Testing versus Estimation of Graph Properties
Extended Abstract + Appendix

Eldar Fischer
Technion
Haifa, Israel
eldar@cs.technion.ac.il

Ilan Newman
University of Haifa
Haifa, Israel
ilan@cs.haifa.ac.il

Abstract

Some properties are testable but not tolerantly testable. We show here that testable graph properties, however, also admit an algorithm that approximates the distance from the property up to any fixed additive constant. This work is also connected to the question of finding a combinatorial characterization of the testable graph properties, and to the question of efficiently finding a regular partition.

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 length 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 length 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 length. Properties that admit such algorithms are called *testable*.

The first time a question formulated in terms of property testing was considered is by Blum, Luby and Rubinfeld [3], and the general notion of property testing was first formally defined by Rubinfeld and Sudan [13]. The first investigation in the combinatorial context is that of Goldreich, Goldwasser and Ron [9], 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 [12] and [6].

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 [9]. Other classes were identified as testable using the Regularity Lemma of Szemerédi, in [2] and [5]. 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 the understanding of the essence of graph property testing.

The question of providing a complete structural characterization result for the testable graph properties seems hard. A partial result that in some sense correlates the graph properties that consist of only one graph according to their testability is found in [7], also making use of the Regularity Lemma. In a different angle of the characterization problem, the canonical testers of [10] 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 is in fact identical to a class of properties that admit algorithms with much stringer requirements than those of property testers (more connections of our result to the general characterization question are discussed in the end).

An investigation that goes beyond the original definition of testable properties was initiated by Parnas, Ron and Rubinfeld [11], when considering tolerant testers. These are 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 [8] 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), and 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 is an $O(1)$-query algorithm that can estimate the relative distance of an input to the property within any fixed additive constant.

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 a testable property. For non-graph properties, this is not always true as shown in [8].

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 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 so can be viewed as an NC_0 circuit generated at random), and $O(\log n)$ random bits, that approximates the signature of a regular partition of a graph to within any additive constant $\epsilon > 0$.

The rest of the paper is organized as follows. Section 2 contains the basic definitions and the formal statement of the main result. Section 3 contains definitions and lemmas concerning Szemerédi's Regularity Lemma and some variations thereof which are central to the proof, and Section 4 contains definitions and lemmas connecting regular partitions with graph property testing, as well as the proof of the main result. The two main lemmas used for the proof of the main result are then proven in Appendix A and Section 5 respectively. The final Section 6 contains some concluding comments, and some implications of the presented proofs.

2 The main result

In the following we formally state our main result, but first we state 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 definitions 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* n*, which is given to the algorithm in advance), and distinguishes with probability* $\frac{2}{3}$ *between the case that an input satisfies the property, and the case that an 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 [11] 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 graph 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 accept any input whose distance from the property is significantly 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), and distinguishes with probability* $\frac{2}{3}$ *between the case that an 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.*

3 Distributions, regularity and robustness

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

The following exercise is useful to what follows.

Observation 3.1. *Suppose that* μ *and* ν *are probability distributions over graphs with the labeled set of vertices* $\{v_1, \ldots, v_q\}$*, where each pair* $v_i v_j$ *is independently chosen to be an edge with probability* $\mu_{i,j}$ *and* $\nu_{i,j}$ *respectively. If* $|\mu_{i,j} - \nu_{i,j}| \leq \epsilon/\binom{q}{2}$ *for every* $1 \leq i < j \leq q$*, then the variation distance between* μ *and* ν *is bounded by* ϵ. ∎

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* $|U'| \geq \epsilon|U|$ *and* $|V'| \geq \epsilon|V|$*, the edge densities satisfy* $|d(U', V') - d(U, V)| < \epsilon$.

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

Lemma 3.2 (see [7] for a formal proof). *For every* k *and* ϵ *there exists* $\gamma = \gamma_{3.2}(k, \epsilon)$*, 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.*

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 as above is called ϵ-regular *if at least $\epsilon\binom{k}{2}$ of the pairs V_i, V_j are ϵ-regular.*

An equipartition $\mathcal{B} = \{W_1, \ldots, W_l\}$ is said to be an ϵ-refinement *of $\mathcal{A}$ as above if all but at most ϵl of the sets W_i are each fully contained in some set of $\mathcal{A}$. A 0-refinement is also simply called a* refinement.

Regular partitions are found using the famed Regularity Lemma of Szemerédi [14] (see [4, Chapter 7] for a good exposition of the proof).

Lemma 3.3 (Szemerédi's Regularity Lemma [14]). *For every k and ϵ there exists $T = T_{3.3}(k, \epsilon)$, such that for every equipartition $\mathcal{A}$ of a graph G with $n \geq N_{3.3}(k, \epsilon)$ vertices into k sets, there exists a refinement $\mathcal{B}$ of $\mathcal{A}$ into $t \leq T$ sets which is ϵ-regular.*

However, for our purpose we will also need some strengthening of the above lemma. For this end let us delve a little into the details of its proof, starting with some more definitions.

Definition 6. *For an equipartition $\mathcal{A}$ of a graph G into t sets, we define its* index $\mathrm{ind}(\mathcal{A})$ *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* $\mathrm{ind}(\mathcal{B}) \geq \mathrm{ind}(A) + \gamma$.

The main lemma used in proving Szemerédi's lemma can be paraphrased as the following (note that in the proof of Lemma 3.3 presented in [4], instead of $\mathrm{ind}(\mathcal{A})$, use is made of a similar function that is denoted by "q").

Lemma 3.4 ([14]). *For every ϵ there exists $\gamma = \gamma_{3.4}(\epsilon)$ and $f = f^{(\gamma)}_{3.4} : \mathbb{N} \to \mathbb{N}$, such that every (f, γ)-robust partition is also ϵ-regular.*

In the original formulation there, it is proven that a non-ϵ-regular partition into t sets has a refinement into $\exp(t)$ sets whose index is larger by at least some $\mathrm{poly}(\epsilon)$. With either formulation, the move from Lemma 3.4 to Lemma 3.3 is made through the following simple observation.

Observation 3.5. *For every k, γ and $f : \mathbb{N} \to \mathbb{N}$ there exists $T = T_{3.5}(k, \gamma, f)$, such that every equipartition $\mathcal{A}$ of a graph G with $n \geq N_{3.5}(k, \gamma, f)$ 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 proving this, repeating the procedure as many times as necessary. Since the index of a partition is always between 0 and 1, the process will terminate after at most $1/\gamma$ iterations. ■

In the following we will also need to 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 7. *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 up to $f(t)$ sets for which* $\mathrm{ind}(\mathcal{B}) \geq \mathrm{ind}(A) + \gamma$.

The following is an analogue of Observation 3.5 to final partitions. The price however is that now we cannot demand that the final partition will be a refinement of a given equipartition $\mathcal{A}$.

Observation 3.6. *For every k, γ and $f : \mathbb{N} \to \mathbb{N}$ there exists $T = T_{3.6}(k, \gamma, f)$, such that for every graph G with $n \geq N_{3.6}(k, \gamma, f)$ 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.

4 Signatures, approximability and statistics

Definition 8. *For an equipartition $\mathcal{A} = \{V_1, \ldots, V_t\}$ of G, an* ϵ-signature *of $\mathcal{A}$ is a sequence $\mathcal{S} = (\eta_{i,j})_{1 \leq i < j \leq t}$, such that $|d(V_i, V_j) - \eta_{i,j}| \leq \epsilon$ for every $i < j$ but at most $\epsilon\binom{t}{2}$ of the pairs. We also use just the term* signature *for $\mathcal{S}$ as above when we do not commit to any specific error parameter ϵ.*

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

A testing algorithm can actually compute a signature of the final partition that Observation 3.6 guarantees for a graph G, as the following lemma shows.

Lemma 4.1. *For every k, γ and $f : \mathbb{N} \to \mathbb{N}$ there exists $q = q_{4.1}(k, \gamma, f)$, such that there exists an algorithm that makes up to q queries to a graph G with $n \geq N_{3.6}(k, \frac{1}{2}\gamma, f)$ vertices, and computes 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 $T_{3.6}(k, \frac{1}{2}\gamma, f)$ sets.*

This lemma is proven in Appendix A, and brings us half-way towards our estimability result. To compute the distance to a property from a signature of a robust partition, we need the notion of combinatorial statistics.

Definition 9. *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 that is induced on all (labeled) graphs with q vertices by picking at random q distinct vertices of G and considering the induced subgraph.*

Given an ϵ-signature *$\mathcal{S} = (\eta_{i,j})_{1 \leq i < j \leq t}$ of an equipartition $\mathcal{A}$ of a graph G, 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 $v_1, \ldots, v_q$, we first choose a uniformly random sequences without repetitions of indices $i_1, \ldots, i_q$ from $1, \ldots, t$. We then independently take every $v_k v_l$ to be an edge of H with probability $\eta_{\min\{i_k, i_l\}, \max\{i_k, i_l\}}$.*

Given a family $\mathcal{H}$ of graphs with q vertices, we denote the probability for obtaining a member of $\mathcal{H}$ according to the q-statistic of G by $\Pr_G(\mathcal{H})$, and 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 4.2. *For every q and ϵ there exists $\gamma = \gamma_{4.2}(q, \epsilon)$ and $r = r_{4.2}(q, \epsilon)$, so that for every γ-regular partition $\mathcal{A}$ of G into $t \geq r$ sets (where G has $n \geq N_{4.2}(q, \epsilon, t)$ vertices) and γ-signature $\mathcal{S}$ of $\mathcal{A}$, the perceived q-statistic of G according to $\mathcal{S}$ is at most ϵ-far from the (actual) q-statistic of G in the variation distance.*

Proof. We set $r = 10\binom{q}{2}/\epsilon$ and $\gamma = \min\{\epsilon/5\binom{q}{2}, \gamma_{3.2}(q, \epsilon/5)\}$. Let $v_1, \ldots, v_q$ 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 $1 - \epsilon/5$, $i_1, \ldots, i_q$ are distinct, and moreover all the pairs V_{i_j}, V_{i_k} are γ-regular, and satisfy $|\eta_{i_j,i_k} - d(V_{i_j}, V_{i_k})| \leq \epsilon/5\binom{q}{2}$. Also, note that $\sum_{i=1}^{t} |(|V_i|/n) - 1/t| \leq \epsilon/5$ (for an appropriate choice of $N_{4.2}(q, \epsilon, t)$).

Finally, for a specific fixed sequence $i_1, \ldots, i_q$ for which the above event holds, Lemma 3.2 guarantees that the conditional distribution of the induced graph on $v_1, \ldots, v_q$ is not more than $\epsilon/5$ far 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 $|d(V_{i_j}, V_{i_k}) - \eta_{i_j,i_k}| \leq \frac{\epsilon}{5}\binom{q}{2}$ and using Observation 3.1, it follows that the difference between the q-statistic of G and the perceived one is at most ϵ. ∎

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 [10].

Lemma 4.3 (Canonical Testers [10]). *If there exists 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 $\mathcal{H}$ 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 $\Pr_G(\mathcal{H}) \leq \frac{1}{3}$.*

So now, 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, 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, but we can calculate 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 the appropriate function.

Lemma 4.4. *For every q and δ there exists $\gamma = \gamma_{4.4}(q, \delta)$, $s = s_{4.4}(q, \delta)$ and $f = f_{4.4}^{(q,\delta)} : \mathbb{N} \to \mathbb{N}$ with the following property. For every family $\mathcal{H}$ of graphs with q (labeled) vertices there exists an 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 $n \geq N_{4.4}(q, \delta, t)$ vertices, and distinguishes (using only $\mathcal{S}$ 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 5. Lemma 4.4 and Lemma 4.1 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 4.3 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 $\mathcal{P}$, $Prob_G(\mathcal{H}) \geq 2/3$ and for any graph G that is $\frac{1}{2}\delta$-far from $\mathcal{P}$, $Prob_G(\mathcal{H}) \leq 1/3$.

Set $\gamma = \gamma_{4.4}(q, \frac{1}{2}\delta)$, $f = f_{4.4}^{(q,\delta/2)}$, and $k = s_{4.4}(q, \frac{1}{2}\delta)$, and apply the algorithm provided by Lemma 4.1 with the parameters k, γ, f, on the input graph G. This algorithm makes up to $q_{4.1}(k, \gamma, f)$ 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 $s_{4.4}(q, \delta)$ sets that is (f, γ)-final.

We now apply the algorithm that is provided by Lemma 4.4, with parameters q, $\frac{1}{2}\delta$ and $\epsilon - \frac{1}{2}\delta$ to the signature $\mathcal{S}$ (note that this is actually a deterministic algorithm making no additional queries). Due to the choice of parameters, it is guaranteed by Lemma 4.4 that we can distinguish between the case that there is a graph G' that is at most $(\epsilon - \delta)$-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. We now claim that this algorithm is in fact an (ϵ, δ)-estimation algorithm for $\mathcal{P}$.

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 $\mathcal{P}$). 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

We first need some definitions about distances of signatures and how signatures behave under refinements of equipartitions.

Definition 10. *The distance between the signatures $\mathcal{S} = (\eta_{i,j})_{1 \leq i < j \leq t}$ and $\mathcal{S}' = (\eta'_{i,j})_{1 \leq i < j \leq t}$ is defined as $\sum_{1 \leq i < j \leq t} |\eta_{i,j} - \eta'_{i,j}| / \binom{t}{2}$.*

Given a signature $\mathcal{S} = (\eta_{i,j})_{1 \leq i < j \leq t}$ 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 $\mathcal{S}' = (\eta'_{i,j})_{1 \leq i < j \leq s}$ 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 there exist no such $k \neq l$.*

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

Lemma 5.1. *For every ϵ and t there exists $\gamma = \gamma_{5.1}(\epsilon)$ and $N = N_{5.1}(t)$, such that if G is a graph with $n \geq N$ vertices and a γ-regular partition $\mathcal{A}$ into t sets, $\mathcal{S} = (\eta_{i,j})_{1 \leq i < j \leq t}$ is a γ-signature of $\mathcal{A}$, and $\mathcal{S}' = (\eta'_{i,j})_{1 \leq i < j \leq t}$ is a signature that is δ-close to $\mathcal{S}$ for some δ, then G is $(\delta + \epsilon)$-close to a graph G' (with the same vertex set) for which $\mathcal{S}'$ is an ϵ-signature of $\mathcal{A}$, and $\mathcal{A}$ is an ϵ-regular partition for G'.*

Proof. We set $\gamma = \frac{1}{4}\epsilon$. Given G, $\mathcal{A} = \{V_1, \ldots, V_t\}$ and $\mathcal{S}$ 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. N is chosen large enough so that all the following will occur together with positive probability.

- $|d'(V_i, V_j) - \eta'_{i,j}| \leq \gamma < \epsilon$ for every $1 \leq i < j \leq t$. In particular this also implies that the fraction of pairs changed between G and G' on this pair does not exceed $|d(V_i, V_j) - \eta'_{i,j}| + \gamma$.
- For every $i < j$ for which V_i, V_j is a γ-regular pair in G, it is also an ϵ-regular pair in G' (this will occur with high probability for a large enough N, just by using large deviation inequalities for $d(U', V')$ for any large enough $U' \subset U$ and $V' \subset V$).

The above conditions imply the required conditions on $\mathcal{S}'$ and $\mathcal{A}$ with respect to G, as well as that the distance between G and G' (for N large enough) is at most

$$\gamma + \sum_{1 \leq i < j \leq t} (|d(V_i, V_j) - \eta'_{i,j}| + \gamma)/\binom{t}{2} \leq 4\gamma + (\sum_{1 \leq i < j \leq t} |\eta - \eta'_{i,j}|)/\binom{t}{2} \leq \delta + \epsilon$$

■

Unfortunately, if G' is the closest graph to G that satisfies a given property, it may be the case that the regular partition of G is not 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], although in essence it was also already implicitly proven in [14], in the proof of Lemma 3.4.

Lemma 5.2 ([2]). *For every ϵ there exists $\gamma = \gamma_{5.2}(\epsilon)$, so that for every equipartition $\mathcal{A}$ of G into s sets and γ-refinement $\mathcal{B}$ of $\mathcal{A}$ into t sets, where G has at least $N_{5.2}(\epsilon, t)$ vertices, either the extension of any γ-signature of $\mathcal{A}$ to $\mathcal{B}$ is an ϵ-signature thereof, or* $\mathrm{ind}(\mathcal{B}) > \mathrm{ind}(\mathcal{A}) + \gamma$.

The following lemma about the index of a refinement never decreasing too much is also implicitly proven in the course of the above.

Lemma 5.3. *For every ϵ there exists $\gamma = \gamma_{5.3}(\epsilon)$, so that for every equipartition $\mathcal{A}$ of G into s sets and γ-refinement $\mathcal{B}$ of $\mathcal{A}$ into t sets, where G has at least $N_{5.3}(\epsilon, t)$ vertices,* $\mathrm{ind}(\mathcal{B}) \geq \mathrm{ind}(\mathcal{A}) - \epsilon$.

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

Lemma 5.4. *For every ϵ there exists $\gamma = \gamma_{5.4}(\epsilon)$ and $f = f^{(\epsilon)}_{5.4} : \mathbb{N} \to \mathbb{N}$ satisfying the following. Suppose that $\mathcal{A}$ is an (f,γ)-robust partition of a graph G into s sets (where G has $n \geq N_{5.4}(\epsilon, s)$ vertices), that G' is a graph that shares the same vertex set as G, and that $\mathcal{S}$ is a γ-signature of $\mathcal{A}$. Then there exists a refinement $\mathcal{B}$ of $\mathcal{A}$ into $t \leq T_{3.3}(\epsilon, s)$ sets which is ϵ-regular for both G and G', and for which the corresponding extension of $\mathcal{S}$ is an ϵ-signature with respect to G.*

Proof. We set $\gamma = \min\{\frac{1}{2}\gamma_{3.4}(\epsilon), \gamma_{5.2}(\epsilon)\}$, and for every $k \in \mathbb{N}$ we set $f(k) = f^{(\epsilon)}_{3.4}(T_{3.3}(\epsilon, k))$. We set N to be the maximum over the respective functions of all lemmas that are used in the following.

Given a partition $\mathcal{A}$ as above, we use Lemma 3.3 to find a refinement $\mathcal{B}$ of $\mathcal{A}$ that is ϵ-regular with respect to G'. Lemma 5.3 (assuming that $N \geq N_{5.3}(\frac{1}{2}\gamma_{3.4}(\epsilon))$) and the robustness of $\mathcal{A}$ imply that $\mathcal{B}$ is now $(f^{(\epsilon)}_{3.4}, \gamma_{3.4}(\epsilon))$-robust with respect to G, and so it is ϵ-regular also with respect to G. In addition, the original robustness requirement ensures that the index of $\mathcal{B}$ with respect to G is no more than $\mathrm{ind}(\mathcal{A}) + \gamma_{5.2}(\epsilon)$, so Lemma 5.2 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 5.5. *If $\mathcal{A}$ is an $(f^{(\epsilon)}_{5.4}, \gamma_{5.4}(\epsilon))$-robust partition of a graph G into s sets (where G has $n \geq N_{5.4}(\epsilon, s)$ vertices), and $\mathcal{B}$ is any refinement of $\mathcal{A}$ with $t \leq T_{3.3}(\epsilon, s)$ sets, then the extension of any $\gamma_{5.4}(\epsilon)$-signature of $\mathcal{A}$ to $\mathcal{B}$ is an ϵ-signature of $\mathcal{B}$ (over G).* ■

We are now ready for the conclusion of this section.

Proof of Lemma 4.4. We set $\gamma = \gamma_{5.4}(\gamma_0)$, $s = r_{4.2}(q, \frac{1}{6})$, and $f(k) = f^{(\gamma_0)}_{5.4}(k)$, where $\gamma_0 = \min\{\frac{1}{3}\delta, \gamma_{4.2}(q, \frac{1}{6}), \gamma_{5.1}(\min\{\frac{1}{2}\delta, \gamma_{4.2}(q, \frac{1}{7})\})\}$. We set N to be the maximum over all respective functions of the participating lemmas and arguments.

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 $T_{3.3}(\gamma_0, t)$ 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. We now prove the two directions that tie the existence of such a $\mathcal{T}'$ with the existence of a corresponding graph G'.

First direction. Suppose that G' is any graph that is $(\epsilon - \delta)$-close G, and for which $\Pr_{G'}(\mathcal{H}) \geq \frac{2}{3}$. Let $\mathcal{B}$ be the refinement of $\mathcal{A}$ that is $\gamma_{4.2}(q, \frac{1}{6})$-regular for both G and G' and for which the extension $\mathcal{T}$ of $\mathcal{S}$ is a $\frac{1}{3}\delta$-signature over G, as guaranteed by Lemma 5.4. Let $\mathcal{T}'$ be the 0-signature of $\mathcal{B}$ over G'.

Now $\mathcal{T}'$ is $(\epsilon - \frac{1}{2}\delta)$ close to $\mathcal{T}$ on account of the two being the appropriate signatures of $(\epsilon - \delta)$-close graphs (provided that n is large enough). Also, by Lemma 4.2, $\Pr_{\mathcal{T}'}(\mathcal{H}) \geq \frac{1}{2}$, and so summing up we obtain that $\mathcal{T}$ is at most $(\epsilon - \frac{1}{2}\delta)$-far from some signature whose perceived probability for $\mathcal{H}$ is at least $\frac{1}{2}$, as required. ■

Second direction. Assume that $\mathcal{T}$ is $(\epsilon - \frac{1}{2}\delta)$-close to some other signature $\mathcal{T}'$ for $\mathcal{B}$, for which $\Pr_{\mathcal{T}'}(\mathcal{H}) \geq \frac{1}{2}$. As $\mathcal{T}$ is a $\gamma_{5.1}(\min\{\frac{1}{2}\delta, \gamma_{4.2}(q, \frac{1}{7})\})$-signature for $\mathcal{B}$ with respect to G (by Observation 5.5), by Lemma 5.1 we can now construct a graph G' that is ϵ-close to G, and for which $\mathcal{T}'$ is a $\gamma_{4.2}(q, \frac{1}{7})$-signature. By Lemma 4.2 about the closeness of the q-statistic of G' to the perceived one, we now have that $\Pr_{G'}(\mathcal{H}) > \frac{1}{3}$ as required. ■

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

6 Concluding comments

Efficient calculation of regular partitions

The main result of [1] is an algorithm that, for a fixed ϵ, can calculate for an input graph G a regular partition thereof. The algorithm is proven to be in NC_1, and with deterministic time (in the non-parallel version) that is the same as that of matrix multiplication. Lemma 4.1 (together with Lemma 3.4) asserts that there is a randomized algorithm that uses only $O(1)$ queries to the input and only $O(\log n)$ coins (that is, an NC_0 circuit chosen at random from a polynomial sized family of such circuits), that for an input graph G will provide with probability $\frac{2}{3}$ an ϵ-signature of an ϵ-regular partition (the success probability can be amplified to be arbitrary close to 1).

Moreover, the property tester given in [9] can also with high probability produce an oracle for the graph partition it finds that is in essence an NC_0 circuit (the GGR-test works by first producing several candidate oracles, and then sampling them to verify whether at least one of them produces a viable partition). It seems that using this, a TC_0 algorithm for finding an ϵ-regular partition (where ϵ is fixed in advance) of an input graph G may be constructed. This would improve upon the NC_1 complexity class of the algorithm in [1], but not improve upon the running time of its non-parallel version.

Combinatorial characterization of testable graph properties

The original intent of the work presented here was to find a combinatorial characterization of all testable graph properties. Such a characterization seems to follow: We will say that a property $\mathcal{Q}$ *ϵ-approximates* a property $\mathcal{P}$ if every graph that satisfies one of $\mathcal{P}$ and $\mathcal{Q}$ is also ϵ-close to satisfying the other property. Lemma 4.1 and Lemma 4.4 may imply that a property $\mathcal{P}$ is testable if and only if for every ϵ it can be ϵ-approximated by the property of having a final partition with the appropriate parameters (some fine details may still need ironing out here). Lemma 4.1 suggests that this in turn may be approximated by a property that consists of a Boolean combination of GGR-properties and their negations. However, this characterization by itself is somewhat cumbersome and its main interest lies in that it implies the estimability result.

Robust partitions and previous variants of regularity

In [2] and [5] use was made of a variant of the regularity lemma that required the existence of both a partition and a regular refinement thereof in the graph G. That variant can also be proven using the notion of robust partitions; in fact, the proof in [2] of the variant is similar in essence to some of the methods used here for proving Observation 3.5 and Lemma 5.4.

Reducing the number of queries

One can reduce somewhat the number of queries in our test, if instead of Lemma 5.4 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 still require a number of queries that is a tower in some function of the number of queries of the original testing algorithm. For this reason we optimized here for proof simplicity instead. It would be interesting if this (and virtually any other graph testing result whose proof depends on the regularity lemma) can be proven by 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] 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).

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

[5] E. Fischer, Testing graphs for colorability properties, *Random Structures and Algorithms*, to appear (a preliminary version appeared in Proc. 12^{th} SODA, 2001).

[6] 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.

[7] E. Fischer, The difficulty of testing for isomorphism against a graph that is given in advance, *Proceedings of the* 36^{th} *ACM STOC* (2004).

[8] E. Fischer and L. Fortnow, Tolerant versus intolerant testing for boolean properties, in preparation.

[9] 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. 37^{th} FOCS, 1996).

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

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

[12] 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).

[13] 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).

[14] 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.

A Proof of Lemma 4.1

Our strategy is rather simple and outlined here. Let k, γ, f be as in the formulation of the lemma and let $T = T_{3.6}(k, \gamma/2, f)$. For every s, $k \leq s \leq f(T)$ we quantize all possible signatures of equipartitions into s parts, so they form an η-net in $[0,1]^{\binom{s}{2}}$ equipped with ℓ_∞ norm, for some η to be chosen later.

We consider all possible signatures of equipartitions into s sets where $k \leq s \leq f(T)$ that are different enough (namely, we consider only members of the finite η-net defined above). For every signature we test whether there exists a partition into s sets with densities in the appropriate ranges. This could be done using the test of Goldreich, Goldwasser and Ron for generalized graph partitions [9], which will be denoted here as the GGR-test. For every positive answer (namely, that such a partition exists) we record the signature and compute the index of the partition. Having all this information, we set for each s the quantity $\mathrm{ind}(s)$ that is the largest index of any of the partitions into s parts. We then set s^* to be such that for every s for which $s^* < s \leq s + f(s)$, we have $\mathrm{ind}(s) \leq \mathrm{ind}(s^*) + \gamma$. Finally, we output the signature that achieves the index 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 $\mathcal{B}$ that is guaranteed by Observation 3.6, for k, γ, f. $\mathcal{B}$ is a partition into $b \leq T$ sets with some signature $\mathcal{S}_B$. Thus, while passing through all possible equipartitions into b sets, in the process above, the closest signature to $\mathcal{S}_B$ must have been considered and the corresponding index, which is a good approximation of $\mathrm{ind}(\mathcal{B})$, was computed. Now, as $\mathcal{B}$ is (f, γ)-final, it follows by the definitions that $s^* = b$ is a valid answer to the output above, assuming that the approximation of the index computation is good enough. The formal details now follow.

Set $\eta = \frac{\gamma}{10 f^2(T)}$. For each $k \leq s \leq f(T)$ let $Q(s, \eta) = \{0, 1, \ldots, 1/\eta\}^{\binom{s}{2}}$. Each member of $q \in Q(s, \eta)$ is associated with a signature $\bar{d} = \eta \cdot q$ of a possible equipartition into s sets. Let $\mathcal{S}(s, \eta)$ denote all such signatures.

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

Observation A.1. *Let $\mathcal{A}$ be an equipartition into s sets and assume that $\{d_{i,j}, 1 \leq i < j \leq s\}$ is an η-signature of $\mathcal{A}$. Then $|\mathrm{ind}(\mathcal{A}) - \mathrm{ind}(d)| \leq 2\eta$.* ■

Let G be a graph on n vertices and let r be fixed. Let $0 \leq \alpha_{i,j} < \beta_{i,j} \leq 1$, $1 \leq i < j \leq r$ be two sequences of numbers. Then the following is a special case of a theorem proved by Goldreich, Goldwasser and Ron [9] (in [9], there are lower and upper bounds on the sizes of the vertex sets too, but having them does not make an essential difference.)

Lemma A.2 (GGR-test of partitions [9]). *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 A.3. *Let s be fixed, let $a \in \mathcal{S}(s, \eta)$ be a signature, and let $\alpha, \beta \in \mathcal{S}(s, \eta)$ be defined by $\alpha_{i,j} = a_{i,j} - \eta, \beta_{i,j} = a_{i,j} + \eta$, $1 \leq i < j \leq s$. Then applying the GGR test on a graph G with s, α, β and distance parameter η 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 a is an $s^2\eta$-signature.*

- *If there is an equipartition $\mathcal{A}$ of G into s sets for which a is an η-signature, then the test accepts with probability at least $\frac{2}{3}$.*

Proof. Assume that the test accepts with probability at least $\frac{1}{3}$ when applied with s, α, β. Then there must be a graph G' that is η-close to G and that has an equipartition $\mathcal{A}$ for which a is an η-signature. Thus $\mathcal{A}$, considered as an equipartition of G, must have $|d_G(V_i, V_j) - d_{G'}(V_i, V_j)| \leq \frac{1}{2}s^2\eta$ for every $1 \leq i < j \leq s$ (as otherwise G' will be more than η-far from G), and so a is an $s^2\eta$-signature for G'. The other direction is similar. ■

We now are ready to conclude this section.

Proof of Lemma 4.1. Suppose that the parameters f, γ and k are given. For $s \in \{k, \ldots, f(T)\}$, let η and $\mathcal{S}(s, \eta)$ be as defined above, and let $m = \Sigma_{s=k}^{f(T)} \frac{1}{\eta}^{\binom{s}{2}}$ be the total number of members in the union of all $\mathcal{S}(s, \eta)$, $s = k, \ldots, f(T)$.

We perform 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 $a \in \mathcal{S}(s, \eta)$ set $\alpha, \beta \in \mathcal{S}(s, \eta)$ to be $\alpha_{i,j} = a_{i,j} - \eta, \beta_{i,j} = a_{i,j} + \eta,\ 1 \leq i < j \leq s$ (just as in Lemma A.3).

  Apply the GGR-test on G with parameters α, β and distance parameter η for $100 \log m$ times. If the majority of the runs accept, then we set $M(s) := \max\{M(s), \mathrm{ind}(a)\}$, and record the signature a 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 a was rejected.

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

Let s^* be the smallest $s \leq T$ such that $M(s) + \gamma - 4f(s)^2\eta \geq M(s')$ for every $s' = s+1, \ldots, f(s)$. If there exists such an s^*, output the signature a 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 first argue that with probability at least $\frac{2}{3}$ the following holds. For every $s \in \{k, \ldots, f(T)\}$ and every $a \in \mathcal{S}(s, \eta)$ which the algorithm did not reject, there is an equipartition $\mathcal{A}_a$ into s sets, with $|\mathrm{ind}(\mathcal{A}_a) - \mathrm{ind}(a)| \leq 2s^2\eta$ and with a as its $s^2\eta$-signature; and for every such s and a which were rejected by the algorithm, there exists no equipartition $\mathcal{A}_a$ for which a is an η-signature.

Proof. Indeed, for a fixed s and a, if a run of the GGR-test accepts a with probability at least $\frac{1}{3}$, then by Lemma A.3 a is an $s^2\eta$-signature of an equipartition $\mathcal{A}$. Since the GGR-test was run $100 \log m$ times then, By a large deviation inequality, even if the rejection probability of a is at least $\frac{1}{6m}$ then there is such an $\mathcal{A} = \mathcal{A}_a$. By Observation A.1, $|\mathrm{ind}(\mathcal{A}_a) - \mathrm{ind}(a)| \leq 2s^2\eta$. Thus by the union bound over all signatures that were considered by the algorithm, we get that with probability at least $\frac{5}{6}$, for all s and a's that were accepted, it holds simultaneously that an $\mathcal{A}_a$ exists such that $|\mathrm{ind}(\mathcal{A}_a) - \mathrm{ind}(a)| \leq 2s^2\eta$ and a is its $s^2\eta$-signature.

The other direction, of showing that with probability at least $\frac{5}{6}$ no a which is an η-signature for some $\mathcal{A}_a$ was rejected, is proven similarly. Together we obtain that with probability at least $\frac{2}{3}$ both directions hold. ■

If the above occurs and the sampler outputs a signature a^* with index s^*, then there exists an equipartition $\mathcal{A}^*$ into s^* sets with a^* as its $s^2\eta$-signature, and by the triangle inequality for all $s \in \{s^*, \ldots, f(s^*)\}$, for every equipartition into s sets $\mathcal{A}$, $\mathrm{ind}(\mathcal{A}^*) + \gamma \geq \mathrm{ind}(\mathcal{A})$. In particular, if this happens, we conclude that a^* is indeed an (f, γ)-final partition.

We still need to show that if the acceptance and rejection of signatures follows the above claim with regards to the existence of corresponding equipartitions, then the algorithm will not fail in the last step.

Proof. 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 $s_1 < T$ exists is asserted in Observation 3.6. 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 in the lexicographic order of its signature). Then, by the fact that $\mathcal{A}$ is $(f, \gamma/2)$-final, $\mathrm{ind}(\mathcal{A}) + \gamma/2 \geq \mathrm{ind}(\mathcal{S})$ 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 $\mathrm{ind}(\mathcal{A}) \geq \mathrm{ind}(\mathcal{A}')$ for any equipartition $\mathcal{A}'$ into s_1 sets. Let $a \in \mathcal{S}(s_1, \eta)$ be the first in lexicographic order such that a is an η-signature of $\mathcal{A}$. There exists such an a since $\mathcal{S}(s_1, \eta)$ forms an η-net in the space of all signatures (with l_∞).

Thus, assuming that the sampler accepted all signatures which were η-signatures of a corresponding partition, a was in particular accepted. By Observation A.1 together with the fact that $\mathrm{ind}(\mathcal{A}) \geq \mathrm{ind}(\mathcal{A}')$ or any equipartition $\mathcal{A}'$ into s_1 sets, it follows that

$$\mathrm{ind}(\mathcal{A}) - 2s^2\eta \leq M(s_1) \leq \mathrm{ind}(\mathcal{A}) + 2s^2\eta$$

Moreover, by combining the inequalities above and Observation A.1 we get that, as long as all signatures that were not $s^2\eta$-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 η-signature $b \in \mathcal{S}(s, \eta)$ that was not rejected by the algorithm,

$$\mathrm{ind}(b) \leq \mathrm{ind}(\mathcal{B}) + 2s^2\eta \leq \mathrm{ind}(\mathcal{A}) + \gamma/2 + 2s^2\eta \leq M(s_1) + \gamma/2 + 4s^2\eta$$

Now this implies that $\mathrm{ind}(b) \leq M(s_1) + \gamma - 4f(s)^2\eta$ by the choosing $\eta = \gamma/8f(s)^2$. Thus s_1 is recognized as a candidate for s^*, and hence the sampler will not fail to output some s^* (we do not claim that the sampler actually outputs s_1 as s^*, but only that existence of s_1 will prevent it from failing in the last step). ■

To summarize, we proved first that with probability at least $\frac{2}{3}$, the sampler accepts all signatures under consideration that are η-signatures of some corresponding equipartition, and rejects all signatures that are not $s^2\eta$-signatures of any equipartition. Then we proved that 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 4.1. ■