Graph limits and testing hereditary graph properties

László Lovász and Balázs Szegedy
Microsoft Research

July 2005, revised August 2005
MSR Technical Report No. TR-2005-110

Abstract

We show that an important recent result of Alon and Shapira on testing hereditary graph properties can be derived from the existence of a limit object for convergent graph sequences.

1 Introduction

A graph property $\mathcal{P}$ is *hereditary*, if for every graph G with property $\mathcal{P}$, every induced subgraph also has property $\mathcal{P}$.

A graph property $\mathcal{P}$ is *testable*, if there is another graph property $\mathcal{P}'$ such that for every $\varepsilon > 0$ there is a $k = k(\varepsilon)$ such that

(a) if changing at most εn^2 edges in any way the obtained graph has property $\mathcal{P}$, then at least a $1 - \varepsilon$ fraction of its k-node induced subgraphs have property $\mathcal{P}'$, and

(b) if changing at most εn^2 edges in any way the obtained graph does not have property $\mathcal{P}$, then at least a $1 - \varepsilon$ fraction of its k-node induced subgraphs do not have property $\mathcal{P}'$.

We can view such a property $\mathcal{P}'$ as a local test for $\mathcal{P}$. We draw a random sample of size k and see if it has property $\mathcal{P}'$; if it does, we guess that G has property $\mathcal{P}$, else we guess that it does not. If a graph has property $\mathcal{P}$ in a robust way so that even changing εn^2 edges it remains valid, then the test will give the correct answer with large probability. The outcome is similar if the graph fails to have the property in a robust way. We have a grey area inbetween, where we can guess arbitrarily. In other words, whatever we guess, we should be able to change at most εn^2 edges to make out guess right.

Alon and Shapira [1] prove the following very general result:

Theorem 1 *Every hereditary graph property is testable.*

They in fact prove more. Let f be a graph parameter (i.e., a real valued function defined on simple graphs, invariant under isomorphism). Assume that the parameter is normalized so that $0 \le f \le 1$. We say that this parameter is *testable* if for every $\varepsilon > 0$ there is a positive integer

$k(\varepsilon)$ such that if G is a graph with at least $k(\varepsilon)$ nodes and we select a set X of $k(\varepsilon)$ independent uniform random nodes of G, then for the subgraph G' induced by them

$$\mathsf{P}(|f(G) - f(G')| > \varepsilon) < \varepsilon.$$

For two graphs G and G' on the same set of nodes, we denote by $D_1(G, G') = |E(G) \triangle E(G')|$ their "edit distance", and we set $d_1(G, G') = D_1(G, G')/|V(G)|^2$. For every graph G and graph property $\mathcal{P}$, let $d_1(G, \mathcal{P}) = \min_{G' \in \mathcal{P}} d_1(G, G')$. We call $d_1(., \mathcal{P})$ the *distance* from property $\mathcal{P}$. (We assume that for every $n > 0$, there is a graph with n nodes with property $\mathcal{P}$; this is clearly the case for hereditary properties that hold for infinitely many graphs.)

Theorem 2 (Alon–Shapira [1]) *The distance from a hereditary graph property is testable.*

Alon and Shapira prove and use a strengthened form of the Szemerédi Regularity Lemma. The goal of this note is to describe an alternate proof that uses instead the existence of a limit of graph sequences.

2 Preliminaries

2.1 Subgraph densities

Let F and G be simple graphs. Let $t(F, G)$ denote the probability that a random map $V(F) \to V(G)$ preserves adjacency; let $t_{\mathrm{inj}}(F, G)$ denote the probability that a random injective map $V(F) \to V(G)$ preserves adjacency; and let $t_{\mathrm{ind}}(F, G)$ denote the probability that a random injective map $V(F) \to V(G)$ preserves adjacency as well as nonadjacency. It is easy to see that t, t_{inj} and t_{ind} are related [4]:

$$|t_{\mathrm{inj}}(F, G) - t(F, G)| \le \binom{|V(F)|}{2} \frac{1}{|V(G)|},$$

and

$$t_{\mathrm{inj}}(F, G) = \sum_{\substack{F' \supseteq F \\ V(F') = V(F)}} t_{\mathrm{ind}}(F, G).$$

From here we get by inclusions-exclusion that

$$t_{\mathrm{ind}}(F, G) = \sum_{\substack{F' \supseteq F \\ V(F') = V(F)}} (1-)^{|E(F') \setminus E(F)|} t_{\mathrm{inj}}(F, G).$$

2.2 2-variable functions

Let $\mathcal{W}$ denote the space of bounded symmetric measurable functions $W : [0, 1]^2 \to \mathbb{R}$, and let $\mathcal{W}_0$ be the set of all functions in $\mathcal{W}$ with range on $[0, 1]$. For every $W \in \mathcal{W}$, we define

$$\|W\|_\square = \sup_{S,T} \Big| \int_{S \times T} W(x, y)\, dx\, dy \Big|,$$

where S and T range over all measurable subsets of $[0, 1]$. It is clear that $\|.\|_\square$ is a norm. For $W, W_1, W_2, \cdots \in \mathcal{W}_0$, we write $W_n \xrightarrow{\square} W$ if $\|W - W_n\| \to 0$.

We can think of $W \in \mathcal{W}_0$ as a graph on node set $[0,1]$, where $W(x,y)$ is the edge density between an infinitesimal neighborhood of x and an infinitesimal neighborhood of y. The subgraph densities t and $t_{\rm ind}$ can be extended to subgraph densities in functions in $W \in \mathcal{W}_0$ as follows: if $F = (V,E)$ is a simple graph, then

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

and

$$t_{\rm ind}(F,W) = \int_{[0,1]^V} \prod_{ij \in E} W(x_i,x_j) \prod_{\substack{ij \notin E \\ i \neq j}} (1 - W(x_i,x_j))\,dx.$$

(There would be no difference between t and $t_{\rm inj}$ in this setting.) Expanding the product in the last formula, we get the identity

$$t_{\rm ind}(F,W) = \sum_{\substack{F' \supseteq F \\ V(F')=V(F)}} (1-)^{|E(F')\setminus E(F)|} t(F,W). \tag{1}$$

This "subgraph density" parameter is continuous in the $\|.\|_\square$ norm [2]; more precisely, for every simple graph F and $U,W \in \mathcal{W}_0$ we have

$$|t(F,U) - t(F,W)| \le |E(F)|\|U - W\|_\square. \tag{2}$$

With every graph G on n nodes we can associate a stepfunction $W_G \in \mathcal{W}_0$: we consider the adjacency matrix (a_{ij}) of G, and replace each entry a_{ij} by a square of size $(1/n) \times (1/n)$ with the constant function a_{ij} on this square. Then it is easy to check that

$$t(F,G) = t(F,W_G) \tag{3}$$

and

$$|t_{\rm ind}(F,G) - t_{\rm ind}(F,W_G)| \le \binom{|V(F)|}{2} \frac{1}{|V(G)|}. \tag{4}$$

Let us conclude this section with a remark about the proof from the analytic point of view. Property testing involves looking at small graphs, which by (2), is connected to the $\|.\|_\square$ norm. On the other hand, the definition of $d_1(G,\mathcal{P})$ is in terms of the "edit distance", which corresponds to the L_1 norm. To go from one norm to the other is the main difficulty.

2.3 Convergent graph sequences

Let (G_n) be a sequence of unweighted simple graphs. We say that this sequence is *convergent*, if $|V(G_n)| \to \infty$, and the sequence $(t(F,G_n))$ has a limit for every simple graph F as $n \to \infty$. It follows by an easy diagonalization argument that every infinite sequence of graphs for which $|V(G_n)| \to \infty$ has a convergent subsequence.

It was proved in [4] that every convergent graph sequence has a limit in the form of a function $W \in \mathcal{W}_0$ such that $t(F,G_n) \to t(F,W)$ for every simple graph F. By (1) it follows that in this case $t_{\rm ind}(F,G_n) \to t_{\rm ind}(F,W)$.

Another characterization of the limit function is the following [2]: For an an appropriate labeling of the nodes of the graphs,

$$\|W - W_{G_n}\|_\square \to 0.$$

See [5, 2] for more on this norm and its connections with Szemerédi partitions and quasirandom graphs.

Let f be a graph parameter. We say that f is *continuous at infinity*, if for every convergent graph sequence (G_n), $f(G_n)$ tends to a limit. It is not hard to prove [2] that

Proposition 3 *A graph parameter is testable if and only if it is continuous at infinity.*

3 Hereditary properties and the $\|.\|_\square$ norm

Lemma 4 *Let $W, W_1, W_2 \cdots \in \mathcal{W}_0$, and suppose that $W_n \xrightarrow{\square} W$. Then for every integrable function $Z : [0,1]^2 \to \mathbb{R}$, we have*

$$\int_{[0,1]^2} Z(x,y)W_n(x,y)\,dx\,dy \longrightarrow \int_{[0,1]^2} Z(x,y)W(x,y)\,dx\,dy.$$

In particular,

$$\int_S W_n \longrightarrow \int_S W$$

for every measurable set $S \subseteq [0,1]^2$.

Proof. If U is the indicator function of a rectangle, this follows from the definition of the $\|.\|_\square$ norm. Hence the conclusion follows for stepfunctions, since they are linear combinations of a finite number of indicator functions of rectangles. Then it follows for for all integrable functions, since they are approximable in $L_1([0,1]^2)$ by stepfunctions. □

Let $\mathcal{P}$ be a hereditary graph property. Let $\mathcal{H}$ denote the set of all functions $U \in \mathcal{W}_0$ such that for all $n \in \mathbb{Z}_+$ and almost all $x_1, \dots, x_n \in [0,1]$, all graphs G on $[n]$ such that

$$\begin{cases} ij \in E(G) & \text{if } i \neq j \text{ and } U(x,y) = 1, \\ ij \notin E(G) & \text{if } i \neq j \text{ and } U(x,y) = 0 \end{cases}$$

have property $\mathcal{P}$. In other words, whenever a graph G does not have property $\mathcal{P}$, then $t_{\mathrm{ind}}(G,U) = 0$. Note that if $U \in \mathcal{H}$, then changing its value in points (x,y) with $0 < U(x,y) < 1$ results in another function in $\mathcal{H}$.

Lemma 5 *The set $\mathcal{H}$ is closed in $\mathcal{W}$ with respect to the $\|.\|_\square$ norm.*

Proof. Suppose that $W_n \xrightarrow{\square} W$. Then for every F, $t(F,W_n) \to t(F,W)$, hence $t_{\mathrm{ind}}(F,W_n) \to t_{\mathrm{ind}}(F,W)$, hence if $t_{\mathrm{ind}}(F,W_n) = 0$ for every $F \notin \mathcal{P}$, then also $t_{\mathrm{ind}}(F,W) = 0$ for these graphs F. □

Let

$$d_1(W,\mathcal{H}) = \min_{U \in \mathcal{H}} \|W - U\|_\square$$

denote the L_1-distance of W from $\mathcal{H}$. The following lemma is the main step in the proof.

Lemma 6 *Let $\mathcal{P}$ be a hereditary graph property, and let $\mathcal{H} \subseteq \mathcal{W}$ denote the set of functions with this property. Then $d_1(W, \mathcal{H})$ is a continuous function of W in the $\|.\|_\square$ norm.*

It is trivial that $d_1(W, \mathcal{H})$ is a continuous function of W in the $\|.\|_1$ norm.

Proof. Suppose that $W_n \xrightarrow{\square} W$. First we prove that

$$\limsup_{n\to\infty} d_1(W_n, \mathcal{H}) \le d_1(W, \mathcal{H}). \tag{5}$$

Let $U \in \mathcal{H}$ such that $d_1(W, \mathcal{H}) = \|W - U\|_1$. Let

$$S_0 = \{(x, y) \in [0, 1]^2 : \ U(x, y) = 0\}, \qquad S_1 = \{(x, y) \in [0, 1]^2 : \ U(x, y) = 1\}.$$

For almost every point $(x, y) \in [0, 1]^2 \setminus S_0 \setminus S_1$ we must have $U(x, y) = W(x, y)$ (else, we could change the value and decrease $\|W - U\|_1$).

Consider the functions $U_n \in \mathcal{W}_0$ defined by

$$U_n(x, y) = \begin{cases} 0, & \text{if } (x, y) \in S_0, \\ 1, & \text{if } (x, y) \in S_1, \\ W_n(x, y), & \text{otherwise.} \end{cases}$$

Since U_n is obtained from U by changing values strictly between 0 and 1, we have $U_n \in \mathcal{H}$. Hence, using Lemma 4, we have

$$\begin{aligned} d_1(W_n, \mathcal{H}) \le \|W_n - U_n\|_1 &= \int_{[0,1]^2} |W_n - U_n| = \int_{S_0} W_n + \int_{S_1} (1 - W_n) \\ &\to \int_{S_0} W + \int_{S_1} (1 - W) = \int_{[0,1]^2} |W - U| = \|W - U\|_1 = d_1(W, \mathcal{H}). \end{aligned}$$

This implies (5).

To prove the converse, let $X_n \in \mathcal{H}$ be chosen so that $\|W_n - X_n\|_1 = d_1(W_n, \mathcal{H})$. We may assume that X_n is convergent in the $\|.\|_\square$ norm, then by Lemma 5, its limit X belongs to $\mathcal{H}$. Let

$$S_0 = \{(x, y) \in [0, 1]^2 : \ X(x, y) = 0\}, \qquad S_1 = \{(x, y) \in [0, 1]^2 : \ X(x, y) = 1\},$$

and let $Y \in \mathcal{W}_0$ be defined by

$$Y(x, y) = \begin{cases} 0, & \text{if } (x, y) \in S_0, \\ 1, & \text{if } (x, y) \in S_1, \\ W(X, Y), & \text{otherwise.} \end{cases}$$

Then clearly $Y \in \mathcal{H}$, and so by Lemma 4 again,

$$\begin{aligned} d_1(W_n, \mathcal{H}) = \|W_n - X_n\|_1 &= \int_{[0,1]^2} |W_n - X_n| \ge \int_{S_0} W_n + \int_{S_1} (1 - W_n) \\ &\to \int_{S_0} W + \int_{S_1} (1 - W) = \int_{[0,1]^2} |W - Y| = \|W - Y\|_1 \ge d_1(W, \mathcal{H}). \end{aligned}$$

This proves that

$$\liminf_{n\to\infty} d_1(W_n, \mathcal{H}) \ge d_1(W, \mathcal{H}),$$

which competes the proof of the Lemma. □

Lemma 7 *If (G_n) is a convergent sequence of graphs with property $\mathcal{P}$, and $G_n \xrightarrow{\square} W$, then $W \in \mathcal{H}$.*

Proof. Suppose $W \notin \mathcal{H}$, then there is a graph $F \notin \mathcal{P}$ such that $t_{\rm ind}(F, W) > 0$. But we know that

$$t_{\rm ind}(F, G_n) \to t_{\rm ind}(F, W),$$

and hence $t_{\rm ind}(F, G_n) > 0$ if n is large enough. Since $\mathcal{P}$ is hereditary, this implies that G_n cannot have property $\mathcal{P}$, a contradiction. □

Lemma 8 *For every graph G,*

$$d_1(G, \mathcal{P}) \le d_1(W_G, \mathcal{H}).$$

Proof. Let $U \in \mathcal{H}$ be such that $\|W_G - U\|_1 = d_1(W_G, \mathcal{H})$. As before, we may assume that U is a $\{0,1\}$-valued function. Let $V(G) = [n]$, and let X_i be a uniform random element of the interval $L_i = [\frac{i-1}{n}, \frac{i}{n}]$. Let G_X denote the graph on $[n]$ in which i and j are adjacent if and only if $U(X_i, X_j) = 1$. Then with probability 1, G_X has property $\mathcal{P}$. We have

$$\begin{aligned}\mathsf{E}(d_1(G, G_X)) &= \frac{1}{n^2}\mathsf{E}(|E(G) \triangle E(G_X)|) = \frac{1}{n^2}\sum_{i,j=1}^{n} \Pr(W_G(X_i, X_j) \neq U(X_i, X_j)) \\ &= \sum_{i,j=1}^{n} \int_{L_i \times L_j} |W_G(X_i, X_j) - U(X_i, X_j)| = \|W_G - U\|_1 = d_1(W_G, \mathcal{H}).\end{aligned}$$

Hence there is a choice of X for which $G_X \in \mathcal{P}$ and $d_1(G, G_X) \le d_1(W_G, \mathcal{H})$. This proves the lemma. □

4 Proof of Theorem 2

Proof. Let (G_n) be a convergent graph sequence, and let $W \in \mathcal{W}_0$ be its limit. We want to show that $d_1(G_n, \mathcal{P})$ is convergent. We in fact show that

$$d_1(G_n, \mathcal{P}) \to d_1(W, \mathcal{H}).$$

We know that $W_{G_n} \xrightarrow{\square} W$, so by Lemma 6, $d_1(W_{G_n}, \mathcal{H}) \to d_1(W, \mathcal{H})$. So by Lemma 8, we have

$$d_1(G_n, \mathcal{P}) \le d_1(W_{G_n}, \mathcal{H}) = d_1(W, \mathcal{H}) + o(1),$$

and so

$$\limsup_{n\to\infty} d_1(G_n, \mathcal{P}) \le d_1(W, \mathcal{H}).$$

The converse inequality

$$\liminf_{n\to\infty} d_1(G_n, \mathcal{P}) \ge d_1(W, \mathcal{H})$$

is somewhat more difficult. Suppose that the liminf is smaller, then we may assume (by keeping only a subsequence) that

$$d_1(G_n, \mathcal{P}) \to c < d_1(W, \mathcal{H}).$$

For each G_n, let H_n be a graph on the same nodeset such that H_n has property $\mathcal{P}$ and $d_1(G_n, H_n) = d_1(G_n, \mathcal{P})$. We may assume that H_n is convergent; let $U \in \mathcal{W}_0$ be its limit. By Lemma 7, we have $U \in \mathcal{H}$. Clearly

$$d_1(W_{G_n}, W_{H_n}) = d_1(G_n, H_n) = d_1(G_n, \mathcal{P}).$$

Hence

$$d_1(W_{G_n}, \mathcal{H}) \le d_1(W_{G_n}, W_{H_n}) + d_1(W_{H_n}, \mathcal{H}) = d_1(G_n, \mathcal{P}) + d_1(W_{H_n}, \mathcal{H}).$$

In this inequality, $d_1(W_{G_n}, \mathcal{H}) \to d_1(W, \mathcal{H})$ and $d_1(W_{H_n}, \mathcal{H}) \to d_1(U, \mathcal{H}) = 0$ by Lemma 6. Hence $d_1(W, \mathcal{H}) \le c$, a contradiction. □

References

[1] N. Alon and A. Shapira: A Characterization of the (natural) Graph Properties Testable with One-Sided Error, Proc. FOCS 2005, to appear.
`http://www.math.tau.ac.il/~asafico/heredit.pdf`

[2] C. Borgs, J. Chayes, L. Lovász, V.T. Sós, K. Vesztergombi: Convergent sequences of dense graphs, in preparation

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

[4] L. Lovász and B. Szegedy: Limits of dense graph sequences (MSR Tech Report No. TR-2004-79).
`ftp://ftp.research.microsoft.com/pub/tr/TR-2004-79.pdf`

[5] L. Lovász and B. Szegedy: Szemerédi's Lemma for the Analyst (MSR Tech Report No. TR-2005-09).
`ftp://ftp.research.microsoft.com/pub/tr/TR-2005-09.pdf`