Estimating Parameters Associated with Monotone Properties*

Carlos Hoppen[1], Yoshiharu Kohayakawa[2], Richard Lang[3], Hanno Lefmann[4], and Henrique Stagni[5]

1 **Universidade Federal do Rio Grande do Sul, Porto Alegre, Brazil**
choppen@ufrgs.br
2 **Universidade de São Paulo, São Paulo, Brazil**
yoshi@ime.usp.br
3 **Universidad de Chile, Santiago, Chile**
rlang@dim.uchile.cl
4 **Technische Universität Chemnitz, Chemnitz, Germany**
Lefmann@Informatik.TU-Chemnitz.de
5 **Universidade de São Paulo, São Paulo, Brazil**
stagni@ime.usp.br

Abstract

There has been substantial interest in estimating the value of a graph parameter, i.e., of a real function defined on the set of finite graphs, by sampling a randomly chosen substructure whose size is independent of the size of the input. Graph parameters that may be successfully estimated in this way are said to be *testable* or *estimable*, and the *sample complexity* $q_z = q_z(\varepsilon)$ of an estimable parameter z is the size of the random sample required to ensure that the value of $z(G)$ may be estimated within error ε with probability at least $2/3$. In this paper, we study the sample complexity of estimating two graph parameters associated with a monotone graph property, improving previously known results. To obtain our results, we prove that the vertex set of any graph that satisfies a monotone property $\mathcal{P}$ may be partitioned equitably into a constant number of classes in such a way that the cluster graph induced by the partition is not far from satisfying a natural weighted graph generalization of $\mathcal{P}$. Properties for which this holds are said to be *recoverable*, and the study of recoverable properties may be of independent interest.

1998 ACM Subject Classification G.2.2 Graph Theory

Keywords and phrases parameter estimation, parameter testing, edit distance to monotone graph properties, entropy of subgraph classes, speed of subgraph classes

Digital Object Identifier 10.4230/LIPIcs.APPROX-RANDOM.2016.35

1 Introduction

In the last two decades, a lot of effort has been put into finding constant-time randomized algorithms (conditional on sampling) to gauge whether a combinatorial structure satisfies

* C. Hoppen acknowledges the support of FAPERGS (Proc. 2233-2551/14-0) and CNPq (Proc. 448754/2014-2 and 308539/2015-0). C. Hoppen and H. Lefmann acknowledge the support of CAPES and DAAD via PROBRAL (CAPES Proc. 408/13 and DAAD 56267227 and 57141126 and 57245206). C. Hoppen, Y. Kohayakawa and H. Stagni thank FAPESP (Proc. 2013/03447-6) and NUMEC/USP (Project MaCLinC/USP) for their support. Y. Kohayakawa was partially supported by FAPESP (2013/07699-0) and CNPq (310974/2013-5 and 459335/2014-6). H. Stagni was supported by FAPESP (2015/15986-4) and CNPq (141970/2015-4 and 459335/2014-6).

some property, or to estimate the value of some numerical function associated with this combinatorial structure. In this paper, we focus on the graph case and, as usual, we consider algorithms that have the ability to query whether any desired pair of vertices in the input graph is adjacent or not. Let $\mathcal{G}$ be the set of finite simple graphs and let $\mathcal{G}(V)$ be the set of such graphs with vertex set V. We shall consider subsets $\mathcal{P}$ of $\mathcal{G}$ that are closed under isomorphism, which we call *graph properties*. To avoid technicalities, we restrict ourselves to graph properties $\mathcal{P}$ such that $\mathcal{P} \cap \mathcal{G}(V) \neq \emptyset$ whenever $V \neq \emptyset$. For instance, this includes all nontrivial *monotone* and *hereditary* graph properties, which are graph properties that are inherited by subgraphs and by induced subgraphs, respectively. Here, we will focus on monotone properties. The prototypical example of a monotone property is $\mathit{Forb}(F)$, the class of all graphs that do not contain a fixed graph F as a subgraph. More generally, if $\mathcal{P}$ is a monotone property and $\mathcal{F}$ contains all minimal graphs that are not in $\mathcal{P}$, then the graphs that lie in $\mathcal{P}$ are precisely those that do not contain an element of $\mathcal{F}$ as a subgraph. This class of graphs will be denoted by $\mathcal{P} = \mathit{Forb}(\mathcal{F})$. The elements of $\mathit{Forb}(\mathcal{F})$ are said to be *$\mathcal{F}$-free*.

A graph property $\mathcal{P}$ is said to be *testable* if, for every $\varepsilon > 0$, there exist a positive integer $q_{\mathcal{P}} = q_{\mathcal{P}}(\varepsilon)$, called the *query complexity*, and a randomized algorithm $\mathcal{T}_{\mathcal{P}}$, called a *tester*, which may perform at most $q_{\mathcal{P}}$ queries in the input graph, satisfying the following property. For an n-vertex input graph Γ, the algorithm $\mathcal{T}_{\mathcal{P}}$ distinguishes with probability at least $2/3$ between the cases in which Γ satisfies $\mathcal{P}$ and in which Γ is ε-far from satisfying $\mathcal{P}$, that is, in which no graph obtained from Γ by the addition or removal of at most $\varepsilon n^2/2$ edges satisfies $\mathcal{P}$. This may be stated in terms of graph distances: given two graphs Γ and Γ' on the same vertex set V, we may define the *normalized edit distance* between Γ and Γ' by $d_1(\Gamma, \Gamma') = \frac{2}{|V|^2} |E(\Gamma) \triangle E(\Gamma')|$, where $E(\Gamma) \triangle E(\Gamma')$ denotes the symmetric difference of their edge sets. If $\mathcal{P}$ is a graph property, we let the distance between a graph Γ and $\mathcal{P}$ be

$$d_1(\Gamma, \mathcal{P}) = \min\{d_1(\Gamma, \Gamma') \colon V(\Gamma') = V(\Gamma) \text{ and } \Gamma' \in \mathcal{P}\}.$$

For instance, if $\Gamma = K_n$ and $\mathcal{P} = \mathit{Forb}(K_3)$, Turán's Theorem ensures that $\binom{n}{2} - \lfloor n^2/4 \rfloor$ edges need to be removed to produce a graph that is K_3-free. In particular, $d_1(K_n, \mathit{Forb}(K_3)) \to 1/2$. Thus a graph property is testable if there is a tester with bounded query complexity that distinguishes with probability at least $2/3$ between the cases $d_1(\Gamma, \mathcal{P}) = 0$ and $d_1(\Gamma, \mathcal{P}) > \varepsilon$.

The systematic study of property testing was initiated by Goldreich, Goldwasser and Ron [19], and there is a very rich literature on this topic. For instance, regarding testers, Goldreich and Trevisan [20] showed that it is sufficient to consider simpler canonical testers, namely those that randomly choose a subset X of vertices in Γ and then verify whether the induced subgraph $\Gamma[X]$ satisfies some related property $\mathcal{P}'$. For example, if the property being tested is having edge density $1/2$, then the algorithm will choose a random subset X of appropriate size and check whether the edge density of $\Gamma[X]$ is within, say, $\varepsilon/2$ of $1/2$. Regarding testable properties, Alon and Shapira [5] proved that every monotone graph property is testable, and, more generally, that the same holds for hereditary graph properties [4]. For more information about property testing, we refer the reader to [18] and the references therein.

In a similar vein, a function $z \colon \mathcal{G} \to \mathbb{R}$ from the set $\mathcal{G}$ of finite graphs into the real numbers is called a *graph parameter* if it is invariant under relabeling of vertices. A graph parameter $z \colon \mathcal{G} \to \mathbb{R}$ is *estimable* if for every $\varepsilon > 0$ and every large enough graph Γ, the value of $z(\Gamma)$ can be approximated up to an additive error of ε by an algorithm that only has access to a subgraph of Γ induced by a set of vertices of size $q_z = q_z(\varepsilon)$, chosen uniformly at random. The query complexity of such an algorithm is $\binom{q_z}{2}$ and the size q_z is called its

sample complexity. Estimable parameters have been considered in [14] and were defined in the above level of generality in [9]. They are often called *testable parameters*. Borgs et al. [9, Theorem 6.1] gave a complete characterization of the estimable graph parameters which, in particular, also implies that the distance from monotone graph properties is estimable. Their work uses the concept of graph limits and does not give explicit bounds on the query complexity required for this estimation.

Estimable parameters are closely related with the notion of *tolerant testing*, which was introduced by Parnas, Ron and Rubinfeld [23], and is a generalization of standard property testing. Let $0 \leq \varepsilon_1 < \varepsilon_2 \leq 1$. An $(\varepsilon_1, \varepsilon_2)$-tolerant tester for a graph property $\mathcal{P}$ is an algorithm that receives a graph Γ as input and distinguishes between the cases $d_1(\Gamma, \mathcal{P}) \leq \varepsilon_1$ and $d_1(\Gamma, \mathcal{P}) \geq \varepsilon_2$ with probability at least $2/3$ and constant query complexity. Fischer and Newman [14] proved that every testable graph property $\mathcal{P}$ has a $(d - \varepsilon, d)$-tolerant tester, for every $d, \varepsilon > 0$. The distance from a graph to $\mathcal{P}$ can then be estimated by successively running such tolerant testers. Since every monotone graph property is testable, it follows that the distance to such a property is estimable. Later, Alon, Shapira and Sudakov [6, Theorem 1.2] proved that the distance to every monotone graph property $\mathcal{P}$ is estimable using a more natural algorithm, which simply computes the distance from the induced sampled graph to $\mathcal{P}$. However, one disadvantage of these approaches is that their analysis relies heavily on stronger versions of the Szemerédi Regularity Lemma [24, 2]. Therefore, their algorithms to estimate the distance from monotone graph properties have a query complexity of order at least $\texttt{TOWER}(\texttt{poly}(1/\varepsilon))$, by which we mean a tower of twos of height that is polynomial in $1/\varepsilon$. Moreover, it follows from a result of Gowers [21] that any approach based on the Szemerédi Regularity Lemma cannot lead to a bound that is better than $\texttt{TOWER}(\texttt{poly}(1/\varepsilon))$.

In this paper, we introduce the concept of *recoverable* graph properties (Definition 10). Roughly speaking, given a function $f\colon (0,1] \to \mathbb{R}$, we say a graph property $\mathcal{P}$ is f-recoverable if every large graph $G \in \mathcal{P}$ is ε-close to admitting a partition $\mathcal{V}$ of its vertex set into at most $f(\varepsilon)$ classes that witnesses pertinence in $\mathcal{P}$ (i.e., such that any graph that can be partitioned in the same way must be in $\mathcal{P}$). We prove the following result for recoverable properties.

▶ **Theorem 1.** *Let $\mathcal{P}$ be an f-recoverable graph property, for some function $f\colon (0,1] \to \mathbb{R}$. Then, for all $\varepsilon > 0$ there is n_0 such that, for any graph Γ with $|V(\Gamma)| \geq n_0$, the graph parameter*

$$z(\Gamma) = d_1(\Gamma, \mathcal{P})$$

can be estimated within an additive error of ε with sample complexity $2^{\texttt{poly}(f(\varepsilon/6)/\varepsilon)}$.

We also show (Theorem 16) that every monotone graph property $Forb(\mathcal{F})$ is f-recoverable for some function f that depends only on the bounds for the weighted graph Removal Lemma (Lemma 12) for the family $\mathcal{F}$ – the Removal Lemma states that if a graph is far from being $\mathcal{F}$-free, then it must contain many copies of some element of $\mathcal{F}$ of bounded size. Thus, our approach can improve the required sample complexity for estimating $d_1(\cdot, Forb(\mathcal{F}))$ for families $\mathcal{F}$ for which there are better bounds for the Removal Lemma. In particular, as a consequence of Theorem 1, Theorem 16 and recent improvements by Fox [15] on the bounds for the Removal Lemma, we have the following result.

▶ **Corollary 2.** *Let $\mathcal{F}$ be a* finite *family of graphs. Then, for all $\varepsilon > 0$ there is n_0 such that, for any graph Γ with $|V(\Gamma)| \geq n_0$, the graph parameter*

$$z(\Gamma) = d_1(\Gamma, Forb(\mathcal{F}))$$

can be estimated within an additive error of ε with sample complexity $\texttt{TOWER}(\texttt{poly}(\log(1/\varepsilon)))$.

We obtain similar results for another bounded graph parameter, which, for a graph family $\mathcal{F}$, counts the number of $\mathcal{F}$-free subgraphs of the input graph Γ. Formally, given a graph $\Gamma \in \mathcal{G}$ and a family $\mathcal{F}$ of graphs, we denote the set of all $\mathcal{F}$-free subgraphs of Γ by $Forb(\Gamma, \mathcal{F}) = \{G \in Forb(\mathcal{F}) : G \text{ is a subgraph of } \Gamma\}$, and we consider the parameter

$$z(\Gamma) = \frac{1}{|V(\Gamma)|^2} \log_2 |Forb(\Gamma, \mathcal{F})|. \tag{1}$$

For example, if $\mathcal{F} = \{K_3\}$ and $\Gamma = K_n$, computing z requires estimating the number of K_3-free subgraphs of K_n, which was done by Erdős, Kleitman and Rothschild for $\mathcal{F} = \{K_k\}$ [13] (see also Erdős, Frankl and Rödl [12] for F-free subgraphs):

$$z(K_n) = \frac{1}{n^2} \log_2 |Forb(\Gamma, \mathcal{F})| = \frac{1}{n^2} \log_2 2^{\frac{1}{2}\binom{n}{2}+o(n^2)} \to \frac{1}{4}.$$

Counting problems of this type were considered by several people. (See, for instance, the logarithmic density in Bollobás [8].)

▶ **Theorem 3.** *Let $Forb(\mathcal{F})$ be an f-recoverable graph property, for some function $f\colon (0,1] \to \mathbb{R}$. Then, for all $\varepsilon > 0$ there is n_0 such that, for any graph Γ with $|V(\Gamma)| \geq n_0$, the graph parameter z defined in (1) can be estimated within an additive error of ε with sample complexity $2^{\mathtt{poly}(f(\varepsilon/6)/\varepsilon)}$.*

▶ **Corollary 4.** *Let $\mathcal{F}$ be a* finite *family of graphs. Then, for all $\varepsilon > 0$ there is n_0 such that, for any graph Γ with $|V(\Gamma)| \geq n_0$, the graph parameter z defined in (1) can be estimated within an additive error of ε with sample complexity $\mathtt{TOWER}(\mathtt{poly}(\log(1/\varepsilon)))$.*

We should mention that the statement of Theorem 3 does not hold for arbitrary non-monotone properties $\mathcal{P}$. For instance, if $\mathcal{P}$ is the hereditary property of graphs having no independent sets of size three, then K_n and $K_n - E(K_3)$ have quite a different number of subgraphs satisfying $\mathcal{P}$, although their distance is negligible. It follows from [9, Theorem 6.1] that this parameter is not estimable.

The remainder of the paper is structured as follows. In Section 2, we provide preliminary definitions that lead to the concept of a recoverable graph property, which is used to prove Theorems 1 and 3. Indeed, these two theorems are consequences of Theorem 18 and Theorem 19, respectively, which are stated in Section 3.

2 Recoverability

The main objective of this section is to introduce the concept of ε-recoverability and to restate our main results in terms of it.

2.1 Estimation over cluster graphs

A *weighted graph* R over a (finite) set of vertices V is a symmetric function from $V \times V$ to $[0,1]$. A weighted graph R may be viewed as a complete graph (with loops) in which a weight $R(i,j)$ is given to each edge $(i,j) \in V(R) \times V(R)$, where $V(R)$ denotes the vertex set of R. The set of all weighted graphs with vertex set V is denoted by $\mathcal{G}^*(V)$ and we define $\mathcal{G}^*$ as the union of all $\mathcal{G}^*(V)$ for V finite. In particular, a graph G is a rational weighted graph such that $G(i,i) = 0$, for every $i \in V(G)$, and either $G(i,j) = 1$ or $G(i,j) = 0$ for every $(i,j) \in V(G) \times V(G)$, $i \neq j$. For a weighted graph $R \in \mathcal{G}^*(V)$ and for sets $A, B \subset V$, we denote $e_R(A,B) = \sum_{(i,j)\in A\times B} R(i,j)$ and $e(R) = e(V,V)$.

Let $k > 0$ and let $R \in \mathcal{G}^*(V)$ be a weighted graph. We define a weighted graph $\mathbb{G}(k, R) \in \mathcal{G}^*([k])$ by assigning weight $R(x_i, x_j)$ to each edge $(i, j) \in [k] \times [k]$, where $\{x_i\}_{i=1}^k$ is a multiset of k vertices of V such that each x_i is chosen with uniform probability, independently of the others (with repetition). With this, we may define estimable parameters in the context of weighted graphs. Henceforth we write $b = a \pm x$ for $a - x \leq b \leq a + x$.

▶ **Definition 5.** We say that a function $z\colon \mathcal{G}^* \to \mathbb{R}$ (also called a *weighted graph parameter*) is *estimable* with *sample complexity* $q\colon (0,1) \to \mathbb{N}$ if, for every $\varepsilon > 0$ and every weighted graph $\Gamma^* \in \mathcal{G}^*(V)$ with $|V| \geq q(\varepsilon)$, we have $z(\Gamma^*) = z(\mathbb{G}(q, \Gamma^*)) \pm \varepsilon$ with probability at least $2/3$.

Given a graph G and vertex sets $U, W \subseteq V(G)$, let $E_G(U, W) = \{(u, w) \in V(G) \times V(G) : u \in U, w \in W\}$ and $e_G(U, W) = |E_G(U, W)|$. An *equipartition* $\mathcal{V} = \{V_i\}_{i=1}^k$ of a weighted graph R is a partition of its vertex set $V(R)$, such that $|V_i| \leq |V_j| + 1$ for all $(i, j) \in [k] \times [k]$. We often abuse terminology and say that $\mathcal{V}$ is a partition of R.

Let $\mathcal{V} = \{V_1, \ldots, V_k\}$ be an equipartition of a graph G. The *cluster graph* of G by $\mathcal{V}$ is a weighted graph $G/\mathcal{V} \in \mathcal{G}^*([k])$ such that $G/\mathcal{V}\,(i, j) = e_G(V_i, V_j)/(|V_i||V_j|)$ for all $(i, j) \in [k] \times [k]$. For a fixed integer $K > 0$, the set of all equipartitions of a vertex set V into at most K classes will be denoted by $\Pi_K(V)$. We also define the set $G/\Pi_K = \{G/\mathcal{V} : \mathcal{V} \in \Pi_K(V(G))\}$ of all cluster graphs of G of vertex size at most K. The following result states that graph parameters that can be expressed as the optimal value of some optimization problem over G/Π_K can be estimated with a query complexity that is only exponential in K and in the error parameter.

▶ **Theorem 6.** *Let $z\colon \mathcal{G} \to \mathbb{R}$ be a graph parameter and suppose that there is a weighted graph parameter $z^*\colon \mathcal{G}^* \to \mathbb{R}$ and constants $K > 0$ and $c > 0$ such that:*

1. *$z(\Gamma) = \min_{R \in \Gamma/\Pi_K} z^*(R)$, for every $\Gamma \in \mathcal{G}$ and*
2. *$|z^*(R) - z^*(R')| \leq c \cdot d_1(R, R')$, for all weighted graphs $R, R' \in \mathcal{G}^*$ on the same vertex set.*

Then z is estimable with sample complexity $\varepsilon \mapsto 2^{poly(K, c/\varepsilon)}$.

The proof of Theorem 6 is based on the following lemma, which asserts that the set of cluster graphs of a graph Γ is very 'similar' to the set of cluster graphs of 'large enough' samples of Γ.

▶ **Lemma 7.** *Given $K > 0$, $\varepsilon > 0$ there is $q = 2^{poly(K, 1/\varepsilon)}$ and n_0 such that the following holds. Consider a graph Γ on $n \geq n_0$ vertices and a random sample $\overline{\Gamma} = \mathbb{G}(q, \Gamma)$ with vertex sets V and $\overline{V}$, respectively. Then, with probability at least $2/3$, we have*

1. *for each $\mathcal{V} \in \Pi_K(V)$, there is a $\overline{\mathcal{V}} \in \Pi_K(\overline{V})$ with $d_1(\Gamma/\mathcal{V}\,, \overline{\Gamma}/\overline{\mathcal{V}}) \leq \varepsilon$;*
2. *for each $\overline{\mathcal{V}} \in \Pi_K(\overline{V})$, there is a $\mathcal{V} \in \Pi_K(V)$ with $d_1(\Gamma/\mathcal{V}\,, \overline{\Gamma}/\overline{\mathcal{V}}) \leq \varepsilon$.*

We now deduce Theorem 6 from Lemma 7.

Proof of Theorem 6. Fix $\varepsilon > 0$ and an input graph $\Gamma \in \mathcal{G}(V)$. Let q be as in Lemma 7 with input K and ε/c. We will show that if $\overline{\Gamma} = \mathbb{G}(q, \Gamma)$, then $z(\Gamma) = z(\overline{\Gamma}) \pm \varepsilon$ with probability at least $2/3$.

Let $\mathcal{V} \in \Pi_K(V)$ be an equipartition of Γ such that $z(\Gamma) = z^*(\Gamma/\mathcal{V})$. By Lemma 7, with probability at least $2/3$, there is a partition $\overline{\mathcal{V}}$ of $\overline{\Gamma}$ such that $d_1(\Gamma/\mathcal{V}\,, \overline{\Gamma}/\overline{\mathcal{V}}) < \varepsilon/c$. By the second condition on z^* in the statement of Theorem 6, we have $|z^*(\overline{\Gamma}/\overline{\mathcal{V}}) - z^*(\Gamma/\mathcal{V})| \leq \varepsilon$, and therefore $z(\overline{\Gamma}) \leq z^*(\overline{\Gamma}/\overline{\mathcal{V}}) \leq z^*(\Gamma/\mathcal{V}) + \varepsilon = z(\Gamma) + \varepsilon$.

A symmetric argument shows that $z(\Gamma) \leq z(\overline{\Gamma}) + \varepsilon$. ◀

In Section 3 we show how to express the parameters we are interested in, namely, $d_1(\Gamma, \mathit{Forb}(\mathcal{F}))$ and $|\mathit{Forb}(\Gamma, \mathcal{F})|$, as solutions of suitable optimization problems over the set Γ/Π_K of cluster graphs of Γ.

2.2 Recovering partitions

The *distance* between two weighted graphs $R, R' \in \mathcal{G}^*(V)$ on the same vertex set V is given by

$$d_1(R, R') = \frac{1}{|V|^2} \sum_{(i,j)\in V\times V} |R(i,j) - R'(i,j)|.$$

Let $\mathcal{H} \subseteq \mathcal{G}^*$ be a property of weighted graphs, i.e., a subset of weighted graphs which is closed under isomorphisms. We define

$$d_1(R, \mathcal{H}) = \min_{\substack{R'\in\mathcal{H}:\\ V(R')=V(R)}} d_1(R, R').$$

We assume that $\mathcal{H}$ contains weighted graphs with vertex sets of all possible sizes.

We are interested in the property of graphs that are free of *copies* of members of a (possibly infinite) family $\mathcal{F}$ of graphs. To relate this property to a property of cluster graphs, we introduce some preliminary definitions. Let $\varphi\colon V(F) \to V(R)$ be a mapping from the set of vertices of a graph $F \in \mathcal{G}$ to the set of vertices of a weighted graph $R \in \mathcal{G}^*$. The *homomorphism weight* $hom_\varphi(F, R)$ of φ is defined as

$$hom_\varphi(F, R) = \prod_{(i,j)\in E(F)} R(\varphi(i), \varphi(j)).$$

The *homomorphism density* $t(F, R)$ of $F \in \mathcal{G}$ in $R \in \mathcal{G}^*$ is defined as the average homomorphism weight of a mapping in $\Phi := \{\varphi\colon V(F) \to V(R)\}$, that is,

$$t(F, R) = \frac{1}{|\Phi|} \sum_{\varphi\in\Phi} hom_\varphi(F, R).$$

Note that, if F and R are graphs, then $t(F, R)$ is roughly the subgraph density of F in R (and converges to this quantity when the size of R tends to infinity). Since weighted graphs will represent cluster graphs associated with a partition of the vertex set of the input graph, it will be convenient to work with the following property of weighted graphs:

$$\mathit{Forb}^*_{hom}(\mathcal{F}) = \{R \in \mathcal{G}^* : t(F, R) = 0 \text{ for every } F \in \mathcal{F}\}.$$

Let $R, S \in \mathcal{G}^*(V)$ be weighted graphs on the same set V of vertices. We say that S is a *subgraph* of R, which will be denoted by $S \leq R$, if $S(i,j) \leq R(i,j)$ for every $(i,j) \in V \times V$. Moreover, for a subset $Q \subseteq V$, let $R[Q]$ denote the induced weighted subgraph of R with vertex set Q. We also define $\mathit{Forb}^*_{hom}(R, \mathcal{F}) = \{S \in \mathit{Forb}^*_{hom}(\mathcal{F}) : S \leq R\}$.

The following result shows that having a cluster graph in $\mathit{Forb}^*_{hom}(\mathcal{F})$ witnesses pertinence in $\mathit{Forb}(\mathcal{F})$.

▶ **Proposition 8.** *Let $\mathcal{F}$ be a family of graphs and let $\mathcal{V}$ be an equipartition of a graph G. If $G/\mathcal{V} \in \mathit{Forb}^*_{hom}(\mathcal{F})$, then $G \in \mathit{Forb}(\mathcal{F})$.*

Proof. Let $\mathcal{V} = \{V_i\}_{i=1}^k$ be an equipartition of G and let $R = G/\mathcal{V}$. Fix an arbitrary element $F \in \mathcal{F}$ and an arbitrary injective mapping $\varphi\colon V(F) \hookrightarrow V(G)$. Define the function $\psi\colon V(F) \to V(R)$ by $\psi(v) = i$ if $\varphi(v) \in V_i$. Now, if $t(F,R) = 0$, there must be some edge $(u,w) \in E(F)$ such that $R(\psi(u),\psi(w)) = 0$, which implies that $G(\varphi(u),\varphi(v)) = 0$. Hence, $hom_\varphi(F,G) = 0$. Since φ and F were taken arbitrarily, we must have $G \in Forb(\mathcal{F})$. ◀

It is easy to see that the converse of Proposition 8 does not hold in general. Indeed, there exist graph families $\mathcal{F}$ and graphs $G \in Forb(\mathcal{F})$ such that $G/\mathcal{V}$ is actually very *far* from being in $Forb^*_{hom}(\mathcal{F})$ for some equipartition $\mathcal{V}$ of G. For one such example, let G be the n-vertex bipartite Turán graph $T_2(n)$ for K_3 with partition $V(G) = A \cup B$ and consider $\mathcal{V} = \{V_i\}_{i=1}^t$ with $V_i = A_i \cup B_i$, $i = 1,\dots,t$, where $\{A_i\}_{i=1}^t$ and $\{B_i\}_{i=1}^t$ are equipartitions of A and B respectively. Then $G/\mathcal{V}$ is a complete graph with weight $1/2$ on every edge, so that it is $1/4$-far from being in $Forb^*_{hom}(\{K_3\})$ by Turán's Theorem. More generally, if $\mathcal{V}$ is a random equitable partition of a triangle-free graph $G \in Forb(\{K_3\})$ with large edge density, then with high probability the cluster graph $G/\mathcal{V}$ is still $1/4$-far from being in $Forb^*_{hom}(\{K_3\})$.

On the other hand, we will prove that there exist partitions for graphs in $Forb(\mathcal{F})$ with respect to which an approximate version of the converse of Proposition 8 does hold, that is, we will prove that every graph in $Forb(\mathcal{F})$ is not too far from having a partition of bounded size that witnesses pertinence in $Forb(\mathcal{F})$. We say that such a partition is *recovering* with respect to $Forb(\mathcal{F})$. In what follows, we define recovering partitions formally and in a more general setting.

For every weighted graph $S \in \mathcal{G}^*$, let $\mathcal{G}_S \subseteq \mathcal{G}$ be the graph property of being *reducible* to S, that is,

$$\mathcal{G}_S = \{G \in \mathcal{G} : S = G/\mathcal{V} \text{ for some equipartition } \mathcal{V} \text{ of } G\}.$$

Moreover, let $\mathcal{P}^*$ be the weighted graph property consisting of all cluster graphs that witness pertinence in $\mathcal{P}$, i.e., $\mathcal{P}^* = \{S \in \mathcal{G}^*\colon \emptyset \neq \mathcal{G}_S \subseteq \mathcal{P}\}$. The following observation motivates this definition: if $S \in \mathcal{P}^*$, then verifying that $G \in \mathcal{G}_S$ is a way of determining that $G \in \mathcal{P}$. As a consequence, if we could find a size $K = K(\mathcal{P})$ such that every $G \in \mathcal{P}$ has an equipartition $\mathcal{V}$ of size at most K such that $G/\mathcal{V} \in \mathcal{P}^*$, then we would be able to decide whether $G \in \mathcal{P}$ by simply testing whether it is reducible to some $S \in \mathcal{P}^*$ of order at most K. Also note that, in the case of monotone properties $\mathcal{P} = Forb(\mathcal{F})$, we have $\mathcal{P}^* = Forb^*_{hom}(\mathcal{F})$.

▶ **Definition 9.** An equipartition $\mathcal{V}$ of a graph $G \in \mathcal{P}$ is *ε-recovering* for $\mathcal{P}$ if

$$d_1(G/\mathcal{V}\,, \mathcal{P}^*) \leq \varepsilon.$$

For monotone properties, this means that an equipartition $\mathcal{V}$ of a graph $G \in Forb(\mathcal{F})$ is ε-recovering for $Forb(\mathcal{F})$ if $d_1(G/\mathcal{V}\,, Forb^*_{hom}(\mathcal{F})) \leq \varepsilon$, which is the approximate converse of Proposition 8 mentioned above. With this, we say that a graph property $\mathcal{P}$ is recoverable if, for every $\varepsilon > 0$, large graphs satisfying $\mathcal{P}$ admit a constant size ε-recovering partition for $\mathcal{P}$.

▶ **Definition 10.** Let $\mathcal{P}$ be a graph property. For a fixed function $f\colon (0,1] \to \mathbb{R}$, we say that the class $\mathcal{P}$ is *f-recoverable* if, for every $\varepsilon > 0$, there exists $n_0 = n_0(\varepsilon)$ such that the following holds. For every graph $G \in \mathcal{P}$ on $n \geq n_0$ vertices, there is an equipartition $\mathcal{V}$ of G of size $|\mathcal{V}| \leq f(\varepsilon)$ which is ε-recovering for $\mathcal{P}$.

As a simple example, one can verify that the graph property $\mathcal{P}$ of being r-colorable is f-recoverable for $f(\varepsilon) = r/\varepsilon$; here and in what follows, for simplicity, we ignore divisibility

conditions and drop floor and ceiling signs. Let G be a graph in $\mathcal{P}$, with color classes $C_1, \ldots, C_r$. Let $k = r/\varepsilon$. Start by fixing parts $V_1, \ldots, V_t$ of size n/k each, with each V_i contained in some C_j ($j = j(i)$), and leaving out fewer than n/k vertices from each C_j ($1 \leq j \leq r$). The sets V_i ($1 \leq i \leq t$) cover a subset C'_j of C_j and $X_j = C_j \setminus C'_j$ is left over. We then complete the partition by taking arbitrary parts $U_1, \ldots, U_{k-t}$ of size n/k each, forming a partition of $\bigcup_{1 \leq j \leq r} X_j$. The cluster graph $G/\mathcal{V}$ can be made r-partite by giving weight zero to every edge incident to vertices corresponding to $U_1, \ldots, U_{k-t}$. Therefore $G/\mathcal{V}$ is at distance at most $r/k \leq \varepsilon$ from being r-partite. But since every r-partite weighted graph S clearly satisfies $\mathcal{G}_S \subseteq \mathcal{P}$, we get that $d_1(G/\mathcal{V}, \mathcal{P}^*) \leq \varepsilon$, as required.

Another interesting easy example is the property of *tournaments* that are *transitive*. A tournament — i.e., a complete graph whose edges are given an orientation — is said to be transitive if it does not contain any cycle or, equivalently, if there is a linear ordering $v_1, \ldots, v_n$ of its vertices such that (v_i, v_j) is an arc for every $i < j$. Computing the distance of a tournament T from being transitive, also called the *Slater index* of T, is an interesting problem which has received some attention in the past (see [10] for a survey) and has applications in many areas like psychometrics and voting theory (cf. [7]). Tournaments do not fit exactly into the framework presented here; it would be necessary to make some minor generalizations. However it is easy to see that, given a tournament T with a linear ordering $v_1, \ldots, v_n$, any equipartition $\mathcal{V} = \{V_i\}_{i=1}^k$ respecting this order (i.e., such that for every $1 \leq i < j \leq k$ and $u \in V_i, v \in V_j$ it holds that (u, v) is an arc) is such that $T/\mathcal{V}$ is a transitive directed graph with a loop on every vertex and, therefore, at distance at most $1/k$ from being a transitive tournament. We conclude that the property of being transitive is f-recoverable, with $f(\varepsilon) = 1/\varepsilon$. By Theorem 1, this is sufficient to show that one can estimate the distance of a tournament from being transitive with sample complexity that is *only exponential* in the error parameter ε. We shall elaborate on this in the full version of this paper.

We end this section by noting that the definition of f-recoverable properties has some similarity with the notion of *regular-reducible* properties $\mathcal{P}$ defined by Alon, Fischer, Newman and Shapira [3]. The main difference is that the notion of being regular-reducible requires that every graph $G \in \mathcal{P}$ should have a *regular* partition such that $G/\mathcal{V}$ is close to some property $\mathcal{R}^*$ of weighted graphs, while the definition of f-recoverable properties does not require the partitions to be regular. Another difference is that $\mathcal{R}^*$ must be such that having a (regular) cluster graph in $\mathcal{R}^*$ witnesses only *proximity* (and not pertinence) to $\mathcal{P}$.

2.3 Monotone graph properties are recoverable

Szemerédi's Regularity Lemma [24] can be used to show that every monotone (and actually every hereditary) graph property is f-recoverable, for $f(\varepsilon) = \texttt{TOWER}(\texttt{poly}(1/\varepsilon))$. In the remainder of this section, we prove that monotone properties $\mathcal{P} = Forb(\mathcal{F})$ are recoverable using a weaker version of regularity along with the Removal Lemma, which leads to an improvement on the growth of f for families $\mathcal{F}$ where the Removal Lemma is known to hold with better bounds than the Regularity Lemma.

The Removal Lemma was first stated explicitly in the literature by Alon *et al.* [1] and by Füredi [17]. The following version, which holds for possibly infinite families of graphs was first proven in [5].

▶ **Lemma 11** (Removal Lemma). *For every $\varepsilon > 0$ and every (possibly infinite) family $\mathcal{F}$ of graphs, there exist $M = M(\varepsilon, \mathcal{F})$, $\delta = \delta(\varepsilon, \mathcal{F}) > 0$ and $n_0 = n_0(\varepsilon, \mathcal{F})$ such that the following holds. If a graph G on $n \geq n_0$ vertices satisfies $d_1(G, Forb(\mathcal{F})) \geq \varepsilon$, then there is $F \in \mathcal{F}$ with $|F| \leq M$ such that $t(F, G) \geq \delta$.*

We derive, from Lemma 11, a slightly stronger version of the Removal Lemma, that deals with weighted graphs and homomorphic copies.

▶ **Lemma 12.** *For every $\varepsilon > 0$ and every (possibly infinite) family $\mathcal{F}$ of graphs, there exist $\delta = \delta(\varepsilon, \mathcal{F})$, $M = M(\varepsilon, \mathcal{F})$ and $n_0 = n_0(\varepsilon, \mathcal{F})$ such that the following holds. If a weighted graph R such that $|V(R)| > n_0$ satisfies $d_1(R, Forb^*_{hom}(\mathcal{F})) \geq \varepsilon$, then there is a graph $F \in \mathcal{F}$ with $|F| \leq M$ such that $t(F, R) \geq \delta$.*

Next, to introduce the version of regularity that we use in this work, we use a second well-known distance between weighted graphs. Let $R_1, R_2 \in \mathcal{G}^*(V)$ be weighted graphs with $|V| = n$. The *cut-distance* between R_1 and R_2 is defined as

$$d_\square(R_1, R_2) = \frac{1}{n^2} \max_{S,T \subseteq V} |e_{R_1}(S,T) - e_{R_2}(S,T)|.$$

Let $\Gamma \in \mathcal{G}(V)$ and $\mathcal{V} = \{V_i\}_{i=1}^k$ be a partition of V. We define the weighted graph $\Gamma_\mathcal{V} \in \mathcal{G}^*(V)$ as the weighted graph such that $\Gamma_\mathcal{V}(u,v) = \Gamma/_\mathcal{V}(i,j)$ if $u \in V_i$ and $v \in V_j$. Graph regularity lemmas ensure that, for any large graph Γ, there exists an equitable partition $\mathcal{V}$ of constant size such that $\Gamma_\mathcal{V}$ is a faithful approximation of Γ. Here, we use the regularity introduced by Frieze and Kannan [16].

▶ **Definition 13.** A partition $\mathcal{V} = \{V_i\}_{i=1}^k$ of a graph Γ is *γ-FK-regular* if $d_\square(\Gamma, \Gamma_\mathcal{V}) \leq \gamma$, or, equivalently if for all $S, T \subseteq V(\Gamma)$ it holds that

$$e(S,T) = \sum_{(i,j) \in [k] \times [k]} |S \cap V_i||T \cap V_j| \, \Gamma/_\mathcal{V}(i,j) \pm \gamma |V(\Gamma)|^2.$$

▶ **Lemma 14** (Frieze-Kannan Regularity Lemma)**.** *For every $\gamma > 0$ and every $t_0 > 0$, there is $T = t_0 \cdot 2^{poly(1/\gamma)}$ such that every graph Γ on $n \geq T$ vertices admits a γ-FK-regular equipartition into t classes, where $t_0 \leq t \leq T$.*

Conlon and Fox [11] found instances where the number t of classes in any γ-FK-regular equipartition is at least $t \geq 2^{1/(2^{60}\gamma^2)}$ (for a previous result, see Lovász and Szegedy [22]).

We will also need the following result, which states that a graph has homomorphism densities close to the ones of the cluster graphs with respect to FK-regular partitions.

▶ **Lemma 15** ([9, Lemma 2.7(a)])**.** *Let $\mathcal{V}$ be a γ-FK-equipartition of a graph $G \in \mathcal{G}$. Then, for any graph $F \in \mathcal{G}$ it holds that $t(F,G) = t(F, G_\mathcal{V}) \pm 4e(F)\gamma = t(F, G/_\mathcal{V}) \pm 4e(F)\gamma$.*

We are now ready to show that every monotone graph property is f-recoverable.

▶ **Theorem 16.** *For every family $\mathcal{F}$ of graphs, the property $Forb(\mathcal{F})$ is f-recoverable for $f(\varepsilon) = n_0 2^{poly(1/\delta, M)}$, where δ, M and n_0 are as in Lemma 12 with input $\mathcal{F}$ and ε.*

Proof. Let δ, M and n_0 be as in Lemma 12 with input $\mathcal{F}$ and ε and let $\gamma = \delta/(3M)^2$. By Lemma 14, it suffices to show that any γ-FK-regular partition $\mathcal{V} = \{V_i\}_{i=1}^k$ of a graph $G \in Forb(\mathcal{F})$ into $k \geq n_0$ classes is ε-recovering.

Let $R = G/_\mathcal{V}$ and suppose by contradiction that $d_1(R, Forb^*_{hom}(\mathcal{F})) \geq \varepsilon$. Then, by Lemma 12, we have $t(F,R) \geq \delta$ for some graph $F \in \mathcal{F}$ such that $|F| \leq M$. By Lemma 15, we would have $t(F,G) \geq \delta - 2\gamma M^2 > 0$, a contradiction to $G \in Forb(\mathcal{F})$. ◀

3 Estimation of $d_1(\Gamma, \mathcal{F})$ and $|Forb(\Gamma, \mathcal{F})|$

The objective of this section is to prove Theorems 1 and 3. For that, we shall use the following fact about equipartitions, whose simple proof is omitted.

▶ **Lemma 17.** *Let $\Gamma, G \in \mathcal{G}(V)$ for some vertex set V and let $\mathcal{V}$ be any equipartition of V. Then $d_1(\Gamma/\mathcal{V}\,, G/\mathcal{V}) \le d_1(\Gamma, G) + |\mathcal{V}|/|V|$.*

The final ingredient needed for Theorem 1 is the result below, which, for a recoverable property $\mathcal{P}$, relates the parameter $d_1(\cdot, \mathcal{P})$ with a parameter to which Theorem 6 may be applied.

▶ **Theorem 18.** *Let $\mathcal{P}$ be an f-recoverable graph property for some function $f\colon (0,1] \to \mathbb{R}$. Fix $\varepsilon > 0$ and let $K = f(\varepsilon/2)$. Then every graph $\Gamma \in \mathcal{G}(V)$ such that $|V| > 2K/\varepsilon$ satisfies*

$$d_1(\Gamma, \mathcal{P}) = \min_{R \in \Gamma/\Pi_K} d_1(R, \mathcal{P}^*) \pm \varepsilon.$$

Proof. Fix $0 < \varepsilon < 1$, $K = f(\varepsilon/2)$. Let $V = [n]$ and let $d = d_1(\Gamma, \mathcal{P})$ and $\widehat{d} = \min_{R \in \Gamma/\Pi_K} d_1(R, \mathcal{P}^*)$.

We first show that $\widehat{d} \le d + \varepsilon$. Let $G \in \mathcal{P}$ be a graph such that $d_1(\Gamma, G) = d$. Since $\mathcal{P}$ is f-recoverable, we can fix an $\varepsilon/2$-recovering equipartition $\mathcal{V}$ of size $1 \le k \le K$ of G, i.e., an equipartition satisfying

$$d_1(G/\mathcal{V}\,, \mathcal{P}^*) \le \frac{\varepsilon}{2}.$$

By Lemma 17 we have

$$d_1(\Gamma/\mathcal{V}\,, G/\mathcal{V}) \le d_1(\Gamma, G) + \frac{k}{n} \le d + \frac{\varepsilon}{2}.$$

Now we add the last two inequalities and apply the triangle inequality to obtain

$$d + \varepsilon \ge d_1(\Gamma/\mathcal{V}\,, \mathcal{P}^*) \ge \widehat{d}.$$

Next, we proceed to show that $d \le \widehat{d} + \varepsilon$. Let $R \in \Gamma/\Pi_K$ and $S \in \mathcal{P}^*$ be such that $d_1(R, S) = \widehat{d}$. Let $k = |V(R)|$ and fix an equipartition $\mathcal{V} = \{V_1, \dots, V_k\}$ of Γ such that $R = \Gamma/\mathcal{V}$. Consider a graph G with vertex set $V(\Gamma)$ such that $G/\mathcal{V} = S$, obtained as follows. For each $(i,j) \in [n] \times [n]$ such that $R(i,j) > S(i,j)$, we remove exactly $(R(i,j) - S(i,j))|V_i||V_j|$ edges from Γ between V_i and V_j; if $S(i,j) > R(i,j)$, we add exactly $(R(i,j) - S(i,j))|V_i||V_j|$ between V_i and V_j to Γ, thus

$$\begin{aligned}
d_1(\Gamma, G) &= \frac{1}{n^2} \sum_{(i,j) \in [k] \times [k]} |E_\Gamma(V_i, V_j) \triangle E_G(V_i, V_j)| \\
&= \frac{1}{n^2} \sum_{(i,j) \in [k] \times [k]} |S(i,j) - R(i,j)||V_i||V_j| \\
&\le \frac{1}{n^2} \sum_{(i,j) \in [k] \times [k]} |S(i,j) - R(i,j)| \frac{(n+k)}{k} \frac{(n+k)}{k} \\
&\le \widehat{d} + \frac{k}{n} + \frac{k^2}{n^2} \le \widehat{d} + \varepsilon. \qquad (\text{as } n > 2K/\varepsilon)
\end{aligned}$$

Since, by construction, G is reducible to $S \in \mathcal{P}^*$, we must have $G \in \mathcal{P}$. Hence, $d \le d_1(\Gamma, G) \le \widehat{d} + \varepsilon$. ◀

Proof of Theorem 1. Let $\mathcal{P}$ be an f-recoverable graph property. Fix $\varepsilon > 0$ and let $K = f(\varepsilon/6)$, so that by Theorem 18 we have

$$\left| d_1(\Gamma, \mathcal{P}) - \min_{R \in \Gamma/\Pi_K} d_1(R, \mathcal{P}^*) \right| \leq \frac{\varepsilon}{3}, \tag{2}$$

whenever $|V(\Gamma)| > 12K/\varepsilon$.

Let $\widehat{z}\colon \mathcal{G} \to \mathbb{R}$ be the graph parameter defined by $\widehat{z}(\Gamma) = \min_{R \in \Gamma/\Pi_K} z^*(R)$, where $z^*(R) = d_1(R, \mathcal{P}^*)$. By the triangle inequality, given R and R' in $\mathcal{G}^*(V)$, we have $z^*(R) \leq d_1(R, R') + z^*(R')$ and $z^*(R') \leq d_1(R, R') + z^*(R)$, so that $|z^*(R) - z^*(R')| \leq d_1(R, R')$. Theorem 6 applies, and $\widehat{z}$ is estimable with sample complexity $q(\varepsilon) = 2^{\mathtt{poly}(K/\varepsilon)}$. Hence, with probability at least $2/3$, a sample $\overline{\Gamma} = \mathbb{G}(q(\varepsilon/3), \Gamma)$ of Γ is such that $|\widehat{z}(\overline{\Gamma}) - \widehat{z}(\Gamma)| \leq \varepsilon/3$. By (2) we have $|\, d_1(\Gamma, \mathcal{P}) - \widehat{z}(\Gamma)| \leq \varepsilon/3$. On the other hand, we can also apply (2) to $\overline{\Gamma}$ to obtain $|\widehat{z}(\overline{\Gamma}) - d_1(\overline{\Gamma}, \mathcal{P})| \leq \varepsilon/3$. Using the triangle inequality along with the last three inequalities, we obtain $|\, d_1(\Gamma, \mathcal{P}) - d_1(\overline{\Gamma}, \mathcal{P})| \leq \varepsilon$. ◀

Proof of Corollary 2. Fox [15] proved Lemma 11 when $\mathcal{F} = \{F\}$ avoiding the Szemerédi Regularity Lemma and thus obtained better bounds on the size of $\delta > 0$ (and n_0). More specifically his result implies the following. For every fixed *finite* family $\mathcal{F}$ of graphs, Lemma 11 holds with both $1/\delta$ and n_0 bounded by $\mathtt{TOWER}(O(\log(1/\varepsilon)))$ as $M = M(\mathcal{F})$ is a constant. Hence, by Theorem 16 we have that if $\mathcal{F}$ is finite, then $\mathit{Forb}(\mathcal{F})$ is f-recoverable, where $f(\varepsilon) = \mathtt{TOWER}(\mathtt{poly}(\log(1/\varepsilon)))$. ◀

The structure of the proof of Theorem 3 is analogous to that of Theorem 1. Recall that $\mathit{Forb}^*_{hom}(R, \mathcal{F}) = \{S \leq R : t(F, S) = 0, \text{for every } F \in \mathcal{F}\}$, and set

$$\mathrm{ex}^*(R, \mathcal{F}) = \frac{1}{2|V(R)|^2} \max_{S \in \mathit{Forb}^*_{hom}(R, \mathcal{F})} e(S),$$

which measures the largest edge density of a subgraph of R not containing a copy of any $F \in \mathcal{F}$.

We shall derive Theorem 3 from the following auxiliary result, whose proof is omitted.

▶ **Theorem 19.** *Let $\mathcal{F}$ be a family of graphs such that $\mathit{Forb}(\mathcal{F})$ is f-recoverable for some function $f\colon (0,1] \to \mathbb{R}$. Then, for any $\varepsilon > 0$, there exists $K = f(\mathtt{poly}(1/\varepsilon))$ and $N = \mathtt{poly}(K)$ such that for any graph Γ of size $n \geq N$ it holds that*

$$\frac{\log_2 |\, \mathit{Forb}(\Gamma, \mathcal{F})|}{n^2} = \max_{R \in \Gamma/\Pi_K} \mathrm{ex}^*(R, \mathcal{F}) \pm \varepsilon.$$

Proof of Theorem 3. Let $\mathcal{F}$ be a family of graphs such that $\mathit{Forb}(\mathcal{F})$ is f-recoverable. Fix $\varepsilon > 0$ and let $K = f(\varepsilon/6)$, so that by Theorem 19 we have

$$\left| \frac{\log_2 |\, \mathit{Forb}(\Gamma, \mathcal{F})|}{n^2} - \max_{R \in \Gamma/\Pi_K} \mathrm{ex}^*(R, \mathcal{F}) \right| \leq \frac{\varepsilon}{3},$$

whenever $|V(\Gamma)| > N$.

Let $\widehat{z}\colon \mathcal{G} \to \mathbb{R}$ be the graph parameter defined by $\widehat{z}(\Gamma) = \max_{R \in \Gamma/\Pi_K} z^*(R)$, where $z^*(R) = \mathrm{ex}^*(R, \mathcal{F})$. We claim that, given R and R' in $\mathcal{G}^*(V)$, we have $|z^*(R) - z^*(R')| \leq d_1(R, R')$. Indeed, assume without loss of generality that $z^*(R) \geq z^*(R')$ and fix a subgraph $S \leq R$ such that $S \in \mathit{Forb}^*_{hom}(R, \mathcal{F})$ and $z^*(R) = e(S)/(2|V(R)|^2)$. If $S \in \mathit{Forb}^*_{hom}(R', \mathcal{F})$,

we are done, so assume that this is not the case. Let S' be a subgraph of S and R' maximizing $e(S')$. Clearly,

$$e(S') \geq e(S) - \frac{1}{2} \sum_{(i,j)\in V\times V} |R(i,j) - R'(i,j)| \geq e(S) - |V|^2 \, d_1(R,R'),$$

so that $0 \leq z^*(R) - z^*(R') \leq \frac{1}{2}|V|^{-2}\,(e(S) - e(S')) \leq d_1(R,R')$.

We now apply Theorem 6 to conclude that $\widehat{z}$ is estimable with sample complexity $q(\varepsilon) = 2^{\mathtt{poly}(K/\varepsilon)}$. It follows that, with probability at least $2/3$, a sample $\overline{\Gamma} = \mathbb{G}(q(\varepsilon/3), G)$ of G is such that $|\widehat{z}(\overline{\Gamma}) - \widehat{z}(\Gamma)| \leq \varepsilon/3$. By (2) we have $\left|n^{-2}\log_2 |\mathit{Forb}(\Gamma,\mathcal{F})| - \widehat{z}(\Gamma)\right| \leq \varepsilon/3$. On the other hand, we can also apply (2) to $\overline{\Gamma}$ to obtain $\left|\widehat{z}(\overline{\Gamma}) - q(\varepsilon/3)^{-2}\log_2 |\mathit{Forb}(\overline{\Gamma},\mathcal{F})|\right| \leq \varepsilon/3$. By adding the last three inequalities, we get that

$$\left|\frac{1}{n^2}\log_2 |\mathit{Forb}(\Gamma,\mathcal{F})| - \frac{1}{q(\varepsilon/3)^2}\log_2 |\mathit{Forb}(\overline{\Gamma},\mathcal{F})|\right| \leq \varepsilon,$$

as required. ◀

Corollary 4 follows directly from Theorem 3, just as Corollary 2 is a direct consequence of Theorem 1.

4 Concluding remarks

Here, we have restricted ourselves to graphs and graph properties. No substantial problems arise if one wishes to cover tournaments or directed graphs: it suffices to consider ordered graphs, that is, graphs whose vertex sets are linearly ordered, with weights on the edges, with negative weights allowed (in fact, one can consider matrices with entries in $[-1,1]$). Details are worked out in the journal version of this extended abstract.

We believe it would be interesting to investigate in more detail the notion of recoverability. For instance, when is a property $f(\varepsilon)$-recoverable for $f(\varepsilon)$ polynomial in $1/\varepsilon$?

Acknowledgements. The authors thank the referees for their helpful comments.

References

1 Noga Alon, Richard A. Duke, Hanno Lefmann, Vojtěch Rödl, and Raphael Yuster. The algorithmic aspects of the regularity lemma. *J. Algorithms*, 16(1):80–109, 1994. doi:10.1006/jagm.1994.1005.

2 Noga Alon, Eldar Fischer, Michael Krivelevich, and Mario Szegedy. Efficient testing of large graphs. *Combinatorica*, 20(4):451–476, 2000. doi:10.1007/s004930070001.

3 Noga Alon, Eldar Fischer, Ilan Newman, and Asaf Shapira. A combinatorial characterization of the testable graph properties: it's all about regularity. *SIAM J. Comput.*, 39(1):143–167, 2009. doi:10.1137/060667177.

4 Noga Alon and Asaf Shapira. A characterization of the (natural) graph properties testable with one-sided error. *SIAM J. Comput.*, 37(6):1703–1727, 2008. doi:10.1137/06064888X.

5 Noga Alon and Asaf Shapira. Every monotone graph property is testable. *SIAM J. Comput.*, 38(2):505–522, 2008. doi:10.1137/050633445.

6 Noga Alon, Asaf Shapira, and Benny Sudakov. Additive approximation for edge-deletion problems. *Ann. of Math. (2)*, 170(1):371–411, 2009. doi:10.4007/annals.2009.170.371.

7 Jean-Pierre Barthélémy and Bernard Monjardet. The median procedure in cluster analysis and social choice theory. *Math. Social Sci.*, 1(3):235–267, 1980/81. doi:10.1016/0165-4896(81)90041-X.

8 Béla Bollobás. Hereditary properties of graphs: asymptotic enumeration, global structure, and colouring. *Doc. Math.*, pages 333–342 (electronic), 1998. Extra Vol. III.

9 Christian Borgs, Jennifer T. Chayes, László Lovász, Vera T. Sós, and Katalin Vesztergombi. Convergent sequences of dense graphs. I. Subgraph frequencies, metric properties and testing. *Adv. Math.*, 219(6):1801–1851, 2008. doi:10.1016/j.aim.2008.07.008.

10 Irène Charon and Olivier Hudry. An updated survey on the linear ordering problem for weighted or unweighted tournaments. *Ann. Oper. Res.*, 175:107–158, 2010. doi:10.1007/s10479-009-0648-7.

11 David Conlon and Jacob Fox. Bounds for graph regularity and removal lemmas. *Geom. Funct. Anal.*, 22(5):1191–1256, 2012. doi:10.1007/s00039-012-0171-x.

12 Paul Erdős, Péter Frankl, and Vojtěch Rödl. The asymptotic number of graphs not containing a fixed subgraph and a problem for hypergraphs having no exponent. *Graphs and Combinatorics*, 2(1):113–121, 1986. doi:10.1007/BF01788085.

13 Paul Erdős, Daniel J. Kleitman, and Bruce L. Rothschild. Asymptotic enumeration of K_n-free graphs. In *Colloquio Internazionale sulle Teorie Combinatorie (Rome, 1973), Tomo II*, pages 19–27. Atti dei Convegni Lincei, No. 17. Accad. Naz. Lincei, Rome, 1976.

14 Eldar Fischer and Ilan Newman. Testing versus estimation of graph properties. *SIAM J. Comput.*, 37(2):482–501 (electronic), 2007. doi:10.1137/060652324.

15 Jacob Fox. A new proof of the graph removal lemma. *Ann. of Math. (2)*, 174(1):561–579, 2011. doi:10.4007/annals.2011.174.1.17.

16 Alan Frieze and Ravi Kannan. Quick approximation to matrices and applications. *Combinatorica*, 19(2):175–220, 1999. doi:10.1007/s004930050052.

17 Z. Füredi. Extremal hypergraphs and combinatorial geometry. In S. D. Chatterji, editor, *Proceedings of the International Congress of Mathematicians: August 3–11, 1994 Zürich, Switzerland*, pages 1343–1352. Birkhäuser Basel, 1995. doi:10.1007/978-3-0348-9078-6_65.

18 Oded Goldreich, editor. *Property Testing – Current Research and Surveys [outgrow of a workshop at the Institute for Computer Science ITCS) at Tsinghua University, January 2010]*, volume 6390 of *Lecture Notes in Computer Science*. Springer, 2010. doi:10.1007/978-3-642-16367-8.

19 Oded Goldreich, Shafi Goldwasser, and Dana Ron. Property testing and its connection to learning and approximation. *J. ACM*, 45(4):653–750, 1998. doi:10.1145/285055.285060.

20 Oded Goldreich and Luca Trevisan. Three theorems regarding testing graph properties. *Random Structures Algorithms*, 23(1):23–57, 2003. doi:10.1002/rsa.10078.

21 William T. Gowers. Lower bounds of tower type for Szemerédi's uniformity lemma. *Geom. Funct. Anal.*, 7(2):322–337, 1997. doi:10.1007/PL00001621.

22 László Lovász and Balázs Szegedy. Szemerédi's lemma for the analyst. *Geom. Funct. Anal.*, 17(1):252–270, 2007. doi:10.1007/s00039-007-0599-6.

23 Michal Parnas, Dana Ron, and Ronitt Rubinfeld. Tolerant property testing and distance approximation. *J. Comput. System Sci.*, 72(6):1012–1042, 2006. doi:10.1016/j.jcss.2006.03.002.

24 Endre Szemerédi. Regular partitions of graphs. In *Problèmes combinatoires et théorie des graphes (Colloq. Internat. CNRS, Univ. Orsay, Orsay, 1976)*, volume 260 of *Colloq. Internat. CNRS*, pages 399–401. CNRS, Paris, 1978.