SIAM J. COMPUT.
Vol. 37, No. 3, pp. 959–976

EFFICIENT TESTING OF BIPARTITE GRAPHS FOR FORBIDDEN INDUCED SUBGRAPHS*

NOGA ALON†, ELDAR FISCHER‡, AND ILAN NEWMAN§

Abstract. Alon et. al. [N. Alon, E. Fischer, M. Krivelevich, and M. Szegedy, *Combinatorica*, 20 (2000), pp. 451–476] showed that every property that is characterized by a finite collection of forbidden induced subgraphs is ϵ-testable. However, the complexity of the test is *double-tower* with respect to $1/\epsilon$, as the only tool known to construct such tests uses a variant of Szemerédi's regularity lemma. Here we show that any property of *bipartite* graphs that is characterized by a finite collection of forbidden induced subgraphs is ϵ-testable, with a number of queries that is polynomial in $1/\epsilon$. Our main tool is a new "conditional" version of the regularity lemma for binary matrices, which may be interesting on its own.

Key words. property testing, graph algorithms, approximation, regularity lemma

AMS subject classifications. 05C85, 05C35, 68Q10

DOI. 10.1137/050627915

1. Introduction.

Property testing, first started in [6] and [17], deals with the following general question: Given a property P and an input which is assumed to come in the form of an oracle, how many queries to the input are required to distinguish between an input which satisfies P and an input which is ϵ-far (in the normalized Hamming distance) from any input that satisfies P? Property testing in general, and the investigation of graph testing that was started in [14], in particular, has become an active research area in recent years (see, for example, [14, 3, 8, 15, 1, 4] and the surveys [16, 9]). In particular, it was shown in [3] that every property that is characterized by a finite collection of forbidden induced subgraphs is ϵ-testable, that is, one can distinguish between graphs that satisfy it and graphs that are ϵ-far from satisfying it, with a number of queries that is bounded by a function of ϵ only, and is independent of the size of the input graph. However, the complexity of the test is *double-tower* with respect to $1/\epsilon$, as the only tool known to prove this testability is a variant of Szemerédi's regularity lemma.

More recently, Alon and Shapira [1, 4] initiated a study of those graph properties that are characterized by forbidden subgraphs and can be tested "very efficiently" in the sense that they can be tested with only $poly(1/\epsilon)$ many queries. In [1] it is shown that the property of not containing a given subgraph (where the subgraph is not necessarily induced) is testable with a number of queries polynomial in $1/\epsilon$ if and

*Received by the editors March 29, 2005; accepted for publication (in revised form) February 20, 2007; published electronically August 29, 2007. A preliminary (and weaker) version of these results formed part of [10].

http://www.siam.org/journals/sicomp/37-3/62791.html

†Schools of Mathematics and Computer Science, Raymond and Beverly Sackler Faculty of Exact Sciences, Tel Aviv University, Tel Aviv 69978, Israel and School of Mathematics, Institute for Advanced Study, Princeton, NJ 08540 (nogaa@tau.ac.il). This author's research was supported in part by a grant from the Israel Science Foundation, by the Hermann Minkowski Minerva Center for Geometry at Tel Aviv University, and by the Von Neumann Fund.

‡Faculty of Computer Science, Technion—Israel Institute of Technology, Haifa 32000, Israel (eldar@cs.technion.ac.il). This author's research was supported in part by grant 55/03 from the Israel Science Foundation.

§Department of Computer Science, University of Haifa, Haifa 31905, Israel (ilan@cs.haifa.ac.il). This author's research was supported in part by grant 55/03 from the Israel Science Foundation.

only if the forbidden subgraph is bipartite. In the context of testing digraphs for a forbidden structure, [4] contains a similar (but more complex) classification. The only known upper bounds for the cases where the number of queries is not polynomial are the tower (or worse) functions that result from Szemerédi's regularity lemma and its variants.

Here we concentrate on graph properties that are characterized by a finite family of forbidden induced subgraphs. For general graphs, the only known upper bound is the tower of towers; it was obtained from the proof in [3] that this is testable at all. We consider here the special case of *bipartite* input graphs and show, in contrast to the above, that any property of bipartite graphs that is characterized by a finite collection of forbidden induced subgraphs is ϵ-testable with a number of queries that is polynomial in $1/\epsilon$.

Our main tool is a new "conditional" version of the regularity lemma for binary matrices (Lemma 1.6 below), which may be interesting on its own. We combine this with some methods similar to those of [11] to obtain the desired result ([11] is an expanded version of the results from [10] about matrix-poset properties, while this paper expands the results from [10] about testing of bipartite graphs; the original bounds in [10] for bipartite graphs, while better than the previously known tower of towers, were not polynomial in $1/\epsilon$).

Our results are stated for graphs that are already given with a bipartition of their vertices (with the definition of a forbidden subgraph also relating to subgraphs with a compatible bipartition). However, in the case of bipartite input graphs whose bipartition is not given in advance (and general induced forbidden subgraphs), we can first use the approximate bipartition oracle given in [14] to reduce that setting to our setting.

We now note that the study of such bipartite graph properties is an extension of the poset model studied in [11], in which the testability of properties is related to the logical complexity of their description (for the purpose here a *model* is the language in which the properties are expressed, so a model is essentially identifiable with its family of expressible properties). In this case the poset is the 2-dimensional $n \times n$ grid, which as a poset is the product of two n-size total orders (lines). The language (syntax) includes the poset relation, the label unary relation (being labeled "1"), and in addition, the relations $row(x_1, x_2)$ which state that x_1 is on the same row as x_2, and similarly $col(x_1, x_2)$ for columns. $\forall$-properties in this model are properties that can be described by a finite formula over a fixed number of variables with only $\forall$-quantifiers in prenex normal form. Such properties would then correspond to exactly the properties that are characterized by a finite collection of forbidden submatrices (in a manner similar to what was done in [11] for the $\forall$-poset model). We call this model the "submatrix model." The submatrix model is closely related to a submodel of the (not always testable) $\forall\exists$-poset model, defined in [11].

The model "submatrix" includes some interesting properties. In particular, the permutation-invariant properties in it are tightly connected to bipartite graph properties that are characterized by a collection of forbidden induced subgraphs.

DEFINITION 1.1. *For a finite collection F of 0/1 matrices, we denote by $\mathcal{S}_F$ all 0/1-matrices that do not contain as a submatrix any row and/or column permutation of a member of F.*

OBSERVATION 1.2. *Every bipartite graph property (where a bipartite graph is identified with its adjacency matrix in the usual way) that is characterized by a finite collection of forbidden induced subgraphs is equivalent to a property $\mathcal{S}_F$ for some finite set F of matrices. In addition, every $\mathcal{S}_F$-property in the "submatrix" model is*

equivalent to a bipartite graph property as above.

It is important to note that here we discuss forbidden *induced* subgraphs. Not having a forbidden subgraph (rather than induced subgraph) is a monotone decreasing property. In this case, the test for the property is trivial, by density. For a large enough density, a Zarankiewicz (see [21], [13]) type theorem asserts that the answer "No" is correct (as the graph will have a large enough complete bipartite graph), while if the density is low then the answer is trivially "Yes," as the graph is close to the empty (edgeless) one. A thorough treatment of this case is found in [1]. The main result in the present paper is the following.

THEOREM 1.3. *Let F be a fixed finite collection of 0/1 matrices. Property $\mathcal{S}_F$ is $(\epsilon, poly(\frac{1}{\epsilon}))$-testable for every $\epsilon > 0$, by a 2-sided error algorithm.*

The test above, however, is not only 2-sided but also very computation-intensive (despite this computation using only a relatively small set of queries as data). Using some additional tools we then derive a 1-sided error test which is also efficient in terms of its running time.

THEOREM 1.4. *Let F be a fixed finite collection of 0/1 matrices. Property $\mathcal{S}_F$ is $(\epsilon, poly(\frac{1}{\epsilon}))$-testable for every $\epsilon > 0$, by a one sided error algorithm whose running time is polynomial in the time it takes to make the queries.*

The derivation of Theorem 1.4 from the main tool used in Theorem 1.3 is done in two stages, in sections 5 and 6. To present the test proving Theorem 1.3, we will need some machinery.

Let M be a 0/1-labeled, $n \times n$ matrix (to simplify notation we restrict ourselves to square matrices, but all arguments and theorems in this paper hold word-for-word for rectangular $n \times m$ matrices as well). We denote by $R(M)$ and $C(M)$ the set of rows and the set of columns of M, respectively. For an integer r, an *r-partition* of M is a partition of the set $R(M)$ into $r' \leq r$ parts $\{R_1, \ldots, R_{r'}\}$ and a partition of the set $C(M)$ into $r'' \leq r$ parts $\{C_1, \ldots, C_{r''}\}$. Each submatrix of the form $R_i \times C_j$ will be called a block (note that the coordinate sets defining the blocks do not necessarily consist of consecutive matrix coordinates). The weight of the (i, j) block is defined as $\frac{1}{n^2}|R_i||C_j|$. We also define similar weights for the R_i's and C_j's, e.g., $w(R_i) = \frac{1}{n}|R_i|$.

For a block B of a 0/1-matrix M and $\delta \geq 0$, we say that B is *δ-homogeneous* if all but a δ-fraction of its values are identical. If B is δ-homogeneous we call the value that appears in at least a $1-\delta$ fraction of the places the *δ-dominant* value of B. Note that this value is also α-dominant for any $\delta < \alpha < 1/2$. We say that a value is the *dominant* value of B if it is simply the majority value in B.

DEFINITION 1.5. *Let $\mathcal{P} = \{R_1, \ldots R_{r'}\} \times \{C_1, \ldots C_{r''}\}$ be an r-partition of M, and let $\delta > 0$. We say that $\mathcal{P}$ is a* (δ, r)-partition *if the total weight of the δ-homogeneous blocks is at least $1 - \delta$.*

The key result is that an input that does not admit some (δ, r)-partition can be rejected easily, because it will then contain many copies of every possible $k \times k$ matrix (including the forbidden ones) as submatrices.

LEMMA 1.6. *Let k be fixed. For every $\delta > 0$ and an $n \times n$, 0/1-matrix M with $n > (k/\delta)^{O(k)}$, either M has a (δ, r)-partition for $r = r(\delta, k) \leq (k/\delta)^{O(k)}$, or for every 0/1-labeled $k \times k$ matrix B, a $(g(\delta, k) \geq (\delta/k)^{O(k^2)})$-fraction of the $k \times k$ submatrices of M are B.*

This lemma allows us to reduce the testing problem to matrices that admit a (δ, r)-partition for certain δ, r; as for matrices that do not admit such partitions, the lemma asserts that querying a random submatrix will find a counterexample with sufficiently high probability. We note that the lemma is essentially a conditional version

of Szemerédi's regularity lemma ([19]; see also [7, Chapter 7]), as a (δ, r)-partition is in particular a regular partition in the sense of Szemerédi of the corresponding bipartite graph. The improvement over directly using the regularity lemma is achieved because of this conditioning. The proof of the lemma will be presented in section 4.

We then construct a test for matrices admitting a (δ, r)-partition. This test will be very similar to the 2-sided boolean matrix poset test in [11]. However, the situation in the poset test is that the partition can be fixed in advance, while in our case there is the problem of "learning" enough of the partition by sampling. The main tool for doing so is Lemma 2.3 below. For stating it we need some more definitions, which are described in section 2 along with the framework of the proof of Theorem 1.3.

The plan of the paper is as follows. Section 2 includes some preliminaries, as well as a proof of Theorem 1.3 from two main lemmas—Lemma 1.6 above and Lemma 2.3 which is stated there. The lemmas themselves are proven in sections 4 and 3, respectively. We then turn to proving Theorem 1.4. This is done in two stages. First, a special case is proven in section 5, and then this case is used as a lemma in section 6 to prove the full result. In both stages we need the main tool that was used in the proof of Theorem 1.3, namely, Lemma 1.6. Finally, section 7 contains some concluding open problems.

2. Partitions, signatures, and Theorem 1.3. Assume that M has a (δ, r)-partition. We have no hope, of course, of finding it using $O(1)$ many queries, as we cannot even sample a single point from every matrix row. Hence, we will need to define the "high-level features" of the (δ, r)-partitions of M that can be detected by sampling.

In the following, whenever we refer to a δ-fraction of the members of a *weighted* set Q, we mean a subset Q', the total weight of whose members is δ (where we assume that the total weight of the members of Q is normalized to be 1). Let M be a matrix with a (δ, r)-partition $\mathcal{P}$ defined by the row partition $\{R_1, \ldots, R_s\}$ and the column partition $\{C_1, \ldots, C_t\}$, $s, t \leq r$. Then $\mathcal{P}$ naturally defines a high-level pattern which is an $s \times t$ matrix of the dominant labels of the blocks.

DEFINITION 2.1. *Let $\mathcal{P}$ be a partition as above, and let P be a 0/1-labeled, $s \times t$ matrix. A block $R_i \times C_j$ is called* δ-good *with respect to P if it is δ-homogeneous and its dominant label is $P_{i,j}$. P is called a* δ-pattern *of $\mathcal{P}$ if all but at most a δ-fraction of the weighted blocks in $\mathcal{P}$ are δ-good with respect to P.*

It is immediate from the definition that if a partition has a δ-good pattern of size $s \times t$, then it is a (δ, r)-partition with $r = \max\{s, t\}$. Conversely, if $\mathcal{P}$ is a (δ, r)-partition, then it has an $r \times r$ δ-pattern (by possibly introducing empty blocks). As the block sizes of a (δ, r)-partition need not be fixed, we will also need information about the weights of R_i and C_j, $(i, j) \in [s] \times [t]$.

DEFINITION 2.2. *Let M be an $n \times n$ matrix with a (δ, r)-partition $\mathcal{P}$ defined by the row partition $\{R_1, \ldots, R_s\}$ and the column partition $\{C_1, \ldots, C_t\}$. Then a* δ-signature *of $\mathcal{P}$ is an $s \times t$, 0/1-labeled matrix P and two sequences $\{\alpha_i\}_1^s, \{\beta_i\}_1^t$, where P is a δ-pattern of $\mathcal{P}$, and in addition $\sum_{i=1}^s |\frac{|R_i|}{n} - \alpha_i| \leq \delta$ and $\sum_{j=1}^t |\frac{|R_j|}{n} - \beta_j| \leq \delta$.*

Note that the signature of a partition is closed under permutations of rows and columns; namely, any row/column permutation of P with the respective permutations of $\{\alpha_i\}_1^s$ and $\{\beta_i\}_1^t$ is also a δ-signature of any matrix for which P is a δ-signature. Moreover, a signature of M is also a signature of all row/column permutations of M.

The signature of a partition has sufficient properties for constructing a test as we shall see in the proof of Theorem 1.3. The following also asserts that it can be approximated by sampling.

LEMMA 2.3. *Let $\delta < 1/81$ and assume that an $n \times n$, 0/1-matrix M has a (δ, r)-partition. By making $q = (r/\delta)^{O(1)}$ many queries, a $26\delta^{1/6}$-signature of a $(16\delta^{1/6}, 10r^2/(4\delta^{1/3}) + 1)$-partition can be found, with success probability $\frac{3}{4}$.*

We note that a test for a much closer approximation of the original (δ, r)-partition can also be deduced from [14], with exponentially worse running time and query complexity. The proof of Lemma 2.3 is given in section 3. We end the discussion by showing that together with Lemma 1.6 this indeed implies a 2-sided error test.

Proof of Theorem 1.3. Assume that we want to ϵ-test M for a permutation-invariant collection of forbidden induced $k \times k$ submatrices. Blocks will now correspond to partition-blocks: Let $\delta = (\frac{\epsilon}{300})^6$, and let $g = g(\delta, k)$, $r = r(\delta, k)$ be those of Lemma 1.6. For $4/g = (k/\epsilon)^{O(k^2)}$ iterations, independently, we choose k random rows and k random columns of M and query all k^2 points in the $k \times k$ matrix that is defined by them. If we find a counterexample in the queried points we answer "No" and terminate the algorithm, and otherwise we continue. Let E_1 denote the event that M has no (δ, r)-partition and yet the algorithm continues. For inputs with a (δ, r)-partition, this event (by definition) never happens, while for other inputs, by Lemma 1.6, the probability of this event is bounded by $\frac{1}{12}$.

We now work under the assumption that M has a (δ, r)-partition and use the algorithm given in Lemma 2.3 to try finding an $\frac{\epsilon}{8}$-signature of an $(\frac{\epsilon}{8}, 10r^2/4(\frac{\epsilon}{300})^2+1)$-partition by sampling $(r/\delta)^{O(1)} = (k/\epsilon)^{O(k)}$ queries. Let P with $\{\alpha_i\}_1^s$ and $\{\beta_i\}_1^t$ be the signature obtained by the algorithm, and let E_2 be the event that it is not an $\frac{\epsilon}{8}$-signature of an $(\frac{\epsilon}{8}, 10r^2/4(\frac{\epsilon}{300})^2 + 1)$-partition of M. If M in fact did not have a (δ, r)-partition, then this event has the same probability as E_1 (which is bounded by $\frac{1}{12}$), and otherwise by Lemma 2.3 the probability of E_2 is bounded by $\frac{1}{4}$.

We now form an $n \times n$ matrix M_Q that represents our knowledge of M: We partition the rows of M_Q into s parts of weights $\{\alpha_i\}_1^s$ and the columns into t parts of weights $\{\beta_i\}_1^t$. For every block of P, we set every entry of the corresponding block of M_Q to have the same label as in P. Now, let $\mathcal{M}_{Q,\epsilon}$ be the set of all matrices that can be obtained from M_Q by changing at most $\epsilon n^2/2$ entries in any possible way. We check if any of the members of $\mathcal{M}_{Q,\epsilon}$ has the property $\mathcal{S}_F$. If there is such a member, the algorithm answers "Yes." Otherwise, if every member $\mathcal{M}_{Q,\epsilon}$ contains a permutation of a forbidden submatrix, then the answer is "No." Note that this last phase of the algorithm involves no additional queries and is just a computation phase.

To see that the algorithm is correct we first note that if a counterexample is found in the first phase of the algorithm, then the input M does not have the property with probability 1. Hence the algorithm can err only in the second phase.

We claim that unless E_2 happened the following hold: (a) some row/column permutation of M is a member of $\mathcal{M}_{Q,\epsilon}$, and (b) every two members of $\mathcal{M}_{Q,\epsilon}$ are of distance at most ϵn^2. Indeed, assume that the signature that has been found is an $\frac{\epsilon}{8}$-signature of an $(\frac{\epsilon}{8}, 10r^2/4(\frac{\epsilon}{300})^2 + 1)$-partition of M. Then M_Q can be obtained from M by changing at most an $\frac{\epsilon}{8}$-fraction of the entries in each $\frac{\epsilon}{8}$-good block, followed by changing any of the entries in the non–$\frac{\epsilon}{8}$-homogeneous blocks, and finally changing entries that are in strips around every block to compensate for the inaccuracy of the size sequences of the signature (whose sizes sum up to no more than $\frac{\epsilon}{8}$ for the rows and $\frac{\epsilon}{8}$ for the columns). The first two types of changes contribute at most an $\frac{\epsilon}{8}$-fraction of changes to the whole matrix each, and the last type contributes at most an $\frac{\epsilon}{4}$-fraction of changes. Thus M is at most $\epsilon n^2/2$-far from M_Q, and, in particular, M is in $\mathcal{M}_{Q,\epsilon}$. This proves (a), while (b) follows automatically from the definition of $\mathcal{M}_{Q,\epsilon}$ and the triangle inequality.

Hence, we may assume that with probability at least $\frac{3}{4}$ (which is the lower bound on E_2 not happening), the $\frac{\epsilon}{8}$-signature is computed correctly and (a) and (b) above are satisfied. We conclude that if M has the property then certainly some member of $\mathcal{M}_{Q,\epsilon}$ will have the property (as M itself is such a member by (a)), and thus the algorithm will accept. On the other hand, if M is more than ϵn^2-far from having the property, then no member of $\mathcal{M}_{Q,\epsilon}$ can have the property by (b).

Clearly the query complexity of the test is $O(k/\epsilon)^{O(k^2)}$, which for a fixed family F (and hence a fixed k) is polynomial in ϵ. □

The above test, while using only a constant number of queries, has a bad dependence of the calculation time on the input size (this can be alleviated somewhat, but in light of the following we omit the details). Unfortunately, this dependence is such that the automatic conversion by Alon of 2-sided tests to 1-sided ones, described in [15, Appendix D], will not work here. Instead we will go on a different route to show that a (δ, r)-partition of the matrix not only contains the necessary information about its farness from our property, but also implies the existence of many witnesses. But first, we turn back to the proofs of Lemmas 2.3 and 1.6.

3. (δ, r)-partitions, row similarity, and the proof of Lemma 2.3. Our goal here is to show that by sampling $(r/\delta)^{O(1)}$ entries in M, one can detect the signature of a (δ', r')-partition, if a (δ, r)-partition exists. For this we need a representation of a partition in a "local" way, which is asserted by Claims 3.2 and 3.3. To do this, we relate the notion of a (δ, r)-partition to relative distances between rows and columns. For the rest of this section we assume that δ is smaller than $1/81$.

For two vectors $u, v \in \{0,1\}^m$ let $\mu(u,v) = \frac{1}{m}|\{i|\ u_i \neq v_i\}|$; namely, $\mu(u,v)$ is the normalized Hamming distance between the two vectors. We will use the following definitions.

DEFINITION 3.1. *Let M be an $n \times n$ matrix. We set $E^R(\mu(r_i, r_j))$ to be the expected value of $\mu(r_i, r_j)$, where r_i, r_j are two rows of M chosen at random. Similarly let $E^C(\mu(c_i, c_j))$ denote the respective quantity where c_i, c_j are two columns chosen at random.*

Given a set of vectors V (usually either the set of rows or the set of columns of M) and a partition $V_0, \ldots, V_s$ of V, we say that the partition is a (δ, r)-clustering of *V if $s \leq r$, $|V_0| \leq \delta|V|$, and for every $1 \leq i \leq r$ and $u, v \in V_i$ we have $\mu(u,v) \leq \delta$.*

Finally, for a partition block B and a row u that intersects B, let $u|_B$ be the restriction of u to the columns in B.

There is a close correlation between (δ, r)-partitions of M and (δ, r)-clusterings of its rows and columns, as the following two claims show.

CLAIM 3.2. *Let M be a $0/1$, $m \times m$ matrix, and assume that M has a (δ, r)-partition. Then there exist a $(4\delta^{1/3}, r)$-clustering of the rows of M as well as a $(4\delta^{1/3}, r)$-clustering of the columns of M.*

CLAIM 3.3. *Let M be a $0/1$, $m \times m$ matrix, and assume that $\{R_0, \ldots, R_s\}$ and $\{C_0, \ldots, C_t\}$ are (δ^2, r)-clusterings, for $r = \max\{s, t\}$, of the set of rows and the set of columns, respectively. Then these clusterings also form a $(4\delta, r+1)$-partition of M.*

Moreover, for the above $R_0, \ldots, R_s$ and $C_0, \ldots, C_t$, a 4δ-signature for the partition is given by the sequences $\alpha_i = w(R_i)$, $i = 0, \ldots, s$, $\beta_i = w(C_i)$, $i = 0, \ldots, t$, and the $s \times t$ matrix P, where the (i,j) entry of P corresponds to the block $R_i \times C_j$ and its label is the dominant label of this block.

Before we prove the two claims we need two simple observations that in some sense correspond to the case "$r = 1$" of the claims.

OBSERVATION 3.4. *Let A be a 0/1 matrix. If A is δ-homogeneous, then $E^R(\mu(r_i, r_j)) \leq 2\delta$ and $E^C(\mu(r_i, r_j)) \leq 2\delta$.*

Proof. As A is δ-homogeneous, we may assume without loss of generality that A contains less than a δ fraction of 0's. Hence, choosing two rows at random and picking a random place i in both, the probability that they are not both "1" in this place is at most 2δ. Thus the expectation of the fraction of the number of places where they differ is bounded by 2δ, and this expectation is exactly $E^R(\mu(r_i, r_j))$. The proof for $E^C(\mu(r_i, r_j))$ is analogous. □

OBSERVATION 3.5. *If A is a 0/1 matrix such that $E^R(\mu(r_i, r_j)) < \delta$ and $E^C(\mu(c_i, c_j)) < \delta$, then A is 4δ-homogeneous.*

Proof. Assume on the contrary that A is not 4δ-homogeneous. This implies that when choosing two points from A independently and uniformly at random, with probability at least 4δ, they will not have the same label. This is also a lower bound on the fraction of the 2×2 submatrices that contain both 0's and 1's, as any two points with different labels can be extended to such a submatrix. On the other hand, if $E^R(\mu(r_i, r_j)) < \delta$, then with probability more than $1 - 2\delta$ both rows of a uniformly random 2×2 submatrix are identical, as this matrix can be expressed as choosing two random places from two random rows. By the same token, if $E^R(\mu(c_i, c_j)) < \delta$, then with probability more than $1 - 2\delta$ the two columns of a random 2×2 matrix are identical. Together these would have implied that less than a 4δ fraction of the 2×2 submatrices have both 0's and 1's, which is a contradiction. □

Proof of Claim 3.2. Assume that M has a (δ, r)-partition defined by the row partition $R_1, \ldots, R_s$ and the column partition $C_1, \ldots, C_t$, $s, t \leq r$. Assume that B is a δ-homogeneous block that contains the rows of R_i. Then by Observation 3.4, $E^R(u|_B, v|_B) \leq 2\delta$ for two rows chosen at random from R_i. For a non– δ-homogeneous block, this expression is at most 1. Let $w_i = w(R_i) = |R_i|/m$, $i = 1, \ldots, s$, and let $E_i(\mu(u, v))$ be the expectation of $\mu(u, v)$, where u, v are two rows chosen uniformly at random from R_i. Then the above implies that $\Sigma_{i=1}^{r} w_i E_i(\mu(u, v)) \leq (1 - \delta)2\delta + \delta \cdot 1 \leq 3\delta$, as this sum goes over all blocks and there is at least a $(1 - \delta)$ fraction of 0/1-blocks contributing at most 2δ each.

Now this implies that the total weight of the R_i's for which $E_i(\mu(u, v)) \geq \delta^{2/3}$ is at most $3\delta^{1/3}$. Let R_0 be the union of all these R_i's. Let $R_1, \ldots, R_{r'}$ be all other R_i's, after renumbering. For every $i = 1, \ldots, r'$, by our assumption, $E_i(\mu(u, v)) < \delta^{2/3}$ for randomly chosen u, v, so there is an $r_i \in R_i$ for which for at least a $(1 - \delta^{1/3})$ fraction of the v's in R_i, $\mu(r_i, v) < \delta^{1/3}$. Hence if we define for $1 \leq i \leq r'$ the set $R_i' = \{v \in R_i | \mu(v, r_i) < \delta^{1/3}\}$ and then define $R_0' = \bigcup_{i=1}^{r'} (R_i \setminus R_i') \cup R_0$, we obtain that $R_0', \ldots, R_{r'}'$ is indeed a $(4\delta^{1/3}, r)$-clustering for the rows of M. The proof for the existence of a clustering of the columns is analogous. □

Proof of Claim 3.3. By the assumptions of the claim, $|R_0| < \delta^2 n$. Also, for any $i \geq 1$ and any two rows $u, v \in R_i$, $\mu(u, v) \leq \delta^2$. Thus for $i = 1, \ldots, s$, $E_i(\mu(u, v)) \leq \delta^2$, where E_i is the expectation when u, v are chosen at random from R_i. Hence for the above partition into rows, $\Sigma_{i=0}^{s} \frac{|R_i|}{m} E_i(\mu(u, v)) \leq 2\delta^2$ (as for each $i > 1$ the corresponding term in this average is at most δ^2, and for $i = 0$ the weight of the term is at most δ^2). Similarly we get the analogous inequality for columns. Let $\mathcal{P}$ be the partition of M into blocks that is defined by the cross product of the two partitions above.

Recall that $\frac{|R_i|}{m}, \frac{|C_i|}{m}$ are the weights $w(R_i), w(C_i)$ of the corresponding sets. Also, for a block B, let $E_R(\mu(u|_B, v|_B))$, respectively, $E_C(\mu(u|_B, v|_B))$, be the expectation of $\mu(\cdot, \cdot)$ for two rows u, v, respectively, columns, chosen at random from B. By the

law of complete probability, $\Sigma_{i=0}^{s} w(R_i) \cdot E_i(\mu(u,v)) = E_B(E_R(\mu(u|_B, v|_B)))$, where in the right-hand side the outer expectation is on blocks of $\mathcal{P}$ chosen according to their weights, and the inner expectation is on rows chosen at random in the block. Hence, the fact that $\Sigma_{i=0}^{s} w(R_i) E_i(\mu(u,v)) \leq 2\delta^2$ implies that the total weight of all blocks B for which $E_R(\mu(u|_B, v|_B)) > \delta$ is bounded by 2δ. By the same argument, for at most a 2δ fraction of the blocks $E_C(\mu(u|_B, v|_B)) > \delta$. Hence, for at least a $1 - 4\delta$ fraction of the blocks (weighted by the block weights), both $E_R(\mu(u|_B, v|_B)) \leq \delta$ and $E_C(\mu(u|_B, v|_B)) \leq \delta$. However, by Observation 3.5 above, each such block is 4δ-homogeneous, and hence at most a 4δ fraction of the blocks (measured by weights) are not 4δ-homogeneous. This implies that $\mathcal{P}$ is a $(4\delta, r+1)$-partition. Also, by definition, a pattern for this partition is any one that has, for each block, the $(1-4\delta)$-dominant label of this block if there is one, or an arbitrary value otherwise. Moreover, as α_i, β_i are the exact weights of the parts in the partition, we get a 4δ-signature for it by definition. □

We are now ready to present the testing algorithm that yields Lemma 2.3. We start with a trivial observation about approximating distances.

CLAIM 3.6. *Let $u, v \in \{0,1\}^n$, $\gamma < 1$. Choose randomly and independently (with repetitions) m elements of $[n]$, naming the resulting (multi)set $L = \{l_1, \ldots, l_m\}$. Let $\tilde{\mu}(u,v) = \frac{1}{m} \sum_{k=1}^{m} |u(l_k) - v(l_k)|$, where $u(i)$ and $v(i)$ are the ith coordinates of u and v, respectively. Then $|\mu(u,v) - \tilde{\mu}(u,v)| \leq \gamma$ with probability at least $1 - 2exp(-\gamma^2 m)$.*

Proof. The proof is immediate by a Chernoff-type inequality (see, e.g., [5, Corollary A.1.7]). □

We next construct a testing algorithm for an approximate notion of clustering. Testing algorithms for clustering were already investigated in [2]; here we will use a simple self-contained proof for an algorithm that gives an approximation in a very weak sense.

LEMMA 3.7. *There exists an approximate oracle algorithm that makes $(r/\delta)^{O(1)}$ bit queries (queries of one coordinate of one vector) to a set V of vectors over $\{0,1\}^n$, such that if V has a (δ, r)-clustering then the algorithm provides a $(4\delta, 10r^2/\delta)$-clustering of V as follows.*

The algorithm makes $(r/\delta)^{O(1)}$ queries in a preprocessing step, and with probability at least 0.9 provides a clustering oracle for V in the following sense: There exists a $(4\delta, 10r^2/\delta)$-clustering $V_0', \ldots, V_t'$ of V, such that for every specified $v \in V$ the algorithm can make $(r/\delta)^{O(1)}$ additional queries to provide an index $0 \leq i_v \leq t$, where it is guaranteed that for at least a $(1-4\delta)$ fraction of the vectors $v \in V$ the provided i_v will satisfy $v \in V_{i_v}$.

Proof. Suppose that $V_0, \ldots, V_s$ is a (δ, r)-clustering of V. The algorithm starts by selecting uniformly at random $r' = 10r^2/\delta$ vectors $v_1, \ldots, v_{r'}$ from V. With probability at least 0.95 (assuming that r is large enough) the situation is that for every $1 \leq i \leq r$ for which $|V_i| \geq \delta|V|/r$, we have picked at least one vector from V_i.

We now pick uniformly at random (with repetitions) $l = (10r' \log r')/\delta$ coordinates from $1, \ldots, n$, and let $\tilde{\mu}(\cdot,\cdot)$ denote the corresponding approximated distance. Claim 3.6 implies that for every $v, v' \in V$, the probability for $|\mu(v,v') - \tilde{\mu}(v,v')| > \frac{1}{2}\delta$ is bounded by $\delta/20r'$, and so with probability at least 0.95 the situation is that for at least a $(1-\delta)$ fraction of the vectors $v \in V$, $|\mu(v,v_i) - \tilde{\mu}(v,v_i)| \leq \frac{1}{2}\delta$ for every $1 \leq i \leq r'$.

Assuming that both of the above events occurred (which is the case with probability at least 0.9), we define $V_0', \ldots, V_{r'}'$ as follows. Every vector v that belongs to V_0, or that belongs to a V_i of size $|V_i| < \delta/r$, or such that there exists some v_i for which

$|\mu(v,v_i) - \tilde{\mu}(v,v_i)| > \frac{1}{2}\delta$, is placed in V_0'. For every other vector we let i be the index for which $\tilde{\mu}(v,v_i)$ is minimal (or the smallest such index if there exist several values that minimize $\tilde{\mu}(v,v_i)$), and define v to be in V_i'.

We claim that $V_0', \ldots, V_{r'}'$ is indeed a $(4\delta, r')$-clustering. First, it is easy to see that $|V_0'| \leq 3\delta|V| < 4\delta|V|$ from the assumption on the size of V_0, and the guarantee that we have on the number of vectors for which the distance was not well approximated. Now, if $u, v \in V_i'$ for some $1 \leq i \leq r'$, then we first note that $\mu(u,v_i) \leq 2\delta$. This is because if we denote by $1 \leq j \leq r$ the index for which $u \in V_j$, then we have $\mu(u,v_i) \leq \tilde{\mu}(u,v_i) + \frac{1}{2}\delta \leq \tilde{\mu}(u,v_j) + \frac{1}{2}\delta \leq \mu(u,v_j) + \delta \leq 2\delta$. The same goes for proving that $\mu(v,v_i) \leq 2\delta$, and so by the triangle inequality $\mu(u,v) \leq 4\delta$. This concludes the claim about $V_0', \ldots, V_{r'}'$.

We now describe the remainder of the algorithm: After choosing $v_1, \ldots, v_{r'}$ and the l coordinates as above, the algorithm now queries each of these coordinates from each v_i, and by this concludes the preprocessing stage. For the oracle stage, given a vector $v \in V$ the algorithm queries all the l chosen coordinates of v, and then calculates $\tilde{\mu}(v,v_i)$ for every i. The algorithm then outputs the index i that minimizes this, or the smallest such index in case there is more than one. It is clear that the algorithm gives the correct index for every vector that is not in V_0', whose size is bounded by 4δ, concluding the proof. □

We note here that we could also use the above to find an approximate oracle for a $(4\delta, r)$-clustering (instead of a $(4\delta, 10r^2/\delta)$-clustering), by trying to get from the set of queried vectors a subset V' for which all but at most a 3δ fraction of the members of V are δ-close to a member of V' (and verifying the validity of V' using a polynomial number of additional queries). This would also improve the dependencies in Lemma 2.3, but we omit it as our proofs already ensure the polynomial dependence on ϵ without this improvement.

We are now ready to describe the algorithm that proves Lemma 2.3, by finding with probability $\frac{3}{4}$ a signature of a $(16\delta^{1/6}, 10r^2/(4\delta^{1/3})+1)$-partition of M, if M has a (δ, r)-partition.

Algorithm Sig.

- By Claim 3.2, there exists a $(4\delta^{1/3}, r)$-clustering of the rows. We perform the preprocessing stage of the algorithm provided by Lemma 3.7 to obtain an approximate oracle for a $(16\delta^{1/3}, 10r^2/(4\delta^{1/3}))$-clustering of the set of rows of M; we denote it by $R_0', \ldots, R_{r'}'$ for $r' = 10r^2/(4\delta^{1/3})$. Similarly, we obtain an approximate oracle for a $(16\delta^{1/3}, r')$-clustering $C_0', \ldots, C_{r'}'$ of the columns.
- We now choose uniformly and independently at random (with repetitions) a (multi)set R of $l = (100r' \log r')/\delta$ rows of M, and for each of these we use the clustering oracle for $R_0', \ldots, R_{r'}'$. For $1 \leq i \leq r'$, we set α_i to be the number of rows from R for which the oracle answered "i," divided by l. We do the analogous operation for a set C of l columns M that were uniformly and independently chosen (this time with respect to the oracle for $C_0', \ldots, C_{r'}'$), and use it to set β_i for $1 \leq i \leq r'$. Both α_0 and β_0 are set to 0, as the above oracles never correctly detect that a row is in R_0' or a column is in C_0'.
- Finally, for every $1 \leq i \leq r'$ and $1 \leq j \leq r'$ we look at the intersections of all the rows in R which the oracle located in R_i', and all the columns in C which the oracle located in C_j'. We query the entries of M at the intersections of the set of sampled rows R and the set of sampled columns C, and we set $P_{i,j}$ to be the value (0 or 1) that has the majority of appearances in these queries.

We now claim that this algorithm satisfies the assertion of Lemma 2.3. First, we

note that with probability at least 0.8, the oracles for both the clustering of the rows and the clustering of the columns are valid, as guaranteed by Lemma 3.7. In turn this guarantees that $R'_0, \ldots, R'_{r'}$ and $C'_0, \ldots, C'_{r'}$ form a $(16\delta^{1/6}, r'+1)$-partition of M, by Claim 3.3. Also, each of the following occurs with probability at least 0.99:

- The difference between every α_i and the total fraction of the rows of M for which the oracle would output "i" is at most δ/r'. This implies that $\sum_{i=0}^{r'} |\frac{|R'_i|}{n} - \alpha_i| \leq 2 \cdot 16\delta^{1/3} + r' \cdot \delta/r' < 33\delta^{1/3}$.
- Similarly to the above, $\sum_{i=0}^{r'} |\frac{|C'_i|}{n} - \beta_i| < 33\delta^{1/3}$. With the previous item this means that for all but at most a $10\delta^{1/6}$ fraction of the pairs (i,j), both $|\frac{|R'_i|}{n} - \alpha_i| \leq 7\delta^{1/6}$ and $|\frac{|C'_j|}{n} - \beta_j| \leq 7\delta^{1/6}$.
- The fraction of appearances of "1" in the values taken under consideration when calculating $P_{i,j}$ differs from the fraction of appearances in the intersections of all rows assigned to "i" and all columns assigned to "j" (by the oracles) by no more than δ. In addition, by the previous item for all but at most a $10\delta^{1/6}$ fraction of the pairs (i,j), the above fraction differs by no more than $14\delta^{1/6}$ from the fraction of appearances of "1" in $R'_i \times C'_j$, and so (if δ is small enough) for the $16\delta^{1/6}$-homogeneous blocks among these, $P_{i,j}$ will get the correct value. Hence, the (weighted) fraction of wrong $P_{i,j}$ labels is no more than $16\delta^{1/6} + 10\delta^{1/6} = 26\delta^{1/6}$.

Therefore, with probability at least $\frac{3}{4}$ all the above occurs (including the two oracles being valid), and a $26\delta^{1/6}$-signature of a $16\delta^{1/6}$-partition is obtained. □

As a final remark, the proof of Lemma 1.6, given in the next section, also uses an interim lemma about clusterings, Lemma 4.1 below. One could save further on the number of queries in the main theorem if the notion of (δ, r)-clustering would be used throughout instead of the notion of (δ, r)-partitions, but it would still be polynomial (not linear) in ϵ. However, the notion of (δ, r)-partitions is more intuitive and could have applications outside the scope of this work, so we use it instead.

4. Proof of Lemma 1.6. We use the same definition of a (δ, r)-clustering (for sets of rows or columns) as we used in the previous section. Claim 3.3, which was proved above, implies that if A has a $(\delta^2/16, t)$-clustering for both its rows and its columns, then A admits a $(\delta, t+1)$-partition. Therefore, the following lemma immediately implies Lemma 1.6. Moreover, it follows that Lemma 1.6 is true even if we insist on the forbidden submatrices also obeying the order of the rows and the columns of the input matrix (which is ignored for our use of a matrix as representing a bipartite graph).

LEMMA 4.1. *Let k be a fixed integer and let $\delta > 0$ be a small real. For every $n \times n$, 0/1-matrix A, with $n > (k/\delta)^{O(k)}$, either A admits (δ, r)-clusterings for both the rows and columns with $r \leq (k/\delta)^{O(k)}$, or for every $k \times k$, 0/1 matrix F, at least a $(\delta/k)^{O(k^2)}$ fraction of the $k \times k$ (ordered) submatrices of A are copies of F.*

We should also note that the above estimate is essentially tight, as shown by a random $n \times n$ matrix A, where each entry is independently chosen to be 1 with probability 2δ, and 0 with probability $1 - 2\delta$. The expected number of copies of the $k \times k$ all 1 matrix in such a matrix is only a $(2\delta)^{k^2}$ fraction of the total number of $k \times k$ submatrices, and it is not difficult to check that with high probability A does not have a $(\delta, o(n))$-clustering for either its rows or its columns.

We will prove the lemma only for the clustering of the columns, because the proof for rows is virtually identical. We make no attempt to optimize the absolute constants and omit all floor and ceiling signs to simplify the presentation. In order to prove the

above lemma, we first need the following simple corollary of Sauer's lemma [18, 20].

LEMMA 4.2. *For every $t > 10k$, every $t \times t^{2k-1}$ binary matrix M with no two identical columns contains every possible $k \times k$ binary matrix as a submatrix.*

Proof. By Sauer's lemma [18, 20], every set of $s = 1 + \sum_{i=0}^{k-1} \binom{t}{i}$ consecutive columns of M contains a $k \times 2^k$ submatrix that has no two identical columns (and so contains all 2^k possible binary vectors as columns). Note that $s < t^{k-1}$ and $s(1 + (k+1)\binom{t}{k}) \leq t^{2k-1}$. Thus M can be partitioned into at least $1 + (k+1)\binom{t}{k}$ blocks of size $t \times s$, each consisting of s consecutive columns. Considering these $1 + (k+1) \cdot \binom{t}{k}$ pairwise disjoint consecutive blocks, we now find in each of them a $k \times 2^k$ submatrix with no identical columns. Considering now the set of k rows in each such submatrix, we obtain by the pigeonhole principle k such submatrices of size $k \times 2^k$, all having the same set of rows, such that their column sets are contained in disjoint intervals (according to the column order of M), one following the other. This implies the desired result, as we can choose from each of the submatrices a desired column and thus construct any given $k \times k$ matrix. □

We now turn to the proof of Lemma 4.1. Fix δ and k, and suppose that n is large enough (as a function of δ and k, to be chosen later). Let t be the smallest integer for which $(1 - \frac{1}{2}\delta)^t t^{4k-2} < 0.1$. A simple computation shows that $t = O(\frac{k}{\delta} \log(\frac{k}{\delta}))$. Define $T = t^{2k-1}$ and suppose that A is an $n \times n$ matrix with 0/1 entries which does not have a δ-clustering of the columns of size T. We have to show that in this case A must contain many copies of every $k \times k$ matrix F.

Indeed, let S be a random set of columns of A obtained by choosing, randomly, uniformly, and independently (with repetitions) $\tau = 5T/\delta$ columns of A. We assume that $n > 10(\frac{5T}{\delta})^2$. Note that, in particular, for such an n, with probability at least 9/10 no column is chosen more than once.

CLAIM 4.3. *With probability at least* 0.9*, S contains a subset S' of T columns so that the Hamming distance between any pair of them is at least $\frac{1}{2}\delta n$.*

Proof. Let us choose the members of S one by one and construct, greedily, a subset S' of S consisting of columns so that the Hamming distance between any pair of them is at least $\frac{1}{2}\delta n$ as follows. The first member of S belongs to S', and for all $i > 1$, the ith chosen column of S is added to S' if its Hamming distance from every previous member of S' is at least $\frac{1}{2}\delta n$. Since, by assumption, there is no (δ, T)-clustering of the columns of A, as long as the cardinality of S' is smaller than T, the probability that the next chosen member of S will be added to S' is at least δ (given any history of the previous choices); otherwise it would mean that the balls of radius $\frac{1}{2}\delta n$ around the members of S' form a δ-clustering. It thus follows that the probability that by the end of the procedure the cardinality of S' will still be smaller than T is at most the probability that a binomial random variable with parameters $5T/\delta$ and δ will have value at most T. Hence this probability is smaller than 0.1, which implies the assertion of the claim. □

The usefulness of S' as above is shown by the following claim.

CLAIM 4.4. *Let S' be a fixed set of T columns of A for which the pairwise Hamming distance is at least $\frac{1}{2}\delta n$. Then, if we choose a random set R of t rows of A by choosing them independently and uniformly at random, with probability at least* 0.9 *all the projections of the members of S' on the rows in R are distinct.*

Proof. Let S' be a fixed set of T columns of A so that the Hamming distance between every pair is at least $\frac{1}{2}\delta n$. For any two fixed columns $c_1, c_2 \in S'$ and a random row r we have that the probability that $c_1[r] = c_2[r]$ is at most $1 - \frac{1}{2}\delta$, where $c[j]$ denotes the jth coordinate of c. Hence, the expected number of pairs of members

of S' whose projections on R are identical is at most $\binom{T}{2}(1-\frac{1}{2}\delta)^t < 0.1$, where the last inequality follows from the choice of t. The desired result follows. □

We can now conclude the proof of Lemma 4.1 as follows. Fix F to be any $k \times k$, 0/1 matrix. Choosing a random $t \times \tau$ submatrix C of A is just like choosing a set R of t random rows and a set S of τ random columns. By Claim 4.3, with probability at least 0.9, the set S of τ columns contains a subset of the columns S' of size T that has pairwise distances at least $\frac{1}{2}\delta n$. Given that this happens, by Claim 4.4 with probability 0.9 all the t projections of S' on the t rows of C are distinct. Hence with probability at least 0.8 (the probability that both events above hold) Lemma 4.2 ensures that C contains F as a submatrix.

Now choosing a random $k \times k$ submatrix of A can be viewed as first choosing a random $t \times \tau$ matrix C as above and then choosing a random subset of k columns and k rows in C. Hence the probability that such a random $k \times k$ matrix will be identical to F is at least $0.8/(\binom{t}{k}\binom{\tau}{k}) = (\frac{\delta}{k})^{O(k^2)}$.

5. Unfoldable graphs and 1-sided testing. To construct a 1-sided test that is polynomial in ϵ, one would like to use the following scheme. First, the case where there is no (δ, r)-partition (for the appropriate parameters) is covered also for 1-sided algorithms by Lemma 1.6. Now, assuming that M is ϵ-far from $\mathcal{S}_F$ and has a (δ, r)-partition, using Lemma 2.3, we can find a submatrix Q that has a (δ', r)-partition with a signature similar to a (δ', r)-partition of M. We would like to show that in this case Q contains a member of F which will provide a witness for rejecting M.

However, having a Q with the same signature as a matrix M that is ϵ-far from $\mathcal{S}_F$ still does not imply that Q contains a member of F, because some of the partition blocks of Q may not be homogeneous and so their behavior may depend on n (this was circumvented in the 2-sided algorithm by checking all $n \times n$ matrices that are compatible with the signature). One way to solve this would be to use a Ramsey-like lemma like the one used in [11] to get rid of nonhomogeneous blocks, but this would create an exponential blow-up in the number of queries.

Here we take a different approach. First, in this section we prove the existence of the test only for the case where it is enough for Q to have only one row and one column from every cluster of the partition of M, and so the issue of homogeneity becomes moot. Later, we will use this special case as a lemma to prove the general case.

DEFINITION 5.1. *A matrix M is called* unfoldable *if it contains no two identical rows and no two identical columns. Equivalently, an unfoldable bipartite graph is one that has no two vertices (on the same side) with exactly the same set of neighbors.*

A family F of matrices is called unfoldable *if all its members are unfoldable.*

The main lemma that we will prove in this section essentially states that properties definable by unfoldable matrices are testable.

LEMMA 5.2. *For every ϵ, k, and a family F of unfoldable $k \times k$ or smaller matrices, there exists $\delta = (\epsilon/k)^{O(k^2)}$ such that if an $n \times n$ matrix M, where $n > (k/\epsilon)^{O(k)}$, is ϵ-far from the property $\mathcal{S}_F$, then M contains at least δn^{2k} distinct submatrices containing members of F (up to permutations).*

What we will need to use for the general case is the following corollary. In the next section we will use it on the *signature* of M to avoid dealing at all with blocks of M that are not homogeneous.

COROLLARY 5.3. *For every ϵ, k, and a family F of unfoldable $k \times k$ or smaller matrices, there exists $\delta = (\epsilon/k)^{O(k^2)}$ such that if an $n \times n$ matrix M, where $n > (k/\epsilon)^{O(k)}$, is ϵ-far from the property $\mathcal{S}_F$, then for every set X of δn^2 entries, M*

contains a member of F (up to permutations) that does not include any entry from X.

Proof. Every set X can clearly intersect at most $|X| \cdot \binom{n-1}{k-1}^2 < |X|n^{2k-2}$ submatrices of M. Hence, if $|X| < \delta n^2$, then Lemma 5.2 implies that, in particular, there exists a copy of a forbidden submatrix which does not intersect X. □

To prove Lemma 5.2, and also for the next section, it is more convenient to work with partitions into equally sized blocks.

DEFINITION 5.4. *An r-partition of an $n \times n$ matrix M is called an* r-equipartition *if the size of all the sets R_i and C_j lie between $\lfloor n/r \rfloor$ and $\lceil n/r \rceil$. In an analogous manner we define a* (δ, r)-equipartition.

Note that for (δ, r)-equipartitions, a δ-signature essentially holds no more information than the δ-pattern it includes. The conditional existence of (δ', r')-equipartitions follows from that of (δ, r)-partitions by the following simple lemma.

LEMMA 5.5. *For $\delta < \frac{1}{4}$, if a matrix M admits a (δ, r)-partition, then it admits also a $(\sqrt{\delta} + 3\delta, r/\delta)$-equipartition.*

Proof. For simplicity we assume that $l = \delta n/r$ is an integer. We repartition the original (δ, r)-partition of M in the following manner. From every R_i whose size is at least l we randomly and uniformly pick $s = \lfloor |R_i|/l \rfloor$ disjoint subsets $R_{i,1}, \ldots, R_{i,s}$ of size l. We call the matrix rows not picked for any $R_{i,x}$ by this procedure *leftover rows*. We now arbitrarily partition the set of leftover rows into disjoint sets of size l. We then perform the analogous procedure for the columns of the matrix M.

Now for every i and j such that $R_i \times C_j$ was δ-homogeneous, every block $R_{i,p} \times C_{j,t}$ will be $\sqrt{\delta}$-homogeneous with probability at least $1 - \sqrt{\delta}$. To see this assume without loss of generality that $R_i \times C_j$ has at most a δ-fraction of 1's. Then, for any fixed p, t, a random submatrix $R_{i,p} \times C_{j,t}$ of $R_i \times C_j$ has the same expected average value of its entries as the average value for $R_i \times C_j$, which is at most δ. Hence, by the Markov inequality, the probability that $R_{i,p} \times C_{j,t}$ will have more than a $\sqrt{\delta}$ fraction of 1's is at most $\sqrt{\delta}$. This probability is, however, the failure probability of $R_{i,p} \times C_{j,t}$ being $\sqrt{\delta}$-homogeneous.

Thus, there is a choice of the repartitions above for which the number of blocks $R_{i,p} \times C_{j,t}$ that come from δ-homogeneous blocks $R_i \times C_j$ but are not themselves $\sqrt{\delta}$-homogeneous is not more than $\sqrt{\delta}(n/l)^2$.

Also, since the original partition was δ-homogeneous, there are no more than $\delta(n/l)^2$ blocks $R_{i,p} \times C_{j,t}$ that come from blocks of the original partition that are not δ-homogeneous. Finally, there are the blocks that are related to leftover rows and columns. From the procedure it follows that there are no more than $lr \leq \delta n$ leftover rows and no more than lr leftover columns. Thus the total number of such blocks is no more than $2\delta(n/l)^2$.

Counting all the above we obtain a total of not more than $(\sqrt{\delta} + 3\delta)(n/l)^2$ blocks that are not $\sqrt{\delta}$-homogeneous, and so the same bound holds also for non-$(\sqrt{\delta} + 3\delta)$-homogeneous blocks. □

LEMMA 5.6. *Let k be fixed. For every $0 < \delta < \frac{1}{4}$ and any $n \times n$, 0/1-matrix M, with $n > (k/\delta)^{O(k)}$, either M has a (δ, t)-equipartition for $t = t(\delta, k) \leq (k/\delta)^{O(k)}$, or for every 0/1-labeled $k \times k$ matrix B, an $h(\delta, k) \geq (\delta/k)^{O(k^2)}$ fraction of the $k \times k$ submatrices of M are B.*

Proof. We set $h(\delta, k) = g(\delta^2/16, k)$ and $t(\delta, k) = 16r(\delta, k)/\delta^2$, where g and r are the functions of Lemma 1.6. If M does not contain an h fraction of $k \times k$ submatrices that are identical to B, then it admits a $(\delta^2/16, r)$-partition as per Lemma 1.6. But

then this implies that M admits a (δ, t)-equipartition by Lemma 5.5. □

The following lemma is the main technical tool, showing that the existence of a (δ, r)-partition (for the appropriate parameters) implies a dichotomy between being close to $\mathcal{S}_F$ and containing many forbidden matrices from F.

LEMMA 5.7. *Let F be an unfoldable family of $k \times k$ or smaller matrices. Furthermore, let M be a matrix, and let P be an $\epsilon/8$-pattern of an $(\epsilon/8, t)$-equipartition of M for $t > 4k^2$. If P is $\epsilon/2$-close to $\mathcal{S}_F$, then M itself is ϵ-close to $\mathcal{S}_F$, while if P is $\epsilon/2$-far from $\mathcal{S}_F$, then M contains at least $\Omega(n/t)^{2k}$ distinct $k \times k$ matrices containing members of F (up to permutations).*

Proof. Let $R_1, \ldots, R_t$ and $C_1, \ldots, C_t$ be the $(\epsilon/8, t)$-equipartition of M, and let P be the corresponding $(\epsilon/8)$-pattern. If P is indeed $\epsilon/2$-close to $\mathcal{S}_F$, then let P' be the $\epsilon/2$-close matrix containing no members of F. Now modify M by setting every entry of M to be identical to the entry of P' corresponding to its block in the $(\epsilon/8, t)$-equipartition. Denote the modified matrix by M'. M' is ϵ-close to M, because the modified entries can only correspond to either entries where P and P' differed (a total of at most $\epsilon/2n^2$ entries), or entries that correspond to blocks that are not good with respect to P (at most $\epsilon/8n^2$), or entries that correspond to good blocks (at most $\epsilon/8n^2$, as in every good block the corresponding entry of P is $\epsilon/8$-dominant). Now since F is unfoldable, M' cannot contain members of F unless all their rows are in distinct R_i and all their columns are in distinct C_j. But then because P' contains no member of F, neither does M'.

We now assume that P is $\epsilon/2$-far from containing no member of F, and calculate the probability that a uniformly random $k \times k$ submatrix A of M is not a member of F. For simplicity we assume that t divides n. Recalling that $t > 4k^2$ we first note that with probability at least $\frac{1}{2}$ this matrix has no two rows in the same R_i and no two columns in the same C_j. Now, we condition the distribution of A on this event and note that it is identical to the one resulting from the following procedure: First choose uniformly, randomly, and independently a row $r_i \in R_i$ for every $1 \leq i \leq t$ and a column $c_j \in C_j$ for every $1 \leq j \leq t$. Denoting this matrix by Q, now let A be a uniformly random $k \times k$ submatrix of Q.

Because P is an $(\epsilon/8)$-pattern of the equipartition, no more than an $\epsilon/8$ fraction of the entries of M that make up Q come from blocks which are not $\epsilon/8$-good with respect to P. For an entry $Q_{i,j}$ of Q that does come from an $\epsilon/8$-good block $R_i \times C_j$, with probability at least $1 - \epsilon/8$ the value of $Q_{i,j}$ is identical to $P_{i,j}$. This implies that for the random set of entries of M that makes up Q, the expectation of the fraction of entries $Q_{i,j}$ that are consistent with the corresponding $P_{i,j}$ is at least $1 - \epsilon/4$. Hence, with probability at least $\frac{1}{2}$ the matrix Q is $\epsilon/2$-close to P, and so contains a member of F. Now conditioned on this event, the probability that A contains the forbidden submatrix is at least t^{-2k}. Putting all the above together using Bayes's law, the unconditional probability that a uniformly random A contains a forbidden submatrix is at least $t^{-2k}/4$, completing the proof. □

We can now put together the proof of Lemma 5.2 that concludes this section.

Proof of Lemma 5.2. If M is ϵ-far from $\mathcal{S}_F$ (where F is unfoldable), then there are two possible cases for M. Either it contains an $(\epsilon/8, t)$-equipartition for $t(\epsilon/8, k)$ as in Lemma 5.6, or M does not contain such an equipartition.

In the second case, Lemma 5.6 ensures that an $(\epsilon/k)^{O(k^2)}$ fraction of the $k \times k$ matrices are identical to an arbitrary member of F, so we are done.

In the first case, let P be an $\epsilon/8$-pattern of the equipartition of M. By Lemma 5.7 P itself cannot be $\epsilon/2$-close to $\mathcal{S}_F$ (as this would contradict the assumption that

M is ϵ-far from $\mathcal{S}_F$), and so P is $\epsilon/2$-far from $\mathcal{S}_F$. But then Lemma 5.7 implies that there is at least an $\Omega(t^{-2k}) = (\epsilon/k)^{O(k^2)}$ fraction of the $k \times k$ submatrices of M, such that each of these $k \times k$ submatrices contains members from F, as required. ∎

6. 1-sided testing for general bipartite graphs. Given a family F of forbidden submatrices that may contain foldable ones, we will first construct a family $\tilde{F}$ that is related to F and is unfoldable.

DEFINITION 6.1. *For a matrix A, we define the* folding *of A as the matrix $\tilde{A}$ resulting from A after removing all duplicate rows and columns, keeping only one of each.*[1]

For a family of matrices F, we define the folding *of F as the family $\tilde{F}$ consisting of all the foldings of the members of A.*

The main technical tool here is proven similarly to Lemma 5.7, but here we actually use Corollary 5.3 for the signature first, to address the possibility of having some nonhomogeneous blocks in our equipartition.

LEMMA 6.2. *Let F be a family of $k \times k$ or smaller matrices, and let $\tilde{F}$ be the folding of F. Furthermore, let M be a matrix, and let P be a δ-pattern of a (δ, t)-equipartition of M, for $t \geq (k/\epsilon)^{O(k)}$ and $\delta = (\epsilon/k)^{O(k^2)}$. If P is $\epsilon/2$-close to $\mathcal{S}_{\tilde{F}}$, then M itself is ϵ-close to $\mathcal{S}_F$, while if P is $\epsilon/2$-far from $\mathcal{S}_{\tilde{F}}$, then M contains at least $\Omega(n/kt)^{2k}$ distinct $k \times k$ matrices containing members of F (up to permutations).*

Proof. Let $R_1, \ldots, R_t$ and $C_1, \ldots, C_t$ be the (δ, t)-equipartition of M. If P is indeed $\epsilon/2$-close to $\mathcal{S}_{\tilde{F}}$, then let P' be the $\epsilon/2$-close matrix containing no members of $\tilde{F}$. Now modify M by setting every entry of M to be identical to the entry of P' corresponding to its block in the (δ, t)-equipartition. Denote the modified matrix by M'. As in the proof of Lemma 5.7, it is not hard to see that M' is ϵ-close to M. Now M' cannot contain a member of F (up to permutations) unless P' contains a folding of this member, which is a contradiction as $\tilde{F}$ is the folding of F.

We now assume that P is $\epsilon/2$-far from containing no member of $\tilde{F}$ and calculate the probability that a uniformly random $k \times k$ submatrix A of M is not a member of F. For simplicity we assume that t divides n. We note that the distribution of picking a uniformly random $k \times k$ submatrix A is identical to the distribution of the following procedure: First choose uniformly, randomly, and independently k distinct rows $r_{i,1}, \ldots, r_{i,k} \in R_i$ for every $1 \leq i \leq t$, and k distinct columns $c_{j,1}, \ldots, c_{j,k} \in C_j$ for every $1 \leq j \leq t$. Denoting this matrix by Q, we now let A be a uniformly random $k \times k$ submatrix of Q.

Since P is a δ-pattern of the equipartition, the probability that a random entry x in M is equal to $P_{i,j}$ given that $x \in R_i \times C_j$ and that $R_i \times C_j$ is δ-good is at least $1 - \delta$. Thus, for a δ-good block, with probability at most δ its intersection with Q is not a $k \times k$ matrix whose entries are all identical to the corresponding label of P. Because P is a δ-pattern of the equipartition, the expectation of the number of blocks $R_i \times C_j$ for which their intersection with Q is not a $k \times k$ matrix whose entries are all identical to the corresponding label of P is no more than $2k^2\delta t^2$. We let X denote the set of entries of P corresponding to all such bad blocks. Let E be the event that $|X| \leq 8k^2\delta t^2$. Clearly E occurs with probability at least $3/4$.

By Corollary 5.3, for X as above and the matrix P, there is a member of $\tilde{F}$ in P whose entries are disjoint from X (for an appropriate choice of the coefficient

[1] Note that if we remove one of two or more identical rows, the identity relations between columns remain exactly the same, and conversely the identity relations between rows remain exactly the same if we remove duplicate columns. Hence, the order in which we remove duplicates does not affect $\tilde{A}$ apart from a possible permutation in its rows and columns.

in the O notation in the expression of δ, and in the lower bound condition on t). However, if P contains a copy of a member $\tilde{B}$ of $\tilde{F}$ whose entries are disjoint from X, then Q contains the member B of F whose folding is $\tilde{B}$. Now conditioned on the event E, the probability that A contains the forbidden submatrix is at least $(kt)^{-2k}$. Putting all of the above together using Bayes's law, the unconditional probability that a uniformly random A contains a forbidden submatrix is at least $(kt)^{-2k}/4$, completing the proof. □

This allows us to conclude with the lemma yielding the 1-sided test.

LEMMA 6.3. *For every ϵ and k there exists $\eta = (\epsilon/k)^{O(k^4)}$ such that if an $n \times n$ matrix M where $n > (k/\epsilon)^{O(k^3)}$ is ϵ-far from the property $\mathcal{S}_F$, where F is a family of $k \times k$ or smaller matrices, then M contains at least ηn^{2k} distinct submatrices containing members of F (up to permutations).*

Proof. We set $\delta = (\epsilon/k)^{O(k^2)}$ as required from Lemma 6.2 and set $t = t(\delta, k) = (k/\epsilon)^{O(k^3)}$ as per Lemma 5.6. Now if M is ϵ-far from $\mathcal{S}_F$, then either M contains a (δ, t)-equipartition or it does not.

In the second case, Lemma 5.6 ensures that there is a $(\delta/k)^{O(k^2)} = (\epsilon/k)^{O(k^4)}$ fraction of the $k \times k$ matrices, such that each of these matrices is identical to an arbitrary member of F, so we are done.

In the first case, let P be a δ-pattern of the equipartition of M. By Lemma 6.2 P itself cannot be $\epsilon/2$-close to $\mathcal{S}_{\tilde{F}}$ (as this would contradict the assumption that M is ϵ-far from $\mathcal{S}_F$), and so P is $\epsilon/2$-far from $\mathcal{S}_{\tilde{F}}$. But then Lemma 6.2 implies that M contains at least an $\Omega((tk)^{-2k}) \geq (\epsilon/k)^{O(k^4)}$ fraction of the $k \times k$ submatrices of M that contain members from F, as required. □

COROLLARY 6.4. *The property $\mathcal{S}_F$ is ϵ-testable with $(\epsilon/k)^{O(k^4)}$ many queries.*

Proof. Using the η of Lemma 6.3, select independently $3/\eta$ uniformly random $k \times k$ submatrices of M, and for each of them, check whether it contains a member of F. □

7. Open problems.

More general combinatorial structures. A long standing question in graph property testing is that of whether there exists a test for the property of a (general) graph being triangle-free, whose number of queries is less than a tower function in ϵ. Noting the "conditional regularity" nature of Lemma 1.6 here, one would hope for an analogue that will work for triangles. However, formulating such an analogue is not as simple as it seems: Gowers [12] constructed a bipartite (hence triangle-free) graph in which there is a tower lower bound on the size of the smallest regular partition. Hence, the only hope would be of finding a partition in which most of the nonregular pairs are somehow labeled as "irrelevant" for the existence of a triangle in the graph. This still remains open; we already know, however, by [1] that, unlike the case of bipartite graphs, a polynomial dependency (in $1/\epsilon$) is not possible for this case.

Another interesting open question would be to formulate a lemma in the spirit of Lemma 1.6 for higher dimensional matrices that would in turn correspond to r-partite r-uniform hypergraphs. Here too there is probably no avoiding the existence of "irrelevant" portions for which there is no regularity. Take, for example, any 3-dimensional matrix which is constant along the last dimension; it does not contain, for example, the $2 \times 2 \times 2$ matrix that is all zero apart from exactly one entry, while it may still not admit any relatively small regular partition.

Matrices with row and column order. This direction seems at the moment more accessible than those outlined above. It would be interesting to test a matrix

for the property of not containing a member of a forbidden family of submatrices, with the same row and column orders (i.e., containing a nontrivial row or column permutation of a forbidden matrix is now allowed). Lemma 1.6 also holds for this framework, so the missing part would be "untangling" the sets of rows and columns in the resulting partition, in order to prove from this partition that one need only consider a set of possible input matrices that can be calculated from a small sample (as in the proof of Theorem 1.3).

Nonbinary matrices. It would also be interesting to prove the result for matrices that are not binary. It is enough to look at matrices with a fixed finite alphabet, because one does not need to distinguish between the different labels that do not appear in the finite set of forbidden matrices F.

Again "full conditional regularity" cannot be guaranteed, but this problem might be a little more accessible (though perhaps with a no longer polynomial dependence of the number of queries on ϵ). A possible course of attack could be to start by partitioning into blocks, each containing less than the full set of labels, and continue by recursively classifying each block as either "repartitionable" or "homogeneous" in a way somewhat reminiscent of what was done (more easily) in [11, 10] for poset properties.

Acknowledgment. We wish to thank Eyal Rozenberg for the discussion concerning an inaccuracy in an earlier version of the proof of Theorem 1.4. We also wish to thank two anonymous referees for their thoughtful comments.

REFERENCES

[1] N. Alon, *Testing subgraphs in large graphs*, Random Structures Algorithms, 21 (2002), pp. 359–370.
[2] N. Alon, S. Dar, M. Parnas, and D. Ron, *Testing of clustering*, SIAM J. Discrete Math., 16 (2003), pp. 393–417.
[3] N. Alon, E. Fischer, M. Krivelevich, and M. Szegedy, *Efficient testing of large graphs*, Combinatorica, 20 (2000), pp. 451–476.
[4] N. Alon and A. Shapira, *Testing subgraphs in directed graphs*, J. Comput. System Sci., 69 (2004), pp. 354–382.
[5] N. Alon and J. H. Spencer, *The Probabilistic Method*, 2nd ed., John Wiley, New York, 2000.
[6] M. Blum, M. Luby, and R. Rubinfeld, *Self-testing/correcting with applications to numerical problems*, J. Comput. System Sci., 47 (1993), pp. 549–595.
[7] R. Diestel, *Graph Theory*, 2nd ed., Springer-Verlag, New York, 2000.
[8] E. Fischer, *Testing graphs for colorability properties*, Random Structures Algorithms, 26 (2005), pp. 289–309.
[9] E. Fischer, *The art of uninformed decisions: A primer to property testing*, Bull. Eur. Assoc. Theor. Comput. Sci. EATCS, 75 (2001), pp. 97–126.
[10] E. Fischer and I. Newman, *Testing of matrix properties*, in Proceedings of the 33rd ACM Annual Symposium on Theory of Computing, ACM, New York, 2001, pp. 286–295.
[11] E. Fischer and I. Newman, *Testing of matrix-poset properties*, Combinatorica, to appear.
[12] W. T. Gowers, *Lower bounds of tower type for Szemerédi's Uniformity Lemma*, Geom. Funct. Anal., 7 (1997), pp. 322–337.
[13] T. Kövary, V.T. Sós, and P. Turán, *On a problem of K. Zarankiewicz*, Colloq. Math., 3 (1954), pp. 50–57.
[14] O. Goldreich, S. Goldwasser, and D. Ron, *Property testing and its connection to learning and approximation*, J. ACM, 45 (1998), pp. 653–750.
[15] O. Goldreich and L. Trevisan, *Three theorems regarding testing graph properties*, Random Structures Algorithms, 23 (2003), pp. 23–57.
[16] D. Ron, *Property testing (a tutorial)*, in Handbook of Randomized Computing, Vol. II, S. Rajasekaran, P. M. Pardalos, J. H. Reif, and J. D. P. Rolim, eds., Kluwer Academic Publishers, Boston, 2001, pp. 597–649.

[17] R. Rubinfeld and M. Sudan, *Robust characterizations of polynomials with applications to program testing*, SIAM J. Comput., 25 (1996), pp. 252–271.
[18] N. Sauer, *On the density of families of sets*, J. Combin. Theory Ser. A, 13 (1972), pp. 145–147.
[19] E. Szemerédi, *Regular partitions of graphs*, in Problems combinatoires et théorie des graphes, Colloq. Internat. CNRS 260, J. C. Bermond, J. C. Fournier, M. Las Vergnas, and D. Sotteau, eds., CNRS, Paris, 1978, pp. 399–401.
[20] S. Shelah, *A combinatorial problem: Stability and order for models and theories in infinitary languages*, Pacific J. Math., 41 (1972), pp. 247–261.
[21] K. Zarankiewicz, *Problem P* 101, Colloq. Math., 2 (1951), pp. 116–131.