Graph limits and parameter testing

Christian Borgs*, Jennifer Chayes*, László Lovász*§, Vera T. Sós†,
Balázs Szegedy‡, Katalin Vesztergombi§

October 2005

Abstract

We define a distance of two graphs that reflects the closeness of both local and global properties. We also define convergence of a sequence of graphs, and show that a graph sequence is convergent if and only if it is Cauchy in this distance. Every convergent graph sequence has a limit in the form of a symmetric measurable function in two variables. We use these notions of distance and graph limits to give a general theory for parameter testing. As examples, we provide short proofs of the testability of MaxCut and the recent result of Alon and Shapira about the testability of hereditary graph properties.

1 Introduction

Imagine that we have a huge graph G, so large that we cannot describe it completely in any way. All we can do is sample a bounded number of nodes of G and look at the subgraph induced by them. What can we learn about G?

There are two related, but slightly different ways of asking this question, *property testing* and *parameter testing*. Our main concern here will be parameter testing; we'll show one way to obtain results about property testing from this, by considering the "edit distance" from the property as a parameter.

Parameter testing is easier to state. We may want to determine some parameter of G, say what is the edge density? Or how large is the density of the maximum cut? Of course, we'll not be able to determine the exact value of this parameter; the best we can hope for is that if we take a sufficiently large sample, we can find the approximate value of the parameter with large probability.

To be precise, a graph parameter f is *testable*, if for every $\varepsilon > 0$ there is a positive integer k such that if G is a graph with at least k nodes and we select a set X of k independent uniform random nodes of G, then from the subgraph $G[X]$ induced by them we can compute an estimate $\widetilde{f}(G[X])$ of f such that

$$\mathsf{P}(|f(G) - \widetilde{f}(G[X])| > \varepsilon) < \varepsilon.$$

It is an easy observation that we can always use $\widetilde{f}(G[X]) = f(G[X])$.

As a basic example, consider the density of maximum cuts (i.e., the number $\mathsf{maxcut}(G) = \mathsf{MaxCut}(G)/|V(G)|^2$, where $\mathsf{MaxCut}(G)$ is the size of

*Microsoft Research, Redmond, WA 98052
†Rényi Institute, Budapest, Hungary
‡Institute for Advanced Study, Princeton, NJ
§Eötvös Loránd University, Budapest, Hungary

the maximum cut in G). It was already in early papers on property testing [12, 2] that this parameter is testable. It is relatively easy to see (using high concentration results like Azuma's inequality) that if S is a sufficiently large random subset of nodes of G, then $\mathsf{maxcut}(G[S]) \geq \mathsf{maxcut}(G) - \varepsilon$: a large cut in G, when restricted to S, gives a large cut in $G[S]$. It is much harder, and in fact quite surprising, that if most subgraphs $G[S]$ have a large cut, then so does G.

Instead of estimating a numerical parameter, we may want to determine some property of G: Is G 3-colorable? Is it connected? Does it have a triangle? The answer will of course have some uncertainty. A precise definition was given by Goldreich, Goldwasser and Ron [12], who also proved several fundamental results about this problem. (In the slightly different context of "additive approximation", closely related problems were studied by Arora, Karger and Karpinski [2].) The standard definition is as follows. A property $\mathcal{P}$ is said to be testable if, for every $\varepsilon > 0$, there is a positive integer k such that for all graphs G on at least k nodes, the following holds: If X is a k-node random subset of G, then, with high probability, the induced subgraph $G[X]$ has the property $\mathcal{P}$ whenever G has the property, and $G[X]$ does not have the property whenever G is ε-far from having the property. Here we say that a graph G on n nodes is ε-far from having the property $\mathcal{P}$ if all graphs G' which differ from G on at most εn^2 edges do not have the property $\mathcal{P}$.

Graph property testing has a large literature; see e.g. [10] for a survey. Many extensions deal with situations where we are allowed to sample more than a constant number of nodes of the large graph G; our concern will be the original setup, where the sample size is bounded. In this direction, a surprisingly general result was proved very recently by Alon and Shapira [5, 6]: they showed that every hereditary graph property is testable; more generally, the "edit distance" to a hereditary property is a testable parameter.

We develop a general theory of parameter testing, based on the notions of convergent graph sequences (introduced by Lovász and Szegedy [13]) and graph distances (introduced by Borgs, Chayes, Lovász, Sós and Vesztergombi [8]).

In Section 3, we introduce a notion of "distance" of two graphs that is suitable for the study of parameter testing. In Section 4, we define the notion of convergence of graphs and characterizations of the limit object. Informally, convergent sequences will be just those sequences that are Cauchy in this metric, and testable parameters will be just those graph parameters that are continuous in this metric.

As an illustration of our results, we show how the testability of the density of maximum cut (and extensions to multiway cuts) follows from our results, and sketch a reasonably simple way to derive the recent result of Alon and Shapira mentioned above.

2 Preliminaries: graph homomorphisms

A *graph parameter* is a function defined on simple graphs that is invariant under isomorphism.

For two graphs G and H, a homomorphism from G to H is an adjacency preserving map $V(G) \to V(H)$. When G is large and H is small, a homomorphism from G to H is often called an *H-coloring* of G. Let $\hom(G, H)$ denote the number of homomorphisms from G to H.

Let G be a graph on n nodes, and F be a graph on k

nodes. We define $t(F,G)$ to be the probability that a random map of $V(F)$ into $V(G)$ is a homomorphism, i.e.,

$$t(F,G) = \frac{\hom(F,G)}{n^k}.$$

We call $t(F,G)$ the *homomorphism density* (of F in G).

We extend these notions to the case when the target graph G is weighted. A *weighted graph* G is a graph with a weight $\alpha_i(G)$ associated with each node and a weight $\beta_{ij}(G)$ associated with each edge ij. In this paper we assume that $\alpha_i(G) > 0$ and $0 \le \beta_{ij}(G) \le 1$. An edge with weight 0 will play the same role as no edge between those nodes. The adjacency matrix of a weighted graph is obtained by replacing the 1's in the adjacency matrix by the weights of the edges. Let $\alpha_G = \sum_{i \in V(G)} \alpha_i(G)$ denote the total nodeweight of G. An *unweighted graph* is a weighted graph where all the node- and edgeweights are 1.

The notion of homomorphisms can be easily extended to the case when the target graph G is weighted:

$$\hom(F,G) = \sum_{\phi:\ V(F)\to V(G)} \prod_{u \in V(F)} \alpha_{\phi(u)}(G) \times \prod_{uv \in E(F)} \beta_{\phi(u),\phi(v)}(G)$$

where the sum runs over all maps from $V(F)$ to $V(G)$. The homomorphism density is now defined as

$$t(F,G) = \frac{\hom(F,G)}{\alpha_G{}^{|V(F)|}}.$$

3 Distances of graphs

3.1 Definition of graph distances

We want to define a measure of similarity of two large graphs. There is more than one very reasonable definition; but one of these, the "rectangle" or "cut" distance, will be particularly useful.

We have to proceed in two steps. It is easier to define the distance when the two graphs have the same set of nodes; but then we have to deal with finding the optimal overlay. It turns out that it is worth extending the arguments to "fractional overlays" (in the spirit of fractional solutions to integer programs, often used in combinatorial optimization).

3.1.1 Labeled graphs

Let G and G' be two unweighted labeled graphs with the same set of n of nodes. We want to define a notion of distance between them that reflects structural similarity. A natural distance measure is the "edit distance":

$$d_{\mathrm{edit}}(G,G') = \frac{|E(G) \triangle E(G')|}{n^2},$$

where for convenience we divided by n^2, so that the distance of two graphs is always between 0 and 1.

This distance notion is, however, often too restrictive: For example, the distance of two random graphs with the same density is of constant order (with large probability), even though two random graphs are structurally very similar. For our purposes, the following distance function will be more useful:

$$d_\square(G,G') = \frac{1}{n^2} \max_{S,T \subseteq \{1,\dots,n\}} |e_G(S,T) - e_{G'}(S,T)|,$$

where $e_G(S,T)$ is the number of edges joining S and T (with the edges in $S \cap T$ counted twice, so that $e_G(S,S)$ is twice the number of edges in G). Note that we are dividing by n^2 and not by $|S| \cdot |T|$, so the contribution of a pair S,T is at most $|S| \cdot |T|/n^2$. Thus small sets of size $o(n)$ play no role when measuring the distance.

If G is weighted, let

$$e_G(S,T) = \sum_{i\in S}\sum_{j\in T} \alpha_i(G)\beta_{ij}(G)\alpha_j(G).$$

For the case when G and G' have the same nodeweights, we define

$$d_\square(G,G') = \frac{1}{\alpha_G^2} \max_{S,T\subseteq\{1,\dots,n\}} |e_G(S,T) - e_{G'}(S,T)|.$$

There are several versions of the $d_\square$ distance that differ from each other only in absolute constant factors. For example, we could restrict the maximization in the definition to $S = T$. Another notable variation is obtained by considering an appropriate semidefinite relaxation of the maximization problem involved in the definition. The fact that this only changes by a factor of at most $2.78\dots$ is equivalent to Grothendieck's inequality in functional analysis. This shows that $d_\square(G,G')$ can be computed in polynomial time up to an absolute constant factor (see [4]).

3.1.2 Unlabeled graphs with the same number of nodes

Now assume that G and G' are unlabeled graphs on n nodes with nodeweights 1. We then define

$$\widehat{\delta}_\square(G,G') = \min_{\tilde{G},\tilde{G}'} d_\square(\tilde{G},\tilde{G}'),$$

where $\tilde{G}$ and $\tilde{G}'$ range over all labelings of G and G', respectively.

Consider any labeling that attains the minimum in this definition, and identify the nodes of G and G' with the same label. In this case, we say that G and G' are *optimally overlaid.*

3.1.3 Unlabeled graphs with different number of nodes

To define the distance of two unlabeled graphs with different number of nodes, say G with n nodes and G' with n' nodes, a first idea is to blow up each node of G into n' nodes, and each node of G' into n nodes, so that both graphs now will have nn' nodes. An improved version of this idea is to match up the nodes "fractionally".

It will be convenient to consider weighted graphs G and G' right away, and assume that the sum of nodeweights is 1 (just scale the nodeweights of each graph). Let X be a nonnegative $n \times n'$ matrix such that

$$\sum_{u=1}^{n'} X_{iu} = \alpha_i \quad \text{and} \quad \sum_{i=1}^{n} X_{iu} = \alpha'_u.$$

We think of X_{iu} as the portion of node i that is mapped onto node u. We call such a matrix X a *fractional overlay* (of G and G'). Let $\mathcal{X}(G,G')$ denote the set of all fractional overlays. If we view $\alpha(G)$ and $\alpha(G')$ as probability distributions on $V(G)$, then every $X \in \mathcal{X}(G,G')$ is a *coupling* of these distributions.

For each fractional overlay, we construct the following two weighted graphs, $G[X]$ and $G'[X^\top]$. The nodes of $G[X]$ are all pairs (i,u) where $1 \le i \le n$ and $1 \le u \le n'$. The weight of the node (i,u) is X_{iu}, the weight of the edge $((i,u),(j,v))$ is β_{ij}. The other graph $G'[X^\top]$ is defined similarly, except that the roles of i and u are interchanged.

Now the node sets of $G[X]$ and $G'[X^\top]$ are labeled by the same set of pairs (i,u), and they have the same nodeweights, so their $d_\square$ distance is well defined. Minimizing over all fractional overlays, we therefore obtain a well-defined distance between any two weighted unlabeled graphs G and G' with total nodeweight 1:

$$\delta_\square(G,G') = \min_{X\in\mathcal{X}(G,G')} d_\square(G[X],G'[X^\top]).$$

This distance can be expressed in terms of the origi-

nal graphs G and G' according to

$$\delta_\square(G,G') = \min_{X\in\mathcal{X}(G,G')} \max_{S,T\subseteq V\times V'} \Big|\sum_{\substack{(i,u)\in S\\(j,v)\in T}} X_{iu}X_{jv}(\beta_{ij}(G)-\beta_{uv}(G'))\Big|.$$

Of course, this definition also applies if G and G' are simple graphs with the same number of nodes, but it may give a different value than discussed above in section 3.1.2. It is trivial that $\delta_\square(G,G') \le \widehat{\delta}_\square(G,G')$. An inequality in the other direction also holds [8], so that

$$\delta_\square(G,G') \le \widehat{\delta}_\square(G,G') \le C\delta_\square(G,G')^{1/4} \qquad (1)$$

with an absolute constant C. We believe that a much stronger bound holds, probably

$$\delta_\square(G,G') = \Omega(\widehat{\delta}_\square(G,G')).$$

This notion of graph distance allows us to give a very simple formulation of Szemerédi's Regularity Lemma, at least in its weaker (but more effective) form given by Frieze and Kannan [11]. For a graph G and partition $\mathcal{P} = (V_1,\dots,V_k)$ of $V(G)$, define the weighted complete graph $G_\mathcal{P}$ for which $V(G_\mathcal{P}) = V(G)$ and every $u\in V_i$ and $v\in V_j$ is connected by an edge with weight $e_G(V_i,V_j)/(|V_i|\cdot|V_j|)$.

Lemma 3.1 (Frieze-Kannan) *For every graph G and every $\varepsilon>0$, there exists a partition of $V(G)$ with at most $2^{2/\varepsilon^2}$ classes such that $d_\square(G,G_\mathcal{P})\le\varepsilon$.*

(At the cost of increasing the bound on $|\mathcal{P}|$, we could assume that the classes of $\mathcal{P}$ have equal size.)

Corollary 3.2 *For every graph G and every $\varepsilon>0$, there exists a weighted graph H with at most $2^{2/\varepsilon^2}$ nodes such that $\delta_\square(G,H)\le\varepsilon$.*

The following inequality was proved in [13]; it is also closely related to the "Counting Lemma" in the theory of Szemerédi partitions:

Lemma 3.3 *For any any pair of weighted graphs G,G' and any simple graph F,*

$$|t(F,G)-t(F,G')| \le |E(F)|\cdot\delta_\square(G,G').$$

3.2 The distance of a sample

The following lemma is a consequence of a more general result in [1]; see also [8] for a simpler proof of a weaker version (which would still be enough for us).

Lemma 3.4 *Let G_1 and G_2 be two weighted graphs on the same set of nodes V with nodeweights 1, and let $\varepsilon = d_\square(G_1,G_2)$. Let $\delta>0$, $k\ge 10^{10}\log(2/\varepsilon)/(\varepsilon^4\delta^5)$, and let S be a random k-subset of V. Then with probability at least $1-\delta$,*

$$2^{-10}\varepsilon \le d_\square(G_1[S],G_2[S]) \le 10^7\varepsilon/\sqrt{\delta}.$$

Now we come to the main theorem about sampling.

Theorem 3.5 *Let G be a simple graph, let $\varepsilon,\delta>0$ and $k\ge 2^{\theta(1/(\delta\varepsilon^2))}$. Let S be a random k-subset of V. Then with probability at least $1-\delta$,*

$$\delta_\square(G,G[S]) \le \varepsilon.$$

Proof. We use the weak Regularity Lemma 3.1. Let $\mathcal{P}$ be a partition of $V(G)$ with $q = 2^{2/(\varepsilon\sqrt{\delta}10^{-8})^2}$ classes so that $d_\square(G,G_\mathcal{P}) \le \varepsilon\sqrt{\delta}10^{-8}$. Consider a random subset S of $V(G)$ with k elements. Then

$$\begin{aligned}\delta_\square(G,G[S]) \le& \delta_\square(G,G_\mathcal{P}) + \delta_\square(G_\mathcal{P},G_\mathcal{P}[S])\\ &+ \delta_\square(G_\mathcal{P}[S],G[S]). \qquad (2)\end{aligned}$$

On the right hand side, the first term is at most $\varepsilon\sqrt{\delta}10^{-8}\le\varepsilon 10^{-8}$. To estimate the second term, notice that both $G_\mathcal{P}$ and $G_\mathcal{P}[S]$ are obtained by blowing

up the same weighted graph H with q nodes (we may assume that $\alpha_H = 1$): in $G_{\mathcal{P}}$, we blow up each node $i \in V(H)$ into a set V_i with $|V_i| = \alpha_i(H)|V(G)|$; in $G_{\mathcal{P}}[S]$, we blow up each node into $|V_i \cap S|$ nodes. Since S is a random sample, with large probability all these sizes will be close to their expectation, which in turn is closed to $\alpha_i k$ provided k/n is small. Blowing up each node of $G_{\mathcal{P}}[S]$ into $|V(G)|/k$ copies, we get a graph H' that can also be obtained from H by blowing up each node i into approximately $\alpha_i(H)|V(G)|$ copies. This argument can be made precise by choosing appropriate fractional overlays, leading to the estimate

$$\delta_\square(G_{\mathcal{P}}, G_{\mathcal{P}}[S]) = \delta_\square(G_{\mathcal{P}}, H') \le \sum_{i=1}^{q} \Big| \alpha_i - \frac{1}{k}|S \cap V_i| \Big|.$$

Using standard second moment bounds, the right hand side can be made smaller than $\varepsilon/3$ with probability at least $1 - \delta/2$ provided $k = O(\epsilon n)$ and $q = O(\delta\varepsilon^2 k)$.

Finally, to see that the third term in (2) is at most $\varepsilon/3$, we invoke Lemma 3.4: this says that the samples $G[S]$ and $G_{\mathcal{P}}[S]$ are close in the $d_\square$ metric, completing the proof for $k = O(\epsilon n)$. The case $k \ge \varepsilon n$ is much easier and left to the reader, see also [8]. □

Corollary 3.6 *For every simple graph G and $\varepsilon > 0$ there is simple graph H with at most $2^{\Theta(1/\varepsilon^2)}$ nodes such that*

$$\delta_\square(G, H) < \varepsilon.$$

This corollary can be thought of as a strengthening of the (weak) Szemerédi lemma in two directions. First, it says that the approximating weighted graph can have 0-1 weights; second, that it can be obtained just by drawing a random sample. The theorem also explains why parameter testing works: if a parameter is "continuous" in the $\delta_\square$ metric, then with large probability, it will not change much if G is replaced by the random sample $G[S]$.

4 Convergent graph sequences and their limits

Let (G_n) be a sequence of unweighted simple graphs, and assume again that $|V(G_n)| \to \infty$. We say that this sequence is *convergent*, if the sequence $t(F, G_n)$ has a limit for every simple graph F. Note that it would be enough to assume this for connected graphs F.

As an example, let $\mathbf{G}(n,p)$ be a random graph on n nodes with edge-density p; then it is not hard to prove that the sequence $(\mathbf{G}(n,p), n = 1, 2, \dots)$ is convergent with probability 1. By definition, every quasirandom graph sequence with density p is also convergent (cf. [9]).

This notion of convergence corresponds to the notion of distance we have introduced [8]:

Theorem 4.1 *A graph sequence (G_n) is convergent if and only if it is Cauchy in the $\delta_\square$ metric.*

One can define a "limit object" for every convergent graph sequence; in fact, there are several quite different and useful descriptions of the limit object [13]. In this extended abstract, we are going to use the following. Let $\mathcal{W}_0$ denote the set of all bounded measurable functions $W : [0,1]^2 \to [0,1]$ such that $W(x,y) = W(y,x)$ for all $x, y \in [0,1]$. Let $W \in \mathcal{W}_0$, and let F be a simple graph with $V(F) = \{1, \dots, k\}$. We define

$$t(F, W) = \int_{[0,1]^k} \prod_{ij \in E(F)} W(x_i, x_j)\, dx.$$

It is easy to see that for every weighted graph H with $\alpha_H = 1$, the graph parameter $t(., H)$ is a special

case: we consider the adjacency matrix of H, and replace each entry (i,j) by a square of size $\alpha_i(H) \times \alpha_j(H)$ with the constant function β_{ij} on this square. Then $W_H \in \mathcal{W}_0$, and

$$t(F,H) = t(F,W_H)$$

for every simple graph F.

It was proved in [13] that

Proposition 4.2 *For every convergent graph sequence (G_n) there is a function $W \in \mathcal{W}_0$ such that*

$$\lim_{n\to\infty} t(F,G_n) = t(F,W)$$

for every finite simple graph F.

We call this function W the *limit* of the sequence (G_n). Every function in $\mathcal{W}_0$ arises as a limit. One proof of this fact is through the construction of certain random graphs that are of interest on their own right. Given a function $W \in \mathcal{W}_0$ and an integer $n > 0$, we can generate a *W-random graph* $\mathbf{G}(n,W)$ on nodes $\{1,\dots,n\}$ as follows: We generate n independent samples $X_1,\dots,X_n$ from the uniform distribution on $[0,1]$, and for all $i,j \in \{1,\dots,n\}$, we connect i and j by an edge ij with probability $W(X_i,X_j)$ (making an independent decision for every pair).

Theorem 4.3 *The graph sequence $\mathbf{G}(n,W)$ is convergent with probability 1, and its limit is the function W.*

The $d_\square$ distance can be extended to functions $U,W \in \mathcal{W}_0$:

$$d_\square(U,W) = \sup_{S,T\subseteq[0,1]} \Big| \int_{S\times T} U(x,y) - W(x,y)\,dx\,dy \Big|.$$

This metric has the following easy but basic property:

Lemma 4.4 *Let $W, W_1, W_2, \dots \in \mathcal{W}_0$, and suppose that $d_\square(W_n,W) \to 0$. Then for every integrable function $Z : [0,1]^2 \to \mathbb{R}$, we have*

$$\int_{[0,1]^2} Z\cdot W_n \longrightarrow \int_{[0,1]^2} Z\cdot W.$$

One can also extend the $\delta_\square$ distance:

$$\delta_\square(U,W) = \inf d_\square(U^\phi, W^\psi),$$

where ϕ,ψ are measure preserving transformations of $[0,1]$, and $U^\phi(x,y) = U(\phi(x),\phi(y))$. Using the (weak) Regularity Lemma, one can prove that *the metric space $(\mathcal{W}_0,\delta_\square)$ is compact.*

5 Characterizations of testable parameters

Theorem 5.1 [8] *A simple graph parameter f is testable if and only if any of the following equivalent conditions holds.*

(a) *For every convergent graph sequence (G_n), the limit of $f(G_n)$ exists as $n \to \infty$.*

(b) *For every $\varepsilon > 0$ there is an integer k_0 such that for every $k > k_0$ and every graph G on at least k nodes, a random set X of k nodes of G satisfies*

$$|f(G) - \mathsf{E}(f(G[X]))| < \varepsilon. \tag{3}$$

(c) *For every $\varepsilon > 0$ there is an $\varepsilon_0 > 0$ and a positive integer n_0 so that if G_1 and G_2 are two graphs with $|V(G_i)| \ge n_0$ and $\delta_\square(G_1,G_2) < \varepsilon_0$, then $|f(G_1) - f(G_2)| < \varepsilon$.*

(d) *There exists a functional $\widehat{f}(W)$ on $\mathcal{W}_0$ that is continuous in the $\delta_\square$ distance of functions, and extends f in the sense that $|\widehat{f}(W_G) - f(G)| \to 0$ if $|V(G)| \to \infty$.*

The definition of testability is very similar to condition (b): it says, in this language, that a random set X of k nodes of G as in (b) satisfies

$$|f(G) - f(G[X])| < \varepsilon$$

with large probability. The equivalence of (b) with testability says that for a parameter satisfying (b), the random variable $f(G[X])$ is concentrated around its expectation.

If we want to use (c) to prove that a certain parameter is testable, then the complicated definition of the $\delta_\square$ distance may cause a difficulty. So it is useful to show that (c) can be replaced by a weaker condition, which consists of three special cases of (c):

Supplement 5.2 *The following three conditions together are also equivalent to the testability of f:*

(c.1) *For every $\varepsilon > 0$ there is an $\varepsilon' > 0$ such that if G and G' are two simple graphs on the same node set and $d_\square(G, G') \le \varepsilon'$ then $|f(G) - f(G')| < \varepsilon$.*

(c.2) *For every simple graph G, $f(G(m))$ has a limit as $m \to \infty$, where $G(m)$ denotes the graph obtained from G by replacing each node by m twins.*

(c.3) *If G^+ is obtained from G by adding a single isolated node, then $f(G^+) - f(G) \to 0$ if $|V(G)| \to \infty$.*

Some of the implications between conditions (a)-(d) in the Theorem are easy, some others follow from the general theory sketched above. To illustrate the use of this theorem, let us consider the density $\mathsf{maxcut}(G)$ of the maximum cut. Of the conditions above, (a) and (b) are more-or-less reformulations of testability. Conditions (c.1–3), on the other hand, are easy to verify: (c.1) is immediate from the definition of the $d_\square$ distance, and (c.3) is trivial. To prove (c.2), notice that every cut in G yields a cut in $G(m)$ with the same density, so $\mathsf{maxcut}(G(m)) \ge \mathsf{maxcut}(G)$. Conversely, consider a maximum cut C in $G(m)$. For each node $i \in V(G)$, select a node i' from the set of m twins in $G(m)$ corresponding to i. The nodes i' induce a copy of G, in which the cut C induces a cut C'. It is easy to check that the expected density of C' is the density of C. Hence $\mathsf{maxcut}(G(m)) = \mathsf{maxcut}(G)$.

We could also use (d): the parameter maxcut extends to functions in a natural way, so that the conditions in (d) are easily verified.

6 Hereditary properties

Let $\mathcal{P}$ be a graph property (i.e., a family of simple graphs invariant under isomorphism). For every graph G, let

$$d_{\mathrm{edit}}(G, \mathcal{P}) = \min\left\{ \frac{|E(G') \triangle E(G)|}{|V(G)|^2} : \ V(G') = V(G),\ G' \in \mathcal{P} \right\}.$$

A graph property is *testable*, if for every $\varepsilon > 0$ there is a positive integer k such that if if $G \in \mathcal{P}$, then all but an ε fraction of k-node induced subgraphs of G have property $\mathcal{P}$; and if $d_{\mathrm{edit}}(G, \mathcal{P}) > \varepsilon$, then all but an ε fraction of k-node induced subgraphs of G do not have property $\mathcal{P}$.

Let $\mathcal{P}$ be a hereditary graph property (i.e., a property closed under removal of vertices). The following surprisingly general result was proved by Alon and Shapira [6].

Theorem 6.1 (Alon and Shapira) *Every hereditary property is testable.*

We sketch a proof of this theorem, based on our framework (see [14] for details). One should point out that while the proof of Alon and Shapira is constructive (even though, through the use of a strengthened version of the Szemerédi Regularity Lemma, it gives

a "tower of towers" dependence of the test size on the error). Our proof is not constructive.

We start with extending the definition of the property $\mathcal{P}$ to $\mathcal{W}_0$, by saying that $W \in \mathcal{W}_0$ has the property if and only if the W-random graph $\mathbf{G}(n, W)$ has the property for every positive integer n with probability 1. Let $\overline{\mathcal{P}}$ denote the set of functions in $\mathcal{W}_0$ with property $\mathcal{P}$.

It is not hard to see that if W is the limit of a convergent graph sequence (G_n), and all graphs G_n have the property, then so does their limit. Together with Theorem 4.3, this implies that $\overline{\mathcal{P}}$ is the closure of $\mathcal{P}$ in the metric $\delta_\square$.

Suppose that $\mathcal{P}$ is not testable. Since $\mathcal{P}$ is hereditary, testing a random k-node subgraph of a graph G for $\mathcal{P}$ cannot give a false negative, so we must have a false positive: there is a positive $\varepsilon > 0$ and a sequence (G_n) of graphs such that an ε fraction of n-node subgraphs of G_n is in $\mathcal{P}$, but $d_{\mathrm{edit}}(G_n, \mathcal{P}) > \varepsilon$. It is not hard to see that by considering smaller samples, we may strengthen the condition so that at least a fraction of $1 - o(1)$ of n-node subgraphs of G_n is in $\mathcal{P}$.

We may assume that the sequence (G_n) is convergent, and let W be its limit. By [13], we may further assume, by taking subsequences if necessary, that W_{G_n} converges to W in the $d_\square$ metric. Finally, is easy to see that the assumption that "most" n-node subgraphs of G_n are in $\mathcal{P}$ implies that $W \in \overline{\mathcal{P}}$. Consider the functions $U_n \in \mathcal{W}_0$ defined by

$$U_n(x,y) = \begin{cases} 0, & \text{if } W(x,y) = 0, \\ 1, & \text{if } W(x,y) = 1, \\ W_{G_n}(x,y), & \text{otherwise.} \end{cases}$$

Then

$$\begin{aligned} \|W_{G_n} - U_n\|_1 &= \int_{[0,1]^2} |W_{G_n} - U_n| \\ &= \int_{W=0} W_{G_n} + \int_{W=1} (1 - W_{G_n}). \end{aligned}$$

By Lemma 4.4, this implies that

$$\|W_{G_n} - U_n\|_1 \to 0. \tag{4}$$

For each $u \in V(G_n)$, let J_u be the subset of $[0,1]$ that represents u in W_{G_n}, and let X_u be a uniformly chosen point from J_u. Let $\mathbf{H}_n$ be the (random) graph on $V(G_n)$ obtained as follows: we connect u and v if and only if either $U_n(X_u, X_v) = 1$ or $U_n(X_u, X_v) > 0$ and u, v are connected in G_n.

Since U_n is obtained from W by changing values strictly between 0 and 1, we have $U_n \in \overline{\mathcal{P}}$, and hence it easily follows that $\mathbf{H}_n \in \mathcal{P}$ with probability 1. Furthermore, $|E(G_n) \triangle E(\mathbf{H}_n)|$ is the number of edges in G_n with $U_n(X_u, X_v) = 0$ plus the number of non-edges in G_n with $U_n(X_u, X_v) = 1$. Let L_{uv} be the set of points in $J_u \times J_v$ where $U_n = 0$ if $uv \in E(G_n)$, the set of points in $J_u \times J_v$ where $U_n = 1$ if $uv \notin E(G_n)$. The probability that a pair u, v is counted in $|E(G_n) \triangle E(\mathbf{H}_n)|$ is $\lambda(L_{uv})/\lambda(J_u \times J_v) = |V(G_n)|^2 \lambda(L_{uv})$, where $\lambda(\cdot)$ denotes Lebesgue measure. Summing over all u, v, and dividing by $|V(G_n)|^2$, we get that

$$\mathsf{E}(d_{\mathrm{edit}}(G_n, \mathbf{H}_n)) = \sum_{u,v} \lambda(L_{uv}) \le \|W_{G_n} - U_n\|_1.$$

By (4), this tends to 0 as $n \to \infty$, a contradiction[1].

References

[1] N. Alon, W. Fernandez de la Vega, R. Kannan and M. Karpinski: Random sampling and ap-

[1] We are grateful to Vojta Rödl, Christian Avart and Mathias Schacht for pointing out an error in a previous version of this proof in [14].

proximation of MAX-CSPs, *J. Comput. System Sci.* **67** (2003), 212–243.

[2] S. Arora, D. Karger and M. Karpinski: Polynomial time approximation schemes for dense instnces of NP-hard problems, *Proc. 27-th ACM STOC* (1995), 284–293.

[3] N. Alon and M. Krivelevich: Testing k-colorability, *SIAM J. Discrete Math. 15* (2002), 211–227.

[4] N. Alon, A. Naor: Approximating the Cut-Norm via Grothendieck's Inequality (preprint)

[5] N. Alon and A. Shapira, Every monotone graph property is testable, Proc. of the 37 ACM STOC, Baltimore, ACM Press (2005), to appear.

[6] N. Alon and A. Shapira: A Characterization of the (natural) Graph Properties Testable with One-Sided Error, Proc. FOCS 2005, to appear.

[7] C. Borgs, J. Chayes, L. Lovász, V.T. Sós, K. Vesztergombi: Counting graph homomorphisms (preprint)

[8] C. Borgs, J. Chayes, L. Lovász, V.T. Sós, K. Vesztergombi: Convergent sequences of dense graphs (manuscript)

[9] F. Chung, R.L. Graham and R.M. Wilson: Quasi-random graphs, *Combinatorica* **9** (1989), 345–362.

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

[11] A. Frieze and R. Kannan: Quick approximation to matrices and applications, *Combinatorica* **19**, 175–220.

[12] O. Goldreich, S. Goldwasser and D. Ron: Property testing and its connection to learning and approximation, *J. ACM* **45** (1998), 653–750.

[13] L. Lovász and B. Szegedy: Limits of dense graph sequences (MSR Tech Report # MSR-TR-2004-79) `ftp://ftp.research.microsoft.com/pub/tr/TR-2004-79.pdf`

[14] L. Lovász and B. Szegedy: Graph limits and testing hereditary graph properties, Microsoft Research Technical Report MSR-TR-2005-110, `ftp://ftp.research.microsoft.com/pub/tr/TR-2005-110.pdf`

[15] R. Rubinfeld and M. Sudan: Robust characterization of polynomials with applications to program testing, *SIAM J. on Computing* **25** (1996), 252–271.