

Non-Deterministic Graph Property Testing

LÁSZLÓ LOVÁSZ and KATALIN VESZTERGOMBI

Institute of Mathematics, Eötvös Loránd University, Budapest, Hungary
(e-mail: lovasz@cs.elte.hu)

Received 29 September 2012; revised 27 April 2013; first published online 3 July 2013

A property of finite graphs is called non-deterministically testable if it has a ‘certificate’ such that once the certificate is specified, its correctness can be verified by random local testing. In this paper we study certificates that consist of one or more unary and/or binary relations on the nodes, in the case of dense graphs. Using the theory of graph limits, we prove that non-deterministically testable properties are also deterministically testable.

2010 *Mathematics subject classification*: Primary 05C85
Secondary 68R10, 05C82

1. Introduction

Let \mathcal{P} be a property of finite simple graphs (*i.e.*, a class of finite simple graphs closed under isomorphism). We say that \mathcal{P} is *testable* if, for every $\varepsilon > 0$, there is an $r_\varepsilon \geq 1$ and another property \mathcal{T} (called a *test property*) satisfying the following conditions:

- if G is a graph with property \mathcal{P} , then for all $r_\varepsilon \leq r \leq |V(G)|$, a random induced subgraph on r nodes (chosen uniformly among all such induced subgraphs) has property \mathcal{T} with probability at least $2/3$, and
- if G is a graph whose edit distance from \mathcal{P} is at least $\varepsilon|V(G)|^2$, then for all $r_\varepsilon \leq r \leq |V(G)|$, a random induced subgraph on r nodes has property \mathcal{T} with probability at most $1/3$.

(The edit distance of a graph from \mathcal{P} is the minimum number of adjacencies one has to change to obtain a graph in \mathcal{P} .)

This notion of testability is often called *oblivious testing*, which refers to the fact that no information about the size of G is assumed. It is easy to see that if \mathcal{P} is a testable property such that arbitrarily large graphs can have the property, then there must exist a graph with property \mathcal{P} with any sufficiently large number of nodes. The definition extends trivially to graphs whose edges are oriented, and whose nodes and/or edges are coloured with a fixed finite number k of colours.

Testability of graph properties was introduced by Rubinfeld and Sudan [19] and Goldreich, Goldwasser and Ron [13]. There are many graph properties that are known to be testable [13, 2]; see, e.g., [9] for a survey and [12] for a collection of more recent surveys. One surprisingly general sufficient condition was found by Alon and Shapira [1]: every hereditary graph property is testable. (A graph property is called *hereditary* if it is inherited by induced subgraphs.)

Various characterizations of testable properties are known [3, 17], but they are not simple to state. The goal of this paper is to prove a characterization which is useful as a sufficient condition in a number of cases.

A *complete digraph* is a directed graph in which any two distinct nodes are connected by two oppositely directed edges. We call a complete digraph whose edges are coloured with $1, \dots, k$ a *k-coloured digraph* for short. A *property* of *k-coloured digraphs* is any class of *k-coloured digraphs* closed under isomorphism (where we allow permutation of vertices but not permutation of colours). A property of simple graphs can be viewed as the special case $k = 2$ (where the two colours indicate whether or not the edge is present, independent of the orientation). The *edit distance* of two *k-coloured digraphs* on the same node set is defined as the number of edges that have different colours in the two graphs. *Testability* of a property \mathcal{P} of *k-coloured digraphs* can be defined in the natural way. For every error-bound ε , we need a test property \mathcal{T} of *k-coloured digraphs*. We take a random induced subgraph F of the given *k-coloured digraph* G , in which all the edge-colours are inherited from G , and make a prediction of the property \mathcal{P} of G based on the test property \mathcal{T} of F . The conditions on the error above can be extended to this case almost verbatim.

Given a *k-coloured digraph* L and a positive integer $m \leq k$, we can get an ordinary graph from L by keeping only the edges with colours $1, \dots, m$, and then forgetting the colouring and the orientation, and merging duplicate edges. We call this graph L'_m the *m-shadow* of L . If \mathcal{Q} is a property of coloured directed graphs, then we define $\mathcal{Q}'_m = \{L'_m : L \in \mathcal{Q}\}$.

A graph property \mathcal{P} is *non-deterministically testable* if there exist two integers $k \geq m \geq 1$ and a property \mathcal{Q} of *k-coloured digraphs* such that \mathcal{Q} is testable and $\mathcal{Q}'_m = \mathcal{P}$. In other words, G has property \mathcal{P} if and only if we can orient its edges (in one or both directions), colour them with m colours, add all the missing oriented edges, and colour them with a further $k - m$ colours, so that the resulting *k-coloured digraph* has property \mathcal{Q} . We call such an orientation and colouring a *certificate* for \mathcal{P} .

Clearly every testable property is non-deterministically testable (choosing $k = 2, m = 1$). Our main result asserts the converse.

Theorem 1.1. *A graph property is non-deterministically testable if and only if it is testable.*

One could say that this theorem shows that ‘ $P = NP$ ’ for property testing in dense graphs. The proof uses the theory of graph limits as developed in [6, 15], and its connection with property testing [17].

We could consider more general certificates in the definition of non-deterministic testability. Most of these would not change this notion. We could consider shadows of (not necessarily simple) directed graphs whose nodes and edges are coloured. We can

reduce this to the case when every pair of nodes is connected by two oppositely directed edges: if some edges are missing, we can add them coloured with an additional colour; if some edges have larger (but bounded) multiplicity, we can use colours to indicate this. Furthermore, we may get rid of the node colours by colouring every edge e with the pair (a, b) , where a is the original colour of the edge and b is the colour of its tail. Edge-colourings obtained in this way have a special property (each edge starting at a node v should carry the same information about the colour of v). This consistency is a testable property, and using this it is easy to see that any property that can be certified by an edge- and node-colouring can also be certified by an edge-colouring.

We could allow loops in the k -coloured digraphs certifying a property. (When taking the shadow, we delete the loops.) This is trivially equivalent to the node- and edge-coloured case, where the colours of the loops are considered to be colours of the corresponding nodes. (It is less obvious to see why we have to get rid of the loops instead of treating them just like the other edges. The reason for this is technical: we are going to use graph limit theory, and loops disappear in the limit.)

Alternatively, instead of a k -colouring we could specify k binary relations on $V(G)$ as a certificate (this would be more in the spirit of mathematical logic). The fact that in a colouring they are disjoint and partition the set $\{(i, j) : i, j \in V(G), i \neq j\}$ can be easily tested. Conversely, a set of k irreflexive relations defines a 2^k -colouring. As long as we are not concerned with efficiency, these two ways of looking at certificates are equivalent. (Further, possibly stronger extensions will be discussed in the last section.)

We close this Introduction with an example. One of the first non-trivial results about property testing concerned the maximum cut [4, 13]. Let us show how one version of this can be derived using Theorem 1.1. Let $0 < c < 1/4$, and let \mathcal{P} be the property of a graph G that its maximum cut contains at least $c|V(G)|^2$ edges. We claim that this property is testable.

A cut in a graph G can be encoded as a 2-colouring α of the nodes. We can encode this as a 4-colouring of the edges of a complete digraph, in which we colour an edge \vec{ij} with

$$\left\{ \begin{array}{l} 1 \text{ if } ij \in E(G) \text{ and } \alpha(i) = 1, \\ 2 \text{ if } ij \in E(G) \text{ and } \alpha(i) = 2, \\ 3 \text{ if } ij \notin E(G) \text{ and } \alpha(i) = 1, \\ 4 \text{ if } ij \notin E(G) \text{ and } \alpha(i) = 2. \end{array} \right.$$

The 2-shadow of this 4-coloured digraph is G .

Let \mathcal{Q} be the following property of 4-coloured digraphs L :

- there are at least $c|V(L)|^2$ edges \vec{ij} coloured with 1 such that \vec{ji} is coloured with 2, and
- for any three distinct nodes i, j and k , either both or neither of \vec{ij}, \vec{ik} have colour in $\{1, 3\}$, and
- for any two distinct nodes i and j , either both or neither of \vec{ij} and \vec{ji} have colour in $\{1, 2\}$.

(The last two conditions guarantee the consistency of the encoding of the node-2-colouring as an edge-4-colouring.) This property is trivially testable, and its 2-shadow is \mathcal{P} . So our theorem implies that \mathcal{P} is testable.

2. Preliminaries

2.1. Convergence and limits

Convergence of a sequence of dense finite graphs was defined by Borgs, Chayes, Lovász, Sós and Vesztegombi [5, 6]. Graphons were introduced by Lovász and Szegedy in [15] as limits of convergent sequences of finite graphs. We have to extend these notions to coloured digraphs; this was essentially done in [18], but we use here a somewhat different (simpler) terminology.

For a graph G and integer $r \leq |V(G)|$, let us select an ordered r -tuple of nodes of G randomly and uniformly (without repetition). Let $\mathbb{G}(r, G)$ denote the subgraph induced by these nodes. If G is coloured and directed, then $\mathbb{G}(r, G)$ is also a coloured digraph in the obvious way.

We say that a sequence of k -coloured digraphs L_n is *convergent* if $|V(G_n)| \rightarrow \infty$, and for every $r \geq 1$, the distribution of $\mathbb{G}(r, L_n)$ tends to a limit as $n \rightarrow \infty$. Note that this distribution is over a finite set, so it does not matter in which norm we want it to converge.

The limit object of a convergent sequence of simple graphs can be described as a symmetric measurable function $W : [0, 1]^2 \rightarrow [0, 1]$, called a *graphon*. We will need the more general notion of a *kernel*, a bounded symmetric measurable function $W : [0, 1]^2 \rightarrow \mathbb{R}$. Dropping the condition of symmetry, we get *digraphons* and *dikernels*.

For a sequence of k -coloured digraphs, the limit object is a bit more complicated: it consists of k digraphons (W^1, \dots, W^k) such that $\sum_h W^h = 1$. We call the k -tuple $\mathbf{W} = (W^1, \dots, W^k)$ a *k-digraphon*. We define a *k-dikernel* analogously.

Let L be a k -coloured digraph with $V(L) = [n]$. Let S_n denote the partition of $[0, 1]$ into n intervals S_1, \dots, S_n of equal length. We associate with L a *k-digraphon* $\mathbf{W}_L = (W_L^1, \dots, W_L^k)$, where

$$\mathbf{W}_L^h(x, y) = \begin{cases} 1 & \text{if } x \in S_i, y \in S_j \text{ (} i \neq j \text{), and the colour of } \vec{ij} \text{ is } h, \\ 1/k & \text{if } x, y \in S_i, \\ 0 & \text{otherwise,} \end{cases}$$

where the case $i = j$ is defined in this way only to make sure that $\sum_h \mathbf{W}_L^h = 1$ (this will play no subsequent role).

More generally, we can consider a *fractionally k-coloured digraph* H in which we have k non-negative weights $\beta^1(i, j), \dots, \beta^k(i, j)$ for every ordered pair (i, j) of nodes, with $\sum_h \beta^h(i, j) = 1$. It will be convenient to allow $i = j$, *i.e.*, loops with weights, but these will play a minor role in the proofs only. We consider every k -coloured digraph as a special case, where β^h is the indicator function of edge colour h , and $\beta^h(i, i) = 1/k$ for all $h \in [k]$ and $i \in V(H)$. For a fractionally k -coloured digraph H , we define the *k-digraphon* \mathbf{W}_H in the obvious way.

We can sample a k -coloured digraph $\mathbb{G}(r, \mathbf{W})$ on node set $[r]$ from a k -digraphon \mathbf{W} as follows. We choose r independent random points $X_1, \dots, X_r \in [0, 1]$ uniformly, and we colour a pair (i, j) ($i \neq j$) with colour h with probability $W^h(X_i, X_j)$ (independently for different pairs of nodes).

The following fact is proved in [18].

Proposition 2.1. *Let L_n be a convergent sequence of k -coloured digraphs. Then there is a k -digraphon \mathbf{W} such that $\mathbb{G}(r, L_n) \rightarrow \mathbb{G}(r, \mathbf{W})$ in distribution.*

We write $L_n \rightarrow \mathbf{W}$ if this holds.

Let W be a dikernel, and let $\mathcal{J} = \{S_1, \dots, S_m\}$ be a partition of $[0, 1]$ into measurable sets with positive measure. We denote by $W_{\mathcal{J}}$ the dikernel obtained by averaging W in every rectangle $S_i \times S_j$. More precisely, for $x \in S_i$ and $y \in S_j$ we define

$$W_{\mathcal{J}}(x, y) = \frac{1}{\lambda(S_i)\lambda(S_j)} \int_{S_i \times S_j} W(u, z) \, du \, dz$$

(where λ is the Lebesgue measure).

We quote a well-known fact (see, e.g., [20, (6.3) Theorem, p. 118]).

Proposition 2.2. *For every dikernel W , we have $W_{S_n} \rightarrow W$ ($n \rightarrow \infty$) almost everywhere.*

2.2. Many distances

2.2.1. Cut norm and cut distance.. Convergence to a k -digraphon can be described in more explicit forms. Let us start by recalling the *cut-norm-distance* of two graphs G and G' on the same node set V (introduced by Frieze and Kannan [11]),

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

where $e_G(S, T)$ denotes the number of edges with one endpoint in S and the other in T . A related notion for dikernels is the *cut norm*

$$\|W\|_{\square} = \sup_{S, T \subseteq [0, 1]} \left| \int_{S \times T} W(x, y) \, dx \, dy \right|.$$

The cut norm defines a distance function between two dikernels in the usual way by $d_{\square}(U, W) = \|U - W\|_{\square}$.

The cut norm has many nice properties (see [6]), for which we need the following result [17, Lemma 2.2].

Proposition 2.3. *Let W_n ($n = 1, 2, \dots$) be a sequence of uniformly bounded dikernels such that $\|W_n\|_{\square} \rightarrow 0$. Then, for every integrable function $Z : [0, 1]^2 \rightarrow \mathbb{R}$, we have $\|W_n Z\|_{\square} \rightarrow 0$.*

Proof. If Z is the indicator function of a rectangle, the conclusion follows from the definition of the $\|\cdot\|_{\square}$ norm. Hence the conclusion follows for step-functions, since they are linear combinations of a finite number of indicator functions of rectangles. Then it

follows for all integrable functions, since they are approximable in $L_1([0, 1]^2)$ by step-functions. □

From the point of view of graph limits, however, a kernel is only relevant up to a measure-preserving transformation of $[0, 1]$. Hence it is often more natural to consider the following distance notion, which we call the *cut distance*:

$$\delta_{\square}(U, W) = \inf_{\phi, \psi} \|U^{\phi} - W^{\psi}\|_{\square},$$

where $\phi, \psi : [0, 1] \rightarrow [0, 1]$ are measure-preserving maps, and $U^{\phi}(x, y) = U(\phi(x), \phi(y))$. This defines a pseudometric on the set of kernels (it is only a pseudometric, since different kernels may have distance 0). An important fact is that endowing the space of all graphons with this pseudometric makes it compact [16].

For two graphs G and G' (not necessarily with the same number of nodes) we define

$$\delta_{\square}(G, G') = \delta_{\square}(W_G, W_{G'}).$$

(There is a finite description of this quantity in terms of the value of a nonlinear min-max program [6, 14], but it is quite complicated, and for us this less explicit definition will be sufficient.) It is easy to see that if $V(G) = V(G')$, then

$$\delta_{\square}(G, G') \leq d_{\square}(G, G').$$

These distance notions can be extended to coloured graphs and kernels. For two fractionally k -coloured digraphs H and H' on the same node set V , with edgeweights $\beta_H^h(i, j)$ and $\beta_{H'}^h(i, j)$ ($h = 1, \dots, k$), let

$$d_{\square}(H, H') = \frac{1}{|V|^2} \sum_{h=1}^k \max_{S, T \subseteq V} \left| \sum_{i \in S, j \in T} (\beta_H^h(i, j) - \beta_{H'}^h(i, j)) \right|.$$

(We could interchange the summation and the maximization at the cost of decreasing d_{\square} by an irrelevant bounded factor. The definition above will be more convenient for us.)

We generalize the cut-norm-distance to two k -digraphons $\mathbf{U} = (U^1, \dots, U^k)$ and $\mathbf{W} = (W^1, \dots, W^k)$:

$$d_{\square}(\mathbf{U}, \mathbf{W}) = \sum_{h=1}^k \|U^h - W^h\|_{\square}. \tag{2.1}$$

As above, we need the ‘unlabelled’ cut distance

$$\delta_{\square}(\mathbf{U}, \mathbf{W}) = \inf_{\phi, \psi} d_{\square}(\mathbf{U}^{\phi}, \mathbf{W}^{\psi}), \tag{2.2}$$

where \mathbf{U}^{ϕ} is obtained from \mathbf{U} by applying ϕ to each of the $2k$ variables of the k functions constituting \mathbf{U} .

For two fractionally k -coloured digraphs H and H' (not necessarily with the same number of nodes), we define

$$\delta_{\square}(H, H') = \delta_{\square}(\mathbf{W}_H, \mathbf{W}_{H'}).$$

The following result is proved (in a more general form) in [18].

Proposition 2.4. *Let L_n be a sequence of k -coloured digraphs, and let \mathbf{W} be a k -digraphon. Then $L_n \rightarrow \mathbf{W}$ if and only if $\delta_{\square}(\mathbf{W}_{L_n}, \mathbf{W}) \rightarrow 0$.*

We cannot claim convergence in the d_{\square} distance, since \mathbf{W}_{L_n} depends on the labelling of the nodes of L_n , while the convergence $L_n \rightarrow \mathbf{W}$ does not. However, at least in the case of simple graphs, the following stronger version is true [6, Theorem 4.16].

Proposition 2.5. *Let G_n be a sequence of graphs, and let U be a graphon such that $G_n \rightarrow U$. Then the graphs G_n can be labelled so that $\|W_{G_n} - U\|_{\square} \rightarrow 0$.*

Edit distance. From the point of view of property testing, the ‘edit distance’ is very important (in fact, from the point of view of analysis, property testing is about the interplay between the edit distance and the cut distance). For two graphs G and G' on the same set of nodes $V = V(G) = V(G')$, their *edit distance* is defined by

$$d_1(G, G') = \frac{|E(G) \Delta E(G')|}{|V|^2}.$$

We generalize this for two k -coloured digraphs L and L' on the same node set:

$$d_1(L, L') = \frac{D_2}{|V|^2},$$

where D_2 is the number of edges coloured differently, in L and L' .

For two kernels, their edit distance is just their L_1 -distance as functions. For two k -digraphons, their edit distance is defined by a formula very similar to (2.1):

$$d_1(\mathbf{U}, \mathbf{W}) = \sum_{h=1}^k \|U^h - W^h\|_1.$$

Similarly to (2.2), we could define the unlabelled version of the edit distance, but we do not need it in this paper.

The following (easy) characterization of testability of graph properties was formulated in Theorem 3.21 of [17]. For a graph property \mathcal{P} , we write $d_1(G, \mathcal{P}) = \min\{d_1(G, F) : F \in \mathcal{P}, V(F) = V(G)\}$, and similarly for the d_{\square} distance. Analogously, we define $\delta_{\square}(G, \mathcal{P}) = \inf\{\delta_{\square}(G, F) : F \in \mathcal{P}\}$ (since the δ_{\square} distance is defined between any two graphs, we may have an infinite number of graphs F to consider here).

Proposition 2.6. *A graph property \mathcal{P} is testable if and only if, for any two sequences (G_n) and (H_n) of graphs with $|V(G_n)|, |V(H_n)| \rightarrow \infty$, $H_n \in \mathcal{P}$, and $\delta_{\square}(G_n, H_n) \rightarrow 0$, we have $d_1(G_n, \mathcal{P}) \rightarrow 0$.*

We note that the condition says that G_n is close to some graph in \mathcal{P} (not necessarily with the same number of nodes, but with size tending to infinity) in the δ_{\square} distance, while the conclusion is that it must be close to a graph in \mathcal{P} on the same node set in the edit distance.

The following extension of this fact to coloured digraphs can be proved along the same lines, and we state it without proof.

Proposition 2.7. *A property \mathcal{Q} of k -coloured digraphs is testable if and only if for any two sequences (L_n) and (J_n) of k -coloured digraphs with $|V(L_n)|, |V(J_n)| \rightarrow \infty, J_n \in \mathcal{Q}$, and $\delta_{\square}(L_n, J_n) \rightarrow 0$, we have $d_1(L_n, \mathcal{Q}) \rightarrow 0$.*

3. Main proof

We start with a randomized construction to obtain a k -coloured digraph from a fractionally k -coloured digraph H : we colour directed edge \vec{ij} ($i \neq j$) with colour h with probability $\beta^h(i, j)$. (The weights of the loops play no role here.) For different pairs i, j we make an independent decision. We denote this random k -coloured digraph by $\mathbb{L}(H)$.

Lemma 3.1. *Let H be a fractionally k -coloured digraph on n nodes. Then*

$$d_{\square}(H, \mathbb{L}(H)) \leq \frac{10k}{\sqrt{n}}$$

with probability at least $1 - ke^{-n}$.

Proof. For two colours, this is just Lemma 4.3 in [6]. For general k , it follows by applying this fact to each colour separately. □

The main step in the proof of Theorem 1.1 is the following lemma.

Lemma 3.2. *Let $\mathbf{W} = (W^1, \dots, W^k)$ be a k -digraphon, and suppose that $U = \sum_{h=1}^m W^h$ is symmetric (where $1 \leq m \leq k$). Let F_n be a sequence of simple graphs such that $F_n \rightarrow U$. Then there exist k -coloured digraphs J_n on $V(F_n)$ such that the m -shadow J'_n of J_n is F_n , and $J_n \rightarrow \mathbf{W}$.*

Proof. First, we construct a fractionally k -coloured digraph H_n on $V(F_n)$. To keep the notation simple, assume that $V(F_n) = [n]$. By Proposition 2.5, we can choose the labelling of the nodes of each F_n so that $\|W_{F_n} - U\|_{\square} \rightarrow 0$.

For every pair $i, j \in [n]$, we define

$$\beta^h(i, j) = \begin{cases} (A_n)_{ij} \frac{W_{S_n}^h(x, y)}{U_{S_n}(x, y)} & \text{if } 1 \leq h \leq m, \\ (1 - (A_n)_{ij}) \frac{W_{S_n}^h(x, y)}{1 - U_{S_n}(x, y)} & \text{if } m + 1 \leq h \leq k, \end{cases}$$

where A_n denotes the adjacency matrix of G_n and $x \in S_i$ and $y \in S_j$ (these numbers are independent of the choice of x and y). Note that if $U_{S_n}(x, y) = 0$, then necessarily $W_{S_n}^h(x, y) = 0$ for all $1 \leq h \leq m$, and we can define $W_{S_n}^h(x, y)/U_{S_n}(x, y)$ to mean $1/m$. Similarly, if $U_{S_n}(x, y) = 1$, then $W_{S_n}^h(x, y) = 0$ for all $m + 1 \leq h \leq k$, and we can define $W_{S_n}^h(x, y)/U_{S_n}(x, y)$ to mean $1/(k - m)$.

It is easy to check that $\sum_h \beta^h(i, j) = 1$ for all i and j . We claim that for the fractionally k -coloured digraph H_n constructed this way, we have

$$d_{\square}(\mathbf{W}_{H_n}, \mathbf{W}) \rightarrow 0 \quad (n \rightarrow \infty). \tag{3.1}$$

Here we have

$$d_{\square}(\mathbf{W}_{H_n}, \mathbf{W}) = \sum_{h=1}^k \|W_{H_n}^h - W^h\|_{\square},$$

and so it suffices to prove that $\|W_{H_n}^h - W^h\|_{\square} \rightarrow 0$ for every fixed h . We describe the proof for $h \leq m$; the other case is analogous. Since $0 \leq W^h \leq U$, we can write $W^h = UZ$, where $0 \leq Z \leq 1$, and $Z = 0$ if $U = 0$. Then we have

$$\|W_{H_n}^h - W^h\|_{\square} = \sup_{S, T \subseteq [0,1]} \left| \int_{S \times T} W_{H_n}^h - W^h \right|.$$

Substituting from the definition,

$$\int_{S \times T} (W_{H_n}^h - W^h) = \int_{S \times T} \left(W_{F_n}(x, y) \frac{W_{S_n}^h(x, y)}{U_{S_n}(x, y)} - W^h(x, y) \right) dx dy.$$

We split this integral as follows:

$$\int_{\substack{S \times T \\ U=0}} W_{F_n} \frac{(UZ)_{S_n}}{U_{S_n}} + \int_{\substack{S \times T \\ U \neq 0}} W_{F_n} \left(\frac{(UZ)_{S_n}}{U_{S_n}} - Z \right) + \int_{S \times T} (W_{F_n} - U)Z. \tag{3.2}$$

The first term, which is non-negative, can be estimated as follows:

$$\int_{\substack{S \times T \\ U=0}} W_{F_n} \frac{(UZ)_{S_n}}{U_{S_n}} \leq \int_{S \times T} W_{F_n} \mathbb{1}_{U=0} = \int_{S \times T} (W_{F_n} - U) \mathbb{1}_{U=0} \leq \|(W_{F_n} - U) \mathbb{1}_{U=0}\|_{\square}.$$

Here the right-hand side tends to 0 by Proposition 2.3. The second term can be estimated as follows. By Proposition 2.2, we have $(UZ)_{S_n} \rightarrow UZ$ and $U_{S_n} \rightarrow U$ almost everywhere. Hence $(UZ)_{S_n}/U_{S_n} \rightarrow Z$ in almost every point where $U \neq 0$. Since the integrand is bounded, this implies that

$$\left| \int_{\substack{S \times T \\ U \neq 0}} W_{F_n} \left(\frac{(UZ)_{S_n}}{U_{S_n}} - Z \right) \right| \leq \int_{U \neq 0} W_{F_n} \left| \frac{(UZ)_{S_n}}{U_{S_n}} - Z \right| \rightarrow 0.$$

Finally, for the third term in (3.2) we have

$$\left| \int_{S \times T} (W_{F_n} - U)Z \right| \leq \|(W_{F_n} - U)Z\|_{\square},$$

and here the right-hand side tends to 0, again by Proposition 2.3. This proves (3.1).

To complete the proof of the lemma, we consider the k -coloured digraphs $J_n = \mathbb{L}(H_n)$. By Lemma 3.1, we have

$$d_{\square}(J_n, H_n) \leq \frac{10k}{\sqrt{n}} \tag{3.3}$$

with probability at least $1 - ke^{-n}$. Since $\sum_n e^{-n}$ is convergent, the Borel–Cantelli lemma implies that almost surely (3.3) holds for all but a finite number of indices n . Choosing the

J_n so that this occurs, we have $d_{\square}(J_n, H_n) = d_{\square}(\mathbf{W}_{J_n}, \mathbf{W}_{H_n}) \rightarrow 0$, and hence $d_{\square}(\mathbf{W}_{J_n}, \mathbf{W}) \rightarrow 0$. □

Proof of Theorem 1.1. Let \mathcal{P} be a non-deterministically testable property; we show that it is testable. By Proposition 2.6 it suffices to prove that if (F_n) and (G_n) are two sequences of graphs such that $|F_n|, |G_n| \rightarrow \infty$, $G_n \in \mathcal{P}$, and $\delta_{\square}(F_n, G_n) \rightarrow 0$, then $d_1(F_n, \mathcal{P}) \rightarrow 0$.

Since \mathcal{P} is non-deterministically testable, there are integers $1 \leq m \leq k$ and a testable property \mathcal{Q} of k -coloured digraphs such that $\mathcal{P} = \mathcal{Q}'_m$. (We consider k and m fixed, and will suppress the subscript m hereafter.) Since $G_n \in \mathcal{P}$, there are k -coloured digraphs $L_n \in \mathcal{Q}$ such that $G_n = L'_n$.

We may assume that the union of colours $1, \dots, m$ contains every edge of G_n in both directions; otherwise, we refine the colouring so that no edge in G and in its complement gets the same colour in any direction (this doubles the number of colours at most). By selecting a subsequence, we may assume that the sequence (L_n) is convergent. By Lemma 2.1, there exists a k -digraphon $\mathbf{W} = (W^1, \dots, W^k)$ representing its limit. Let $U = \sum_{h=1}^m W^h$; then $G_n \rightarrow U$. From $\delta_{\square}(G_n, F_n) \rightarrow 0$ we see that $F_n \rightarrow U$.

Now we invoke Lemma 3.2, and construct k -coloured digraphs J_n such that $J'_n = F_n$ and $J_n \rightarrow \mathbf{W}$. Hence $d_{\square}(J_n, \mathcal{Q}) \rightarrow 0$. Since \mathcal{Q} is testable, this implies by Proposition 2.7 that $d_1(J_n, \mathcal{Q}) \rightarrow 0$, and so we can change the colour of $o(n^2)$ edges in J_n so that the resulting k -coloured digraph M_n belongs to \mathcal{Q} . But then $M'_n \in \mathcal{P}$, and M'_n differs from F_n in $o(n^2)$ edges only, so $d_1(F_n, \mathcal{P}) \leq d_1(F_n, M'_n) \rightarrow 0$. □

4. Applications

There are many graph properties that can be certified by a node-colouring: k -colourable graphs, split graphs, *etc.* Many of these properties are hereditary, and so their testability follows also by the Alon–Shapira theorem mentioned in the Introduction. Here we formulate some consequences for non-hereditary graph properties.

We have shown in the Introduction how to derive from our theorem that the property of a graph G that it has a cut with at least $c|V(G)|^2$ edges is testable. Similar arguments can be applied to the maximum bisection problem. Goldreich, Goldwasser and Ron [13, Theorem 9.1] prove the testability of more general properties, namely the existence of multiway cuts with upper and lower bounds on the sizes of partition classes as well as on edge densities between parts. The existence of such a cut can be certified by a node-colouring, and so this property is trivially non-deterministically testable. So their general result (without explicit bounds on the sample size) follows from Theorem 1.1.

Alon, Fischer, Krivelevich and Szegedy [2] prove that a graph property is testable, provided that it is expressible in the form $\exists x_1 \dots \exists x_a \forall y_1 \dots \forall y_b \Phi(x_1, \dots, x_a, y_1, \dots, y_b)$, where the x_i and y_j are variables ranging over nodes, and Φ is a (quantifier-free) Boolean expression involving equality and adjacency of the variables x_i and y_j . They also give an example showing that graph properties defined by more general first-order sentences (with more quantifier alternations) are not necessarily testable.

Combining our result with their method, we get a more general testability condition in terms of logical formulas.

Corollary 4.1. *Let \mathcal{P} be a graph property expressible by a second-order formula of the form $\Psi = \exists S_1 \dots \exists S_c \exists x_1 \dots \exists x_a \forall y_1 \dots \forall y_b \Phi(S_1, \dots, S_c, x_1, \dots, x_a, y_1, \dots, y_b)$, where the S_i are variables ranging over unary or binary relations, x_i and y_j are variables ranging over nodes, and Φ is a (quantifier-free) Boolean expression involving equality, adjacency, and the relations S_i of the variables x_i and y_j . Then \mathcal{P} is testable.*

Sketch of proof. For every graph G satisfying the given formula Ψ , we construct a complete digraph L on $V(G)$ with nodes and edges coloured, so that G is a shadow of L . The relations S_1, \dots, S_c can be encoded into edge- and node-colourings of the complete digraph on $V(G)$, as described at the end of the Introduction. Using the main trick from the proof of Alon, Fischer, Krivelevich and Szegedy, we encode the (existentially quantified) nodes x_1, \dots, x_a into a node-colouring: the colour of a node expresses which of the nodes x_i it is equal to or connected to, along with the full description of adjacency, equality, and the relations S_j , between the nodes x_i . In this way we obtain the node- and edge-coloured complete digraph L . The validity of Φ (for this choice of $S_1, \dots, S_c, x_1, \dots, x_a$) is equivalent to forbidding a finite number of coloured induced subgraphs in L . Let \mathcal{Q} be the property of node- and edge-coloured complete digraphs that they do not have any of these forbidden subgraphs. This property is testable (this follows, for example, from a rather straightforward generalization of Theorem 6.1 in [2] to edge-coloured graphs). The property defined by Ψ is a shadow of \mathcal{Q} , and hence it is testable as well. \square

In [17], the *upward closure* of a graph property \mathcal{P} was defined as the graph property \mathcal{P}^\uparrow consisting of those graphs that have a spanning subgraph in \mathcal{P} .

Corollary 4.2. *The upward closure of a testable graph property is testable.*

Proof. Obtaining a spanning subgraph in \mathcal{P} can be formulated as a 3-edge-colouring of the complete graph, in which colour 1 forms a graph in \mathcal{P} , and G is the 2-shadow of this 3-edge-coloured graph. If \mathcal{P} is testable, then the property of 3-edge-coloured graphs that edges with colour 1 form a graph with property \mathcal{P} is testable, and so the upward closure of \mathcal{P} is testable. \square

Fischer and Newman [10] proved that the edit distance to a testable graph property is a testable parameter. This is equivalent to the following corollary of our theorem.

Corollary 4.3. *If \mathcal{P} is a testable property, then the property that $d_1(G, \mathcal{P}) \leq c$ is also testable for every $c \geq 0$.*

Proof. Being closer than c to \mathcal{P} can be formulated as follows: the complete graph on $V(G)$ has a 4-edge-colouring such that edges with colours 1 and 2 form the graph G , edges with colours 1 and 3 form a graph with property \mathcal{P} , and edges with colours 2 and 3 are fewer than $c|V(G)|^2$. If \mathcal{P} is testable, then this property of 4-coloured graphs is testable, and hence the property that $d_1(G, \mathcal{P}) \leq c$ is non-deterministically testable, and so by Theorem 1.1, it is testable. \square

5. Parameter estimation

Let f be a bounded graph parameter (i.e., a function defined on simple graphs, invariant under isomorphism). We say that f is *estimable* if, for every $\varepsilon, \delta > 0$, there is a positive integer k such that if G is a graph with at least k nodes and we select a random k -set $X \subseteq V(G)$, then

$$\mathbb{P}(|f(G) - f(G[X])| > \varepsilon) < \delta. \quad (5.1)$$

We define an estimable parameter of edge-coloured digraphs similarly. If g is such a parameter, then we can define

$$g'(G) = \max\{g(L) : L' = G\}.$$

An argument very similar to the proof of Theorem 1.1 above gives the following theorem.

Theorem 5.1. *If g is an estimable parameter of k -coloured digraphs, then g' is estimable as well.*

We could of course replace the maximum in the definition of g' by minimum.

As an example, let us consider graphs L whose nodes are 2-coloured red and blue, and let $g(L)$ denote the number of 2-coloured edges, divided by $|V(G)|^2$. Then g is trivially estimable. The simple graph parameter g' is the maximum cut (normalized), so this is estimable.

6. Concluding remarks

There are several possible analogues and extensions of our results. One could consider certificates in the form of t -ary relations for any t . One could then allow hypergraphs instead of the original graphs. A limit theory for hypergraphs is available (Elek and Szegedy [8]), and we expect our main result to generalize to hypergraphs; however, several of the auxiliary results we have used have not been extended, and a full proof will take further research.

A generalization in a different direction would be to consider, instead of colouring, node and edge decorations from a compact topological space. For example, the property of being a threshold graph can be certified by a decoration of the nodes by numbers from $[0, 1]$. The limit theory for graphs has been extended to compact decorations [18]; perhaps our main result would extend too, but this would take further research.

We should point out that our results are non-effective, in that they do not provide any explicit bound on how large a sample size must be chosen for a given error bound. In this sense, what we have given is a pure existence proof of an algorithm. From a practical point of view, this does not make much difference from related results based on the Regularity Lemma, but from a theoretical point of view, it would be interesting to determine whether Theorem 1.1 can be proved with an effective bound.

Finally, let us mention that the situation is quite different in the case of graphs with bounded degree (for which a limit theory analogous to the dense case is available, and

property testing has been extensively studied). Here the sampling method is to select r random nodes (uniformly), and explore their neighbourhoods of depth r . The property of a graph G that ‘ G is the disjoint union of two graphs on at least $|V(G)|/3$ nodes’ can be certified by colouring the nodes in these two graphs with different colours, so this property is non-deterministically testable. On the other hand, sampling will not distinguish between an expander graph and the disjoint union of two copies of it, so this property is not testable.

Acknowledgements

We thank the referees of this paper for their very helpful and thorough remarks.

Research was supported by the European Research Council grant 227701 and by the National Science Foundation under agreement DMS-0835373. Hospitality of the Institute for Advanced Study is gratefully acknowledged. Any opinions and conclusions expressed in this material are those of the authors and do not necessarily reflect the views of the NSF or of the ERC.

References

- [1] Alon, N. and Shapira, A. (2008) A characterization of the (natural) graph properties testable with one-sided error. *SIAM J. Comput.* **37** 1703–1727.
- [2] Alon, N., Fischer, E., Krivelevich, M. and Szegedy, M. (2000) Efficient testing of large graphs. *Combinatorica* **20** 451–476.
- [3] Alon, N., Fischer, E., Newman, I. and Shapira, A. (2006) A combinatorial characterization of the testable graph properties: It’s all about regularity. In *Proc. 38th ACM Symposium on the Theory of Computing: STOC*, pp. 251–260.
- [4] Arora, S., Karger, D. and Karpinski, M. (1995) Polynomial time approximation schemes for dense instances of NP-hard problems. In *Proc. 27th ACM Symposium on the Theory of Computing*, pp. 284–293.
- [5] Borgs, C., Chayes, J., Lovász, L., Sós, V. T. and Vesztergombi, K. (2006) Counting graph homomorphisms. In *Topics in Discrete Mathematics* (M. Klazar *et al.*, eds), Springer, pp. 315–371.
- [6] Borgs, C., Chayes, J. T., Lovász, L., Sós, V. T. and Vesztergombi, K. (2008) Convergent graph sequences I: Subgraph frequencies, metric properties, and testing. *Adv. Math.* **219** 1801–1851.
- [7] Borgs, C., Chayes, J. T., Lovász, L., Sós, V. T. and Vesztergombi, K. (2012) Convergent graph sequences II: Multiway cuts and statistical physics. *Ann. of Math.* **176** 151–219.
- [8] Elek, G. and Szegedy, B. Limits of hypergraphs, removal and regularity lemmas: A non-standard approach. [arXiv:0705.2179](https://arxiv.org/abs/0705.2179)
- [9] Fischer, E. (2001) 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** 97–126.
- [10] Fischer, E. and Newman, I. (2005) Testing versus estimation of graph properties. In *Proc. 37th ACM Symposium on the Theory of Computing*, pp. 138–146.
- [11] Frieze, A. and Kannan, R. (1999) Quick approximation to matrices and applications. *Combinatorica* **19** 175–220.
- [12] Goldreich, O., ed. (2010) *Property Testing: Current Research and Surveys*, Vol. 6390 of *Lecture Notes in Computer Science*, Springer.
- [13] Goldreich, O., Goldwasser, S. and Ron, D. (1998) Property testing and its connection to learning and approximation. *J. Assoc. Comput. Mach.* **45** 653–750.

- [14] Lovász, L. (2012) *Large Networks and Graph Limits*, AMS.
- [15] Lovász, L. and Szegedy, B. (2006) Limits of dense graph sequences. *J. Combin. Theory Ser. B* **96** 933–957.
- [16] Lovász, L. and Szegedy, B. (2007) Szemerédi's Lemma for the analyst. *Geom. Func. Anal.* **17** 252–270.
- [17] Lovász, L. and Szegedy, B. (2010) Testing properties of graphs and functions. *Israel J. Math.* **178** 113–156.
- [18] Lovász, L. and Szegedy, B. Limits of compact decorated graphs. [arXiv:1010.5155](https://arxiv.org/abs/1010.5155)
- [19] Rubinfeld, R. and Sudan, M. (1996) Robust characterization of polynomials with applications to program testing. *SIAM J. Comput.* **25** 252–271.
- [20] Saks, S. (1964) *Theory of the Integral*, Dover.