Linear Equations, Arithmetic Progressions and Hypergraph Property Testing [*]

Noga Alon [†] Asaf Shapira [‡]

Abstract

For a fixed k-uniform hypergraph D (k-graph for short, $k \geq 3$), we say that a k-graph H satisfies property $\mathcal{P}_D$ (resp. $\mathcal{P}^*_D$) if it contains no copy (resp. induced copy) of D. Our goal in this paper is to classify the k-graphs D for which there are property-testers for testing $\mathcal{P}_D$ and $\mathcal{P}^*_D$ whose query complexity is polynomial in $1/\epsilon$. For such k-graphs we say that $\mathcal{P}_D$ (resp. $\mathcal{P}^*_D$) is *easily testable*.

For $\mathcal{P}^*_D$, we prove that aside from a single 3-graph, $\mathcal{P}^*_D$ is easily testable **if and only if** D is a single k-edge. We further show that for large k, one can use more sophisticated techniques in order to obtain better lower bounds for any large enough k-graph. These results extend and improve previous results about graphs [5] and k-graphs [18].

For $\mathcal{P}_D$, we show that for any k-partite k-graph D, $\mathcal{P}_D$ is easily testable, by giving an efficient one-sided error-property tester, which improves the one obtained by [18]. We further prove a nearly matching lower bound on the query complexity of such a property-tester. Finally, we give a sufficient condition for inferring that $\mathcal{P}_D$ is not easily testable. Though our results do not supply a complete characterization of the k-graphs for which $\mathcal{P}_D$ is easily testable, they are a natural extension of the previous results about graphs [1].

Our proofs combine results and arguments from additive number theory, linear algebra and extremal hypergraph theory. We also develop new techniques, which we believe are of independent interest. The first is a construction of a dense set of integers, which does not contain a subset that satisfies a certain set of linear equations. The second is an algebraic construction of certain extremal hypergraphs. These techniques have already been applied in two recent papers [6], [26].

1 Definitions and Background

All the hypergraphs considered here are finite and have no parallel edges. A k-uniform hypergraph (=k-graph for short) $H = (V, E)$, is a hypergraph in which each edge contains precisely k distinct vertices of V. As usual, we will call a 2-graph, a graph. Let $\mathcal{P}$ be a property of k-graphs, that is, a

[*] A preliminary version of this paper appeared in the Proc. of the 16^{th} Annual ACM-SIAM Symposium on Discrete Algorithms (SODA), ACM Press (2005), to appear.

[†] Schools of Mathematics and Computer Science, Raymond and Beverly Sackler Faculty of Exact Sciences, Tel Aviv University, Tel Aviv, Israel, and IAS, Princeton, NJ. Email: nogaa@tau.ac.il. Research supported in part by a USA-Israeli BSF grant, by the Israel Science Foundation and by the Hermann Minkowski Minerva Center for Geometry at Tel Aviv University.

[‡] School of Computer Science, Raymond and Beverly Sackler Faculty of Exact Sciences, Tel Aviv University, Tel Aviv, Israel. Email: asafico@tau.ac.il. This work forms part of the author's Ph.D. thesis. Research supported in part by a Charles Clore Foundation Fellowship and by an IBM Ph.D. Fellowship.

family of k-graphs closed under isomorphism. A k-graph H with n vertices is *ϵ-far from satisfying* $\mathcal{P}$ if one must add or delete at least ϵn^k edges in order to turn H into a k-graph satisfying $\mathcal{P}$. An *ϵ-tester*, or *property tester*, for $\mathcal{P}$ is a randomized algorithm which, given the quantity n and the ability to make queries whether a desired set of k vertices spans an edge in H, distinguishes with high probability (say, 2/3) between the case of H satisfying $\mathcal{P}$ and the case of H being ϵ-far from satisfying $\mathcal{P}$. Such an ϵ-tester is a *one-sided* ϵ-tester if when H satisfies $\mathcal{P}$ the ϵ-tester determines that this is the case (with probability 1). The ϵ-tester is a *two-sided* ϵ-tester if it may determine that H does not satisfy $\mathcal{P}$ even if H satisfies it. The property $\mathcal{P}$ is called *strongly-testable*, if for every fixed $\epsilon > 0$ there exists a one-sided ϵ-tester for $\mathcal{P}$ whose total number of queries is bounded only by a function of ϵ, which is independent of the size of the input graph. This means that the running time of the algorithm is also bounded by a function of ϵ only, and is independent of the input size. In this paper we measure query-complexity by the number of vertices sampled, assuming we always examine all edges spanned by them. For a fixed k-graph D, let $\mathcal{P}^*_D$ denote the property of being induced D-free. Therefore, H satisfies $\mathcal{P}^*_D$ if and only if it contains no induced sub-hypergraph isomorphic to D. We define $\mathcal{P}_D$ to be the property of being (not necessarily induced) D-free. Therefore, H satisfies $\mathcal{P}_D$ if and only if it contains no copy of D.

The general notion of property testing was first formulated by Rubinfeld and Sudan [25], who were motivated mainly by its connection to the study of program checking. The study of the notion of testability for combinatorial objects, and mainly for labelled graphs, was introduced by Goldreich, Goldwasser and Ron [14]. See [11], [13] and [24] for surveys and additional references on the topic.

2 The Main Results

2.1 Previous results

In [3] it is shown that every first order graph property without a quantifier alternation of type "$\forall\exists$" has ϵ-testers whose query complexity is independent of the size of the input graph. It follows from the main result of [3] that for every fixed graph D, the property $\mathcal{P}^*_D$ is strongly testable. Although the query complexity is independent of n, it has a huge dependency on $1/\epsilon$ (the fourth function in the Ackerman Hierarchy, which is a tower of towers of exponents of height polynomial in $1/\epsilon$). In [2] it was shown, using Szemerédi's Regularity Lemma, that for every fixed graph D, the property $\mathcal{P}_D$ is also strongly testable. This result was generalized to the case of directed graphs (=digraphs) in [4], by first proving a directed version of the regularity lemma. In the above two cases the query complexity is also huge, a tower of 2's of height polynomial in $1/\epsilon$.

As in many cases, moving from graphs to hypergraphs has many unexpected difficulties. While for graphs, the strong testability of $\mathcal{P}_D$ and $\mathcal{P}^*_D$ follows quite easily from an appropriate regularity lemma [3], [27], until very recently there was no strong enough regularity lemma suitable for proving that $\mathcal{P}_D$ and $\mathcal{P}^*_D$ are strongly testable for any hypergraph D. The only results for k-graphs were obtained by Frankl and Rödl [12], who (implicitly) showed that for any 3-graph D, $\mathcal{P}_D$ is strongly testable (see also [20]), and by Kohayakawa, Nagle and Rödl in [18], where it was shown that for any 3-graph D, $\mathcal{P}^*_D$ is strongly testable. Recent works of Gowers [17] and independently of Nagle, Rödl, Schacht and Skokan [23], [21], suggest that a powerful new hypergraph regularity lemma implies that $\mathcal{P}^*_D$ and $\mathcal{P}_D$ are both strongly testable for any k-graph D, for arbitrary value of k.

For some k-graphs, however, there are obviously much more efficient property testers than the ones guaranteed by the general results. For example, for any k, if D is a single k-edge, then there is obviously a one-sided error property tester for $\mathcal{P}_D = \mathcal{P}^*_D$, whose query complexity is $\Theta(1/\epsilon)$. We simply sample $\Theta(1/\epsilon)$ vertices, and check if they span an edge. A natural question is therefore, to decide for which k-graphs D, there is a one-sided error property tester for $\mathcal{P}_D$ or $\mathcal{P}^*_D$, whose query complexity is bounded by a *polynomial* of $1/\epsilon$. We call a property $\mathcal{P}$ *easily testable* if there is a one-sided error property tester for $\mathcal{P}$ whose query complexity is polynomial in $1/\epsilon$. If no such property tester exists, we say that $\mathcal{P}$ is *hard to test.*

In [1] it is shown that for an undirected graph D, property $\mathcal{P}_D$ is easily testable if and only if D is bipartite. The authors of [4] obtain a precise characterization of all the directed graphs D for which $\mathcal{P}_D$ is easily testable. In [5] it is shown that for any graph D other than the paths of length 1,2,3 (which have 2,3,4 vertices respectively) the cycle of length 4, and their complements, $\mathcal{P}^*_D$ is not easily testable. A similar result was also proved for directed graphs. For $k > 2$, the only result in the direction of classifying the k-graphs for which $\mathcal{P}_D$ and $\mathcal{P}^*_D$ are easily testable was obtained in [18], where it was shown that for any k, the complete k-graph on $k+1$ vertices is not easily testable. A natural step is therefore to classify all the k-graphs D for which $\mathcal{P}^*_D$ and $\mathcal{P}_D$ are easily testable.

2.2 The new results

Our first two results concern testing $\mathcal{P}^*_D$. In what follows we denote by $D_{3,2}$ the unique 3-graph on 4-vertices that has 2 edges.

Theorem 1 *For any $k \geq 3$ and any k-graph D other than a single k-edge and $D_{3,2}$, there exists a constant $c = c(D) > 0$ such that the query-complexity of any one-sided error ϵ-tester for $\mathcal{P}^*_D$ is at least*

$$\left(\frac{1}{\epsilon}\right)^{c\log(1/\epsilon)}.$$

As noted above, for any k, there is an obvious one-sided error property tester for the case of D being a single k-edge, whose query complexity is $\Theta(1/\epsilon)$. We therefore get that Theorem 1 gives a complete characterization of the k-graphs D, for which $\mathcal{P}^*_D$ is easily testable, besides the case of $D_{3,2}$.

Our second result states that for large k, we can significantly improve the lower bounds for testing $\mathcal{P}^*_D$ for almost all k-graphs.

Theorem 2 *For any k there is a constant $r(k)$, such that for any k-graph D on at least $r(k)$ vertices, there is a constant $c = c(D) > 0$, such that any one-sided error property tester for testing $\mathcal{P}^*_D$ has query complexity at least*

$$\left(\frac{1}{\epsilon}\right)^{c(\log 1/\epsilon)^{\lfloor \log k \rfloor}}.$$

In fact, the lower bounds in the above theorem apply also to some k-graphs on less than $r(k)$ vertices, amongst them all the k-graphs that contain F^k, which is the complete k-graph on $k+1$ vertices. As a special case, we thus improve the lower bound for the case of F^k obtained in [18], which was similar to the lower bound in Theorem 1. Moreover, our technique supplies a slightly inferior lower bound (namely, with exponent $\lfloor \log\lceil k/2+1\rceil \rfloor$ instead of $\lfloor \log k \rfloor$) for **any** k-graph D on more than k vertices (see discussion following the proof of Theorem 2 in Subsection 6.2). Note, that the bounds of Theorem 2 are *super-polynomial* in the bounds of Theorem 1, thus for large k we obtain substantially better lower bounds.

Our next two results concern testing $\mathcal{P}_D$. We first give an efficient one-sided error property tester for any k-partite k-graph. Recall, that a k-graph is k-partite if its vertex set can be partitioned into k sets, such that each edge has precisely one vertex in each of the partition classes.

Theorem 3 *(i) Let $t_1 \leq \ldots \leq t_k$, put $t^* = t_1 \cdot \ldots \cdot t_k$ and let D be any k-partite k-graph with partition classes of sizes $t_1, \ldots, t_k$. Then, there is a one sided-error ϵ-tester for $\mathcal{P}_D$ with query complexity*

$$O\left(\frac{1}{\epsilon}\right)^{t^*/t_k}.$$

(ii) For any k-partite k-graph D on d vertices, which contains $|E|$ edges, the query complexity of any one-sided error ϵ-tester for $\mathcal{P}_D$ is

$$\Omega\left(\frac{1}{\epsilon}\right)^{|E|/d}.$$

The upper bound in the above theorem improves the one obtained by [18] in which the exponent was t^*. See Section 8 for more details. Observe, that when D is the complete k-partite k-graph $K_{t,\ldots,t}$, the exponent in the upper bound is t^{k-1} while the one in the lower bound is t^{k-1}/k, which are very close. The proof of this theorem appears in Section 8.

For the next result we need some definitions. A homomorphism from a k-graph D to a k-graph K is a mapping $\varphi : V(D) \mapsto V(K)$, which maps edges to edges, namely, if $(v_1, \ldots, v_k) \in E(D)$ then $(\varphi(v_1), \ldots, \varphi(v_k)) \in E(K)$.

Definition 2.1 (Core) *The core of a k-graph D, is the smallest (in terms of edges)* **subgraph** *of D, K, for which there exists a homomorphism from D to K. A k-graph D is called a core if it is the core of itself.*

We also need to define a generalization of cycles in graphs;

Definition 2.2 (Hyper-Cycle) *A k-graph on d vertices $1, \ldots, d$ is called a hyper-cycle if it contains $d - k + 2$ edges $e_1, \ldots, e_{d-k+2}$ and one can arrange its vertices on a cycle such that every edge e_i contains the vertices $\{i \pmod{d}, \ldots, i + k - 1 \pmod{d}\}$.*

Observe, that for $k = 2$ the above definition boils down to the definition of a cycle. Also, a single k-edge is not a hyper-cycle, as it contains $1 < k - k + 2 = 2$ edges. The next theorem gives a sufficient condition for inferring that for a k-graph D, property $\mathcal{P}_D$ is not easily testable.

Theorem 4 *If the core of a k-graph D spans a hyper-cycle (not necessarily as an induced subgraph), then there exists a constant $c = c(D) > 0$ such that the query-complexity of any one-sided error ϵ-tester for $\mathcal{P}_D$ is at least*

$$\left(\frac{1}{\epsilon}\right)^{c\log(1/\epsilon)}.$$

Observe that the core of any k-partite k-graph is a single edge, which does not satisfy the definition of a hyper-cycle. It is important to note that though Theorem 4 establishes that for a large family of non k-partite k-graphs D, property $\mathcal{P}_D$ is not easily testable, it does not cover all the non k-partite k-graphs, as the core of some of them does not contain a hyper-cycle. However, for $k = 2$, Theorem 4 does cover all the non-bipartite graphs, as it is easy to see that the core of any non-bipartite graph must contain a cycle, namely, one of the shortest odd cycles of the graph. As we have mentioned above, for $k = 2$, this is precisely the definition of a hyper-cycle. Hence, Theorems 3 and 4 imply that for $k = 2$, property $\mathcal{P}_D$ is easily testable if and only if D is bipartite, thus extending the result of [1], where the characterization for graphs was first obtained. We finally mention that using the main ideas of the proof of Theorem 4 one can slightly extend it by showing that it holds even if in the definition of a hyper-cycle one only requires that the first two vertices of e_i would be $i \pmod{d}, i+1 \pmod{d}$. As the proof with this definition is more involved (mainly due to cumbersome notations), and still does not cover all the cases of non k-partite k-graphs, we preferred to give the proof of the slightly less general case, which contains all the important ideas.

We can finally show that the lower bounds of Theorems 1, 2 and 4 can also be extended to the cases of two-sided error ϵ-testers.

Theorem 5 *The lower bounds of Theorems 1, 2 and 4 hold for two-sided error ϵ-testers as well.*

2.3 Techniques

Our main results in this paper, Theorems 1, 2 and 4, are based on two novel constructions. All the previous results on testing $\mathcal{P}_D$ and $\mathcal{P}^*_D$ ([1],[4],[5],[18]) were based on constructions of sets of integers, which do not contain small subsets that satisfy a certain *single* equation. All these constructions were based on Behrebd's construction [7] of a large set of integers containing no 3-term arithmetic progression. In our case, however, we consider sets of integers that do not contain small subsets that satisfy a certain *set* of equations. The key benefit of this consideration is that requiring the set of integers to satisfy a set of equations, rather than a single one, allows us to construct much denser sets than the ones used in previous papers. This benefit translates to significantly improved lower bounds. The proof of this new construction appears in Section 3. Some of the techniques we apply in the proof of this result are motivated by the work of Laba and Lacey in [19], where they reproved a result of Rankin [22] by constructing large sets of integers without k-term arithmetic progressions. The ideas used in our number theoretic construction have been further applied in another recent paper [26].

Our second technical contribution is an algebraic construction of certain extremal k-graphs. The goal of this construction is to resolve the main technical difficulty in the proof of our main results. The main benefit of this construction is that it allows us to infer certain linear equations between the integers that are used in the definition of these k-graph. In previous papers about testing subgraphs in graphs, ([1],[4],[5]) inferring these linear equations was trivial. This construction can be viewed as an extension of a construction of Frankl and Rödl [12], but ours is far more complicated to analyze. It is also much more applicable than the construction of [12], which, for example, can only be used to show that the complete k-graph on $k+1$ vertices is not easily testable and with a lower bound as in Theorem 1, rather than the one in Theorem 2. Our new algebraic technique is applied in Sections 4 and 7. The ideas used in the algebraic construction of extremal k-graphs have been further applied in another recent paper [6].

2.4 Organization

In Sections 3 and 4 we develop the main machinery needed to prove Theorems 1 and 2. In Section 3 we describe a new number theoretic construction. In Section 4 we describe a new algebraic construction of extremal k-graphs. In Section 5 we prove two useful lemmas, which use the constructions of Sections 3 and 4 in order to obtain the lower bounds of Theorems 1 and 2. The results of Sections 3, 4 and 5 are essentially independent, and thus these sections can be read independently. To further simplify the reading of these sections, each of them starts with a short subsection in which we state the important definitions and state the main results proved in that section.

The proofs of Theorems 1 and 2, which follow quite easily by combining the main results of Sections 3, 4 and 5, are given in Section 6. In Section 7 we apply our algebraic technique again, this time to construct extremal k-graphs, which are a central tool in the proof of Theorem 4. The proof of Theorem 4 also appears in Section 7. In Section 8 we prove Theorem 3. As the proof of Theorem 5 uses ideas similar to the ones used in [5] (in addition to the ideas of this paper) we omit it. Section 9 contains some concluding remarks and open problems.

Throughout this paper we assume, whenever this is needed, that the number of vertices n of the k-graph considered is sufficiently large, and that the error parameter ϵ is sufficiently small. In order to simplify the presentation, we omit all floor and ceiling signs whenever these are not crucial, and make no attempt to optimize the absolute constants. All the logarithms appearing in the paper are in base 2. When we later refer to a graph D as being easy/hard to test, we mean that $\mathcal{P}^*_D$ (or $\mathcal{P}_D$) is easy/hard to test.

3 Arithmetic Progressions and Linear Equations

3.1 The main results of this section

In this section we give our new number theoretic construction, which will be later used in Section 6. We start with some definitions.

Definition 3.1 (**(k,h)-Gadget**) *Call a set of $k-2$ linear equations $\mathcal{E} = \{e_1, \ldots, e_{k-2}\}$ with integer coefficients in k unknowns $x_1, \ldots, x_k$ a (k,h)-gadget if it satisfies the following properties:*

1. *Each of the unknowns $x_1, \ldots, x_k$ appears in at least one of the equations.*
2. *For $1 \le t \le k-2$ equation e_t is of the form*
$$p_t x_i + q_t x_j = (p_t + q_t) x_\ell,$$
where $0 < p_t, q_t \le h$ and x_i, x_j, x_ℓ are distinct.
3. *Equations $e_1 \ldots, e_{k-2}$ are linearly independent.*

We say that $z_1, \ldots, z_k$ satisfy a (k,h)-gadget $\mathcal{E}$ if they satisfy the $k-2$ equations of $\mathcal{E}$. Note, that any gadget $\mathcal{E}$ has a trivial solution $x_1 = \ldots = x_k$.

Definition 3.2 (**(k,h)-Gadget-Free**) *A set of integers Z, is called (k,h)-gadget-free if there are no k* **distinct** *integers $z_1, \ldots, z_k \in Z$ that satisfy an arbitrary (k,h)-gadget.*

Our main goal in this section is to prove the following theorem, which will be a key ingredient in the lower-bounds for $\mathcal{P}^*_D$.

Theorem 6 *For every h and k there is an integer $c = c(k, h)$, such that for every n there is a $(k+1, h)$-gadget-free subset $Z \subset [n] = \{1, 2, \ldots, n\}$ of size at least*

$$|Z| \geq \frac{n}{e^{c(\log n)^{1/\lfloor \log 2k \rfloor}}} \tag{1}$$

As we have explained before, note that for larger k the above theorem guarantees the existence of a substantially larger set Z. The special case of the above theorem, where $k = 2$, was proved and used in [5] and [9]. As the details of the proof of Theorem 6 will reveal, the main idea is to somehow reduce the construction required to prove Theorem 6 to a construction related to a notion very similar to arithmetic progressions. The main idea of this reduction will be to show that integers satisfying the linear equations of a gadget, nearly form an arithmetic progression. Our notion of "near" arithmetic progression is the following:

Definition 3.3 ((k, h)-Progression) *A set of k integers $z_1 \leq z_2 \leq \ldots \leq z_k$ is said to form a (k, h)-progression if there are integers $d, n_2, \ldots, n_k$ with $n_i \leq h$, such that for $2 \leq i \leq k$, we have*

$$z_i = z_{i-1} + n_i d \tag{2}$$

The fact that a (k, h)-progression is "nearly" an arithmetic progressions comes from the fact that in an arithmetic progression one requires $n_2 = \ldots = n_k = 1$. Also, d is analogous to the difference between consecutive integers in an arithmetic progression. In other words, a k-term arithmetic progression is a $(k, 1)$-progression of distinct elements. The proof of Theorem 6 appears in the following two subsections. In the first subsection we show how to transform the problem from one that deals with linear equations and gadgets to an analogous problem about (k, h)-progressions. We also show how the solution of the problem about (k, h)-progressions implies Theorem 6. In the second subsection we solve the problem about (k, h)-progressions.

Definition 3.4 (Trivial (k, h)-Progression) *A (k, h)-progression is said to be trivial if some of its elements are identical.*

Therefore, a (k, h)-progression as defined in Definition 3.3 is non-trivial if d is non-zero and the integers n_i are positive. We also call the integers n_i the *coefficients* of the progression and d the *difference.*

3.2 Gadgets and (k, h)-Progressions

We start this subsection by reducing gadgets to (k, h)-progressions. This is done in the next three claims.

Claim 3.1 *If $z_1 < \ldots < z_k$ are positive integers, which satisfy a (k, h)-gadget $\mathcal{E}$, then for $2 \leq i \leq k-1$ there are positive integers $a_i, b_i \leq (h^2k)^{k/2}$ such that $a_i z_{i-1} + b_i z_{i+1} = (a_i + b_i) z_i$.*

Proof: As there is nothing to prove for $k = 3$, we assume $k \geq 4$. In order to simplify the notation, we show that there are positive integers $a, b \leq (h^2k)^{k/2}$ such that

$$az_1 + bz_3 = (a + b)z_2. \tag{3}$$

The other $k-3$ cases are identical. We first substitute $z_1, \ldots, z_k$ into the set $\mathcal{E}$, and obtain $k-2$ linear equations of the form $p_t z_i + q_t z_j = (p_t + q_t) z_k$. Henceforth, when we refer to equation $e_t \in \mathcal{E}$ we will refer to the equation after we have substituted the suitable z_is into it. Our goal is simply to show that there are $\alpha_1, \ldots, \alpha_{k-2}$ not all of them equal to zero, such that in the linear combination $C = \alpha_1 e_1 + \ldots + \alpha_{k-2} e_{k-2}$ the coefficients of the integers $z_4, \ldots, z_k$ vanish. We first claim that this will give us (3). Indeed, note that as $e_1, \ldots, e_{k-2}$ are by assumption linearly independent, it cannot be the case that all the coefficients of the integers z_i vanish. Also, as for each of the equations in $\mathcal{E}$ the sum of the coefficients in the left hand side is equal to the coefficient in the right hand side, this must also hold for C, hence, it cannot be the case that precisely one of coefficients of z_1, z_2, z_3 does not vanish. Similarly, if precisely two of coefficients of z_1, z_2, z_3 do not vanish, this would imply that they are equal, which contradicts our assumption that $z_1 < \ldots < z_k$. Finally as we assume that each of the z_is appears at least once, we are guaranteed to get (3).

In order to make sure that in a linear combination with coefficients $\alpha_1, \ldots, \alpha_{k-2}$ the integers $z_4, \ldots, z_k$ vanish, we may write $k-3$ homogenous linear equation, which require that. This is a set of $k-3$ homogenous equations in $k-2$ unknowns with coefficients bounded by h. Therefore, by Lemma 4.3 (see Section 4) it has a non-zero solution with integer coefficients of size at most $(h^2(k-2))^{k/2-1}$. This means that the coefficients of C are bounded by $(k-3)(h^2(k-2))^{k/2-1} \leq (h^2 k)^{k/2}$, as needed. ■

Claim 3.2 *Suppose z_1, z_2, z_3, a, b are positive integers, such that $z_1 \leq z_2 \leq z_3$ and $a, b \leq h$. If the following equation holds*

$$az_1 + bz_3 = (a+b)z_2,$$

then z_1, z_2, z_3 form a $(3, h)$-progression.

Proof: We first assume that a and b are co-prime, as otherwise we can divide them by their gcd, and obtain a new equation $a'z_1 + b'z_3 = (a' + b')z_2$, with $a' < a, b' < b$. Rearranging the equation we get that $a(z_2 - z_1) = b(z_3 - z_2)$. As a and b are co-prime $d = (z_3 - z_2)/a = (z_2 - z_1)/b$ is an integer. Thus, we can write $z_2 = z_1 + bd$ and $z_3 = z_2 + ad$, and take $n_2 = b \leq h$ and $n_3 = a \leq h$ in the definition of the $(3, h)$-progression. ■

Claim 3.3 *Suppose $z_1 \leq z_2 \leq \ldots \leq z_k$ are positive integers, such that for every $2 \leq i \leq k-1$ there are integers $a_i, b_i \leq h$, such that*

$$a_i z_{i-1} + b_i z_{i+1} = (a_i + b_i) z_i$$

holds. Then $z_1, z_2, \ldots, z_k$ form a (k, h^{k-2})-progression.

Proof: By induction on k. The base case $k = 3$ follows from Claim 3.2. Assuming the claim holds for k we prove it for $k+1$. By the induction hypothesis, for $2 \leq i \leq k$ we can write $z_i = z_{i-1} + m_i t$ for some integer t and $m_i \leq h^{k-2}$. By assumption $a_k z_{k-1} + b_k z_{k+1} = (a_k + b_k) z_k$. Rearranging this gives

$$z_{k+1} - z_k = \frac{a_k}{b_k}(z_k - z_{k-1}). \tag{4}$$

Put $g = gcd(b_k, t)$ $(\leq h)$ and $d = t/g$, and observe that for $1 \leq i \leq k$ we can write $z_i = z_{i-1} + g \cdot m_i \cdot d$, and thus take $n_i = m_i \cdot g \leq hh^{k-2} = h^{k-1}$. As in Claim 3.2, we may assume that a_k and b_k are co-prime, and conclude from (4) that b_k divides $z_k - z_{k-1} = m_k t$. We may thus write

$$z_{k+1} = z_k + \frac{a_k m_k t}{b_k} = z_k + \frac{a_k m_k g}{b_k} \cdot d = z_k + n_{k+1} d.$$

As $a_k g/b_k \leq a_k \leq h$ and $m_k \leq h^{k-2}$, we have $n_{k+1} \leq h^{k-1}$, and the proof is complete. ■

Combining Claims 3.1 and 3.3 we immediately obtain the following corollary, which is our sought after transformation from gadgets to (k, h)-progressions.

Corollary 3.1 *For every k and h there is an integer $c = c(k, h)$ such that if $z_1, \ldots, z_k$ satisfy a (k, h)-gadget then they form a (k, c)-progression.*

Though we do not need this here, it is worth mentioning that the converse of Corollary 3.1 is also true. Indeed, if $z_1, \ldots, z_k$ form a (k, h)-progression, then for every $2 \leq i \leq k-1$ we have $z_i = z_{i-1} + n_i d$, and $z_{i+1} = z_i + n_{i+1} d$. This implies that $(n_i + n_{i+1}) z_i = n_{i+1} z_{i-1} + n_i z_{i+1}$. Hence, $z_1, \ldots, z_k$ satisfy the $k-2$ linear equations $(n_i + n_{i+1}) x_i = n_{i+1} x_{i-1} + n_i x_{i+1}$ that are easily checked to satisfy the three requirements of a (k, h)-gadget.

The proof of Theorem 6 will follow by combining Corollary 3.1 and the following lemma.

Lemma 3.1 *For every h and $p \geq 2$, there is an integer $c = c(p, h)$ such that for every n there is a subset $Z \subset [n] = \{1, 2, \ldots, n\}$ of size at least*

$$|Z| \geq \frac{n}{e^{c \log^{1/p} n}} \tag{5}$$

that does not contain any non-trivial $(1 + 2^{p-1}, h)$-progression.

Proof of Theorem 6: Let p be the largest integer satisfying $1 + 2^{p-1} \leq 1 + k$, namely, $p = \lfloor \log 2k \rfloor$. Let $c' = c(k+1, h)$ be the constant appearing in Corollary 3.1. Now, by Lemma 3.1, there is a constant $c = c(p, c')$, such that for every n there is a subset $Z \subseteq [n]$ of size as in (5), which contains no non-trivial $(1 + 2^{p-1}, c')$-progression. By our choice of p, this set contains no non-trivial $(k+1, c')$-progression. By Corollary 3.1, Z does not contain $k+1$ distinct integers, which satisfy a $(k+1, h)$-gadget. As $p = \lfloor \log 2k \rfloor$, Z satisfies the requirements of Theorem 6. ■

It is easy to see that the elements of a $(1 + 2^{p-1}, h)$-progression must be taken from an arithmetic progression of length at most $h2^{p-1}$, whose difference is the integer d from the definition of the $(1 + 2^{p-1}, h)$-progression in Definition 3.3. Thus, another way to look at Lemma 3.1 is as a construction of a set Z with the following property: not only doesn't Z contain arithmetic progressions of length $1 + 2^{p-1}$, but it does not even contain $1 + 2^{p-1}$ numbers out of some other not too large arithmetic progression, whose other elements need not even belong to Z.

In order to prove Lemma 3.1, we will first show that it holds for every *fixed* set of coefficients $n_2, \ldots, n_{1+2^{p-1}}$. Namely, we show that there is a subset of $[n]$ of the same size as in (5) that does not contain any $(1 + 2^{p-1}, h)$-progression $z_1, \ldots, z_{1+2^{p-1}}$ such that $z_i = z_{i-1} + n_i d$ for every $2 \leq i \leq 1 + 2^{p-1}$. Note, that the difference d, may be arbitrary. To be precise, we want to show the following:

Claim 3.4 *For every fixed distinct $n_2, \ldots, n_{1+2^{p-1}} \le h$ there is an integer $c = c(p,h)$ such that for every n there is a subset $Z \subset [n] = \{1, 2, \ldots, n\}$ of size at least*

$$|Z| \ge \frac{n}{e^{c \log^{1/p} n}} \tag{6}$$

that does not contain any non-trivial $(1 + 2^{p-1}, h)$-progression with coefficients $n_2, \ldots, n_{2^p-1}$.

The proof of this claim appears in the next subsection. We first show how to obtain Lemma 3.1 using the above claim.

Proof of Lemma 3.1: For every set s, of distinct 2^{p-1} integers $n_2, \ldots, n_{1+2^{p-1}} \le h$, let Z_s be the largest subset of $[n]$, which does not contain any non-trivial $(1+2^{p-1}, h)$-progression with coefficients $n_2, \ldots, n_{1+2^{p-1}}$. By Claim 3.4 we have that for any s the set Z_s has size at least $n/e^{c \log^{1/p} n}$, where c depends only on p and h. Denote the number of sets s by m, and observe that as the coefficients in each set s are bounded by h there are less than 2^{ph} choices for the set s.

Randomly and uniformly at random pick m integers $t_1, \ldots, t_m$ from $\{-n, \ldots, n\}$, and consider the set $Z = \bigcap_{i=1}^{m} (Z_i + t_i)$ (where, $Z + t$ denotes the translate of Z by t, i.e. $Z + t = \{z + t : z \in Z\}$). Clearly Z contains no $(1 + 2^{p-1}, h)$-progressions with arbitrary coefficients bounded by h. For every integer $z \in [n]$ the probability that it belongs to $Z_i + t_i$ is $1/e^{c \log^{1/p} n}$, hence the probability that it belongs to all the sets $Z_i + t_i$, and therefore also to Z, is $(1/e^{c \log^{1/p} n})^m = 1/e^{c' \log^{1/p} n}$ for a possibly larger c' that still depends only on p and h. By linearity of expectation we get that the expected size of Z is $n/e^{c' \log^{1/p} n}$, and therefore there is some choice of $t_1, \ldots, t_m$ for which the resulting set Z is at least this large. ■

3.3 Large sets of integers without a given (k, h)-Progression

In this subsection we apply the method of [19] in order to prove Claim 3.4. The proof will require some more definitions. We first need to further extend the notion of arithmetic progressions as follows: we call a set of p integers $z_0, \ldots, z_{p-1}$ a (p, t, h)-progression if there are $t + 1$ integers $d_0, \ldots, d_t$ and integers $n_0 = 0, n_1, \ldots, n_{p-1} \le h$ such that for $0 \le i \le p - 1$

$$z_i = d_0 + n_i \cdot d_1 + n_i^2 \cdot d_2 + \ldots + n_i^t \cdot d_t. \tag{7}$$

To avoid confusion, note that by definition $d_0 = z_0$, thus, we did not really need d_0 and n_0, which is fixed to be zero, in the above definition. However, this way of defining the integers of the set will make subsequent notation more compact. We call a (p, t, h)-progression *non-trivial* if at least one of $d_1, \ldots, d_t$ is non-zero and $n_0, \ldots, n_{p-1}$ are distinct. Observe, that the non-triviality condition for $(p, 1, h)$-progression is equivalent to the non-triviality condition for (p, h)-progression, namely, that they contain distinct integers. On the other hand, note that for $t > 1$ a non-trivial (p, t, h)-progression may contain identical integers. Note, that unlike the definition of (k, h)-progressions in Definition 3.3, here we define each element of the sequence with respect to the smallest number $z_0 = d_0$, rather than the preceding one as in Definition 3.3. This will further simplify the presentation.

For a set s of p integers $n_0 = 0, n_1 \ldots, n_{p-1} \le h$, define $R_s(p, t, n)$ to be the largest possible size of a subset of $[n]$, which does not contain any non trivial (p, t, h)-progression whose coefficients are the integers of s. The proof of Lemma 3.1 will follow by combining the following two claims.

Claim 3.5 *For every set s, of $2t+1$ distinct integers bounded by h, there is an integer $c = c(t,h)$, such that*

$$R_s(2t+1,t,n) \geq \frac{n}{e^{c\sqrt{\log n}}} \tag{8}$$

Claim 3.6 *For every set s, of p distinct integers bounded by h, there is an integer $c = c(p,h)$, such that if $n = g^b$ and $p \geq t+1$, then*

$$R_s(p,t,n) \geq \frac{n \cdot R_s(p,2t,g^2 b)}{c^b g^2 b} \tag{9}$$

Note that a (p,h)-progression as defined in Definition 3.3 is also a $(p,1,h(p-1))$-progression as defined in (7). Hence, we can prove Claim 3.4 by showing that for every set s, of coefficients $n_2,\ldots,n_{1+2^{p-1}} \leq h(1+2^{p-1})$ there is a set that contains no $(p,1,h(1+2^{p-1}))$-progression with given coefficients from s. As the value of h only affects the hidden constant c, we will prove Claim 3.4 for every set of coefficients bounded by h. We first show how to obtain Lemma 3.4 from the above two claims.

Proof of Lemma 3.1: Consider any set s, of distinct integers bounded by h. Given integers n and p, we prove by induction on ℓ that for every $2 \leq \ell \leq p$ there is a constant $c = c(p,h)$, such that

$$R_s(1+2^{p-1},2^{p-\ell},n) \geq \frac{n}{e^{c(\log n)^{1/\ell}}}. \tag{10}$$

The case $\ell = 2$ follows from Claim 3.5 with $t = 2^{p-2}$. Assuming the claim holds for ℓ we prove it for $\ell+1$. Set $b = (\log n)^{1/(\ell+1)}$, and let g satisfy $n = g^b$, namely $g = e^{(\log n)^{1-1/(\ell+1)}}$. A short calculation shows that in this case

$$(\log g^2 b)^{1/\ell} \leq c(\log n)^{1/(\ell+1)}, \tag{11}$$

where c depends only on p. We get that

$$R(1+2^{p-1},2^{p-\ell-1},n) \geq \frac{n \cdot R(1+2^{p-1},2^{p-\ell},g^2 b)}{c^b g^2 b} \geq$$

$$\frac{n}{c^b g^2 b} \cdot \frac{g^2 b}{e^{c(\log g^2 b)^{1/\ell}}} \geq \frac{n}{c^b e^{c(\log n)^{1/(\ell+1)}}} \geq \frac{n}{e^{c(\log n)^{1/(\ell+1)}}},$$

where the first inequality follows from Claim 3.6, the second from the induction hypothesis in (10) with $n = g^2 b$, the third from (11), and the last from our choice of b and the fact that c depends only on p and h. Also, note that by the reasonings we used to derive each of these inequalities, all the above constants depend only on p and h (we called all of them c in order to simplify the notation). This completes the proof of (10). The proof of the lemma now follows upon setting $\ell = p$ in (10). ∎

We now turn to prove Claims 3.5 and 3.6, which will require (yet again) several additional definitions. Given a set of integers S we denote by $S+r$ the *translate* of S by r, that is, $S+r = \{x+r : x \in S\}$. Note, that if S does not contain any non trivial (p,t,h)-progression than so does any translate of S. For reasons that will soon become clear, we prefer to prove Claims 3.5 and 3.6

with respect to the set of integers $\{-n/2, \ldots, n/2\}$ rather than $[n] = \{1, \ldots, n\}$. We also consider representations of integers from $\{-n/2, \ldots, n/2\}$ in base g, where g will depend on n and will be much smaller than n. If $n = g^b$ we define, for an integer $c \geq 2$,

$$Q_c = \{x \in Z : x = \sum_{i=0}^{b-1} x_i g^i, -g/c \leq x_i \leq g/c\},$$

namely, all the integers whose "digits" in base g belong to $-g/c, \ldots, g/c$. As $Q_c \subseteq \{-n/2, \ldots, n/2\}$ we may and will construct our sought after sets from integers belonging to Q_c for an appropriate large enough constant c. Note, that somewhat unconventionally, we allow for negative digits. This representation, however, is well defined in the sense that given $x \in Q_c$, there are unique integers $-g/c \leq x_0, \ldots, x_{b-1} \leq g/c$ such that $x = \sum_{i=0}^{b-1} x_i g^i$. Given an integer $x \in Q_c$ we will denote by $\overline{x} = (x_0, \ldots, x_{b-1})$ the unique b dimensional vector in Z^b such that $x = \sum_{i=0}^{b-1} x_i g^i$. We will also denote $||x||^2 = ||\overline{x}||^2 = \sum_{i=0}^{b-1} x_i^2$. Our argument will critically rely on the observation that if c is sufficiently large then addition, and more generally linear combinations with small coefficients, of numbers from Q_c is equivalent to linear combinations of their corresponding vectors. For example, observe that if $x, y, z \in Q_2$, then $x + y = z$ if and only if $\overline{x} + \overline{y} = \overline{z}$. The reason for that is simply that there is no carry in the base g addition of the number. More generally, if c is sufficiently large with respect to integers $\alpha_1, \ldots, \alpha_t$, then for $x, x_1, \ldots, x_t \in Q_c$,

$$x = \sum_{i=1}^{t} \alpha_i x_i \Longleftrightarrow \overline{x} = \sum_{i=1}^{t} \alpha_i \overline{x_i}. \tag{12}$$

Also, note that if c is sufficiently large with respect to integers $\alpha_1, \ldots, \alpha_t$, then for $x_1, \ldots, x_t \in Q_c$,

$$\overline{x} = \sum_{i=1}^{t} \alpha_i \overline{x_i} \in Q_{c'}, \tag{13}$$

for another (possibly smaller) constant c'. It should be noted that had we chosen to work with the set $[n]$ rather than $-n/2, \ldots, n/2$ and represented integers using positive digits, then (12) and (13) would only hold for positive coefficients. The reason is that the difference of two numbers with small digits may contain very large digits. As we also allow for negative digits, the difference also contains small digits. Finally, given integers $p_1, \ldots, p_t$ we denote by $V(p_1, \ldots, p_t)$ the Vandermonde matrix satisfying for $1 \leq i, j \leq t$, $V_{i,j} = p_i^j$.

Proof of Claim 3.5: Consider any set s of $2t + 1$ distinct integers $n_0 = 0, n_1, \ldots, n_{2t} \leq h$. For an integer r define $S_r = \{x \in Q_c : ||x||^2 = r\}$. We claim that if c is large enough in terms of t and h, then S_r contains no non-trivial $(2t + 1, t, h)$-progression with coefficients taken from s. Suppose to the contrary that there are such $2t + 1$ integers $z_0, z_1, \ldots, z_{2t}$. By (7) we have that for $0 \leq i \leq 2t$

$$z_i = d_0 + n_i \cdot d_1 + n_i^2 \cdot d_2 + \ldots + n_i^t \cdot d_t, \tag{14}$$

where $d_0 = z_0, d_1, \ldots, d_t$ are arbitrary integers. Recall, that for this set to be non-trivial at least one of $d_1, \ldots, d_t$ must be non-zero (the integers $n_i \in s$ are already assumed to be distinct). Denote by D the determinant of the Vandermonde matrix $V = V(n_0, \ldots, n_t)$, and for $0 \leq i \leq t$ denote by D_i the determinant of the matrix obtained from V by replacing the i^{th} column with the column vector $(z_0, \ldots, z_t)$. Observe, that we can view the first $t+1$ equations in (14) as $t+1$ equations in unknowns $d_0, d_1, \ldots, d_t$. It follows from Cramer's rule, that for $0 \leq i \leq t$ we have $Dd_i = D_i$. From the definition of the determinant we can view D_i as a linear combination of $z_0, \ldots, z_t$ with integer coefficients bounded by a constant that depends only on t and $n_0, n_1, \ldots, n_t$. As $n_0, n_1, \ldots, n_t \leq h$, these coefficients are bounded by a constant that depends only on t and h. Hence, by (12), if c was chosen large enough in terms of t and h then for $0 \leq i \leq t$, we get that $\overline{Dd_i}$ (the b dimensional vector representing Dd_i) is a linear combination of $\overline{z_0}, \ldots, \overline{z_t}$. Moreover, by (13) we may conclude that $\overline{Dd_i} \in Q_{c'}$ for some $c' < c$. As by (14), $z_0, \ldots, z_{2t}$ are defined as linear combinations of $d_0, \ldots, d_t$, we conclude that if c is large enough (so that c' is large enough), we can use (12) again to write (14) as

$$D\overline{z_i} = \overline{Dd_0} + n_i \cdot \overline{Dd_1} + n_i^2 \cdot \overline{Dd_2} + \ldots + n_i^t \cdot \overline{Dd_t}. \tag{15}$$

Define the following polynomial of degree $2t$

$$P(x) := \sum_{q=0}^{b-1} \left((\overline{Dd_0})_q + (\overline{Dd_1})_q \cdot x + (\overline{Dd_2})_q \cdot x^2 + \ldots + (\overline{Dd_t})_q \cdot x^t \right)^2,$$

where $(\overline{Dd_i})_q$ denotes the q^{th} entry of the vector $\overline{Dd_i}$. The key observation now is that by (15) we have for $0 \leq j \leq 2t$ that $P(n_j) = ||D\overline{z_j}||^2 = D^2||z_j||^2$. Hence, as by assumption all the z_is belong to S_r, we have that P is a polynomial of degree $2t$ with $2t+1$ distinct values (namely $n_0, n_1, \ldots, n_{2t}$) for which it is equal to D^2r. Therefore, P must be identical to D^2r, which can be easily seen to imply that $(\overline{Dd_i})_q = 0$ for all $0 \leq q \leq d-1$ and $1 \leq i \leq t$. Hence, $d_1 = \ldots = d_t = 0$, contradicting our assumption that this is a non-trivial $(2t+1, t, h)$-progression. We conclude that if c is large enough in terms of h and t then S_r contains no non-trivial $(2t+1, t, h)$-progression.

The claim now follows by averaging. As the absolute value of each digit in Q_c is bounded by g/c, we have $r \leq b(g/c)^2 \leq bg^2$. Similarly, we conclude that Q_c is of size $(2g/c)^b > (g/c)^b$. As the union of the sets S_r covers the entire set Q_c there must be one r for which $|S_r| \geq (g/c)^b/bg^2 = n/bg^2c^b$. Setting $b = \sqrt{\log n}$, and hence $g = e^{\sqrt{\log n}}$, gives (8) for an appropriate constant $c = c(t, h)$. ■

Proof of Claim 3.6: We again consider an arbitrary set s, of distinct integers $n_0 = 0, n_1, \ldots, n_{p-1}$ bounded by h. As in the previous proof, we continue to write $n = g^b$ and represent numbers in base g with possibly negative digits. We will also construct our sought after set from Q_c for a large enough constant c that will only depend on p, t and h. Let D denote the determinant of the Vandermonde matrix $V = V(n_0, \ldots, n_t)$. Let $R \subseteq \{1, \ldots, D^2b(g/c)^2\}$ be a set of size $R_s(p, 2t, D^2b(g/c)^2)$ that contains no non-trivial $(p, 2t, h)$-progression with coefficients from s, and recall that any translate of R also satisfies this property. Let $L = \{-D^2b(g/c)^2, \ldots, D^2b(g/c)^2\}$. For any $\ell \in L$ define

$$A_\ell = \{x \in Q_c : ||Dx||^2 \in R + \ell\}.$$

We claim that A_ℓ does not contain any non-trivial (p, t, h)-progression, with coefficients from s, provided c in the definition of Q_c is large enough. Suppose this is not case, and let $z_0, \ldots, z_{p-1}$ be such a non trivial (p, t, h)-progression. Namely, suppose there are $d_0, d_1, \ldots, d_t$ not all equal to zero such that $z_j = d_0 + \sum_{i=1}^t n_j^t d_t$ holds for $0 \leq i \leq p-1$. As by assumption $p \geq t+1$ we can still write the $t+1$ linear equations as in (14). We can then argue as in the proof of Claim 3.5 that provided c is large enough, we may conclude that for $0 \leq j \leq p-1$ one can write

$$D\overline{z_i} = \overline{Dd_0} + n_i \cdot \overline{Dd_1} + n_i^2 \cdot \overline{Dd_2} + \ldots + n_i^t \cdot \overline{Dd_t}. \tag{16}$$

This implies, as in Claim 3.5, that for every $0 \leq j \leq p-1$ we can write

$$||Dz_j|| = ||D\overline{z_j}||^2 = \sum_{q=0}^{b-1} \left(\sum_{i=0}^{t} (\overline{Dd_i})_q \cdot n_j^i \right)^2 = d'_0 + n_j \cdot d'_1 + n_j^2 \cdot d'_2 + \ldots + n_j^{2t} \cdot d'_{2t}, \tag{17}$$

where $d'_0, \ldots, d'_{2t}$ are *identical* to all $0 \leq j \leq p-1$. It is easy to see that as $d_0, \ldots, d_t$ are by assumption not all zero, then so are $d'_0, \ldots, d'_{2t}$. As $d'_0, \ldots, d'_{2t}$ are common to all $||Dz_j||^2$, the right hand side of (17) has the structure of a non-trivial $(p, 2t, h)$-progression with coefficients from s. This means that $||Dz_0||^2, \ldots, ||Dz_{t-1}||^2$ form a non-trivial $(p, 2t, h)$-progression with coefficients from s. This contradicts our choice of R and A_ℓ.

We conclude that for any $\ell \in L$, the set A_ℓ contains no non-trivial (p, t, h)-progression with coefficients from s. It is thus enough to show that for some $\ell \in L$ the set A_ℓ is large enough. We do this again by averaging. As the absolute value of the digits of numbers from Q_c is bounded by g/c we have $0 \leq ||Dx||^2 \leq D^2 b(g/c)^2$ for any $x \in Q_c$. Therefore, for any $r \in R$ and $x \in Q_c$ there is an $\ell \in L$ such that $||Dx||^2 = r + \ell$. Hence, for any $x \in Q_c$ there are $|R|$ integers $\ell \in L$ such that $x \in A_\ell$. In other words, $\sum_{\ell=1}^{|L|} A_\ell \geq |R||Q|$, and therefore for some choice of $\ell \in L$ we have $|A_\ell| \geq |R||Q_c|/|L|$. As $|Q_c| > (2g/c)^b$, the proof follows as for some $\ell \in L$ we must have

$$|A_b| \geq \frac{R(p, 2t, h, D^2bg^2) \cdot (2g/c)^b}{D^2 b(g/c)^2} \geq \frac{R(p, 2t, h, bg^2) \cdot g^b}{D^2 c^b b g^2} \geq n \frac{R(p, 2t, h, bg^2)}{c^b b g^2}, \tag{18}$$

where we used the fact that by definition $n = g^b$ and D is bounded by a function of t and h only, therefore, we can use a slightly larger constant c to "absorb" D^2. ■

4 Linear Equations and Extremal Hypergraphs

4.1 The main results of this section

In this section we describe the first algebraic construction of extremal k-graphs, which will play an important role in the proofs of Theorems 1 and 2 about testing $\mathcal{P}_D^*$ in Section 6. The second construction, related to $\mathcal{P}_D$ and Theorem 4, appears in Section 7. Denote by T^k the family of k-graphs on $k+1$ vertices, which contain **at least** three edges. Let m be an integer, T a member of T^k, and Z an arbitrary subset of $[m]$. Let also $P_d = \{p_1, \ldots, p_{k+1}\}$ be a set of $k+1$ *distinct* integers bounded by d (thus $d > k$). Consider the following definition of a k-graph $S = S(m, Z, T, P_d)$: The vertex set of S consists of $k+1$ pairwise disjoint sets of vertices $V_1, \ldots, V_{k+1}$, where, with a slight abuse of notation, we think of each of these sets as being the set of integers $1, \ldots, d^k m$. Define

$$E(z_0, z_1, \ldots, z_{k-1}, p) = z_0 + p \cdot z_1 + p^2 \cdot z_2 + p^3 \cdot z_3 + \ldots + p^{k-1} \cdot z_{k-1}. \tag{19}$$

We define the edge set of S by specifying the edge sets of $|Z|^k$ copies of T that we put in S. In what follows we refer to the $k+1$ vertices of T as integers in $\{1, \ldots, k+1\}$. For every set of (not necessarily distinct) integers $z_0, \ldots, z_{k-1} \in Z$, we put a copy of T in S spanned by the vertices $v_1 \in V_1, \ldots, v_{k+1} \in V_{k+1}$, where for $1 \leq i \leq k+1$ we choose $v_i = E(z_0, \ldots, z_{k-1}, p_i)$. In order to specify the edges of this copy, we simply regard the vertices $v_1,\dots,v_{k+1}$ as if they were the vertices $1,\dots,k+1$ of a regular copy of T and put the corresponding edges. Namely, for every edge $(t_1,\dots,t_k)\in E(T)$, we put in S an edge that contains the vertices

$$E(z_0,\dots,z_{k-1},p_{t_1})\in V_{t_1},\quad E(z_0,\dots,z_{k-1},p_{t_2})\in V_{t_2},\ \dots\ ,E(z_0,\dots,z_{k-1},p_{t_k})\in V_{t_k}.$$

In what follows we denote by $C(z_0,\dots,z_{k-1})$, the copy of T defined using the integers $z_0,\dots,z_{k-1}$. Note, that each of these $|Z|^k$ copies of D has precisely one vertex in each of the sets $V_1,\dots,V_{k+1}$. Note also, that for every $z_0,\dots,z_{k-1}$ and p_i, the function E satisfies $1\le E(z_0,\dots,z_{k-1},p_i)\le kd^{k-1}m\le d^km$, thus the vertices "fit" into the sets $V_1,\dots,V_{k+1}$. The reader should also observe that we treat the set of integers P_d, as an *ordered* set, as when choosing the vertex from V_i we use the integer $p_i\in P_d$. Our first goal in this section is to prove the following lemma.

Lemma 4.1 (The Key Lemma) *Let T be a member of T^k, m an arbitrary integer, Z a subset of $[m]$ and P_d a set of $k+1$ distinct integers bounded by d. Define $S=S(m,Z,T,P_d)$, and suppose E_1,E_2,E_3 are three edges that belong to a copy of T in S. If $E_1\in C(a_0,\dots,a_{k-1})$, $E_2\in C(b_0,\dots,b_{k-1})$ and $E_3\in C(c_0,\dots,c_{k-1})$, and if $a_i\le c_i\le b_i$ for some i, $0\le i\le k-1$, then either $a_i=b_i=c_i$ or there are positive integers $\beta_1,\beta_2\le d^{3d^2}$ such that*

$$\beta_1a_i+\beta_2b_i=(\beta_1+\beta_2)c_i.$$

Using the above lemma, we will construct the following extremal k-graph, which will be a key ingredient in the lower bounds of Theorem 1 and 2.

Lemma 4.2 *For every fixed k-graph D on d vertices that contains a copy of $T\in T^k$ with $r\ge 3$ edges, an integer m and a (r,d^{3d^2})-gadget-free set $Z\subset[m/d^{k+2}]$, there is a k-graph F on m vertices with the following properties:*

1. *F contains $|Z|^k$ induced copies of D, which are singled out from the rest of the copies of D and are called the* essential copies *of D in F.*
2. *Each pair of these essential copies share at most $k-1$ common vertices.*
3. *Every copy of T belongs to one of the essential copies of D.*

It is important to note that we do not claim that F does not contain any copies of D other than the $|Z|^k$ essential copies, nor will we claim so later on in this section. As the statement of the above lemma is rather technical, the reader can find in Subsection 4.3 a short intuitive explanation of it.

4.2 Proof of Lemma 4.1

The main idea of the proof is very simple; as T has only $k+1$ vertices, the 3 edges spanned by these vertices must have many common vertices. As the vertices of each set were chosen using the function E defined in (19), we get from each vertex that is common to, say, E_1 and E_2, a linear equation that relates the integers $a_0,\dots,a_{k-1}$, which were used to define E_1 and the integers $b_0,\dots,b_{k-1}$, which were used to define E_2. We then show that for every i either $a_i=b_i=c_i$ or the linear equations induced by the intersections of the edges are "reach" enough to enable us to extract a linear equation of the form $\beta_1a_i+\beta_2b_i=(\beta_1+\beta_2)c_i$.

Let E_1, E_2 and E_3 be three edges that belong to a copy of a member of $T \in T^k$. As T has $k+1$ vertices and any k-graph on $k+1$ vertices that contains at least 3 edges is a core (recall Definition 2.1), the $k+1$ vertices must belong to distinct sets V_i. Call these vertices $v_1 \in V_1, \ldots, v_{k+1} \in V_{k+1}$. Assume, without loss of generality, that $E_1 = \{v_1, \ldots, v_{k+1}\} \setminus v_{k+1}$, $E_2 = \{v_1, \ldots, v_{k+1}\} \setminus v_k$ and $E_3 = \{v_1, \ldots, v_{k+1}\} \setminus v_{k-1}$. Recall, that every edge in S belongs to one of the copies of T, defined using some k integers from Z. Suppose $E_1 \in C(a_0, \ldots, a_{k-1})$, $E_2 \in C(b_0, \ldots, b_{k-1})$, and $E_3 \in C(c_0, \ldots, c_{k-1})$. As $v_1 \in V_1, \ldots, v_{k-1} \in V_{k-1}$, are common to both E_1 and E_2 we conclude that for every $i \in [k+1] \setminus \{k, k+1\}$, the following equation holds:

$$E(a_0, \ldots, a_{k-1}, p_i) = v_i = E(b_0, \ldots, b_{k-1}, p_i).$$

As $v_1 \in V_1, \ldots, v_{k-2} \in V_{k-2}, v_k \in V_k$, are common to both E_1 and E_3 we conclude that for every $i \in [k+1] \setminus \{k-1, k+1\}$, the following equation holds:

$$E(a_0, \ldots, a_{k-1}, p_i) = v_i = E(c_0, \ldots, c_{k-1}, p_i).$$

We could have written $k-1$ equations for the common vertices of E_2 and E_3, however, all but one of them follow from the previous equations. The only independent equation is due to v_{k+1}:

$$E(b_0, \ldots, b_{k-1}, p_{k+1}) = v_{k+1} = E(c_0, \ldots, c_{k-1}, p_{k+1}).$$

We get a set of $2k-1$ equations in $3k$ unknowns, $a_0, \ldots, a_{k-1}$, $b_0, \ldots, b_{k-1}$ and $c_0, \ldots, c_{k-1}$. In order to simplify the rest of this subsection, we substitute the definition of E from (19) and write our set of equations as follows:

$$\begin{aligned}
a_0 + p_1 a_1 + p_1^2 a_2 + \ldots + p_1^{k-1} a_{k-1} &= b_0 + p_1 b_1 + p_1^2 b_2 + \ldots + p_1^{k-1} b_{k-1} \\
a_0 + p_2 a_1 + p_2^2 a_2 + \ldots + p_2^{k-1} a_{k-1} &= b_0 + p_2 b_1 + p_2^2 b_2 + \ldots + p_2^{k-1} b_{k-1} \\
&\vdots \\
a_0 + p_{k-1} a_1 + p_{k-1}^2 a_2 + \ldots + p_{k-1}^{k-1} a_{k-1} &= b_0 + p_{k-1} b_1 + p_{k-1}^2 b_2 + \ldots + p_{k-1}^{k-1} b_{k-1} \\
\\
a_0 + p_1 a_1 + p_1^2 a_2 + \ldots + p_1^{k-1} a_{k-1} &= c_0 + p_1 c_1 + p_1^2 c_2 + \ldots + p_1^{k-1} c_{k-1} \\
a_0 + p_2 a_1 + p_2^2 a_2 + \ldots + p_2^{k-1} a_{k-1} &= c_0 + p_2 c_1 + p_2^2 c_2 + \ldots + p_2^{k-1} c_{k-1} \\
&\vdots \\
a_0 + p_{k-2} a_1 + p_{k-2}^2 a_2 + \ldots + p_{k-2}^{k-1} a_{k-1} &= c_0 + p_{k-2} c_1 + p_{k-2}^2 c_2 + \ldots + p_{k-2}^{k-1} c_{k-1} \\
a_0 + p_k a_1 + p_k^2 a_2 + \ldots + p_k^{k-1} a_{k-1} &= c_0 + p_k c_1 + p_k^2 c_2 + \ldots + p_k^{k-1} c_{k-1} \\
\\
b_0 + p_{k+1} b_1 + p_{k+1}^2 b_2 + \ldots + p_{k+1}^{k-1} b_{k-1} &= c_0 + p_{k+1} c_1 + p_{k+1}^2 c_2 + \ldots + p_{k+1}^{k-1} c_{k-1}
\end{aligned}$$

In what follows we denote by Φ the above set of equations. The main idea of the proof will be to show that either $a_0 = b_0 = c_0$ or there is a linear combination of Φ with *integer coefficients* $\alpha_1, \ldots, \alpha_{2k-1}$, which results in the required linear equation relating a_0, b_0 and c_0. The other cases relating a_i, b_i, c_i with $i > 0$ are completely identical. The main idea is to find a linear combination in which for $1 \leq i \leq k-1$ the coefficients of a_i, b_i and c_i vanish. To this end, we introduce a set of equations whose solution will be our desired α_is. Observing Φ, we see that all the a_is appear in the left hand side of the first $2k-2$ equations. Thus, in order for the coefficient of a_i to vanish in a linear combination of Φ with coefficients $\alpha_1, \ldots, \alpha_{2k-1}$, the following equation must hold

$$A_i: \qquad \alpha_1 \cdot p_1^i + \alpha_2 \cdot p_2^i + \ldots + \alpha_{2k-3} \cdot p_{k-2}^i + \alpha_{2k-2} \cdot p_k^i = 0.$$

Each of the b_is appears only in the right hand side of the first $k-1$ equations and in the left hand side of the last equation. Therefore, in order for the coefficient of b_i to vanish the following equation must hold

$$B_i: \qquad \alpha_1 \cdot p_1^i + \alpha_2 \cdot p_2^i + \ldots + \alpha_{k-1} \cdot p_{k-1}^i - \alpha_{2k-1} \cdot p_{k+1}^i = 0.$$

Finally, each of the c_is appears only in the right hand side of the last $k-1$ equations. Hence, in order for the coefficient of c_i to vanish the following equation must hold

$$C_i: \qquad \alpha_k \cdot p_1^i + \alpha_{k+1} \cdot p_2^i + \ldots + \alpha_{2k-3} \cdot p_{k-2}^i + \alpha_{2k-2} \cdot p_k^i + \alpha_{2k-1} \cdot p_{k+1}^i = 0.$$

Observe, that we can write the analogous linear equations A_0, B_0 and C_0 that will require the coefficients of a_0, b_0 and c_0 to vanish. Though we apparently don't need these equations, they will be useful for the proof. In what follows we denote by Υ the set of equations $A_1, \ldots, A_{k-1}$, $B_1, \ldots, B_{k-1}$, and $C_1, \ldots, C_{k-1}$. We will need the following well known result that follows from Cramer's rule and Hadamard Inequality (see, e.g., [16]).

Lemma 4.3 *Let Ψ be a set of p homogenous linear equations in q variables. If $p < q$ and each of the coefficients in these equations has absolute value at most r, then Ψ has a* non zero *solution $\{\alpha_1, \ldots, \alpha_q\}$, where all the α_is are* integers *with absolute value at most $(r^2p)^{p/2}$.*

The set Υ consists of $3k-3$ homogenous linear equations in $2k-1$ unknowns $\alpha_1, \ldots, \alpha_{2k-1}$. Observe, however, that for $1 \leq i \leq k-1$,

$$A_i = B_i + C_i.$$

Therefore, Υ is equivalent to a set of $3k-3-(k-1) = 2k-2$ linear homogenous equations in $2k-1$ unknowns, which consists of equations B_i, C_i . Observe also, that each of the coefficients in Υ has absolute value at most d^k (recall that $1 \leq p_1, \ldots, p_{k+1} \leq d$). By Lemma 4.3, we are guaranteed that there are *integers* $\alpha_1, \ldots, \alpha_{2k-1}$ not all of them equal to zero, whose absolute value is upper bounded by $(d^{2k}(2k-2))^{k-1} \leq d^{2d^2}$, such that in a linear combination of the above equations the coefficients of all the variables but a_0, b_0, c_0 vanish. We now claim that in such a combination either the coefficient of b_0 or the coefficient of c_0 does not vanish. An important observation is that as the integers $p_1, \ldots, p_{k+1}$ are distinct, the k linear equations $B_0, \ldots, B_{k-1}$ that require the coefficients of $b_0, \ldots, b_{k-1}$ to vanish are linearly independent. Hence, their only solution is $\alpha_1 = 0, \ldots, \alpha_{k-1} = 0, \alpha_{2k-1} = 0$. Similarly, the k linear equations $C_0, \ldots, C_{k-1}$ that require the coefficients of $c_0, \ldots, c_{k_1}$ to vanish are linearly independent. Hence, their only solution is $\alpha_k = 0, \ldots, \alpha_{2k-1} = 0$. Thus, if the coefficients of b_0 and c_0 vanish we may conclude that we must have used a linear combination with $\alpha_1 = \ldots = \alpha_{2k-1} = 0$, which contradicts our choice.

Note, that as in each of the equations of Φ the sum of the coefficients in the right hand side is equal to the sum of the coefficients in the left hand side, this property must also hold in a linear combination of Φ. Hence, there is no linear combination in which the coefficient of precisely one of a_0, b_0, c_0 does not vanish. It also follows that if the coefficients of precisely two of a_0, b_0, c_0 do not vanish, then they must be equal. However, if for example $a_0 = b_0$, then we can "replace" b_0 with a_0 in the last equation of Φ, and use the last k equations of Φ to infer that for $1 \leq i \leq k-1$ we have $a_i = c_i$. We would thus get that $a_0 = b_0 = b_0$, which satisfies the requirement of the lemma. The other two cases are similar. As in the previous paragraph we have ruled out the possibility that the coefficients of a_0, b_0 and c_0 vanish, the remaining possibility is that the coefficients a_0, b_0 and c_0 do not vanish. In this case, we can use again the fact that in the resultant equation the sums of the coefficients in each side are equal to infer that we must get the required equation. Finally, as the coefficients α_i are bounded by d^{2d^2}, the coefficients in the linear combination are bounded by $(2k-1)d^{2d^2} < d^{3d^2}$. ■

4.3 Intuition for Lemma 4.2

We give some explanation as to why, or more precisely *when*, Lemma 4.2 is not trivial. Consider for simplicity the case of $k = 2$, that is, when T^k is simply a triangle, and D is a K_4 (a clique of size 4). In this case, the lemma says that we can construct a graph on m vertices that contains $|Z|^2$ essential copies of K_4 that are pairwise edge disjoint, and such that each triangle in the graph belongs to one of these copies of K_4. Note, that if $|Z| = 1$ this statement is trivial as we can simply take a single copy of K_4 in order to construct such a graph. However, if $|Z| = m^{1-o(1)}$, the lemma claims that we can construct the following non trivial graph: It has m vertices and $|Z|^2 = m^{2-o(1)}$ essential copies of K_4 that are pairwise edge disjoint, such that each triangle in the graph belongs to one of these copies of K_4. As each K_4 contains at most 4 triangles, this graph contains less than m^2 triangles. As any triangle appears in at most m copies of K_4 such a graph has at most m^3 copies of K_4. Note that any trivial such construction (e.g. random) will contain roughly $m^{4-o(1)}$ copies of K_4. The fact that we can construct graphs that contain many induced copies of a graph, where each two copies have at most 1 common vertex (or $k-1$ vertices in the case of k-graphs) while containing relatively few copies of it, will be crucial in the proofs of Theorems 1 and 2.

4.4 Proof of Lemma 4.2

We define a k-graph F, similar to the one used in Lemma 4.1. The vertex set of F consists of d pairwise disjoint sets of vertices $V_1, \ldots, V_d$, where, with a slight abuse of notation, we think of each of these sets as being the set of integers $1, \ldots, m/d$. We define the edge set of F by specifying the edge sets of $|Z|^k$ copies of D that we put in F. In what follows we refer to the d vertices of D as integers in $\{1, \ldots, d\}$.

For every set of (not necessarily distinct) integers $z_0, \ldots, z_{k-1} \in Z$, we put a copy of D in F spanned by the vertices $v_1 \in V_1, \ldots, v_d \in V_d$, where for $1 \leq i \leq d$ we choose $v_i = E(z_0, \ldots, z_{k-1}, i)$. In order to specify the edges of this copy, we simply regard the vertices $v_1, \ldots, v_d$ as if they were the vertices of a regular copy of D and put the corresponding edges. Namely, for every edge $(p_1, \ldots, p_k) \in E(T)$, we put in F an edge that contains the vertices

$$E(z_0, \ldots, z_{k-1}, p_1) \in V_{p_1}, \quad E(z_0, \ldots, z_{k-1}, p_2) \in V_{p_2}, \quad \ldots \quad , E(z_0, \ldots, z_{k-1}, p_k) \in V_{p_k}.$$

In what follows we denote by $C(z_0,\ldots,z_{k-1})$, the copy of D defined using the integers $z_0,\ldots,z_{k-1}$. This defines $|Z|^k$ copies of D. These $|Z|^k$ copies of D will be the essential copies of D in F in the statement of the lemma (but we will still have to show that they are induced copies of D in F). Observe, that any essential copy D has precisely one vertex in each of the sets $V_1,\ldots,V_d$. Note also, that as $Z \subseteq [m/d^{k+2}]$, for every $z_0,\ldots,z_{k-1}$ and $1 \leq i \leq d$, the function E satisfies $1 \leq E(z_0,\ldots,z_{k-1},i) \leq kd^{k-1}m/d^{k+2} \leq m/d$, thus the vertices "fit" into the sets $V_1,\ldots,V_d$.

We now turn to prove the assertions of the lemma. Let $v_1,\ldots,v_k$ be k vertices that belong to one of the essential copies of D in F. As the vertices of an essential copy belong to different sets V_i, there are *distinct* integers $1 \leq p_1,\ldots,p_k \leq d$, such that $v_1 \in V_{p_1},\ldots,v_k \in V_{p_k}$. From the definitions of F and the function E in (19), there are $z_0,\ldots,z_{k-1}$, such that the following equations hold:

$$z_0 + p_1 z_1 + p_1^2 z_2 + \ldots + p_1^{k-1} z_{k-1} = E(z_0,\ldots,z_{k-1},p_1) = v_1,$$

$$\vdots$$

$$z_0 + p_k z_1 + p_k^2 z_2 + \ldots + p_k^{k-1} z_{k-1} = E(z_0,\ldots,z_{k-1},p_k) = v_k.$$

If we view the following equations as a set of k linear equations with unknowns $z_0,\ldots,z_{k-1}$, they correspond to the matrix equation $Ax = b$, where $b = \{v_1,\ldots,v_k\}$, $x = \{z_0,\ldots,z_{k-1}\}$, and $A_{i,j} = p_i^{j-1}$. As A is an invertible Vandermonde matrix (here we use the fact that the p_is are distinct), we conclude that $z_0,\ldots,z_{k-1}$ are uniquely defined by this set of equations. Hence, they belong to precisely one of the essential copies of D, namely, $C(z_0,\ldots,z_{k-1})$. We have thus shown that each pair of essential copies share at most $k-1$ common vertices. As F is a k-graph, the essential copies of D are in particular edge disjoint. As by definition, every edge in D belongs to one of the essential copies of D, we conclude that the essential copies of D in F are in fact *induced*. We have thus proved items (1) and (2).

We now turn to prove item (3). Suppose $v_1,\ldots,v_{k+1}$ are $k+1$ vertices that span a copy of T, namely, they span $r \geq 3$ edges. As any member of T^k contains at least 3 edges, T is a core (recall Definition 2.1). Hence, there are distinct $p_1,\ldots,p_{k+1}$ such that $v_1 \in V_{p_1},\ldots,v_{k+1} \in V_{p_{k+1}}$. Suppose the r edges of T are $e_1 \in C(z_{0,1},\ldots,z_{k-1,1}),\ldots,e_r \in C(z_{0,r},\ldots,z_{k-1,r})$. In order to show that each copy of T belongs to one of the essential copies of D we may show that for $0 \leq i \leq k-1$ we have $z_{i,1} = \ldots = z_{i,r}$. This will mean that the r edges belong to $C(z_0,\ldots,z_{k-1})$. For ease of notation we show that $z_{1,1} = \ldots = z_{1,r}$. The other cases are completely identical.

An important observation at this point is that the subgraph of F induced on $V_{p_1},\ldots,V_{p_{k+1}}$ is *precisely* the k-graph S defined in Lemma 4.1 with $P_d = \{p_1,\ldots,p_{k+1}\}$. Consider any three distinct integers $j_1,j_2,j_3 \in \{1,\ldots,r\}$, such that $z_{1,j_1} \leq z_{1,j_2} \leq z_{1,j_3}$. By Lemma 4.1, either $z_{1,j_1} = z_{1,j_2} = z_{1,j_3}$ or there are positive integers $\beta_1,\beta_2 \leq d^{3d^2}$ such that the following equation holds

$$\beta_1 z_{1,j_1} + \beta_2 z_{1,j_3} = (\beta_1 + \beta_2) z_{1,j_2}.$$

Assume first that for some choice of $j_1,j_2,j_3 \in \{1,\ldots,r\}$ we have $z_{1,j_1} = z_{1,j_2} = z_{1,j_3}$ and assume for simplicity that $j_1 = 1, j_2 = 2, j_3 = 3$. Consider any other $4 \leq j \leq r$ and assume without loss of generality that $z_{1,1} \leq z_{1,j} \leq z_{1,2}$. By the above, either $z_{1,1} = z_{1,2} = z_{1,j}$ or there are positive integers $\beta_1,\beta_2 \leq d^{3d^2}$ such that $\beta_1 z_{1,1} + \beta_2 z_{1,2} = (\beta_1 + \beta_2) z_{1,j}$. However, as by assumption $z_{1,1} = z_{1,2}$ and $\beta_1,\beta_2 > 0$ we can conclude that in this case we also have $z_{1,1} = z_{1,2} = z_{1,j}$. We thus conclude that in this case we have $z_{1,1} = z_{1,2} = \ldots = z_{1,r}$.

Assume now that none of $j_1, j_2, j_3 \in \{1, \ldots, r\}$ are such that $z_{1,j_1} = z_{1,j_2} = z_{1,j_3}$. Suppose we rename the integers $z_{1,1}, \ldots, z_{1,r}$ such that $z_{1,1} \leq \ldots \leq z_{1,r}$. By Lemma 4.1, we have for every $2 \leq i \leq r-1$ that there are positive integers $\beta_{i_1}, \beta_{i_2} \leq d^{3d^2}$ such that

$$\beta_{i_1} z_{1,i-1} + \beta_{i_2} z_{1,i+1} = (\beta_{i_1} + \beta_{i_2}) z_{1,i} \tag{20}$$

holds (Note that by our ordering of $z_{1,1}, \ldots, z_{1,r}$ we satisfy the requirement of Lemma 4.1 that $a_i \leq c_i \leq b_i$). But this means that $z_{1,1}, \ldots, z_{1,r}$ satisfy the (r, d^{3d^2})-gadget $\mathcal{E} = \{e_2, \ldots, e_{r-1}\}$ where

$$e_i := \beta_{i_1} x_{i-1} + \beta_{i_2} x_{i+1} = (\beta_{i_1} + \beta_{i_2}) x_i$$

(it is easy to verify that this is indeed a (r, d^{3d^2})-gadget). However, as by assumption Z is (r, d^{3d^2})-gadget-free, the integers $z_{1,1}, \ldots, z_{1,r}$ cannot be distinct. Assume, without loss of generality, that $z_{1,1} = z_{1,2}$. As $z_{1,1}, z_{1,2}, z_{1,3}$ satisfy the linear equation given in (20) and as by assumption $z_{1,1} = z_{1,2}$ it must be the case that $z_{1,3} = z_{1,1} = z_{1,2}$. This contradicts our assumption that there is no triple of equal integers $z_{1,j_1}, z_{1,j_2}, z_{1,j_3}$. ■

5 Extremal Hypergraphs and Lower Bounds for $\mathcal{P}^*_D$

5.1 The main results of this section

Our main goal in this section is to prove the following lemma

Lemma 5.1 *Let D be a fixed k-graph on d vertices, which contains a copy of $T \in T^k$ with r edges. Suppose we can find, for every integer m, a (r, d^{3d^2})-gadget-free subset $Z \subseteq [m/d^{k+2}]$ of size $m/f(m)$. Then, for every small enough $\epsilon > 0$, and every large enough integer n, there is a k-graph H on n vertices that is ϵ-far from being induced D-free, and yet contains only $O(n^d/q(\epsilon))$ copies of D, where $q(\epsilon) = \max\{m : (1/f(m))^k \geq \epsilon\}$.*

We also need the following lemma, which follows from the *canonical graph property tester* of Goldreich and Trevisan in [15] (see also [3]).

Lemma 5.2 *Suppose there is a k-graph on n vertices that is ϵ-far from satisfying $\mathcal{P}^*_D$ (or $\mathcal{P}_D$) and yet contains $O(n^d/q(\epsilon))$ copies of D. Then the query complexity of any one-sided error property-tester for $\mathcal{P}^*_D$ (or $\mathcal{P}_D$) is $\Omega(q(\epsilon)^{1/d})$, where d is the size of D. In particular, if $q(\epsilon)$ is super-polynomial in $1/\epsilon$, then so is the query complexity of any one-sided error property-tester for $\mathcal{P}^*_D$ (or $\mathcal{P}_D$).*

As is evident from the statement of Lemma 5.2, in order to obtain a high lower bound for testing $\mathcal{P}^*_D$, one would want to apply it to a k-graph H that is ϵ-far from satisfying $\mathcal{P}^*_D$ and contains $O(n^d/q(\epsilon))$ copies of D with q growing as fast as possible. Inspecting the statement of Lemma 5.1 we see that it supplies such a k-graph, but in this case the function f should grow as slow as possible (in some sense q is f^{-1}). Note, that one can use the output of Theorem 6 as an input to Lemma 5.1. Finally, requiring f in Lemma 5.1 to grow slowly, means requiring the set Z in Theorem 6 to be as large as possible. Finally, observe that we can use the number theoretic construction of Theorem 6, which supplies such a set of size $n/f(n)$ with f being sub-polynomial. This will give a super-polynomial q, and thus super-polynomial lower bounds, which are our ultimate goals. In

Section 6 we indeed apply the above two lemmas, along with Lemma 4.2 and the number theoretic construction of Theorem 6, In order to prove Theorems 1 and 2. The reader can find further intuition for Lemma 5.1 in the following Subsection. The proofs of Lemmas 5.1 and 5.2 appear in the following subsections.

5.2 Intuition for Lemma 5.1

Going back to the discussion following the statement of Lemma 4.2 we see that using Lemma 4.2 with a set Z of size $n^{1-o(1)}$ gets us very close to the requirements of Lemma 5.2, with two important differences. Returning to the example of K_4 from Subsection 4.3, we see that on the one hand the k-graph of Lemma 4.2 contains at most m^3 copies of K_4 on m vertices, which is far better than the $n^4/q(\epsilon)$ copies on n vertices, which Lemma 5.2 expects to get. On the other hand, however, the input k-graph to Lemma 5.2 must be ϵ-far from being induced K_4-free while the k-graph in Lemma 4.2 is only $m^{-o(1)}$-far from being induced K_4-free as it contains only $m^{2-o(1)}$ copies of K_4. Thus, Lemma 5.1 can be viewed as a bridge between the extremal hypergraph construction of Lemma 4.2 and the lower bounds that we can obtain using Lemma 5.2.

5.3 Proof of Lemma 5.1

An *s-blow-up* of a k-graph $T = (V(T), E(T))$ on t vertices is the k-graph obtained from T by replacing each vertex $v_i \in V(T)$ by an independent set I_i of size s, and each edge $(v_{i_1}, \ldots, v_{i_k}) \in E(T)$, by a complete k-partite k-graph whose vertex classes are $I_{i_1}, \ldots, I_{i_k}$ (A complete k-partite k-graph has as its vertex set k sets $V_1, \ldots, V_k$, and its edge set is $\{\{v_1, \ldots, v_k\} : v_1 \in V_1, \ldots, v_k \in V_k\}$). Note that if we take an s-blow-up of T, we get a k-graph on st vertices that contains s^t induced copies of T, where each vertex of the copy belongs to a different blow-up of a vertex from T (simply pick one vertex from each independent set). We call these copies the *special copies* of the blow-up. As each set of k vertices in the blow-up is contained in at most s^{t-k} special copies of T, it follows that adding or removing an edge from the k-graph can destroy at most s^{t-k} special copies of T. We conclude that one must add or remove at least $s^t/s^{t-k} = s^k$ edges from the blow-up in order to destroy all its special copies of T.

Proof of Lemma 5.1: Given a small $\epsilon > 0$, define

$$m = q(\epsilon). \tag{21}$$

Let $Z \subseteq [m/d^{k+2}]$ be a (r, d^{3d^2})-gadget-free, and let F be the output of Lemma 4.2, given D, T, m and Z. Recall that F has m vertices. Let H be an s-blow-up of F, where

$$s = \left\lfloor \frac{n}{|V(F)|} \right\rfloor = \left\lfloor \frac{n}{m} \right\rfloor. \tag{22}$$

If necessary, add some more isolated vertices to make sure that the number of vertices of H is precisely n. Claims 5.1 and 5.3 below complete the proof of this lemma ■

Claim 5.1 *The k-graph H defined in the proof of Lemma 5.1 is ϵ-far from being induced D-free.*

Proof: Consider two essential copies of D in F, D_1 and D_2. By item (2) in Lemma 4.2, D_1 and D_2 share at most $k-1$ vertices $v_{i_1}, \ldots, v_{i_{k-1}}$ in F. It follows that their corresponding blow-ups in H will share at most $k-1$ independent sets $I_{i_1}, \ldots, I_{i_{k-1}}$, which replace the vertices $v_{i_1}, \ldots, v_{i_{k-1}}$ from F. Now, consider the blow-ups of D_1 and D_2 in H, denoted T_1 and T_2. As T_1 and T_2 share at most $k-1$ common independent set, and each of the special copies of D in T_1/T_2 has *precisely* one vertex in each of the independent sets that replaced the vertices of F, we get that a special copy of D in T_1 and a special copy of D in T_2 share at most $k-1$ vertices. Thus, adding or removing an edge from H, can either destroy special copies of D that belong to T_1, or special copies of D that belong to T_2 (or not destroy any induced copies at all). By item (2) in Lemma 5.1 each of the essential copies of D in F is induced, thus, as we explained above, in order to destroy all the special copies of an s-blow-up of D, one needs to add or remove at least s^k edges from the blow-up. As $|Z| = m/f(m)$ we have by Lemma 4.2 item (1) that F contains $m^k/f^k(m)$ essential copies of D. Therefore, H contains $m^k/f^k(m)$ blow-ups of copies of D. We may thus infer that one has to add or delete at least

$$\frac{s^k m^k}{f^k(m)} = \frac{n^k}{f^k(m)} \geq \epsilon n^k \tag{23}$$

edges in order to turn H into an induced D-free k-graph, where the inequality follows from our choice of m in (21) and the definition of $q(\epsilon)$. Thus, H is ϵ-far from being induced D-free. ■

In what follows we denote by I_v the independent set of vertices in H that replaced vertex v from F. As H is a blow-up of F it is clear that $\{v_1 \in I_{t_1}, \ldots, v_k \in I_{t_k}\}$ is an edge in H if and only if $\{t_1, \ldots, t_k\}$ is an edge in F. We remind the reader that by assumption D contains a copy of $T \in T^k$, which contains $r \geq 3$ edges. We need the following simple claim:

Claim 5.2 *The number of copies of T in H is s^{k+1} times the number of copies of T in F.*

Proof: Assume $v_1 \in I_{t_1}, \ldots, v_{k+1} \in I_{t_{k+1}}$ span a copy of T in H. As T is a core the sets $I_{t_1}, \ldots, I_{t_{k+1}}$ are all distinct. As H is a blow-up of F we get that $t_1, \ldots, t_{k+1}$ span a copy of T in F. We conclude that a copy of T in H is obtained only by picking a single vertex from each one of the $k+1$ sets $I_{t_1}, \ldots, I_{t_{k+1}}$ such that $t_1, \ldots, t_{k+1}$ span a copy of T in F. As H is an s blow-up of F, we conclude that the number of copies of T in H is precisely s^{k+1} times the number of copies of T in F. ■

Claim 5.3 *The k-graph H defined in the proof of Lemma 5.1 has $O(n^d/q(\epsilon))$ copies of D.*

Proof: Note, that as D contains at least one copy of T, and each copy of T belongs to at most $\binom{n}{d-k-1} \leq n^{d-k-1}$ copies of D, it is enough to show that H contains at most $n^{k+1}/q(\epsilon)$ copies (induced or not) of T. By Claim 5.2, the number of copies of T in H is precisely s^{k+1} times the number of copies of T in F. By item (3) in Lemma 4.2 each copy of T belongs to one of the essential copies of D. As each copy of D can contain at most $\binom{d}{k+1} \leq d^{k+1}$ copies of T, and F contains *precisely* $m^k/f^k(m)$ essential copies of D, we get that H contains at most

$$\frac{d^{k+1} \cdot m^k \cdot s^{k+1}}{f^k(m)} = \frac{d^{k+1} \cdot m^k \cdot n^{k+1}}{f^k(m) \cdot m^{k+1}} \leq \frac{d^{k+1} \cdot n^{k+1}}{m} = O(n^{k+1}/q(\epsilon)) \tag{24}$$

copies of T, where the first equality is due to our choice of s in (22), and in the last equality we used the definition of m in (21). ■

5.4 Proof of Lemma 5.2

We need the following result of [15], mentioned already in [3].

Lemma 5.3 ([3],[15]) *If there exists an ϵ-tester for a graph property that makes q queries, then there exists such an ϵ-tester that makes its queries by uniformly and randomly choosing a set of $2q$ vertices and querying all their pairs. In particular, it is a non-adaptive ϵ-tester making $\binom{2q}{2}$ queries.*

In [15] the authors measure the query complexity of a property tester by counting the number of edge queries. As we measure query complexity by the the number of vertices sampled, assuming we always query all possible edges within the sample, we infer from Lemma 5.3 that we can simply assume that the property tester first samples a set of vertices, queries about all the edges, and then proceeds to perform some other computation. Also, the proof of Lemma 5.3 was described in [15] for graphs, however, precisely the same argument carries over to k-graphs.

Proof of Lemma 5.2: We describe the proof with respect to $\mathcal{P}^*_D$. The proof for $\mathcal{P}_D$ is identical. By Lemma 5.3, given a one-sided error ϵ-tester for testing $\mathcal{P}^*_D$ we may assume, without loss of generality, that it queries about all k-subsets of a randomly chosen set of vertices. As the algorithm is a one-sided-error algorithm, it can report that H is not induced D-free only if it finds an induced copy of D in it. Observe, that if the tester picks a random subset of x vertices, and an input k-graph contains only $O(n^d/q(\epsilon))$ copies (induced or not) of D, then the expected number of induced copies of D spanned by x is $O(x^d/q(\epsilon))$, which is $o(1)$ unless $x = \Omega(q(\epsilon)^{1/d})$. By Markov's inequality, unless $x = \Omega(q(\epsilon)^{1/d})$, the tester finds an induced copy of D with negligible probability. ■

6 Proofs of Theorems 1 and 2

6.1 A lower bound for (almost) all k-graphs

In this section we apply the number theoretic construction of Theorem 6, the construction of the extremal k-graphs of Lemma 4.2 as well as Lemmas 5.1 and 5.2 in order to prove Theorem 1. We first need the following claim in which we denote by $\overline{D}$ the complement of D, that is, a k-graph that contains an edge if and only if D does not. We also call a k-graph D, *strongly* T^k-free, if neither D nor $\overline{D}$ contains a copy of T^k.

Claim 6.1 *There are no strongly T^3-free 3-graphs on at least 7 vertices. For any $k > 3$, there are no strongly T^k-free k-graphs on at least $k+1$ vertices.*

Proof: The case of $k > 3$ is simple. As in this case $\binom{k+1}{k} \geq 5$, on *any* set of $k+1$ vertices either D or $\overline{D}$ spans a copy of T^k. For the case of $k = 3$, observe that D is strongly T^3-free, if and only if each set of 4 vertices spans precisely 2 edges. Fixing any set of 7 vertices, this set must span precisely $\binom{7}{4}2/4$ edges, where we count the number of 4-sets, multiply by 2 as each 4-set by assumption spans 2 edges, and divide by 4, because we count each edge 4 times. Since this is not an integer it is impossible. Thus, on *any* set of 7 vertices either D or $\overline{D}$ spans a copy of T^3. ■

Proof of Theorem 1: Let D be a fixed k-graph on d vertices. A simple yet crucial observation is that for every k-graph D, testing $\mathcal{P}^*_D$ is equivalent to testing $P^*_{\overline{D}}$. Note, that this relation does not hold for testing $\mathcal{P}_D$. It follows that in order to prove a lower bound for testing $\mathcal{P}^*_D$, we may prove a lower bound for testing $P^*_{\overline{D}}$. By Claim 6.1 all the k-graphs in the statement of Theorem 1 (besides some 3-graphs on 4,5 and 6 vertices. See comment below on how to deal with them) are not strongly T^k-free, hence we may assume that D contains a copy of $T \in T^k$ with at least 3 edges. By Theorem 6 (with $k = 2$ and $h = d^{3d^2}$), there is a $(3, d^{3d^2})$-gadget-free set $Z \subseteq m/d^{k+2}$ of size $(m/d^{k+2})/e^{c\sqrt{\log(m/d^{k+2})}} = m/e^{c\sqrt{\log m}}$ for an appropriate $c = c(d)$. This means that we can use Lemma 5.1 with $f(m) = e^{c\sqrt{\log m}}$. It is easy to check that in this case $q(\epsilon)$ in the statement of Lemma 5.1 satisfies

$$q(\epsilon) \geq \left(\frac{1}{\epsilon}\right)^{c' \log 1/\epsilon}, \tag{25}$$

for an appropriate constant $c' = c'(d)$. By Lemma 5.1 we get a k-graph that is ϵ-far from being induced D-free, and contains only $O(n^d/q(\epsilon))$ copies of D. By Lemma 5.2 the query complexity of any one-sided error property tester for $\mathcal{P}^*_D$ is lower bounded by $q(\epsilon)^{1/d}$, which is (25), with c' replaced by c'/d. ■

It is worth mentioning that there are strongly T^3-free 3-graphs on 4,5, and 6 vertices. For 4 vertices there is a unique such 3-graph, which is $D_{3,2}$ (which contains 2 edges) mentioned in the statement of the Theorem 2. This is the only k-graph for which we do not know whether $\mathcal{P}^*_D$ is easily testable. For 5 vertices, it is easy to verify that the only strongly T^3-free 3-graph has the edges $\{(1,2,3),\ (2,3,4),\ (3,4,5),\ (4,5,1),\ (5,1,2)\}$. This 3-graph is better understood if one considers the 5 vertices on a cycle, and the edges as all triples consisting of three consecutive vertices on the cycle. In what follows we call it $D_{3,5}$. It is easy to check that $D_{3,5}$ is a hyper-cycle (see Definition 2.2), thus we can prove a version of Lemma 5.1 (namely, constructing a k-graph, which is ϵ-far from being $D_{3,5}$-free and yet contains only $O(n^5/(1/\epsilon)^{c \log 1/\epsilon})$ copies of $D_{3,5}$) that instead of using Lemmas 4.2 and Theorem 6, uses Lemmas 7.1 and 7.2, which are proved below. The details are very similar. For 6 vertices there are also some 3-graphs that are strongly T^3-free, however, every 5 vertices of such a 3-graph must span a copy of $D_{3,5}$ thus we can again use the same arguments as for $D_{3,5}$ to prove that any such 3-graph is not easily testable.

6.2 The improved lower bound

Proof of Theorem 2: Observing, as in the proof of Theorem 1, that we may either prove a lower bound for D or $\overline{D}$, we recall that by Ramsey's Theorem, for any integer k there is an integer $r(k)$ such that for any k-graph D on at least $r(k)$ vertices, either D or $\overline{D}$ contains a copy of F^k. Hence, we may assume that D contains a copy of F^k, which is a member of T^k with $k+1$ edges. By Theorem 6, there is a $(k+1, d^{3d^3})$-gadget-free set $Z \subseteq m/d^{k+2}$ of size

$$|Z| \geq (m/d^{k+2})/e^{c'(\log(m/d^{k+2}))^{1/\lfloor \log 2k \rfloor}} = m/e^{c(\log m)^{1/\lfloor \log 2k \rfloor}}$$

for an appropriate $c = c(d,k)$. This means that we can use Lemma 5.1 with $f(m) = e^{c(\log m)^{1/\lfloor \log 2k \rfloor}}$. It is easy to check that in this case $q(\epsilon)$ in the statement of Lemma 5.1 satisfies

$$q(\epsilon) \geq \left(\frac{1}{\epsilon}\right)^{c'(\log 1/\epsilon)^{\lfloor \log k \rfloor}}, \tag{26}$$

for an appropriate constant $c' = c'(d)$. By Lemma 5.1 we get a k-graph, which is ϵ-far from being induced D-free, and contains only $O(n^d/q(\epsilon))$ copies of D. By Lemma 5.2 the query complexity of any one-sided error property tester for $\mathcal{P}^*_D$ is lower bounded by $q(\epsilon)^{1/d}$, which is (26), with c' replaced by c'/d. ■

Note, that though the statement of Theorem 2 states the improved lower bounds only for k-graphs on at least $r(k)$ vertices, it should be clear that the same lower bound also applies to any k-graph on less than $r(k)$ vertices such that either the k-graph or its complement spans a copy of F^k. This, in particular, applies to F^k itself, thus, as mentioned after the statement of Theorem 2, we indeed get an improvement on the lower bound for testing $\mathcal{P}^*_{F^k}$ from [18]. It is worth mentioning that if one is willing to replace $\lfloor \log k \rfloor$ with $\lfloor \log \lceil k/2+1 \rceil \rfloor$ in the statement of Theorem 2, then one can obtain this slightly weaker lower bound for **any** k-graph on at least $k+1$ vertices, instead of k-graphs on at least $r(k)$ vertices. One just has to note that for any set of $k+1$ vertices, either the k-graph or its complement spans at least $\lceil k/2+1 \rceil$ edges. One then proceeds as in the proof of Theorem 2 by taking a set Z, which is $(\lceil k/2+1 \rceil, d^{3d^2})$-gadget-free instead of $(k+1, d^{3d^2})$-gadget-free.

7 More on Linear Equations and Extremal Hypergraphs

In this section we prove Theorem 4. Analogously to our proof technique for $\mathcal{P}^*_D$, the first step in the proof of Theorem 4 is to show that given a hyper-cycle $D = (V, E)$ on d vertices we can construct a k-graph that contains many copies of D such that from each copy of D we can infer a certain linear equation. The main idea, as in Lemma 4.1, is to give an algebraic construction of such a graph, but as we explain below, in this case we have some additional difficulties.

Let m be an integer, $Z \subseteq [m]$ and D a hyper-cycle of size d, whose vertices are numbered $\{1, \ldots, d\}$ as in the definition of a hyper-cycle. We define the following k-graph $F = F(D, Z)$ as follows: the vertex set of F consists of d pairwise disjoint sets of vertices $V_1, \ldots, V_d$, where, with a slight abuse of notation, we think of each of these sets as being the set of integers $1, \ldots, d^{k+1}m$. We define the edge set of F by specifying the edge sets of $|Z|^k$ copies of D that we put in F. For every set of (not necessarily distinct) integers $z_0, \ldots, z_{k-1} \in Z$, we define a copy of D denoted $C = C(z_0, \ldots, z_{k-1})$: As the vertex set of C, we choose d vertices $v_1 \in V_1, \ldots, v_d \in V_d$, where for $1 \leq t \leq d$ we choose $v_t = E(z_0, \ldots, z_{k-1}, t)$, and E is the function defined in (19). Note, that for any choice of $z_0, \ldots, z_{k-1} \in Z$ we have $E(z_0, \ldots, z_{k-1}, t) \in [d^{k+1}m]$, thus the vertices "fit" into the sets $V_1, \ldots, V_d$. As for the edges of C, we simply regard the vertices $v_1 \in V_1, \ldots, v_d \in V_d$ as if they were the vertices $1, \ldots, d$ in D, namely, if $(t_1, \ldots, t_k) \in E(D)$, we put in F an edge that contains the vertices

$$E(z_0, \ldots, z_{k-1}, t_1) \in V_{t_1},\ E(z_0, \ldots, z_{k-1}, t_2) \in V_{t_2},\ \ldots\ , E(z_0, \ldots, z_{k-1}, t_k) \in V_{t_k}.$$

The main technical step in this section is the proof of the following lemma, whose role in the proof of Theorem 4 is analogous to the role of Lemma 4.1 in the proof of Theorems 1 and 2.

Lemma 7.1 *Let m be an arbitrary integer, $Z \subseteq [m]$ and D a hyper-cycle on d vertices. Construct $F = F(D, Z)$ as above. Suppose $v_1 \in V_1, \ldots, v_d \in V_d$ span a copy of D, with v_t playing the role of vertex t in D. Suppose that for $1 \leq i \leq d-k+2$ edge e_i belongs to $C(z_{0,i}, \ldots, z_{k-1,i})$. Then, for every $1 \leq j \leq k-1$ there are* **positive** *integers $a_1, \ldots, a_{d-k+1} \leq c = c(d)$ such that*

$$a_1 \cdot z_{j,1} + a_2 \cdot z_{j,2} + \ldots + a_{d-k+1} \cdot z_{j,d-k+1} = (a_1 + a_2 + \ldots + a_{d-k+1}) \cdot z_{j,d-k+2}$$

In order to apply Lemma 7.1, we need another notion of linear equations suitable for it, which we formulate as follows: a set $Z \subseteq [m] = \{1, 2, \ldots, m\}$ is called *(k, h)-linear-free* if for every k **positive** integers $a_1, \ldots, a_k \leq h$, the only solution of the equation

$$a_1 z_1 + \ldots + a_k z_k = (a_1 + \ldots + a_k) z_{k+1}, \tag{27}$$

which uses $k+1$ integers from Z satisfies $z_1 = \ldots = z_{k+1}$. That is, whenever $a_1, \ldots, a_k \leq h$, the only solution to (27) from Z, is one of the $|Z|$ trivial solutions. Similar to our proof technique of Theorems 1 and 2, in this case we will also need a dense (k, h)-linear-free sets of integers, with which we will apply Lemma 7.1.

The main difficulty in proving Lemma 7.1 is two fold; While we still have to show that we can extract a linear combination of the integers, as we did in Lemma 4.1, we are faced with the following problem; suppose we manage to extract a linear equation but it is of the form $z_1 + z_2 - z_3 = z_4$. In Lemma 4.1 this was not an issue, as in that lemma the required equation only relates 3 integers, thus if we get an equation of the form, say, $3a - 2b = c$, we can simply "shift" $2b$ to the other side and get the required equation. This is not possible in our case. The problem is even more serious; as we mentioned above (and analogously to our proof technique for $\mathcal{P}^*_D$), our ultimate goal will be to apply Lemma 7.1 with a (k, h)-linear-free set of size $m^{1-o(1)}$. However, it follows from the pigeon-hole principle that the largest size of a subset of $[m]$ without solutions to $z_1 + z_2 - z_3 = z_4$ is $O(\sqrt{m})$. Thus, we must make sure that all the coefficients in the linear equation we extract are positive. One may also ask, why we cannot prove our lower bounds for $\mathcal{P}_D$ by using only linear equations with 3 unknowns, like we use for $\mathcal{P}^*_D$. The main reason for that is that for $\mathcal{P}^*_D$ we can prove a lower bound either for D or its complement, and one of them must contain a copy of T^k. For $\mathcal{P}_D$, however, we cannot use this reasoning and have to deal with D itself, which does not necessarily contain a copy of T^k.

The proof of Theorem 4 will follow by using Lemma 7.1 together with arguments similar to those used in the proofs of Lemmas 4.2, 5.1 and 5.2. The proof Lemma 7.1 appears in the following subsection, and the proof of Theorem 4 appears in Subsection 7.2.

7.1 Proof of Lemma 7.1

For the proof of Lemma 7.1 we need the following simple observation:

Claim 7.1 *For a given set of $p \leq r$ distinct integers $t_1, \ldots, t_p$ bounded by r, let A be the matrix $(A)_{i,j} = t_i^{p+1-j} - (t_i - 1)^{p+1-j}$ $(1 \leq i, j \leq p)$. Then, there is a non-zero integer vector v, all of whose entries are bounded (in absolute value) by r^{2r}, such that for $1 \leq i \leq p-1$ $(Av)_i = 0$, while $(Av)_p > 0$.*

Proof: As the integers $t_1, \ldots, t_p$ are distinct, the Vandermonde matrix $(V)_{i,j} = t_i^{j-1}$ is invertible. As A can be obtained from V by rank preserving operations, A is also invertible. By Claim 4.3 there is a non-zero integer vector v, all of whose entries are bounded by $(r^2 p)^{p/2} \leq (rp)^p \leq r^{2r}$, such that for $1 \leq i \leq p-1$ we have $(Av)_i = 0$. As A is invertible and v is non zero, it cannot be the case that $(Av)_p = 0$, and if $(Av)_p < 0$ we can replace v by $-v$. ■

As a first step towards the proof of Lemma 7.1 we prove the following claim.

Claim 7.2 *Let m be an arbitrary integer, $Z \subseteq [m]$ and D a hyper-cycle on d vertices. Construct $F = F(D, Z)$ as in Lemma 7.1, and denote by $\overline{i}$ the vector $(i, i^2, \ldots, i^{k-1})$ and by $\overline{z_i}$ the vector $(z_{1,i}, z_{2,i}, \ldots, z_{k-1,i})$. Then the following equation holds*

$$\overline{z_1} \cdot (\overline{2} - \overline{1}) + \ldots + \overline{z_{d-k+1}} \cdot (\overline{d-k+2} - \overline{d-k+1}) = \overline{z_{d-k+2}} \cdot (\overline{d-k+2} - \overline{1}). \tag{28}$$

Also, for every $1 \le i \le d-k+1$ and $i+1 \le t \le i+k-2$ the following equation holds

$$\overline{z_{i+1}} \cdot (\overline{t+1} - \overline{t}) - \overline{z_i} \cdot (\overline{t+1} - \overline{t}) = 0 \tag{29}$$

Proof: Let $v_1 \in V_1, \ldots, v_d \in V_d$ be d vertices spanning a copy of D, with $v_i \in V_i$ playing the role of vertex i in D. For every $1 \le i \le d-k+1$ consider the vertices $v_i \in V_i$ and $v_{i+1} \in V_{i+1}$ and recall that by the definition of a hyper-cycle they belong to $e_i \in C(z_{0,i}, \ldots, z_{k-1,i})$. If we regard v_i and v_{i+1} as integers we get from the definition of F that $v_i = E(z_{0,i}, \ldots, z_{k-1,i}, i)$ and that $v_{i+1} = E(z_{0,i}, \ldots, z_{k-1,i}, i+1)$. From the definition of E in (19) this means that (note that $z_{0,i}$ disappears)

$$v_{i+1} - v_i = z_{1,i} \cdot ((i+1) - i) + z_{2,i} \cdot ((i+1)^2 - i^2) + \ldots + z_{k-1,i} \cdot ((i+1)^{k-1} - i^{k-1}). \tag{30}$$

As in the statement of the claim, let the vector $\overline{i}$ denote $(i, i^2, \ldots, i^{k-1})$ and let the vector $\overline{z_i}$ denote $(z_{1,i}, z_{2,i}, \ldots, z_{k-1,i})$. Therefore, we can write for every $1 \le i \le d-k+1$ the vector equation

$$v_{i+1} - v_i = \overline{z_i} \cdot (\overline{i+1} - \overline{i}). \tag{31}$$

As e_{d-k+2} contains the vertices $v_{d-k+2} \in V_{d-k+2}, \ldots, v_d \in V_d$ we have for every $d-k+2 \le i \le d-1$

$$v_{i+1} - v_i = \overline{z_{d-k+2}} \cdot (\overline{i+1} - \overline{i}). \tag{32}$$

Summing (31) for $1 \le i \le d-k+1$ and (32) for $d-k+2 \le i \le d-1$ we obtain

$$v_d - v_1 = \overline{z_1} \cdot (\overline{2} - \overline{1}) + \ldots + \overline{z_{d-k+1}} \cdot (\overline{d-k+2} - \overline{d-k+1}) + \overline{z_{d-k+2}} \cdot (\overline{d} - \overline{d-k+2}). \tag{33}$$

As e_{d-k+2} contains the vertices $v_1 \in V_1$ and $v_d \in V_d$, we also have by the same reasoning

$$v_d - v_1 = \overline{z_{d-k+2}} \cdot (\overline{d} - \overline{1}). \tag{34}$$

Combining (33) and (34) we obtain (28).

In order to obtain the other equations, for any $1 \le i \le d-k+1$ consider edge e_i and recall that it contains the vertices $i, \ldots, i+k-1$. Note that for every $i+1 \le t \le i+k-2$ vertices t and $t+1$ belong to both e_i and e_{i+1}. By the same reasoning used to obtain (30) and (31) we can write for every $i+1 \le t \le i+k-2$

$$v_{t+1} - v_t = \overline{z_i} \cdot (\overline{t+1} - \overline{t}) \tag{35}$$

$$v_{t+1} - v_t = \overline{z_{i+1}} \cdot (\overline{t+1} - \overline{t}) \tag{36}$$

Combining these equations we get (29) for every $i+1 \le t \le i+k-2$, thus completing the proof. ■

For the rest of the proof let us use the following notation: for every $i+1 \le t \le i+k-2$ denote by $\mathcal{E}_{i,t}$ the equation of (29). Note, that for every $1 \le i \le d-k+1$ we have $k-2$ equations $\mathcal{E}_{i,t}$. To illustrate the main ideas of the proof the reader may want to consider the special case where $d = 6$ and $k = 4$ depicted in Figure 1.

We also need the following claim. For its proof, the reader may find it useful to refer to the example given in Figure 1.

$$\begin{aligned}
z_{1,1} + 3z_{2,1} + 7z_{3,1} + z_{1,2} + 5z_{2,2} + 19z_{3,2} + z_{1,3} + 7z_{2,3} + 37z_{3,3} &= 3z_{1,4} + 15z_{2,4} + 63z_{3,4}\\
z_{1,1} + 5z_{2,1} + 19z_{3,1} - z_{1,2} - 5z_{2,2} - 19z_{3,2} &= 0\\
z_{1,1} + 7z_{2,1} + 37z_{3,1} - z_{1,2} - 7z_{2,2} - 37z_{3,2} &= 0\\
z_{1,2} + 7z_{2,2} + 37z_{3,2} - z_{1,3} - 7z_{2,3} - 37z_{3,3} &= 0\\
z_{1,2} + 9z_{2,2} + 61z_{3,2} - z_{1,3} - 9z_{2,3} - 61z_{3,3} &= 0\\
z_{1,3} + 9z_{2,3} + 61z_{3,3} &= z_{1,4} + 9z_{2,4} + 61z_{3,4}\\
z_{1,3} + 11z_{2,3} + 191z_{3,3} &= z_{1,4} + 11z_{2,4} + 191z_{3,4}
\end{aligned}$$

Figure 1: The linear equations (28), $\mathcal{E}_{1,2}, \mathcal{E}_{1,3}, \mathcal{E}_{2,3}, \mathcal{E}_{2,4}, \mathcal{E}_{3,4}, \mathcal{E}_{3,5}$ when $d = 6$ and $k = 4$.

Claim 7.3 *For every $1 \leq i \leq d - k + 1$ there is a linear combination of (28) and equations $\mathcal{E}_{i,i+1}, \ldots, \mathcal{E}_{i,i+k-2}$ with integer coefficients, in which the coefficient of $z_{i,1}$ is positive, while the coefficients of $z_{i,2}, \ldots, z_{i,k-1}$ vanish.*

Proof: Let $\alpha_1, \ldots, \alpha_{k-1}$ denote the unknown coefficients of (28) and $\mathcal{E}_{i,i+1}, \ldots, \mathcal{E}_{i,i+k-2}$, respectively, in the linear combination, which we seek. Suppose we write $k - 1$ linear equations $e_1, \ldots, e_{k-1}$ in unknowns $\alpha_1, \ldots, \alpha_{k-1}$, where equation e_i requires the coefficient of $z_{i,1}$ to vanish in the linear combination of (28), $\mathcal{E}_{i,i+1}, \ldots, \mathcal{E}_{i,i+k-2}$ with coefficient $\alpha_1, \ldots, \alpha_{k-1}$. Observing the coefficients of $z_{1,i}, \ldots, z_{k-1,i}$ in (28) and in $\mathcal{E}_{i,i+1}, \ldots, \mathcal{E}_{i,i+k-2}$ it is easy to see that the entries of the $(k-1) \times (k-1)$ matrix A, whose i^{th} row contains equation e_i satisfies the properties of Claim 7.1. We can now take the entries of the vector v, whose exitance is guaranteed by Claim 7.1, to be the required integer coefficients $\alpha_1, \ldots, \alpha_{k-1}$. ■

Proof of Lemma 7.1: We first observe that (28) is an equation in $z_{j,i}$, where for $1 \leq j \leq k-1$ and $1 \leq i \leq d-k+1$ we have $z_{j,i}$ in the left hand side of the equation, and for every $1 \leq j \leq k-1$ we have $z_{j,d-k+2}$ on the right hand side. Furthermore, all the coefficients in this equation are positive. Finally, for every $1 \leq j \leq k-1$ the sum of the coefficients of $z_{j,1}, \ldots, z_{j,d-k+1}$ is equal to $(d-k+2)^j - 1$, which is precisely the coefficient of $z_{j,d-k+2}$. It thus follows that (28) is the **sum** of the $k - 1$ equations that we need to obtain in order to prove the lemma. In order to simplify the notation we now turn to show how to obtain the linear equation relating $z_{1,1}, \ldots, z_{1,d-k+2}$. The other cases are completely identical.

To simplify the rest of the proof, when we later refer to *fixing* i we mean obtaining a linear equation in which $z_{2,i}, \ldots, z_{k-1,i}$ do not appear, while the coefficient of $z_{1,i}$ is positive. Our main idea of extracting from (28) the required linear equation relating $z_{1,1}, \ldots, z_{1,d-k+2}$ is the following: For $1 \leq i \leq d-k+2$, equation (28) contains the variables $z_{1,i}, \ldots, z_{k-1,i}$, while we want an equation in which only $z_{1,i}$ appears. We thus need to fix i for every $1 \leq i \leq d-k+2$. By Claim 7.3 we know that for every $1 \leq i \leq d-k+1$ we can find a linear combination of (28) and $\mathcal{E}_{i,i+1}, \ldots, \mathcal{E}_{i,i+k-2}$, which fixes i. The main problem is that we need a linear combination which simultaneously fixes all the is. Suppose we first use Claim 7.3 in order to obtain a new equation, denoted $\mathcal{E}$, which fixes $i = 1$. We would now want to reapply Claim 7.3 in order to fix $i = 2$. The only difficulty is that we would now want to take a linear combination of $\mathcal{E}_{2,3}, \ldots, \mathcal{E}_{2,k}$ with $\mathcal{E}$, and not with (28) as taking a linear combination with (28) might "bring back" $z_{2,1}, \ldots, z_{k-1,1}$.

However, it is easy to see that we can also find a linear combination of $\mathcal{E}_{2,3}, \ldots, \mathcal{E}_{2,k}$ and $\mathcal{E}$, which fixes $i = 2$. By Claim 7.3, we know that there is a linear combination of $\mathcal{E}_{2,3}, \ldots, \mathcal{E}_{2,k}$ and (28), which fixes $i = 2$. Consider now the coefficients of $z_{1,2}, \ldots, z_{k-1,2}$ in equations (28), $\mathcal{E}_{1,2}, \ldots, \mathcal{E}_{1,k-1}$ and $\mathcal{E}_{2,3}, \ldots, \mathcal{E}_{2,k}$. Note, that the coefficients, which appear in equations (28), $\mathcal{E}_{1,2}, \ldots, \mathcal{E}_{2,k-2}$ also appear in equations (28), $\mathcal{E}_{2,3}, \ldots, \mathcal{E}_{2,k}$. To be more precise, the coefficients of $z_{1,2}, \ldots, z_{k-1,2}$ in equation $\mathcal{E}_{1,2}$ are precisely the coefficients of $z_{1,2}, \ldots, z_{k-1,2}$ in (28) and for every $3 \leq i \leq k-1$ the coefficients of $z_{1,2}, \ldots, z_{k-1,2}$ in equation $\mathcal{E}_{1,i}$ are precisely the coefficients of $z_{1,2}, \ldots, z_{k-1,2}$ in equation $\mathcal{E}_{2,i-1}$. Thus, as $\mathcal{E}$ is a linear combination of (28) and $\mathcal{E}_{1,2}, \ldots, \mathcal{E}_{2,k-1}$ we infer that if there is a linear combination of (28) and $\mathcal{E}_{2,3}, \ldots, \mathcal{E}_{2,k}$, which fixes $i = 2$, then there must be such a linear combination of $\mathcal{E}$ and $\mathcal{E}_{2,3}, \ldots, \mathcal{E}_{2,k}$. It is finally important to note that as $z_{1,1}, \ldots, z_{1,k-1}$ do not appear in equations $\mathcal{E}_{2,3}, \ldots, \mathcal{E}_{2,k}$ then in the new linear equation $i = 1$ remains fixed.

Note that the above argument can be generalized to any $2 \leq i \leq d-k+1$ as equations $\mathcal{E}_{i,i+1}, \ldots, \mathcal{E}_{i,i+k-2}$ do not contain the unknowns $z_{1,p}, \ldots, z_{k-1,p}$ for any $p < i$, and the coefficients of $z_{i,1}, \ldots, z_{i,k-1}$ appearing in (28) and $\mathcal{E}_{i-1,i}, \ldots, \mathcal{E}_{i-1,i+k-3}$ also appear in equations (28) and $\mathcal{E}_{i,i+1}, \ldots, \mathcal{E}_{i,i+k-2}$. Hence, we can apply an iterative procedure, where in the i^{th} step we add to (28) an appropriate linear combination of equations $\mathcal{E}_{i,i+1}, \ldots, \mathcal{E}_{i,i+k-2}$, which fixes i. Moreover, later iterations of this procedure will not change the coefficients of $z_{1,p}, \ldots, z_{k-1,p}$ for any $p < i$. In particular, if in iteration p we fixed $i = p$, then we will also have this property at the end of the process. We have thus established that for $1 \leq i \leq d-k+1$ our process obtains in iteration i a linear combination in which for every $p \leq i$ the coefficient of $z_{1,p}$ is positive, while the coefficients of $z_{2,p}, \ldots, z_{k-1,p}$ have vanished. We now observe that as both in (28) and (29) the coefficient of $z_{i,d-k+2}$ is equal to the sum of the coefficients of $z_{i,1}, \ldots, z_{i,d-k-1}$, it must be the case that after iteration $d-k+1$ the coefficient of $z_{1,d-k+2}$ is positive while the coefficients of $z_{2,d-k+2}, \ldots, z_{k-1,d-k+2}$ have vanished, thus $i = d-k+2$ is also fixed. This means that we have obtained the required equation relating $z_{1,1}, z_{1,2}, \ldots, z_{1,d-k+2}$. As for the size of the integers in this linear equation, note that the coefficients of (28) and (29) are bounded by $d^k \leq d^d$. As we apply the above iterative process $d-k+1 < d$ times, Claim 7.1 guarantees that when we are done the coefficients are bounded by a function of d only. ∎

Corollary 7.1 *For every d, there is $c = c(d)$ such that if we construct the k graph F in Lemma 7.1 with a $(d-k+2, c)$-linear-free set Z, then F contains precisely $|Z|^d$ copies of D spanned by vertices $v_1 \in V_1, \ldots, v_d \in V_d$, with v_t playing the role of vertex t in D.*

Proof: The main idea is simply to show that the only such copies of D belong to the same copy of D defined for some choice of integers $z_0, \ldots, z_{k-1} \in Z$. Consider any copy of D spanned by vertices $v_1 \in V_1, \ldots, v_d \in V_d$, with v_t playing the role of vertex t in D. Suppose for every $1 \leq i \leq d-k+2$ edge e_i of D belongs to $C(z_{0,i}, \ldots, z_{k-1,i})$. Lemma 7.1 guarantees that for every $1 \leq j \leq k-1$ there are positive integers $a_1, \ldots, a_{d-k+1} \leq c = c(d)$ such that the the following equation is satisfied

$$a_1 \cdot z_{j,1} + a_2 \cdot z_{j,2} + \ldots + a_{d-k+1} \cdot z_{j,d-k+1} = (a_1 + a_2 + \ldots + a_{d-k+1}) \cdot z_{j,d-k+2}.$$

Therefore, if we use a set Z, which is $(d-k+2, c)$-linear-free it must be the case that for every $1 \leq j \leq k-1$, we have $z_{j,1} = z_{j,2} = \ldots = z_{j,d-k+2}$. To complete the proof we just have to show that we also have $z_{0,1} = z_{0,2} = \ldots = z_{0,d-k+2}$ as this will imply that all the edges of D belong to the same copy defined using $z_{0,1}, z_{1,1}, \ldots, z_{k-1,1}$. To show this we observe that for every $2 \leq t \leq d-k+2$, vertex $v_t \in V_t$ is common to both e_{t-1} and e_t. This means that

$$E(z_{0,t-1}, \ldots, z_{k-1,t-1}, t) = v_t = E(z_{0,t}, \ldots, z_{k-1,t}, t).$$

As we already know that for every $1 \leq j \leq k-1$ we have $z_{j,i} = z_{j,i+1}$, the above equation implies that $z_{0,t-1} = z_{0,t}$ holds for every $2 \leq t \leq d-k+2$, thus completing the proof. ∎

7.2 Proof of Theorem 4

Given Lemma 7.1, the proof of Theorem 4 follows by going along the lines of the proofs of Lemmas 4.2 and 5.1, with one key difference, which we shall explain. In order to avoid repeating the same arguments we will just sketch them, while assuming that the reader is familiar with the proofs of Lemmas 4.2 and 5.1. As in Lemma 5.1, we will also need a large set of integers that does not satisfy linear equations similar to the one we extract by using Lemma 7.1. We will thus need the following:

Lemma 7.2 *For every k and h there is $c = c(k,h)$, such that for every n, there is a (k,h)-linear-free subset $Z \subset [n] = \{1, 2, \ldots, n\}$ of size at least*

$$|Z| \geq \frac{m}{e^{c\sqrt{\log m}}} \tag{37}$$

By using Behrend's technique [7], this lemma has been proved in [5] and [9] for the case of $k = 3$ and arbitrary h, and in [1] for the case $h = 1$ and arbitrary k. As the proof of the above lemma simply combines the ideas of [1] and [5], we do not include it here.

Proof of Theorem 4: (sketch) To further simplify the proof, we will use c to indicate (possibly distinct) constants that depend only on d. Let D be a fixed k-graph on d vertices, whose core L, contains a hyper-cycle R, of size r ($\leq d$). Denote by ℓ ($\leq d$) the size of L and assume we rename its vertices such that a copy of R is spanned by the first r vertices of L, with every vertex $1 \leq i \leq r$ playing the role of vertex i of R. As in the proof of Theorem 1, the main idea is to apply Lemma 5.2 by constructing a k-graph H that is ϵ-far from satisfying $\mathcal{P}_D$, and contains only $n^d/q(\epsilon)$ copies of D, with $q(\epsilon) \geq (1/\epsilon)^{c \log 1/\epsilon}$. To this end, we will first construct a k-graph F (as in Lemma 4.2), and then take an appropriate blow-up of it (as in Lemma 5.1).

Given ϵ, let m be the largest integer satisfying

$$e^{c\sqrt{\log m}} < 1/\epsilon. \tag{38}$$

It is easy to see that

$$m > (1/\epsilon)^{c \log 1/\epsilon}. \tag{39}$$

Let Z be a $(r-k+2, c)$-linear-free subset of $[m]$. Note, that by Lemma 7.2 we have

$$|Z| \geq \frac{m}{e^{c\sqrt{\log m}}}$$

Define a k-graph F as follows: It has ℓ clusters of vertices $V_1, \ldots, V_\ell$ of size $d^{\ell+2}m$ each (thus, F has $\ell d^{\ell+2}m$ vertices). For each set of k integers $z_0, \ldots, z_{k-1} \in Z$ we put a copy of L in F spanned by the vertices $v_1 \in V_1, \ldots, v_\ell \in V_\ell$ with v_i playing the role of i, and $v_i = E(z_0, \ldots, z_{k-1}, i)$, with the function E define in (19) (note, that the vertices fit into the sets $V_1, \ldots, V_\ell$). As in Lemma 4.2 item (2), one can easily show that we have thus defined

$$|Z|^k \geq \frac{m^k}{e^{c\sqrt{\log m}}}$$

copies of L, with each pair sharing at most $k-1$ vertices. In particular these copies are edge disjoint. It will also be important for the rest of the proof to note that the subgraph of F, which is spanned by the first r vertices, is precisely the k-graph defined in Lemma 7.1 (with R being the hyper-cycle D in the statement of the lemma). We thus get by Corollary 7.1 that if we took an $(r-k+2, c)$-linear-free set Z, with a sufficiently small c (in terms of d), then there are $|Z|^r$ choices of vertices $v_1 \in V_1, \ldots, v_r \in V_r$ such that $v_1, \ldots, v_r$ span a copy of R with v_t playing the role of vertex t or R. In what follows we call such copies of R *nice*.

Suppose we construct an n vertex k-graph H, by taking an $n/(\ell d^{\ell+2}m)$ blow-up of F (recall that F has $\ell d^{\ell+2}m$ vertices). By repeating the argument of Claim 5.1, it is not difficult to see that as F contains at least $m^k/e^{c\sqrt{\log m}}$ edge disjoint copies of L, we may infer that H contains at least $n^k/e^{c\sqrt{\log m}}$ edge-disjoint copies of L. By our choice of m in (38) we get that H is ϵ-far from being L-free. It can be easily shown that as L is the core of D, in this case H is also ϵ-far from being D-free.

We are thus left with showing that H contains relatively few copies of D. By repeating the argument of Claim 5.3, it can be shown that as F spans at most $|Z|^k$ nice copies of R, then H spans at most

$$|Z|^k \left(\frac{n}{\ell d^{\ell+2}m}\right)^r \leq m^k \left(\frac{n}{\ell d^{\ell+2}m}\right)^r = O(n^r/m)$$

nice copies of R (observe that we always have $r > k$). Assume we prove that every copy of D spanned by H contains a nice copy of R. It would thus follow that as each copy of R is contained in at most $\binom{n}{d-r} \leq n^{d-r}$ copies of D, that H spans at most $O(n^d/m)$ copies of D. By (39) we would get the required upper bound on the number of copies of D spanned by H.

We thus only have to show that every copy of D spanned by H contains a nice copy of R. Given a copy of D in H, consider the following homomorphism $\varphi : V(D) \mapsto V(L)$: suppose $v \in V(D)$ is one of the vertices (in H) of the independent set that replaced vertex $i' \in V_i$, then we map v to i. Note that this is indeed a mapping from $V(D)$ to $V(L)$. Also, note that if $(i_1, \ldots, i_k) \notin E(L)$ then in F there are no edges connecting vertices of $V_{i_1}, \ldots, V_{i_k}$. As H is a blow-up we infer that φ is indeed a homomorphism. As L is by definition a subgraph of D, φ induces a homomorphism φ', from L to itself. By the minimality of L (recall Definition 2.1), we may infer that φ' is in fact an automorphism, that is $(i_1, \ldots, i_k) \in E(L) \Leftrightarrow (\varphi'(i_1), \ldots, \varphi'(i_k)) \in E(L)$. This means that φ' maps some copy of $R \subset D$ to the subgraph of L spanned by vertices $1, \ldots, r$. Finally, by our definition of φ this means that this is a nice copy of R. ■

8 Proof of Theorem 3

We start this section with the proof of Theorem 3 part (i). To this end, we need the following well known lemma of Erdős and Simonovits.

Lemma 8.1 ([10]) *For every t and k, there are constants $n_0 = n_0(t,k)$, $c = c(t,k)$ and $\gamma = \gamma(t,k) > 0$ with the following properties: For every $t_1,\ldots,t_k \leq t$, every k-graph on at least n_0 vertices, which contains $\delta(n) \cdot n^k > n^{k-\gamma}$ edges, contains at least $c\delta(n)^{t^*} n^t$ copies of $K_{t_1,\ldots,t_k}$, where $t^* = t_1 \cdot \ldots \cdot t_k$ and $t = t_1 + \ldots + t_k$.*

We comment that the proof of this lemma is described in [10] for the case $t_1 = \ldots = t_k$. However, simple modifications of the argument give the above lemma. Observe, that a k-graph, which is ϵ-far from being D-free, where $D = K_{t_1,\ldots,t_k}$, must contain at least $\epsilon n^k \gg n^{k-\gamma}$ edges. From the above lemma we infer that such a k-graph must contain $c\epsilon^{t^*} n^{\bar{t}}$ copies of K. Hence, as observed in [18], there is a one-sided error property-tester for $\mathcal{P}_D$ that simply samples $O((1/\epsilon)^{t^*})$ sets of $\bar{t}$ vertices, and accepts iff it finds no copy of D. By the above claim it finds a copy of D with high probability. As we now show, we can improve this simple upper bound and show a lower bound, which is nearly tight in many cases.

Proof of Theorem 3, part (i): Let c and n_0 be the constants of Lemma 8.1. Given an input k-graph H on $n > n_0$ vertices, the algorithm samples $10(t_1 + \ldots + t_k)/(c\epsilon^{t^*/t_k})$ vertices and declares H to be D-free iff it finds no copy of D in the subgraph spanned by the set of vertices. Clearly, if H is D-free, the algorithm accepts H with probability 1. So assume H is ϵ-far from being D-free. We wish to show that with high probability the set of vertices spans a copy of D. Recall that such a graph must contain at least ϵn^k edges.

For a vertex v denote by $\delta(v)$ the degree of v, namely, the number of edges of H to which v belongs. For a vertex v in H denote by $H(v)$ the following $(k-1)$-graph: we take all the edges to which v belongs and remove v from them. Note that the number of edges of $H(v)$ is precisely $\delta(v)$, and that $H(v)$ obviously has at most n vertices. It follows from Lemma 8.1, that for some fixed $\gamma > 0$, if $\delta(v) > n^{k-1-\gamma}$, then $H(v)$ contains at least

$$c\left(\frac{\delta(v)}{n^{k-1}}\right)^{t^*/t_k} n^t$$

copies of the $(k-1)$-partite $(k-1)$-graph $K_{t_1,\ldots,t_{k-1}}$, where $t = t_1 + \ldots + t_{k-1}$ and $t^* = t_1 \cdot \ldots \cdot t_k$. On the other hand, if $\delta(v) < n^{k-1-\gamma}$, then it might be the case that $H(v)$ contains no copies of $K_{t_1,\ldots,t_{k-1}}$ at all. In any case, however, $H(v)$ contains at least

$$c\left(\left(\frac{\delta(v)}{n^{k-1}}\right)^{t^*/t_k} - \left(\frac{1}{n^\gamma}\right)^{t^*/t_k}\right) n^t \tag{40}$$

copies of $K_{t_1,\ldots,t_{k-1}}$. Hence, all vertices v belong to at least this many copies of the k-partite k-graph $K = K_{t_1,\ldots,t_{k-1},1}$, where v plays the role of the single vertex in the last vertex class of K. Suppose we sample t vertices uniformly at random from H. Let X_v be an indicator random variable for the event that these vertices form a copy of K along with vertex v, such that v plays the role of the single vertex in the last vertex class of K. By (40),

$$Prob[X_v = 1] \geq \max\left(0, c\left(\frac{\delta(v)}{n^{k-1}}\right)^{t^*/t_k} - c\left(\frac{1}{n^\gamma}\right)^{t^*/t_k}\right).$$

Define $X = \sum_v X_v$. The expectation of $|X|$ thus satisfies

$$E(|X|) = \sum_v Prob[X_v = 1] \geq c\sum_v \left(\frac{\delta(v)}{n^{k-1}}\right)^{t^*/t_k} - c\sum_v n^{-\gamma t^*/t_k} \geq$$

$$cn\left(\frac{\sum_v \delta(v)}{n^k}\right)^{t^*/t_k} - cn^{1-\gamma t^*/t_k} \geq cn(k\epsilon)^{t^*/t_k} - o(n) \geq 2cn\epsilon^{t^*/t_k},$$

where in the second inequality we have applied Jensen's inequality to the first summation, and in the third we have used the fact that H must contain at least ϵn^k edges. Observing that $|X| \leq n$, we conclude that

$$2cn\epsilon^{t^*/t_k} \leq E(|X|) \leq cn\epsilon^{t^*/t_k} + n \cdot Prob[|X| \geq cn\epsilon^{t^*/t_k}].$$

Therefore,

$$Prob[|X| \geq cn\epsilon^{t^*/t_k}] \geq \epsilon^{t^*/t_k}.$$

Hence, by Markov's inequality, after sampling $10/\epsilon^{t^*/t_k}$ sets of t vertices, with probability at least $9/10$ we find at least one set of t vertices, which forms a copy of K with at least $cn\epsilon^{t^*/t_k}$ of the vertices of H. After finding this set of t vertices, all we need is t_k vertices that form a copy of K with this set of vertices, as together they would form a copy of D. By assumption, there are at least $cn\epsilon^{t^*/t_k}$ vertices that form a copy of K with the set of t vertices. By Markov's inequality, after sampling $10t_k/(c\epsilon^{t^*/t_k})$ vertices, with probability at least $9/10$ we find the required set of t_k vertices. In total, we sampled $10(t_1 + \ldots + t_k)/(c\epsilon^{t^*/t_k})$ vertices, as needed. ■

Proof of Theorem 3, part (ii): As in the proof of Lemma 5.2, we use Lemma 5.3 in order to assume, without loss of generality that a one sided error property tester for $\mathcal{P}_D$ samples non-uniformly a set of vertices, queries all edges in the sample, and then proceeds to execute some computation.

Consider the random k-graph $H(n, 2k^k\epsilon)$, that is, a k-graph on n vertices, where each set of k vertices forms an edge randomly and independently with probability $2k^k\epsilon$. The expected number of edges in H is obviously $2k^k\epsilon\binom{n}{k} \geq 2\epsilon n^k$, hence, by a standard Chernoff bound, the number of edges in H is at least $\frac{3}{2}\epsilon n^k$ with probability at least $3/4$ (in fact, the probability is $1 - 2^{-\Theta(n^k)}$ but we do not need this stronger estimate here). As by Lemma 8.1, every k-graph with $\frac{\epsilon}{2}n^k$ edges contains a copy of D, we get that with probability at least $3/4$ H is ϵ-far from being D-free.

Fix a set of $d = t_1 + \ldots + t_k$ vertices. The number of ways to partition this set into k subsets of size k is at most $d!$. The probability that any of these partitions spans a copy of D is at most $\binom{t^*}{|E|}(2k^k\epsilon)^{|E|}$, where $t^* = t_1 \cdot \ldots \cdot t_k$. Therefore, the expected number of copies of D in $H(n, 2k^k\epsilon)$ is at most

$$\binom{n}{d} d! \binom{t^*}{|E|} 2k^k\epsilon^{|E|} \leq n^d(t^*2k^k\epsilon)^{|E|}.$$

By Markov's inequality, the probability that the number of copies of D is 4 times its expectation is at most $1/4$. We conclude that there is a k-graph, which is both ϵ-far from being D-free, and yet contains less than

$$4n^d/(1/t^*2k^k\epsilon)^{|E|}$$

copies of D. By Lemma 5.2, the query complexity of a one-sided error property tester for $\mathcal{P}_D$ is $\Omega((1/\epsilon)^{|E|/d})$. ■

9 Concluding Remarks and Open Problems

- The most interesting problem related to this paper is to give a complete characterization of the k-graphs D for which $\mathcal{P}_D$ is easily testable. We believe that the techniques presented in this paper should be useful in resolving this problem. It is known that for $k = 2$, $\mathcal{P}_D$ is easily testable iff D is bipartite. It seems likely that the "right" characterization is that for larger k, $\mathcal{P}_D$ is easily testable iff D is k-partite. Using Theorem 1, we can rule out the possibility of extending the characterization of $k = 2$ to, "$\mathcal{P}_D$ is easily testable iff D is 2-colorable". Indeed, note that for $k > 2$, F^k, the complete k-graph on $k+1$-vertices, is 2-colorable. On the other hand, as $\mathcal{P}_{F^k}$ is equivalent to $\mathcal{P}^*_{F^k}$, we get from Theorem 1 that $\mathcal{P}_{F^k}$ is not easily testable.

- In light of Theorem 1 one may hope to show that the only k-graphs D, for which $\mathcal{P}^*_D$ is easily testable are the single k-edges. This, however, is false. As shown in [5], in case D is a path of length 2, $\mathcal{P}^*_D$ has a one-sided error tester, whose query complexity is $O(\log(1/\epsilon)/\epsilon)$. It would thus be interesting to decide if $D_{3,2}$ (see Theorem 1) is easily testable.

- It would also be very interesting to improve the lower bounds obtained in Theorem 2. It should be noted that using our techniques, one cannot obtain lower bounds that match the current upper bounds. For example, the best known upper bound for testing $\mathcal{P}^*_D$, for D being a triangle, has query complexity that is a tower of exponents of height polynomial in $1/\epsilon$. As is evident from the statement of Lemma 5.1, in order to prove a matching lower bound using our methods, one would have to use an $(3, h)$-gadget-free subset of the first m integers of size $\Omega(m/\log^* m)$ (and observe that such a set contains no 3-term arithmetic progressions). However, by a result of Bourgain [8], every subset of the first m integers of size $\Omega(m/\sqrt{\log m/\log\log m})$ contains a 3-term arithmetic progression. Thus, the best lower bound one might hope to prove using these techniques is roughly $2^{\log(1/\epsilon)/\epsilon^2}$, which is very far from the current upper bound. Also, any one sided error property-tester for $\mathcal{P}_{K_3} = \mathcal{P}^*_{K_3}$ with query complexity $2^{O((1/\epsilon)^2)}$ would imply an improvement of Bourgain's result.

References

[1] N. Alon, Testing subgraphs in large graphs, Random Structures and Algorithms 21 (2002), 359-370. Also, Proc. 42^{nd} IEEE FOCS, IEEE (2001), 434-441.

[2] N. Alon, R. A. Duke, H. Lefmann, V. Rödl and R. Yuster, The algorithmic aspects of the Regularity Lemma, J. of Algorithms 16 (1994), 80-109. Also, Proc. 33^{rd} IEEE FOCS, Pittsburgh, IEEE (1992), 473-481.

[3] N. Alon, E. Fischer, M. Krivelevich and M. Szegedy, Efficient testing of large graphs, Combinatorica 20 (2000), 451-476. Also, Proc. of 40^{th} FOCS, New York, NY, IEEE (1999), 656–666.

[4] N. Alon and A. Shapira, Testing subgraphs in directed graphs, JCSS 69/3 (2004), 354-382. Also, Proc. of the 35^{th} STOC, 2003, 700–709.

[5] N. Alon and A. Shapira, A characterization of easily testable induced subgraphs, Proc. of the 15^{th} Annual ACM-SIAM SODA, ACM Press (2004), 935-944. Also, Combinatorics, Probability and Computing, to appear.

[6] N. Alon and A. Shapira, On an extremal hypergraph problem of Brown, Erdős and Sós, Combinatorica, to appear.

[7] F. A. Behrend, On sets of integers which contain no three terms in arithmetic progression, *Proc. National Academy of Sciences USA* 32 (1946), 331–332.

[8] J. Bourgain, On triples in arithmetic progression. Geom. Funct. Anal. 9 (1999) 968–984.

[9] P. Erdős, P. Frankl and V. Rödl, The asymptotic number of graphs not containing a fixed subgraph and a problem for hypergraphs having no exponent, Graphs Combin. 2 (1986) 113-121.

[10] P. Erdős and M. Simonovits, Supersaturated graphs and hypergraphs, Combinatorica 3 (1983), 181-192, 29.

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

[12] P. Frankl and V. Rödl, Extremal problems on set systems, Random Struct. Algorithms 20 (2002), no. 2, 131-164.

[13] O. Goldreich, Combinatorial property testing - a survey, In: Randomization Methods in Algorithm Design (P. Pardalos, S. Rajasekaran and J. Rolim eds.), AMS-DIMACS (1998), 45-60.

[14] O. Goldreich, S. Goldwasser and D. Ron, Property testing and its connection to learning and approximation, JACM 45(4): 653-750 (1998).

[15] O. Goldreich and L. Trevisan, Three theorems regarding testing graph properties, Random Structures and Algorithms, 23(1):23-57, 2003.

[16] I. S. Gradshteyn and I. M. Ryzhik, *Tables of Integrals, Series, and Products*, 6th Edition, Academic Press (2000), p. 1110.

[17] W. T. Gowers, Hypergraph regularity and the multidimensional Szemerédi theorem, manuscript.

[18] Y. Kohayakawa, B. Nagle and V. Rödl, Efficient testing of hypergraphs, Proc. of 29^{th} ICALP, (2002), 1017–1028.

[19] I. Laba and M. Lacey, On sets of integers not containing long arithmetic progressions, unpublished manuscript. Available at http://arxiv.org/abs/math.CO/0108155.

[20] B. Nagle and V. Rödl, Regularity properties for triple systems, Random Structures and Algorithms 23 (2003), 3, 264-332.

[21] B. Nagle, V. Rödl and M. Schacht, The counting lemma for regular k-uniform hypegraphs, manuscript.

[22] R. A. Rankin, Sets of integers containing not more than a given number of terms in arithmetical progression. Proc. Roy. Soc. Edinburgh Sect. A, 65:332-344, 1962.

[23] V. Rödl and J. Skokan, Regularity lemma for k-uniform hypergraphs, Random Structures and Algorithms, 25 (2004), 1, 1-42.

[24] D. Ron, Property testing, in: P. M. Pardalos, S. Rajasekaran, J. Reif and J. D. P. Rolim, editors, *Handbook of Randomized Computing*, Vol. II, Kluwer Academic Publishers, 2001, 597–649.

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

[26] A. Shapira, Behrend-type constructions for sets of linear equations, submitted.

[27] E. Szemerédi, Regular partitions of graphs, In: *Proc. Colloque Inter. CNRS* (J. C. Bermond, J. C. Fournier, M. Las Vergnas and D. Sotteau, eds.), 1978, 399–401.