$. It is not hard to see that for general equipartitions these quantities do not differ my much.

▶ **Lemma 24.** *Given a graph G on n vertices, and an equipartition $A = \{V_1, \ldots, V_k\}$, let G_A and S_A be defined as above. Then $|ind(H, G_A) - ind(H, S_A)| \leq \frac{2h^2}{k} + \frac{2kh}{n}$ for every graph H on h vertices.*

We now combine the above facts to conclude that a signature of a γ-FK-partition of a graph gives a good approximation of $ind(H, G)$. The proof of the next lemma will appear in the journal version of the paper.

▶ **Lemma 25.** *For every h, k and $\delta > 0$ there are*

$$\gamma = \gamma_{25}(h, \delta) = \text{poly}(\delta/h), \quad r = r_{25}(h, \delta) = \text{poly}(h/\delta), \quad N = N_{25}(h, k, \delta) = \text{poly}(hk/\delta)\ ,$$

so that if G is a graph on at least N vertices, and A is a γ-FK-regular partition of G with at least r and up to k parts, then for every γ-signature S of A, we have $|ind(H, G) - ind(H, S)| \leq \delta$ for every H on h vertices.

▶ **Definition 26** (Extension)**.** *Given a signature $S = (\eta_{ij})_{1 \leq i < j \leq t}$ of an equipartition A, and a refinement $B = \{W_1, \ldots, W_s\}$ of A, the extension of S to B is the sequence $S' = (\eta'_{ij})_{1 \leq i < j \leq s}$ defined as $\eta'_{i,j} = \eta_{k,l}$ if there exist $k \neq l$ such that $W_i \subseteq V_k$ and $W_j \subseteq V_l$, and setting $\eta'_{i,j} = 0$ if W_i and W_j are both subsets of the same V_k.*

The proof of the next claim will appear in the journal version of the paper.

▷ Claim 27. For every ε and s there exists $r = r_{27}(\varepsilon) = \text{poly}(1/\varepsilon)$ and $N = N_{27}(\varepsilon, s) = \text{poly}(s/\varepsilon)$ so that the following holds for every pair of graphs G, G' on the same set of $n \geq N$ vertices. If G, G' are α-close and S, S' are γ, γ'-signatures of G, G' respectively, of the same equipartition A of the vertex set of G, G' into $s \geq r$ sets, then $d_1(S, S') \leq \alpha + \varepsilon + 2(\gamma + \gamma')$.

The proof of the next lemma will appear in the journal version of the paper.

▶ **Lemma 28.** *For every ε and t there exists $\gamma = \gamma_{28}(\varepsilon) = \text{poly}(\varepsilon)$ and $N = N_{28}(t, \varepsilon) = \text{poly}(t/\varepsilon)$, so that for every graph G on $n \geq N$ vertices, if S is a γ-signature of a γ-FK-regular* partition A of G with t sets, then for every signature S' satisfying $d_1(S, S') \leq \delta$ for some δ, there is a graph G' that is $(\delta + \varepsilon)$-close to G, so that A is an ε-FK-regular partition of G', and S' is an ε-signature of A.*

We will also need the following lemmas.

▶ **Lemma 29** ([2] Lemma 3.7)**.** *For every ε, t there exists $\gamma = \gamma_{29}(\varepsilon) = \text{poly}(\varepsilon)$ and $N = N_{29}(t, \varepsilon) = \text{poly}(t/\varepsilon)$ satisfying the following. Assume A is an equipartition into s sets of a graph G with $n \geq N$ vertices, and that B is a refinement of A into at most t sets. Assume further that S is any γ-signature of A, and that T is its extension to B. If B satisfies $ind(B) \leq ind(A) + \gamma$, then T is an ε-signature for B.*

▶ **Lemma 30** ([16] Lemma 6.6). *For every ε, t there exists $N = N_{30}(t,\varepsilon) = \mathrm{poly}(t/\varepsilon)$ so that for every equipartition A of G with $n \geq N$ vertices into s sets, and every refinement B of A into at most t sets, $ind(B) \geq ind(A) - \varepsilon$.*

The next observation is implicit in the proof of the Frieze–Kannan Regularity Lemma (i.e. Lemma 21). The main step of the proof involves showing that if A is an equipartition of G into t parts and A is not ε-FK-regular*, then A has a refinement B into $k \leq 16t/\varepsilon^4$ sets so that $ind(B) \geq ind(A) + \frac{\varepsilon^4}{2}$ (see, e.g., the proof of Theorem 1.1 in [33] and the proof of Theorem 6 in [29]).

▶ **Lemma 31.** *For every $\varepsilon > 0$ there exists $\gamma = \gamma_{31}(\varepsilon) = \mathrm{poly}(\varepsilon)$ and $f = f_{31}^{(\varepsilon)} : \mathbb{N} \to \mathbb{N}$ satisfying $f(x) = \mathrm{poly}(1/\varepsilon) \cdot x$ and such that every (f,γ)-final partition of a graph is also ε-FK-regular*.*

The proof of the next lemma will appear in the journal version of the paper.

▶ **Lemma 32.** *For every s and $\varepsilon > 0$ there are $\gamma = \gamma_{32}(\varepsilon)$, $T = T_{32}(s,\varepsilon)$, $f = f_{32}^{(\varepsilon)}$ and $N = N_{32}(\varepsilon, s)$ so that*

$$\gamma = \mathrm{poly}(\varepsilon), \quad T = s \cdot 2^{\mathrm{poly}(1/\varepsilon)}, \quad f(x) = x \cdot 2^{\mathrm{poly}(1/\varepsilon)}, \quad N = \mathrm{poly}(s) \cdot 2^{\mathrm{poly}(1/\varepsilon)}$$

and the following holds. Suppose G has at least N vertices and A is an (f,γ)-final partition of G into at most s sets and that S is a γ-signature of A. Then for every G' on the same vertex set of G, there exists a refinement A' of A into $t \leq T$ sets so that

- **(i)** *A' is an ε-FK-regular* partition of G'.*
- **(ii)** *Every refinement A'' of A with $t \leq T$ sets (and in particular A'), is an ε-FK-regular* partition of G.*
- **(iii)** *For every refinement A'' of A with $t \leq T$ sets, the extension S'' of S (in the sense of Definition 26) with respect to A'' is an ε-signature of A'' with respect to G (note that A' is such an A'').*

4.2 Proof of Lemma 15

Given h, ε and δ we first choose

$$\gamma_0 = \min\{\varepsilon/10, \gamma_{25}(h,\delta/6), \gamma_{28}(\min\{\varepsilon/2, \gamma_{25}(h,\delta/6)\})\} = \mathrm{poly}(\varepsilon\delta/h)\ ,$$

and then define

$$\gamma = \gamma_{32}(\gamma_0) = \mathrm{poly}(\varepsilon\delta/h), \quad s = \max\{r_{25}(h,\delta/6), r_{27}(\varepsilon/10), 20h^2/\delta\} = \mathrm{poly}(h/\varepsilon\delta)\ ,$$

$$f(x) = f_{32}^{(\gamma_0)}(x) = x \cdot 2^{\mathrm{poly}(1/\gamma_0)} = x \cdot 2^{\mathrm{poly}(h/\varepsilon\delta)}\ ,$$

to be the constants and function in the statement of Lemma 15, noting that they satisfy the guarantees of that lemma. Given t as in the statement of Lemma 15, we set

$$T = T_{32}(t, \gamma_0)$$

and define

$$N = \max\{N_{25}(h,T,\delta/6), N_{27}(\varepsilon/10,T), N_{32}(\gamma_0,s), N_{28}(t,\gamma_0)\} = \mathrm{poly}(t) \cdot 2^{\mathrm{poly}(h/\varepsilon\delta)}\ ,$$

to be the constant in Lemma 15.

Given a family of graphs $\mathcal{H}$ on at most h vertices, we define a family of signatures as follows

$$\mathcal{C}_{\delta,\mathcal{H},T} = \{C : |C| \leq T \text{ and } ind(H,C) \leq \delta/2 \text{ for every } H \in \mathcal{H}\} .$$

In order for $\mathcal{C}_{\delta,\mathcal{H},T}$ to be finite, we only put in it signatures C with edge weights $\eta_{i,j}$ that are integer multiples of $\beta = \min\{\varepsilon/10, \delta/10h^2\}$. Intuitively, this is the set of signatures "certifying" (hence C) that a graph with that signature is close to being induced $\mathcal{H}$-free. We also define $\mathcal{S}_T$ to be the set of all signatures on up to T parts, that are extensions[5] of S. Intuitively, these are the signatures one can obtain by refining A into at most T sets (recall that the crucial point is that the algorithm only has access to S and not to G).

Suppose now that we are given a γ-signature S of some (f,γ)-final (with the above defined f,γ) partition A of a graph G, so that S has $t \geq s$ parts and G has at least N vertices. The algorithm checks if there are $S' \in \mathcal{S}_T$ and $C \in \mathcal{C}_{\delta,\mathcal{H},T}$ satisfying $d_1(S',C) \leq \alpha - \frac{\varepsilon}{2}$. If there is such a pair, the algorithm says that case (i) holds, otherwise it says that case (ii) holds. We now prove the correctness of the algorithm.

Proof of first direction

Suppose there is a graph G' which is $(\alpha-\varepsilon)$-close to G, and satisfies $ind(H,G') = 0$ for every $H \in \mathcal{H}$. We will show that the algorithm will declare that case (i) holds.

Recall that A is an (f,γ)-final partition of G into $t \geq s$ sets and that S is a γ-signature of A. By Lemma 32, there exists a refinement A' of A into at most T sets so that A' is γ_0-FK-regular* for both G **and** G'. Moreover, denoting by S' the corresponding extension of S to A', we have that S' is a γ_0-signature of A' with respect to G. Note that $S' \in \mathcal{S}_T$. By the choice of γ_0, this implies that A' is $\gamma_{25}(h,\delta/6)$-FK-regular* for both G and G', and that S' is a $\frac{1}{10}\varepsilon$-signature of A' with respect to G. Let C' be the 0-signature of A' over G'. Lemma 25 (using A' and G') implies that $|ind(H,G') - ind(H,C')| \leq \delta/6$ for all $H \in \mathcal{H}$. Thus $ind(H,C') \leq \delta/6$ for all $H \in \mathcal{H}$. Clearly there is a signature C of size C' so that all of C's weights are constant multiples of β and $d_1(C',C) \leq \beta$. Since $d_\square(C',C) \leq d_1(C',C) \leq \delta/10h^2$ we infer from Lemma 23 (applied on $\frac{\delta}{10h^2}$, as $s \geq \frac{20h^2}{\delta}$) that $ind(H,C) \leq \delta/6 + \delta/5 < \delta/2$ for all $H \in \mathcal{H}$, so $C \in \mathcal{C}_{\delta,\mathcal{H},T}$. In addition, by Claim 27 (since A' has at least $r_{27}(\varepsilon/10)$ parts and assuming that n is large enough), we infer that $d_1(S',C) \leq \alpha - \frac{\varepsilon}{2}$ (since G and G' are $(\alpha-\varepsilon)$-close and $d_1(C,C') \leq \varepsilon/10$). Thus, S' and C provide a witness that the algorithm will indeed declare that case (i) holds.

Proof of second direction

Suppose the algorithm declares that case (i) holds. We show that in this case there is a graph G', which is α-close to G, and satisfies $ind(H,G') < \delta$ for all $H \in \mathcal{H}$.

Indeed, if the algorithm declared that case (i) holds then there are signatures $S' \in \mathcal{S}_T$ and $C \in \mathcal{C}_{\delta,\mathcal{H},T}$ satisfying $d_1(S',C) \leq \alpha - \frac{\varepsilon}{2}$. As $S' \in \mathcal{S}_T$, there is a refinement A' of A, so that S' is the extension of S according to A'. Lemma 32 (regarding A' as a possible

[5] Note that strictly speaking, an extension per Definition 26 must be relative to a partition A and its refinement B, while here we only have the signature S. So what we mean here is that if one takes some graph that has a partition A whose 0-signature is S, then $\mathcal{S}_T$ is the family of all signatures that one obtains by taking all refinements of A into at most T sets, and then taking the extension of S to these refinements. Of course we do not need any graph in order to produce $\mathcal{S}_T$; we just break the "parts" of S into a total of at most T new "parts", and then define the densities $\eta'_{i,j}$ between the new vertices as in Definition 26.

refinement of A with respect to G) asserts that S' is a γ_0-signature of A' (with respect to G), which by the choice of γ_0 means that it is a $\gamma_{28}(\min\{\frac{\varepsilon}{2}, \gamma_{25}(h, \delta/6)\})$-signature for A' with respect to G. Now, Lemma 28 (applied with A' as the γ_0-FK-regular* partition of G, and with S' as S and C as S') implies that there is a graph G' that is $(\alpha - \frac{\varepsilon}{2} + \frac{\varepsilon}{2})$-close to G, namely α-close to G, and for which C is a $\gamma_{25}(h, \delta/6)$-signature of A', which in turn is $\gamma_{25}(h, \delta/6)$-FK-regular over G' . Lemma 25 implies that $|ind(H, G') - ind(H, C)| \leq \delta/6$ for all $H \in \mathcal{H}$. Thus, $ind(H, G') < \delta/2 + \delta/6 < \delta$ for all $H \in \mathcal{H}$ as required. Hence we have found the required G'.

5 Proof of Theorem 9

The proof of Theorem 9 is very similar to that of Theorem 4. In order to assist the reader who is already familiar with the proof of Theorem 4, we mention in several places where certain lemmas are analogous to lemmas we introduced in one of the previous sections. The idea is the following: by a theorem of Goldreich and Trevisan [22], every testable property is testable by a canonical tester, which samples a set of vertices of size $q = q_{\mathcal{P}}(\varepsilon)$ and accepts/rejects based on the graph induced by these q vertices. Hence the acceptance/rejection of the algorithm only depends on the number of induced copies in G of graphs on q vertices. Hence, turning a graph into a graph satisfying $\mathcal{P}$ is equivalent to turning it into a graph with a certain number of copies of certain graphs on q vertices. As evident, this is very similar to the case of Theorem 4 where we wanted to have a very small number of copies of graphs not in $\mathcal{P}$. The reason why there is an additional exponential factor is that we need to control the number of induced copies of *all* graphs on q vertices.

We now state the key lemmas, which are variants of lemmas we used in the proof of Theorem 4.

▶ **Definition 33.** *Given two distributions μ and ν over a finite family $\mathcal{H}$ of combinatorial structures, their variation distance is defined as:* $|\mu - \nu| = \frac{1}{2}\sum_{H\in\mathcal{H}} |\Pr_\mu(H) - \Pr_\nu(H)|$

▶ **Lemma 34.** *If two distributions μ and ν over a finite family $\mathcal{H}$ of combinatorial structures satisfy $|\mu - \nu| \leq \delta$, then for any set $A \subset \mathcal{H}$ we have $|\Pr_\mu(A) - \Pr_\nu(A)| \leq \delta$*

▶ **Lemma 35.** *Suppose that μ and ν are two probability distributions over graphs with set of vertices $\{v_1, \ldots, v_q\}$, where each edge v_iv_j is independently chosen to be an edge with probability $\mu_{i,j}$ and $\nu_{i,j}$ respectively. If $|\mu_{i,j} - \nu_{i,j}| \leq \varepsilon/\binom{q}{2}$ for every $1 \leq i < j \leq q$, then the variation distance between μ and ν is bounded by ε.*

▶ **Definition 36** (q-statistic)**.** *The q-statistic of a graph G is the probability distribution over all (labeled) graphs with q vertices that result from picking at random q distinct vertices of G and considering the induced subgraph. For a given graph H we denote the probability for obtaining H when drawing a graph according to the q-statistic by $\Pr_G(H)$.*

▶ **Definition 37.** *For an equipartition $A = \{V_1, \ldots, V_t\}$ of G, and a signature $S = (\eta_{i,j})_{1\leq i<j\leq t}$ of A, the perceived q-statistic according to S is the following distribution $\Pr_S$ over labeled graphs with q vertices $v_1, \ldots, v_q$. Start by choosing a uniformly random sequence without repetitions of indices $i_1, \ldots, i_q$ from $1, \ldots, t$. Then, independently, take every v_kv_l for $k < l$ to be an edge with probability η_{i_k,i_l} if $i_k < i_l$ and with probability η_{i_l,i_k} if $i_l < i_k$. Then $\Pr_S(H)$ is defined as the probability that the resulting labeled graph equals H.*

The following lemma will replace Lemma 1 in the proof of Theorem 9.

▶ **Lemma 38** (see [22]). *If there is an ε-test for a graph property $\mathcal{P}$ that makes $Q = Q(\varepsilon)$ edge queries, then there exists an appropriate family $\mathcal{H}$ of labeled graphs on $q = 2Q$ vertices such that any graph G which satisfies $\mathcal{P}$, satisfies also $\Pr_G(\mathcal{H}) \geq \frac{2}{3}$, and any graph G that is ε-far from satisfying $\mathcal{P}$, satisfies also $\Pr_G(\mathcal{H}) < \frac{1}{3}$.*

We now introduce a variant of Lemma 25 that is suited for the proof of Theorem 9. The proof of the lemma will appear in the journal version of the paper.

▶ **Lemma 39.** *For every q, ε there are $\gamma = \gamma_{39}(q,\varepsilon)$, $r = r_{39}(q,\varepsilon)$ so that*

$$\gamma = \mathrm{poly}(\varepsilon \cdot 2^{-q^2}), \quad r = \mathrm{poly}(1/\varepsilon \cdot 2^{q^2})$$

and for every γ-signature S of a γ-FK-regular equipartition A into $t \geq r$ sets, of a graph G on $n \geq N_{39}(q,\varepsilon,t) = \mathrm{poly}(t/\varepsilon)2^{\mathrm{poly}(q)}$ vertices, we have $|\Pr_S - \Pr_G| \leq \varepsilon$, where $\Pr_G$ is the q-statistic and $\Pr_S$ is the perceived q-statistic according to S.

We now introduce a variant of Lemma 15 that is suited for the proof of Theorem 9. The proof of the lemma will appear in the journal version of the paper.

▶ **Lemma 40.** *For every q and ε there exist $\gamma = \gamma_{40}(q,\varepsilon)$, $s = s_{40}(q,\varepsilon)$ and $f_{40}^{(q,\varepsilon)} : \mathbb{N} \to \mathbb{N}$, such that*

$$\gamma = \mathrm{poly}(\varepsilon \cdot 2^{-q^2}), \quad s = \mathrm{poly}\Big(\frac{2^{q^2}}{\varepsilon}\Big), \quad f_{40}^{(q,\varepsilon)}(x) = x \cdot 2^{\mathrm{poly}\left(\frac{2^{q^2}}{\varepsilon}\right)}$$

with the following property. For every family $\mathcal{H}$ of graphs with q vertices, there exists a deterministic algorithm, that receives as an input a γ-signature S of an (f,γ)-final partition A into $t \geq s$ sets of a graph G with $n \geq N_{40}(q,\varepsilon,t) = t \cdot 2^{\mathrm{poly}(1/\varepsilon)\cdot 2^{\mathrm{poly}(q)}}$ vertices and distinguishes given any α between the following two cases:

- **(i)** *G is $(\alpha-\varepsilon)$-close to some graph G' for which $\Pr_{G'}(\mathcal{H}) \geq \frac{2}{3}$.*
- **(ii)** *G is α-far from every G' for which $\Pr_{G'}(\mathcal{H}) \geq \frac{1}{3}$.*

Theorem 9 is derived from Lemmas 14 and 40, similarly to how Theorem 4 is derived from Lemmas 14 and 15. This will appear in the journal version of the paper.

References

1. N. Alon, B. Chazelle, S. Comandur, and D. Liue. Estimating the distance to a monotone function. *Random Struct Algorithms*, 31:371–383, 2007.
2. N. Alon, E. Fischer, M. Krivelevich, and M. Szegedy. Efficient testing of large graphs. *Combinatorica*, 20:451–476, 2000.
3. N. Alon, E. Fischer, I. Newman, and A. Shapira. A combinatorial characterization of the testable graph properties: it's all about regularity,. *SIAM J Comput*, 39:143–167, 2009.
4. N. Alon and J. Fox. Easily testable graph properties. *Combin Probab. Comput*, 24:646–657, 2015.
5. N. Alon and A. Shapira. A characterization of the (natural) graph properties testable with one-sided error. *SIAM J Comput.*, 37:1703–1727, 2008.
6. T. Batu, F. Ergun, J. Kilian, A. Magen, S. Raskhodnikova, R. Rubinfeld, and R. Sami. A sublinear algorithm for weakly approximating edit distance. *ACM Comput Surv.*, 35:316–324, 2003.
7. T. Batu, L. Fortnow, R. Rubinfeld, W. Smith, and P. White. Testing closeness of discrete distributions. *Journal of the ACM*, 60:1–25, 2013.
8. P. Berman, M. Murzabulatov, and S. Raskhodnikova. Tolerant testers of image properties. *Proc. of ICALP*, pages 1–14, 2016.

9 E. Blais, C. Canonne, T. Eden, A. Levi, and D. Ron. Tolerant junta testing and the connection to submodular optimization and function isomorphism. *ACM Trans Comput. Theory*, 11, 2019.

10 C. Borgs, J. Chayes, L. Lovász, V. T. Sós, B. Szegedy, and K. Vesztergombi. Graph limits and parameter testing. *Proc. of STOC*, pages 261–270, 2006.

11 A. Campagna, A. Guo, and R. Rubinfeld. Local reconstructors and tolerant testers for connectivity and diameter. *Proc. of APPROX*, pages 411–424, 2013.

12 D. Conlon and J. Fox. Bounds for graph regularity and removal lemmas. *Geom Funct. Anal*, 22:1191–1256, 2012.

13 T. Eden, R. Levi, and D. Ron. Testing bounded arboricity. *Proc. of SODA*, pages 2081–2092, 2018.

14 N. Fiat and D. Ron. On efficient distance approximation for graph properties. *Proc. of SODA*, pages 1618–1637, 2021.

15 E. Fischer and L. Fortnow. Tolerant versus intolerant testing for boolean properties. *Theory Comput.*, 2:173–183, 2006.

16 E. Fischer and I. Newman. Testing versus estimation of graph properties. *SIAM J Comput.*, 37:482–501, 2007.

17 A. Frieze and R. Kannan. The regularity lemma and approximation schemes for dense problems. *Proc. of FOCS*, pages 12–20, 1996.

18 A. Frieze and R. Kannan. Quick approximation to matrices and applications. *Combinatorica*, 19:175–220, 1999.

19 L. Gishboliner and A. Shapira. Removal lemmas with polynomial bounds. *Proc. of STOC*, pages 510–522, 2017.

20 O. Goldreich. *Introduction to Property Testing*. Cambridge University Press, 2017.

21 O. Goldreich, S. Goldwasser, and D. Ron. Property testing and its connection to learning and approximation. *Journal of the ACM*, 45:653–750, 1998.

22 O. Goldreich and L. Trevisan. Three theorems regarding testing graph properties. *Random Struct Algorithms*, 23:23–57, 2003.

23 V. Guruswami and A. Rudra. Tolerant locally testable codes. *Proc. of RANDOM*, pages 306–317, 2005.

24 C. Hoppen, Y. Kohayakawa, R. Lang, H. Lefmann, and H. Stagni. Estimating parameters associated with monotone properties,. *Combin. Probab. Comput.*, 29(2020):616–632, 2016.

25 C. Hoppen, Y. Kohayakawa, R. Lang, H. Lefmann, and H. Stagni. On the query complexity of estimating the distance to hereditary graph properties. *SIAM J Discret. Math.*, 35:1238–1251, 2021.

26 S. Kopparty and S. Saraf. Tolerant linearity testing and locally testable codes. *Proc. of RANDOM*, pages 601–614, 2009.

27 L. Lovász and B. Szegedy. Szemerédi's lemma for the analyst. *Geom. Funct. Anal*, 17:252–270, 2007.

28 S. Marko and D. Ron. Distance approximation in bounded-degree and general sparse graphs. *ACM Trans. Algorithms*, 5:22:1–22:28, 2009.

29 G. Moshkovitz and A. Shapira. A sparse regular approximation lemma. *Trans. Amer. Math. Soc.*, 371:6779–6814, 2019.

30 M. Parnas, D. Ron, and R. Rubinfeld. Tolerant property testing and distance approximation. *J. Comput. Syst. Sci.*, 72:1012–1042, 2006.

31 V. Rödl and R. Duke. On graphs with small subgraphs of large chromatic number. *Graphs and Combinatorics*, 1:91–96, 1985.

32 V. Rödl and M. Schacht. Generalizations of the removal lemma. *Combinatorica*, 29:467–501, 2009.

33 V. Rödl and M. Schacht. Regularity lemmas for graphs. *Fete of Combinatorics and Computer Science, vol. 20 series, Bolyai Soc Math. Stud*, pages 287–325, 2010.

34 A. Shapira and H. Stagni. A tight bound for testing partition properties. 2023.

35 E. Szemerédi. Regular partitions of graphs, 1978. In *Proc. Colloque Inter CNRS (J. C. Bermond, J. C. Fournier, M. Las Vergnas and D. Sotteau, eds.)*.