Testing versus estimation of graph properties[*][†]

Eldar Fischer [‡] Ilan Newman [§]

May 7, 2016

Abstract

Tolerant testing is an emerging topic in the field of property testing, which was defined in [15] and has recently become a very active topic of research. In the general setting, there exist properties that are testable but are not tolerantly testable [12]. On the other hand, we show here that in the setting of the dense graph model, all testable properties are not only tolerantly testable (which was already implicitly proven in [2] and [14]), but also admit a constant query size algorithm that estimates the distance from the property up to any fixed additive constant.

In the course of the proof we develop a framework for extending Szemerédi's Regularity Lemma, both as a prerequisite for formulating what kind of information about the input graph will provide us with the correct estimation, and as the means for efficiently gathering this information. In particular, we construct a probabilistic algorithm that finds the parameters of a regular partition of an input graph using a constant number of queries, and an algorithm to find a regular partition of a graph using a TC_0 circuit. This, in some ways, strengthens the results of [1].

1 Introduction

Combinatorial property testing deals with the following task: For a fixed $\epsilon > 0$ and a fixed property $\mathcal{P}$, distinguish using as few queries as possible (and with probability at least $\frac{2}{3}$) between the case that an input of size m satisfies $\mathcal{P}$, and the case that the input is ϵ-far from satisfying $\mathcal{P}$. In our context the inputs are boolean, and the distance from $\mathcal{P}$ is measured by the minimum number of bits that have to be modified in the input in order to make it satisfy $\mathcal{P}$, divided by the input size m. For the purpose here we are mainly interested in tests that have a number of queries that depends only on the approximation parameter ϵ and is independent of the input size. Properties that admit such algorithms are called *testable*.

[*] Research supported in part by an Israel Science Foundation grant number 55/03.

[†] A preliminary version of this paper appeared in the proceedings of the 37th ACM STOC (2005).

[‡] Faculty of Computer Science, Technion – Israel institute of technology, Technion City, Haifa 32000, Israel. Email: eldar@cs.technion.ac.il

[§] Department of Computer Science, Haifa University, Haifa 31905, Israel. Email: ilan@cs.haifa.ac.il

The first time a question formulated in terms of property testing was considered is by Blum, Luby and Rubinfeld [7], and the general notion of property testing was first formally defined by Rubinfeld and Sudan [17]. The first investigation in the combinatorial context is that of Goldreich, Goldwasser and Ron [13], where the testing of combinatorial graph properties (in the "dense" graph model) is first formalized; their framework will also be the one used here. In recent years the field of property testing has enjoyed rapid growth, as witnessed in the surveys [16] and [10].

One of the main goals in the study of graph property testing is the finding of structural characterization results, or failing that, results that identify large classes of properties that are testable. An example of such a large class is the class of partition properties that was identified in [13]. Other classes were identified as testable using the Regularity Lemma of Szemerédi, in [2] and [9]. The Regularity Lemma is a very useful tool that guarantees the existence of a sort of a "short summary" for graphs with any number of vertices. The price is that the involved constants will not have a practical bound, but for theoretical results this lemma is the most powerful tool up to date for understanding the essence of graph property testing.

The question of providing a complete structural characterization result for the testable graph properties was one of the central themes in the research on testing graph properties. A partial result that in some sense characterizes graph properties that consist of only one graph according to their testability is found in [11], also making use of the Regularity Lemma. Concerning 1-sided graph property testing, where the algorithm is also required to be independent of the number of vertices of the graph, a recent work of Alon and Shapira [5] approaches what is in essence a full characterization. Very recently a complete characterization of the properties that are testable by 2-sided error tests was provided in [3]. In a different angle of the characterization problem, the canonical testers of [14] can be considered as a first hint that testable graph properties are more than just testable. Here we investigate this further, showing that the class of all testable graph properties (with 1-sided or 2-sided algorithms) is in fact identical to a class of properties that admit algorithms with much stringer requirements than those of property testers.

An investigation that goes beyond the original definition of testable properties was initiated by Parnas, Ron and Rubinfeld [15], concerning tolerant testers. These are property testers that reject all instances that are far enough from the property $\mathcal{P}$, and accept every instance that is close enough to $\mathcal{P}$ (and not just instances that are in $\mathcal{P}$). Recently, Fischer and Fortnow [12] showed that not all testable properties are also tolerantly testable. Here we prove a positive general result on testable graph properties that involves a much tighter concept. We say that a property is (ϵ, δ)-*estimable* if there exists a probabilistic algorithm making a constant number of queries on any input (independently of the input size), that distinguishes with probability $\frac{2}{3}$ between the case that the input is $(\epsilon - \delta)$-close to some input that satisfies the property, and the case that it is ϵ-far from any input satisfying the property. We call a property *estimable* if it is (ϵ, δ)-estimable for every fixed $\epsilon > 0$ and $\delta > 0$. Thus, if a property is estimable, then there exists an $O(1)$-query algorithm that can estimate the relative distance of an input from the property within any fixed additive constant.

Obviously estimability (and also tolerant testing, where we only demand an (ϵ, δ)-estimation for some $\delta > 0$ that may depend on ϵ), is a generalization of the standard testing and the two notions coincide when we take $\delta = \epsilon$.

Our main result is a proof that all testable graph properties are also estimable. Equivalently, we obtain that for every testable property $\mathcal{P}$ and every $\epsilon > 0$, the property of being ϵ-close to $\mathcal{P}$ is in itself testable. For non-graph properties this is not always true, as shown in [12].

While the famed Regularity Lemma of Szemerédi is not very applicable in practice, it is quite important theoretically, and not only for property testing. Alon, Duke, Lefmann, Rödl and Yuster [1] have shown that a regular partition can be found in asymptotically the same time complexity as that of matrix multiplication. For many applications of the Regularity Lemma, one does not need to know the regular partition itself, but only its signature (the pairwise edge densities between its sets). A lemma towards our main result asserts the existence of a randomized algorithm that uses only $O(1)$ queries to the input and approximates the signature of an ϵ-regular partition of a graph to within an additive error of ϵ, for any fixed constant $\epsilon > 0$.

As it turns out, this proof also implies a new algorithm that allows the finding of a regular partition using a very low complexity class algorithm (namely TC_0, as opposed to NC_1 which was previously known from [1]).

The rest of the paper is organized as follows. Section 2 contains the most basic definitions and the formal statement of the main result. Section 3 contains definitions and lemmas concerning Szemerédi's Regularity Lemma, the essence of property testing algorithms in the dense graph model, and the connection between them. Section 4 contains a framework for extending Szemerédi's Regularity Lemma, leading to the proof of the main result. This proof is based on two main lemmas: One lemma states that knowing the parameters of a certain partition of the graph (which is guaranteed to exist by the extension of Szemerédi's lemma) is enough for knowing how far is the input graph from a graph which a property testing algorithm would accept, and the other lemma states that an approximation of the parameters of such a partition can indeed be calculated with high probability from a small sample of the graph. These two main lemmas are then proven in Section 5 (approximating a partition) and Section 6 (estimating the distance from the property). The final Section 7 contains some concluding comments, including a description of the low complexity algorithm for finding an ϵ-regular partition of a graph.

2 The main result

In the following we formally state our main result. We start with the most basic definition of property testing of graphs (in the "dense model" context).

Definition 1. *We say that two graphs G and G' with the same vertex set of size n are* ϵ-close*, if the number of vertex pairs that form an edge for one of G and G' but not the other does not exceed $\epsilon\binom{n}{2}$. For a property $\mathcal{P}$ of graphs, we say that G is* ϵ-close to $\mathcal{P}$ *if there exists a graph G'*

that satisfies $\mathcal{P}$ and is ϵ-close to G. If there exists no such G' then we say that G is ϵ-far from $\mathcal{P}$*. For properties of combinatorial objects other than graphs, we replace "$\binom{n}{2}$" in the definition above with the size of the corresponding input.*

We call a property ϵ-testable *if there exists a probabilistic algorithm making a constant number of queries on any input (independently of the input size, which is given to the algorithm in advance), that distinguishes with probability $\frac{2}{3}$ between the case that the input satisfies the property, and the case that the input is ϵ-far from any input that satisfies the property. We call a property* testable *if it is ϵ-testable for every fixed constant $\epsilon > 0$.*

Parnas, Ron and Rubinfeld [15] have started investigating properties (of various combinatorial objects and not just graphs) for which there exists a probabilistic algorithm, that apart from being an ϵ-test is also guaranteed to accept (with probability at least $\frac{2}{3}$) any input that is sufficiently close to satisfying the property. In the following we concern ourselves with the strictest possible definition of such properties, in that we want to also accept any input whose distance from the property is only somewhat smaller than the guaranteed rejection distance.

Definition 2. *We call a property* (ϵ, δ)-estimable *if there exists a probabilistic algorithm making a constant number of queries on any input (independently of the input size), that distinguishes with probability $\frac{2}{3}$ between the case that the input is $(\epsilon - \delta)$-close to some input that satisfies the property, and the case that it is ϵ-far from any input satisfying the property. We call a property* estimable *if it is (ϵ, δ)-estimable for every fixed $\epsilon > 0$ and $\delta > 0$.*

We prove that for graph properties (in the dense model), estimation algorithms exist for any property for which there exists a test in the usual sense.

Theorem 2.1. *Every testable property of graphs is also estimable.*

As a corollary from the proofs, we also find an algorithm for constructing an ϵ-regular partition (or a strengthening thereof) of an input graph G using a low complexity (TC_0) algorithm. As the required definitions for stating this result and its motivations are only presented in Section 3 and Section 4, it is discussed in full in Section 7.

3 The building blocks

In this section we prepare some tools that are needed for the following discussion. We define and explain the role of regular partitions, and show their relevance to predicting the behavior of a given testing algorithm when applied to the input graph.

Starting with this section and throughout the paper, we use the convention that a function defined by the statement of a lemma is indexed with the lemma's number. We make no attempt anywhere to minimize the constants involved, and ignore floor and ceiling signs when these make no essential difference for the argument.

For some of the proofs we use the following standard Chernoff-type large deviation inequality (see e.g. [6, Appendix A]).

Lemma 3.1. *Suppose that $X_1, \ldots, X_m$ are m independent Boolean random variables, satisfying $\Pr(X_i = 1) = p_i$. Let $E = \sum_{i=1}^m p_i$. Then, $\Pr(|\sum_{i=1}^m X_i - E| \geq \delta m) \leq 2e^{-2\delta^2 m}$.*

In the following we often use one distribution to approximate another. For this the following is handy.

Definition 3. *Given two distributions μ and ν over a finite family $\mathcal{H}$ of combinatorial structures, their* variation distance *is defined as $|\mu - \nu| = \frac{1}{2}\sum_{H \in \mathcal{H}} |\Pr_\mu(H) - \Pr_\nu(H)|$.*

We note that the variation distance is just a normalized distance in ℓ_1. In particular $0 \leq |\mu - \nu| \leq 1$ for any ν, μ. The importance of this measure is that if $|\mu - \nu|$ is small then ν approximates μ well, as asserted by the following well-known lemma for which we provide a proof for completeness.

Lemma 3.2. *If two distributions μ, ν over a finite family $\mathcal{H}$ of combinatorial structures satisfy $|\mu - \nu| \leq \delta$, then for any set $\mathcal{A} \subseteq \mathcal{H}$ we have $|\Pr_\mu(\mathcal{A}) - \Pr_\nu(\mathcal{A})| \leq \delta$.*

Proof. Set $\mathcal{B} = \mathcal{H} \setminus \mathcal{A}$. Because these are probability spaces we have $\Pr_\mu(\mathcal{B}) - \Pr_\nu(\mathcal{B}) = \Pr_\nu(\mathcal{A}) - \Pr_\mu(\mathcal{A})$. Therefore,

$$\begin{aligned} |\Pr_\mu(\mathcal{A}) - \Pr_\nu(\mathcal{A})| &= \frac{1}{2}|\Pr_\mu(\mathcal{A}) - \Pr_\nu(\mathcal{A})| + \frac{1}{2}|\Pr_\mu(\mathcal{B}) - \Pr_\nu(\mathcal{B})| \\ &\leq \frac{1}{2}\sum_{H \in \mathcal{A}} |\Pr_\mu(H) - \Pr_\nu(H)| + \frac{1}{2}\sum_{H \in \mathcal{B}} |\Pr_\mu(H) - \Pr_\nu(H)| = |\mu - \nu| \end{aligned}$$

■

The following is also a well-known probabilistic lemma that we will use.

Lemma 3.3. *Let μ be a product distribution over $\{0,1\}^k$, where for $(\alpha_1, \ldots, \alpha_k) \in \{0,1\}^k$ we have $\Pr_\mu((\alpha_1, \ldots, \alpha_k)) = \prod_{i=1}^k (p_i)^{\alpha_i}(1 - p_i)^{1-\alpha_i}$ for a fixed sequence $p_1, \ldots, p_k$. Similarly let ν be a product distribution over $\{0,1\}^k$, with $q_1, \ldots, q_k$ replacing $p_1, \ldots, p_k$ in the definition above. Then, $|\mu - \nu| \leq \sum_{i=1}^k |p_i - q_i|$.*

Proof. The proof is by induction on k. For $k = 1$ this is immediate from the definition, and for $k > 1$ we use the definition of the variation distance to reduce it to the expression for $k - 1$, using extensively the simple inequality $|ab - cd| \leq |a - c|b + c|b - d|$ for $a, b, c, d \geq 0$.

$$\begin{aligned}
|\mu-\nu| &= \frac{1}{2}\Bigg(\sum_{\alpha_1,\ldots,\alpha_{k-1}}\Big|p_k\prod_{i=1}^{k-1}(p_i)^{\alpha_i}(1-p_i)^{1-\alpha_i}-q_k\prod_{i=1}^{k-1}(q_i)^{\alpha_i}(1-q_i)^{1-\alpha_i}\Big| \\
&+ \sum_{\alpha_1,\ldots,\alpha_{k-1}}\Big|(1-p_k)\prod_{i=1}^{k-1}(p_i)^{\alpha_i}(1-p_i)^{1-\alpha_i}-(1-q_k)\prod_{i=1}^{k-1}(q_i)^{\alpha_i}(1-q_i)^{1-\alpha_i}\Big|\Bigg) \\
&\le \frac{1}{2}\sum_{\alpha_1,\ldots,\alpha_{k-1}}\Bigg(|p_k-q_k|\prod_{i=1}^{k-1}(p_i)^{\alpha_i}(1-p_i)^{1-\alpha_i} \\
&+ q_k\Big|\prod_{i=1}^{k-1}(p_i)^{\alpha_i}(1-p_i)^{1-\alpha_i}-\prod_{i=1}^{k-1}(q_i)^{\alpha_i}(1-q_i)^{1-\alpha_i}\Big| \\
&+ |(1-p_k)-(1-q_k)|\prod_{i=1}^{k-1}(p_i)^{\alpha_i}(1-p_i)^{1-\alpha_i} \\
&+ (1-q_k)\Big|\prod_{i=1}^{k-1}(p_i)^{\alpha_i}(1-p_i)^{1-\alpha_i}-\prod_{i=1}^{k-1}(q_i)^{\alpha_i}(1-q_i)^{1-\alpha_i}\Big|\Bigg) \\
&= |p_k-q_k|+\frac{1}{2}\sum_{\alpha_1,\ldots,\alpha_{k-1}}\Big|\prod_{i=1}^{k-1}(p_i)^{\alpha_i}(1-p_i)^{1-\alpha_i}-\prod_{i=1}^{k-1}(q_i)^{\alpha_i}(1-q_i)^{1-\alpha_i}\Big| \le \sum_{i=1}^{k}|p_i-q_i|
\end{aligned}$$

■

An immediate corollary of Lemma 3.3 is the following (by taking $k=\binom{q}{2}$).

Lemma 3.4. *Suppose that μ and ν are two probability distributions over graphs with the set of vertices $\{v_1,\ldots,v_q\}$, where each pair v_iv_j is independently chosen to be an edge with probability $\mu_{i,j}$ and $\nu_{i,j}$ respectively. If $|\mu_{i,j}-\nu_{i,j}|\le\epsilon/\binom{q}{2}$ for every $1\le i<j\le q$, then the variation distance between μ and ν is bounded by ϵ.* ■

A crucial notion to the following arguments (as is the case with many other graph property-testing results) is Szemerédi's notion of regularity.

Definition 4. *For two nonempty disjoint vertex sets U and V of a graph G, we define the* density *$d(U,V)$ of the pair to be the number of edges of G between U and V, divided by $|U||V|$.*

We say that the pair U,V is ϵ-regular, *if for any two subsets U' of U and V' of V, satisfying $|U'|\ge\epsilon|U|$ and $|V'|\ge\epsilon|V|$, the edge densities satisfy $|d(U',V')-d(U,V)|<\epsilon$.*

Although Definition 4 bounds the deviation in densities for any two subsets U',V' that are at least as large as their respective thresholds, it is easy to see that it is enough to require the above only for every two subsets U',V' of size exactly $|U'|=\lceil\epsilon|U|\rceil$ and $|V'|=\lceil\epsilon|V|\rceil$.

Regular pairs behave much like random graphs, as seen in the following well-known lemma.

Lemma 3.5 (see e.g. [11, Lemma 4.2] for a proof)**.** *For every k and ϵ there exists $\gamma = \gamma_{3.5}(k, \epsilon)$, so that if $U_1, \ldots, U_k$ are disjoint sets of vertices of G such that every two sets form a γ-regular pair, then the following two distributions for picking a graph H with vertices $v_1, \ldots, v_k$ have variation distance at most ϵ between them.*

- *For every $1 \leq i < j \leq k$, independently take v_iv_j to be an edge of H with probability $d(V_i, V_j)$.*
- *Pick uniformly and independently a vertex $u_i \in U_i$ for every i, and let v_iv_j be an edge of H if and only if u_iu_j is an edge of G.*

What we need is a "cover" of an entire graph with regular pairs. This idea is formalized in the following.

Definition 5. *Given a graph G, an* equipartition *$\mathcal{A} = \{V_1, \ldots, V_k\}$ of G is a partition of its vertex set for which the sizes of any two sets differ by at most 1. An equipartition $\mathcal{B} = \{W_1, \ldots, W_l\}$ is said to be a* refinement *of $\mathcal{A}$ if all of the sets W_i are each fully contained in some set of $\mathcal{A}$.*

*An equipartition $\mathcal{B}$ as above is called ϵ-*regular *if at least $(1 - \epsilon)\binom{l}{2}$ of the pairs W_i, W_j are ϵ-regular.*

Regular partitions are found using the famed Regularity Lemma of Szemerédi [18] (see [8, Chapter 7] for a good exposition of the proof).

Lemma 3.6 (Szemerédi's Regularity Lemma [18])**.** *For every k and ϵ there exists $T = T_{3.6}(k, \epsilon)$, such that for every equipartition $\mathcal{A}$ of a graph G with $n \geq N_{3.6}(k, \epsilon)$ vertices into k sets, there exists a refinement $\mathcal{B}$ of $\mathcal{A}$ into $t \leq T$ sets which is ϵ-regular.*

We now turn to the behavior of property testers when applied to an input graph G. The most important feature of G would be the count of its small induced subgraphs of any kind, as exemplified in the following.

Definition 6. *The q-*statistic *of a graph G is the following probability space over (labeled) graphs with q vertices: Given a labeled graph H with the vertex set $\{v_1, \ldots, v_q\}$, the probability for H is exactly the probability that the edge relation of G, when restricted to a uniformly random sequence of q vertices (without repetitions) $w_1, \ldots, w_k$, is identical to that of H where each w_i plays the role of v_i. Namely, the q-statistic is just the probability distribution over all (labeled) graphs with q vertices that results from picking at random q distinct vertices of G and considering the induced subgraph.*

Given a family $\mathcal{H}$ of graphs with q vertices, we denote the probability for obtaining a member of $\mathcal{H}$ when drawing a graph according to the q-statistic of G by $\Pr_G(\mathcal{H})$.

Note that in the definition above one could work with isomorphic copies of H rather than labeled graphs. This however, brings in the extra complication of having to take into account the automorphism group size of H. When dealing with labeled graphs the analysis is simpler.

The importance of knowing the q-statistic of a graph G is in its close connection with the distance of G from a given testable property, proven in [14].

Lemma 3.7 (Canonical Testers [14]). *If there is an ϵ-test for a graph property $\mathcal{P}$ that makes a constant number of queries, then there exists such a test that makes its queries by choosing uniformly q distinct vertices of G (for an appropriate constant q) and querying the induced subgraph. In particular, there exists an appropriate family $\mathcal{H}$ of labeled graphs such that any graph G that 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 $\Pr_G(\mathcal{H}) \leq \frac{1}{3}$.*

The above motivates us to try deducing the q-statistic of the graph from the densities of one of its regular partitions, as per the following definition.

Definition 7. *For an equipartition $\mathcal{A} = \{V_1, \dots, V_t\}$ of G, a (γ, ϵ)-*signature *of $\mathcal{A}$ is a sequence $\mathcal{S} = (\eta_{i,j})_{1 \leq i < j \leq t}$, such that $|d(V_i, V_j) - \eta_{i,j}| \leq \gamma$ for every $i < j$ but at most $\epsilon\binom{t}{2}$ of the pairs. A (γ, γ)-signature is simply referred to as a γ-*signature*. We use just the term* signature *for $\mathcal{S}$ as above when we do not commit to any specific error parameters.*

Given a signature $\mathcal{S} = (\eta_{i,j})_{1 \leq i < j \leq t}$ as above, the perceived q-statistic *of G according to $\mathcal{S}$ is the following probability distribution over labeled graphs with q vertices: To choose H with the vertex set $v_1, \dots, v_q$, we first choose a uniformly random sequences without repetitions of indices $i_1, \dots, i_q$ from $1, \dots, t$. We then independently take every $v_k v_l$ for $k < l$ to be an edge of H with probability η_{i_k, i_l} if $i_k < i_l$, and with probability η_{i_l, i_k} if $i_k > i_l$.*

Given a family $\mathcal{H}$ of graphs with q vertices, we denote the probability for obtaining a member of $\mathcal{H}$ according to the perceived q-statistic by $\Pr_{\mathcal{S}}(\mathcal{H})$.

The following lemma shows that for a regular partition, the perceived statistic is indeed close to the statistic of the graph.

Lemma 3.8. *For every q and ϵ there exist $\gamma = \gamma_{3.8}(q, \epsilon)$ and $r = r_{3.8}(q, \epsilon)$, so that for every γ-regular partition $\mathcal{A}$ of G into $t \geq r$ sets, where G has $n \geq N_{3.8}(q, \epsilon, t)$ vertices, and every γ-signature $\mathcal{S}$ of $\mathcal{A}$, the variation distance between the perceived q-statistic according to $\mathcal{S}$ and the (actual) q-statistic of G is at most ϵ.*

Proof. Recall Definition 3 for the variation distance between two distributions over a combinatorial structure. Here the structure is the set of labeled graphs on q vertices. The perceived statistic distribution is as defined above, and the actual statistic is as defined by the process of picking a random q-size labeled subgraph of G in Definition 6.

Set $r = 7\binom{q}{2}/\epsilon$ and $\gamma = \min\{\epsilon/7\binom{q}{2}, \gamma_{3.5}(q, \epsilon/7)\}$. Let $v_1, \dots, v_q$ be a uniformly random set of q distinct vertices, and let i_j for every $1 \leq j \leq q$ denote the index for which $v_i \in V_{i_j}$. With probability at least $1 - 4\epsilon/7$, $i_1, \dots, i_q$ are distinct, and moreover all the pairs V_{i_j}, V_{i_k} are γ-regular, and satisfy $|\eta_{i_j, i_k} - d(V_{i_j}, V_{i_k})| \leq \epsilon/7\binom{q}{2}$. Also, note that $\sum_{i=1}^{t} |(|V_i|/n) - 1/t| \leq \epsilon/7$ for an appropriate choice of $N_{3.8}(q, \epsilon, t)$.

Finally, for a specific fixed sequence $i_1, \ldots, i_q$ for which the above event holds, Lemma 3.5 guarantees that the conditional distribution of the induced graph on $v_1, \ldots, v_q$ is not more than $\epsilon/7$-far (in the variation distance) from the distribution on a random graph over $v_1, \ldots, v_q$ for which every edge $v_i v_j$ is independently selected with probability $d(V_{i_j}, V_{i_k})$. Noting that $|d(V_{i_j}, V_{i_k}) - \eta_{\min\{i_j,i_k\},\max\{i_j,i_k\}}| \leq \epsilon/7\binom{q}{2}$ and using Lemma 3.4, it follows that the variation distance between the q-statistic of G and the perceived one is (after summing all the above error terms) at most ϵ. ■

By now we note that knowing an accurate enough signature of a regular partition enables us to estimate the q-statistics of a graph, which in turn enables us to predict the behavior of a property tester (by Lemma 3.7), and thus distinguish between graphs that satisfy the property and graphs which are ϵ-far from satisfying it.

However, for estimability we would like to know more than that. It is not enough to know whether the input graph G has a regular partition that indicates its acceptance by the property tester; we also need to know how far is our input graph G from a graph G' that has such a partition. The problem is that the regular partition that indicates the acceptance of G' may be different from the regular partition found for G. Our technique is to find in G a partition that, in addition to being regular, will be able to "withstand" a repartitioning according to such a G'. This issue, and the issue of efficiently finding a signature for such a partition, are addressed in the next section.

4 Robust and final partitions and proving Theorem 2.1

To prove the main result, we must first define a framework that allows us to extend the notion of regular partitions. To this end let us first delve a little into the details of the proof of the original regularity lemma. We start with the basic function defined in [18] to track graph partitions with respect to their possible regularity.

Definition 8. *For an equipartition $\mathcal{A}$ of a graph G into t sets, its* index $\mathrm{ind}(\mathcal{A})$ *is defined as $t^{-2}\sum_{1\leq i<j\leq t} d^2(V_i, V_j)$. For a function $f : \mathbb{N} \to \mathbb{N}$ and a constant γ, we say that $\mathcal{A}$ as above is* (f,γ)-robust *if there exists no refinement $\mathcal{B}$ of $\mathcal{A}$ with up to $f(t)$ sets for which* $\mathrm{ind}(\mathcal{B}) \geq \mathrm{ind}(\mathcal{A}) + \gamma$.

The main lemma used in [18] for proving Szemerédi's lemma can be paraphrased as the following (note that in the proof of Lemma 3.6 as presented in [8, Chapter 7], instead of $\mathrm{ind}(\mathcal{A})$ a similar function that is denoted there by "q" is used, and the equipartitions are allowed to have a small number of "exceptional vertices" not in any set).

Lemma 4.1 ([18], see also [8, Lemma 7.2.4]). *For every ϵ there exist $\gamma = \gamma_{4.1}(\epsilon)$ and $f = f^{(\epsilon)}_{4.1} : \mathbb{N} \to \mathbb{N}$, such that every (f,γ)-robust partition is also ϵ-regular.*

In the original formulation of [18], it is proven that a non-ϵ-regular partition into t sets has a refinement into $\max\{\exp(t), \exp(1/\epsilon)\}$ many sets whose index is larger by at least some $\mathrm{poly}(\epsilon)$

(without explicitly stating Lemma 4.1). With either formulation, the move from Lemma 4.1 to Lemma 3.6 is made through the following simple observation.

Observation 4.2. *For every k, γ and $f : \mathbb{N} \to \mathbb{N}$ there exists $T = T_{4.2}(k, \gamma, f)$, such that every equipartition $\mathcal{A}$ of a graph G with $n \geq N_{4.2}(k, \gamma, f)$ vertices into at most k sets, has a refinement $\mathcal{B}$ into at most T sets that is (f, γ)-robust.*

Proof. We start by setting $\mathcal{B} = \mathcal{A}$, but if it is not (f, γ)-robust then we replace it with the refinement showing this, repeating the procedure as many times as necessary. Since the index of a partition is always between 0 and 1, this process will terminate after at most $1/\gamma$ iterations. ■

In the following we will also consider robust partitions for choices of f that grow faster than what is required for ϵ-regularity. This means that in some sense we will use a strengthening of the original regularity lemma.

The following definition is clearly a strengthening of the definition of robustness.

Definition 9. *For a function $f : \mathbb{N} \to \mathbb{N}$ and a constant γ, we say that $\mathcal{A}$ as above is (f, γ)-*final *if there exists no partition $\mathcal{B}$ (even one that is not a refinement of $\mathcal{A}$) with at least t and up to $f(t)$ sets for which* $\text{ind}(\mathcal{B}) \geq \text{ind}(\mathcal{A}) + \gamma$.

The following is an analogue of Observation 4.2 to final partitions. The price is that now we can no longer demand that the final partition will be a refinement of a given equipartition.

Observation 4.3. *For every k, γ and $f : \mathbb{N} \to \mathbb{N}$ there exists $T = T_{4.3}(k, \gamma, f)$, such that for every graph G with $n \geq N_{4.3}(k, \gamma, f)$ vertices there exists an equipartition $\mathcal{A}$ into at least k and at most T sets that is (f, γ)-final.* ■

In fact we do not need the stronger but less flexible condition of finality for our combinatorial statements, but we use it because the parameters of a final partition are easier to detect than those of a robust one. A testing algorithm can actually compute a signature of a final partition like the one that Observation 4.3 guarantees for a graph G, as the following lemma shows.

Lemma 4.4. *For every k, γ and $f : \mathbb{N} \to \mathbb{N}$ there exists $q = q_{4.4}(k, \gamma, f)$, such that there exists an algorithm that makes up to q queries to a graph G with $n \geq N_{4.3}(k, \frac{1}{2}\gamma, f)$ vertices, computing with probability at least $\frac{2}{3}$ a γ-signature for an (f, γ)-final partition of G into at least k and at most $T_{4.3}(k, \frac{1}{2}\gamma, f)$ sets.*

This lemma is proven in Section 5, and brings us half-way towards our estimability result.

At this point, if from a signature of a regular partition of G we could estimate how far is G from having a regular partition with a different given signature, then we could use it to estimate how far is G from having a statistic that will cause a canonical tester to accept it with high probability. This we cannot do directly, but we can instead estimate such a difference if we are provided with a signature of a partition that is somewhat more than regular, that is, robust with respect to an appropriate function. We explain in Section 6 why a regular partition is not enough while a robust one is. We now present the formal statement of the appropriate result and show how it implies Theorem 2.1.

Lemma 4.5. *For every q and δ there exist $\gamma = \gamma_{4.5}(q, \delta)$, $s = s_{4.5}(q, \delta)$ and $f = f_{4.5}^{(q,\delta)} : \mathbb{N} \to \mathbb{N}$ with the following property. For every family $\mathcal{H}$ of graphs with q (labeled) vertices there exists a deterministic algorithm, that receives as an input only a γ-signature $\mathcal{S}$ for an (f, γ)-robust partition $\mathcal{A}$ with $t \geq s$ sets of a graph G with $n \geq N_{4.5}(q, \delta, t)$ vertices, and distinguishes (using no information on G apart from $\mathcal{S}$ and t) given any ϵ between the case that G is $(\epsilon - \delta)$-close to some graph G' for which $\Pr_{G'}(\mathcal{H}) \geq \frac{2}{3}$, and the case that G is ϵ-far from every graph G' for which $\Pr_{G'}(\mathcal{H}) > \frac{1}{3}$.*

This lemma is proven in Section 6. Lemma 4.4 and Lemma 4.5 together imply the main result.

Proof of Theorem 2.1. Suppose that $\mathcal{P}$ is a testable graph property, and let ϵ and δ be constants for which we want to (ϵ, δ)-estimate $\mathcal{P}$. As $\mathcal{P}$ is in particular $\frac{1}{2}\delta$-testable, Lemma 3.7 asserts that there exists a constant q and a family $\mathcal{H}$ of graphs on q vertices, such that for every graph G that is in $\mathcal{P}$, $\Pr_G(\mathcal{H}) \geq 2/3$, and for any graph G that is $\frac{1}{2}\delta$-far from $\mathcal{P}$, $\Pr_G(\mathcal{H}) \leq 1/3$.

Set $\gamma = \gamma_{4.5}(q, \frac{1}{2}\delta)$, $f = f_{4.5}^{(q,\delta/2)}$, and $k = s_{4.5}(q, \frac{1}{2}\delta)$, and apply the algorithm provided by Lemma 4.4, with the parameters k, γ and f, on the input graph G. This algorithm makes up to $q_{4.4}(k, \gamma, f)$ queries to the graph G, and with probability at least $\frac{2}{3}$ returns a γ-signature $\mathcal{S}$ of an equipartition of G into at least $s_{4.5}(q, \delta)$ sets and at most $T_{4.3}(k, \frac{1}{2}\gamma, f)$ sets that is (f, γ)-final.

We now apply the algorithm that is provided by Lemma 4.5, with parameters q, $\frac{1}{2}\delta$ and $\epsilon - \frac{1}{2}\delta$, to the signature $\mathcal{S}$ (remember that this is a deterministic algorithm making no additional queries). Due to the choice of parameters, it is guaranteed by Lemma 4.5 that we can distinguish between the case that there is a graph G' that is $(\epsilon - \delta)$-close to G and for which $\Pr_{G'}(\mathcal{H}) \geq \frac{2}{3}$, and the case that G is $(\epsilon - \frac{1}{2}\delta)$-far from every graph G' for which $\Pr_{G'}(\mathcal{H}) > \frac{1}{3}$. In the first case G is accepted, and in the second case it is rejected.

For the above to work we require that $n \geq \max\{N_{4.3}(k, \frac{1}{2}\gamma, f), N_{4.5}(q, \frac{1}{2}\delta, T_{4.3}(k, \frac{1}{2}\gamma, f))\}$. For a smaller n we can just read the entire input and compute its distance from the property to be estimated. We now claim that the above algorithm is indeed an (ϵ, δ)-estimation algorithm for $\mathcal{P}$ for every large enough n.

If G is $(\epsilon - \delta)$-close to $\mathcal{P}$, then by the premises above, it is also $(\epsilon - \delta)$-close to a graph G' for which $\Pr_{G'}(\mathcal{H}) \geq \frac{2}{3}$, and so the first case above will hold as long as $\mathcal{S}$ is in fact a γ-signature of an (f, γ)-robust partition of G, which happens with probability at least $\frac{2}{3}$. Thus G is accepted with probability at least $\frac{2}{3}$.

On the other hand, if G is ϵ-far from $\mathcal{P}$, then by the triangle inequality it is $(\epsilon - \frac{1}{2}\delta)$-far from any graph G' for which $\Pr_{G'}(\mathcal{H}) > \frac{1}{3}$ (because such a G' would be $\frac{1}{2}\delta$-close to satisfying $\mathcal{P}$, as q was chosen to suffice for testing $\mathcal{P}$ with distance parameter $\frac{1}{2}\delta$). Thus, if $\mathcal{S}$ is indeed a γ-signature of an (f, γ)-robust partition, then the algorithm rejects G, and this again happens with probability at least $\frac{2}{3}$.

With both cases covered, the proof is concluded. ■

5 Proof of Lemma 4.4

Our strategy as outlined here is rather simple. Let k, γ and f be as in the formulation of the lemma, and set $T = T_{4.3}(k, \gamma/2, f)$. For every $s \in \{k, \ldots, f(T)\}$ we quantize all possible signatures of equipartitions into s parts, choosing such a finite family of signatures so that every possible signature of an s-partition is close enough to one of the chosen signatures. For every such chosen signature we test whether there exists a partition into s sets with densities that are as determined by the signature, allowing for a small slack. This is done using the test of Goldreich, Goldwasser and Ron for generalized graph partitions [13]. For every positive answer (namely, that such a partition exists) we record the signature and estimate the index of the partition. Having all this information, we set for every s the quantity $M(s)$ that is the largest index of any of the partitions into s sets that we (approximately) know about. We then set s^* to be such that for every s for which $s^* \leq s \leq f(s^*)$, the records indicate that $M(s) \leq M(s^*) + \frac{3}{4}\gamma$. Finally, we output the signature that achieves $M(s^*)$, and claim that it is a signature of an (f, γ)-final equipartition.

To see that such an s^* indeed exists, consider the $(f, \frac{1}{2}\gamma)$-final equipartition $\mathcal{A}$ that is guaranteed by Observation 4.3, for k, γ and f. $\mathcal{A}$ is a partition into $b \leq T$ sets with some signature $\mathcal{S}$. Thus, while passing through all possible signatures of equipartitions into b sets in the process above, the closest signature to $\mathcal{S}$ must have been considered and the corresponding index, which is a good approximation of $\operatorname{ind}(\mathcal{A})$, was computed. Now, as $\mathcal{A}$ is $(f, \frac{1}{2}\gamma)$-final, it follows by the definitions that $s^* = b$ is a valid answer to the output above, assuming that all the index estimations are good enough. Let us now proceed with the formal details.

Set $\epsilon = \gamma/(24 \cdot f^2(T))$. We assume that ϵ^{-1} is an integer without loss of generality, as otherwise we can decrease it a little more (by a factor of less than 2) without changing the essence of the arguments. For every $k \leq s \leq f(T)$ set $S(s, \epsilon) = \{0, \epsilon, 2\epsilon, \ldots, 1\}^{\binom{s}{2}}$. Every $\mathcal{S} \in S(s, \epsilon)$ is clearly associated with a signature of a possible equipartition of G into s sets.

As we only have signatures to work with, we have to use them to estimate the index of a partition.

Definition 10. *In an analogue manner to the definition of the index of a partition, we define the* index *of a signature $\mathcal{S} = (\eta_{i,j})_{1 \leq i < j \leq t}$ to be $\operatorname{ind}(\mathcal{S}) = t^{-2} \sum_{1 \leq i < j \leq t} (\eta_{i,j})^2$*

Following is an obvious observation (by a simple calculation) that relates the index of any ϵ-signature of a partition to the index of the partition.

Observation 5.1. *Let $\mathcal{A}$ be an equipartition into s sets and assume that $\mathcal{S} = (\eta_{i,j})_{1 \leq i < j \leq s}$ is an ϵ-signature of $\mathcal{A}$. Then $|\operatorname{ind}(\mathcal{A}) - \operatorname{ind}(\mathcal{S})| \leq 3\epsilon$.* ■

Let G be a graph with n vertices and let s be fixed. Let $0 \leq \alpha_{i,j} < \beta_{i,j} \leq 1$, $1 \leq i < j \leq s$ be two sequences of numbers. The following is a special case of a theorem proved by Goldreich, Goldwasser and Ron [13] (in [13], there are lower and upper bounds on the sizes of the vertex sets too, but having them here does not make an essential difference).

Lemma 5.2 (GGR-test of graph partitions [13])**.** *For a fixed s, let $\mathcal{P}$ be the property of a graph G with n vertices having an equipartition $V_1, \ldots, V_s$ of its vertex set, such that $\alpha_{i,j} \leq d(V_i, V_j) \leq \beta_{i,j}$ for every $1 \leq i < j \leq s$ (for fixed, given $\alpha_{i,j} < \beta_{i,j}$).*

Property $\mathcal{P}$ is testable, with a number of queries that is polynomial in ϵ (for every fixed s) and is independent of n.

We use the following guarantee on the approximation of a signature given by a GGR-test.

Lemma 5.3. *Let $s \geq 2/\epsilon$ be fixed, let $\mathcal{S} = (\eta_{i,j})_{1 \leq i < j \leq s}$ be a signature, and let $\alpha = (\alpha_{i,j})_{1 \leq i < j \leq s}$ and $\beta = (\beta_{i,j})_{1 \leq i < j \leq s}$ be defined by $\alpha_{i,j} = \eta_{i,j} - \epsilon$ and $\beta_{i,j} = \eta_{i,j} + \epsilon$ for $1 \leq i < j \leq s$. Then applying the* GGR*-test on a graph G with s, α, β and distance parameter ϵ results in the following.*

- *If the test accepts with probability more than $\frac{1}{3}$, then there exists an equipartition $\mathcal{A}$ of G into s sets for which $\mathcal{S}$ is an $s^2\epsilon$-signature.*
- *If there is an equipartition $\mathcal{A}$ of G into s sets for which $\mathcal{S}$ is an $(\epsilon, 0)$-signature, then the test accepts with probability at least $\frac{2}{3}$.*

Proof. The first thing to note is that the GGR-property to be tested is exactly the property that $\mathcal{S}$ is an $(\epsilon, 0)$-signature for some partition of G. This immediately yields the second item in the assertion of the lemma.

For the first item, assume that the test accepts with probability more than $\frac{1}{3}$ when applied with s, α and β. Then there must be a graph G' that is ϵ-close to G and that has an equipartition $\mathcal{A}$ for which $\mathcal{S}$ is an $(\epsilon, 0)$-signature. Thus $\mathcal{A}$, considered as an equipartition of G, must have $|d_G(V_i, V_j) - d_{G'}(V_i, V_j)| \leq \frac{1}{2}s^2\epsilon$ for every $1 \leq i < j \leq s$ (as otherwise G' will be more than ϵ-far from G), and so $\mathcal{S}$ is an $s^2\epsilon$-signature for G'. ■

We are now ready to conclude this section.

Proof of Lemma 4.4. Suppose that the parameters f, γ and k are given, and set $T = T_{4.3}(k, \frac{1}{2}\gamma, f)$. For $s \in \{k, \ldots, f(T)\}$, let ϵ and $S(s, \epsilon)$ be as defined above, and let $m = \sum_{s=k}^{f(T)} \epsilon^{-\binom{s}{2}}$ be the total number of members in the union of all $S(s, \epsilon)$ for $k \leq s \leq f(T)$.

We use the following procedure for every $s \in \{k, \ldots, f(T)\}$.

- Initialize $M(s) = 0$. This variable will contain the supposed maximum index of any equipartition into s sets.

- for every $\mathcal{S} = (\eta_{i,j})_{1\le i<j\le t} \in S(s,\epsilon)$, define α and β by $\alpha_{i,j} = \eta_{i,j} - \epsilon$ and $\beta_{i,j} = \eta_{i,j} + \epsilon$ for $1 \le i < j \le s$ (just as in Lemma 5.3).

  Apply the GGR-test on G with parameters α, β and distance parameter ϵ for $100 \log m$ times. If the majority of the runs accept then we say that $\mathcal{S}$ was *accepted*. In this case we take $\max\{M(s), \text{ind}(\mathcal{S})\}$ to be the new value of $M(s)$, and record the signature $\mathcal{S}$ if it is the one for which this maximum is obtained. If the test rejects on the majority of the runs then we do nothing, and say that $\mathcal{S}$ was *rejected*.

Note that in the second step above we need to go over all signatures $\mathcal{S} \in S(s,\epsilon)$. It is not hard to generate and go over them in a lexicographic order.

Let s^* be the smallest number in $\{k,\dots,T\}$ such that $M(s^*) + \frac{3}{4}\gamma \ge M(s')$ for every $s' \in \{s^*+1,\dots,f(s^*)\}$. If there exists such an s^*, output the signature $\mathcal{S}^*$ that achieves the maximum for s^*. Otherwise, the algorithm fails.

It is clear that the algorithm above uses a constant number of queries (on account of using a constant number of GGR-tests). We now need to show that with probability at least $\frac{2}{3}$, the algorithm indeed produces a γ-signature of an (f,γ)-final partition of G into at least k and at most T sets. We conclude the proof using the following claims.

Claim 5.4. *With probability at least $\frac{2}{3}$ the following holds. For every $s \in \{k,\dots,f(T)\}$ and every $\mathcal{S} \in S(s,\epsilon)$ which the algorithm accepted, there is an equipartition $\mathcal{A}_{\mathcal{S}}$ into s sets, with $|\text{ind}(\mathcal{A}_{\mathcal{S}}) - \text{ind}(\mathcal{S})| \le 3s^2\epsilon$ and with $\mathcal{S}$ as its $s^2\epsilon$-signature; and for every such s and $\mathcal{S}$ which were rejected by the algorithm, there exists no equipartition $\mathcal{A}_{\mathcal{S}}$ for which $\mathcal{S}$ is an $(\epsilon,0)$-signature.*

Proof. We prove for each of the two parts of the claim that it occurs with probability at least $\frac{5}{6}$, and so it follows that the entire claim holds with probability at least $\frac{2}{3}$. We start with the second part.

Let s and $\mathcal{S}$ be such that $\mathcal{S}$ is an $(\epsilon,0)$-signature for some $\mathcal{A}$. Then by Lemma 5.3 it will be accepted by any one run of the GGR-test (with the corresponding parameters) with probability at least $2/3$. Thus, it will be rejected by the test only if it is rejected by the majority of the $100 \log m$ runs, which by Lemma 3.1 will occur with probability at most $1/(6m)$. Hence, with probability at least $5/6$ the test will accept all such $\mathcal{S}$ as above. This proves that the second part of the claim occurs with probability at least $5/6$.

For the first part of the claim, let us assume now that $\mathcal{S}$ is not an $s^2\epsilon$-signature for any possible equipartition of G into s sets. By Lemma 5.3 this means that every run of the GGR-test will reject $\mathcal{S}$ with probability at least $2/3$. Hence, by Lemma 3.1, the probability that $\mathcal{S}$ is accepted by the majority of the runs is no more than $1/6m$. This implies that with probability at least $5/6$, every signature $\mathcal{S}$ that was accepted by our algorithm is an $s^2\epsilon$-signature of some equipartition $\mathcal{A}_{\mathcal{S}}$ of G into s sets, and then by Observation 5.1 this means that $\mathcal{S}$ and $\mathcal{A}_{\mathcal{S}}$ satisfy $|\text{ind}(\mathcal{A}_{\mathcal{S}}) - \text{ind}(\mathcal{S})| \le 3s^2\epsilon$.

We have proven that each of the parts occurs with probability at least $5/6$, and so the claim that both of them hold with probability at least $2/3$ follows. ■

Claim 5.5. *If the event of Claim 5.4 occurred, then the algorithm succeeds in the following sense: The algorithm does not fail in its last step, and the signature it outputs is an $s^2\epsilon$-signature of some (f,γ)-final partition.*

Proof. We assume that the event of Claim 5.4 indeed occurred, and first show that the algorithm does not fail in the last step.

Set s_1 to be the smallest s for which G has an $(f,\gamma/2)$-final partition into s_1 sets. The fact that such an $s_1 \in \{k,\dots,T\}$ exists is asserted in Observation 4.3. Let $\mathcal{A}$ be the corresponding $(f,\gamma/2)$-final equipartition with the largest index (if there are more than one then let $\mathcal{A}$ be the first one in the lexicographic order of its signature). Then, by the fact that $\mathcal{A}$ is $(f,\gamma/2)$-final, we have that $\operatorname{ind}(\mathcal{A})+\gamma/2 \geq \operatorname{ind}(\mathcal{S})$ for any equipartition $\mathcal{S}$ into at least s_1 and at most $f(s_1)$ sets. Also by our choice of $\mathcal{A}$ we have $\operatorname{ind}(\mathcal{A}) \geq \operatorname{ind}(\mathcal{A}')$ for any equipartition $\mathcal{A}'$ into s_1 sets. Let $\mathcal{S} \in S(s_1,\epsilon)$ be the first in lexicographic order such that $\mathcal{S}$ is an $(\epsilon,0)$-signature of $\mathcal{A}$. Obviously there exists such an $\mathcal{S}$ by the choice of $S(s_1,\epsilon)$.

Thus, assuming that the sampler accepted all signatures which were $(\epsilon,0)$-signatures of a corresponding partition, $\mathcal{S}$ was in particular accepted. By Observation 5.1, together with the fact that $\operatorname{ind}(\mathcal{A}) \geq \operatorname{ind}(\mathcal{A}')$ for any equipartition $\mathcal{A}'$ into s_1 sets, it follows that

$$\operatorname{ind}(\mathcal{A}) - 3s^2\epsilon \leq M(s_1) \leq \operatorname{ind}(\mathcal{A}) + 3s^2\epsilon$$

Moreover, by combining the inequalities above and Observation 5.1, we get that as long as all signatures that were not $s^2\epsilon$-signatures of some equipartition were rejected, the following holds. For any equipartition $\mathcal{B}$ into s sets with $s_1 \leq s \leq f(s_1)$, that has a corresponding $s^2\epsilon$-signature $\mathcal{T} \in S(s,\epsilon)$ that was accepted by the algorithm, we have $\operatorname{ind}(\mathcal{T}) \leq \operatorname{ind}(\mathcal{B}) + 3s^2\epsilon \leq \operatorname{ind}(\mathcal{A}) + \gamma/2 + 3s^2\epsilon \leq M(s_1) + \gamma/2 + 6s^2\epsilon$.

Now this implies that $\operatorname{ind}(\mathcal{T}) \leq M(s_1) + \gamma/2 + 6f(s_1)^2\epsilon \leq M(s_1) + \frac{3}{4}\gamma$ by our choice of $\epsilon = \gamma/24f(T)^2$. Thus s_1 is recognized as a candidate for s^*, and hence the sampler will not fail to output some s^* and $\mathcal{S}^*$ (we do not claim that the sampler actually outputs s_1 as s^*, but only that the existence of s_1 ensures that the algorithm does not fail to output anything in the last step).

It remains to show that if the event of Claim 5.4 occurs and the sampler outputs a signature $\mathcal{S}^*$ with index s^*, then there exists a corresponding (f,γ)-final equipartition. Indeed, this event implies that there exists an equipartition $\mathcal{A}^*$ into s^* sets so that $\mathcal{S}^*$ is its $s^2\epsilon$-signature. This also means that for all $s \in \{s^*,\dots,f(s^*)\}$ and all signatures $\mathcal{S} \in S(s,\epsilon)$, no such signature satisfying $\operatorname{ind}(\mathcal{S}) > M(s^*) + \frac{3}{4}\gamma$ is an $(\epsilon,0)$-signature of any equipartition of G (as these signatures were rejected by the algorithm). Now if $\mathcal{A}^*$ was not (f,γ)-final, then there would be an equipartition $\mathcal{B}$ with $s \in \{s^*,\dots,f(s^*)\}$ sets for which $\operatorname{ind}(\mathcal{B}) \geq \operatorname{ind}(\mathcal{A}^*) + \gamma \geq M(s^*) + \gamma - 3(s^*)^2\epsilon$. But if we set $\mathcal{S}$ to be an $(\epsilon,0)$-signature of $\mathcal{B}$ (by approximating each pair density of $\mathcal{B}$ by its closest multiple of ϵ). This would imply, by Observation 5.1, that $\operatorname{ind}(\mathcal{S}) \geq M(s^*) + \gamma - 3(s^*)^2\epsilon - 3\epsilon > M(s^*) + \frac{3}{4}\gamma$, a contradiction since such an $\mathcal{S}$ (which would have been accepted by the algorithm) means that $\mathcal{S}^*$ would not be a valid output. ■

To summarize, by Claim 5.4, with probability at least $\frac{2}{3}$ the sampler accepts all signatures under consideration that are $(\epsilon, 0)$-signatures of some corresponding equipartition, and rejects all signatures that are not $s^2\epsilon$-signatures of any equipartition. Then, by Claim 5.5, whenever this event occurs the algorithm will output without fail a γ-signature for some (f, γ)-final equipartition. Together this means that with probability at least $\frac{2}{3}$ the algorithm will supply the desired output, concluding the proof of Lemma 4.4. ■

6 Proof of Lemma 4.5

By Lemma 3.8 (using Lemma 3.7 about canonical testing), if we know a signature of a regular partition of a graph G, then this is enough to distinguish whether the graph satisfies a given testable property, or is δ-far from satisfying it. For estimability we would like to go a step further, and use a signature of G to approximate its distance from any graph G' that the δ-test may accept.

However, knowing just the signature of a regular partition of G is insufficient, since regular partitions of two graphs of small relative distance might still be quite different (and have quite different signatures). Thus, if G does not satisfy a testable property, but is close to satisfying it as witnessed by a graph G', then a regular partition of G with a corresponding signature may still not provide us with information about the regular partition of G' and thus about the distance of G from the property. Instead, our strategy will be to ask for a signature of a partition $\mathcal{A}$ of G, that is robust enough to ensure that G' will have a regular partition that is a refinement of $\mathcal{A}$ which is still regular for G. With this setting, we will also be able to calculate a signature in G for the new partition of G', using only the signature of $\mathcal{A}$ in G. This will enable us to compare possible signatures for estimating the distance between G and the hypothetical G'.

We now turn to the formal proof. We need first some definitions about distances of signatures, and about how signatures behave under refinements of equipartitions.

Definition 11. *The* distance *between the signatures* $\mathcal{S} = (\eta_{i,j})_{1\le i<j\le t}$ *and* $\mathcal{S}' = (\eta'_{i,j})_{1\le i<j\le t}$ *is defined as the average density difference* $\sum_{1\le i<j\le t} |\eta_{i,j} - \eta'_{i,j}| / \binom{t}{2}$.

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

The following follows directly from the above definition (for any equipartition, disregarding the regularity conditions).

Observation 6.1. *For every ϵ and s there exist $r = r_{6.1}(\epsilon)$ and $N = N_{6.1}(\epsilon, s)$ satisfying the following. Suppose that G and G' are α-close graphs on the same vertex set of size $n \geq N$, and that $\mathcal{S}$ and $\mathcal{S}'$ are γ and γ' signatures respectively, of the same equipartition $\mathcal{A}$ of the vertex set of G and G' into $s \geq r$ sets. Then the distance between $\mathcal{S}$ and $\mathcal{S}'$ is at most $\alpha + \epsilon + 2(\gamma + \gamma')$.*

Proof sketch. Setting $r = 2/\epsilon$, it is clear that for n large enough the 0-signatures (i.e. the sequences of actual densities) of $\mathcal{A}$ over G and G' differ by no more than $\alpha + \epsilon$. Also, it is not hard to see that the 0-signature and any γ-signature of $\mathcal{A}$ over G differ by no more than 2γ, and similarly the 0-signature and any γ'-signature of $\mathcal{A}$ over G' differ by no more than $2\gamma'$. We conclude the proof using the triangle inequality. ■

Given a signature for a regular partition of G, we can use it to bound the distance of G from some other graph that shares the same regular partition.

Lemma 6.2. *For every ϵ and t there exists $\gamma = \gamma_{6.2}(\epsilon)$ and $N = N_{6.2}(t, \epsilon)$, such that for every graph G with $n \geq N$ vertices, if $\mathcal{S}$ is a γ-signature of a γ-regular partition $\mathcal{A}$ of G with t sets, then for every signature $\mathcal{S}'$ that is δ-close to $\mathcal{S}$ for some δ, there is a graph G' (with the same vertex set) that is $(\delta + \epsilon)$-close to G, so that $\mathcal{A}$ is an ϵ-regular partition of G', and $\mathcal{S}'$ is an ϵ-signature thereof.*

Before we continue, we note that the converse is false, as there could be two graphs that share exactly the same signature but are quite far. For example, two graphs chosen uniformly at random from the set of all graphs with a fixed labeled set of n vertices will be with high probability far from each other, and still share the same signature for the same regular partition, namely the all-$\frac{1}{2}$ signature.

Proof of Lemma 6.2. We set $\gamma = \frac{1}{4}\epsilon$. Given G, $\mathcal{A} = \{V_1, \ldots, V_t\}$, $\mathcal{S} = (\eta_{i,j})_{1 \leq i < j \leq t}$ and $\mathcal{S}' = (\eta'_{i,j})_{1 \leq i < j \leq t}$, as above, we create G' from G in the following manner.

- For every i, the edges within V_i are unchanged.
- For $i < j$ such that $\eta'_{i,j} < d(V_i, V_j)$, every edge of G between V_i and V_j is removed with probability $1 - \eta'_{i,j}/d(V_i, V_j)$, independently of all other probabilistic actions in this construction.
- For $i < j$ such that $\eta'_{i,j} > d(V_i, V_j)$, every vertex pair of G between V_i and V_j that is not an edge becomes one with probability $1 - (1 - \eta'_{i,j})/(1 - d(V_i, V_j))$, independently of all other probabilistic actions in this construction.

Let G' be the resulting graph. For every $X \subseteq V_i$, $Y \subseteq V_j$ let $d'(X, Y) = d_{G'}(X, Y)$ be the pairwise density with regards to G' (Definition 4). We choose $N > 8t^4/(\gamma^3)$. Making extensive use of Lemma 3.1, we now prove two claims. We first prove that with high probability we will get in G' the correct densities.

Claim 6.3. *For every $1 \leq i < j \leq t$, $|d'(V_i, V_j) - \eta'_{i,j}| > 2\gamma$ with probability at most $1/(2t^2)$.*

Proof. Suppose first that $\eta'_{i,j} < d(V_i, V_j)$. Then, we have $m = d(V_i, V_j) \cdot (n/t)^2$ edges, where each edge is now removed with probability $p = 1 - \eta'_{i,j}/d(V_i, V_j)$ (independently of other edges). Note that the expected number of removed edges is $E = (d(V_i, V_j) - \eta'_{i,j}) \cdot (n/t)^2$ and thus the expected value of $d'(V_i, V_j)$ is exactly $\eta'_{i,j}$. Hence for the event above to occur, the deviation of the number of edges removed from E has to be more than $2\gamma \cdot (n/t)^2$. Now, if $d(V_i, V_j) > 2\gamma$ then m is large enough (assuming that n is large enough) for Lemma 3.1 to ensure that the probability that the deviation above is more than $\gamma \cdot (n/t)^2$ is below the claimed bound and thus imply the statement. For $d(V_i, V_j) < 2\gamma$ the number of removed edges is at most $d(V_i, V_j)$ and thus the event above occurs with probability 1. If $\eta'_{i,j} > d(V_i, V_j)$ then the argument is analogous so we omit it here. ■

Note that if $|d'(V_i, V_j) - \eta'_{i,j}| \leq 2\gamma$ for a pair (i, j), then we have $|d'(V_i, V_j) - d(V_i, V_j)| \leq |d'(V_i, V_j) - \eta'_{i,j}| + |\eta'_{i,j} - \eta_{i,j}| + |\eta_{i,j} - d(V_i, V_j)| \leq 2\gamma + |\eta'_{i,j} - \eta_{i,j}| + |\eta_{i,j} - d(V_i, V_j)|$, and by the assumption on the distance between $\mathcal{S}$ and $\mathcal{S}'$ we also know that $\sum_{1 \leq i < j \leq t} |\eta'_{i,j} - \eta_{i,j}| \leq \binom{t}{2}\delta$. We now prove a claim about the regularity of the pairs in G'.

Claim 6.4. *For every $i < j$ for which V_i, V_j is a γ-regular pair in G, this will not be an ϵ-regular pair in G' with probability at most* $1/(2t^2)$.

Proof. Again we assume that $\eta'_{i,j} < d(V_i, V_j)$, as the argument for the complement case is analogous. Then, for V_i, V_j not to be ϵ-regular with respect to G' there must be some subsets $X \subseteq V_i$ and $Y \subseteq V_j$ of size $\epsilon n/t$ for which $|d'(X, Y) - d'(V_i, V_j)| > \epsilon$. We call such sets a *violation* at (X, Y). However, since $\mathcal{A}$ is γ-regular for G, we have that $|d(X, Y) - d(V_i, V_j)| \leq \gamma$ (over G). Thus a violation at (X, Y) can occur only when the number of removed edges from $e(X, Y)$ deviates from its expectation by more than $(\epsilon - \gamma) \cdot (\epsilon n/t)^2$. Note also that the number of possible edges between X and Y is $m = d(X, Y) \cdot (\epsilon n/t)^2$.

If $d(V_i, V_j) > 2\gamma = \epsilon/2$ then m is large enough (assuming that n is large enough), for Lemma 3.1 to ensure that the probability that the deviation above is more than $(\epsilon - \gamma) \cdot (\epsilon n/t)^2$ is less than $1/(2t)^2 \cdot 2^{-2n/t}$. Thus, by the union bound, the probability that there exist a pair (X, Y) for which a violation occurs is bounded above by $(1/(2t)^2 \cdot 2^{-2n/t})2^{|V_i|+|V_j|} \leq 1/(2t^2)$ as claimed.

If $d(V_i, V_j) < 2\gamma$ then the number of removed edges is at most $2\gamma(n/t)^2$ and thus a violation at (X, Y) cannot occur at all (recall that no edges are added by our procedure in the case $\eta'_{i,j} < d(V_i, V_j)$. ■

By the analysis above, the union bound (for every $1 \leq i < j \leq t$) implies that there is such a G' for which the assertions of both claims hold simultaneously for every $1 \leq i < j \leq t$. Thus by the statement of Claim 6.4, $\mathcal{S}'$ is a signature for an ϵ-regular partition of G', being an ϵ-signature thereof by the statement of Claim 6.3. In addition, Claim 6.3 implies (as noted right after its proof) that for every pair V_i, V_j, at most a $2\gamma + |\eta'_{i,j} - \eta_{i,j}| + |\eta_{i,j} - d(V_i, V_j)|$ fraction of edges are removed or added while moving from G to G'.

Summing this for all pairs, and recalling that $|\eta_{i,j} - d(V_i, V_j)| \leq \gamma$ for all but a γ-fraction of the pairs (due to $\mathcal{S}$ being a γ-signature of G) as well as that $\mathcal{S}$ and $\mathcal{S}'$ are are δ-close, we get that the total distance between G and G' is bounded by $2\gamma + \delta + (1-\gamma)\gamma + \gamma \cdot 1 \leq \delta + 4\gamma \leq \delta + \epsilon$. ■

In general, even if G' and G are close enough graphs (but not too close), a regular partition of G is not necessarily regular for G'. Instead, we will look at a refinement of the partition of G that is regular for G'. However, a refinement of a regular partition is not necessarily in itself regular, or is its signature close to the corresponding extension of the original signature. For this we turn to robustness, with the aid of a lemma about the index of a refinement. The following lemma was proven in [2, Lemma 3.7] (using the Cauchy-Schwartz inequality), although in essence it was also already implicitly proven in [18], in the proof of Lemma 4.1.

Lemma 6.5 ([2, Lemma 3.7])**.** *For every ϵ and t there exist $\gamma = \gamma_{6.5}(\epsilon)$ and $N_{6.5}(t, \epsilon)$ satisfying the following. Assume that $\mathcal{A}$ is an equipartition of a graph G with $n \geq N_{6.5}(t, \epsilon)$ vertices into s sets, and that $\mathcal{B}$ is a refinement of $\mathcal{A}$ into at most t sets. Assume further that $\mathcal{S}$ is any γ-signature of $\mathcal{A}$, and that $\mathcal{T}$ is its extension to $\mathcal{B}$. If $\mathcal{B}$ satisfies* $\mathrm{ind}(\mathcal{B}) \leq \mathrm{ind}(\mathcal{A}) + \gamma$, *then $\mathcal{T}$ is an ϵ-signature for $\mathcal{B}$.*

The following lemma about the index of a refinement never decreasing too much was also implicitly proven in the course of several regularity-related proofs. See for example [8, Lemma 7.2.2].

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

Proof sketch. If t divides n (and hence so does s), then we would have $\mathrm{ind}(\mathcal{B}) \geq \mathrm{ind}(\mathcal{A})$ as a direct consequence of the Cauchy-Schwartz inequality (see e.g. [8, Lemma 7.2.2]): Set $\mathcal{A} = \{V_i | 1 \leq i \leq s\}$ and $\mathcal{B} = \{W_{i,k} | 1 \leq i \leq s, 1 \leq k \leq t/s\}$, where $\{W_{i,1}, \ldots, W_{i,t/s}\}$ are assumed to be exactly the members of $\mathcal{B}$ that are contained in V_i. It is clear that for all $1 \leq i < j \leq s$ we have that $d(V_i, V_j)$ is the average of the sequence $\langle d(W_{i,k}, W_{j,l}) | 1 \leq k, l \leq t/s \rangle$. Hence, the square of $d(V_i, V_j)$ is at most the average of the squares of $\langle d(W_{i,k}, W_{j,l}) | 1 \leq k, l \leq t/s \rangle$, and from here it is easy to see that $\mathrm{ind}(\mathcal{B}) \geq \mathrm{ind}(\mathcal{A})$.

If t does not divide n then we may lose on the difference between $\mathrm{ind}(\mathcal{A})$ and $\mathrm{ind}(\mathcal{B})$ on account of rounding errors, but for an appropriate choice of N this loss would be less than ϵ. ■

We can now prove the existence of a refinement for $\mathcal{A}$ that is also regular with respect to G', provided that $\mathcal{A}$ is robust enough.

Lemma 6.7. *For every ϵ there exist $\gamma = \gamma_{6.7}(\epsilon)$ and $f = f^{(\epsilon)}_{6.7} : \mathbb{N} \to \mathbb{N}$ satisfying the following. Suppose that $\mathcal{A}$ is an (f, γ)-robust partition of a graph G into s sets and that $\mathcal{S}$ is a γ-signature of $\mathcal{A}$, where G has $n \geq N_{6.7}(s, \epsilon)$ vertices. Then for every G' that shares the same vertex set as*

G, there exists a refinement $\mathcal{B}$ of $\mathcal{A}$ into $t \le T_{3.6}(s,\epsilon)$ sets which is ϵ-regular for both G and G'. Moreover, the corresponding extension of $\mathcal{S}$ to $\mathcal{B}$ is an ϵ-signature with respect to G.

Proof. We set $\gamma = \min\{\frac{1}{2}\gamma_{4.1}(\epsilon), \gamma_{6.5}(\epsilon)\}$, and for every $k \in \mathbb{N}$ we set $f(k) = f^{(\epsilon)}_{4.1}(T_{3.6}(k,\epsilon))$. We set N to be the maximum over the respective functions of all lemmas that are used in the following (this will be explained later on).

Given a partition $\mathcal{A}$ as above, and assuming that $N \ge N_{3.6}(s,\epsilon)$, Lemma 3.6 asserts that there is a refinement $\mathcal{B}$ of $\mathcal{A}$ that partitions $V(G')$ into at most $t \le T_{3.6}(s,\epsilon)$ sets and is ϵ-regular with respect to G'.

Lemma 6.6, assuming that $N \ge N_{6.6}(T_{3.6}(s,\epsilon),\gamma)$, asserts that $\mathrm{ind}(\mathcal{B}) \ge \mathrm{ind}(\mathcal{A}) - \gamma \ge \mathrm{ind}(\mathcal{A}) - \frac{1}{2}\gamma_{4.1}(\epsilon)$ over G (the last inequality is by the choice of γ). This implies that $\mathcal{B}$ is $(f^{(\epsilon)}_{4.1}, \gamma_{4.1}(\epsilon))$-robust with respect to G, as otherwise, it would mean that there is a refinement $\mathcal{C}$ with at most $f^{(\epsilon)}_{4.1}(t)$ sets for which $\mathrm{ind}(\mathcal{C}) > \mathrm{ind}(\mathcal{B}) + \gamma_{4.1}(\epsilon)$, but this would imply that $\mathrm{ind}(\mathcal{C}) > \mathrm{ind}(\mathcal{A}) + \gamma_{4.1}(\epsilon) - \frac{1}{2}\gamma_{4.1}(\epsilon)$, which contradicts the robustness requirement of $\mathcal{A}$. Hence we conclude by Lemma 4.1 (applied to $\mathcal{B}$) that $\mathcal{B}$ is also ϵ-regular with respect to G. This proves that the refinement $\mathcal{B}$ is as needed.

In addition, the original robustness requirement for $\mathcal{A}$ ensures that the index of $\mathcal{B}$ with respect to G is no more than $\mathrm{ind}(\mathcal{A}) + \gamma_{6.5}(\epsilon)$. Hence, Lemma 6.5 ensures that the extension of $\mathcal{S}$ is an ϵ-signature for $\mathcal{B}$ with respect to G, as required. ■

In the course of the proof of the above, we also make the following observation.

Observation 6.8. *If $\mathcal{A}$ is an $(f^{(\epsilon)}_{6.7}, \gamma_{6.7}(\epsilon))$-robust partition of a graph G into s sets, where G has $n \ge N_{6.7}(s,\epsilon)$ vertices, and $\mathcal{B}$ is any refinement of $\mathcal{A}$ with $t \le T_{3.6}(s,\epsilon)$ sets, then the extension of any $\gamma_{6.7}(\epsilon)$-signature of $\mathcal{A}$ to $\mathcal{B}$ is an ϵ-signature of $\mathcal{B}$ (over G).* ■

We are now ready for the proof of Lemma 4.5. The intuition of the proof is the following: Assume that $\mathcal{S}$ is a γ-signature of an equipartition $\mathcal{A}$ that is (f,γ)-robust for G, for a small enough γ and a fast enough growing f. Our decision whether to accept or reject G is based on checking whether there is a refinement $\mathcal{B}$ of $\mathcal{A}$ (with not too many sets) for which the extension $\mathcal{T}$ of $\mathcal{S}$ is close enough to a signature $\mathcal{T}'$ for which the perceived q-statistic satisfies $\Pr_{\mathcal{T}'}(\mathcal{H}) \ge \frac{1}{2}$. If such a refinement exists then we accept G, and otherwise we reject G.

Now, if there is an $(\epsilon - \delta)$-close graph G' for which $\Pr_{G'}(\mathcal{H}) \ge \frac{2}{3}$ then G will be accepted, as close enough graphs have close signatures (Observation 6.1), and f and γ will be chosen so that $\mathcal{B}$ will be regular enough for both G and G' (as implied by Lemma 6.7), so that the signature $\mathcal{T}'$ of $\mathcal{B}$ with respect to G' (which is close to $\mathcal{T}$) is such that $\Pr_{\mathcal{T}'}(\mathcal{H})$ approximates $\Pr_{G'}(H)$ well enough so it does not fall below $1/2$. On the other hand, if G is accepted on account of some signature $\mathcal{T}'$ close to $\mathcal{T}$ for which $\Pr_{\mathcal{T}'}(\mathcal{H}) \ge \frac{1}{2}$, then Lemma 6.2 asserts that there is a close enough G' to G, for which the partition $\mathcal{B}$ is is regular enough, and for which $\mathcal{T}'$ is indeed a signature ensuring that $\Pr_{G'}(H)$ is close enough to $\Pr_{\mathcal{T}'}(\mathcal{H})$ so that it is larger than $\frac{1}{3}$. We now choose the actual parameters and present the formal proof.

Proof of Lemma 4.5. We set the values $\gamma = \gamma_{6.7}(\gamma_0)$, $s = \max\{r_{3.8}(q, \frac{1}{6}), r_{6.1}(\frac{1}{6}\delta)\}$, and $f(k) = f_{6.7}^{(\gamma_0)}(k)$, where

$$\gamma_0 = \min\{\frac{1}{6}\delta,\ \gamma_{3.8}(q, \frac{1}{6}),\ \gamma_{6.2}(\min\{\frac{1}{2}\delta, \gamma_{3.8}(q, \frac{1}{7})\})\}.$$

We set N to be the maximum over all respective functions of the lemmas and arguments used in the following (we omit here the exact details of the implicitly assumed lower bounds on n).

Given a γ-signature $\mathcal{S}$ for an (f, γ)-robust partition $\mathcal{A}$ into $t \geq s$ sets, we do the following. We check whether there could be any refinement $\mathcal{B}$ of $\mathcal{A}$ with at most $T_{3.6}(t, \gamma_0)$ sets, for which the extension $\mathcal{T}$ of $\mathcal{S}$ to $\mathcal{B}$ is $(\epsilon - \frac{1}{2}\delta)$-close to any signature $\mathcal{T}'$ such that the perceived q-statistic according to $\mathcal{T}'$ satisfies $\Pr_{\mathcal{T}'}(\mathcal{H}) \geq \frac{1}{2}$. If there exists such a signature then we accept G, and otherwise we reject it. Note that the existence of the refinement $\mathcal{B}$ depends only on the provided signature $\mathcal{S}$, so we do not make here any additional queries to the graph G. We now prove the two directions that tie the existence of such a $\mathcal{T}'$ with the existence of a corresponding graph G'.

Proof of the first direction. Suppose that G' is any graph that is $(\epsilon - \delta)$-close to G, and for which $\Pr_{G'}(\mathcal{H}) \geq \frac{2}{3}$. We will show that G is accepted by the above procedure. We only use here that $\gamma_0 \leq \min\{\frac{1}{6}\delta, \gamma_{3.8}(q, \frac{1}{6})\}$ in the expressions for γ, s and f.

Indeed let $\mathcal{A}$ be an (f, γ)-robust partition of G into $t \geq s$ sets and let $\mathcal{S}$ be a γ-signature of $\mathcal{A}$. By Lemma 6.7, there exists a refinement $\mathcal{B}$ of $\mathcal{A}$ into at most $T_{3.6}(t, \gamma_0)$ sets, so that $\mathcal{B}$ is γ_0-regular for both G and G'. Moreover, denoting by $\mathcal{T}$ the corresponding extension of $\mathcal{S}$, we have that $\mathcal{T}$ is a γ_0-signature of $\mathcal{B}$ with respect to G. By the upper bound on γ_0 this implies that $\mathcal{B}$ is $\gamma_{3.8}(q, \frac{1}{6})$-regular for both G and G', and that $\mathcal{T}$ is a $\frac{1}{6}\delta$-signature of $\mathcal{B}$ with respect to G.

Let $\mathcal{T}'$ be the 0-signature of $\mathcal{B}$ over G'. Lemma 3.8 implies (using $\mathcal{B}$ and G') that the perceived statistics with respect to $\mathcal{T}'$ and the actual statistics of G' are of variation distance at most $\frac{1}{6}$. Thus, Lemma 3.2 implies that $\Pr_{\mathcal{T}'}(\mathcal{H}) \geq \frac{2}{3} - \frac{1}{6} = \frac{1}{2}$. In addition, by Observation 6.1 (since $\mathcal{B}$ has at least $r_{6.1}(\frac{1}{6}\delta)$ sets and assuming that n is large enough), $\mathcal{T}'$ is $(\epsilon - \frac{1}{2}\delta)$ close to $\mathcal{T}$ on account of G and G' being $(\epsilon - \delta)$-close graphs. Thus, $\mathcal{T}$ and $\mathcal{T}'$ provide a witness that the procedure above accepts G. ■

Proof of the second direction. Let $\mathcal{A}$ be an (f, γ)-robust partition of G into $t \geq s$ sets and let $\mathcal{S}$ be a γ-signature of $\mathcal{A}$. Assume that there is a refinement $\mathcal{B}$ of $\mathcal{A}$ into at most $T_{3.6}(t, \gamma_0)$ sets, for which the extension $\mathcal{T}$ of $\mathcal{S}$ to $\mathcal{B}$ is $(\epsilon - \frac{1}{2}\delta)$-close to a signature $\mathcal{T}'$ such that the perceived q-statistic according to $\mathcal{T}'$ satisfies $\Pr_{\mathcal{T}'}(\mathcal{H}) \geq \frac{1}{2}$.

We will show that there is a graph G' that is ϵ-close to G and for which $\Pr_{G'}(\mathcal{H}) > \frac{1}{3}$. We use here the fact that $\gamma_0 \leq \gamma_{6.2}(\min\{\frac{1}{2}\delta, \gamma_{3.8}(q, \frac{1}{7})\})$ in the expressions for γ, s and f.

Indeed, Observation 6.8 (regarding $\mathcal{B}$ as a possible refinement of $\mathcal{A}$ with respect to G) asserts that $\mathcal{T}$ is a γ_0-signature of $\mathcal{B}$ (with respect to G), which by the upper bound on γ_0 means that it is a $\gamma_{6.2}(\min\{\frac{1}{2}\delta, \gamma_{3.8}(q, \frac{1}{7})\})$-signature for $\mathcal{B}$ with respect to G.

Now Lemma 6.2 (applied on $\mathcal{T}$ as an appropriate signature of $\mathcal{B}$ and the relatively close signature $\mathcal{T}'$) implies that there is a graph G' that is $(\epsilon - \frac{1}{2}\delta + \frac{1}{2}\delta)$-close to G, namely ϵ-close to G, and for which

$\mathcal{T}'$ is a $\gamma_{3.8}(q, \frac{1}{7})$-signature of $\mathcal{B}$, which in turn is $\gamma_{3.8}(q, \frac{1}{7})$-regular over G'. By Lemma 3.8 about the closeness of the q-statistic of G' to the perceived one, and Lemma 3.2, $\Pr_{G'}(\mathcal{H}) \geq \frac{1}{2} - \frac{1}{7} > \frac{1}{3}$ as required. ■

With both directions proven, the correctness of the above algorithm is now established. ■

7 Concluding comments

Efficient calculation of regular partitions

The main result of [1] is an algorithm that, for a fixed ϵ, calculates for an input graph G an ϵ-regular partition thereof. The algorithm is proven there to be in NC_1, and with deterministic time (in its non-parallel version) that is the same as that of matrix multiplication. By carefully reviewing the proof of Lemma 4.4 we can strengthen the first part of their result. First we give a formal definition for the computational complexity of our algorithms.

Definition 12. *A* TC_0 solution *for a problem is an efficient (polynomial time in n) algorithm for constructing a polynomial size (in n) circuit for every n, that gives a correct answer for every input instance of this size, where the height of the circuit is independent of n, and the circuit consists solely of unlimited fan-in AND ($\wedge$) gates, OR ($\vee$) gates, threshold gates (a threshold gate, for inputs $y_1, \ldots, y_m$ and a given in advance parameter t, checks whether $\sum_{i=1}^{m} y_i \geq t$), and negations ($\neg$). By contrast, an* NC_1 solution *allows circuits with only negations and fanin* 2 *AND/OR gates, but whose height can be logarithmic in n.*

By our methods we are able to prove the following.

Theorem 7.1. *For every k, γ and $f : \mathbb{N} \to \mathbb{N}$ there exists a* TC_0 *solution, that for an input graph G with $n \geq N_{4.3}(k, \frac{1}{2}\gamma, f)$ vertices computes an (f, γ)-final partition of G into at least k and at most $T_{4.3}(k, \frac{1}{2}\gamma, f)$ sets.*

Proof sketch. First we show how to calculate only a signature for such a partition. We follow the proof of Lemma 4.4. We recall that whenever the algorithm in the proof needs to accept or reject a signature $\mathcal{S}$, it makes a constant number of iterations of a GGR-test. Here we will instead reject or accept $\mathcal{S}$ based on an estimation of the acceptance probability of one GGR-test. For this end we first construct a deterministic circuit for every possible choice of the queries from G that the GGR-test can make. The queries of a GGR-test are made by first uniformly choosing a constant number of vertices of G, so there is a polynomial number of such choices, and for each one of them we can use a constant size circuit to know whether the test would have accepted had it made these queries. Then we collect all the outputs of all these circuits through one threshold gate, setting the threshold to be equal to half of the number of the inputs of the gate. Thus we will (deterministically) accept

$\mathcal{S}$ if and only if the corresponding GGR-test would have accepted with probability at least $\frac{1}{2}$, and we can clearly state and prove for this procedure a (deterministic) replacement for Claim 5.4.

In the original algorithm of Lemma 4.4 there were no other queries made apart from those coming from the constant number of instances of the GGR-test. Given all the acceptance and rejection decisions of the signatures above, whose number is independent of n, we can now find s^* and $\mathcal{S}^*$ as in the algorithm of Lemma 4.4 using an additional constant number of gates. A claim analogous to Claim 5.5 will also work here to ensure that this output is correct.

To find the actual final partition of G, we turn again to the proof in [13] of Lemma 5.2. In addition to the test itself, it is proven in [13] that it is possible with high probability to find a constant query size oracle for placing every vertex of G in its correct set of the partition. In our case we will go over all possible oracles (again there is a polynomial number of such oracles, as the randomized oracle was built in [13] using a constant number of queries to the graph), and for every possibility we use threshold gates to check whether its densities are indeed within the parameters of the corresponding s^* and $\mathcal{S}^*$ (noting that there is only a constant number of possibilities for the values of s^* and $\mathcal{S}^*$). ■

Comparing the above theorem to the main result of [1], it is a strengthening both in the types of partitions it can find (finding ϵ-regular partitions through Lemma 4.1), and in the complexity class of the algorithm (TC_0 as compared to NC_1). On the other hand, if we consider the running time of the non-parallel version of the algorithm and are concerned only with regular partitions, then the algorithm of [1] still performs significantly better than the one here.

Robust partitions and variants of regularity

A variant of the regularity lemma that required the existence of both a partition and a regular refinement thereof in the graph G played a central role in [2], [9] and [5]. That variant can also be proven using the notion of robust partitions; in fact, the proof in [2] of the corresponding variant is similar in essence to some of the methods used here for proving Observation 4.2 and Lemma 6.7, so the framework here can be viewed as a generalization of the previous frameworks.

Reducing the number of queries

One can reduce somewhat the number of queries of our test, if instead of Lemma 6.7 a more complicated lemma (but with better parameters) about the existence of a partition that is final for both G and G' is proven (rather than starting with a partition $\mathcal{A}$ that is only final for G). However, such an approach would make for a more complicated proof, and for a more complicated estimation algorithm that will have to find the parameters for all possible final partitions.

This improvement in the number of queries still would not have made it practical, since as long as the Regularity Lemma is used in such a form, the estimation will require a number of queries that is at least a tower in some function of the number of queries of the original testing algorithm. For this reason we aimed here for proof simplicity instead. It would be interesting if this (or any other graph testing result whose only known proof depends on the Regularity Lemma) can be proven using alternative methods that would provide a saner dependency of the parameters.

References

[1] N. Alon, R. A. Duke, H. Lefmann, V. Rödl and R. Yuster, The algorithmic aspects of the Regularity Lemma, *Journal of Algorithms* 16 (1994), 80–109.

[2] N. Alon, E. Fischer, M. Krivelevich and M. Szegedy, Efficient testing of large graphs, *Combinatorica* 20 (2000), 451–476.

[3] N. Alon, E. Fischer, I. Newman and A. Shapira, A combinatorial characterization of the testable graph properties: It's all about regularity, *Proceedings of the 38th STOC* (2006), 251–260

[4] N. Alon and A. Shapira, Every monotone graph property is testable, *Proceedings of the* 37^{th} *ACM STOC* (2005), 128–137.

[5] N. Alon and A. Shapira, A characterization of the (natural) graph properties testable with one-sided error, *Proceedings of the* 46^{th} *IEEE FOCS* (2005), 429–438.

[6] N. Alon and J. H. Spencer, *The Probabilistic Method*, Second Edition, Wiley, New York, 2000.

[7] M. Blum, M. Luby and R. Rubinfeld, Self-testing/correcting with applications to numerical problems. *Journal of Computer and System Sciences* 47 (1993), 549–595 (a preliminary version appeared in Proc. 22^{nd} STOC, 1990).

[8] R. Diestel, *Graph Theory* (2^{nd} edition), Springer (2000).

[9] E. Fischer, Testing graphs for colorability properties, *Random Structures and Algorithms* 26 (2005), 289–309.

[10] E. Fischer, The art of uninformed decisions: A primer to property testing, *The Bulletin of the European Association for Theoretical Computer Science* 75 (2001), 97–126.

[11] E. Fischer, The difficulty of testing for isomorphism against a graph that is given in advance, *SIAM Journal on Computing* 34 (2005), 1147–1158.

[12] E. Fischer and L. Fortnow, Tolerant versus intolerant testing for boolean properties, *Proceedings of the* 20^{th} *IEEE Conference on Computational Complexity* (2005), 135–140.

[13] O. Goldreich, S. Goldwasser and D. Ron, Property testing and its connection to learning and approximation, *Journal of the ACM* 45 (1998), 653–750 (a preliminary version appeared in Proc. 37th FOCS, 1996).

[14] O. Goldreich and L. Trevisan, Three theorems regarding testing graph properties, *Random Structures and Algorithms* 23 (2003), 23–57.

[15] M. Parnas, D. Ron and R. Rubinfeld, Tolerant property testing and distance approximation, available as *ECCC Report TR04-010*.

[16] D. Ron, Property testing (a tutorial), In: *Handbook of Randomized Computing* (S. Rajasekaran, P. M. Pardalos, J. H. Reif and J. D. P. Rolim eds), Kluwer Press (2001), Vol. II, 597–649.

[17] R. Rubinfeld and M. Sudan, Robust characterization of polynomials with applications to program testing, *SIAM Journal of Computing* 25 (1996), 252–271 (first appeared as a technical report, Cornell University, 1993).

[18] E. Szemerédi, Regular partitions of graphs, In: *Proc. Colloque Inter. CNRS* No. 260 (J. C. Bermond, J. C. Fournier, M. Las Vergnas and D. Sotteau eds.), 2978, 399–401.