The Raymond and Beverly Sackler
Faculty of Exact Sciences
School of Mathematical Sciences

Efficient Removal Lemmas and Related Problems

**Thesis submitted in partial fulfillment of the requirements
for the degree "Doctor of Philosophy" in mathematics**

by

Lior Gishboliner

**Under the supervision of
Prof. Asaf Shapira**

June 2020

Acknowledgments

First and foremost, I would like to express my deepest gratitude to my advisor, Prof. Asaf Shapira, for patiently guiding me in the various stages of my studies, from an enthusiastic (though, admittedly, mostly clueless) master's student to a young and confident researcher. I look up to you both as a mathematician and as a human being, and have taken example from you in various ways. I feel grateful for the way you have treated me in our many years of mutual work.

Second, I would like to thank the professors in the combinatorics department of Tel Aviv University: Noga Alon, Michael Krivelevich, Asaf Shapira and Wojciech Samotij, for many invaluable classes and for their kind treatment of me through the years.

Third, I am indebted to my fellow graduate students and postdocs in Tel Aviv Univeristy: Omri Ben-Eliezer, Gal Kronenberg, Eoin Long, Peleg Michaeli, Yinon Spinka, Henrique Stagni and Misha Tyomkyn, for many useful and stimulating discussions and for their generous help in many situations.

Finally, I would like to thank my parents for their constant love and support.

Abstract

This thesis studies quantitative aspects of graph property testing and removal lemmas, as well as several related questions. The general problem is as follows: given that a graph G violates a given graph property $\mathcal{P}$ (say, induced C_4-freeness) in a strong way – in the sense that many edges need to be added to/deleted from G in order to turn it into a graph which satisfies $\mathcal{P}$ – how abundant are subgraphs of G which violate $\mathcal{P}$? Or, more concretely, how many vertices of G should we sample so that with high probability we come across a witness to the fact that G violates $\mathcal{P}$? Such local-vs.-global problems lie at the heart of extremal graph theory, and are fundamental to the field of graph property testing. Our main results are as follows.

- We obtain new general sufficient and necessary conditions for a graph-family $\mathcal{F}$ to have a removal lemma with polynomial dependence between the parameters. These results prove in a unified way (almost) all previously known results of this type, as well as many new ones. As a consequence of our method, we establish a conjecture of Alon stating that every *semi-algebraic graph property* has a removal lemma with polynomial bounds.
- We initiate the study of efficient removal lemmas for tournaments, and give a characterization of the tournaments F for which the F-removal-lemma has polynomial bounds.
- We prove an exponential upper bound for the induced-C_4 removal lemma, thus greatly improving the previous tower-type bound, and making progress on a problem raised by Alon and Fox [8] and by Conlon and Fox [35].
- We find a surprising way of exploiting generalized Turán results to prove complexity-theoretic theorems in graph property testing. In particular, we prove a hierarchy theorem for the one-sided-error query-complexity of monotone graph properties, and show a separation between the one-sided-error and two-sided-error query-complexity of such properties. To this end, we prove tight bounds for the generalized Turán function of pairs of cycles.
- We study a *distribution-free* model of testing dense graphs which was recently introduced by Goldreich [58]. We provide a complete characterization of the graph properties testable in this model, thus answering in a strong form an open question raised in [58].
- We study the testability of graph properties definable by a linear inequality involving subgraph densities. Disproving a conjecture of Goldreich and Shinkar [62], we show that there are properties of this form which are not testable with a number of queries independent of the size of the input.
- We obtain the first asymptotic improvement over a 15-year-old bound of Sárközy and Selkow [99] for the Brown–Erdős–Sós problem.

Contents

Chapter 1

Background and Overview of Results

The triangle removal lemma is one of the cornerstones of modern extremal graph theory. This result, established in the seminal paper of Ruzsa and Szemerédi [98], asserts that if at least εn^2 edges must be deleted from an n-vertex graph G in order to destroy all of its triangles, then the number of triangles in G is at least δn^3 for some $\delta = \delta(\varepsilon) > 0$. While perhaps innocent-looking at first glance, this statement has profound and deep implications. Its original purpose was to prove the $(6,3)$-conjecture of Brown, Erdös and Sós [26, 27], which asserted that an n-vertex 3-uniform hypergraph with $\Omega(n^2)$ edges must contain a so-called $(6,3)$-configuration, i.e. a set of 3 edges on most 6 vertices. In the same paper [98], Ruzsa and Szemerédi also established an intriguing connection between the triangle removal lemma and problems in additive combinatorics, showing that the former can be used to derive Roth's theorem [95] on sets avoiding 3-term arithmetic progressions, and using Behrend's construction [20] of dense progression-free sets in order to construct n-vertex 3-uniform hypergraphs with $n^{2-o(1)}$ edges and no $(6,3)$-configuration, thus giving a nearly-tight matching lower bound for the $(6,3)$-conjecture. This link between extremal graph theory and additive combinatorics has since had a major influence on the development of both fields, attesting yet again to the historical importance of the removal lemma.

In the last few decades, the triangle removal lemma has been generalized in a number of groundbreaking ways. First, it was extended to the setting of induced subgraphs by the *induced removal lemma* of Alon, Fischer, Krivelevich and Szegedy [5]. This result states that if an n-vertex graph G is such that one must add/delete at least εn^2 edges of G in order to destroy all induced[1] copies of some fixed graph H, then G contains at least $\delta n^{v(H)}$ induced copies of H, where $\delta = \delta_H(\varepsilon) > 0$. This line of research culminated in the remarkably general result of Alon and Shapira [10], stating that *every* hereditary graph property admits a removal lemma. This result is sometimes known as the *infinite removal lemma*. Recall that a graph property $\mathcal{P}$ is hereditary if it is closed under taking induced subgraphs. Equivalently, $\mathcal{P}$ is hereditary if and only if there is a family $\mathcal{F} = \mathcal{F}(\mathcal{P})$ of *minimal forbidden induced subgraphs* for $\mathcal{P}$, such that a graph satisfies $\mathcal{P}$ if and only if it is induced F-free for every $F \in \mathcal{F}$. An n-vertex graph G is *ε-far* from $\mathcal{P}$ if one must add/delete at least εn^2 edges of G in order to make it satisfy $\mathcal{P}$. The Alon–Shapira theorem then states that for every hereditary property $\mathcal{P}$ and $\varepsilon > 0$ there exist $\delta = \delta_{\mathcal{P}}(\varepsilon) > 0$ and $M = M_{\mathcal{P}}(\varepsilon)$ such that if an n-vertex graph G is ε-far from $\mathcal{P}$ then there exists $F \in \mathcal{F}$ with $v(F) \leq M$ such that G contains at least $\delta n^{v(F)}$ induced copies of F. It is not hard to see that this result has the following equivalent formulation:

[1] The analogous statement for not-necessarily-induced copies, which is usually known as the graph removal lemma, is a straightforward extension of the triangle removal lemma.

for every hereditary property $\mathcal{P}$ there exists $w = w_{\mathcal{P}}(\varepsilon)$ such that if a graph G is ε-far from satisfying $\mathcal{P}$, then a random sequence of w vertices of G, sampled uniformly and independently, induces a graph which does not satisfy $\mathcal{P}$ with probability at least $\frac{2}{3}$ (say). This second formulation will be more convenient for us. For further information about removal lemmas for graphs, we refer the reader to the survey [35].

A second major generalization of the triangle removal lemma was its extension to hypergraphs of higher uniformity in the *hypergraph removal lemma*, proven independently by Gowers [67] and by Nagle, Rödl, Schacht and Skokan [86, 89, 90] (see also [112]). This fundamental result has found numerous applications in combinatorics and number theory, see [92] and the references therein. Subsequent to the proof of the hypergraph removal lemma, an *infinite hypergraph removal lemma* was established by Rödl and Schacht [93] and by Austin and Tao [17], thus generalizing the aforementioned result of [10].

Removal-type statements are usually proven using *regularity lemmas*, which, rougly speaking, are theorems asserting that any graph (or hypergraph) can be broken into a bounded number of pieces, almost all of which "behave randomly". The first statement of this type, namely the celebrated regularity lemma of Szemerédi [110], was the tool used in the original proof of the triangle removal lemma. The induced removal lemma requires the more sophisticated *strong regularity lemma* [5], while the hypergraph removal lemma required a generalization of Szemerédi's regularity lemma to higher uniformities [67, 89].

A major open problem concerning the triangle removal lemma is to understand the "correct" dependence of δ on ε. As mentioned above, the original proof of the triangle removal lemma [98] relied on Szemerédi's regularity lemma [110]. As a result, the lower bound on $\delta = \delta(\varepsilon)$ supplied by this proof is of the form $1/\delta \leq \text{tower}(\text{poly}(1/\varepsilon))$, where tower is the tower of exponents function, defined as $\text{tower}(0) = 1$ and $\text{tower}(i) = 2^{\text{tower}(i-1)}$. Thanks to the work of Gowers see [66], it is known this bound on $\delta(\varepsilon)$ cannot be improved by an argument which uses Szemerédi's regularity lemma. Fox [49] managed to improve the aforementioned bound by giving a new proof of the triangle removal lemma, obtaining the current best known bound of $1/\delta \leq \text{tower}(\log(1/\varepsilon))$ (see also [85] for a different proof). Strikingly, the best known upper bound on $\delta = \delta(\varepsilon)$ is only slightly superpolynomial, namely $1/\delta \geq (1/\varepsilon)^{\Omega(1/\varepsilon)}$ (see [98]).

This limited understanding of the dependence of δ on ε persists (sometimes even in greater capacity) in the other removal-type statements mentioned above. For the induced removal lemma, the original proof of [5] supplied a bound of the form $1/\delta_H(\varepsilon) \leq \text{wowzer}(\text{poly}(1/\varepsilon))$, where the wowzer function is one level above the tower function in the Ackermann hierarchy, namely $\text{wowzer}(i) = \text{tower}(\text{wowzer}(i-1))$. This bound was subsequently improved to $1/\delta_H(\varepsilon) \leq \text{tower}(\text{poly}(1/\varepsilon))$ by Conlon and Fox [34]. As for the infinite removal lemma of Alon and Shapira [10], it is known (see [12]) that there is no uniform bound on $w_{\mathcal{P}}(\varepsilon)$ which works for every hereditary property. Namely, for every function $f : (0,1) \to \mathbb{N}$ there is a hereditary property $\mathcal{P}$ such that $w_{\mathcal{P}}(\varepsilon) \geq f(\varepsilon)$. Still, one might expect that at least for natural graph properties, the aforementioned tower-type bounds could be improved. This remains a major open problem in the area.

In recent years, several prominent researchers [56, 8] raised the problem of characterizing the hereditary properties $\mathcal{P}$ for which $w_{\mathcal{P}}(\varepsilon)$ is polynomial in $1/\varepsilon$. A considerable portion of this thesis is dedicated to making progress on this goal. Before stating our results, we first describe the connection between graph removal lemmas and the area of computer science known as graph property testing.

Property testing is concerned with the study of very fast, randomized, algorithms, called testers, whose goal is to distinguish, with some high enough probability (say, 2/3), between objects satisfying some fixed property $\mathcal{P}$ and those that are ε-*far* from satisfying it. Here, ε-far means that an ε-fraction of the input

object should be modified in order to obtain an object satisfying $\mathcal{P}$. The study of such problems originated in the seminal papers of Rubinfeld and Sudan [96], Blum, Luby and Rubinfeld [21], and Goldreich, Goldwasser and Ron [59]. For a thorough introduction to property testing, we refer the reader to the book of Goldreich [57].

In the *dense graph model* of property testing, which is the concrete setting we consider here, the input graph G is given as an $n \times n$ adjacency matrix, and the algorithm may query the matrix to find out if a certain pair of vertices are adjacent or not. Recall that G is said to be *ε-far* from a property $\mathcal{P}$ if one must add/delete at least εn^2 edges of G in order to turn it into a graph satisfying $\mathcal{P}$ (note that adding/deleting edges corresponds to changing entries in the adjacency matrix of G). We now give the precise definition of testability of graph properties.

Definition 1. *A graph property $\mathcal{P}$ is* testable *if there is an algorithm $\mathcal{T}$ which works as follows. The input given to $\mathcal{T}$ consists of a graph G and a proximity parameter $\varepsilon > 0$. Upon receiving the input, $\mathcal{T}$ samples a sequence of $f = f_{\mathcal{P}}(\varepsilon)$ vertices uniformly and independently, and queries the graph on all pairs of sampled vertices. The algorithm then either accepts or rejects; it must accept with probability at least $\frac{2}{3}$ if G satisfies $\mathcal{P}$, and reject with probability at least $\frac{2}{3}$ if G is ε-far from $\mathcal{P}$. The algorithm's decision is based solely on the isomorphism class of the subgraph $G[S]$ induced by the sample S. The algorithm $\mathcal{T}$ is called a* tester *for $\mathcal{P}$.*

The quantity $f_{\mathcal{P}}(\varepsilon)$ is called the *sample complexity* (or *query complexity*) of the tester $\mathcal{T}$. Note, crucially, that $f_{\mathcal{P}}(\varepsilon)$ is assumed to be *independent of the size of the input graph* (which we denote by n). I.e., in the setting we consider in this thesis, the sample complexity may depend only on ε and the given property.

The reader may wonder why in Definition 1 we only allow for algorithms which operate in a very specific manner. The answer is that, as it turns out, we do not lose generality by posing such restrictions; indeed, it was shown by Goldreich and Trevisan [63] that any tester can be converted to a tester which operates as described in Definition 1 at the price of a very minor loss in query complexity (which will be insignificant for the problems we consider here).

We say that a tester $\mathcal{T}$ has *one-sided error* if $\mathcal{T}$ accepts with probability 1 whenever G satisfies $\mathcal{P}$. Otherwise, $\mathcal{T}$ has *two-sided error*. It is not hard to see that if $\mathcal{T}$ is a one-sided error tester for a hereditary property $\mathcal{P}$, then $\mathcal{T}$ must always accept if the subgraph $G[S]$ induced by the sample seen by the algorithm satisfies $\mathcal{P}$. It follows that if $\mathcal{P}$ admits a one-sided error tester with sample complexity $f_{\mathcal{P}}(\varepsilon)$, then for every graph G which satisfies $\mathcal{P}$, a random sample $f_{\mathcal{P}}(\varepsilon)$ vertices of G induces a subgraph which does not satisfy $\mathcal{P}$ with probability at least 2/3. Therefore, the minimal sample complexity of a one-sided-error tester for a hereditary property $\mathcal{P}$ exactly equals the minimal integer $w = w_{\mathcal{P}}(\varepsilon)$ satisfying the conclusion of the infinite removal lemma, stated above. Thus, the notion of testability generalizes the study of removal lemmas. We now give the precise definition of $w_{\mathcal{P}}(\varepsilon)$, as it will be used throughout this thesis.

Definition 2. *For a hereditary property $\mathcal{P}$ and $\varepsilon > 0$, we define $w_{\mathcal{P}}(\varepsilon)$ to be the minimal integer $w > 0$ such that whenever a graph G is ε-far from $\mathcal{P}$, a sample of w vertices of G, taken uniformly and independently, induces a graph which does not satisfy $\mathcal{P}$ with probability at least $\frac{2}{3}$.*

1.1 Efficient Removal Lemmas for Graphs and Tournaments (Chapters 2 and 3)

As mentioned above, a central open problem in the field of graph property testing is characterizing the hereditary properties $\mathcal{P}$ which satisfy $w_{\mathcal{P}}(\varepsilon) = \text{poly}(1/\varepsilon)$; or, in other words[2], the hereditary properties which have a one-sided-error tester whose sample complexity is (only) polynomial in the proximity parameter. We therefore introduce the following definition.

Definition 3. *A hereditary property $\mathcal{P}$ is* easily testable *if $w_{\mathcal{P}}(\varepsilon) = \text{poly}(1/\varepsilon)$, and* hard *otherwise.*

In this section we describe our main results in the pursuit of a characterization of the easily testable graph properties, as well as similar results in the setting of tournaments. The problem of characterizing the easily testable graph properties was first raised by Goldreich [56], and later also by Alon and Fox [8]. Several partial answers to this problem have appeared in the literature. For example, Goldreich, Goldwasser and Ron [59] have shown that k-colorability and, more generally, any so-called *partition property*, is easily testable[3]. A famous theorem of Alon [2] states that the property of being (not necessarily induced) H-free (for a given graph H) is easily testable if and only if H is a bipartite graph. Alon and Shapira [9] proved a similar result for *induced* H-freeness; they showed that induced H-freeness is easily testable if $H \in \{P_2, P_3, \overline{P_2}, \overline{P_3}\}$ (where P_k denotes the path with k vertices, and $\overline{F}$ is the complement of F), and hard if $H \notin \{P_2, P_3, \overline{P_2}, \overline{P_3}, P_4, C_4, \overline{C_4}\}$ (where C_k denotes the cycle of length k). This left open the cases of P_4 and C_4. The case of $H = P_4$ was settled recently by Alon and Fox [8].

Our first set of results in this context gives very simple yet general combinatorial sufficient and necessary conditions for a hereditary property to be easily testable. Our results establish in a unified manner (almost) all previously known results of this type (in particular, the results of [2, 9, 8] mentioned in the previous paragraph), as well as many new ones. In particular, we obtain polynomially bounded removal lemmas for many natural graph properties for which it was not previously known how to obtain a removal lemma without using the regularity lemma.

From this point on, it will be more natural to think of a hereditary property in terms of its forbidden subgraphs. Given a family of graphs $\mathcal{F}$, let $\mathcal{P}^*_{\mathcal{F}}$ be the property[4] of being induced $\mathcal{F}$-free, i.e. not containing an induced copy of any of the graphs of $\mathcal{F}$. When $\mathcal{F}$ consists of a single graph F we will use the notation $\mathcal{P}^*_F$. Recall that a co-bipartite graph is a graph whose complement is bipartite, and a split graph is a graph whose vertex set can be partitioned into two parts: one spanning a clique and the other an independent set. Our main results are as follows.

Theorem 4. *If $\mathcal{F}$ is a finite family of graphs that contains a bipartite graph, a co-bipartite graph and a split graph then $\mathcal{P}^*_{\mathcal{F}}$ is easily testable.*

Theorem 5. *Let $\mathcal{F}$ be a finite family for which $\mathcal{P}^*_{\mathcal{F}}$ is easily testable. Then $\mathcal{F}$ contains a bipartite graph and a co-bipartite graph.*

[2] We will use the removal lemma language and property testing language interchangeably.

[3] This result of [59] dramatically improved upon a previous tower-type bound due to Rödl and Duke [91] on the query-complexity of testing k-colorability.

[4] Note that a property can be represented as $\mathcal{P}^*_{\mathcal{F}}$ (for some graph-family $\mathcal{F}$) if and only if it is hereditary.

We also establish a conjecture of Alon [3], stating that every *semi-algebraic* graph property is easily testable. Roughly speaking, a property is semi-algebraic if it can be defined by semi-algebraic relations (i.e., polynomial inequalities) in real variables. For the precise definitions, see Chapter 2.

Theorem 6. *Every semi-algebraic graph property is easily testable.*

As mentioned above, the 4-cycle C_4 is the only graph H for which it is not known whether induced H-freeness is easily testable. In particular, C_4 satisfies neither the sufficient condition given by Theorem 4, nor the necessary condition given by Theorem 5. The problem of deciding whether induced C_4-freeness is easily testable has been raised by several authors, such as Alon and Fox [8] and Conlon and Fox [35]. It seems that C_4 constitutes the first major obstacle to obtaining a characterization of the easily testable graph properties. In fact, we believe that resolving the C_4-problem would allow one to make significant progress on the problem of characterizing the *finite* families $\mathcal{F}$ for which $\mathcal{P}^*_{\mathcal{F}}$ is easily testable. In the following result we make significant progress on the C_4-problem, showing that $w_{\mathcal{P}^*_{C_4}}(\varepsilon) \leq 2^{\text{poly}(1/\varepsilon)}$. The previous best bound was of tower-type. We conjecture that $\mathcal{P}^*_{C_4}$ is easily testable.

Theorem 7. *If an n-vertex graph G is ε-far from being induced C_4-free, then G contains at least $n^4/2^{\text{poly}(1/\varepsilon)}$ induced copies of C_4.*

An interesting aspect of Theorem 7 is that it gives a fairly efficient removal lemma for a property that is satisfied by graphs which have (only) extremely inefficient regular partitions. Let us elaborate on this point. Recall that Szemerédi's regularity lemma [110] states that for every $\varepsilon > 0$, every graph has an *ε-regular equipartition* of its vertex-set into at most $M(\varepsilon)$ parts, where an equipartition is ε-regular if all but an ε-fraction of the pairs of its parts induce "random-like" bipartite graphs (we will not define this notion precisely, but instead refer the reader to [94] for an overview of graph regularity). In applications, the regularity lemma becomes stronger the smaller the size of the partition is. Unfortunately, the proof of the regularity lemma [110] only gave a tower-type bound of $M(\varepsilon) \leq \text{tower}(\text{poly}(1/\varepsilon))$, and we know, thanks to the celebrated work of Gowers [66], that this dependence is unavoidable in general graphs. For special graph classes, however, one can often guarantee much smaller regular partitions. It turns out that virtually all[5] of the properties $\mathcal{P}$ which are known to be easily testable are such that every graph satisfying $\mathcal{P}$ has an ε-partition *with only* $\text{poly}(1/\varepsilon)$ *parts*. In fact, this approach (of proving that a property $\mathcal{P}$ is easily testable by utilizing the fact that graphs satisfying $\mathcal{P}$ have polynomial-sized regular partitions) is used explicitly in our proofs of Theorems 4 and 6.

On the negative side, there are natural graph properties $\mathcal{P}$ for which there exist graphs satisfying $\mathcal{P}$ that only have ε-regular partitions of size $\text{tower}(\text{poly}(1/\varepsilon))$. In fact, $\mathcal{P}$ is of this type whenever $\mathcal{P}$ contains all bipartite graphs or all co-bipartite graphs or all split graphs. The reason is that one can take a bipartite version of Gowers' construction [66], and then put cliques on some of the sides to obtain a bipartite/co-bipartite/split graph which still only has ε-regular partitions of size $\text{tower}(\text{poly}(1/\varepsilon))$. Prime examples of such properties $\mathcal{P}$ are triangle-freeness (as every bipartite graph is triangle-free) and induced C_4-freeness (as every split graph is induced C_4-free). Recall that the best bounds for the triangle removal lemma [49, 85] are still tower-type. For induced C_4-freeness, however, Theorem 7 gives a much better bound. Therefore, Theorem 7 can be thought of as the first example showing that one *can* obtain an efficient

[5] An exception to this rule is the result of [59] that partition properties (such as k-colorability) are easily testable.

removal lemma for a ("subgraph-freeness") property $\mathcal{P}$ despite the fact that graphs satisfying $\mathcal{P}$ might have only regular partitions of tower-type size.

Next we consider similar problems in the setting of tournaments. Here, the distance of a tournament from a tournament property is defined similarly, with the key difference being that one is only allowed to reverse the direction of edges (and not delete edges). It is known that a removal-type statement holds in this setting as well: for every oriented graph F and $\varepsilon > 0$ there is $\delta = \delta_F(\varepsilon)$ such that if an n-vertex tournament T is ε-far from being F-free (i.e., if one must reverse at least εn^2 edges in order to destroy all copies of F in T), then T contains at least $\delta n^{v(F)}$ copies of F. An oriented graph F is called *easy* if $\delta_F(\varepsilon) = \text{poly}(\varepsilon)$, and *hard* otherwise.

Theorem 8. *An oriented graph F is easy if and only if $V(F)$ can be partitioned into two sets, each spanning an acyclic directed graph.*

We also address the (complexity-theoretic) problem of deciding whether a given oriented graph F is easy. An oriented graph F is called *k-colorable* if there is a partition $V(F) = X_1 \cup \cdots \cup X_k$ such that $F[X_i]$ is an acyclic digraph for every $1 \leq i \leq k$. Theorem 8 can then be restated as saying that F is easy if and only if F is 2-colorable. It is natural to ask if the characterization given in Theorem 8 is "efficient", that is, how hard it is to decide if an oriented graph F is easy. It follows from the work of Bokal et al. [22] that this task is in fact NP-hard. The following theorem strengthens the result of Bokal et al. [22] by showing that the problem is hard even in the case that F is a tournament.

Theorem 9. *For every $k \geq 2$, the problem of deciding if a tournament is k-colorable is NP-hard.*

References: The results of these chapters appeared as:

- L. Gishboliner and A. Shapira. Removal Lemmas with Polynomial Bounds. International Math Research Notices (IMRN), to appear. Also in Proceedings of STOC 2017, 510-522.
- J. Fox, L. Gishboliner, A. Shapira and R. Yuster. The Removal Lemma for Tournaments. Journal of Combinatorial Theory Ser. B 136 (2019), 110-134.
- L. Gishboliner and A. Shapira. Efficient Removal without Efficient Regularity. Combinatorica 39 (2019), 639-658. Also in Proceedings of ITCS 2018, 1-15.

1.2 Various Results on Graph Property Testing

1.2.1 A Query-Complexity Hierarchy Theorem via a Generalized Turán Result (Chapter 4)

We establish several complexity-theoretic results regarding one- and two-sided error testing of hereditary graph properties. Our first result shows that for every (decreasing) function $f : (0, 1) \to \mathbb{N}$, there is a hereditary property whose one-sided-error sample complexity essentially equals $f(\varepsilon)$. This can be considered a *hierarchy theorem* for the sample complexity of one-sided-error testers, somewhat reminiscent of the famous time/space hierarchy theorems in computational complexity theory. A special case of this result partially resolves a problem raised by Goldreich [56, 57], of exhibiting a natural property whose (two-sided-error) sample complexity is exponential in $1/\varepsilon$.

Theorem 10. *For every decreasing function $f\colon (0,1) \to \mathbb{N}$ with $f(x) \geq 1/x$, there is a monotone[6] graph property $\mathcal{P}$ satisfying $f(\varepsilon) \leq w_{\mathcal{P}}(\varepsilon) \leq \mathrm{poly}(1/\varepsilon) \cdot f(\Omega(\varepsilon))$.*

Our second result shows that two-sided-error testers can be *arbitrarily stronger* than one-sided-error ones, even for hereditary properties. More precisely, we show that there are hereditary properties whose two-sided-error sample complexity is (only) polynomial in $1/\varepsilon$, while their one-sided-error sample complexity can be arbitrarily large.

Theorem 11. *For every decreasing function $f\colon (0,1) \to \mathbb{N}$ satisfying $f(x) \geq 1/x$, there is a monotone graph property $\mathcal{P}$ so that*

- *$\mathcal{P}$ has one-sided-error sample complexity $w_{\mathcal{P}}(\varepsilon) \geq f(\varepsilon)$;*
- *For every $\varepsilon > 0$ there is $n_0(\varepsilon)$ such that $\mathcal{P}$ has a (two-sided-error) tester whose sample complexity is $\mathrm{poly}(1/\varepsilon)$ when invoked with input graphs that have at least $n_0(\varepsilon)$ vertices.*

Prior to this work, it was not even known that two-sided-error testers can be super-polynomially stronger than one-sided-error testers.

Let us give another perspective on Theorem 11. Observe that a property $\mathcal{P}$ can be ε-tested (possibly with two-sided error) using query complexity $q = q(\varepsilon)$ if and only if the distribution of induced subgraphs on q vertices obtained by drawing q vertices from a graph in $\mathcal{P}$ is distinguishable[7] from the distribution obtained by drawing these vertices from a graph that is ε-far from $\mathcal{P}$. So Theorem 11 implies that there is a monotone graph property $\mathcal{P}$ and a graph G that is ε-far from $\mathcal{P}$ so that even though *almost all* (in fact, *all*) subsets of vertices of G of size (say) $2^{1/\varepsilon}$ do satisfy $\mathcal{P}$, the distribution of induced subgraphs on $\mathrm{poly}(1/\varepsilon)$ vertices drawn from G is distinguishable from the one drawn from a graph satisfying $\mathcal{P}$. In other words, we can detect that G does not satisfy $\mathcal{P}$ without actually finding a proof of this fact.

Theorems 10 and 11 are proven using a *generalized Turán result* for cycles, which is of independent interest. For graphs T, H and an integer n, let $\mathrm{ex}(n, T, H)$ denote the maximum number of copies of T in an n-vertex H-free graph. The systematic study of the function $\mathrm{ex}(n, T, H)$ was initiated by Alon and Shikhelman [15], and has since received considerable attention (see Chapter 4 for further discussion and references). Our main result is the determination of the order of magnitude[8] of $\mathrm{ex}(n, C_k, C_\ell)$ for *all* $k, \ell \geq 4$.

Theorem 12. *For distinct $k, \ell \geq 4$, we have*

$$\mathrm{ex}(n, C_k, C_\ell) = \begin{cases} \Theta_k\big(n^{k/2}\big) & k \geq 5, \ell = 4, \\ \Theta_k\big(\ell^{\lceil k/2 \rceil} n^{\lfloor k/2 \rfloor}\big) & \ell \geq 6 \text{ even}, k \geq 4, \\ \Theta_k\big(\ell^{\lceil k/2 \rceil} n^{\lfloor k/2 \rfloor}\big) & k, \ell \text{ odd}, 5 \leq k < \ell. \end{cases}$$

A careful reader may notice that the statement of Theorem 12 does not in fact cover every possible choice of distinct $k, \ell \geq 4$. The reason is that in all remaining cases, one (trivially) has $\mathrm{ex}(n, C_k, C_\ell) = \Theta_k(n^k)$, since in these cases a blowup of C_k is C_ℓ-free (see Section 4.1 for more details). We note that the special case of Theorem 12 in which k, ℓ are even was independently established in [54]. As a biproduct of proving Theorem 12, we also obtain tight bounds for $\mathrm{ex}(n, P_k, C_\ell)$, $k \geq 2$, where P_k is the path with k edges.

[6] A graph property is *monotone* if it is closed under the removal of vertices and edges. In particular, every monotone property is hereditary.

[7] This interpretation is reminiscent of the way one studies limits of dense and sparse graph sequences, see [79].

[8] The notation $O_k/\Omega_k/\Theta_k$ used in Theorem 12 (and elsewhere in Chapter 4) means that the implicit multiplicative constant depends on k. We will write $O/\Omega/\Theta$ to mean that the implicit constant is absolute.

References: The results of this chapter appeared as:

- L. Gishboliner and A. Shapira. A Generalized Turan Problem and its Applications. International Math Research Notices (IMRN) 11 (2020), 3417-3452. Also in Proceedings of STOC 2018, 760-772.

1.2.2 Testing Graphs against an Unknown Distribution (Chapter 5)

The definition of testability as stated in Definition 1 assumes that one can uniformly sample entries of the input (or, in our setting, vertices of the input graph). In *distribution-free* testing one assumes that the input is endowed with some *arbitrary and unknown* distribution $\mathcal{D}$, which also affects the way one defines the distance to satisfying a property. As discussed in [58], one motivation for this model is that it can handle settings in which one cannot produce uniformly distributed entries from the input. Another motivation is that the distribution $\mathcal{D}$ can assign higher weight/importance to parts of the input which we want to have higher impact on the distance to satisfying the given property. Until very recently, problems of this type were studied almost exclusively in the setting of testing properties of functions, see [29, 30, 38, 55, 71].

Here we study the *vertex-distribution-free dense graph model* (VDF model, for short), which was recently introduced by Goldreich [58]. In this model, the input to the algorithm is a pair $(G, \mathcal{D})$, where G is a graph and $\mathcal{D}$ is some *arbitrary and unknown* distribution on $V(G)$. For a pair of graphs G_1, G_2 on the same vertex-set V and a distribution $\mathcal{D}$ on V, the (edit) distance between G_1 and G_2 with respect to $\mathcal{D}$ is defined as $\text{dist}_{\mathcal{D}}(G_1, G_2) = \sum_{\{x,y\} \in E(G_1) \triangle E(G_2)} \mathcal{D}(x)\mathcal{D}(y)$. We say that $(G, \mathcal{D})$ is ε-far from satisfying a graph property $\mathcal{P}$ if for every $G' \in \mathcal{P}$, the distance between G and G' with respect to $\mathcal{D}$ is at least ε. Note that if $\mathcal{D}$ is the uniform distribution, then these definitions boil down to the usual definition of distance/farness, as described in the beginning of Chapter 1.

The definition of testability in the VDF model is similar to that given in Definition 1, but with two crucial differences: first, the vertices sampled by the algorithm are distributed according to $\mathcal{D}$ (to be precise, the algorithm is given access to a device that produces random vertices of G distributed according to $\mathcal{D}$); and second, the distance of G to the given property $\mathcal{P}$ is defined with respect to $\mathcal{D}$; that is, the tester is required to reject with probability at least $\frac{2}{3}$ if $(G, \mathcal{D})$ is ε-far from $\mathcal{P}$ (and accept with probability at least $\frac{2}{3}$ if G satisfies $\mathcal{P}$). We note that in the VDF model, the algorithm does not receive $|V(G)|$ as part of the input (while in the usual "uniform" model, one sometimes assumes that it does). For further discussion of the subtleties of the VDF model, we refer the reader to [58].

A very elegant result proved in [58], states that if $\mathcal{P}$ is testable in the VDF model then it is testable in the "standard" (i.e., uniform) model with one-sided error. A natural follow-up question, raised by Goldreich in [58], asks whether the converse is also true. We answer Goldreich's question in the negative by providing a *complete* characterization of the properties which are testable (possibly with two-sided error) in the VDF model. To state this characterization, we need the following definition: say that graph property $\mathcal{P}$ is *extendable* if for every graph G satisfying $\mathcal{P}$ there is a graph G' on $|V(G)|+1$ vertices which satisfies $\mathcal{P}$ and contains G as an induced subgraph. In other words, $\mathcal{P}$ is extendable if whenever G is a graph satisfying $\mathcal{P}$ and v is a "new" vertex (i.e. $v \notin V(G)$), one can connect v to $V(G)$ in such a way that this larger graph will also satisfy $\mathcal{P}$. Our characterization then states the following:

Theorem 13. *A graph property is testable in the VDF model if and only if it is hereditary and extendable.*

Immediate corollaries of Theorem 13 are that induced H-freeness is testable in the VDF model for every graph H, and (not necessarily induced) H-freeness is testable in the VDF model if and only if H has no

isolated vertices. It is interesting to compare the above (rather) simple characterization of the properties that are testable in the VDF model, with the (very) complicated characterization of [7] of the properties that are testable in the standard model.

Next, we consider variants of the VDF model in which one of the following restrictions is posed:

- Only "large enough" inputs (as a function of ε) can be fed to the tester.
- The weight assigned by $\mathcal{D}$ to any vertex of the input graph is $o(1)$.
- The weight assigned by $\mathcal{D}$ to any vertex of the input graph is $\Omega(1/n)$.
- $|V(G)|$ is given to the tester as part of the input.

We show that in each of these four models, every hereditary property is testable (cf. Theorem 13).

References: The results of this chapter appeared as:

- L. Gishboliner and A. Shapira. Testing Graphs Against an Unknown Distribution. Proceedings of STOC 2019, 535-546.

1.2.3 Testing Linear Inequalities of Subgraph Statistics (Chapter 6)

Goldreich and Shinkar [62] initiated the study of the testability of graph properties defined by a linear inequality involving subgraph densities. Let us now give the precise definitions. For graphs H, G, the *density* of H in G, denoted by $p(H,G)$, is the fraction of induced subgraphs of G of order $v(H)$ which are isomorphic to H. In other words, $p(H,G) = \#\{U \in \binom{V(G)}{v(H)} : G[U] \cong H\}/\binom{v(G)}{v(H)}$ (where $\cong$ denotes graph isomorphism). Given an integer $h \geq 2$, a rational number b and rational numbers $w_H \geq 0$, where H runs over all h-vertex graphs, we define $\Pi_{h,w,b}$ to be the property of all graphs G satisfying

$$\sum_H w_H \cdot p(H,G) \leq b.$$

Throughout this chapter, a tuple (h,w,b) will always consist of an integer $h \geq 2$, a rational number b, and a function $w\colon \{H : v(H) = h\} \to \mathbb{Q}^+$ from the set of all (unlabeled) h-vertex graphs to the positive rationals. The value assigned by w to a graph H is denoted by w_H. Note that if $b = 0$ then $\Pi_{h,w,b}$ is the property of being induced $\{H : w_H > 0\}$-free, which is testable since it is hereditary (see [5, 10]). So we see that the family of properties $\Pi_{h,w,b}$ constitutes a strict generalization of the family of properties of the form "induced $\mathcal{H}$-freeness" for a finite graph-family $\mathcal{H}$, since the former can encode the latter[9].

The $\Pi_{h,w,b}$ properties are closely related to a special type of testers, called *proximity oblivious testers* (POTs for short), which are defined as follows.

Definition 14. *A* proximity oblivious tester *(POT) for a graph property* Π *is an algorithm which makes a* constant *(i.e. independent of* n and ε*) number of queries to the input and satisfies the following. There is a constant* $c \in (0,1]$ *and a function* $f : (0,1] \to (0,1]$ *such that:*

[9]Indeed, if all graphs in $\mathcal{H}$ have the same size h then we can simply set $b = 0$, $w_H = 1$ for each $H \in \mathcal{H}$, and $w_H = 0$ for each h-vertex graph H which is not in $\mathcal{H}$. If graphs in $\mathcal{H}$ have varying sizes, then we reduce to the previous case by taking advantage of the fact that for every pair of graphs F, G and $h \geq v(F)$, it holds that $p(F,G) = \sum_H p(F,H) \cdot p(H,G)$, where the sum is over all h-vertex graphs H.

1. *If the input graph satisfies* Π *then the tester accepts with probability at least* c.

2. *If the input graph is* ε*-far from* Π *then the tester accepts with probability at most* $c - f(\varepsilon)$.

Proximity oblivious testers were introduced by Goldreich and Ron [60], who studied the special case of one-sided-error POTs (this corresponds to having $c = 1$ in Definition 14). Later, Goldreich and Shinkar [62] studied general (namely, two-sided-error) POTs in several settings, including those of boolean functions, dense graphs and bounded degree graphs.

It is not hard to see that the results of [5] imply not only that induced $\mathcal{H}$-freeness is testable for every finite graph-family $\mathcal{H}$, but also that every property of this type has a POT. It is then natural to ask whether in fact every property of the form $\Pi_{h,w,b}$ also has a POT. Such a conjecture has indeed been raised by Goldreich and Shinkar in [62].

Conjecture 15 ([62, Open Problem 3.11])**.** *Every property* $\Pi_{h,w,b}$ *has a POT.*

Our main result in this chapter, Theorem 16, disproves the above conjecture in a strong sense, by showing that there are properties $\Pi_{h,w,b}$ that are not testable at all (let alone testable using a POT). For a graph H, denote by $\overline{H}$ the complement of H.

Theorem 16. *Let* K_4 *denote the complete graph on* 4 *vertices,* D_4 *the diamond graph (i.e.* K_4 *minus an edge),* P_3 *the graph on* 4 *vertices containing a path on* 3 *vertices and an isolated vertex,* C_4 *the* 4*-cycle,* P_4 *the path on* 4 *vertices, and* $K_{1,3}$ *the star on* 4 *vertices. Set* $h = 4$*, and let* w_H *be the following weight-function assigning a non-negative weight to each graph on* 4 *vertices.*

H :	K_4	$\overline{K_4}$	D_4	$\overline{D_4}$	P_3	$\overline{P_3}$	C_4	$\overline{C_4}$	$K_{1,3}$	$\overline{K_{1,3}}$	P_4
w_H :	1	$\frac{1}{2}$	$\frac{5}{12}$	$\frac{5}{12}$	$\frac{1}{3}$	$\frac{1}{6}$	$\frac{1}{2}$	$\frac{1}{3}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$

Set $b = 5/16$*. Then, the property* $\Pi_{h,w,b} = \left\{G \ : \ \sum_{H:\, v(H)=4} w_H \cdot p(H,G) \leq \frac{5}{16}\right\}$ *is* not *testable.*

Given Theorem 16, it is natural to ask if every property of the form $\Pi_{h,w,b}$ can at least be tested using $o(n^2)$ edge-queries. We leave this as an open problem.

To state our second main result, we first need to introduce the following important definition.

Definition 17. *A tuple* (h, w, b) *has the* removal property *if there is a function* $f : (0, 1] \to (0, 1]$ *such that for every* $\varepsilon \in (0, 1)$ *and for every graph* G*, if* G *is* ε*-far from* $\Pi_{h,w,b}$ *then*

$$\sum_H w_H \cdot p(H, G) \geq b + f(\varepsilon).$$

Goldreich and Shinkar [62] observed that if (h, w, b) has the removal property then $\Pi_{h,w,b}$ admits a *size-oblivious* POT, where a tester is called size-oblivious if it does not know $|V(G)|$; that is, if its function depends only on the proximity parameter ε (and not on the size of the input). Goldreich and Shinkar's size-oblivious POT works as follows: given an input graph G, the POT samples a random induced subgraph of G of order h, and then rejects with probability w_H if the sampled subgraph is isomorphic to H, for each H on h vertices. If G satisfies $\Pi_{h,w,b}$ then by the definition of this property, G is rejected with probability $\sum_H w_H \cdot p(H, G) \leq b$. On the other hand, if G is ε-far from $\Pi_{h,w,b}$ then by the removal property, G is rejected with probability $\sum_H w_H \cdot p(H, G) \geq b + f(\varepsilon)$. Thus, Definition 14 is satisfied with $c = 1 - b$.

Our second result, Theorem 18, establishes the converse of the observation described in the previous paragraph, by showing that the removal property is *necessary* to having a size-oblivious POT.

Theorem 18. *If $\Pi_{h,w,b}$ has a size-oblivious POT then (h,w,b) has the removal property.*

Theorem 18 readily implies that if one "representation" of a given property as $\Pi_{h,w,b}$ has the removal property, then all such representations have the removal property.

References: The results of this chapter appeared as:

- L. Gishboliner, A. Shapira and H. Stagni. Testing Linear Inequalities of Subgraph Statistics. Random Structures and Algorithms, to appear. Also in Proceedings of ITCS 2020, 1-9.

1.3 A New Bound for the Brown–Erdős–Sós Problem (Chapter 7)

As mentioned in the beginning of Chapter 1, the triangle removal lemma is closely related to the $(6,3)$-conjecture of Brown, Erdős and Sós [26, 27], now known as the $(6,3)$-theorem of Ruzsa and Szemerédi [98]. This theorem states that every n-vertex 3-uniform hypergraph with $\Omega(n^2)$ edges must contain a $(6,3)$-configuration, where a (v,e)-*configuration* is a set of e edges on at most v vertices. The famous *Brown–Erdős–Sós conjecture* asserts that the $(6,3)$-theorem is a special case of a much more general phenomenon:

Conjecture 19 (Brown–Erdős–Sós Conjecture)**.** *For every $e \geq 3$, every large enough n-vertex 3-uniform hypergraph with $\Omega(n^2)$ edges contains an $(e+3,e)$-configuration.*

Despite much effort by many researchers, Conjecture 19 is wide open and seems very challenging even for $e = 4$. Consequently, it is natural to look for approximate versions. Namely, given $e \geq 3$, find the smallest $d = d(e)$ so that every large enough n-vertex 3-graph with $\Omega(n^2)$ edges contains e edges on at most $e + d$ vertices. Conjecture 19 then states that $d(e) \leq 3$ for every $e \geq 3$. The best known general bound on $d(e)$ was obtained 15 years ago by Sárközy and Selkow [99], who proved that $d(e) \leq 2 + \lfloor \log_2 e \rfloor$. Since the result of [99], the only advance was obtained by Solymosi and Solymosi [107], who improved the bound $d(10) \leq 2 + \lfloor \log_2 10 \rfloor = 5$ of [99] to $d(10) \leq 4$.

The main result of this chapter, Theorem 20, gives the first general improvement over the aforementioned result of [99]. Crucially, it shows that one can replace the $2 + \lfloor \log_2 e \rfloor$ bound on $d(e)$ proven in [99] by a much smaller, sub-logarithmic, term.

Theorem 20. *$d(e) \leq 18 \log e / \log \log e$ for every $e \geq 3$.*

An interesting feature of the proof of Theorem 20 is that although this theorem is concerned with 3-uniform hypergraphs, its proof relies on an application of the r-uniform removal lemma for *all* values of r.

References: The results of this chapter appeared as:

- D. Conlon, L. Gishboliner, Y. Levanzov and A. Shapira. A New Bound for the Brown–Erdős–Sós Problem. arXiv preprint: arXiv:1912.08834, 2019.

Chapter 2

Removal Lemmas with Polynomial Bounds

This chapter is concerned with polynomially-bounded removal lemmas for graphs and tournaments. In particular, it contains the proofs of Theorems 4, 5, 6, 8 and 9, as well as several other related results. A detailed overview of our results for (undirected) graphs and their applications is given in Section 2.1.

2.1 Detailed Overview of Results (for Undirected Graphs)

Recall that for a family of graphs $\mathcal{F}$, we denote by $\mathcal{P}^*_{\mathcal{F}}$ the graph property of being induced $\mathcal{F}$-free. Recall that a hereditary graph property $\mathcal{P}$ is *easily testable* if $w_{\mathcal{P}}(\varepsilon) = \text{poly}(1/\varepsilon)$, and *hard* otherwise.

Theorem 4 states that if a *finite* graph-family $\mathcal{F}$ contains a bipartite graph, a co-bipartite graph and a split graph, then $\mathcal{P}^*_{\mathcal{F}}$ is easily testable. And Theorem 5 states that if a *finite* graph-family $\mathcal{F}$ contains no bipartite graph or no co-bipartite graph, then $\mathcal{P}^*_{\mathcal{F}}$ is hard. Let us now mention some immediate applications of Theorems 4 and 5, starting with the former. Let P_k denote the path on k vertices. Alon and Shapira [9] proved that $\mathcal{P}^*_{P_3}$ is easily testable by relying on the fact that a graph satisfies $\mathcal{P}^*_{P_3}$ if and only if it is a disjoint union of cliques. Observing that P_3 is bipartite, co-bipartite and split, Theorem 4 gives the same result. In the same paper [9], it was shown that for any F other than P_2, P_3, P_4, C_4 and their complements, the property $\mathcal{P}^*_F$ is not easily testable. The two cases that were left open were $\mathcal{P}^*_{P_4}$ and $\mathcal{P}^*_{C_4}$. The case of $\mathcal{P}^*_{P_4}$ was settled only very recently by Alon and Fox [8] who used the structural characterization of induced P_4-free graphs in order to show that $\mathcal{P}^*_{P_4}$ is easily testable. As in the case of P_3, since P_4 is bipartite, co-bipartite and split, Theorem 4 gives the result of Alon and Fox [8] as a special case. Finally, a famous theorem of Alon [2] states that the property of being (not necessarily induced) F-free is easily testable if and only if F is bipartite. It is easy to see that the 'if part' of this theorem follows immediately from Theorem 4. Indeed, this follows from the simple observation that being F-free is equivalent to satisfying $\mathcal{P}^*_{\mathcal{F}}$, where $\mathcal{F}$ consists of all supergraphs of F on $|V(F)|$ vertices.

We now turn to derive some new testability results from Theorem 4. It is well known that the property of being a *line graph* is equivalent to $\mathcal{P}^*_{\mathcal{F}}$, where $\mathcal{F}$ is a family of 9 graphs, each having at most 6 vertices (see [72]). One of these graphs is $K_{1,3}$, which is both bipartite and split, and another one is a complete graph on 5 vertices minus a single edge, which is co-bipartite. Hence, Theorem 4 implies that the property

of being a line graph is easily testable. Two other graph properties which can be shown to be easily testable via Theorem 4 are being a *threshold graph* and a *trivially perfect graph*. Since both properties are equivalent to $\mathcal{P}^*_{\mathcal{F}}$ for an appropriate finite $\mathcal{F}$, where in both cases $P_4 \in \mathcal{F}$ (see [64, 65]), we immediately deduce from Theorem 4 that both are easily testable.

As we mentioned above, Alon [2] proved that being F-free is easily testable if and only if F is bipartite. It is now easy to see that the 'only if' part of Alon's result follows from Theorem 5. As we mentioned above, Alon and Shapira [9] proved that $\mathcal{P}^*_F$ is not easily testable for every F other than P_2, P_3, P_4, C_4 and their complements. Again, this result follows as a special case of Theorem 5.

Having given both a sufficient condition (i.e., Theorem 4) and a necessary condition (i.e., Theorem 5) for being easily testable, it is natural to ask whether one of these conditions in fact characterizes the finite families $\mathcal{F}$ for which $\mathcal{P}^*_{\mathcal{F}}$ is easily testable. Unfortunately, none do. It is known that being a split graph is equivalent to $\mathcal{P}^*_{\mathcal{F}}$ where $\mathcal{F} = \{C_5, C_4, \overline{C_4}\}$ (see [64]). While $\mathcal{F}$ does not satisfy the condition of Theorem 4 (it does not contain a split graph), the property of being a split graph is easily testable since it is one of the partition properties that were shown to be easily testable in [59]. Therefore, the sufficient condition in Theorem 4 is not necessary. Showing that the necessary condition of Theorem 5 is not sufficient is a bit harder, and is stated in the following theorem.

Theorem 2.1.1. *There is a bipartite F_1 and a co-bipartite F_2 such that $\mathcal{P}^*_{\{F_1,F_2\}}$ is not easily testable.*

Thus the above theorem also implies that in Theorem 4 we cannot drop the requirement that $\mathcal{F}$ should contain a split graph. The fact that we cannot drop the requirement that $\mathcal{F}$ should contain a bipartite graph follows from [98] where it was (implicitly) proved that triangle-freeness is not easily testable. By symmetry, the same holds for the co-bipartite graph.

The following theorem, which may be of independent interest, is the key technical step in the proof of Theorem 5.

Theorem 2.1.2. *For every $h \geq 3$ there is $\varepsilon_0 = \varepsilon_0(h)$ such that the following holds for every $\varepsilon < \varepsilon_0$ and every non-bipartite graph H on h vertices. For every $n \geq n_0(\varepsilon)$ there is a graph on n vertices which is ε-far from being induced H-free and yet contains at most $\varepsilon^{\Omega(\log(1/\varepsilon))}n^h$ (not necessarily induced) copies of H.*

Thus far, we have only considered properties of the form $\mathcal{P}^*_{\mathcal{F}}$ for a *finite* set of forbidden induced subgraphs $\mathcal{F}$. We now move on to consider the case where $\mathcal{F}$ may be infinite. Here the situation is somewhat more complicated. We start by introducing an important feature of a hereditary graph property.

Definition 2.1.3. *Let F be a graph with vertex set $V(F) = \{1, \ldots, p\}$ and let $g : V(F) \to \{0, 1\}$. We say that a graph G is a g-*blowup *of F if G admits a vertex partition $V(G) = P_1 \cup \cdots \cup P_p$ such that:*

1. *For every $1 \leq i < j \leq p$, if $(i, j) \in E(F)$ then (P_i, P_j) is a complete bipartite graph, and if $(i, j) \notin E(F)$ then (P_i, P_j) is an empty bipartite graph;*

2. *For every $1 \leq i \leq p$, if $g(i) = 1$ then P_i is a clique and if $g(i) = 0$ then P_i is an independent set.*

Definition 2.1.4. *We say that a graph property $\mathcal{P}$ is* closed under blowups *if for every graph F which satisfies $\mathcal{P}$ there is a function $g : V(F) \to \{0, 1\}$ such that every g-blowup of F satisfies $\mathcal{P}$.*

Our main result regarding hereditary properties characterized by an infinite family of forbidden subgraphs $\mathcal{F}$ is the following.

Theorem 2.1.5. *Let $\mathcal{F}$ be a graph family such that*

1. *$\mathcal{F}$ contains a bipartite graph, a co-bipartite graph and a split graph.*
2. *$\mathcal{P}^*_{\mathcal{F}}$ is closed under blowups.*

Then $\mathcal{P}^*_{\mathcal{F}}$ is easily testable.

We now recall the definition of a *semi-algebraic graph property*, which appears in the statement of Theorem 6. A semi-algebraic graph property $\mathcal{P}$ is given by an integer $k \geq 1$, a set of real $2k$-variate polynomials $f_1, \ldots, f_t \in \mathbb{R}[x_1, \ldots, x_{2k}]$ and a Boolean function $\Phi : \{\text{true}, \text{false}\}^t \to \{\text{true}, \text{false}\}$. A graph G satisfies the property $\mathcal{P}$ if one can assign a point $p_v \in \mathbb{R}^k$ to each vertex $v \in V(G)$ in such a way that a pair of distinct vertices u, v are adjacent if and only if

$$\Phi\Big(f_1(p_u, p_v) \geq 0, \ldots, f_t(p_u, p_v) \geq 0\Big) = \text{true}.$$

In the expression $f_i(p_u, p_v)$, we substitute p_u into the first k variables of f_i and p_v into the last k variables of f_i. In what follows, we call the points p_v *witnesses*[1] to the fact that G satisfies $\mathcal{P}$.

Some examples of semi-algebraic graph properties are those that correspond to being an intersection graph of certain semi-algebraic sets in $\mathbb{R}^k$. For example, a graph is an *interval graph* if one can assign an interval in $\mathbb{R}$ to each vertex so that u, v are adjacent iff their intervals intersect. Similarly, a graph is a *unit disc graph* if it is the intersection graph of unit discs in $\mathbb{R}^2$.

The family of semi-algebraic graph properties has been extensively studied by many researchers, see e.g. [51] and its references. Alon [3] conjectured that every semi-algebraic graph property is easily testable. This conjecture is resolved by Theorem 6 which, as we now show, is a special case of Theorem 2.1.5.

Proof sketch for Theorem 6. Fix a semi-algebraic graph property $\mathcal{P}$. Let $\mathcal{F}$ be the family of all graphs which do not satisfy $\mathcal{P}$. As $\mathcal{P}$ is a hereditary property, we have $\mathcal{P} = \mathcal{P}^*_{\mathcal{F}}$. To prove the theorem, it is enough to show that $\mathcal{P} = \mathcal{P}^*_{\mathcal{F}}$ satisfies Conditions 1 and 2 in Theorem 2.1.5. The fact that $\mathcal{F}$ satisfies Condition 1 of Theorem 2.1.5 follows directly from the well-known fact that every graph satisfying $\mathcal{P}$ has a bounded VC-dimension (we will give the definition of the VC-dimension of a graph in the detailed proof of Theorem 6, see Section 2.5). As for Condition 2, assume F satisfies $\mathcal{P}$, and $\{p_v : v \in V(F)\}$ are points witnessing this fact. Then setting $g(v) = 1$ if and only if $\Phi\Big(f_1(p_v, p_v) \geq 0; \ldots; f_t(p_v, p_v) \geq 0\Big) = \text{true}$, it is easy to see that every g-blowup of F satisfies $\mathcal{P}$. Indeed, the points witnessing the fact that a g-blowup of F satisfies $\mathcal{P}$ are obtained by taking each of the points p_v an appropriate number of times. ■

The reader can find a more detailed proof of Theorem 6 in Section 2.5. Observe that an immediate corollary of Theorem 6 is that for every semi-algebraic graph property $\mathcal{P}$ and $\varepsilon > 0$ there is $w^* = w^*_{\mathcal{P}}(\varepsilon) = \text{poly}(1/\varepsilon)$, so that if G is ε-far from satisfying $\mathcal{P}$, then G contains an induced subgraph on w^* vertices which does not satisfy $\mathcal{P}$.

The concept of VC-dimension (implicitly) plays a key role in our proofs of Theorems 4, 2.1.5 and 6 (see [16, Chapter 14] for an overview of this concept). In fact, as we (implicitly) show later in the chapter, a hereditary property $\mathcal{P}$ satisfies Condition 1 of Theorem 2.1.5 (i.e., it forbids a bipartite graph, a co-bipartite

[1] Note that a graph G might have many assignments of points witnessing the fact that it satisfies $\mathcal{P}$.

graph and a split graph), if and only if it has bounded VC dimension[2], in the sense that the VC-dimension of any graph satisfying $\mathcal{P}$ is bounded from above by some constant depending only on $\mathcal{P}$. Another aspect of the role played by VC-dimension in our results is the fact that the main tool we use, i.e. the "conditional" regularity lemma of [6] (stated here as Lemma 2.2.3), can be roughly stated as saying that graphs with bounded VC-dimension have small and highly-structured regular partitions (see [80] for a similar result). The proof of this lemma in [6] uses properties of VC-dimension.

It is worth mentioning that by now there are several works concerning efficient (i.e. polynomial) regularity lemmas for special classes of graphs, such as graphs with bounded VC-dimension [6, 80] (as mentioned above, the regularity lemma of [6] plays a key role in some of the proofs in this chapter); semi-algebraic graphs and hypergraphs [50, 51, 112] and more generally distal graphs [31, 32, 103]; and graphs excluding an induced bipartite half-graph [82].

Given Theorem 4, it is natural to ask if Condition 1 in Theorem 2.1.5 already guarantees that a property is easily testable. In light of the above discussion, this is equivalent to the (aesthetically pleasing) statement that every hereditary property of bounded VC dimension is easily testable. As our final theorem shows, this is regretfully not the case.

Theorem 2.1.6. *There is a family of graphs $\mathcal{F}$ that contains a bipartite graph, a co-bipartite graph and a split graph, for which $\mathcal{P}^*_{\mathcal{F}}$ is not easily testable.*

2.1.1 Some Nuggets from the Proofs

Here we give a rough overview of the proofs of Theorems 4, 5, 2.1.2 and 2.1.5. One key observation needed for proving Theorems 4 and 2.1.5 is that given a bipartite graph A_1, a co-bipartite graph A_2, and a split graph A_3, there is a bipartite graph B with sides X, Y, so that no matter which graphs one puts on X and on Y, one always gets a graph containing an *induced* copy of either A_1, A_2 or A_3 (see Lemma 2.3.6). This means that if a graph-family $\mathcal{F}$ satisfies the assumption of Theorem 4 and G is induced $\mathcal{F}$-free, then G has no induced copy of any graph obtained by adding edges to the two sides X, Y of B. If this is the case, then one can apply a "conditional regularity lemma" of Alon, Fischer and Newman [6] in order to find a highly structured partition of G (even more structured than the one produced by Szemerédi's regularity lemma [110]), which is of size only poly$(1/\varepsilon)$. This is in sharp contrast to the general argument of [10] that relied on Szemerédi's regularity lemma (or, more precisely, strengthenings thereof), which can only produce partitions of size tower(poly$(1/\varepsilon)$), see [66].

The proof of Theorem 2.1.5 is similar to that of Theorem 4, but involves an additional twist. The difference between these two theorems is that in Theorem 2.1.5, the graph-family $\mathcal{F}$ is allowed to be infinite. What usually considerably complicates proofs of this type is the need to embed multiple vertices into the same part of the partition mentioned above. The difficulty arises from the fact that parts of the partition are not highly structured (as opposed to the bipartite graphs between them). The purpose of Condition 2 of Theorem 2.1.5 is precisely to overcome this difficulty. Indeed, when dealing with properties satisfying this condition, it is enough to embed at most one vertex into each part. This feature is what makes it possible to prove Theorem 2.1.5.

[2]What we show (see Lemma 2.3.6) is that Condition 1 of Theorem 2.1.5 implies that every graph G satisfying $\mathcal{P}$ has no induced bipartite copy of some $k \times k$ bipartite graph. It is easy to see that this in turn implies that such a graph must have VC dimension at most $2k$.

As we mentioned above, the construction described in Theorem 2.1.2 is the key step in the proof of Theorem 5. We note that the novelty of Theorem 2.1.2 is that it constructs a graph G such that on the one hand, G is far from being *induced* H-free, and on the other hand, even the number of *not necessarily induced* copies of H in G is small. In comparison, constructions given in prior works [2, 9] were either far from being induced H-free but contained many (non-induced) copies of H, or contained few copies of H but were close to being induced H-free.

To prove Theorem 2.1.2, we too use a Ruzsa-Szemerédi-type construction based on Behrend's example [20] of a large set of integers S without 3-term arithmetic progressions. However, our argument involves the following twists. First, we take a set S that does not contain a (non-trivial) solution to *any* convex[3] linear equation with small coefficients. Second, we carefully label the vertices/clusters in this construction in such a way that any copy of H in the construction will necessarily contain a *monotone* cycle, i.e. a cycle whose labels increase in value. This property guarantees that such a cycle corresponds to a solution of a convex linear equation with integers from S, but we know that S has no such solution.

2.2 Regularity in Graphs, Tournaments and Matrices

In this section we introduce some definitions related to the regularity method. We then state the Alon–Fischer–Newman "conditional regularity lemma" [6], which is the key tool used in the proofs of Theorems 4, 2.1.5 and 8. Finally, we use the Alon–Fischer–Newman lemma to derive some efficient ("conditional") regularity lemmas for graphs and tournaments. Throughout this section and Sections 2.4 and 2.6, we assume that n, i.e. the number of vertices of the host graph G, is large enough as a function of the other parameters (i.e. the property $\mathcal{P}$ and the approximation parameter ε). We note that the minimal n for which our arguments work is (only) polynomial in $1/\varepsilon$ (where the polynomial depends on $\mathcal{P}$). To keep the presentation clean, we will often implicitly assume[4] that n is divisible by various integers that are bounded from above by a function of $\mathcal{P}$ and ε (which is polynomial in $1/\varepsilon$).

2.2.1 Regularity in Graphs

Let G be a graph on n vertices. For a set $X \subseteq V(G)$, we denote by $G[X]$ the subgraph of G induced by X. We say that X is *homogeneous* if it is either a clique or an independent set.

For a pair of disjoint sets $X, Y \subseteq V(G)$, let $e(X,Y)$ denote the number of edges with one endpoint in X and one endpoint in Y, and set $d(X,Y) = \frac{e(X,Y)}{|X||Y|}$. The number $d(X,Y)$ is called the *density* of the pair (X,Y). Note that $d(X,Y) = 1$ (resp. $d(X,Y) = 0$) if and only if the bipartite graph between X and Y is complete (resp. empty). We say that the pair (X,Y) is *homogeneous* if either $d(X,Y) = 1$ or $d(X,Y) = 0$. For $\delta \in (0,1)$, we say that (X,Y) is *δ-homogeneous* if either $d(X,Y) \geq 1-\delta$ or $d(X,Y) \leq \delta$. In cases where we consider several graphs at the same time, we write $d_G(X,Y)$ to refer to the density in G. The *weight* of (X,Y) is defined as $\frac{|X||Y|}{n^2}$.

Let $\mathcal{U} = \{U_1, \ldots, U_r\}$ be a vertex-partition of G, i.e. $V(G) = U_1 \uplus \cdots \uplus U_r$. We say that $\mathcal{U}$ is an *equipartition* if $||U_i| - |U_j|| \leq 1$ for every $1 \leq i,j \leq r$. Evidently, if r divides n (which we will assume to be the case, as mentioned above) then all parts $U_1, \ldots, U_r$ have the same size.

[3] A linear equation is convex if it is of the form $a_1x_1 + \ldots + a_kx_k = (a_1 + \ldots + a_k)x_{k+1}$ with all $a_i > 0$.

[4] if one wishes to discard this assumption, then it may be necessary to slightly change some of the constants chosen in the course of the proofs appearing in Sections 2.2, 2.4 and 2.6.

We say that $\mathcal{U}$ is *δ-homogeneous* if the sum of weights of non-δ-homogeneous pairs (U_i, U_j), $1 \leq i \neq j \leq r$, is at most δ. Note that if all parts in $\mathcal{U}$ have the same size then $\mathcal{U}$ is δ-homogeneous if and only if the number of (ordered) non-δ-homogeneous pairs (U_i, U_j) is at most δr^2.

2.2.2 Regularity in Tournaments

The definitions of Section 2.2.1 have natural analogues in the setting of tournaments. For a pair of vertices x, y in a digraph D, we write (x, y) for the edge directed from x to y. For a pair of disjoint subsets $X, Y \subseteq V(D)$, we write $X \to Y$ to mean that $(x, y) \in E(D)$ for every $x \in X, y \in Y$. If $X = \{x\}$ and $Y = \{y\}$, we write $x \to y$ instead of $\{x\} \to \{y\}$. We write $E(X, Y)$ for the set of edges going from X to Y. Note that $E(X, Y)$ is not the same as $E(Y, X)$. Evidently, if the digraph D is a tournament then $E(Y, X) = \emptyset$ if and only if $X \to Y$.

Let T be a tournament and let $X, Y \subseteq V(T)$ be disjoint. We set $e(X, Y) = |E(X, Y)|$ and $d(X, Y) = \frac{e(X,Y)}{|X||Y|}$. Note that $d(X, Y) + d(Y, X) = 1$, as T is a tournament. We have $X \to Y$ if and only if $d(X, Y) = 1$, and $Y \to X$ if and only if $d(X, Y) = 0$. For a constant $\delta < \frac{1}{2}$, we say that (X, Y) is δ-homogeneous if either $d(X, Y) \geq 1 - \delta$ or $d(X, Y) \leq \delta$. We say that the *dominant direction* of (X, Y) is $X \to Y$ if $d(X, Y) \geq \frac{1}{2}$ and is $Y \to X$ if $d(X, Y) < \frac{1}{2}$. The *weight* of the pair (X, Y) is $\frac{|X||Y|}{n^2}$. As in the graph case, an equipartition $\mathcal{U} = \{U_1, \ldots, U_r\}$ of $V(T)$ is *δ-homogeneous* if the sum of weights of non-δ-homogeneous pairs (U_i, U_j), $1 \leq i \neq j \leq r$, is at most δ.

2.2.3 Regularity in Matrices and the Alon–Fischer–Newman Regularity Lemma

Let A be an $n \times n$ matrix with 0/1 entries whose rows and columns are indexed by $1, ..., n$. For two sets $R, C \subseteq [n]$, the *block* $R \times C$ is the submatrix of A whose rows are the elements of R and whose columns are the elements of C. The *density* of the block $R \times C$, denoted by $d(R \times C)$, is the fraction of 1's in the block. For $\delta \in (0, 1)$, we say that $R \times C$ is *δ-homogeneous* if either $d(R \times C) \geq 1 - \delta$ or $d(R \times C) \leq \delta$. The *weight* of $R \times C$ is $\frac{|R||C|}{n^2}$. A partition of A is a pair $(\mathcal{R}, \mathcal{C})$, where $\mathcal{R}$ and $\mathcal{C}$ are partitions of $[n]$. We think of $\mathcal{R}$ as a partition of the rows of A, and of $\mathcal{C}$ as a partition of the columns of A. We say that $(\mathcal{R}, \mathcal{C})$ is *δ-homogeneous* if the sum of weights of non-δ-homogeneous blocks $R \times C$, where $R \in \mathcal{R}$ and $C \in \mathcal{C}$, is at most δ. In the case that A is the adjacency matrix of a graph G or of a tournament[5] T, these definitions are analogous to the definitions given in Sections 2.2.1 and 2.2.2, respectively. Indeed, every pair of disjoint sets X, Y (in a graph or a tournament) satisfies $d(X, Y) = d(X \times Y)$ (where the quantity on the left-hand side is the edge density in the bipartite graph/tournament with sides X, Y, and the quantity on the right-hand side is the density of the block $X \times Y$ in A). Moreover, if $\mathcal{P}$ is a partition of $[n]$ such that $(\mathcal{P}, \mathcal{P})$ is a δ-homogeneous partition of A, then[6] $\mathcal{P}$ is a δ-homogeneous partition of the corresponding graph or tournament.

A partition $(\mathcal{R}', \mathcal{C}')$ is a *refinement* of a partition $(\mathcal{R}, \mathcal{C})$ if every block of $\mathcal{R}' \times \mathcal{C}'$ is contained in some block of $\mathcal{R} \times \mathcal{C}$. We will need the following two lemmas.

Lemma 2.2.1. *Let $\delta \in (0, 1)$. If $(\mathcal{R}, \mathcal{C})$ is a $\frac{\delta^2}{2}$-homogeneous partition of an $n \times n$ matrix A, then every refinement of $(\mathcal{R}, \mathcal{C})$ is a δ-homogeneous partition of A.*

[5] The adjacency matrix of a tournament $T = (V, E)$ is the $V \times V$ matrix in which the (i, j)th entry is 1 if $(i, j) \in E$ and 0 otherwise.

[6] The other direction is not necessarily true, because the definition of a δ-homogeneous partition of a matrix takes into account the "diagonal" blocks $X \times X$, while the definition of a δ-homogeneous partition of a graph/tournament does not.

Proof. Let $(\mathcal{R}',\mathcal{C}')$ be a refinement of $(\mathcal{R},\mathcal{C})$. Let $\mathcal{N}$ be the set of non-δ-homogeneous blocks of $(\mathcal{R}',\mathcal{C}')$. Our goal is to show that the sum of weights of blocks $R'\times C'\in\mathcal{N}$ is at most δ. Let $\mathcal{N}_1$ (resp. $\mathcal{N}_2$) be the set of blocks $R'\times C'\in\mathcal{N}$ that are contained in a $\frac{\delta^2}{2}$-homogeneous (resp. non-$\frac{\delta^2}{2}$-homogeneous) block of $(\mathcal{R},\mathcal{C})$. Since $(\mathcal{R},\mathcal{C})$ is a $\frac{\delta^2}{2}$-homogeneous partition, the sum of weights of blocks $R'\times C'\in\mathcal{N}_2$ is at most $\frac{\delta^2}{2}$. Since $\mathcal{N}=\mathcal{N}_1\cup\mathcal{N}_2$ and $\frac{\delta}{2}+\frac{\delta^2}{2}\leq\delta$, it is enough to show that the sum of weights of blocks $R'\times C'\in\mathcal{N}_1$ is at most $\frac{\delta}{2}$.

Let $R\times C$ be a $\frac{\delta^2}{2}$-homogeneous block of $(\mathcal{R},\mathcal{C})$ and suppose without loss of generality that $d(R\times C)\leq\frac{\delta^2}{2}$ (the case that $d(R\times C)\geq 1-\frac{\delta^2}{2}$ is symmetrical). Let $R'_1,\ldots,R'_k$ (resp. $C'_1,\ldots,C'_\ell$) be the parts of $\mathcal{R}'$ (resp. $\mathcal{C}'$) which are contained in R (resp. C). By averaging we have

$$d(R\times C)=\sum_{i=1}^{k}\sum_{j=1}^{\ell}\frac{|R'_i||C'_j|}{|R||C|}\cdot d(R'_i\times C'_j).$$

By Markov's inequality, the total weight of blocks $R'_i\times C'_j$ for which $d(R'_i,C'_j)>\delta$ is less than $\frac{\delta}{2}\cdot\frac{|R||C|}{n^2}$. In conclusion, for every $\frac{\delta^2}{2}$-homogeneous block $R\times C$ of $(\mathcal{R},\mathcal{C})$ it holds that the total weight of blocks $R'\times C'\in\mathcal{N}_1$ contained in $R\times C$ is less than $\frac{\delta}{2}\cdot\frac{|R||C|}{n^2}$. By summing over all $\frac{\delta^2}{2}$-homogeneous blocks of $(\mathcal{R},\mathcal{C})$ we get that the total weight of blocks $R'\times C'\in\mathcal{N}_1$ is less than $\frac{\delta}{2}$, as required. ■

Lemma 2.2.2. *Let A be an $n\times n$ matrix, let $\delta\in(0,1)$, let $\mathcal{P}_0$ be an equipartition of $[n]$, and let $(\mathcal{R},\mathcal{C})$ be a $\frac{\delta^2}{8}$-homogeneous partition of A. Then there is an equipartition $\mathcal{U}$ of $[n]$ such that $(\mathcal{U},\mathcal{U})$ is a δ-homogeneous partition of A, and such that $\mathcal{U}$ refines $\mathcal{P}_0$ and has $r:=\lceil 4/\delta\rceil\cdot|\mathcal{P}_0|\cdot|\mathcal{R}|\cdot|\mathcal{C}|$ parts.*

Proof. Let $\mathcal{S}$ be the common refinement of $\mathcal{R}$, $\mathcal{C}$ and $\mathcal{P}_0$, i.e. $\mathcal{S}=\{R\cap C\cap P : R\in\mathcal{R},C\in\mathcal{C},P\in\mathcal{P}_0\}$. Partition every $S\in\mathcal{S}$ into equal parts of size $\frac{n}{r}$ and an additional part of size less than $\frac{n}{r}$. Denote the resulting partition by $\mathcal{T}$. For each $P\in\mathcal{P}_0$, let Z_P be the union of all additional parts contained in P, and note that $|Z_P|<|\mathcal{R}|\cdot|\mathcal{C}|\cdot\frac{n}{r}$. Set $Z=\bigcup_{P\in\mathcal{P}_0}Z_P$, noting that $|Z|\leq|\mathcal{P}_0|\cdot|\mathcal{R}|\cdot|\mathcal{C}|\cdot\frac{n}{r}\leq\frac{\delta n}{4}$. As $(\mathcal{T},\mathcal{T})$ is a refinement of $(\mathcal{R},\mathcal{C})$ and $(\mathcal{R},\mathcal{C})$ is a $\frac{\delta^2}{8}$-homogeneous partition of A, Lemma 2.2.1 (with $\frac{\delta}{2}$ in place of δ) implies that $(\mathcal{T},\mathcal{T})$ is a $\frac{\delta}{2}$-homogeneous partition of A.

Let $\mathcal{U}$ be the equipartition obtained from $\mathcal{T}$ by partitioning each of the sets Z_P ($P\in\mathcal{P}_0$) into parts of size $\frac{n}{r}$. It is clear that $\mathcal{U}$ refines $\mathcal{P}_0$ and has r parts. We claim that $(\mathcal{U},\mathcal{U})$ is a δ-homogeneous partition of A. Observe that if $X\times Y$ is a non-δ-homogeneous block of $(\mathcal{U},\mathcal{U})$, then either $X\times Y$ is a non-δ-homogeneous block of $(\mathcal{T},\mathcal{T})$, or one of the sets X,Y is contained in Z. Since $|Z|\leq\frac{\delta n}{4}$, the sum of weights of blocks $X\times Y$ for which X or Y is contained in Z is at most $\frac{2|Z|n}{n^2}\leq\frac{\delta}{2}$. Combining this with the fact that $(\mathcal{T},\mathcal{T})$ is $\frac{\delta}{2}$-homogeneous, we get that $(\mathcal{U},\mathcal{U})$ is δ-homogeneous, as required. ■

Let B be a 0/1-valued $h\times h$ matrix. A *copy* of B in a matrix A is a sequence of rows $r_1<\cdots<r_h$ and a sequence of columns $c_1<\cdots<c_h$ such that $A_{r_i,c_j}=B_{i,j}$ for every $1\leq i,j\leq h$. We are now ready to state the Alon–Fischer–Newman Regularity Lemma [6].

Lemma 2.2.3 (Alon–Fischer–Newman [6])**.** *There is a constant c_0 such that the following holds for every integer $h\geq 1$ and $\delta\in(0,1)$. For every 0/1-valued matrix A of size $n\times n$ with $n>(h/\delta)^{c_0h}$, either A has a δ-homogeneous partition $(\mathcal{R},\mathcal{C})$ with $|\mathcal{R}|,|\mathcal{C}|\leq(h/\delta)^{c_0h}$, or for every 0/1-valued $h\times h$ matrix B, A contains at least $(\delta/h)^{c_0h^2}n^{2h}$ copies of B.*

2.2.4 Applications of the Alon–Fischer–Newman Regularity Lemma

In this section we apply Lemma 2.2.3 to the adjacency matrices of graphs and tournaments. We will give proofs only for the graph case, since the proofs in the tournament case are essentially the same. In what follows, we assume that the vertex set of the graph G is $[n]$.

An *induced bipartite copy* of a bipartite graph $H = (S \cup T, E)$ in a graph G is an injection $\varphi : V(H) \to V(G)$ such that for every $s \in S$ and $t \in T$ we have $(s,t) \in E(H)$ if and only if $(\varphi(s), \varphi(t)) \in E(G)$. Notice that there is no restriction on the subgraphs of G induced by $\varphi(S)$ or by $\varphi(T)$ (in other words, the definition only "cares" about the edges between $\varphi(S)$ and $\varphi(T)$).

Lemma 2.2.4. *There is a function $\rho_{2.2.4} : \mathbb{N} \times (0,1) \to (0,1)$ such that[7] $\rho_{2.2.4}(h,\delta) = \text{poly}(\delta)$, and such that for every integer $h \geq 1$, for every $h \times h$ bipartite graph $H = (S \cup T, E)$ and for every $\delta \in (0,1)$, the following holds: let G be a graph on $n \geq n_0(h,\delta) = \text{poly}(1/\delta)$ vertices and let $\mathcal{P}_0$ be an equipartition of $V(G) = [n]$. Then G either contains at least $\rho_{2.2.4}(h,\delta)n^{2h}$ induced bipartite copies of H or admits a δ-homogeneous equipartition $\mathcal{U}$ which refines $\mathcal{P}_0$ and has at most $|\mathcal{P}_0| \cdot \rho_{2.2.4}(h,\delta)^{-1}$ parts.*

Proof. We prove the lemma with $\rho = \rho_{2.2.4}(h,\delta) := (\frac{\delta^2}{8h})^{3c_0h^2}$ (where c_0 is from Lemma 2.2.3). Let $A = A(G)$ be the adjacency matrix of G. Let B be the *bipartite adjacency matrix* of H; that is, B is an $h \times h$ matrix, indexed by $S \times T$, in which $B_{s,t} = 1$ if $(s,t) \in E(H)$ and $B_{s,t} = 0$ otherwise. Suppose first that A contains at least $(\frac{\delta^2}{8h})^{c_0h^2}n^{2h}$ copies of B. Observe that a copy of B which does not intersect the main diagonal of A corresponds to an induced bipartite copy of H in G. The number of $h \times h$ submatrices of A which intersect its main diagonal is $O(n^{2h-1})$. Thus, G contains at least $(\frac{\delta^2}{8h})^{c_0h^2}n^{2h} - O(n^{2h-1}) \geq (\frac{\delta^2}{8h})^{2c_0h^2}n^{2h} \geq \rho n^{2h}$ induced bipartite copies of H, as required.

Now suppose that A contains less than $(\frac{\delta^2}{8h})^{c_0h^2}n^{2h}$ copies of B. By Lemma 2.2.3, applied with approximation parameter $\frac{\delta^2}{8}$, A admits a $\frac{\delta^2}{8}$-homogeneous partition $(\mathcal{R},\mathcal{C})$ with $|\mathcal{R}|, |\mathcal{C}| \leq (\frac{8h}{\delta^2})^{c_0h}$. By Lemma 2.2.2, there is an equipartition $\mathcal{U}$ of $[n]$ which refines $\mathcal{P}_0$, has

$$\lceil 4/\delta \rceil \cdot |\mathcal{P}_0| \cdot |\mathcal{R}| \cdot |\mathcal{C}| \leq 8\delta^{-1}|\mathcal{P}_0| \cdot \left(\frac{8h}{\delta^2}\right)^{2c_0h} \leq |\mathcal{P}_0| \cdot \left(\frac{8h}{\delta^2}\right)^{3c_0h} \leq |\mathcal{P}_0| \cdot \rho^{-1}$$

parts, and satisfies that $(\mathcal{U},\mathcal{U})$ is a δ-homogeneous partition of A. This implies that $\mathcal{U}$ is a δ-homogeneous partition of G. The lemma follows. ■

Lemma 2.2.6 below is a "conditional" variant of a well-known corollary of Szemerédi's regularity lemma (see e.g. [5]). For its proof we will need the following standard quantitative version of Ramsey's theorem.

Claim 2.2.5 (see e.g. [23])**.** *Every graph on 4^k vertices contains a homogeneous set of size k.*

Lemma 2.2.6. *There is a function $\zeta_{2.2.6} : \mathbb{N}^2 \times (0,1) \to (0,1)$ such that $\zeta_{2.2.6}(h,m,\gamma) = \text{poly}(\gamma)$, and such that the following holds for every $h,m \geq 1$, for every $h \times h$ bipartite graph H, and for every $\gamma \in (0,1)$. Every graph G on $n \geq n_0(h,m,\gamma) = \text{poly}(1/\gamma)$ vertices either contains at least $\zeta_{2.2.6}(h,m,\gamma)n^{2h}$ induced bipartite copies of H or there are pairwise-disjoint subsets $W_1, \ldots, W_m \subseteq V(G)$ with the following properties:*

[7] By $\rho_{2.2.8}(h,\delta) = \text{poly}(\delta)$ we mean that $\rho_{2.2.8}(h,\delta)$ is at least polynomial in δ. The particular polynomial may (and usually will) depend on h, but we omit this from the notation because in what follows, h will depend only on the property $\mathcal{P}$ (and not on ε). Similarly, the notation $\zeta_{2.2.8}(h,m,\delta,\gamma) = \text{poly}(\delta,\gamma)$ in Lemma 2.2.8 means that $\zeta_{2.2.8}(h,m,\delta,\gamma)$ is (at least) polynomial in δ,γ, where the polynomial may depend on h,m.

1. *Either* $d(W_i, W_j) \geq 1 - \gamma$ *for every* $1 \leq i < j \leq m$, *or* $d(W_i, W_j) \leq \gamma$ *for every* $1 \leq i < j \leq m$.

2. $|W_i| \geq n \cdot \zeta_{2.2.6}(h, m, \gamma)$ *for every* $1 \leq i \leq m$.

Proof. Set $\delta := \min\{4^{-m-1}, \gamma\}$. We prove the lemma with $\zeta = \zeta_{2.2.6}(h, m, \gamma) := 4^{-m-1} \cdot \rho_{2.2.4}(h, \delta)$, where $\rho_{2.2.4}$ is from Lemma 2.2.4. We assume that G contains less than ζn^{2h} (and hence also less than $\rho_{2.2.4}(h, \delta)n^{2h}$) induced bipartite copies of H and prove that the other alternative in the lemma holds. Let $\mathcal{P}_0$ be an arbitrary equipartition of $V(G)$ into 4^{m+1} parts. Apply Lemma 2.2.4 with δ as defined above, to obtain a δ-homogeneous equipartition $\mathcal{W}$ of G which refines $\mathcal{P}_0$ and has at most $|\mathcal{P}_0| \cdot \rho_{2.2.4}(h, \delta)^{-1} = 4^{m+1} \cdot \rho_{2.2.4}(h, \delta)^{-1} = \zeta^{-1}$ parts. Then every $W \in \mathcal{W}$ satisfies $|W| \geq \zeta n$.

Set $w := |\mathcal{W}|$, noting that $w \geq 4^{m+1}$ because $\mathcal{W}$ refines $\mathcal{P}_0$. Define an auxiliary graph J on the set $\mathcal{W}$ in which (W, W') is an edge if and only if the pair (W, W') is δ-homogeneous. Since $\mathcal{W}$ is a δ-homogeneous partition, we have

$$e(J) \geq \binom{w}{2} - \delta w^2 \geq \binom{w}{2} - 4^{-m-1}w^2 > \left(1 - \frac{1}{4^m - 1}\right)\frac{w^2}{2}.$$

By Turán's Theorem (see e.g. [23]), there is a subset $\mathcal{W}' \subseteq \mathcal{W}$ of size $|\mathcal{W}'| = 4^m$ which spans a clique in J. Then for every $W, W' \in \mathcal{W}'$, the pair (W, W') is δ-homogeneous and hence also γ-homogeneous. Define a new graph on $\mathcal{W}'$ as follows: for $W, W' \in \mathcal{W}'$, put an edge between W and W' if and only if $d(W, W') \geq 1 - \gamma$ (the other option being that $d(W, W') \leq \gamma$). By Ramsey's theorem (see Claim 2.2.5), this graph contains a homogeneous set of size m, which we denote by $\{W_1, \ldots, W_m\}$. Then we have either $d(W_i, W_j) \geq 1 - \gamma$ for every $1 \leq i < j \leq m$, or $d(W_i, W_j) \leq \gamma$ for every $1 \leq i < j \leq m$, depending on whether $\{W_1, \ldots, W_m\}$ is a clique or an independent set. This completes the proof. ■

In what follows, we will need the following simple claim, whose proof is straightforward and thus omitted.

Claim 2.2.7. *Let* $\gamma, \eta \in (0, 1)$, *let* X, Y *be disjoint vertex-sets and let* $X' \subseteq X$, $Y' \subseteq Y$ *be such that* $|X'| \geq (\eta/\gamma)^{1/2}|X|$ *and* $|Y'| \geq (\eta/\gamma)^{1/2}|Y|$. *If* (X, Y) *is* η*-homogeneous then* $|d(X', Y') - d(X, Y)| \leq \gamma$.

The following lemma is the main tool used in the proofs of Theorems 4 and 2.1.5. It is worth noting that the idea of taking a regular partition and a refinement thereof (with a better measure of regularity) was first introduced in [5]. This approach, tailored to regularity lemmas with polynomial bounds, was also applied in [51].

Lemma 2.2.8. *There are functions* $\rho_{2.2.8} : \mathbb{N} \times (0, 1) \to (0, 1)$ *and* $\zeta_{2.2.8} : \mathbb{N}^2 \times (0, 1)^2 \to (0, 1)$ *such that* $\rho_{2.2.8}(h, \delta) = \text{poly}(\delta)$, $\zeta_{2.2.8}(h, m, \delta, \gamma) = \text{poly}(\delta, \gamma)$, *and the following holds for every pair of integers* $h, m \geq 1$, *for every* $\gamma, \delta \in (0, 1)$ *and for every* $h \times h$ *bipartite graph* H. *Every graph* G *on* $n \geq n_0(h, m, \delta, \gamma) = \text{poly}(1/\delta, 1/\gamma)$ *vertices either contains at least* $\zeta_{2.2.8}(h, m, \delta, \gamma)n^{2h}$ *induced bipartite copies of* H *or satisfies the following. There is an equipartition* $\mathcal{U} = \{U_1, ..., U_r\}$ *of* $V(G)$ *with* $\delta^{-1} \leq r \leq \rho_{2.2.8}(h, \delta)^{-1}$ *parts, and for each* $1 \leq i \leq r$ *there is a set* $W_i \subseteq U_i$ *and pairwise-disjoint sets* $W_{i,1}, \ldots, W_{i,m} \subseteq W_i$ *satisfying*

1. *For all but at most* δr^2 *of the pairs* $1 \leq i < j \leq r$, *it holds that* (U_i, U_j) *is* δ*-homogeneous and* $|d(W_i, W_j) - d(U_i, U_j)| \leq \frac{1}{4}$.

2. *For every* $1 \leq i < j \leq r$, (W_i, W_j) *is* γ*-homogeneous and* $|d(W_{i,s}, W_{j,t}) - d(W_i, W_j)| \leq \gamma$ *for every* $1 \leq s, t \leq m$.

3. *For every* $1 \le i \le r$*, either* $d(W_{i,s}, W_{i,t}) \ge 1-\gamma$ *for every* $1 \le s < t \le m$ *or* $d(W_{i,s}, W_{i,t}) \le \gamma$ *for every* $1 \le s < t \le m$*.*

4. $|W_{i,s}| \ge n \cdot \zeta_{2.2.8}(h, m, \delta, \gamma)$ *for every* $1 \le i \le r$ *and* $1 \le s \le m$*.*

Proof. Put

$$\rho := \frac{\delta}{2} \cdot \rho_{2.2.4}\left(h, \frac{\delta}{5}\right),$$

$$\eta := \min\left\{\rho^4,\ \gamma \cdot \zeta_{2.2.6}(h, m, \gamma)^2\right\},$$

$$\rho_1 := \rho \cdot \rho_{2.2.4}(h, \eta),$$

$$\zeta := \min\left\{\rho,\ \ \zeta_{2.2.6}(h, m, \gamma) \cdot \rho_1^{2h},\ \ (\eta/\gamma)^{1/2} \cdot \rho_1\right\},$$

where $\rho_{2.2.4}$ is from Lemma 2.2.4 and $\zeta_{2.2.6}$ is from Lemma 2.2.6. We prove the lemma with $\rho_{2.2.8}(h, \delta) := \rho$ and $\zeta_{2.2.8}(h, m, \delta, \gamma) := \zeta$. It is easy to check (using the guarantees of Lemmas 2.2.4 and 2.2.6) that ρ is polynomial in δ, and that ζ is polynomial in δ and γ, as required.

We assume that G contains less than ζn^{2h} induced bipartite copies of H and prove that the other alternative in the statement of the lemma holds. Since $\zeta \le \rho \le \rho_{2.2.4}(h, \frac{\delta}{5})$, G contains less than $\rho_{2.2.4}(h, \frac{\delta}{5})n^{2h}$ induced bipartite copies of H. Let $\mathcal{P}_0$ be an arbitrary equipartition of $V(G)$ into $\lceil 1/\delta \rceil$ parts. By Lemma 2.2.4 with approximation parameter $\frac{\delta}{5}$, there is a $\frac{\delta}{5}$-homogeneous equipartition $\mathcal{U} = \{U_1, \ldots, U_r\}$ of G which refines $\mathcal{P}_0$, and satisfies

$$|\mathcal{U}| = r \le |\mathcal{P}_0| \cdot \rho_{2.2.4}\left(h, \frac{\delta}{5}\right)^{-1} \le \frac{2}{\delta} \cdot \rho_{2.2.4}\left(h, \frac{\delta}{5}\right)^{-1} = \rho^{-1} = \rho_{2.2.8}(h, \delta)^{-1}.$$

Note also that $r \ge \delta^{-1}$, as $\mathcal{U}$ is a refinement of $\mathcal{P}_0$.

Since $\zeta \le \rho_1 \le \rho_{2.2.4}(h, \eta)$, our assumption in the beginning of the proof implies that G contains less than $\rho_{2.2.4}(h, \eta)n^{2h}$ induced bipartite copies of H. Thus, by Lemma 2.2.4 with approximation parameter η and with $\mathcal{P}_0 = \mathcal{U}$, G admits an η-homogeneous equipartition $\mathcal{W}$ that refines $\mathcal{U}$ and has at most $|\mathcal{W}| \le |\mathcal{U}| \cdot \rho_{2.2.4}(h, \eta)^{-1} \le \rho^{-1} \cdot \rho_{2.2.4}(h, \eta)^{-1} = \rho_1^{-1}$ parts. Hence, for every $W \in \mathcal{W}$ we have $|W| \ge \rho_1 n$.

For each $1 \le i \le r$ define $\mathcal{W}_i = \{W \in \mathcal{W} : W \subseteq U_i\}$. Sample a part $W_i \in \mathcal{W}_i$ uniformly at random. Let $\mathcal{A}_1$ be the event that all pairs (W_i, W_j) are η-homogeneous. By using the fact that $\mathcal{W}$ is η-homogeneous, we get that for every $1 \le i < j \le r$, the probability that (W_i, W_j) is not η-homogeneous is at most

$$\frac{\eta|\mathcal{W}|^2}{(|\mathcal{W}|/|\mathcal{U}|)^2} = \eta|\mathcal{U}|^2 = \eta r^2 \le \eta\rho^{-2} \le \rho^2.$$

Thus, by the union bound over all $\binom{r}{2} < \frac{1}{2\rho^2}$ pairs $1 \le i < j \le r$, we get that $\mathbb{P}[\mathcal{A}_1] > \frac{1}{2}$.

A pair $1 \le i < j \le r$ is called *good* if (U_i, U_j) is $\frac{\delta}{5}$-homogeneous and $|d(W_i, W_j) - d(U_i, U_j)| \le \frac{1}{4}$; otherwise (i, j) is called *bad*. Let $\mathcal{A}_2$ be the event that there are at most δr^2 bad pairs. Note that if $\mathcal{A}_2$ happened then Item 1 of the lemma is satisfied. We claim that $\mathbb{P}[\mathcal{A}_2] > \frac{1}{2}$. To this end, note that if (U_i, U_j) is $\frac{\delta}{5}$-homogeneous and $|d(W_i, W_j) - d(U_i, U_j)| > \frac{1}{4}$, then either $d(U_i, U_j) \ge 1 - \frac{\delta}{5}$ and $d(W_i, W_j) < \frac{3}{4}$, or $d(U_i, U_j) \le \frac{\delta}{5}$ and $d(W_i, W_j) > \frac{1}{4}$; in either case, the probability that this happens is less than $\frac{\delta/5}{1/4} = \frac{4\delta}{5}$.

It follows that the expected number of pairs $1 \leq i < j \leq r$ for which (U_i, U_j) is $\frac{\delta}{5}$-homogeneous but $|d(W_i, W_j) - d(U_i, U_j)| > \frac{1}{4}$, is less than $\frac{4\delta}{5}\binom{r}{2} < \frac{2\delta}{5}r^2$. By Markov's inequality, the probability that there are more than $\frac{4\delta}{5}r^2$ such pairs is smaller than $\frac{1}{2}$. Now, since all but at most $\frac{\delta}{5}r^2$ of the pairs (U_i, U_j) are $\frac{\delta}{5}$-homogeneous, our assertion that $\mathbb{P}[\mathcal{A}_2] > \frac{1}{2}$ follows. Thus, we proved that $\mathbb{P}[\mathcal{A}_i] > \frac{1}{2}$ for both $i = 1, 2$. This implies that $\mathbb{P}[\mathcal{A}_1 \cap \mathcal{A}_2] > 0$. From now on we fix a choice of $W_1, \ldots, W_r$ for which both $\mathcal{A}_1$ and $\mathcal{A}_2$ happened.

Let $1 \leq i \leq r$, and observe that $G[W_i]$ contains less than $\zeta_{2.2.6}(h, m, \gamma) \cdot |W_i|^{2h}$ induced bipartite copies of H. Indeed, this follows from the fact that $|W_i| \geq \rho_1 n$, the fact that $\zeta \leq \zeta_{2.2.6}(h, m, \gamma) \cdot \rho_1^{2h}$, and our assumption that G contains less than ζn^{2h} induced bipartite copies of H. So by Lemma 2.2.6, applied to the graph $G[W_i]$, there are pairwise-disjoint sets $W_{i,1}, \ldots, W_{i,m} \subseteq W_i$ such that

$$|W_{i,s}| \geq \zeta_{2.2.6}(h, m, \gamma) \cdot |W_i| \geq (\eta/\gamma)^{1/2} \cdot |W_i| \geq (\eta/\gamma)^{1/2} \cdot \rho_1 n \geq \zeta n$$

for every $1 \leq s \leq m$ (where in the second inequality we used our choice of η), and such that either $d(W_{i,s}, W_{i,t}) \geq 1 - \gamma$ for every $1 \leq s < t \leq m$ or $d(W_{i,s}, W_{i,t}) \leq \gamma$ for every $1 \leq s < t \leq m$. This establishes Item 3-4 of the lemma. Item 1 is guaranteed by our choice of $W_1, \ldots, W_r$ (i.e. by the assumption that $\mathcal{A}_2$ happened). It thus remains to establish Item 2. The fact that all pairs (W_i, W_j) are γ-homogeneous follows from our assumption that $\mathcal{A}_1$ happened and the fact that $\eta \leq \gamma$. Now let $1 \leq i < j \leq r$ and $1 \leq s, t \leq m$. Note that (W_i, W_j) is η-homogeneous (as $\mathcal{A}_1$ happened), and that $|W_{i,s}| \geq (\eta/\gamma)^{1/2} \cdot |W_i|$ and $|W_{j,t}| \geq (\eta/\gamma)^{1/2} \cdot |W_j|$. So by Claim 2.2.7 with $X = W_i$, $Y = W_j$, $X' = W_{i,s}$ and $Y' = W_{j,t}$, we have $|d(W_{i,s}, W_{j,t}) - d(W_i, W_j)| \leq \gamma$, as required. ■

Let us now describe the tournament analogue [8] of Lemma 2.2.8, as well as the minor changes that need to be made in the proofs in this section in order to proof this lemma. In the tournament setting, the analogue of induced bipartite graphs is *bipartite tournaments*, i.e. orientations of complete bipartite graphs. A copy of a bipartite tournament H with sides M, N in a tournament T is an injection $\varphi : V(H) \to V(T)$ such that for every $x \in S$ and $y \in T$ we have $(x, y) \in E(H)$ if and only if $(\varphi(x), \varphi(y)) \in E(T)$. When proving the tournament version of Lemma 2.2.4, one needs to consider the bipartite adjacency matrix of the given bipartite tournament $H = (M \cup N, E)$; this is the $M \times N$ matrix in which (x, y)th entry is 1 if $(x, y) \in E(H)$ and 0 if $(y, x) \in E(H)$ (for all $x \in M$ and $y \in N$). All other proofs[9] carry over essentially as is (with induced bipartite graphs replaced with bipartite tournaments). The following is our conditional regularity lemma for tournaments.

Lemma 2.2.9. *There is a function $\rho_{2.2.9} : \mathbb{N} \times (0, 1) \to (0, 1)$ and such that $\rho_{2.2.9}(h, \delta) = \text{poly}(\delta)$ and the following holds for every integer $h \geq 1$, for every $\delta \in (0, 1)$ and for every $h \times h$ bipartite tournament H. Every tournament T on $n \geq n_0(h, \delta) = \text{poly}(1/\delta)$ vertices either contains at least $\rho_{2.2.9}(h, \delta)n^{2h}$ copies of H or satisfies the following. There is an equipartition $\mathcal{U} = \{U_1, ..., U_r\}$ of $V(T)$ with $\delta^{-1} \leq r \leq \rho_{2.2.9}(h, \delta)^{-1}$ parts, and for each $1 \leq i \leq r$ there is a set $W_i \subseteq U_i$, such that the following holds.*

1. *For all but at most δr^2 of the pairs $1 \leq i < j \leq r$, it holds that (U_i, U_j) is δ-homogeneous and $|d(W_i, W_j) - d(U_i, U_j)| \leq \frac{1}{4}$.*

[8] In fact, in order to prove (the if part of) Theorem 8 we do not need the full strength of Lemma 2.2.8, but can settle for the special case $m = 1$ of this lemma (see Lemma 2.2.9).

[9] Since we will only need the (tournament analogue of the) $m = 1$ case of Lemma 2.2.8, we have no need for proving a tournament analogue of Lemma 2.2.6, although this can be done by essentially repeating the proof of that lemma.

2. (W_i, W_j) *is* δ*-homogeneous for every* $1 \leq i < j \leq r$.

3. $|W_i| \geq n \cdot \rho_{2.2.9}(h, \delta)$ *for every* $1 \leq i \leq r$.

2.3 Some Randomized Constructions

In this section we describe some randomized constructions of graphs and tournaments that will be used later on. Here we will start with the tournament case, as this case will require a somewhat more complicated construction. We then explain how to adapt our arguments to give an analogous construction for graphs.

A *k-partite tournament* is an orientation of a complete k-partite graph. A *completion* of a k-partite tournament $H = (V_1 \cup V_2 \cdots \cup V_k, E)$ is any tournament on $V(H)$ that agrees with H on the edges between the sets $V_1, ..., V_k$, i.e. any tournament obtained from H by adding k arbitrary tournaments on the sets $V_1, \ldots, V_k$. Our main lemma is as follows.

Lemma 2.3.1. *For every* $f \geq 2$ *there are* $m_0 = m_0(f)$ *and* $\gamma = \gamma(f) > 0$ *with the following property. Let* F *be an oriented graph on* f *vertices and let* D *be an oriented graph on* $[k]$*, where* $2 \leq k \leq f$*. Suppose that* $V(F)$ *has a partition* $V(F) = X_1 \cup \cdots \cup X_k$ *such that* $F[X_i]$ *is an acylic digraph for every* $1 \leq i \leq k$*, and such that* $E(X_j, X_i) = \emptyset$ *for every* $(i, j) \in E(D)$*. Then for every* $m \geq m_0$ *there exists a* k*-partite tournament* H *with sides* $V_1, \ldots, V_k$ *such that the following holds.*

1. $|V_i| = m$ *for every* $i = 1, ..., k$.

2. $V_i \to V_j$ *for every* $(i, j) \in E(D)$.

3. *Every completion of* H *contains a collection* $\mathcal{C}$ *of at least* γm^2 *copies of* F *with the property that every edge* $e \in E(H)$ *is contained in at most one of the copies of* F *in* $\mathcal{C}$.

In the proof of Lemma 2.3.1 we use the following three claims. Denote by $\mathrm{Bin}(N, p)$ the binomial distribution with parameters N and p. The following is a standard Chernoff-type bound.

Claim 2.3.2 ([16])**.** $\Pr\left[Bin(N, p) < (1 - \alpha)Np\right] \leq e^{-Np\alpha^2/2}$.

The following is the well-known tournament analogue of Ramsey's theorem.

Claim 2.3.3 ([84])**.** *Every tournament on* 2^{k-1} *vertices contains a transitive subtournament on* k *vertices.*

Claim 2.3.4. *Let* $t \geq 1, q \geq 2$ *be integers. Then there is a collection* $\mathcal{S} \subseteq [t]^q$ *of size at least* $(t/q)^2$ *such that every pair of distinct* q*-tuples in* $\mathcal{S}$ *have at most one identical entry.*

Proof. We construct the collection $\mathcal{S}$ greedily: we start with an empty collection, add an arbitrary q-tuple to it, discard all q-tuples that coincide in more than one entry with the q-tuple we added, and repeat. At the beginning we have all t^q of the q-tuples in $[t]^q$. At each step we discard at most $\binom{q}{2}t^{q-2}$ tuples. Therefore, at the end of the process we have a collection of size at least

$$\frac{t^q}{1 + \binom{q}{2}t^{q-2}} \geq \frac{t^q}{q^2 t^{q-2}} = \frac{t^2}{q^2},$$

as required. ■

Proof of Lemma 2.3.1. For every $i = 1, ..., k$ put $F_i = F[X_i]$ and $f_i = |X_i|$. Fix an integer $m > m_0(f)$, where $m_0(f)$ will be chosen later. For convenience of presentation, we assume that m is divisible by $2f$ and by $2f_i$ for every i. Let $V_1, ..., V_k$ be pairwise-disjoint vertex sets of size m each. The edges between the sets $V_1, ..., V_k$ are oriented as follows: for every $(i, j) \in E(D)$ we direct all edges from V_i to V_j. For every $1 \leq i < j \leq k$ for which $(i, j), (j, i) \notin E(D)$, orient the edges between V_i and V_j randomly and independently with probability $1/2$. We will show that with positive probability, the resulting k-partite tournament, H, satisfies the assertion of Item 3 in the lemma, thus finishing the proof.

An *F-partition* is a tuple $(\mathcal{P}_{i,j}, \mathcal{T}_{i,j})_{i,j}$, where $1 \leq i \leq k$ and $1 \leq j \leq \frac{m}{2f_i}$, with the following two properties.

- For each $1 \leq i \leq k$, $\mathcal{P}_{i,1}, ..., \mathcal{P}_{i,\frac{m}{2f_i}}$ are pairwise-disjoint subsets of V_i, each of size $f_i = |X_i|$.
- For each $1 \leq i \leq k$ and $1 \leq j \leq \frac{m}{2f_i}$, $T_{i,j}$ is a labeled transitive tournament on the set $\mathcal{P}_{i,j}$.

Note that $\bigcup_{j=1}^{m/(2f_i)} \mathcal{P}_{i,j}$ is a subset of V_i of size exactly $\frac{m}{2}$ (for each $1 \leq i \leq k$). The number of ways to choose an F-partition is exactly

$$\prod_{i=1}^{k} \frac{m!}{(m/2)!} \leq m^{km}. \tag{2.1}$$

By Claim 2.3.4 with parameters $t = \frac{m}{2f}$ and $q = k$, there is a collection $\mathcal{S} \subseteq \left[\frac{m}{2f}\right]^k \subseteq \left[\frac{m}{2f_1}\right] \times \cdots \times \left[\frac{m}{2f_k}\right]$ such that $|\mathcal{S}| \geq \left(\frac{m}{2fk}\right)^2 \geq \frac{m^2}{4f^4}$, and such that the following holds:

$$\text{For every pair } s = (s_1, ..., s_k), s' = (s'_1, ..., s'_k) \in \mathcal{S}, \quad \#\left\{1 \leq i \leq k : s_i = s'_i\right\} \leq 1. \tag{2.2}$$

For each $i = 1, \ldots, k$ we fix a linear ordering of the vertices of F_i in which all edges point forward, that is, if $(u, v) \in E(F_i)$ then u precedes v in the ordering. Such an ordering exists since F_i is acyclic. Fix an F-partition $\mathcal{Q} = (\mathcal{P}_{i,j}, \mathcal{T}_{i,j})_{i,j}$ and let $s = (s_1, ..., s_k) \in \mathcal{S}$. Since T_{i,s_i} is transitive and F_i is acyclic, F_i can be embedded into T_{i,s_i}. In what follows, when we say that T_{i,s_i} plays the role of F_i we mean that F_i is embedded in T_{i,s_i} in an order-preserving way with respect to our fixed ordering of F_i and the unique ordering of T_{i,s_i} in which all edges point forward. Let $A_{\mathcal{Q}}(s)$ be the event that $T_{1,s_1} \cup \cdots \cup T_{k,s_k}$, together with the edges of H connecting the sets $\mathcal{P}_{1,s_1}, \ldots, \mathcal{P}_{k,s_k}$, contains a copy of F with T_{i,s_i} playing the role of F_i. Then $\mathbb{P}\left[A_{\mathcal{Q}}(s)\right] \geq 2^{-\sum f_i f_j} > 2^{-f^2}$. Observe that by (2.2), the events $\{A_{\mathcal{Q}}(s) : s \in \mathcal{S}\}$ are independent. Since $|\mathcal{S}| \geq \frac{m^2}{4f^4}$, the random variable

$$Z_{\mathcal{Q}} := \sum_{s \in \mathcal{S}} \mathbb{1}_{A_{\mathcal{Q}}(s)}$$

stochastically dominates a binomial random variable with distribution $\mathrm{Bin}\left(\frac{m^2}{4f^4}, 2^{-f^2}\right)$. By Claim 2.3.2 with parameter $\alpha = \frac{1}{2}$ we have:

$$\mathbb{P}\left[Z_{\mathcal{Q}} < \frac{2^{-f^2} m^2}{8f^4}\right] \leq \mathbb{P}\left[\mathrm{Bin}\left(\frac{m^2}{4f^4}, 2^{-f^2}\right) < \frac{2^{-f^2} m^2}{8f^4}\right] \leq \exp\left\{-\frac{2^{-f^2} m^2}{32f^4}\right\} < m^{-fm} \leq m^{-km}.$$

The strict inequality above holds if m is large enough, and we choose $m_0(f)$ accordingly. Set $\gamma = \gamma(f) = 2^{-f^2}/(8f^4)$. As we saw in (2.1), there are at most m^{km} ways to choose an F-partition $\mathcal{Q}$. By the union bound over all F-partitions we get that the following event has positive probability: for every F-partition

$\mathcal{Q}$, the number of $s \in \mathcal{S}$ for which $A_{\mathcal{Q}}(s)$ happened is at least γm^2. We now show that if this event happens then H satisfies the assertion of Item 3 in the lemma.

Let T be a completion of H. For every $1 \leq i \leq k$, we use Claim 2.3.3 to extract from V_i a collection $\mathcal{P}_{i,1}, ..., \mathcal{P}_{i,\frac{m}{2f_i}}$ of pairwise-disjoint sets, each of size f_i, such that $T[\mathcal{P}_{i,j}]$ is transitive for every $1 \leq j \leq \frac{m}{2f_i}$. We extract these sets one by one and stop when there are $\frac{m}{2}$ remaining vertices. By Claim 2.3.3, we can do this as long as there are at least 2^{f_i-1} remaining vertices. By choosing $m_0(f)$ to be large enough we can guarantee that $\frac{m}{2} \geq 2^{f_i-1}$.

For every $1 \leq i \leq k$ and $1 \leq j \leq \frac{m}{2f_i}$, set $T_{i,j} = T[\mathcal{P}_{i,j}]$. Consider this F-partition $\mathcal{Q} = (\mathcal{P}_{i,j}, \mathcal{T}_{i,j})_{i,j}$. By our assumption, the event $A_{\mathcal{Q}}(s)$ happened for at least γm^2 of the elements $s \in \mathcal{S}$. By the definition of the event $A_{\mathcal{Q}}(s)$, if this event happened then the vertex-set $\mathcal{P}_{1,s_1} \cup \cdots \cup \mathcal{P}_{k,s_k}$ contains a copy of F (in the tournament T) with $\mathcal{T}_{i,s_i}$ playing the role of F_i. The collection $\mathcal{C}$ required by Item 3 consists of all such copies of F. By (2.2), every pair of copies of F in $\mathcal{C}$ can share vertices in no more than one of the parts $V_1, \ldots, V_k$. Therefore, every edge $e \in E(H)$ (that is, an edge that connects vertices in two distinct parts V_i, V_j) is contained in at most one of these copies. Thus, Item 3 in the lemma holds, as required. ■

We now establish a useful corollary of Lemma 2.3.1. We say that a bipartite tournament H *forces* an oriented graph F if every completion of H contains a copy of F.

Lemma 2.3.5. *Let F be a 2-colorable oriented graph. Then there is a bipartite tournament that forces F.*

Proof. Let $V(F) = X_1 \cup X_2$ be a proper 2-coloring of F. Apply Lemma 2.3.1 with parameter $f = |V(F)|$ and with D being the empty digraph on 2 vertices. Lemma 2.3.1 implies that there is a bipartite tournament H with sides V_1, V_2, where $|V_1| = |V_2| = m := m_0(f)$, such that every completion of H contains at least $\gamma(f) \cdot m^2$ (and, in particular, a positive number of) copies of F. ■

We now describe the analogue of Lemma 2.3.5 in the setting of undirected graphs. In this context, the main difference between graphs and tournaments is that while tournaments have only one type of "homogeneous structure", namely a transitive tournament, graphs have two types of "homogeneous structures", namely an independent set and a clique (see also Claim 2.2.5 vs. Claim 2.3.3). As a matter of fact, many of our sufficient conditions for having polynomially-bounded removal lemmas can be stated as saying that the objects under consideration (either graphs or tournaments) can be partitioned into two sets, each being homogeneous. For example, the "if" part of Theorem 8 requires the oriented graph F to be partitionable into two sets, each spanning an acylic digraph (which is just a subdigraph of a transitive tournament). And Theorem 4 requires that the graph family $\mathcal{F}$ contains a graph partitionable into two independent sets (i.e. a bipartite graph), a graph partitionable into two cliques (i.e., a co-bipartite graph), and a graph partitionable into an independent set and a clique (i.e., a split graph). In other words, we require $\mathcal{F}$ to contain any possible combination of the two homogeneous graph structures.

Let $H = (S \cup T, E)$ be a bipartite graph. A *completion* of H is any graph on $V(H)$ that agrees with H on the edges between S and T. In other words, a completion of H is any graph obtained by putting two arbitrary graphs on the sets S and T. We say that H is a *bipartite obstruction* for a graph property $\mathcal{P}$ if *no* completion of H satisfies $\mathcal{P}$. The following lemma can be thought of as the graph analogue of the case $k = 2$ of Lemma 2.3.1. The proof strategy is similar to that of Lemma 2.3.1.

Lemma 2.3.6. *Let $\mathcal{F}$ be a graph-family. Then $\mathcal{P}^*_{\mathcal{F}}$ admits a bipartite obstruction if and only if $\mathcal{F}$ contains a bipartite graph, a co-bipartite graph and a split graph.*

Proof. We start with the "only-if" part of the lemma. Let H be a bipartite obstruction for $\mathcal{P}^*_{\mathcal{F}}$ with sides S and T. By putting empty graphs on S and T we get a bipartite graph that does not satisfy $\mathcal{P}^*_{\mathcal{F}}$. This bipartite graph must then contain as an induced subgraph some element of $\mathcal{F}$, which is evidently also bipartite. This shows that $\mathcal{F}$ contains a bipartite graph. Similarly, by putting complete graphs on S and T (resp. a complete graph on S and an empty graph on T) we infer that $\mathcal{F}$ contains a co-bipartite (resp. split) graph, as required.

We now prove the "if" part of the lemma. Let $F_1, F_2, F_3 \in \mathcal{F}$ be such that F_1 is bipartite, F_2 is co-bipartite and F_3 is split, and write $V(F_1) = P_1 \cup Q_1$, $V(F_2) = P_2 \cup Q_2$, $V(F_3) = P_3 \cup Q_3$, where P_1, Q_1, P_3 are independent sets and P_2, Q_2, Q_3 are cliques. Put $f := v(F_1) + v(F_2) + 2v(F_3)$, and let h be some large integer, to be chosen later. Let $H = (S \cup T, E)$ be a random bipartite graph with $|S| = |T| = h$; that is, for each $s \in S, t \in T$, the edge (s,t) is included in H with probability $\frac{1}{2}$, independently. We will show that with positive probability, H is a bipartite obstruction for $\mathcal{F}$, thus proving the lemma. Let us set

$$r := \left\lfloor \frac{h - 4^f}{f} \right\rfloor.$$

An *f-partition* is a $2r$-tuple $(S_1, ..., S_r; T_1, ..., T_r)$ such that $S_1, ..., S_r$ (resp. $T_1, ..., T_r$) are pairwise-disjoint subsets of S (resp. T) of size f each. The number of ways to choose an f-partition is exactly

$$\left(\frac{h!}{(f!)^r (h - fr)!} \right)^2 \leq h^{2h}.$$

Let F and H be graphs and let $V(F) = P \cup Q$ and $V(H) = S \cup T$ be vertex-partitions. An *induced bipartite copy* of $F[P,Q]$ in $H[S,T]$ is an injection $\varphi : V(F) \to V(H)$ such that $\varphi(P) \subseteq S$ and $\varphi(Q) \subseteq T$, and such that for every $p \in P$ and $q \in Q$ we have $(p,q) \in E(F)$ if and only if $(\varphi(p), \varphi(q)) \in E(H)$.

For an f-partition $\mathcal{Q} = (S_1, ..., S_r; T_1, ..., T_r)$ and for $(i,j) \in [r]^2$, let $A_{\mathcal{Q}}(i,j)$ be the event that $H[S_i, T_j]$ contains induced bipartite copies of $F_1[P_1, Q_1]$, $F_2[P_2, Q_2]$, $F_3[P_3, Q_3]$ and $F_3[Q_3, P_3]$. We claim that for every completion H' of H, if S_i and T_j are homogeneous sets in H' and $A_{\mathcal{Q}}(i,j)$ happened, then H' is not induced $\mathcal{F}$-free (and hence does not satisfy $\mathcal{P}^*_{\mathcal{F}}$). Indeed, if S_i, T_j are independent sets (in H') then $H'[S_i \cup T_j]$ contains an induced copy of F_1; if S_i, T_j are cliques (in H') then $H'[S_i \cup T_j]$ contains an induced copy of F_2; and if S_i is a clique and T_j is an independent set or vice versa, then $H'[S_i \cup T_j]$ contains an induced copy of F_3.

Now let $\mathcal{A}$ be the event that for every f-partition $\mathcal{Q}$, there is a pair $(i,j) \in [r]^2$ for which $A_{\mathcal{Q}}(i,j)$ happened. We now show that if $\mathcal{A}$ happened then H is a bipartite obstruction for $\mathcal{F}$. We will then show that $\mathcal{A}$ happens with positive probability. Let H' be a completion of H. By repeatedly applying Claim 2.2.5, we extract from S pairwise-disjoint homogeneous sets $S_1, S_2, \ldots, S_r$ of size f each. This is possible due to our choice of r. Similarly, we extract from T pairwise-disjoint homogeneous sets $T_1, T_2, \ldots, T_r$ of size f each. Consider the f-partition $\mathcal{Q} = (S_1, ..., S_r; T_1, ..., T_r)$. Since $\mathcal{A}$ happened, there is $(i,j) \in [r]^2$ for which $A_{\mathcal{Q}}(i,j)$ happened. Since S_i and T_j are homogeneous in H', we get that H' does not satisfy $\mathcal{P}^*_{\mathcal{F}}$.

So it remains to show that $\mathbb{P}[\mathcal{A}] > 0$. Let $\mathcal{Q} = (S_1, ..., S_r; T_1, ..., T_r)$ be an f-partition. Since $|S_i| = |T_j| = f = v(F_1) + v(F_2) + 2v(F_3)$, it is possible to put a bipartite graph on (S_i, T_j) that will contain induced bipartite copies of $F_1[P_1, Q_1]$, $F_2[P_2, Q_2]$, $F_3[P_3, Q_3]$ and $F_3[Q_3, P_3]$. This implies that $\mathbb{P}\left[A_{\mathcal{Q}}(i,j)\right] \geq 2^{-f^2}$. Since the events $\{A_{\mathcal{Q}}(i,j) : i,j \in [r]\}$ are independent, the probability that $A_{\mathcal{Q}}(i,j)$ did not happen for any $(i,j) \in [r]^2$ is at most $\left(1 - 2^{-f^2}\right)^{r^2} \leq e^{-2^{-f^2} r^2} < h^{-2h}$, with the rightmost inequality holding provided

that we choose h to be large enough (see our choice of r). Recall that there are at most h^{2h} ways to choose an f-partition $\mathcal{Q}$. By the union bound over all f-partitions, we get $\mathbb{P}[\mathcal{A}^c] < 1$, as required. ■

We note that after completing the work presented here, we learned that a statement similar to Lemma 2.3.6 was already proved in [80].

2.4 Proof of Theorems 4 and 2.1.5

We start by proving the following simple counting lemma.

Lemma 2.4.1. *Let F be a graph, say with vertex-set $V(F) = \{1, \ldots, \ell\}$, and let $\lambda \in (0,1)$. Let $W_1, ..., W_\ell$ be pairwise-disjoint vertex sets in an n-vertex graph G such that*

1. *For every $1 \leq i < j \leq \ell$, if $(i,j) \in E(F)$ then $d(W_i, W_j) \geq 1 - \frac{1}{2\ell^2}$ and if $(i,j) \notin E(F)$ then $d(W_i, W_j) \leq \frac{1}{2\ell^2}$.*
2. *$|W_i| \geq \lambda n$ for every $1 \leq i \leq \ell$.*

Then with probability at least $\frac{2}{3}$, a random sequence of $12\ell/\lambda$ vertices of G, sampled uniformly and independently, contains an induced copy of F.

Proof. For each $1 \leq i \leq \ell$, sample a vertex $w_i \in W_i$ uniformly at random. For every $1 \leq i < j \leq \ell$, the assumption of the lemma gives that with probability at least $1 - \frac{1}{2\ell^2}$, if $(i,j) \in E(F)$ then $(w_i, w_j) \in E(G)$ and if $(i,j) \notin E(F)$ then $(w_i, w_j) \notin E(G)$. By the union bound over all pairs $1 \leq i < j \leq \ell$ we get that with probability at least $1 - \binom{\ell}{2}/2\ell^2 \geq \frac{3}{4}$, the set $\{w_1, \ldots, w_\ell\}$ spans an induced copy of F in which w_i plays the role of i.

Now let $u_1, \ldots, u_s \in V(G)$ be a random sequence of vertices, sampled uniformly and independently, where $s = 12\ell/\lambda$. Let $\mathcal{A}$ be the event that $U := \{u_1, \ldots, u_s\}$ contains a vertex of W_i for every $1 \leq i \leq \ell$. What we proved in the previous paragraph implies that conditioned on $\mathcal{A}$ happening, $G[U]$ contains an induced copy of F with probability at least $\frac{3}{4}$. Hence, to finish the proof it is enough to show that $\mathbb{P}[\mathcal{A}^c] \leq \frac{1}{12}$. For $1 \leq i \leq \ell$, the probability that $U \cap W_i = \emptyset$ is $\left(1 - \frac{|W_i|}{n}\right)^s \leq (1-\lambda)^s \leq e^{-\lambda s} \leq \frac{1}{12\ell}$. Here we used the assumption $|W_i| \geq \lambda n$ and our choice of s. By the union bound over all $1 \leq i \leq \ell$ we get that $\mathbb{P}[\mathcal{A}^c] \leq \frac{1}{12}$, as required. ■

We are now ready to prove Theorems 4 and 2.1.5.

Proof of Theorem 4. Our goal is to prove that $w_{\mathcal{P}^*_{\mathcal{F}}}(\varepsilon) = \text{poly}(1/\varepsilon)$. By Lemma 2.3.6, $\mathcal{P}^*_{\mathcal{F}}$ has a bipartite obstruction H. We can assume (by adding additional vertices if needed) that the two sides of H are of the same size, which we denote by h. We set $m := \max_{F \in \mathcal{F}} v(F)$. Given $\varepsilon < \frac{1}{2}$, set

$$\zeta := \zeta_{2.2.8}\left(h, m, \frac{\varepsilon}{3}, \frac{1}{4m^2}\right),$$

noting that $\zeta = \text{poly}(\varepsilon)$ (as h and m depend only on $\mathcal{P}$). Let G be an n-vertex graph which is ε-far from being induced $\mathcal{F}$-free. If G contains at least ζn^{2h} induced bipartite copies of H, then a random sequence of $2h = |V(H)|$ vertices of G (sampled uniformly and independently) spans an induced bipartite copy of

H with probability at least ζ. Hence, a random sequence of $4h \cdot \zeta^{-1}$ vertices of G contains an induced bipartite copy of H with probability at least

$$1 - (1-\zeta)^{1/\zeta} \geq 1 - e^{-2} \geq \frac{2}{3}.$$

Since H is a bipartite obstruction for $\mathcal{P}^*_{\mathcal{F}}$, every graph which contains an induced bipartite copy of H does not satisfy $\mathcal{P}^*_{\mathcal{F}}$. It follows that $w_{\mathcal{P}^*_{\mathcal{F}}}(\varepsilon) \leq 4h \cdot \zeta^{-1} = \text{poly}(1/\varepsilon)$. So we see that the assertion of the theorem holds in the case that G contains at least ζn^{2h} induced bipartite copies of H.

Suppose from now on that G contains less than ζn^{2h} induced bipartite copies of H. We apply Lemma 2.2.8 to G with parameters $\delta = \frac{\varepsilon}{3}$, $\gamma = \frac{1}{4m^2}$ and m as defined above, to get an equipartition $\mathcal{U} = \{U_1, ..., U_r\}$, sets $W_i \subseteq U_i$ and pairwise-disjoint sets $W_{i,1}, \ldots, W_{i,m} \subseteq W_i$ with the properties stated in the lemma.

Let G' be the graph obtained from G by making the following changes.

(a) For every $1 \leq i < j \leq r$, if $d(W_i, W_j) \geq 1 - \frac{1}{4m^2}$ then turn (U_i, U_j) into a complete bipartite graph, and if $d(W_i, W_j) \leq \frac{1}{4m^2}$ then turn (U_i, U_j) into an empty bipartite graph. By Item 2 in Lemma 2.2.8, one of these options holds.

(b) For every $1 \leq i \leq r$, if $d(W_{i,s}, W_{i,t}) \geq 1 - \frac{1}{4m^2}$ (resp. $d(W_{i,s}, W_{i,t}) \leq \frac{1}{4m^2}$) for every $1 \leq s < t \leq m$, then turn U_i into a clique (resp. an independent set). By Item 3 in Lemma 2.2.8, one of these options holds.

We claim that the number of edge-changes made in items (a)-(b) is less than εn^2. To prove this, we define $\mathcal{H}$ to be the set of pairs $1 \leq i < j \leq r$ for which (U_i, U_j) is $\frac{\varepsilon}{3}$-homogeneous and $|d(W_i, W_j) - d(U_i, U_j)| \leq \frac{1}{4}$. Observe that if $(i,j) \in \mathcal{H}$ then at most $\frac{\varepsilon}{3}|U_i||U_j|$ edge-changes were made in the bipartite graph (U_i, U_j) in Item (a) above; indeed, in the case that $d(W_i, W_j) \geq 1 - \frac{1}{4m^2} \geq \frac{3}{4}$ we have $d(U_i, U_j) \geq \frac{1}{2}$ and hence actually $d(U_i, U_j) \geq 1 - \frac{\varepsilon}{3}$; and the case that $d(W_i, W_j) \leq \frac{1}{4m^2}$ is symmetrical. By Item 1 in Lemma 2.2.8, the number of pairs $1 \leq i < j \leq r$ *not* belonging to $\mathcal{H}$ is at most $\frac{\varepsilon}{3}r^2$. It follows that the overall number of changes made in Item (a) is at most $|\mathcal{H}| \cdot \frac{\varepsilon}{3} \cdot \left(\frac{n}{r}\right)^2 + \frac{\varepsilon}{3}r^2 \cdot \left(\frac{n}{r}\right)^2 \leq \frac{2\varepsilon}{3}n^2$. As for item (b), the number of edge-changes made there is at most $r\binom{n/r}{2} < \frac{n^2}{r} \leq \frac{\varepsilon}{3}n^2$, where in the last inequality we used the fact that $r \geq \frac{3}{\varepsilon}$, which is guaranteed by Lemma 2.2.8. In conclusion, the number of edge-changes made when turning G into G' is less than εn^2.

Since G is ε-far from being induced $\mathcal{F}$-free, G' must contain an induced copy of some $F \in \mathcal{F}$. Suppose without loss of generality that $U_1, \ldots, U_p$ are the parts of $\mathcal{U}$ which contain vertices of this copy, and let X_i be the set of vertices of this copy which lie in U_i (for $1 \leq i \leq p$). From the definition of G' it follows that the sets $X_1, \ldots, X_p$ and the bipartite graphs (X_i, X_j), $1 \leq i < j \leq p$, are homogeneous. Note that by our choice of m we clearly have $\ell := v(F) \leq m$, and in particular $|X_i| \leq m$ for every $1 \leq i \leq p$.

We now show that the sets $W_{i,s}$, where $1 \leq i \leq p$ and $1 \leq s \leq |X_i|$, satisfy Condition 1 of Lemma 2.4.1 (with respect to F) *in the graph* G. First, for every $1 \leq i \leq p$, if X_i is a clique (resp. an independent set) then $G'[U_i]$ is a clique (resp. an independent set), which implies that $d_G(W_{i,s}, W_{i,t}) \geq 1 - \frac{1}{4m^2} \geq 1 - \frac{1}{2\ell^2}$ (resp. $d_G(W_{i,s}, W_{i,t}) \leq \frac{1}{4m^2} \leq \frac{1}{2\ell^2}$) for every $1 \leq s < t \leq m$ (see Item (b) above). Second, let $1 \leq i < j \leq p$. If (X_i, X_j) is a complete bipartite graph then $d_{G'}(U_i, U_j) = 1$ and hence $d_G(W_i, W_j) \geq 1 - \frac{1}{4m^2}$ (by Item (a) above). Now Item 2 in Lemma 2.2.8 implies that $d_G(W_{i,s}, W_{j,t}) \geq d_G(W_i, W_j) - \frac{1}{4m^2} \geq 1 - \frac{1}{2m^2} \geq 1 - \frac{1}{2\ell^2}$ for every $1 \leq s, t \leq m$. Similarly, if (X_i, X_j) is an empty bipartite graph then $d_{G'}(U_i, U_j) = 0$ and hence $d_G(W_i, W_j) \leq \frac{1}{4m^2}$. This implies that $d_G(W_{i,s}, W_{j,t}) \leq d_G(W_i, W_j) + \frac{1}{4m^2} \leq \frac{1}{2m^2} \leq \frac{1}{2\ell^2}$ for every pair $1 \leq s, t \leq m$.

We now apply Lemma 2.4.1 to the graph F, the sets $(W_{i,s} : 1 \leq i \leq p,\ 1 \leq s \leq |X_i|)$ and $\lambda = \zeta$, while noting that $|W_{i,s}| \geq \zeta n$ for every i, s, as guaranteed by Item 4 of Lemma 2.2.8. By Lemma 2.4.1, a sample of $\frac{12\ell}{\zeta}$ vertices from G contains an induced copy of F (and hence does not satisfy $\mathcal{P}^*_{\mathcal{F}}$) with probability at least $\frac{2}{3}$. It follows that $w_{\mathcal{P}^*_{\mathcal{F}}}(\varepsilon) \leq \frac{12\ell}{\zeta} \leq \frac{12m}{\zeta} = \text{poly}(1/\varepsilon)$, as required. This completes the proof. ■

Proof of Theorem 2.1.5. Again, our goal is to prove that $w_{\mathcal{P}^*_{\mathcal{F}}}(\varepsilon) = \text{poly}(1/\varepsilon)$. By Lemma 2.3.6, $\mathcal{P}^*_{\mathcal{F}}$ has a bipartite obstruction H. We may and will assume that both sides of H have the same size, h (as otherwise we can just add vertices to one of the sides). Let $\varepsilon < \frac{1}{2}$, and set

$$\gamma := \frac{1}{2} \cdot \rho_{2.2.8}\left(h, \frac{\varepsilon}{3}\right)^2,$$

and

$$\zeta := \zeta_{2.2.8}\left(h, 1, \frac{\varepsilon}{3}, \gamma\right).$$

Note that $\gamma = \text{poly}(\varepsilon)$ and hence also $\zeta = \text{poly}(\varepsilon)$.

Let G be an n-vertex graph which is ε-far from satisfying $\mathcal{P}^*_{\mathcal{F}}$. If G contains at least ζn^{2h} induced bipartite copies of H, then, just as in the proof of Theorem 4, a random sequence of $4h \cdot \zeta^{-1}$ vertices of G (sampled uniformly and independently) contains an induced bipartite copy of H, and hence does not satisfy $\mathcal{P}^*_{\mathcal{F}}$, with probability at least $\frac{2}{3}$. It follows that $w_{\mathcal{P}^*_{\mathcal{F}}}(\varepsilon) \leq 4h \cdot \zeta^{-1} = \text{poly}(1/\varepsilon)$, as required. Thus, in this case the required result holds.

Suppose, then, that G contains less than ζn^{2h} induced bipartite copies of H. We apply Lemma 2.2.8 to G with parameters $\delta = \frac{\varepsilon}{3}$, γ defined as above and $m = 1$, to get an equipartition $\mathcal{U} = \{U_1, ..., U_r\}$ and sets $W_i \subseteq U_i$ with the properties stated in the lemma.

Define a graph F on $[r]$ as follows. For $1 \leq i < j \leq r$, if $d(W_i, W_j) \geq 1 - \gamma$ then $(i, j) \in E(F)$ and if $d(W_i, W_j) \leq \gamma$ then $(i, j) \notin E(F)$ (by Item 2 of Lemma 2.2.8, one of these options must hold). We will show that F does not satisfy $\mathcal{P}^*_{\mathcal{F}}$. Let us first complete the proof based on this fact. By Lemma 2.2.8 we have $v(F) = r \leq \rho_{2.2.8}(h, \frac{\varepsilon}{3})^{-1} = \text{poly}(1/\varepsilon)$ and hence $\gamma \leq \frac{1}{2r^2}$. So by the definition of F, the sets $W_1, \ldots, W_r$ satisfy condition 1 of Lemma 2.4.1. By Item 4 of Lemma 2.2.8 we have $|W_i| \geq \zeta n$ for every $1 \leq i \leq r$. So Lemma 2.4.1 with $\lambda = \zeta$ implies that a sample of $12r/\zeta = \text{poly}(1/\varepsilon)$ vertices of G, sampled uniformly at random and independently, contains an induced copy of F, and hence does not satisfy $\mathcal{P}^*_{\mathcal{F}}$, with probability at least $\frac{2}{3}$.

It thus remains to show that F does not satisfy $\mathcal{P}^*_{\mathcal{F}}$. Assume, by contradiction, that F satisfies $\mathcal{P}^*_{\mathcal{F}}$. Since $\mathcal{P}^*_{\mathcal{F}}$ is closed under blowups (recall Definition 2.1.4), there is a function $g : V(F) \to \{0, 1\}$ such that every g-blowup of F satisfies $\mathcal{P}^*_{\mathcal{F}}$. Now let G' be the graph obtained from G by making the following changes.

(a) For every $1 \leq i < j \leq r$, if $(i, j) \in E(F)$ then turn (U_i, U_j) into a complete bipartite graph and if $(i, j) \notin E(F)$ then turn (U_i, U_j) into an empty bipartite graph.

(b) For every $1 \leq i \leq r$, if $g(i) = 1$ then turn U_i into a clique and if $g(i) = 0$ then turn U_i into an independent set.

Since G' is a g-blowup of F (see Definition 2.1.3), G' satisfies $\mathcal{P}^*_{\mathcal{F}}$. We now show that the number of edge-changes made in Items (a)-(b) is less than εn^2, which will stand in contradiction to the fact that G is ε-far from satisfying $\mathcal{P}^*_{\mathcal{F}}$.

The definitions of F and G' imply the following: for every $1 \leq i < j \leq r$, if the bipartite graph (U_i, U_j) is complete (resp. empty) in G' then $d_G(W_i, W_j) \geq 1 - \gamma$ (resp. $d_G(W_i, W_j) \leq \gamma$). As in the proof of Theorem 4, let $\mathcal{H}$ be the set of pairs $1 \leq i < j \leq r$ such that (U_i, U_j) is $\frac{\varepsilon}{3}$-homogeneous (in G) and such that $|d_G(W_i, W_j) - d_G(U_i, U_j)| \leq \frac{1}{4}$. Observe that if $(i,j) \in \mathcal{H}$ then the number of edge-changes made in the bipartite graph (U_i, U_j) is at most $\frac{\varepsilon}{3}|U_i||U_j|$. Indeed, let $(i,j) \in \mathcal{H}$ and suppose first that (U_i, U_j) is a complete bipartite graph in G'. Then $d_G(W_i, W_j) \geq 1 - \gamma \geq \frac{3}{4}$, implying that $d_G(U_i, U_j) \geq d_G(W_i, W_j) - \frac{1}{4} \geq \frac{1}{2}$. Hence actually $d_G(U_i, U_j) \geq 1 - \frac{\varepsilon}{3}$. The case that (U_i, U_j) is an empty bipartite graph in G' is symmetrical.

By Item 1 in Lemma 2.2.8, there are at most $\frac{\varepsilon}{3}r^2$ pairs $1 \leq i < j \leq r$ which are not in $\mathcal{H}$. It follows that the overall number of edge-changes made in Item (a) is at most $|\mathcal{H}| \cdot \frac{\varepsilon}{3} \cdot \left(\frac{n}{r}\right)^2 + \frac{\varepsilon}{3}r^2 \cdot \left(\frac{n}{r}\right)^2 \leq \frac{2\varepsilon}{3}n^2$. As for item (b), the number of edge-changes made there is at most $r\binom{n/r}{2} < \frac{n^2}{r} \leq \frac{\varepsilon}{3}n^2$, where in the last inequality we used the fact that $r \geq \frac{3}{\varepsilon}$ (as guaranteed by Lemma 2.2.8). Thus, the overall number of edge-changes made in Items (a)-(b) is less than εn^2, as required. ■

2.5 Detailed Proof of Theorem 6

Proof of Theorem 6. Let $\mathcal{P}$ be a semi-algebraic graph property defined by polynomials $f_1, \ldots, f_t \in \mathbb{R}[x_1, \ldots, x_{2k}]$ and a boolean function $\Phi : \{\text{true}, \text{false}\}^t \to \{\text{true}, \text{false}\}$. Let $\mathcal{F}$ be the family of all graphs which do not satisfy $\mathcal{P}$. As $\mathcal{P}$ is a hereditary property, we have $\mathcal{P} = \mathcal{P}^*_{\mathcal{F}}$. To prove the theorem, we only need to show that Conditions 1-2 of Theorem 2.1.5 are satisfied.

We start with Condition 1. The *VC-dimension* of a binary matrix A is the maximal integer $d \geq 0$ for which there is a $d \times 2^d$ submatrix B of A, such that the set of columns of B is the set of all 2^d binary vectors of length d. The VC-dimension of a graph is defined as the VC-dimension of its adjacency matrix. It is known that for every semi-algebraic graph property $\mathcal{P}$ there exists[10] $d = d(\mathcal{P})$ such that every graph which satisfies $\mathcal{P}$ has VC-dimension strictly less than d. Indeed, this follows from Warren's theorem on sign patterns of systems of polynomials, see for example [1]. Now let B be a $d \times 2^d$ binary matrix whose columns are all 2^d binary vectors of length d. Let H be the bipartite graph with sides $X = \{x_1, \ldots, x_d\}$ and $Y = \{y_1, \ldots, y_{2^d}\}$ such that $(x_i, y_j) \in E(H)$ if and only if $B_{i,j} = 1$. It is easy to see that no matter which graphs one puts on X and on Y (without changing the edges between X and Y), the resulting graph on $X \cup Y$ will not satisfy $\mathcal{P}$ since its VC-dimension will be at least $d = d(\mathcal{P})$. This means that H is a bipartite obstruction for $\mathcal{P}$, which implies (via Lemma 2.3.6) that $\mathcal{F}$ contains a bipartite graph, a co-bipartite graph and a split graph, as required.

As for Condition 2, let F be a graph on $V(F) = [p]$ which satisfies $\mathcal{P}$, and let $x_1, \ldots, x_p \in \mathbb{R}^k$ be witnesses to the fact that F satisfies $\mathcal{P}$. That is, for every $1 \leq i \neq j \leq p$ we have $(i,j) \in E(F)$ if and only if $\Phi\Big(f_1(x_i, x_j) \geq 0; \ldots; f_t(x_i, x_j) \geq 0\Big) = \text{true}$. We define a function $g : V(F) \to \{0, 1\}$ as follows: $g(i) = 1$ if

$$\Phi\Big(f_1(x_i, x_i) \geq 0; \ldots; f_t(x_i, x_i) \geq 0\Big) = \text{true}$$

and $g(i) = 0$ otherwise. We now show that every g-blowup of F satisfies $\mathcal{P}$. Let G be a g-blowup of F with a vertex partition $V(G) = P_1 \cup \cdots \cup P_p$ (as in Definition 2.1.3). Then for every $1 \leq i \leq p$, we simply assign the point x_i to every vertex of P_i. From the definition of a g-blowup and from our choice of g, it

[10] In fact, d can be bounded from above by a function of k, t, and the degrees of the polynomials $f_1, \ldots, f_t$.

follows that for every $1 \leq i, j \leq p$ and for every pair of distinct vertices $v_i \in P_i$, $v_j \in P_j$ we have that $(v_i, v_j) \in E(G)$ if and only if $\Phi\Big(f_1(x_i, x_j) \geq 0; \ldots; f_t(x_i, x_j) \geq 0\Big) = \text{true}$. Thus we have shown that $\mathcal{P}$ is closed under blowups, completing the deduction of Theorem 6 from Theorem 2.1.5. ■

2.6 Proof of the "If" Part of Theorem 8

In this section we prove the "if" part of Theorem 8. We start by proving a tournament analogue to the counting lemma 2.4.1.

Lemma 2.6.1. *For every f there is $\alpha = \alpha(f) > 0$ such that the following holds for every oriented graph F on f vertices. Let $X_1, ..., X_\ell$ be a partition of $V(F)$ such that $X_1, ..., X_\ell$ induce acyclic digraphs, and such that for every $1 \leq i < j \leq \ell$, either $E(X_j, X_i) = \emptyset$ or $E(X_i, X_j) = \emptyset$. Let $W_1, ..., W_\ell$ be pairwise-disjoint vertex sets in a tournament T having the following properties:*

1. *$|W_i| \geq 2^{f-1}$ for every $1 \leq i \leq \ell$.*
2. *For every $1 \leq i \neq j \leq \ell$, if $E(X_i, X_j) \neq \emptyset$ then $d(W_i, W_j) \geq 1 - \alpha$.*

Then T contains at least $\alpha \cdot \prod_{i=1}^{\ell} |W_i|^{|X_i|}$ copies of H.

Proof of Lemma 2.4.1. Set $m = 2^{f-1}$. We prove the lemma with $\alpha = \alpha(f) := \min\{\frac{1}{2}(fm)^{-2}, \frac{1}{2}m^{-f}\}$. For each $i = 1, ..., \ell$ we choose a subset $Y_i \subseteq W_i$ of size m uniformly at random. For $1 \leq i < j \leq \ell$, let us say that (Y_i, Y_j) *agrees* with (X_i, X_j) if $Y_i \to Y_j$ whenever $E(X_i, X_j) \neq \emptyset$ and $Y_j \to Y_i$ whenever $E(X_j, X_i) \neq \emptyset$. By the assumption on the pairs (W_i, W_j), the probability that (Y_i, Y_j) does not agree with (X_i, X_j) is at most αm^2. By the union bound, the probability that there is a pair $1 \leq i < j \leq \ell$ for which (Y_i, Y_j) does not agree with (X_i, X_j) is at most $\alpha m^2 \binom{\ell}{2} \leq \alpha m^2 f^2 \leq \frac{1}{2}$, where in the last inequality we used our choice of α.

Consider a choice of $Y_1, \ldots, Y_\ell$ such that (Y_i, Y_j) agrees with (X_i, X_j) for every $1 \leq i < j \leq \ell$. By Claim 2.3.3 and our choice of m, we get that Y_i contains a subset Z_i which induces a transitive tournament and has size $|Z_i| = |X_i|$. It follows that $Y := \bigcup_{i=1}^{\ell} Y_i$ contains a copy of F with Z_i playing the role of X_i. Now note that every copy of F of this form is contained in at most $\prod_{i=1}^{\ell} \binom{|W_i|-|X_i|}{m-|X_i|}$ such sets Y. Therefore the number of copies of F is at least

$$\frac{\frac{1}{2}\prod_{i=1}^{\ell}\binom{|W_i|}{m}}{\prod_{i=1}^{\ell}\binom{|W_i|-|X_i|}{m-|X_i|}} = \frac{\frac{1}{2}\prod_{i=1}^{\ell}\binom{|W_i|}{|X_i|}}{\prod_{i=1}^{\ell}\binom{m}{|X_i|}} \geq \frac{1}{2} \cdot \prod_{i=1}^{\ell}\left(\frac{|W_i|}{m}\right)^{|X_i|} = \frac{1}{2} \cdot m^{-f} \cdot \prod_{i=1}^{\ell}|W_i|^{|X_i|} \geq \alpha \cdot \prod_{i=1}^{\ell}|W_i|^{|X_i|}$$

■

Proof of the "if" part of Theorem 8. Let F be a 2-colorable oriented graph on f vertices. Let $\varepsilon > 0$, and let T be an n-vertex tournament which is ε-far from being F-free. Our goal is to show that T contains at least $\text{poly}(\varepsilon) \cdot n^f$ copies of F.

By Lemma 2.3.5, there is a bipartite tournament H with sides M and N that forces F. Note that we can clearly assume that $|M| = |N|$ (by adding additional vertices if necessary). Put $h := |M| = |N|$. Set

$$\delta := \min\left\{\frac{\varepsilon}{3},\ \alpha(f),\ \frac{1}{4}\right\},$$

where $\alpha(f)$ is from Lemma 2.6.1, and

$$\rho := \rho_{2.2.9}(h, \delta),$$

noting that $\rho = \text{poly}(\varepsilon)$ (as guaranteed by Lemma 2.2.9 and our choice of δ).

Assume first that T contains at least ρn^{2h} copies of H. Since H forces F, every copy of H in T contains a copy of F (as T is a tournament). On the other hand, every copy of F can be contained in at most n^{2h-f} such copies of H. It follows that T contains at least $n^{-(2h-f)} \cdot \rho n^{2h} = \rho n^f = \text{poly}(\varepsilon) \cdot n^f$ copies of F, giving the desired result in this case.

Suppose from now on that T contains less than ρn^{2h} copies of H. Then by Lemma 2.2.9, applied to T with δ as above, there is an equipartition $\mathcal{U} = \{U_1, ..., U_r\}$ and subsets $W_i \subseteq U_i$ with the properties stated in that lemma. Define $\mathcal{N}$ to be the set of pairs $1 \leq i < j \leq r$ for which either (a) (U_i, U_j) is not δ-homogeneous, or (b) $|d(W_i, W_j) - d(U_i, U_j)| > \frac{1}{4}$. By Item 1 in Lemma 2.2.9 we have $|\mathcal{N}| \leq \delta r^2 \leq \frac{\varepsilon}{3} r^2$. Hence,

$$\sum_{(i,j) \in \mathcal{N}} |U_i||U_j| \leq \frac{\varepsilon}{3} r^2 \cdot \left(\frac{n}{r}\right)^2 = \frac{\varepsilon}{3} n^2. \tag{2.3}$$

Note also that if $(i, j) \notin \mathcal{N}$, then either $d(U_i, U_j), d(W_i, W_j) \geq 1 - \frac{\varepsilon}{3}$ or $d(U_i, U_j), d(W_i, W_j) \leq \frac{\varepsilon}{3}$. Indeed, let $(i, j) \notin \mathcal{N}$. Then (U_i, U_j) is δ-homogeneous and $|d(W_i, W_j) - d(U_i, U_j)| \leq \frac{1}{4}$. Recall also that (W_i, W_j) is δ-homogeneous by Item 2 of Lemma 2.2.9. Now, if $d(W_i, W_j) \geq 1 - \delta$, then $d(U_i, U_j) \geq d(W_i, W_j) - \frac{1}{4} \geq \frac{3}{4} - \delta$, which implies, as (U_i, U_j) is δ-homogeneous (and $3/4 - \delta > \delta$), that $d(U_i, U_j) \geq 1 - \delta \geq 1 - \frac{\varepsilon}{3}$. By symmetry, if $d(W_i, W_j) \leq \frac{\varepsilon}{3}$ then $d(U_i, U_j) \leq \frac{\varepsilon}{3}$, as claimed.

Now let T' be the tournament obtained from T by making the following changes.

(a) Make U_i span a transitive tournament for every $i = 1, \ldots, r$.

(b) For every $1 \leq i < j \leq r$, if $d(W_i, W_j) \geq 1 - \delta$ then set $U_i \to U_j$ (i.e. orient all edges from U_i to U_j), and if $d(W_i, W_j) \leq \delta$ then set $U_j \to U_i$ (i.e. orient all edges from U_j to U_i). By Item of Lemma 2.2.9, one of these options has to hold.

Let us estimate the number of edge-reversals made in Items (a)-(b). The number of edge-reversals made in Item (a) is at most $r\binom{n/r}{2} \leq \frac{n^2}{r} \leq \frac{\varepsilon}{3} n^2$, where the last inequality uses the bound $r \geq 1/\delta \geq 3/\varepsilon$, which is guaranteed by Lemma 2.2.9. As for Item (b), note that if $(i, j) \notin \mathcal{N}$ then the number of reversals of edges with one endpoint in U_i and one in U_j is at most $\frac{\varepsilon}{3}|U_i||U_j|$. Using these facts and (2.3), we get that the total number of edge-reversals made in Items (a)-(b) is at most $\frac{\varepsilon}{3} n^2 + \sum_{i<j} \frac{\varepsilon}{3}|U_i||U_j| + \frac{\varepsilon}{3} n^2 < \varepsilon n^2$.

Since T is ε-far from being F-free and T' is obtained from T by reversing less than εn^2 edges, T' must contain a copy of F. Fix such a copy of F, and suppose without loss of generality that $U_1, \ldots, U_\ell$ are the parts of the partition $\mathcal{U}$ which intersect the vertex-set of this copy. For $i = 1, ..., \ell$, let X_i be the set of vertices of U_i which participate in this copy of F. The way we constructed T' from T in Items (a)-(b) implies the following: in the graph F, the sets $X_1, \ldots, X_\ell$ span acyclic digraphs and for every $1 \leq i < j \leq \ell$, either $E(X_j, X_i) = \emptyset$ or $E(X_i, X_j) = \emptyset$. Moreover, for every $1 \leq i \neq j \leq \ell$, if $E(X_i, X_j) \neq \emptyset$ then $U_i \to U_j$ in T', implying that $d(W_i, W_j) \geq 1 - \delta \geq 1 - \alpha(f)$ in T (here we used our choice of δ). Finally, by Item 3 of Lemma 2.2.9 we have $|W_i| \geq \rho n$ for every $i = 1, \ldots, \ell$. So if n is large enough then $|W_i| \geq 2^{f-1}$ for every $i = 1, \ldots, \ell$. We conclude that $W_1, \ldots, W_\ell$ satisfy the conditions of Lemma 2.6.1 *in the tournament* T with respect to the partition $V(F) = X_1 \cup \cdots \cup X_\ell$ of the oriented graph F. Now, By Lemma 2.6.1, the

number of copies of F in T is at least

$$\alpha(f) \cdot \prod_{i=1}^{\ell} |W_i|^{|X_i|} \geq \alpha(f) \cdot (\rho n)^{|X_1|+\cdots+|X_\ell|} = \alpha(f) \cdot \rho^f n^f = \mathrm{poly}(\varepsilon) \cdot n^f,$$

as required. This completes the proof. ■

2.7 A Variant of the Ruzsa-Szemerédi Construction

In this section we describe a Ruzsa-Szemerédi-type construction that will be used in the proofs of Theorems 5, 2.1.1 and 2.1.2, as well as in the proof of the "only-if" part of Theorem 8. We start with the following Behrend-type construction.

Lemma 2.7.1. *For every $k \geq 2$ there is $\alpha = \alpha(k)$ such that for every integer m there is a set $S \subseteq [m]$, $|S| \geq \frac{m}{e^{\alpha\sqrt{\log m}}}$, with the following property: Let $2 \leq \ell \leq k$ and let $a_1, ..., a_\ell \geq 1$ be integers satisfying $a_1 + \cdots + a_\ell \leq k$. Then the only solutions to the equation*

$$a_1 s_1 + a_2 s_2 + \cdots + a_\ell s_\ell = (a_1 + \cdots + a_\ell) s_{\ell+1}$$

with $s_1, ..., s_{\ell+1} \in S$ are the trivial ones, i.e. $s_1 = s_2 = \cdots = s_\ell = s_{\ell+1}$.

Lemma 2.7.1 is a variant of Behrend's construction [20] of a large subset of $[m]$ without a 3-term arithmetic progression (note that the case $k = \ell = 2$ and $a_1 = a_2 = 1$ exactly corresponds to a 3-term arithmetic progression). It is easy to show (see e.g. [97] and [2]) that the same exact proof actually works for any fixed convex equation, and that moreover, it works "simultaneously" for all convex equations (for fixed k), thus establishing the above lemma. We therefore omit its proof.

The following lemma is our variant of the Ruzsa-Szemerédi construction (cf. [98] and [2]).

Lemma 2.7.2. *For every $h \geq 3$ there are $\delta_0 = \delta_0(h)$ and $\beta = \beta(h)$ such that for every $\delta < \delta_0$ there is a graph $R = R(h, \delta)$ with a vertex-partition $V(R) = V_1 \uplus \cdots \uplus V_h$, such that the following holds.*

1. *$|V(R)| \geq (1/\delta)^{\beta \log(1/\delta)}$.*
2. *$E(R)$ is the union of at least $\delta|V(R)|^2$ pairwise edge-disjoint h-cliques, each of the form $\{v_1, \ldots, v_h\}$ with $v_i \in V_i$, $i = 1, \ldots, h$.*
3. *For every $3 \leq t \leq h$ and for every sequence $1 \leq i_1 < i_2 \cdots < i_t \leq h$, R contains at most $|V(R)|^2$ (not necessarily induced) cycles of the form $v_{i_1} v_{i_2} \ldots v_{i_t} v_{i_1}$ with $v_{i_j} \in V_{i_j}$ $(1 \leq j \leq t)$.*

Proof. Let $0 < \delta < \delta_0$ (for $\delta_0 = \delta_0(h)$ to be chosen later), and let m be the largest integer satisfying

$$\delta \leq \frac{1}{(h+1)^4 e^{\alpha\sqrt{\log m}}} \tag{2.4}$$

where $\alpha = \alpha(h-1)$ is from Lemma 2.7.1. It is easy to check that

$$m \geq e^{\alpha^{-2} \log^2\left(\frac{1}{\delta(h+1)^4}\right)} \geq (1/\delta)^{\beta \log(1/\delta)}, \tag{2.5}$$

where the second inequality holds provided that we choose $\beta = \beta(h)$ to be small enough, and provided that δ is sufficiently small with respect to h (we choose δ_0 accordingly).

Let $S \subseteq [m]$ be the set obtained by applying Lemma 2.7.1 with $k = h-1$. For each $j = 1, \ldots, h$ set $V_j = \{1, 2, ..., jm\}$. With a slight abuse of notation, we think of $V_1, ..., V_h$ as disjoint sets. The vertex-set of R is $V(R) = V_1 \uplus \cdots \uplus V_h$. By (2.5) we have $|V(R)| = \binom{h+1}{2} m \geq (1/\delta)^{\beta \log(1/\delta)}$, as required.

We now specify the edges of R. For every $x \in [m]$ and $s \in S$, set

$$A(x,s) := \{x, x+s, x+2s, \ldots, x+(h-1)s\},$$

and put a clique on $A(x,s)$, in which $x+(j-1)s$ is taken from V_j for every $j = 1, \ldots, h$. Note that for every $(x,s), (x',s') \in [m] \times S$, if $(x,s) \neq (x',s')$ then $|A(x,s) \cap A(x',s')| \leq 1$. Indeed, if $|A(x,s) \cap A(x',s')| \geq 2$ then there are $0 \leq i < j \leq h-1$ for which $x + is = x' + is'$ and $x + js = x' + js'$. Solving this system of equations yields $(x,s) = (x',s')$, as required. So the cliques that we defined are edge-disjoint, as required in Item 2 of the lemma. By the lower bound on $|S|$ in Lemma 2.7.1, and by (2.4), the number of these cliques is

$$m \cdot |S| \geq \frac{m^2}{e^{\alpha\sqrt{\log m}}} \geq \delta(h+1)^4 m^2 \geq \delta |V(R)|^2.$$

To finish the proof, it remains to establish Item 3. Fix any $t \geq 3$ and any sequence of indices $1 \leq i_1 < i_2 \cdots < i_t \leq h$. We will show that for every cycle of the form $v_{i_1} v_{i_2} \ldots v_{i_t} v_{i_1}$ with $v_{i_j} \in V_{i_j}$, there are $x \in [m]$ and $s \in S$ such that $v_{i_1}, v_{i_2}, \ldots, v_{i_t} \in A(x,s)$. This will show that the cycles of this form are pair-disjoint, where two subgraphs are called *pair disjoint* if they share at most one vertex. This in turn will imply that there are at most $|V(R)|^2$ such cycles.

Let $v_{i_1} v_{i_2} \ldots v_{i_t} v_{i_1}$ be a cycle in R with $v_{i_j} \in V_{i_j}$ for every $1 \leq j \leq t$. By the construction of R, for every $j = 1, \ldots, t$ there is $(x_j, s_j) \in [m] \times S$ such that $\{v_{i_j}, v_{i_{j+1}}\} \subseteq A(x_j, s_j)$, with indices taken modulo t. This means that

$$v_{i_{j+1}} - v_{i_j} = (i_{j+1} - i_j) s_j \tag{2.6}$$

for every $1 \leq j \leq t-1$, and also that $v_{i_t} - v_{i_1} = (i_t - i_1) s_t$. Setting $a_j := i_{j+1} - i_j$ for $1 \leq j \leq t-1$, we see that

$$a_1 s_1 + a_2 s_2 + \cdots + a_{t-1} s_{t-1} = (a_1 + \cdots + a_{t-1}) s_t.$$

Since S was chosen via Lemma 2.7.1, we must have $s_1 = s_2 = \cdots = s_t$. Hence, by (2.6) we have that $v_{i_j} = v_{i_1} + (i_j - i_1) s_1$ for every $1 \leq j \leq t$. By the definition of $A(x_1, s_1)$ and by the fact that $v_{i_1} \in A(x_1, s_1)$, we get that $v_{i_1}, v_{i_2}, \ldots, v_{i_t} \in A(x_1, s_1) = \cdots = A(x_t, s_t)$, as required. ■

2.8 Homomorphisms and Cores

In this section we survey some properties of homomorphisms and cores of graphs and of ordered graphs. These will be used in subsequent sections.

2.8.1 Homomorphisms and Cores of Graphs

Recall that a *homomorphism* from a graph G_1 to a graph G_2 is a map $f : V(G_1) \to V(G_2)$ such that for every $u, v \in V(G_1)$, if $(u,v) \in E(G_1)$ then $(f(u), f(v)) \in E(G_2)$. We write $G_1 \leq_{\text{hom}} G_2$ — and say that

G_1 is homomorphic to G_2 — if there is a homomorphism from G_1 to G_2. Notice that the relation $\leq_{\text{hom}}$ is transitive. For a graph G, the *core* of G, denoted $C(G)$, is an induced subgraph of G to which there is[11] a homomorphism from G, and which has the smallest number of vertices among all such induced subgraphs of G. We say that a graph G is *a core* if $C(G) = G$. Observe that the core of any graph is a core, and that every homomorphism from a core to itself is an isomorphism. It is now easy to check that for every pair of cores C_1, C_2, if $C_1 \leq_{\text{hom}} C_2$ and $C_2 \leq_{\text{hom}} C_1$ then C_1 and C_2 are isomorphic. This in turn implies that the core of a graph is defined uniquely, up to isomorphism. We refer the reader to [74] for detailed proofs of these claims, as well as an overview of the topic of graph homomorphisms and cores.

Let $\mathcal{F}$ be a finite family of graphs and consider the set $\mathcal{C} = \mathcal{C}(\mathcal{F}) = \{C(F) : F \in \mathcal{F}\}$. As we explained above, $(\mathcal{C}, \leq_{\text{hom}})$ is a poset in the following sense: for every $C_1, C_2 \in \mathcal{C}$, if $C_1 \leq_{\text{hom}} C_2$ and $C_2 \leq_{\text{hom}} C_1$, then C_1 and C_2 are isomorphic. Namely, $\leq_{\text{hom}}$ is a partial order on the set of equivalence classes of $\mathcal{C}$ under the equivalence relation of graph isomorphism. Let $K(\mathcal{F})$ be a minimal element of the poset $(\mathcal{C}, \leq_{\text{hom}})$; i.e., $K(\mathcal{F})$ is an (arbitrary) element of an (arbitrary) minimal equivalence class. The minimality of $K(\mathcal{F})$ implies that for every $C \in \mathcal{C}$, if there is a homomorphism from C to $K(\mathcal{F})$ (namely, if $C \leq_{\text{hom}} K(\mathcal{F})$), then C is isomorphic to $K(\mathcal{F})$. The key property of the graph $K(\mathcal{F})$ is given by the following proposition.

Proposition 2.8.1. *For every $F \in \mathcal{F}$ and for every homomorphism $f : F \to K(\mathcal{F})$, there is a set $X \subseteq V(F)$ such that $f|_X$ is an isomorphism onto $K(\mathcal{F})$.*

Proof. Let $C = C(F)$ be the core of F. Since $f|_{V(C)}$ is a homomorphism from C to $K(\mathcal{F})$, and since $K(\mathcal{F})$ is minimal (in the sense described above), we have that C is isomorphic to $K(\mathcal{F})$. Fix an isomorphism $g : K(\mathcal{F}) \to C$. Then $f|_{V(C)} \circ g$ is a homomorphism from $K(\mathcal{F})$ to itself, and is hence an isomorphism (since $K(\mathcal{F})$ is a core). As g and $f|_{V(C)} \circ g$ are both isomorphisms, $f|_{V(C)}$ must also be an isomorphism. So the assertion of the proposition holds with $X = V(C)$. ■

2.8.2 Homomorphisms and Cores of Ordered Graphs

This section is concerned with ordered graphs, namely graphs having linear ordering of their vertices. For simplicity, we will equip graphs with such an ordering by assuming that the vertex-sets of all graphs are subset of $\mathbb{N}$. We will always assume that no two vertices of a graph are labeled with the same natural number. A subgraph of a graph G is always assumed to inherit its vertex-labeling from G.

Let G, G' be (undirected) graphs. A homomorphism $g : V(G) \to V(G')$ is said to be *order-preserving* if $g(i) \leq g(j)$ for every pair of vertices $i, j \in V(G)$ with $i \leq j$. We write $G \leq_{\text{ord-hom}} G'$ if there is an order-preserving homomorphism from G to G'. Notice that the relation $\leq_{\text{ord-hom}}$ is transitive (the composition of order-preserving homomorphisms is also an order-preserving homomorphism). An *order-preserving isomorphism* is an order-preserving homomorphism which is a graph isomorphism. We write $G \cong_{\text{ord}} G'$ if there is an order-preserving isomorphism between G and G'. [12]

The *ordered core* of G is a smallest (with respect to number of vertices) subgraph of G to which there is an order-preserving homomorphism from G. (Recall that the ordered core of G is assumed to inherit the same vertex-labeling as it had in G.) Notice that by definition, there is no order-preserving homomorphism

[11] We note that our definition of a core is a bit different (but equivalent) to the usual definition of a core, see e.g. [74].

[12] Notice that two isomorphic labeled graphs do not necessarily have an *order-preserving* isomorphism between them. Moreover, if two graphs have an order-preserving isomorphism between them then it is unique, assuming that the vertices in each graph have different labels, which we always do in our setting.

from the ordered core of G to a proper induced subgraph thereof. We say that a graph is an *ordered core* if it is the ordered core of itself.

Proposition 2.8.2. *Let G_1, G_2 be ordered cores. If $G_1 \leq_{ord\text{-}hom} G_2$ and $G_2 \leq_{ord\text{-}hom} G_1$ then $G_1 \cong_{ord} G_2$.*

Proof. By assumption there exist order-preserving homomorphisms $g : G_1 \to G_2$ and $h : G_2 \to G_1$. Then $h \circ g$ is an order-preserving homomorphism from G_1 to itself. Since G_1 is a core, h must be surjective. The same argument shows that g is surjective. So g, h are bijections, and since g, h are order-preserving, we must have $h = g^{-1}$. It follows that g, h are order-preserving isomorphisms, as required. ■

Proposition 2.8.2 shows that the ordered core of a graph is unique up to order-preserving isomorphism.

Proposition 2.8.3. *Let G_1, G_2 be a pair of ordered cores and suppose that $G_1 \cong_{ord} G_2$. Then every order-preserving homomorphism $g : G_1 \to G_2$ is an order-preserving isomorphism.*

Proof. By definition, there is an order-preserving isomorphism $h : G_2 \to G_1$. Now $h \circ g$ is an order-preserving homomorphism from G_1 to itself. By the definition of an ordered core, $h \circ g$ is a bijection. Since g, h are order-preserving, we have that $h \circ g$ is the identity map and hence $g = h^{-1}$. So we see that g is an isomorphism, as required. ■

Let F be an oriented graph, and suppose again that the vertices of F are labeled with (distinct) natural numbers. We say that an edge $(i, j) \in E(F)$ is a *forward-edge* if $i < j$ and *backward-edge* (or *backedge*) otherwise. The *backedge graph* of F is the (undirected, ordered) graph on $V(F)$ in which $\{i, j\}$ is an edge if and only if $i < j$ and $j \to i$. Note that the backedge graph depends on the labeling of the vertices of F; backedge graphs corresponding to different labelings may not be isomorphic (even as unordered graphs). The following simple proposition relates the (directed) chromatic number of an oriented graph to the (undirected) chromatic numbers of its backedge graphs.

Proposition 2.8.4. *An oriented graph F is k-colorable (as a digraph) if and only if there is a labeling of the vertices of F for which the corresponding backedge graph is k-colorable (as an undirected graph).*

Proof. Assume first that there is a labeling of $V(F)$ such that the corresponding backedge graph, G, is k-colorable (as an undirected graph). Let $V(G) = U_1 \cup \cdots \cup U_k$ be a partition of $V(G) = V(F)$ into independent sets. Then for every $i = 1, \ldots, k$, the digraph $F[U_i]$ is acyclic because all of the edges inside U_i are forward-edges. It follows that F is k-colorable (as an digraph).

Now assume that F is k-colorable (as a digraph), and let $V(F) = U_1 \cup \cdots \cup U_k$ be a partition of $V(F)$ such that $F[U_1], \ldots, F[U_k]$ are acyclic digraphs. For every $i = 1, ..., k$, label the vertices of U_i such that there are no backedges of F inside U_i (this is possible because $F[U_i]$ is acyclic). Then $U_1, \ldots, U_k$ are independent set in the backedge graph corresponding to this labeling. It follows that this backedge graph is k-colorable (as an undirected graph). ■

For an oriented f-vertex graph F, we define a family of (undirected) graphs $\mathcal{C} = \mathcal{C}(F)$, all labeled with the numbers $1, \ldots, f$, as follows. We go through all $f!$ vertex-labelings of F using the labels $1, ..., f$, and for each labeling we take the ordered core of the corresponding backedge graph. We then let $\mathcal{C} = \mathcal{C}(F)$ be the set of all these ordered cores.

Proposition 2.8.2 implies that $(\mathcal{C}, \leq_{\text{ord-hom}})$ is a poset in the following sense: for every $C_1, C_2 \in \mathcal{C}$, if $C_1 \leq_{\text{ord-hom}} C_2$ and $C_2 \leq_{\text{ord-hom}} C_1$ then $C_1 \cong_{\text{ord}} C_2$. In other words, $\leq_{\text{ord-hom}}$ is a partial order on the set of equivalence classes of $\mathcal{C}$ under the equivalence relation $\cong_{\text{ord}}$. Now, let $K(F)$ be a minimal element of the poset $(\mathcal{C}, \leq_{\text{ord-hom}})$, i.e. $K(F)$ is an (arbitrary) element of a minimal equivalence class. The minimality of $K(F)$ implies that for every $C \in \mathcal{C}$, if there is an order-preserving homomorphism from C to $K(F)$ (namely if $C \leq_{\text{ord-hom}} K(F)$) then $C \cong_{\text{ord}} K(F)$. The key property of the graph $K(F)$ is as follows.

Proposition 2.8.5. *Let F be an oriented graph. Fix any vertex-labeling of F, and let G be the corresponding backedge graph. Then for every order-preserving homomorphism $g : G \to K(F)$, there is a set $X \subseteq V(F) = V(G)$ such that $g|_X$ is an isomorphism onto $K(F)$.*

Proof. Let C be the ordered core of G. Then $g|_{V(C)}$ is an order-preserving homomorphism from C to $K(F)$. By the minimality of $K(F)$ we have $C \cong_{\text{ord}} K(F)$. Now Proposition 2.8.3 implies that $g|_{V(C)}$ is an order-preserving isomorphism. So we see that the assertion of Proposition 2.8.5 holds with $X = V(C)$. ■

Corollary 2.8.6. *If F is a non-2-colorable oriented graph, then the graph $K(F)$ contains a cycle $c_1, c_2, \ldots, c_\ell, c_1$ such that the following holds. Fix any vertex-labeling of F and let G be the corresponding backedge graph. Then for every order-preserving homomorphism $g : G \to K(F)$, there are vertices $u_1 \in g^{-1}(c_1), \ldots, u_\ell \in g^{-1}(c_\ell)$ such that $u_1 u_2 \ldots u_\ell u_1$ is a cycle in G.*

Proof. By the definition of $K(F)$, there is a vertex-labeling of F such that $K(F)$ is the ordered core of the corresponding backedge graph, G_0. By Proposition 2.8.4, G_0 is not 2-colorable (as an undirected graph) and therefore contains an odd cycle. It is easy to see that the homomorphic image of an odd cycle must itself contain an odd cycle. It follows that $K(F)$ contains an odd cycle, since $K(F)$ is the homomorphic image of G_0 (recall the definition of an ordered core). It is now easy to see that Corollary 2.8.6 follow from Proposition 2.8.1. ■

2.9 Proof of Theorems 5, 2.1.1 and 2.1.2

2.9.1 Proof of Theorems 5 and 2.1.2

Theorems 5 and 2.1.2 will be derived from the following theorem.

Theorem 2.9.1. *For every $h \geq 3$ there is $\varepsilon_0 = \varepsilon_0(h)$ such that the following holds for every $\varepsilon < \varepsilon_0$ and for every non-bipartite graph H on h vertices. Let K be the core of H. For every $n \geq n_0(\varepsilon)$, there is a graph on n vertices with the following properties.*

1. *G is homomorphic to K.*
2. *G is ε-far from being induced-H-free.*
3. *G contains at most $\varepsilon^{\Omega(\log(1/\varepsilon))} n^k$ (not necessarily induced) copies of K, where $k = |V(K)|$.*

Proof. Fix a homomorphism $\varphi : H \to K$. Since H is not bipartite, and since the homomorphic image of a non-bipartite graph is itself non-bipartite, we get that K is not bipartite, and hence contains an odd cycle. Label the vertices of K by $a_1, \ldots, a_k$ so that $a_1 a_2 \ldots a_t a_1$ is an odd cycle. Define $H_i = \varphi^{-1}(a_i)$ for

$i = 1, \ldots, k$. Label the vertices of H by $1, \ldots, h$ so that for each $1 \leq i < j \leq k$, the labels of the vertices in H_i are smaller than the labels of the vertices in H_j.

Let $\varepsilon > 0$. We will assume that ε is small enough where needed (in other words, we will choose $\varepsilon_0(h)$ implicitly). Assuming that $\varepsilon < \delta_0(h)/h^2$ (where $\delta_0(h)$ is from Lemma 2.7.2), let $R = R(h, h^2\varepsilon)$ be the graph from Lemma 2.7.2. Recall that $V(R) = V_1 \uplus \cdots \uplus V_h$, and put $r := |V(R)|$.

We now define a graph S on $V(R)$ as follows. By Item 2 of Lemma 2.7.2, R contains a collection $\mathcal{H}$ of at least $\varepsilon h^2 r^2$ pair-disjoint h-cliques, each of the form $\{v_1, \ldots, v_h\}$ with $v_i \in V_i$ $(1 \leq i \leq h)$. For every $\{v_1, \ldots, v_h\} \in \mathcal{H}$, we let $S[\{v_1, \ldots, v_h\}]$ span an induced copy of H in which v_i plays the role of i for every $i \in [h] = V(H)$. The resulting graph is S. It is clear from the definition that $\mathcal{H}$ is a collection of pair-disjoint induced copies of H in S.

Let n be a large integer which we assume, for simplicity of presentation, to be divisible by $r = |V(S)|$. Let G be the $\frac{n}{r}$-blowup of S; that is, G is the graph obtained by replacing each vertex $v \in V(S)$ with an independent set $B(v)$ of size $\frac{n}{r}$ (where distinct vertices are replaced by disjoint sets), replacing edges with complete bipartite graphs and replacing non-edges with empty bipartite graphs. Clearly $|V(G)| = n$. For $1 \leq i \leq h$ put $B(V_i) := \bigcup_{v \in V_i} B(v)$. Observe that the map which sends $\bigcup_{i \in H_j} B(V_i)$ to a_j for every $1 \leq j \leq k$, is a homomorphism from G to K. This establishes Item 1 in the statement of the theorem.

As we already showed, $\mathcal{H}$ is a collection of at least $\varepsilon h^2 r^2$ pair-disjoint induced copies of H in S. We call these copies the *base copies* of H. For every base copy $\{v_1, \ldots, v_h\} \in \mathcal{H}$, Claim 2.3.4 (with parameters $t = n/r$ and $q = h$) gives a collection of at least $(n/rh)^2$ pair-disjoint induced copies of H in G, each of the form $\{x_1, \ldots, x_h\}$ with $x_i \in B(v_i)$. We say that these copies are *derived* from $\{v_1, \ldots, v_h\}$. Since the base copies are pair-disjoint, two copies which are derived from different base copies are also pair-disjoint. Thus, G contains a collection of at least $|\mathcal{H}| \cdot (n/rh)^2 \geq \varepsilon h^2 r^2 \cdot (n/rh)^2 = \varepsilon n^2$ pair-disjoint induced copies of H. This shows that G is ε-far from being induced H-free.

To finish the proof it remains to show that G contains at most $\varepsilon^{\Omega(\log(1/\varepsilon))} n^k$ copies of K. To this end, we now show that copies of K in G must be of a special form, which will allow us to bound their number. The details follow. Consider a copy of K in G. For each $j = 1, \ldots, k$, let $U_j \subseteq V(G)$ be the set of vertices of this copy that are contained in $\bigcup_{i \in H_j} B(V_i)$. Notice that the map that sends U_j to a_j (for each $j = 1, \ldots, k$) is a homomorphism from K to itself. By the property of a core (see Section 2.8.1), this map is an isomorphism. Thus, $|U_j| = 1$ for every $1 \leq j \leq k$. Write $U_j = \{u_j\}$, and note that for each $1 \leq i < j \leq k$ we have $(u_i, u_j) \in E(G)$ if and only if $(a_i, a_j) \in E(K)$. Now the fact that $a_1 a_2 \ldots a_t a_1$ is a cycle in K implies that $u_1, \ldots, u_t, u_1$ is a cycle in G. For each $1 \leq j \leq k$, let $i_j \in H_j$ be such that $u_j \in B(V_{i_j})$ and let $v_{i_j} \in V_{i_j}$ be such that $u_j \in B(v_{i_j})$. Then $i_1 < i_2 < \cdots < i_t$ due to the way we labeled the vertices of H. Moreover $v_{i_1} v_{i_2} \ldots v_{i_t} v_{i_1}$ is a cycle in S because G is a blowup of S. Finally, it follows from the definition of S that $v_{i_1} v_{i_2} \ldots v_{i_t} v_{i_1}$ must be a cycle in R.

We thus proved that every copy of K in G contains vertices $u_1, \ldots, u_t$ with the following property: there is an increasing sequence $1 \leq i_1 < i_2 < \cdots < i_t \leq h$ and vertices $v_{i_j} \in V_{i_j}$ (for $1 \leq j \leq t$) such that $u_j \in B(v_{i_j})$ and such that $v_{i_1} v_{i_2} \ldots v_{i_t} v_{i_1}$ is a cycle in R. For every increasing sequence $(i_1, i_2, \ldots, i_t)$, Lemma 2.7.2 states that R contains at most r^2 cycles of the form $v_{i_1} v_{i_2} \ldots v_{i_t} v_{i_1}$ with $v_{i_j} \in V_{i_j}$. Therefore, the number of copies of K in G that correspond to a specific increasing sequence is at most $r^2 (n/r)^t n^{k-t} \leq n^k/r$ (here we used the obvious fact that $t \geq 3$). By taking the union bound over all $\binom{h}{t}$ increasing sequences $(i_1, i_2, \ldots, i_t)$ and using the inequality $r \geq (1/h^2\varepsilon)^{\beta(h)\log(1/h^2\varepsilon)} \geq (1/\varepsilon)^{\Omega(\log 1/\varepsilon)}$ (which is guaranteed by Item 1 of Lemma 2.7.2), we get that the number of copies of K in G is at most $\binom{h}{t} n^k/r \leq \varepsilon^{\Omega(\log 1/\varepsilon)} n^k$.

This completes the proof. ■

Proof of Theorem 2.1.2. By Theorem 2.9.1, for every sufficiently small $\varepsilon > 0$ and for every $n \geq n_0(\varepsilon)$, there is a graph G on n vertices which is ε-far from being induced H-free yet contains at most $\varepsilon^{\Omega(\log 1/\varepsilon)} n^k$ (not necessarily induced) copies of K, the core of H. As K is a subgraph of H, G contains at most $\varepsilon^{\Omega(\log 1/\varepsilon)} n^k \cdot n^{h-k} = \varepsilon^{\Omega(\log 1/\varepsilon)} n^h$ (not necessarily induced) copies of H. ■

Proof of Theorem 5. Write $\mathcal{F} = \{F_1, \ldots, F_\ell\}$. By symmetry (with respect to graph complementation), it is enough to prove that there is $1 \leq i \leq \ell$ for which F_i is bipartite. Assume, by contradiction, that F_i is not bipartite for every $1 \leq i \leq \ell$. We will show that for every sufficiently small $\varepsilon > 0$ and for every $n \geq n_0(\varepsilon)$, there is a graph G which is ε-far from being induced $\mathcal{F}$-free and yet contains at most $\varepsilon^{\Omega(\log 1/\varepsilon)} n^{v(F_i)}$ copies of F_i for every $1 \leq i \leq \ell$ (where the implicit constant in the exponent depends only on $\mathcal{F}$). This will imply that $\mathcal{P}^*_{\mathcal{F}}$ is not easily testable, a contradiction.

Let $K = K(\mathcal{F})$ be the graph defined in Section 2.8.1. Then K is the core of one of the graphs $F_1, \ldots, F_\ell$. Let us assume, without loss of generality, that K is the core of F_1. We claim that the graph G, obtained by applying Theorem 2.9.1 with $H = F_1$ (and K), satisfies our requirements. Evidently, G is ε-far from being induced $\mathcal{F}$-free because it is ε-far from being induced F_1-free.

By Item 1 of Theorem 2.9.1, there is a homomorphism $g : G \to K$. Now let $1 \leq i \leq \ell$, and consider an embedding $f : F_i \to G$ of F_i into G. Then $g \circ f$ is a homomorphism from F_i to K. By Proposition 2.8.1, there is a set $X \subseteq V(F_i)$ such that $(g \circ f)|_{V(X)}$ is an isomorphism onto K. This means that $f(V(F_i)) \subseteq V(G)$ contains a copy of K. We conclude that every copy of F_i in G contains a copy of K. By Item 3 of Theorem 2.9.1, G contains at most $\varepsilon^{\Omega(\log 1/\varepsilon)} n^k$ copies of K. It follows that G contains at most $\varepsilon^{\Omega(\log 1/\varepsilon)} n^k \cdot n^{v(F_i)-k} = \varepsilon^{\Omega(\log 1/\varepsilon)} n^{v(F_i)}$ copies of F_i, as required. ■

2.9.2 Proof of Theorem 2.1.1

Let M be the *complement* of the 7-vertex graph with vertex-set $\{1,2,3,4,5,6,7\}$ and edge-set $\{\{1,2\},\{3,4\},\{5,6\}\}$. It is easy to see that M is co-bipartite. We will prove Theorem 2.1.1 with $F_1 = C_8$ (the cycle on 8 vertices) and $F_2 = M$. We need the following lemma, which we prove later.

Lemma 2.9.2. *Let G be a graph admitting a vertex partition $V(G) = X_1 \cup \cdots \cup X_8$ such that*

- *X_1, X_3, X_5, X_7 are cliques and X_2, X_4, X_6, X_8 are independent sets.*
- *The only edges between the parts $X_1, \ldots, X_8$ are between consecutive parts; that is, for every $1 \leq i \neq j \leq 8$, we have $E(X_i, X_j) = \emptyset$ unless $|i - j| \equiv \pm 1 \pmod 8$.*

Then the following holds.

1. *Every induced copy of C_8 in G is of the form $x_1x_2...x_8x_1$, where $x_i \in X_i$.*
2. *G is induced M-free.*

Proof of Theorem 2.1.1. Set $F_1 = C_8$ and $F_2 = M$. We will show that for every sufficiently small $\varepsilon > 0$ and for every $n \geq n_0(\varepsilon)$ there is a graph G on n vertices which is ε-far from being induced $\{F_1, F_2\}$-free

yet contains at most[13] $\varepsilon^{\Omega(\log 1/\varepsilon)}n^{v(F_i)}$ induced copies of F_i for both $i = 1, 2$. This will imply that $\mathcal{P}^*_{\{F_1,F_2\}}$ is not easily testable.

Let $\varepsilon \in \left(0, \frac{\delta_0(8)}{64}\right)$, where $\delta_0(8)$ is from Lemma 2.7.2. Let $R = R(8, 64\varepsilon)$ be the graph obtained by applying Lemma 2.7.2. Recall that $V(R) = V_1 \uplus \cdots \uplus V_8$, and put $r = |V(R)|$. For simplicity of presentation, we assume that n is divisible by r. We define a graph G on an $\frac{n}{r}$-blowup of R; that is, we replace each vertex $v \in V(R)$ with a vertex-set $B(v)$ of size $\frac{n}{r}$, where the sets $(B(v) : v \in V(R))$ are pairwise-disjoint. Put $B(V_i) := \bigcup_{v\in V_i} B(v_i)$ for $1 \leq i \leq 8$. The edges of G are defined as follows: $B(V_1), B(V_3), B(V_5), B(V_7)$ are cliques and $B(V_2), B(V_4), B(V_6), B(V_8)$ are independent sets. To define the edges between the sets $B(V_1), \ldots, B(V_8)$, recall that by Lemma 2.7.2, R contains a collection $\mathcal{H}$ of at least $64\varepsilon r^2$ pairwise edge-disjoint cliques, each of the form $\{v_1, \ldots, v_8\}$ with $v_i \in V_i$. For each such clique $\{v_1, \ldots, v_8\} \in \mathcal{H}$ we put a blowup of C_8 on the sets $B(v_1), \ldots, B(v_8)$; namely, for each $(x_1, \ldots, x_8) \in B(v_i) \times \cdots \times B(v_8)$, $x_1x_2 \ldots x_8x_1$ is an induced 8-cycle in G. Notice that G satisfies the assumptions of Lemma 2.9.2 with $X_i = B(V_i)$. Thus, G is induced M-free, and every induced copy of C_8 in G is of the form $x_1x_2 \ldots x_8x_1$ with $x_i \in B(V_i)$. Let $x_1x_2 \ldots x_8x_1$ be an induced copy of C_8 in G and let $v_i \in V_i$ be such that $x_i \in B(v_i)$. From the construction of G it follows that $v_1v_2 \ldots v_8v_1$ is a (not necessarily induced) cycle in R. By Item 3 in Lemma 2.7.2 (with parameters $t = 8$ and $i_j = j$ for $1 \leq j \leq 8$), the number of such cycles is at most r^2. We conclude that G contains at most $r^2 (n/r)^8 \leq n^8/r$ induced copies of C_8. By Item 1 in Lemma 2.7.2 we have $r \geq (\frac{1}{64\varepsilon})^{\beta \log(1/64\varepsilon)} \geq (\frac{1}{\varepsilon})^{\Omega(\log 1/\varepsilon)}$ (where $\beta = \beta(8)$ is from Lemma 2.7.2). Therefore, the number of induced copies of C_8 in G is at most $\varepsilon^{\Omega(\log 1/\varepsilon)}n^8$, as required.

To finish the proof, we show that G contains εn^2 pair-disjoint induced copies of C_8, which will imply that G is ε-far from being induced $\{C_8, M\}$-free. By Claim 2.3.4 and the construction of G, for every clique $\{v_1, \ldots, v_8\} \in \mathcal{H}$ there is a collection $\mathcal{S}_{v_1,\ldots,v_8}$ of at least $(n/8r)^2$ pair-disjoint induced copies of C_8 of the form $(x_1, \ldots, x_8) \in B(v_i) \times \cdots \times B(v_8)$. (Here we apply Claim 2.3.4 with parameters $t = n/r$ and $q = 8$.) Since the cliques in $\mathcal{H}$ are pair-disjoint, copies of C_8 that come from different cliques are pair-disjoint. In other words, for every pair of distinct $\{v_1^{(1)}, \ldots, v_8^{(1)}\}, \{v_1^{(2)}, \ldots, v_8^{(2)}\} \in \mathcal{H}$ and for every $(x_1^{(i)}, \ldots, x_8^{(i)}) \in \mathcal{S}_{v_1^{(i)},\ldots,v_8^{(i)}}$ (for $i = 1, 2$), it holds that $|\{x_1^{(1)}, \ldots, x_8^{(1)}\} \cap \{x_1^{(2)}, \ldots, x_8^{(2)}\}| \leq 1$. We thus conclude that $\mathcal{S} := \bigcup_{\{v_1,\ldots,v_8\}\in\mathcal{H}} \mathcal{S}_{v_1,\ldots,v_8}$ is a collection of at least $|\mathcal{H}| \cdot (n/8r)^2 \geq 64\varepsilon r^2 (n/8r)^2 = \varepsilon n^2$ pair-disjoint induced copies of C_8 in G (where in the first inequality we used the fact that $|\mathcal{H}| \geq 64\varepsilon r^2$). This completes the proof of the theorem. ■

Proof of Lemma 2.9.2. We start by proving Item 1. Let $C = x_1x_2...x_8x_1$ be an induced copy of C_8 in G. Our goal is to show that $|C \cap X_i| = 1$ for every $1 \leq i \leq 8$. First, assume by contradiction, that $|C \cap X_i| \geq 2$ for some $i \in \{1, 3, 5, 7\}$, say $i = 1$ (without loss of generality). Since X_1 is a clique, there must be some $j \in \{1, ..., 8\}$ for which $x_j, x_{j+1} \in X_1$ (with indices taken modulo 8). We assume, without loss of generality, that $j = 1$, i.e. that $x_1, x_2 \in X_1$. Note that $|C \cap X_1| < 3$, as otherwise C would contain a triangle. So we see that $C \cap X_1 = \{x_1, x_2\}$. As $(x_2, x_3), (x_1, x_8) \in E(G)$ but $x_3, x_8 \notin X_1$, we must have $x_3, x_8 \in X_2 \cup X_8$. First we consider the case that x_3 and x_8 are in the same part, say $x_3, x_8 \in X_2$. Then $x_4, x_7 \in X_3$ because $(x_4, x_3), (x_7, x_8) \in E(G)$, X_2 is an independent set and $x_4, x_7 \notin X_1$. Since X_3 is a clique, we get that $(x_4, x_7) \in E(G)$, in contradiction to the fact that C is an induced cycle. Now we consider the case that x_3 and x_8 are in different parts, say $x_3 \in X_2$, $x_8 \in X_8$. The path $P = x_3x_4...x_8$ cannot go through X_1, and hence it must contain at least one vertex from each of the seven parts $X_2, ..., X_8$. But this is impossible as

[13] In fact, G will be induced F_2-free.

P consists of 6 vertices.

In the previous paragraph we showed that $|C \cap X_i| \leq 1$ for every $i \in \{1,3,5,7\}$. Define the sets $X_{odd} := X_1 \cup X_3 \cup X_5 \cup X_7$ and $X_{even} := X_2 \cup X_4 \cup X_6 \cup X_8$. Since X_{even} is an independent set and $\alpha(C_8) = 4$, we have $|C \cap X_{even}| \leq 4$. Thus $|C \cap X_{odd}| \geq 4$, implying that $|C \cap X_i| = 1$ for every $i \in \{1,3,5,7\}$. In order to finish the proof (of Item 1) it is enough to show that $|C \cap X_i| \geq 1$ for each $i \in \{2,4,6,8\}$. Suppose, by contradiction, that $C \cap X_i = \emptyset$ for some $i \in \{2,4,6,8\}$, say $i = 2$. Let $j, k \in \{1, \ldots, 8\}$ be such that $x_j \in X_1$ and $x_k \in X_3$. In the cycle C there is a path between x_j and x_k with at most 5 vertices (including x_j and x_k). This path cannot intersect X_2, so it must contain at least one vertex from each of the seven parts $X_1, X_3, X_4, ..., X_8$, which is impossible.

We now prove Item 2. Suppose by contradiction that $Y \subseteq V(G)$ spans an induced copy of M. As before, define $X_{odd} = X_1 \cup X_3 \cup X_5 \cup X_7$ and $X_{even} = X_2 \cup X_4 \cup X_6 \cup X_8$, and notice that X_{even} is an independent set and that X_{odd} is a disjoint union of cliques and hence induced P_3-free (where P_3 is the path with 3 vertices). It is easy to check that every set of 5 vertices of M contains an induced copy of P_3. We conclude that $|Y \cap X_{odd}| \leq 4$. Moreover, $|Y \cap X_{even}| \leq 2$ because $\alpha(M) = 2$. All in all we get that $|Y| \leq 6 < 7 = |V(M)|$, a contradiction. ■

2.10 Proof of the "Only-If" Part of Theorem 8

Let F be a non-2-colorable oriented graph on the vertex-set $[f]$. We will show that there is $c = c(f) > 0$ such that for every $0 < \varepsilon < c$ and $n \geq n_0(\varepsilon)$, there is an n-vertex tournament T which is ε-far from being F-free yet contains at most $\varepsilon^{c \log(1/\varepsilon)} \cdot n^f$ copies of F. This will imply that F is hard.

Let $K = K(F)$ be the graph defined in Section 2.8.2. Put $k = |V(K)|$ and write $V(K) = \{a_1, ..., a_k\}$, where $1 \leq a_1 < a_2 < \cdots < a_k \leq f$ (recall that K inherits its vertex-labeling from the backedge graph of F whose ordered core is K). By Corollary 2.8.6, K contains a cycle $(a_{j_1} a_{j_2} ... a_{j_\ell} a_{j_1})$ of length $\ell \geq 3$. Let $m_0 = m_0(f)$ and $\gamma = \gamma(f)$ be from Lemma 2.3.1. We choose $c = c(f)$ to be small enough so that every $0 < \varepsilon < c$ satisfies the inequalities

$$\varepsilon < \gamma \cdot \delta_0(k) \quad \text{and} \quad (\varepsilon/\gamma)^{\beta(k) \cdot \log(\gamma/\varepsilon)} \leq \varepsilon^{c \log(1/\varepsilon)}, \tag{2.7}$$

where $\delta_0(k)$ and $\beta(k)$ are from Lemma 2.7.2. Let $0 < \varepsilon < c$, let $R = R(k, \delta)$ be the graph obtained by applying Lemma 2.7.2 with parameters k and $\delta := \varepsilon/\gamma$, and let $V(R) = V_1 \uplus \cdots \uplus V_k$ be a partition of $V(R)$ as in that lemma. Item 3 of Lemma 2.7.2 guarantees that for every increasing sequence of indices $1 \leq i_1 < \cdots < i_\ell \leq k$, R contains at most $|V(R)|^2$ (not necessarily induced) cycles of the form $v_{i_1} v_{i_2} \ldots v_{i_\ell} v_{i_1}$ with $v_{i_1} \in V_{i_1}, \ldots, v_{i_\ell} \in V_{i_\ell}$. By permuting the names of the sets $V_1, \ldots, V_k$ (if necessary), we may assume that R contains at most $|V(R)|^2$ (not necessarily induced) cycles of the form $v_{j_1} v_{j_2} \ldots v_{j_\ell} v_{j_1}$ with $v_{j_1} \in V_{j_1}, \ldots, v_{j_\ell} \in V_{j_\ell}$, where $j_1, \ldots, j_\ell$ are the indices of the cycle $(a_{j_1} a_{j_2} ... a_{j_\ell} a_{j_1})$ in K, as above. (Note that the sequence $j_1, \ldots, j_\ell$ may not be increasing, but by permuting the names of the sets $V_{j_1}, \ldots, V_{j_\ell}$, we may assume that R has at most $|V(R)|^2$ cycles corresponding to this particular sequence.)

By the definition of K, there is a vertex-labeling of F such that K is the ordered core of the corresponding backedge graph, G_0. Let $g : G_0 \to K$ be an order-preserving homomorphism. Denote $X_i = g^{-1}(a_i)$ for $i = 1, ..., k$. We claim that $X_1, \ldots, X_k$ have the following two properties in the oriented graph F.

(a) $F[X_i]$ is an acyclic digraph for every $i = 1, ..., k$;

(b) For every pair $1 \leq i < j \leq k$, if $\{a_i, a_j\} \notin E(K)$ then $E(X_j, X_i) = \emptyset$.

Item (a) follows from the definition of a backedge graph and the fact that g is a graph homomorphism. For Item (b) we also need to use the fact that g is order-preserving.

Define an oriented graph D on $[k]$ as follows. For every $1 \leq i < j \leq k$, if $\{a_i, a_j\} \notin E(K)$ then $(i, j) \in D$ (that is, there is a directed edge from i to j) and otherwise $(i, j), (j, i) \notin E(D)$. Note that for every $(i, j) \in E(D)$ we have $E(X_j, X_i) = \emptyset$. It follows that F satisfies the conditions of Lemma 2.3.1 with respect to the k-coloring $V(F) = X_1 \cup \cdots \cup X_k$ and the oriented graph D.

Fix a large enough integer n and assume, for simplicity of presentation, that n is divisible by $|V(R)|$. Apply Lemma 2.3.1 with parameter $m := n/|V(R)|$ to obtain a k-partite tournament H with sides $U_1, \ldots, U_k$ which satisfies the properties stated in that lemma. In particular, $|U_i| = n/|V(R)|$ for every $i = 1, \ldots, k$, and the following holds:

$$\text{For every pair } 1 \leq i < j \leq k, \text{ if } \{a_i, a_j\} \notin E(K) \text{ then } U_i \to U_j. \tag{2.8}$$

To make our application of Lemma 2.3.1 valid, we assume that n is large enough so that $n/|V(R)| \geq m_0(f)$.

By Item 2 of Lemma 2.7.2, R contains a collection $\mathcal{K}$ of at least $\delta|V(R)|^2$ pairwise edge-disjoint k-cliques of the form $\{v_1, \ldots, v_k\}$ with $v_i \in V_i$, $i = 1, \ldots, k$. Moreover, every edge of R belongs to (exactly) one of these cliques.

We define a tournament T on an "$\frac{n}{|V(R)|}$-blowup" of $V(R)$; that is, each vertex $v \in V(R)$ is replaced by a vertex-set $B(v)$ of size $n/|V(R)|$, and the union of all resulting sets $B(v)$, $v \in V(R)$, forms the vertex-set of T. For each $1 \leq i \leq k$, put $B(V_i) = \bigcup_{v \in V_i} B(v)$. The edges of T are oriented as follows.

(i) $B(V_i)$ spans a transitive tournament for every $i = 1, ..., k$.

(ii) For every $1 \leq i < j \leq k$ and for every $v_i \in V_i, v_j \in V_j$, if $\{v_i, v_j\} \notin E(R)$ then set $B(v_i) \to B(v_j)$ (i.e., orient all edges from $B(v_i)$ to $B(v_j)$).

(iii) For every $\{v_1, ..., v_k\} \in \mathcal{K}$, put a copy of H on $B(v_1) \cup \cdots \cup B(v_k)$ in which $B(v_i)$ plays the role of U_i for every $i = 1, ..., k$.

Since every edge of R is contained in (exactly) one of the cliques in $\mathcal{K}$, Items (ii)-(iii) together specify the orientations of all edges which go between different parts among $B(V_1), \ldots, B(V_k)$. Therefore, Items (i)-(iii) indeed define a tournament. There is no contradiction in Item (iii) because the cliques in $\mathcal{K}$ are pairwise edge-disjoint.

We now show that T satisfies our requirements, that is, T is ε-far from being F-free yet contains at most $\varepsilon^{c \log(1/\varepsilon)} n^f$ copies of F. We start with the following two observations, that play a central role in the proof. First, notice that by Item (ii) and by the combination of Item (iii) and (2.8), we have the following:

$$\text{For every pair } 1 \leq i < j \leq k, \text{ if } \{a_i, a_j\} \notin E(K) \text{ then } B(V_i) \to B(V_j). \tag{2.9}$$

Second, let $\mathcal{R}$ be the set of all ℓ-tuples $(x_{j_1}, \ldots, x_{j_\ell}) \in B(V_{j_1}) \times \cdots \times B(V_{j_\ell})$ such that for every $i = 1, \ldots, \ell$, if $j_i < j_{i+1}$ then $x_{j_{i+1}} \to x_{j_i}$ and if $j_i > j_{i+1}$ then $x_{j_i} \to x_{j_{i+1}}$, with indices taken modulo ℓ. We claim that

$$|\mathcal{R}| \leq n^\ell / |V(R)|. \tag{2.10}$$

To prove (2.10), fix any given $(x_{j_1},\ldots,x_{j_\ell})\in\mathcal{R}$, and let $v_{j_i}\in V_{j_i}$ be such that $x_{j_i}\in B(v_{j_i})$ (for $i=1,\ldots,\ell$). We claim that $v_{j_1},v_{j_2},\ldots,v_{j_\ell},v_{j_1}$ is a cycle in R. To this end, fix any $1\leq i\leq\ell$, and let us first handle the case that $j_i<j_{i+1}$. By the definition of $\mathcal{R}$ we have $x_{j_{i+1}}\to x_{j_i}$. If we had $\{v_{j_i},v_{j_{i+1}}\}\notin E(R)$ then by Item (ii) above we would have $B(v_{j_i})\to B(v_{j_{i+1}})$, which would contradict $x_{j_{i+1}}\to x_{j_i}$. Therefore $\{v_{j_i},v_{j_{i+1}}\}\in E(R)$ in this case. Similarly, if $j_i>j_{i+1}$ then by the definition of $\mathcal{R}$ we have $x_{j_i}\to x_{j_{i+1}}$. Now, if we had $\{v_{j_i},v_{j_{i+1}}\}\notin E(R)$, then by Item (ii) above we would have $B(v_{j_{i+1}})\to B(v_{j_i})$, in contradiction to $x_{j_i}\to x_{j_{i+1}}$. Therefore $\{v_{j_i},v_{j_{i+1}}\}\in E(R)$ in this case as well, proving our assertion that $v_{j_1},v_{j_2},\ldots,v_{j_\ell},v_{j_1}$ is a cycle in R. By our choice of R, the number of cycles of this form in R is at most $|V(R)|^2$. Now, recalling that T is an $\frac{n}{|V(R)|}$-blowup of R, we see that $|\mathcal{R}|\leq|V(R)|^2\cdot(n/|V(R)|)^\ell\leq n^\ell/|V(R)|$, establishing (2.10).

Next, we claim that every copy of F in T contains vertices $x_{j_1},\ldots,x_{j_\ell}$ with $(x_{j_1},\ldots,x_{j_\ell})\in\mathcal{R}$. To this end, consider an embedding $\varphi:F\to T$; that is, φ is such that $\mathrm{Im}\varphi$ spans a copy of F in T with $\varphi(a)$ playing the role of a for every $a\in V(F)$. For $i=1,...,k$, define $A_i=\varphi^{-1}(B(V_i))$. Then $F[A_i]$ is an acyclic digraph by Item (i) above. Consider a vertex-labeling of F with labels $1,...,f$ in which (a) for every $1\leq i<j\leq k$, the labels given to the vertices of A_i are smaller than the labels given to the vertices of A_j; and (b) for every $i=1,...,k$, the vertices in A_i are labeled in such a way that all edges are forward-edges, that is, for every $a,a'\in A_i$ we have $a\to a'$ only if $a<a'$ (such a vertex-labeling of A_i exists since $F[A_i]$ is acyclic). Let G be the backedge graph of F with respect to this vertex-labeling. Notice that if $\{a,a'\}\in E(G)$ and $a<a'$ then there are $1\leq i<j\leq k$ such that $a\in A_i$ and $a'\in A_j$, as $A_1,\ldots,A_k$ are independent sets in G.

We claim that the function $h:V(F)\to V(K)$ which maps A_i to a_i (for $i=1,\ldots,k$) is an order-preserving homomorphism from G to K. The fact that h is order-preserving is immediate from the definition of the labeling. To see that h is a graph homomorphism, consider any edge $\{a,a'\}\in E\left(G\right)$ and assume without loss of generality that $a<a'$. By the definition of a backedge graph, we have $a'\to a$ in F, implying that $\varphi(a')\to\varphi(a)$ in T. As mentioned before, there are $1\leq i<j\leq k$ such that $a\in A_i$ and $a'\in A_j$. Assume, for the sake of contradiction, that we have $\{h(a),h(a')\}=\{a_i,a_j\}\notin E(K)$. By (2.9), this implies that $B(V_i)\to B(V_j)$. But as $\varphi(a)\in B(V_i)$ and $\varphi(a')\in B(V_j)$, we get a contradiction to $\varphi(a')\to\varphi(a)$. Therefore $\{h(a),h(a')\}=\{a_i,a_j\}\in E(K)$, showing that h is indeed a homomorphism.

Having shown that h is an order-preserving homomorphism, we use Corollary 2.8.6 to infer that there are $u_{j_i}\in h^{-1}(a_{j_i})=A_{j_i}$, $1\leq i\leq\ell$, such that $u_{j_1},u_{j_2},\ldots,u_{j_\ell},u_{j_1}$ is a cycle in G. For each $i=1,\ldots,\ell$, set $x_{j_i}=\varphi(u_{j_i})$ and observe that by the definition of the sets $A_1,\ldots,A_k$, we have $x_{j_i}\in B(V_{j_i})$. We now show that $(x_{j_1},\ldots,x_{j_\ell})\in\mathcal{R}$. Note that $\{u_{j_i},u_{j_{i+1}}\}\in E\left(G\right)$ for every $1\leq i\leq\ell$ (with indices taken modulo ℓ). Fix any $1\leq i\leq\ell$, and assume first that $j_i<j_{i+1}$. Then $u_{j_{i+1}}\to u_{j_i}$ in F by the definition of a backedge graph. Therefore, $x_{j_{i+1}}=\varphi(u_{j_{i+1}})\to\varphi(u_{j_i})=x_{j_i}$, as φ is an embedding. Similarly, if $j_i>j_{i+1}$ then $u_{j_i}\to u_{j_{i+1}}$ in F by the definition of a backedge graph, implying that $x_{j_i}=\varphi(u_{j_i})\to\varphi(u_{j_{i+1}})=x_{j_{i+1}}$. This shows that $(x_{j_1},\ldots,x_{j_\ell})\in\mathcal{R}$, as required.

We are now in position to prove that T contains at most $\varepsilon^{c\log(1/\varepsilon)}n^f$ copies of F. Indeed, above we have shown that every copy of F in T contains vertices $x_{j_1},\ldots,x_{j_\ell}$ with $(x_{j_1},\ldots,x_{j_\ell})\in\mathcal{R}$. It follows that the number of copies of F in T is at most

$$|\mathcal{R}|\cdot n^{f-\ell}\leq n^f/|V(R)|\leq\delta^{\beta(k)\cdot\log(1/\delta)}\cdot n^f=(\varepsilon/\gamma)^{\beta(k)\cdot\log(\gamma/\varepsilon)}\cdot n^f\leq\varepsilon^{c\log(1/\varepsilon)}\cdot n^f,$$

where in the first inequality we used (2.10), in the second we used Item 1 of Lemma 2.7.2, and in the third we used (2.7).

It remains to show that T is ε-far from being F-free. To this end, let us say that an edge $e\in E(T)$ is

a *cluster-edge* if it is contained in $B(V_i)$ for some $i = 1, ..., k$, and a *cut-edge* otherwise; that is, a cut-edge connects vertices from two different clusters among $B(V_1), \ldots, B(V_k)$. Let T' be any tournament obtained from T by reversing less than εn^2 edges. Our goal is to show that T' contains a copy of F. Let T'' be the tournament that agrees with T on all cut-edges and agrees with T' on all cluster-edges. Then T'' and T' disagree on less than εn^2 edges, and the same is true for T'' and T.

Fix any $K = \{v_1, ..., v_k\} \in \mathcal{K}$, and note that the tournament $T''[B(v_1) \cup \cdots \cup B(v_k)]$ is a completion of the k-partite tournament H (this follows from Item (iii) above and the fact that T'' agrees with T on cut-edges). By our choice of H via Lemma 2.3.1, $T''[B(v_1) \cup \cdots \cup B(v_k)]$ contains a collection $\mathcal{C}(K)$ of at least $\gamma \cdot (n/|V(R)|)^2$ copies of F, any two of which do not share cut-edges.

Now let $K = \{v_1, \ldots, v_k\}$ and $K' = \{v'_1, \ldots, v'_k\}$ be distinct cliques in $\mathcal{K}$. Since K and K' are edge-disjoint, $T''[B(v_1) \cup \cdots \cup B(v_k)]$ and $T''[B(v'_1) \cup \cdots \cup B(v'_k)]$ do not share cut-edges. Therefore, copies of F belonging to $\mathcal{C}(K)$ do not share cut-edges with copies of F belonging to $\mathcal{C}(K')$. Setting $\mathcal{C} := \bigcup_{K \in \mathcal{K}} \mathcal{C}(K)$, we see that $\mathcal{C}$ is a collection of copies of F in T'', any two of which do not share cut-edges. As $|\mathcal{K}| \geq \delta |V(R)|^2$ and by our choice of δ, we have $|\mathcal{C}| \geq \delta |V(R)|^2 \cdot \gamma \left(n/|V(R)|\right)^2 = \varepsilon n^2$. Since no two copies of F in $\mathcal{C}$ share cut-edges, if one wishes to destroy all copies of F in T'' by only reversing cut-edges then one must reverse at least $|\mathcal{C}| \geq \varepsilon n^2$ cut-edges. Recall that T' and T'' agree on cluster-edges, and disagree on less than εn^2 edges. Therefore, one of the copies of F in T'' is also present in T'. This completes the proof.

2.11 Proof of Theorem 2.1.6

Let K be a graph with vertex set $[k]$. We say that a graph F is a *blowup* of K if F admits a vertex-partition $V(F) = X_1 \cup \cdots \cup X_k$ such that $X_1, \ldots, X_k$ are independent sets and for every $1 \leq i < j \leq k$, if $(i, j) \in E(K)$ then (X_i, X_j) is a complete bipartite graph and if $(i, j) \notin E(K)$ then (X_i, X_j) is an empty bipartite graph. We say that F is the *s-blowup* of K if $|X_1| = \cdots = |X_k| = s$.

Throughout this section, C_m denotes the cycle of length m. We will need the following simple proposition, whose proof appears at the end of this section.

Proposition 2.11.1. *Let k be an odd integer and let G be a blowup of C_k. Then G is induced C_6-free and (not necessarily induced) C_ℓ-free for every odd $3 \leq \ell < k$.*

Recall the definition of a graph homomorphism from Section 2.8.1. We will use the simple fact that $C_{2\ell+1}$ has a homomorphism into C_{2k+1} if and only if $\ell \geq k$ (this fact accounts for the second part of Proposition 2.11.1). The proof of Theorem 2.1.6 will make use of the following lemma from [11].

Lemma 2.11.2 ([11])**.** *Let K be a graph on k vertices, let F be a graph on f vertices which has a homomorphism into K and let G be the $\frac{n}{k}$-blowup of K where $n \geq n_0(f)$. Then G is $\frac{1}{2k^2}$-far from being (not necessarily induced) F-free.*

For a graph F, denote by $SG(F)$ the set of supergraphs of F (namely, the set of all graphs on $V(F)$ obtained from F by adding edges). Note that being (not necessarily induced) F-free is equivalent to being induced $SG(F)$-free. We are now ready to prove Theorem 2.1.6.

Proof of Theorem 2.1.6. Define a sequence $\{a_i\}_{i \geq 1}$ as follows: set $a_1 = 3$ and $a_{i+1} = 2^{2(a_i+2)^2} + 1$. Note

that a_i is odd for every $i \geq 1$. We prove the theorem with the graph family

$$\mathcal{F} = \{C_6\} \cup \bigcup_{i \geq 1} SG(C_{a_i}).$$

Since $a_1 = 3$ we have $C_3 \in \mathcal{F}$. Note that C_6 is a bipartite graph and that C_3 is both a co-bipartite graph and a split graph. For $i \geq 1$ put $\varepsilon_i = \frac{1}{2(a_i+2)^2}$. We will show that $w_{\mathcal{P}^*_{\mathcal{F}}}(\varepsilon_i) \geq 2^{1/\varepsilon_i}$ for every $i \geq 1$ (recall Definition 2, which would imply that $\mathcal{P}^*_{\mathcal{F}}$ is not easily testable.

Let $i \geq 1$ and put $k = a_i + 2$ and $f = a_{i+1}$. Since a_i is odd and $a_i \geq 3$, we have that k is odd and $k \geq 5$. Fix $n \geq n_0(f)$ which is divisible by k (where $n_0(f)$ is from Lemma 2.11.2), and let G be the $\frac{n}{k}$-blowup of C_k. By our choice of ε_i and k we have $\varepsilon_i = \frac{1}{2k^2}$. Since C_f has a homomorphism into C_k, Lemma 2.11.2 implies that G is ε_i-far from being C_f-free and hence is ε_i-far from being induced $SG(C_f)$-free. As $SG(C_f) \subseteq \mathcal{F}$, we conclude that G is ε_i-far from being induced $\mathcal{F}$-free.

Proposition 2.11.1 implies that G is induced C_6-free and that for every odd $3 \leq \ell < k$, G is C_ℓ-free and hence induced $SG(C_\ell)$-free. By the definition of $\mathcal{F}$, if $F \in \mathcal{F}$ is an induced subgraph of G then $|V(F)| \geq a_{i+1} > 2^{2(a_i+2)^2} = 2^{1/\varepsilon_i}$. Here we used the definition of the sequence $\{a_i\}_{i \geq 1}$ and our choice of ε_i. We conclude that every set $Q \subseteq V(G)$ of size less than $2^{1/\varepsilon_i}$ is induced $\mathcal{F}$-free, implying that $w_{\mathcal{P}^*_{\mathcal{F}}}(\varepsilon_i) \geq 2^{1/\varepsilon_i}$, as required. ■

We remark that using essentially the same proof as above, we could have proven the following strengthening of Theorem 2.1.6. For every function $g : (0,1) \to \mathbb{N}$ there is a graph family $\mathcal{F}$ that contains a bipartite graph, a co-bipartite graph and a split graph, and there is a decreasing sequence $\{\varepsilon_i\}_{i \geq 1}$ with $\varepsilon_i \to 0$, such that $w_{\mathcal{P}^*_{\mathcal{F}}}(\varepsilon_i) > g(\varepsilon_i)$ for every $i \geq 1$.

Proof of Proposition 2.11.1. As G is a blow-up of C_k, it has a partition $V(G) = X_1 \cup \cdots \cup X_k$ into independent sets such that (X_i, X_j) is a complete bipartite graph if $|i - j| \equiv \pm 1 \pmod{k}$ and an empty bipartite graph otherwise. For the first part of the proposition, assume, by contradiction, that there is $Z \subseteq V(G)$ such that $G[Z]$ is isomorphic to C_6. Since C_6 is not a subgraph of C_k, there must be $1 \leq i \leq k$ such that $|Z \cap X_i| \geq 2$. Assume without loss of generality that there are distinct $u, v \in Z \cap X_1$. By the structure of C_6, there are distinct $x, y \in Z$ such that $(u,x), (u,y) \in E(G)$. Then $x, y \in X_2 \cup X_k$, implying that $(v,x), (v,y) \in E(G)$. Thus, $uxvy$ is a 4-cycle, in contradiction to the fact that $G[Z]$ is isomorphic to C_6. For the second part of the proposition, simply observe that every subgraph of G with less than k vertices is bipartite. ■

2.12 The Hardness of Deciding Tournament Colorability

In this section we prove Theorem 9. The main challenge in proving this theorem is the case $k = 2$.

Theorem 2.12.1. *Deciding if a tournament is 2-colorable is NP-hard.*

After proving Theorem 2.12.1, we will derive Theorem 9 by using a simple reduction from the $(k-1)$-Colorability problem to the k-Colorability problem for every $k \geq 3$.

Theorem 2.12.1 is proved by showing a reduction from a known NP-hard problem, namely the *Triangle-Free Cut Problem*, which is defined as follows.

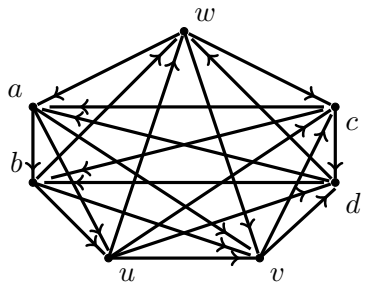

Figure 2.1: the gadget H

Definition 2.12.2 (**Triangle-Free Cut**). *For an (undirected) graph G, a* triangle-free cut *of G is a 2-coloring of $V(G)$ with no monochromatic triangle.*

It is known that the problem of deciding if a given graph has a triangle-free cut is NP-hard (see [83]).

For a vertex v in a digraph, we denote $N^+(v) = \{u : v \to u\}$ and $N^-(v) = \{u : u \to v\}$. If a pair of vertices u, v satisfy $u \to v$ then we say that u *dominates* v and that v is *dominated* by u. For the proof of Theorem 2.12.1 we need the following proposition regarding the gadget H depicted in Figure 2.1.

Proposition 2.12.3. *H has the following properties.*

1. *H has a proper 2-coloring in which u and v have the same color and all the vertices in the set $N^-(u) \cup N^+(v)$ have the other color.*

2. *In every proper 2-coloring of H, the colors of u and v are the same.*

Proof. For Item 1, color u, v, w with one color and a, b, c, d with the other color. We now prove Item 2. Consider a 2-coloring of $V(H)$ in which u and v have different colors, say u is colored red and v is colored blue. If there is a color, red or blue, that appears in both $\{a, b\}$ and $\{c, d\}$, then the coloring is not proper, as we get a monochromatic cyclic triangle by joining either u or v. Therefore, we may assume that either a, b are colored with red and c, d are colored with blue, or vice versa. But in both cases there is no color for w as $\{a, b, w\}$ and $\{c, d, w\}$ are cyclic triangles. ■

Proof of Theorem 2.12.1. Given a graph G with vertices $V(G) = \{x_1, ..., x_n\}$, we construct a tournament $T = T(G)$ and prove that G has a triangle-free cut if and only if T is 2-colorable. T is defined as follows. First, we put in T vertices $y_1, ..., y_n$ and set $y_i \to y_j$ for every $i < j$. We think of y_i as corresponding to the vertex x_i of G. Denote $Y = \{y_1, ..., y_n\}$. Let $C_1, ..., C_m$ be an enumeration of all triangles in G. Fix $1 \leq t \leq m$ and suppose that C_t contains the vertices $x_i, x_j, x_k \in V(G)$, where $i < j < k$. We add to T three new vertices, z_t^i, z_t^j, z_t^k, and set $z_t^i \to z_t^j \to z_t^k \to z_t^i$. So $Z_t := \left\{z_t^i, z_t^j, z_t^k\right\}$ spans a cyclic triangle. Set $Z_s \to Z_t$ for each $1 \leq s < t \leq m$. Denote $Z = \bigcup_{t=1}^m Z_t$ and set $Y \to Z$.

Let $1 \leq t \leq m$, suppose that $Z_t = \left\{z_t^i, z_t^j, z_t^k\right\}$, where $i < j < k$, and fix any $\ell \in \{i, j, k\}$. We add a copy of H (see Figure 2.1), denoted by H_t^ℓ, in which y_ℓ plays the role of u, z_t^ℓ plays the role of v and all other five vertices are new. Notice that this does not contradict $Y \to Z$, as we have $u \to v$ in H. Let K_t^ℓ

be the subtournament of H_t^ℓ spanned by the five "new" vertices, that is $V(K_t^\ell) = V(H_t^\ell) \setminus (Y \cup Z)$. Set $K_t^i \to K_t^j \to K_t^k$ and $K_t^i \to K_t^k$. Denote $K_t = K_t^i \cup K_t^j \cup K_t^k$ and for each $1 \leq s < t \leq m$ set $K_s \to K_t$.

Define $K = \bigcup_{t=1}^m K_t$ and note that we have $|Y| = n$, $|Z| = 3m$ and $|K| = 15m$. The vertex set of the tournament $T(G)$ is $Y \uplus Z \uplus K$. So far we defined the edges of $T(G)$ inside Y, Z and K and we set $Y \to Z$. We also already put some edges between Y and K and between K and Z, namely the edges which are contained in H_t^ℓ for some $1 \leq t \leq m$ and $1 \leq \ell \leq n$. We direct all other edges from Y to K and from K to Z; that is, if a pair $(p, q) \in Y \times K$ is not contained in any H_t^ℓ then we set $p \to q$, and similarly for K and Z. In what follows we use the fact that an edge going from K to Y or from Z to K is contained in H_t^ℓ for some $1 \leq t \leq m$ and $1 \leq \ell \leq n$. This completes the definition of the tournament $T = T(G)$.

It remains to show that G has a triangle-free cut if and only if T is 2-colorable. Assume first that T admits a proper 2-coloring, $c : V(T) \to \{\text{red, blue}\}$. For each $i = 1, ..., n$ set $\phi(x_i) = c(y_i)$. We claim that ϕ is a triangle-free cut of G, that is, for every $1 \leq t \leq m$, the triangle C_t in G is not monochromatic. Fix $1 \leq t \leq m$ and suppose that C_t contains the vertices x_i, x_j, x_k. By Item 2 in Proposition 2.12.3, it must be the case that $c(z_t^i) = c(y_i)$, $c(z_t^j) = c(y_j)$ and $c(z_t^k) = c(y_k)$. Since the set $Z_t = \{z_t^i, z_t^j, z_t^k\} \subseteq V(T)$ spans a cyclic triangle, we deduce that $c(y_i), c(y_j), c(y_k)$ are not all identical. Our choice of ϕ guarantees that C_t is not monochromatic.

Now assume that G admits a triangle-free cut, $\phi : V(G) \to \{\text{red, blue}\}$. We define a 2-coloring c of $V(T)$ as follows. First, set $c(y_i) = \phi(x_i)$ for every $i = 1, \ldots, n$. Next, let $1 \leq t \leq m$ and suppose that $Z_t = \left\{z_t^i, z_t^j, z_t^k\right\}$. For each $\ell \in \{i, j, k\}$ set $c(z_t^\ell) = c(y_\ell)$. Recall that H_t^ℓ is a copy of H in which y_ℓ plays the role of u and z_t^ℓ plays the role of v. Extend the coloring of $\left\{y_\ell, z_t^\ell\right\}$ to a coloring of H_t^ℓ as in Item 1 of Proposition 2.12.3, that is, H_t^ℓ is colored properly and any vertex that dominates y_ℓ or that is dominated by z_t^ℓ has a different color from that of y_ℓ, z_t^ℓ. This guarantees that H_t^ℓ does not contain monochromatic edges going from K to Y or from Z to K. As mentioned before, any edge in T going from K to Y or from Z to K is contained in H_t^ℓ for some $1 \leq t \leq m$ and $1 \leq \ell \leq n$. We conclude that T does not contain monochromatic edges going from K to Y or from Z to K.

It remains to show that the 2-coloring c of $V(T) = Y \cup Z \cup K$, defined in the previous paragraph, is proper. Let S be a cyclic triangle in T. We show by case analysis that S is not monochromatic. First we consider the cases (a) $S \subseteq Y \cup K$ and S intersects both Y and K, (b) $S \subseteq K \cup Z$ and S intersects both K and Z, (c) S has one vertex in each of the sets Y, Z, K. Case (a) implies that S contains an edge going from K to Y. Similarly, case (b) implies that S contains an edge that goes from Z to K. Case (c) also implies that S contains an edge from Z to K because $Y \to Z$. As proven in the previous paragraph, T does not contain any monochromatic edge going from K to Y or from Z to K. Therefore, S is not monochromatic in each of the cases (a), (b) and (c).

Given the previous paragraph, the only remaining cases to consider are $S \subseteq Y \cup Z$ and $S \subseteq K$. First, notice that the only cyclic triangles which are contained in Z are $Z_1, \ldots, Z_m$. Let $1 \leq t \leq m$ and suppose that $Z_t = \left\{z_t^i, z_t^j, z_t^k\right\}$.

By the definition of the coloring c we have $c(z_t^\ell) = c(y_\ell) = \phi(x_\ell)$ for every $\ell \in \{i, j, k\}$. The vertices of the triangle C_t (in G) are x_i, x_j, x_k. Since ϕ is a triangle-free cut, it follows that $\phi(x_i), \phi(x_j), \phi(x_k)$ are not all identical. Therefore $c(z_t^i), c(z_t^j), c(z_t^k)$ are not all identical, namely Z_t is not monochromatic.

Recall that Y is transitive and we have $Y \to Z$. Therefore $Y \cup Z$ does not contain any monochromatic cyclic triangle. Finally, every cyclic triangle inside K is contained in some K_t^ℓ. These triangles are not monochromatic because each K_t^ℓ is colored properly. This finishes the case analysis, showing that T does

not contain a monochromatic cyclic triangle and completing the proof of the theorem. ■

Proof of Theorem 9. We will show that for every $k \geq 3$ there is a simple reduction from the $(k-1)$-Colorability problem to the k-Colorability problem. Given this reduction, we can prove the theorem by induction on k, with the base case $k = 2$ already settled by Theorem 2.12.1.

Given a tournament T, define a tournament T' as follows. The vertex-set of T' consists of two vertex-disjoint copies of T, denoted T_1 and T_2, and an additional vertex z. We set $T_1 \to T_2 \to z \to T_1$. We now show that T is $(k-1)$-colorable if and only if T' is k-colorable. First, if T is $(k-1)$-colorable then clearly T' is k-colorable: we color T_1 and T_2 according to a proper $(k-1)$-coloring of T, using the same $k-1$ colors for both T_1 and T_2, and then color z with the remaining k'th color. It is easy to see that this k-coloring of T' is proper. In the other direction, suppose that there is a proper coloring $c : V(T') \to [k]$ and assume without loss of generality that $c(z) = k$. Then it cannot be the case that both T_1 and T_2 contain a vertex with color k, as that will imply that there is a cyclic triangle in this color. Therefore, there is $i = 1, 2$ such that T_i is colored with $[k-1]$, implying that T is $(k-1)$-colorable. This completes the proof. ■

Chapter 3

The Induced-C_4 Removal Lemma

This chapter is devoted to proving Theorem 7. Our method of proving Theorem 7 gives in fact a more general result; it shows that for every (finite or infinite) graph family $\mathcal{F}$ which includes C_4 and satisfies an additional technical condition (see Item 2 of Lemma 3.3.1), the property $\mathcal{P}^*_{\mathcal{F}}$ of induced $\mathcal{F}$-freeness satisfies $w_{\mathcal{P}^*_{\mathcal{F}}}(\varepsilon) \leq 2^{\text{poly}(1/\varepsilon)}$. In particular, our method shows that $\mathcal{P}_{\text{chordal}} :=$ *chordality* (i.e., the property of being induced C_k-free for every $k \geq 4$) satisfies $w_{\mathcal{P}_{\text{chordal}}}(\varepsilon) \leq 2^{\text{poly}(1/\varepsilon)}$. This bound was subsequently improved to $w_{\mathcal{P}_{\text{chordal}}}(\varepsilon) = \text{poly}(1/\varepsilon)$ by de Joannis de Verclos [75].

In light of the discussion in the previous paragraph, it is natural to ask if one can obtain an exponential bound for (the removal lemma corresponding to) *any* graph-family $\mathcal{F}$ containing C_4. As the following theorem shows, this is not the case in a very strong sense.

Theorem 3.0.1. *For every (decreasing) function $g\colon (0,1/2) \to \mathbb{N}$ there is a family of graphs $\mathcal{F} = \mathcal{F}(g)$ so that $C_4 \in \mathcal{F}$ and yet $w_{\mathcal{P}^*_{\mathcal{F}}}(\varepsilon) \geq g(\varepsilon)$. In fact, for every (small enough) $\varepsilon > 0$ and every $n \geq n_0(\varepsilon)$, there is an n-vertex graph G which is ε-far from being induced $\mathcal{F}$-free, and yet does not contain an induced copy of* **any** *$F \in \mathcal{F}$ on fewer than $g(\varepsilon)$ vertices.*

3.1 Preliminary Lemmas

Our goal in this section is to introduce several definitions and prove Lemma 3.1.4 stated below, regarding graphs not containing induced matchings of size 2 of a specific type, which we now formally define. Let G be a graph and let $X, Y \subseteq V(G)$ be disjoint sets of vertices. An *induced copy of M_2* in (X,Y) is an (unordered) quadruple x, x', y, y' such that $x, x' \in X$, $y, y' \in Y$, $(x,y), (x',y') \in E(G)$ and $(x,y'), (x',y) \notin E(G)$. We say that (X,Y) is *induced M_2-free* if it does not contain induced copies of M_2. Observe that if X and Y are cliques then $G[X \cup Y]$ is induced C_4-free if and only if (X,Y) is induced M_2-free. For $x \in X$, we denote $N_Y(x) = \{y \in Y : (x,y) \in E(G)\}$.

Claim 3.1.1. *(X,Y) is induced M_2-free*[1] *if and only if there is an enumeration $x_1, \ldots, x_m$ of the elements of X such that $N_Y(x_i) \subseteq N_Y(x_j)$ for every $1 \leq i < j \leq m$.*

Proof. Observe that (X,Y) contains an induced M_2 if and only if there are $x, x' \in X$ for which there exist

[1] Let us mention that *half-graphs* are a special case of induced M_2-free bipartite graphs. A half-graph has $2n$ vertices $x_1, \ldots, x_n, y_1, \ldots, y_n$, and x_i is adjacent to y_j if and only if $i \geq j$.

$y \in N_Y(x) \setminus N_Y(x')$ and $y' \in N_Y(x') \setminus N_Y(x)$. Therefore, (X,Y) is induced M_2-free if and only if for every $x, x' \in X$ it holds that either $N_Y(x) \subseteq N_Y(x')$ or $N_Y(x') \subseteq N_Y(x)$. It is now easy to see that the assertion of the claim holds. For example, assuming that (X,Y) is induced M_2-free, consider the preorder on X in which x precedes x' if and only if $N_Y(x) \subseteq N_Y(x')$. This preorder defines a linear order. Enumerate the elements of X from minimal to maximal to get the required enumeration $x_1, \ldots, x_m$. ■

For a pair of disjoint vertex-sets X, Y, we say that (X,Y) is *homogeneous* if the bipartite graph between X and Y is either complete or empty. Throughout this chapter, and in particular in the following lemma, we will avoid floor/ceiling signs, by assuming that the number of vertices in the vertex-set under consideration is divisible by some small integers (ultimately these integers would depend only on the parameter ε). In what follows, when considering partitions of a set, we allow partition classes to be empty.

Lemma 3.1.2. *If (X,Y) is induced M_2-free then for every integer $r \geq 1$ there are partitions $X = X_1 \cup \cdots \cup X_r$ and $Y = Y_1 \cup \cdots \cup Y_{r+1}$ such that $|X_i| = |X|/r$ for every $1 \leq i \leq r$, and (X_i, Y_j) is homogeneous for every $1 \leq i \leq r$ and $1 \leq j \leq r+1$ satisfying $i \neq j$.*

Proof. Let $x_1, \ldots, x_m$ be the enumeration of the elements of X from Claim 3.1.1. For $1 \leq i \leq r$ define $X_i = \{x_j : (i-1)m/r < j \leq im/r\}$. Let now $y_1, \ldots, y_n$ be an enumeration of the elements of Y with the property that for every $x \in X$, the set $N_Y(x)$ is a "prefix" of the enumeration, that is, so that $N_Y(x) = \{y_1, \ldots, y_k\}$ for some $0 \leq k \leq n$. Define $Y_1 = N_Y(x_{m/r})$, $Y_i = N_Y(x_{im/r}) \setminus N_Y(x_{(i-1)m/r})$ for $i = 2, \ldots, r$ and $Y_{r+1} = Y \setminus N_Y(x_m)$.

It remains to show that (X_i, Y_j) is homogeneous for every $i \neq j$. Assume first that $i < j$. Then for every $x \in X_i$ we have $N_Y(x) \subseteq N_Y(x_{im/r}) \subseteq N_Y(x_{(j-1)m/r})$. By the definition of Y_j we have $Y_j \cap N_Y(x_{(j-1)m/r}) = \emptyset$. Thus, $Y_j \cap N_Y(x) = \emptyset$ for every $x \in X_i$, implying that the bipartite graph (X_i, Y_j) is empty. Now assume that $i > j$. For every $x \in X_i$ we have $N_Y(x_{jm/r}) \subseteq N_Y(x_{(i-1)m/r}) \subseteq N_Y(x)$. By the definition of Y_j we have $Y_j \subseteq N_Y(x_{jm/r})$. Thus, $Y_j \subseteq N_Y(x)$ for every $x \in X_i$, implying that the bipartite graph (X_i, Y_j) is complete. ■

For two partitions $\mathcal{P}_1, \mathcal{P}_2$ of the same set, we say that $\mathcal{P}_2$ is a *refinement* of $\mathcal{P}_1$ if every part of $\mathcal{P}_2$ is contained in one of the parts of $\mathcal{P}_1$. A vertex partition $\mathcal{P}$ of an n-vertex graph G is called *δ-homogeneous* if the sum of $|U||V|$ over all non-homogeneous unordered distinct pairs $U, V \in \mathcal{P}$ is at most δn^2. Note that if a δ-homogeneous partition $\mathcal{P}$ refines a partition $\{X_1, \ldots, X_k\}$ such that each X_i is either a clique or an independent set, then every refinement of $\mathcal{P}$ is also δ-homogeneous.

Lemma 3.1.3. *Let $k \geq 1$, let $\delta \in (0,1)$, let G be an n-vertex graph and let $V(G) = X_1 \cup \cdots \cup X_k$ be a partition such that $X_1, \ldots, X_k$ are cliques and (X_i, X_j) is induced M_2-free for every $1 \leq i < j \leq k$. Then there is a δ-homogeneous partition which refines $\{X_1, \ldots, X_k\}$ and has at most $k\,(3/\delta)^k$ parts.*

Proof. The assertion of the lemma is trivial for $k = 1$, so suppose that $k \geq 2$. For every $1 \leq i < j \leq k$, we apply Lemma 3.1.2 to (X_i, X_j) with parameter $r = \lceil 1/\delta \rceil$ to get partitions $\mathcal{P}_{i,j}$ of X_i and $\mathcal{P}_{j,i}$ of X_j, $\mathcal{P}_{i,j} = \{X^1_{i,j}, \ldots, X^r_{i,j}\}$, $\mathcal{P}_{j,i} = \{X^1_{j,i}, \ldots, X^{r+1}_{j,i}\}$, such that $|X^p_{i,j}| = |X_i|/r$ for every $1 \leq p \leq r$, and $(X^p_{i,j}, X^q_{j,i})$ is homogeneous for every $p \neq q$. Note that

$$\sum_{p=1}^{r} |X^p_{i,j}||X^p_{j,i}| = \sum_{p=1}^{r} \frac{1}{r}|X_i||X^p_{j,i}| \leq \frac{1}{r}|X_i||X_j| \leq \delta |X_i||X_j|. \tag{3.1}$$

For every $i = 1, \ldots, k$, define $\mathcal{P}_i$ to be the common refinement of the partitions $(\mathcal{P}_{i,j})_{1 \le j \le k,\, j \ne i}$. We have $|\mathcal{P}_i| \le (r+1)^{k-1} \le (\frac{1}{\delta} + 2)^{k-1} \le (3/\delta)^k$. The partition $\mathcal{P} := \bigcup_{i=1}^k \mathcal{P}_i$ refines $\{X_1, \ldots, X_k\}$ and has at most $k\,(3/\delta)^k$ parts. For every $U, V \in \mathcal{P}$, if (U, V) is not homogeneous, then there are $1 \le i \ne j \le k$ and $1 \le p \le r$ such that $U \subseteq X_{i,j}^p$ and $V \subseteq X_{j,i}^p$. This follows from the fact that $X_1, \ldots, X_k$ are cliques and the property of the partitions $(\mathcal{P}_{i,j})_{1 \le i \ne j \le k}$. By (3.1), we have

$$\sum_{1 \le i < j \le k} \sum_{p=1}^{r} |X_{i,j}^p||X_{j,i}^p| \le \delta \sum_{1 \le i < j \le k} |X_i||X_j| \le \delta n^2,$$

implying that $\mathcal{P}$ is δ-homogeneous, as required. ■

Lemma 3.1.4. *For every $k \ge 2$ and $\delta \in (0,1)$, there is $\rho = \rho_{3.1.4}(k, \delta) \ge (\delta^k/k)^{O(k)}$ such that the following holds. Let G be an n-vertex graph and let $V(G) = X_1 \cup \cdots \cup X_k$ be a partition such that $X_1, \ldots, X_k$ are cliques and (X_i, X_j) is induced M_2-free for every $1 \le i < j \le k$. Then there is a set $Z \subseteq V(G)$ of size $|Z| < \delta n$, a partition $V(G) \setminus Z = Q_1 \cup \cdots \cup Q_q$ which refines $\{X_1 \setminus Z, \ldots, X_k \setminus Z\}$ and subsets $W_i \subseteq Q_i$ such that the following hold.*

1. *The sum of $|Q_i||Q_j|$ over all non-homogeneous pairs (Q_i, Q_j), $1 \le i < j \le q$, is at most δn^2.*
2. *$|W_i| \ge \rho n$ for every $1 \le i \le q$, and (W_i, W_j) is homogeneous for every $1 \le i < j \le q$.*

Proof. The assertion of the lemma is trivial for $k = 1$, so suppose that $k \ge 2$. Apply Lemma 3.1.3 to G with parameter δ to obtain a δ-homogeneous partition $\mathcal{P}$ which refines $\{X_1, \ldots, X_k\}$ and has at most $k\,(3/\delta)^k$ parts. Let us define $\mathcal{Q} = \{U \in \mathcal{P} : |U| \ge \delta n/|\mathcal{P}|\}$ and write $\mathcal{Q} = \{Q_1, \ldots, Q_q\}$. Then Item 1 holds since $\mathcal{P}$ is δ-homogeneous. Setting $Z = \bigcup_{U \in \mathcal{P} \setminus \mathcal{Q}} U$, notice that $\mathcal{Q}$ refines $\{X_1 \setminus Z, \ldots, X_k \setminus Z\}$ and that $|Z| < |\mathcal{P}| \cdot \delta n/|\mathcal{P}| = \delta n$. Apply Lemma 3.1.3 again, this time with $G[V(G) \setminus Z]$ as the input graph, with the partition $\{X_1 \setminus Z, \ldots, X_k \setminus Z\}$ in place of $\{X_1, \ldots, X_k\}$, and with approximation parameter $\delta' := \delta^2/|\mathcal{P}|^4$. Lemma 3.1.3 gives a δ'-homogeneous partition $\mathcal{V}$ of $V(G) \setminus Z$ which refines $\{X_1 \setminus Z, \ldots, X_k \setminus Z\}$ and has at most $k(3|\mathcal{P}|^4/\delta^2)^k$ parts. Let $\mathcal{W}$ be the common refinement of $\mathcal{Q}$ and $\mathcal{V}$. Note that $\mathcal{W}$ is δ'-homogeneous as a refinement of $\mathcal{V}$, since $\mathcal{V}$ is δ'-homogeneous and refines $\{X_1 \setminus Z, \ldots, X_k \setminus Z\}$, and $X_1 \setminus Z, \ldots, X_k \setminus Z$ are cliques. Moreover, we have

$$|\mathcal{W}| \le |\mathcal{Q}| \cdot |\mathcal{V}| \le |\mathcal{P}| \cdot k(3|\mathcal{P}|^4/\delta^2)^k \le (k/\delta^k)^{O(k)}, \tag{3.2}$$

where in the last inequality we used the fact that $|\mathcal{P}| \le k\,(3/\delta)^k$.

For each $1 \le i \le q$, define $\mathcal{W}_i = \{W \in \mathcal{W} : W \subseteq Q_i\}$, choose a vertex $w_i \in Q_i$ uniformly at random and let $W_i \in \mathcal{W}_i$ be such that $w_i \in W_i$. We will show that with positive probability, the sets $W_1, \ldots, W_q$ satisfy the assertion of Item 2. For $1 \le i \le q$, the probability that $|W_i| < \frac{|Q_i|}{2q|\mathcal{W}|}$ is smaller than

$$\left(|\mathcal{W}| \cdot \frac{|Q_i|}{2q\,|\mathcal{W}|}\right)/|Q_i| = \frac{1}{2q},$$

as evidently there are at most $|\mathcal{W}_i| \le |\mathcal{W}|$ sets $W \in \mathcal{W}_i$ of size less than $\frac{|Q_i|}{2q|\mathcal{W}|}$. By the union bound, with probability larger than $\frac{1}{2}$, for every $1 \le i \le q$ we have

$$|W_i| \ge \frac{|Q_i|}{2q|\mathcal{W}|} \ge \frac{\delta n}{2|\mathcal{P}|^2|\mathcal{W}|} \ge n \cdot (\delta^k/k)^{O(k)},$$

where in the second inequality we used the bounds $|Q_i| \geq \delta n/|\mathcal{P}|$ and $q \leq |\mathcal{P}|$, and in the last inequality we used (3.2) and the fact that $|\mathcal{P}| \leq k\,(3/\delta)^k$.

For $1 \leq i < j \leq q$, the probability that the pair (W_i, W_j) is not homogeneous is

$$\sum \frac{|W||W'|}{|Q_i||Q_j|} \leq \frac{|\mathcal{P}|^2}{\delta^2 n^2} \sum |W||W'| \leq \frac{|\mathcal{P}|^2}{\delta^2 n^2} \cdot \delta' n^2 = \frac{1}{|\mathcal{P}|^2},$$

where the sums are taken over all non-homogeneous pairs $(W, W') \in \mathcal{W}_i \times \mathcal{W}_j$, the first inequality uses $|Q_i|, |Q_j| \geq \delta n/|\mathcal{P}|$ and the second the fact that $\mathcal{W}$ is δ'-homogeneous. By the union bound, with probability at least $1 - \binom{q}{2} \cdot \frac{1}{|\mathcal{P}|^2} \geq 1 - \binom{|\mathcal{P}|}{2} \cdot \frac{1}{|\mathcal{P}|^2} > \frac{1}{2}$, all pairs (W_i, W_j) are homogeneous. We conclude that Item 2 holds with positive probability. ■

It is worth mentioning that the bounds in the above lemma are the sole reason why our bound in Theorem 7 is exponential rather than polynomial.

3.2 A Partial Structure Theorem for Induced C_4-Free Graphs

Our main goal in this section is to prove Lemma 3.2.6 stated below, which gives an *approximate partial* structure theorem for induced C_4-free graphs. This structure result is approximate because the graph will only be *close* to having a certain nice structure, and it is partial because there will be a (possibly) large part of the graph about which we will have no control. This partialness is unavoidable, as evidenced by the fact that all split graphs are induced C_4-free (meaning that an induced C_4-free graph may contain a large bipartite graph, namely the bipartite graph between the clique and independent-set parts of a split graph, over which we have no control).

In addition to the lemmas from the previous section, we will also need the following theorems of Goldreich, Goldwasser and Ron [59] and of Gyárfás, Hubenko and Solymosi [69]. In both cases, $\omega(G)$ denotes the maximum size of a clique in G.

Theorem 3.2.1 ([59], Theorem 7.1)**.** *For every $\varepsilon \in (0,1)$ there is $q_{3.2.1}(\varepsilon) \leq (1/\varepsilon)^{O(1)}$ with the following property. Let $\rho \in (0,1)$ be such that $\varepsilon < \rho^2/2$ and let G be a graph which is ε-far from satisfying $\omega(G) \geq \rho n$. Suppose $q \geq q_{3.2.1}(\varepsilon)$ and let $Q \in \binom{V(G)}{q}$ be a randomly chosen set of q vertices of G. Then with probability at least $\frac{2}{3}$ we have $\omega(G[Q]) < (\rho - \frac{\varepsilon}{2})q$.*

Theorem 3.2.2 ([69])**.** *Every induced C_4-free graph G with n vertices and at least αn^2 edges satisfies $\omega(G) \geq 0.4\alpha^2 n$.*

Let us now derive the following important corollary of the above two theorems. For a set $X \subseteq V(G)$ with at least 2 vertices, define $d(X) = e(X)/\binom{|X|}{2}$, where $e(X)$ is the number of edges of G with both endpoints in X.

Lemma 3.2.3. *For every $\alpha, \beta \in (0,1)$, there is $\zeta = \zeta_{3.2.3}(\alpha, \beta) \geq (\alpha\beta)^{O(1)}$ such that the following holds. Let G be a graph on n vertices with at least αn^2 edges. Then either G contains at least ζn^4 induced copies of C_4 or there is a set $X \subseteq V(G)$ with $|X| \geq 0.1\alpha^2 n$ and $d(X) \geq 1 - \beta$.*

In the proof of Lemma 3.2.3 we need the following simple fact, which is proved by a standard application of the second moment method (see [16]). We first prove Claim 3.2.4 and then move on to prove Lemma 3.2.3.

Claim 3.2.4. *Let $\alpha \in (0,1/2)$ and let G be a graph with n vertices and at least αn^2 edges. Then for every $r \geq \frac{240}{\alpha}$, a randomly chosen set $R \in \binom{V(G)}{r}$ satisfies $e(R) \geq \frac{\alpha}{2} r^2$ with probability at least $\frac{2}{3}$.*

Proof. Let $\eta \in (0,1)$ be such that G has exactly ηn^2 edges, noting that $\eta \geq \alpha$. We consider the random variable $e(R)$. For each $e \in E(G)$, let I_e be the indicator of the event $e \subseteq R$. Then $e(R) = \sum_{e \in E(G)} I_e$. Note that $\mathbb{P}[e \subseteq R] = \frac{r(r-1)}{n(n-1)}$, so by linearity of expectation we have $\mathbb{E}[e(R)] = e(G) \cdot \frac{r(r-1)}{n(n-1)} \geq e(G) \cdot \frac{3}{4} \cdot \frac{r^2}{n^2} = \frac{3\eta}{4} r^2$. Now let us estimate the variance of $e(R)$. We have

$$\mathrm{Var}[e(R)] = \sum_{e \in E(G)} \mathbb{P}[e \subseteq R] + \sum_{e,e' \in E(G)} \left(\mathbb{P}[e,e' \subseteq R] - \mathbb{P}[e \subseteq R] \cdot \mathbb{P}[e' \subseteq R] \right),$$

where the second sum is over all ordered pairs e, e' of distinct edges. If e, e' are disjoint then $\mathbb{P}[e,e' \subseteq R] = \frac{r(r-1)(r-2)(r-3)}{n(n-1)(n-2)(n-3)} \leq \left(\frac{r}{n}\right)^4$, and $\mathbb{P}[e \subseteq R] \cdot \mathbb{P}[e' \subseteq R] = \left(\frac{r(r-1)}{n(n-1)}\right)^2 \geq \frac{r^2(r-1)^2}{n^4} \geq \frac{r^4 - 2r^3}{n^4}$. So the term corresponding to the pair e, e' in the above sum is at most $\frac{r^4}{n^4} - \frac{r^4-2r^3}{n^4} = \frac{2r^3}{n^4}$. Since there are at most $e(G)^2 = \eta^2 n^4$ pairs of edges e, e' altogether, the pairs in which e, e' are disjoint contribute at most $\eta^2 n^4 \cdot \frac{2r^3}{n^4} = 2\eta^2 r^3$ to the above sum.

If e, e' intersect (namely, have a vertex in common), then $\mathbb{P}[e,e' \subseteq R] = \frac{r(r-1)(r-2)}{n(n-1)(n-2)} \leq \left(\frac{r}{n}\right)^3$. Since there are at most $e(G) \cdot 2 \cdot n$ pairs of intersecting edges e, e', these pairs contribute at most $e(G) \cdot 2n \cdot \frac{r^3}{n^3} = 2\eta r^3$. Altogether, we have

$$\mathrm{Var}[e(R)] \leq \sum_{e \in E(G)} \mathbb{P}[e \subseteq R] + 2\eta^2 r^3 + 2\eta r^3 \leq \eta r^2 + 2\eta^2 r^3 + 2\eta r^3 \leq 5\eta r^3,$$

where in the second inequality we used the fact that $\sum_{e \in E(G)} \mathbb{P}[e \subseteq R] = \mathbb{E}[e(R)] \leq e(G) \cdot \left(\frac{r}{n}\right)^2 = \eta r^2$. By Chebyshev's inequality (see e.g. [16]), we have

$$\mathbb{P}\left[|e(R) - \mathbb{E}[e(R)]| > \eta r^2/4\right] \leq \frac{\mathrm{Var}[e(R)]}{\eta^2 r^4/16} \leq \frac{5\eta r^3}{\eta^2 r^4/16} = \frac{80}{\eta r} \leq \frac{80}{\alpha r} \leq \frac{1}{3},$$

where in the last inequality we used our choice of r. Finally, notice that if $|e(R) - \mathbb{E}[e(R)]| \leq \eta r^2/4$ then $e(R) \geq \mathbb{E}[e(R)] - \frac{\eta}{4} r^2 \geq \frac{3\eta}{4} r^2 - \frac{\eta}{4} r^2 = \frac{\eta}{2} r^2 \geq \frac{\alpha}{2} r^2$, as required. ■

Proof of Lemma 3.2.3. Set $\rho = 0.1\alpha^2$, $\varepsilon = \rho^2\beta/4 = \alpha^4\beta/400$ and $r = \max\{q_{3.2.1}(\varepsilon), \frac{240}{\alpha}\}$. By Theorem 3.2.1 we have $r \leq (\alpha^{-1}\beta^{-1})^{O(1)}$. We prove the lemma with $\zeta = \zeta_{3.2.3}(\alpha,\beta) := 1/(3r^4) \geq (\alpha\beta)^{O(1)}$.

Let us assume that there is no $X \subseteq V(G)$ with $|X| \geq 0.1\alpha^2 n$ and $d(X) \geq 1-\beta$, and prove that G contains at least ζn^4 induced copies of C_4. Let $X \subseteq V(G)$ be such that $|X| \geq \rho n$. Since $d(X) < 1-\beta$, we have $\binom{|X|}{2} - e(G[X]) > \beta\binom{|X|}{2} \geq \beta\frac{|X|^2}{4} \geq \frac{\rho^2\beta}{4} n^2 = \varepsilon n^2$. This shows that G is ε-far from containing a clique of size ρn or larger. By our choice of r via Theorem 3.2.1, a random sample R of r vertices of G satisfies $\omega(G[R]) < (\rho - \frac{\varepsilon}{2})r < 0.1\alpha^2 r$ with probability at least $\frac{2}{3}$. By Claim 3.2.4, we also have $e(R) > \frac{\alpha}{2} r^2$ with probability at least $\frac{2}{3}$. So with probability at least $\frac{1}{3}$, we have both $\omega(G[R]) < 0.1\alpha^2 r$ and $e(R) > \frac{\alpha}{2} r^2$. If both events happen, then $G[R]$ must contain an induced copy of C_4, by Theorem 3.2.2. We conclude that G contains at least $\frac{1}{3}\binom{n}{r}/\binom{n-4}{r-4} = \frac{1}{3}\binom{n}{4}/\binom{r}{4} \geq n^4/(3r^4) = \zeta n^4$ induced copies of C_4. ■

The last ingredient we need is the following special case of a result of Alon, Fischer and Newman [6]. For a pair of disjoint vertex sets X, Y, we say that (X,Y) is *ε-far* from being induced M_2-free if one has to add/delete at least $\varepsilon|X||Y|$ of the edges *between X and Y* in order to make (X,Y) induced M_2-free. Otherwise, we say that (X,Y) is *ε-close* to being induced M_2-free.

Lemma 3.2.5 ([6]). *For every $\varepsilon \in (0,1)$ there is $\eta = \eta_{3.2.5}(\varepsilon) \geq \varepsilon^{O(1)}$ such that the following holds. If (X,Y) is ε-far from being induced M_2-free then (X,Y) contains at least $\eta|X|^2|Y|^2$ induced copies of M_2.*

We note that an elementary proof of Lemma 3.2.5 was given in [75].

The following is the key lemma of this section. Note that it gives us a lot of information about $G[Y]$ and $G[X_1 \cup \cdots \cup X_k]$ but no information about the bipartite graph connecting $X_1 \cup \cdots \cup X_k$ and Y. This is unavoidable as every split graph is induced C_4-free.

Lemma 3.2.6. *For every $\alpha, \gamma \in (0,1)$ there is $\zeta = \zeta_{3.2.6}(\alpha,\gamma) \geq (\alpha\gamma)^{O(1)}$ such that every n-vertex graph G either contains at least ζn^4 induced copies of C_4, or admits a vertex partition $V(G) = X_1 \cup \cdots \cup X_k \cup Y$ with the following properties.*

1. *$e(Y) < \alpha n^2$.*
2. *$|X_i| \geq 0.1\alpha^3 n$ and $d(X_i) \geq 1-\gamma$ for every $1 \leq i \leq k$.*
3. *For every $1 \leq i < j \leq k$, the pair (X_i, X_j) is γ-close to being induced M_2-free.*

Proof. Set $\eta = \eta_{3.2.5}(\gamma)$ and $\beta = \min\{\gamma, \eta\}$. We prove the lemma with

$$\zeta = \zeta_{3.2.6}(\alpha,\gamma) := \min\left\{\zeta_{3.2.3}(\alpha,\beta)\cdot\alpha^4,\ 0.5\cdot 10^{-4}\alpha^{12}\eta\right\}.$$

The polynomial dependencies in Lemmas 3.2.3 and 3.2.5 imply that $\zeta \geq (\alpha\gamma)^{O(1)}$.

Define inductively two sequences of sets, $(V_i)_{i\geq 0}$ and $(X_i)_{i\geq 1}$, as follows. Set $V_0 = V(G)$. At the ith step (starting from $i=0$), if $e(V_i) < \alpha n^2$ then we stop. Note that if we did not stop then $|V_i| \geq \alpha^{1/2}n > \alpha n$. If $e(V_i) \geq \alpha n^2$ then by Lemma 3.2.3, applied to $G[V_i]$ with parameters α and β as above, either $G[V_i]$ contains at least $\zeta_{3.2.3}(\alpha,\beta)\cdot|V_i|^4 \geq \zeta_{3.2.3}(\alpha,\beta)\cdot\alpha^4 n^4 \geq \zeta n^4$ induced copies of C_4, or there is $X_{i+1} \subseteq V_i$ with $|X_{i+1}| \geq 0.1\alpha^2|V_i| \geq 0.1\alpha^3 n$ and $d(X_i) \geq 1-\beta$. In the former case the assertion of the lemma holds, so we may assume that the latter case happens, in which case we set $V_{i+1} = V_i \setminus X_{i+1}$ and continue. Suppose that this process stops at the kth step for some $k \geq 0$. Set $Y = V_k$. We clearly have $V(G) = X_1 \cup \cdots \cup X_k \cup Y$. For every $1 \leq i \leq k$ we have $|X_i| \geq 0.1\alpha^3 n$ and $d(X_i) \geq 1-\beta \geq 1-\gamma$. Since the process stopped at the kth step, we must have $e(Y) = e(V_k) < \alpha n^2$.

To finish the proof, we show that if Item 3 in the lemma does not hold then G contains at least $0.5\cdot 10^{-4}\alpha^{12}\eta n^4 \geq \zeta n^4$ induced copies of C_4. If Item 3 does not hold, then for some $1 \leq i < j \leq k$, the pair (X_i, X_j) is γ-far from being induced M_2-free. By our choice of η via Lemma 3.2.5, (X_i, X_j) contains at least $\eta|X_i|^2|X_j|^2$ induced copies of M_2. Let (x_i, x_i', x_j, x_j') be such a copy, where $x_i, x_i' \in X_i$ and $x_j, x_j' \in X_j$. If $\{x_i, x_i'\}, \{x_j, x_j'\} \in E(G)$ then x_i, x_i', x_j, x_j' span an induced copy of C_4. Since $d(X_i), d(X_j) \geq 1-\beta \geq 1-\eta$, there are at most $2\cdot\eta\binom{|X_i|}{2}\binom{|X_j|}{2} \leq 0.5\eta|X_i|^2|X_j|^2$ quadruples of distinct vertices $(x_i, x_i', x_j, x_j') \in X_i \times X_i \times X_j \times X_j$ for which either $\{x_i, x_i'\} \notin E(G)$ or $\{x_j, x_j'\} \notin E(G)$. Thus, G contains at least $0.5\eta|X_i|^2|X_j|^2 \geq 0.5\cdot 10^{-4}\alpha^{12}\eta n^4$ induced copies of C_4, as required. ■

We finish this section with the following corollary of the above structure theorem, which will be more convenient to use when proving Theorem 7 in the next section.

Lemma 3.2.7. *For every $\alpha, \gamma \in (0,1)$ there are $\zeta = \zeta_{3.2.7}(\alpha,\gamma) \geq (\alpha\gamma)^{O(1)}$ and $\rho = \rho_{3.2.7}(\alpha) \geq \alpha^{O(\alpha^{-6})}$ such that every n-vertex graph G either contains ζn^4 induced copies of C_4, or satisfies the following:*

there is a graph G' on $V(G)$, a partition $V(G) = X_1 \cup \cdots \cup X_k \cup Y$, where $0 \leq k \leq 10\alpha^{-3}$, a subset $Z \subseteq X := X_1 \cup \cdots \cup X_k$, a partition $X \setminus Z = Q_1 \cup \cdots \cup Q_q$ which refines $\{X_1 \setminus Z, \ldots, X_k \setminus Z\}$, and subsets $W_i \subseteq Q_i$ such that the following holds.

1. *$G'[X_i \setminus Z]$ is a clique for every $1 \leq i \leq k$, and $G'[Y]$ is an independent set.*
2. *$|Z| < \alpha n$ and every $z \in Z$ is an isolated vertex in G'.*
3. *The sum of $|Q_i||Q_j|$, taken over all pairs $1 \leq i < j \leq q$ such that (Q_i, Q_j) is non-homogeneous in G', is at most αn^2.*
4. *$|W_i| \geq \rho|X|$ for every $1 \leq i \leq q$, and (W_i, W_j) is homogeneous in G' for every $1 \leq i < j \leq q$.*
5. *$|E(G') \triangle E(G)| < (2\alpha + \gamma)n^2$ and $|E(G'[X \setminus Z]) \triangle E(G[X \setminus Z])| < \gamma n^2$.*

Proof. We prove the lemma with $\zeta_{3.2.7}(\alpha, \gamma) := \zeta_{3.2.6}(\alpha, \gamma)$ and

$$\rho = \rho_{3.2.7}(\alpha) := \rho_{3.1.4}(10\alpha^{-3}, \alpha).$$

Lemma 3.1.4 guarantees that $\rho \geq (0.1\alpha^{3+10\alpha^{-3}})^{O(\alpha^{-3})} \geq \alpha^{O(\alpha^{-6})}$. Note that we may assume that the function $\rho_{3.1.4}(k, \delta)$ (given by Lemma 3.1.4) is monotone decreasing in k.

Apply Lemma 3.2.6 to G with the given α and γ. If G contains at least $\zeta_{3.2.6}(\alpha, \gamma) \cdot n^4$ induced copies of C_4 then the assertion of Lemma 3.2.7 holds. Otherwise, let $X_1, \ldots, X_k, Y$ be as in the statement of Lemma 3.2.6. Note that $k \leq 10\alpha^{-3}$ since $|X_i| \geq 0.1\alpha^3 n$ for every $1 \leq i \leq k$ (as guaranteed by Lemma 3.2.6). Let G'' be the graph obtained from G by making Y an independent set, making $X_1, \ldots, X_k$ cliques and making (X_i, X_j) induced M_2-free for every $1 \leq i < j \leq k$. By Lemma 3.2.6 we have $|E(G''[Y]) \triangle E(G[Y])| < \alpha n^2$ and $|E(G''[X]) \triangle E(G[X])| < \gamma \sum_{i=1}^{k} \binom{|X_i|}{2} + \gamma \sum_{i<j} |X_i||X_j| < \gamma n^2$, where $X := X_1 \cup \cdots \cup X_k$.

If $X = \emptyset$ (i.e., if $V(G) = Y$) then the assertion of the lemma holds trivially with $G' := G''$. Now apply Lemma 3.1.4 with $G''[X]$ as the input graph, $\{X_1, \ldots, X_k\}$ as the partition of $V(G''[X]) = X$, and $\delta = \alpha$. Lemma 3.1.4 supplies a subset $Z \subseteq X$ of size $|Z| < \alpha|X| \leq \alpha n$, a partition $X \setminus Z = Q_1 \cup \cdots \cup Q_q$ which refines $\{X_1 \setminus Z, \ldots, X_k \setminus Z\}$, and subsets $W_i \subseteq Q_i$ $(i = 1, \ldots, q)$, all satisfying Items 1-2 of Lemma 3.1.4. In particular, for every $1 \leq i \leq q$ we have

$$|W_i| \geq \rho_{3.1.4}(k, \alpha) \cdot |X| \geq \rho_{3.1.4}(10\alpha^{-3}, \alpha) \cdot |X| = \rho|X|. \tag{3.3}$$

Let G' be the graph obtained from G'' by making every $z \in Z$ an isolated vertex. Then Item 2 is satisfied. The second part of Item 5 holds because $G'[X \setminus Z] = G''[X \setminus Z]$ and $|E(G''[X]) \triangle E(G[X])| < \gamma n^2$. For the first part of Item 5, note that $|E(G') \triangle E(G'')| < |Z|n < \alpha n^2$, which implies that $|E(G') \triangle E(G)| \leq |E(G') \triangle E(G'')| + |E(G'') \triangle E(G)| < (2\alpha + \gamma)n^2$. Since $G'[X \setminus Z] = G''[X \setminus Z]$ and $G'[Y] = G''[Y]$, it is enough to establish that the assertions of Items 1, 3 and 4 hold if G' is replaced by G''. For Item 1, this is immediate from the definition of G''; for Items 3 and 4, this follows from our choice of $\mathcal{Q} = \{Q_1, \ldots, Q_q\}$ and $W_1, \ldots, W_q$ via Lemma 3.1.4 and from (3.3). ■

3.3 Proof of Theorem 7

We begin by proving the following lemma.

Lemma 3.3.1. *Let $\mathcal{F}$ be a (finite or infinite) family of graphs such that*

1. *$C_4 \in \mathcal{F}$.*
2. *For every $F \in \mathcal{F}$ and for every $v \in V(F)$, the neighbourhood of v (in F) is of size at least 2 and is not a clique.*

Suppose G is a graph with vertex partition $V(G) = X \cup Y$ such that Y is an independent set and $G[X]$ is induced $\mathcal{F}$-free. If one must add/delete at least $\varepsilon|X||Y|$ of the edges between X and Y to make G induced $\mathcal{F}$-free, then G contains at least $\frac{\varepsilon^4}{2^8}|X|^2|Y|^2$ induced copies of C_4.

Proof. Let us pick for every $y \in Y$ a maximal anti-matching $\mathcal{M}(y)$ in $G[N_X(y)]$, that is, a maximal collection of pairwise-disjoint non-edges contained in $N_X(y)$. For every pair of non-edges $\{u,v\},\{u',v'\} \in \mathcal{M}(y)$, there must be at least one non-edge between the vertices $\{u,v\}$ and the vertices $\{u',v'\}$, as otherwise u,v,u',v' would span an induced C_4 in X, in contradiction to the assumptions that $G[X]$ is induced $\mathcal{F}$-free and $C_4 \in \mathcal{F}$. Therefore, for every y there are at least $\binom{|\mathcal{M}(y)|}{2} + |\mathcal{M}(y)| \geq |\mathcal{M}(y)|^2/2$ non-edges inside the set $N_X(y)$. For every $y \in Y$ let $d_2(y)$ denote the number of unordered pairs of vertices in $N_X(y)$, that are non-adjacent. Then the above discussion implies that every $y \in Y$ satisfies

$$d_2(y) \geq \frac{|\mathcal{M}(y)|^2}{2}\,. \tag{3.4}$$

Let G' be the graph obtained from G by deleting, for every $y \in Y$, all edges going between y and the vertices of $\mathcal{M}(y)$. Since $\mathcal{M}(y)$ is spanned by $2|\mathcal{M}(y)|$ vertices, we have

$$|E(G')\triangle E(G)| = 2\sum_{y\in Y}|\mathcal{M}(y)|. \tag{3.5}$$

We now claim that G' is induced $\mathcal{F}$-free. Indeed, suppose $U \subseteq V(G)$ spans an induced copy of some $F \in \mathcal{F}$. Since by assumption $G[X]$ is induced $\mathcal{F}$-free and since $G'[X] = G[X]$, there must be some $y \in U \cap Y$. Since the neighbourhood of y in F is of size at least 2 and is not a clique, and since $G'[Y] = G[Y]$ is an empty graph, there must be $u,v \in U \cap X$ for which $u,v \in N_X(y)$ and $\{u,v\} \notin E(G')$. Now, the fact that u,v are connected to y in G' means that neither of them participated in one of the non-edges of $\mathcal{M}(y)$. But then the fact that $\{u,v\} \notin E(G')$ implies that also $\{u,v\} \notin E(G)$ (because we did not change $G[X]$) which in turn implies that $\{u,v\}$ could have been added to $\mathcal{M}(y)$, contradicting its maximality.

By the assumption of the lemma we thus have $|E(G')\triangle E(G)| \geq \varepsilon|X||Y|$. Combining this with (3.4), (3.5) and Jensen's inequality thus gives

$$\sum_{y\in Y} d_2(y) \geq \frac{1}{2}\sum_{y\in Y}|\mathcal{M}(y)|^2 \geq \frac{1}{2}|Y|\cdot\left(\frac{\sum_{y\in Y}|\mathcal{M}(y)|}{|Y|}\right)^2 = \frac{1}{2}|Y|\cdot\left(\frac{|E(G')\triangle E(G)|}{2|Y|}\right)^2 \geq \frac{\varepsilon^2}{8}|X|^2|Y|.$$

For a pair of distinct vertices $u,v \in X$ set $t(u,v) = 0$ if $\{u,v\} \in E(G)$ and otherwise set $t(u,v)$ to be the number of vertices $y \in Y$ connected to both u and v. Recalling that Y is an independent set in G, we see that u,v belong to at least $\binom{t(u,v)}{2}$ induced copies of C_4. Hence, G contains at least

$$\sum_{u,v\in X}\binom{t(u,v)}{2} \;\geq\; \binom{|X|}{2}\cdot\binom{\sum_{u,v\in X}t(u,v)/\binom{|X|}{2}}{2}$$

$$= \binom{|X|}{2} \cdot \binom{\sum_{y\in Y} d_2(y)/\binom{|X|}{2}}{2}$$
$$\geq \frac{|X|^2}{4} \cdot \frac{(\varepsilon^2|Y|/4)^2}{4} = \frac{\varepsilon^4}{2^8}|X|^2|Y|^2,$$

induced copies of C_4, where the first inequality is Jensen's, the following equality is double-counting, and the last inequality uses our above lower bound for $\sum_{y\in Y} d_2(y)$. ■

We are now ready to prove Theorem 7.

Proof of Theorem 7. Set $\alpha := 2^{-13} \cdot \varepsilon^6$, $\rho := \rho_{3.2.7}(\alpha)$ and

$$\gamma := \frac{1}{2} \cdot (\varepsilon\rho/2)^4.$$

Lemma 3.2.7 guarantees that $\rho \geq \alpha^{O(\alpha^{-6})} \geq 2^{-\operatorname{poly}(1/\varepsilon)}$ and hence also $\gamma \geq 2^{-\operatorname{poly}(1/\varepsilon)}$.

Let G be an n-vertex graph which is ε-far from being induced C_4-free. We apply Lemma 3.2.7 to G with the α and γ defined above. If G contains at least $\zeta_{3.2.7}(\alpha,\gamma) \cdot n^4$ induced copies of C_4 then we are done, as $\zeta_{3.2.7}(\alpha,\gamma) \geq (\alpha\gamma)^{O(1)} \geq 2^{-\operatorname{poly}(1/\varepsilon)}$. Otherwise, let G', $X = X_1 \cup \cdots \cup X_k$, Y, Z, $\mathcal{Q} = \{Q_1, \ldots, Q_q\}$ and $W_i \subseteq Q_i$ be as in Lemma 3.2.7. Let G'' be the graph obtained from G' by doing the following: for every $1 \leq i < j \leq q$, if (W_i, W_j) is a complete (resp. empty) bipartite graph then we turn (Q_i, Q_j) into a complete (resp. empty) bipartite graph. By Item 4 in Lemma 3.2.7, one of these options holds. By Item 3 in Lemma 3.2.7, the number of changes made is at most αn^2. By Item 5 in Lemma 3.2.7 we have

$$\left|E(G'')\triangle E(G)\right| \leq \left|E(G'')\triangle E(G')\right| + \left|E(G')\triangle E(G)\right| < (3\alpha+\gamma)n^2 < \frac{\varepsilon}{2}n^2,$$

implying that G'' is $\frac{\varepsilon}{2}$-far from being induced C_4-free (as G is ε-far from being induced C_4-free). Note that $|X \setminus Z| \geq \frac{\varepsilon}{2}n$, as otherwise deleting all edges incident to the vertices of $X \setminus Z$ would make G'' an empty graph (which in particular is induced C_4-free) by deleting $|X \setminus Z| \cdot n < \frac{\varepsilon}{2}n^2$ edges.

Let us assume first that $G''[X \setminus Z]$ contains an induced copy of C_4, say on the vertices v_1, v_2, v_3, v_4. For $1 \leq s \leq 4$, let i_s be such that $v_s \in Q_{i_s}$. It is easy to see that by the definition of G'', every quadruple $(w_1, \ldots, w_4) \in W_{i_1} \times W_{i_2} \times W_{i_3} \times W_{i_4}$ spans an induced copy of C_4 *in the graph* G'. Thus, G' contains at least

$$|W_{i_1}| \cdot |W_{i_2}| \cdot |W_{i_3}| \cdot |W_{i_4}| \geq \rho^4|X|^4 \geq \rho^4(\varepsilon/2)^4 n^4 = 2\gamma n^4$$

induced copies of C_4, where in the first inequality we used Item 4 of Lemma 3.2.7. Now, by Item 5 in Lemma 3.2.7, $G[X \setminus Z]$ and $G'[X \setminus Z]$ differ on less than γn^2 edges, each of which can participate in at most n^2 induced copies of C_4. Thus, G contains at least $\gamma n^4 \geq 2^{-\operatorname{poly}(1/\varepsilon)} n^4$ induced copies of C_4, as required.

From now on we assume that $G''[X \setminus Z]$ is induced C_4-free, implying that $G''[X]$ is induced C_4-free (as every $z \in Z$ is isolated in G''). Since G'' is $\frac{\varepsilon}{2}$-far from being induced C_4-free, one cannot make G'' induced C_4-free by adding/deleting less than $\frac{\varepsilon}{2}n^2 \geq \varepsilon|X||Y|$ edges between X and Y. Hence, we have $|X||Y| \geq \frac{\varepsilon}{2}n^2$, as otherwise one could remove all edges between X and Y, thus making G'' induced C_4-free by removing at most $\frac{\varepsilon}{2}n^2$ edges. Notice that the conditions of Lemma 3.3.1 hold (with respect to the family $\mathcal{F} = \{C_4\}$) since $G''[Y] = G'[Y]$ is an independent set (by Item 1 in Lemma 3.2.7) and $G''[X]$ is induced C_4-free by assumption. By Lemma 3.3.1, G'' contains at least $\frac{\varepsilon^4}{2^8}|X|^2|Y|^2 \geq \frac{\varepsilon^6}{2^{10}}n^4 = 8\alpha n^4$ induced copies of C_4. Since $|E(G'')\triangle E(G)| < (3\alpha+\gamma)n^2 < 4\alpha n^2$, at least $4\alpha n^4 = \frac{\varepsilon^6}{2^{11}}n^4$ of these copies are also present in G. This completes the proof of the theorem. ■

3.4 Proof of Theorem 3.0.1

In this section we prove Theorem 3.0.1, which we restate as follows.

Theorem 3.4.1. *For every function $g : (0, \frac{1}{2}) \to \mathbb{N}$ there is a graph family $\mathcal{F}$ which contains C_4 and there is a sequence $\{\varepsilon_k\}_{k=1}^{\infty}$ with $\varepsilon_k > 0$ and $\varepsilon_k \to 0$, such the following holds. For every $k \geq 1$ and $n \geq n_0(k)$ there is an n-vertex graph G which is ε_k-far from being induced $\mathcal{F}$-free, but still every induced subgraph of G on $g(\varepsilon_k)$ vertices is induced $\mathcal{F}$-free.*

We will need the following theorem due to Erdős [39].

Theorem 3.4.2 ([39])**.** *For every integer f there is $n_{3.4.2} = n_{3.4.2}(k, f)$ such that every k-uniform hypegraph with $n \geq n_{3.4.2}$ vertices and $n^{k-f^{1-k}}$ edges contains a complete k-partite k-uniform hypergraph with f vertices in each part.*

For integers $k, f \geq 1$, let $B_{k,f}$ be the graph obtained by replacing each vertex of the cycle C_k by a clique of size f, and replacing each edge by a complete bipartite graph.

Lemma 3.4.3. *For every pair of integers $k \geq 3$ and $f \geq 1$ there is $n_{3.4.3} = n_{3.4.3}(k, f)$ such that for every $n \geq n_{3.4.3}$, the graph $B_{k,n/k}$ is $\frac{1}{2k^2}$-far from being induced $\{C_4, B_{k,f}\}$-free.*

Proof. Let $V_1, \ldots, V_k$ be the sides of $G := B_{k,n/k}$ (each a clique of size n/k). Let G' be a graph obtained from G by adding/deleting at most $\frac{v(G)^2}{2k^2} = \frac{n^2}{2k^2}$ edges. Our goal is to show that G' is not induced $\{C_4, B_{k,f}\}$-free. Let H be the k-partite k-uniform hypergraph with parts $V_1, \ldots, V_k$ whose edges are all k-tuples $(v_1, \ldots, v_k) \in V_1 \times \cdots \times V_k$ such that $v_1 v_2 \ldots v_k v_1$ is an induced cycle in G'. Note that in G, every such k-tuple spans an induced cycle, and that adding/deleting an edge can destroy at most $\left(\frac{n}{k}\right)^{k-2}$ such cycles. Thus, G' contains at least $\left(\frac{n}{k}\right)^k - \frac{n^2}{2k^2}\left(\frac{n}{k}\right)^{k-2} = \frac{1}{2}\left(\frac{n}{k}\right)^k$ of these induced cycles, implying that $e(H) \geq \frac{1}{2}\left(\frac{n}{k}\right)^k$. For a large enough n we have $\frac{1}{2}\left(\frac{n}{k}\right)^k \geq n^{k-f^{1-k}}$ and $n \geq n_{3.4.2}(k, f)$. Thus, by Theorem 3.4.2, H contains a complete k-partite k-uniform hypergraph with parts $U_i \subseteq V_i$, each of size f. This means that in the graph G', (U_i, U_j) is a complete bipartite graph if $j - i \equiv \pm 1 \pmod{k}$ and an empty bipartite graph otherwise. If $G'[U_i]$ is a clique for every $1 \leq i \leq k$ then $U_1 \cup \cdots \cup U_k$ spans an induced copy of $B_{k,f}$ in G'. Suppose then that U_i is not a clique for some $1 \leq i \leq k$, say $i = 1$, and let $x, y \in U_1$ be such that $\{x, y\} \notin E(G')$. Then for every $z \in U_2$ and $w \in U_k$, $\{x, y, z, w\}$ spans an induced copy of C_4 in G'. Thus, in any case G' is not induced $\{C_4, B_{k,f}\}$-free. ■

Proof of Theorem 3.4.1. For $k \geq 5$ put $\varepsilon_k = \frac{1}{2k^2}$ and $f_k = g(\varepsilon_k)$. We will show that the family $\mathcal{F} = \{C_4\} \cup \{B_{k,f_k} : k \geq 5\}$ satisfies the requirement. Let $k \geq 5$, let $n \geq n_{3.4.3}(k, f_k)$ and set $G = B_{k,n/k}$. By Lemma 3.4.3, G is ε_k-far from being induced $\{C_4, B_{k,f_k}\}$-free. Since $C_4, B_{k,f_k} \in \mathcal{F}$, we get that G is ε_k-far from being induced $\mathcal{F}$-free.

We claim that for every $4 \leq \ell < k$, G is induced C_ℓ-free. Suppose, by contradiction, that $x_1, \ldots, x_\ell, x_1$ is an induced ℓ-cycle in G. Let $V_1, \ldots, V_k$ be the sides of $G = B_{k,n/k}$. If $|\{x_1, \ldots, x_\ell\} \cap V_i| \leq 1$ for every $1 \leq i \leq k$ then $x_1, \ldots, x_\ell$ are contained in an induced path, which is impossible. So there is some $1 \leq i \leq k$ for which $|\{x_1, \ldots, x_\ell\} \cap V_i| \geq 2$. Suppose without loss of generality that $x_1, x_2 \in V_1$ (recall that $V_1, \ldots, V_k$ are cliques). Then $x_3 \in V_2$ or $x_3 \in V_k$, and in either case x_1, x_2, x_3 span a triangle, a contradiction.

We conclude that the smallest $F \in \mathcal{F}$ which is an induced subgraph of G, is $F = B_{k,f_k}$. Thus, every induced subgraph of G on less than $v(B_{k,f_k}) = k \cdot g(\varepsilon_k)$ vertices is induced $\mathcal{F}$-free, completing the proof. ■

Chapter 4

A Hierarchy Theorem for Query-Complexity via a Generalized Turán Result

4.1 Background on (Generalized) Turán Problems

Turán's Theorem [113], one of the cornerstone results in graph theory, determines the maximum number of edges in an n-vertex graph that does not contain a copy of K_t (the complete graph on t vertices). Turán's problem is the following more general question: for a fixed graph H and an integer n, what is the maximum number of edges in an n-vertex H-free graph? This quantity is denoted by $\mathrm{ex}(n, H)$. Estimating $\mathrm{ex}(n, H)$ for various graphs H is the fundamental problem of extremal graph theory (see [105] for a survey).

Alon and Shikhelman [15] have recently initiated the systematic study of the following natural generalization of the function $\mathrm{ex}(n, H)$; for fixed graphs H and T, let $\mathrm{ex}(n, T, H)$ denote the maximum number of copies[1] of T in an n-vertex graph that contains no copy of H. Note that $\mathrm{ex}(n, H) = \mathrm{ex}(n, K_2, H)$. Problems involving the estimation of $\mathrm{ex}(n, H, T)$ (for specific graphs T, H) are known as *generalized Turán problems*. Some concrete problems of this type have already been considered by Erdős in the 60's, with [41, 42] establishing bounds on $\mathrm{ex}(n, T, H)$ for various pairs T, H. For the sake of brevity, we refer the reader to [15] for more background and motivation, as well as examples of some well-studied problems in extremal combinatorics which relate to the study of $\mathrm{ex}(n, T, H)$ for various pairs H and T.

One special case of Turán's problem which has been extensively studied is the estimation of $\mathrm{ex}(n, C_k)$ (where C_k denotes the k-cycle, i.e., the cycle of length k). While for odd k it is known [104] that $\mathrm{ex}(n, C_k) = \lfloor n^2/4 \rfloor$ (for large enough n), the problem of determining the order of magnitude of $\mathrm{ex}(n, C_k)$ for even k is a long-standing, major open problem in extremal graph theory, see the survey [114] and its references.

The problem of estimating $\mathrm{ex}(n, C_k, C_\ell)$ has recently received a lot of attention. Bollobás and Győri [24] proved that $\mathrm{ex}(n, C_3, C_5) = \Theta\big(n^{3/2}\big)$. Győri and Li [70] extended this result by considering $\mathrm{ex}(n, C_3, C_{2\ell+1})$. The dependence of their upper bound on ℓ was subsequently improved upon by Alon and Shikhelman [15]. At this moment, the best known bounds imply that

$$\Omega(\mathrm{ex}(n, \{C_4, C_6, \ldots, C_{2\ell}\})) \leq \mathrm{ex}(n, C_3, C_{2\ell+1}) \leq O(\ell \cdot \mathrm{ex}(n, C_{2\ell})), \tag{4.1}$$

[1] Throughout this chapter, unless explicitly stated otherwise, we always count *unlabeled* copies.

where $\mathrm{ex}(n, \{C_4, C_6, \ldots, C_{2\ell}\})$ is the maximal number of edges in an n-vertex graph with no copy of C_{2t} for any $2 \le t \le \ell$. The lower bound in (4.1) was proved in [70], and the upper bound in [15]. These bounds were also independently obtained by Füredi and Özkahya [53]. The lower and upper bounds in (4.1) are known to be of the same order of magnitude, $\Theta\big(n^{1+1/\ell}\big)$, for $\ell \in \{2, 3, 5\}$ (see [114] and the references therein).

Recall that our main generalized-Turán-type result, Theorem 12, extends the aforementioned results of [15, 53, 70] by determining the order of magnitude of $\mathrm{ex}(n, C_k, C_\ell)$ for all fixed $k, \ell \ge 4$. As mentioned in Section 1.2.1, the statement of Theorem 12 does not in fact cover every possible choice of values for distinct $k, \ell \ge 4$, since some cases are (in a sense) trivial. These missing cases are: (a) k is even and ℓ is odd; and (b) k and ℓ are both odd and $k > \ell$. Observe that in either of these cases, a blow-up of C_k is C_ℓ-free, implying immediately that $\mathrm{ex}(n, C_k, C_\ell) = \Theta_k(n^k)$. It is worth noting that in these so-called "dense" cases, while it is trivial to determine the *order of magnitude* of $\mathrm{ex}(n, C_k, C_\ell)$, it may be very challenging to get an *exact* (or even *asymptotic*) result. A prime example of this is the famous problem of the exact determination of $\mathrm{ex}(n, C_5, C_3)$, which was settled in [68, 73].

Next, let us consider the case of $\mathrm{ex}(n, C_3, C_{2\ell})$. Here we will prove the following.

Proposition 4.1.1. *For every $\ell \ge 2$ we have*

$$\Omega(\mathrm{ex}(2n/3, \{C_4, C_6, \ldots, C_{2\ell}\})) \le \mathrm{ex}(n, C_3, C_{2\ell}) \le O_\ell\big(\mathrm{ex}(n, C_{2\ell})\big).$$

As in the case of (4.1), the lower and upper bounds in Proposition 4.1.1 are known to be of the same order of magnitude for $\ell \in \{2, 3, 5\}$. We note that Proposition 4.1.1 was independently proved by [53].

Finally, we consider the problem of estimating $\mathrm{ex}(n, P_k, C_\ell)$, where P_k denotes the path of length k, namely, the path with k edges. As a side-product of our methods, we obtain the following theorem, which gives a tight bound on $\mathrm{ex}(n, P_k, C_{2\ell})$ for all $k, \ell \ge 2$.

Theorem 4.1.2. *For every $k \ge 2$, we have*

$$\mathrm{ex}(n, P_k, C_{2\ell}) = \begin{cases} \Theta_k(n^{k/2+1}) & \ell = 2, \\ \Theta_k(\ell^{\lfloor (k+1)/2 \rfloor} n^{\lceil (k+1)/2 \rceil}) & \ell \ge 3. \end{cases}$$

To complement Theorem 4.1.2, note that $\mathrm{ex}(n, P_k, C_{2\ell+1}) = \Theta_k(n^{k+1})$, since a blowup of P_k does not contain odd cycles.

The rest of this chapter is organized as follows. In Section 4.2 we discuss an additional application of Theorem 12 to the study of graph removal lemmas. Section 4.3 contains the proofs of all lower bounds in Theorems 12 and 4.1.2 and Proposition 4.1.1, and can be read independently of the other sections. In Section 4.4 we give a tight upper bound on $\mathrm{ex}(n, C_{2k+1}, C_{2k+3})$ for every $k \ge 2$. This case of Theorem 12 — that is, where the cycle lengths are consecutive odd integers — turns out to require a different argument than the one we use for all other cases of Theorem 12. The problem of upper-bounding $\mathrm{ex}(n, C_{2k+1}, C_{2k+3})$ for $k \ge 2$ appears to be significantly harder than the case $\mathrm{ex}(n, C_3, C_5)$ which was resolved by Bollobás–Győri [24]. This is best evidenced by the fact that while $\mathrm{ex}(n, C_3, C_5) = \Theta(n^{3/2})$, for $k \ge 2$ we have $\mathrm{ex}(n, C_{2k+1}, C_{2k+3}) = \Theta_k(n^k)$. Section 4.5 is devoted to establishing some key lemmas, which then allow us to prove Theorem 4.1.2 and Proposition 4.1.1 in Section 4.6 and all remaining case of Theorem 12 in Section 4.7. The proof of one key lemma, namely Lemma 4.5.5, relies on a bound for the skew version of the even-cycle Turán problem, due to Naor and Verstraëte [87]. Finally, in Section 4.8 we apply Theorem 12 in order to prove Theorems 10 and 11 (as well as Theorem 4.2.1, which is stated in the next section).

4.2 Removal-Lemma Bounds for Forbidden-Cycles Properties

Here we describe an additional application of Theorem 12, which is closely related to Theorem 10. We start with some background. Recall the function $w_{\mathcal{P}}(\varepsilon)$ defined in Definition 2. In the case where $\mathcal{P}$ is the k-colorability property, we will use $w_{k\text{-col}}(\varepsilon)$ instead of $w_{\mathcal{P}}(\varepsilon)$. The earliest upper bound on $w_{k\text{-col}}(\varepsilon)$ was given by Rödl and Duke [91], whose proof relied on the regularity lemma [110], and thus only supplied tower-type bounds. A significantly better bound was obtained by Goldreich, Goldwasser and Ron [59] who proved that $w_{k\text{-col}}(\varepsilon) = \text{poly}(1/\varepsilon)$. In a recent breakthrough, Sohler [106] obtained the nearly tight bound $w_{k\text{-col}}(\varepsilon) = \tilde{\Theta}(1/\varepsilon)$.

For a set of integers L, let us say that a graph is *L-free* if it is C_ℓ-free for every $\ell \in L$. Let $\mathcal{P}_L$ denote the property of being L-free. In what follows, we will use the notation w_L instead of $w_{\mathcal{P}_L}$. The result of [59] stating that $w_{2\text{-col}}(\varepsilon) = \text{poly}(1/\varepsilon)$ is then equivalent to the statement that if L consists of *all* odd integers then $w_L(\varepsilon) = \text{poly}(1/\varepsilon)$. Another related result is due to Alon [2] who proved that $w_L(\varepsilon) = \text{poly}(1/\varepsilon)$ if L contains at least one even integer, and that $w_L(\varepsilon)$ is super-polynomial whenever L is a *finite* set of odd integers. Alon and Shapira [12] asked if one can extend the above results by characterizing the sets of integers L for which $w_L(\varepsilon) = \text{poly}(1/\varepsilon)$. Our next theorem solves all cases not handled by previous results.

Theorem 4.2.1. *Let $L = \{\ell_1, \ell_2, \ldots\}$ be an infinite increasing sequence of odd integers. Then*

$$w_L(\varepsilon) = \text{poly}(1/\varepsilon) \quad \textit{if and only if} \quad \limsup_{j \longrightarrow \infty} \frac{\log \ell_{j+1}}{\log \ell_j} < \infty.$$

By the above theorem, as long as ℓ_j does not grow faster than 2^{2^j}, one can get a polynomial bound on $w_L(\varepsilon)$, while for any (significantly) faster-growing ℓ_j one cannot get such a bound.

Let us give another perspective on Theorem 4.2.1. As mentioned in Chapter 1, several prior works [56, 8] have raised the problem of characterizing the hereditary properties $\mathcal{P}$ for which $w_{\mathcal{P}}(\varepsilon) = \text{poly}(1/\varepsilon)$. Since this problem seemed (and still seems) to be out of reach, the authors of [12] asked if one can at least solve a (very) special case of this problem by characterizing the sets L for which $w_L(\varepsilon) = \text{poly}(1/\varepsilon)$. This problem is resolved by Theorem 4.2.1.

4.3 Lower Bound on $\text{ex}(n, C_k, C_\ell)$ and $\text{ex}(n, P_k, C_\ell)$

In this section we prove all lower bounds in Theorems 12 and 4.1.2 and Proposition 4.1.1. We start with the following two claims, which handle the case where the forbidden cycle is *not* C_4. For every $\ell \neq 4$, Claim 4.3.1 gives lower bounds on $\text{ex}(n, C_k, C_\ell)$ and $\text{ex}(n, P_k, C_\ell)$ which have the correct dependence on n. To get the correct dependence on ℓ for $\ell \gg k$, we need Claim 4.3.2, which gives a general lower bound for $\text{ex}(n, T, H)$, but is only applicable when H (that is, C_ℓ) is somewhat larger than T (that is, C_k or P_k). To prove the lower bound for all values of k and $\ell \neq 4$, we need to combine these two claims, which is done in Corollary 4.3.3. For a graph G, denote by $\alpha(G)$ the independence number of G.

Claim 4.3.1. *For a pair of distinct $k \geq 3$ and $4 \neq \ell \geq 3$ we have $\text{ex}(n, C_k, C_\ell) = \Omega_k\big(n^{\lfloor k/2 \rfloor}\big)$. For $k \geq 2$ and $4 \neq \ell \geq 3$ we have $\text{ex}(n, P_k, C_\ell) = \Omega_k\big(n^{\lceil (k+1)/2 \rceil}\big)$.*

Proof. We start with the first part of the claim. Let I be a maximum independent set of the k-cycle $1, \ldots, k$. Replace each $i \in I$ with a vertex-set of size m, where different vertices are replaced with disjoint

sets and all of these sets are disjoint from $[k] \setminus I$. Edges of C_k are replaced with complete bipartite graphs. In other words, we take a blowup of C_k in which vertices $i \in [k] \setminus I$ are not blown up, while vertices $i \in I$ are blown up to size m. As $|I| = \alpha(C_k) = \lfloor k/2 \rfloor$, the resulting graph has $n := \lfloor k/2 \rfloor \cdot m + \lceil k/2 \rceil$ vertices and $m^{|I|} = m^{\lfloor k/2 \rfloor} = \Omega_k\left(n^{\lfloor k/2 \rfloor}\right)$ copies of C_k. It is easy to check that this graph is C_ℓ-free by our assumptions that $\ell \neq k$ and $\ell \neq 4$.

We now prove the second part of the claim using a similar construction. Let I be a maximum independent set of the path P_k on the vertices $1, \ldots, k+1$. Replace each $i \in I$ with a vertex-set of size m, where different vertices are replaced with disjoint sets and all of these sets are disjoint from $[k+1] \setminus I$. Edges of P_k are replaced with complete bipartite graphs. As $|I| = \alpha(P_k) = \lceil (k+1)/2 \rceil$, the resulting graph has $n := \lceil (k+1)/2 \rceil \cdot m + \lfloor (k+1)/2 \rfloor$ vertices and $m^{|I|} = m^{\lceil (k+1)/2 \rceil} = \Omega_k(n^{\lceil (k+1)/2 \rceil})$ copies of P_k. It is easy to check that this graph is C_ℓ-free by our assumptions that $\ell \neq 4$. ■

Claim 4.3.2. *Let T, H be graphs on t and h vertices, respectively, such that $h - \alpha(H) - 1 \geq t - \alpha(T)$. Then for every $n \geq h - \alpha(H) - 1 + \alpha(T)$, it holds that $\mathrm{ex}(n, T, H) \geq \Omega_t\left((h - \alpha(H))^{t-\alpha(T)} n^{\alpha(T)}\right)$.*

Proof. Suppose that $V(T) = \{1, \ldots, t\}$ and let I be a maximum independent set of T. Let $U_1, \ldots, U_t$ be disjoint vertex-sets such that $|U_1| + \cdots + |U_t| = n$ and such that the following holds: $\sum_{i \in V(T) \setminus I} |U_i| = h - \alpha(H) - 1$, these $h - \alpha(H) - 1$ vertices are divided as equally as possible among the $t - \alpha(T)$ sets $(U_i)_{i \in V(T) \setminus I}$, and the $n - h + \alpha(H) + 1$ vertices of $\bigcup_{i \in I} U_i$ are divided as equally as possible among $(U_i)_{i \in I}$. Then none of $U_1, \ldots, U_t$ is empty by the assumptions of the claim. Define a graph G on $U_1 \cup \cdots \cup U_t$ by making (U_i, U_j) a complete bipartite graph if $\{i, j\} \in E(T)$, and an empty bipartite graph otherwise (there are no edges inside the sets $U_1, \ldots, U_t$). Then G has $\Omega_t\left((h - \alpha(H))^{t-\alpha(T)} n^{\alpha(T)}\right)$ copies of T. It remains to show that G is H-free. Assume by contradiction that there is a copy of H in G. Then this copy contains two adjacent vertices which are both in $\bigcup_{i \in I} U_i$, since $\sum_{i \in V(T) \setminus I} |U_i| < h - \alpha(H)$. But $\bigcup_{i \in I} U_i$ is an independent set in G, as I is an independent set in T and G is a blowup of T, a contradiction. ■

We are now ready to prove the lower bounds in the last two cases of Theorem 12 and in the second case of Theorem 4.1.2. In other words, we handle all cases in which the forbidden cycle is not C_4.

Corollary 4.3.3. *For a pair of distinct $k \geq 3$ and $4 \neq \ell \geq 3$ we have $\mathrm{ex}(n, C_k, C_\ell) = \Omega_k\left(\ell^{\lceil k/2 \rceil} n^{\lfloor k/2 \rfloor}\right)$. For $k \geq 2$ and $4 \neq \ell \geq 3$ we have $\mathrm{ex}(n, P_k, C_\ell) = \Omega_k(\ell^{\lfloor (k+1)/2 \rfloor} n^{\lceil (k+1)/2 \rceil})$.*

Proof. Note that since our bound hides constants that depend on k, if $\ell < k + 3$ then the assertion of the corollary follows from Claim 4.3.1. So we may assume that $\ell \geq k + 3$, which implies that $\lceil \ell/2 \rceil \geq \lceil k/2 \rceil + 1$. Under this assumption, Claim 4.3.2 is applicable to $(T, H) = (C_k, C_\ell)$, giving $\mathrm{ex}(n, C_k, C_\ell) = \Omega_k(\ell^{\lceil k/2 \rceil} n^{\lfloor k/2 \rfloor})$, and to $(T, H) = (P_k, C_\ell)$, giving $\mathrm{ex}(n, P_k, C_\ell) = \Omega_k(\ell^{\lfloor (k+1)/2 \rfloor} n^{\lceil (k+1)/2 \rceil})$. ■

When excluding C_4, a different construction is required. The construction we use is due to Erdős and Rényi [48]. The case of $\mathrm{ex}(n, C_3, C_4)$ was handled (using the same construction) in [15]. Via the following lemma, we get the lower bound in the first item of Theorem 12 and of Theorem 4.1.2.

Lemma 4.3.4. *Let q be a prime power and set $n = q^2 - 1$. Then there is an n-vertex C_4-free graph which contains at least $\left(\frac{1}{2k} - o(1)\right) n^{\frac{k}{2}}$ copies of C_k for every $4 \neq k \geq 3$, and at least $\left(\frac{1}{2} - o(1)\right) n^{\frac{k}{2}+1}$ copies of P_k for every $k \geq 1$. Here, the $o(1)$ term is a function which depends on k and tends to 0 as n tends to infinity. Hence, $\mathrm{ex}(n, C_k, C_4) \geq \left(\frac{1}{2k} - o(1)\right) n^{\frac{k}{2}}$ for every $4 \neq k \geq 3$, and $\mathrm{ex}(n, P_k, C_4) \geq \left(\frac{1}{2} - o(1)\right) n^{\frac{k}{2}+1}$ for every $k \geq 1$.*

Proof. The last part of the theorem is deduced from the first part as follows. It is known that for every large enough x there is a prime in the interval $[x - x^\theta, x]$ for an absolute constant $\theta \in [\frac{1}{2}, 1)$, see e.g. [19]. Fixing a large enough n, let p be a prime in $[x - x^\theta, x]$ for $x = n^{1/2}$. Now take the construction from the first part of the theorem on $p^2 - 1$ vertices and add isolated vertices to get a graph on n vertices. This graph gives the required lower bounds on $\mathrm{ex}(n, C_k, C_4)$ and $\mathrm{ex}(n, P_k, C_4)$.

From now on we assume that $n = q^2 - 1$, where q is a prime power. Let $\mathbb{F}$ be the field with q elements. The vertex set of G is $\mathbb{F}^2 \setminus \{(0,0)\}$ and a pair of vertices $(a,b), (c,d)$ are adjacent if and only if $ac + bd = 1$. Note that $(a,b) \in V(G)$ has a loop if and only if $a^2 + b^2 = 1$. The number of solutions to $x^2 + y^2 = 1$ is at most $2q$, since for every fixed $x \in \mathbb{F}$ there are at most 2 solutions for y. This implies that the number of loops is at most $2q$. Note that for every $(a,b) \in V(G)$ there are q solutions (x,y) to $ax + by = 1$. Thus, the degree of every $(a,b) \in V(G)$ is either $q-1$ or q, depending on whether or not (a,b) has a loop. This implies that for every $k \geq 1$, G contains at least $\frac{1}{2}n(q-1)(q-2)\dots(q-k) = \left(\frac{1}{2} - o(1)\right) n^{\frac{k}{2}+1}$ paths of length k.

Observe that for every pair of vertices $(a,b), (c,d) \in V(G)$, there is at most one solution to the system $ax + by = cx + dy = 1$, implying that (a,b) and (c,d) have at most one common neighbour. This shows that G is C_4-free. To finish the proof, it remains to show that the number of k-cycles in G is as stated. Since this was proved for $k = 3$ in [15], we may assume from now on that $k \geq 5$.

Note that if $(a,b), (c,d) \in V(G)$ are linearly independent and have no loops then they have a common neighbor. Indeed, by linear independence there is a (unique) solution to the system $ax + by = cx + dy = 1$. As (a,b) and (c,d) do not have loops, this solution is neither (a,b) nor (c,d), and hence it is a common neighbour of (a,b) and (c,d). As the number of loops in G is at most $2q$, the number of pairs of vertices $(a,b), (c,d) \in V(G)$ for which either (a,b) or (c,d) has a loop is at most $2qn$. Furthermore, the number of collinear pairs $(a,b), (c,d) \in V(G)$ is $\frac{(q-1)n}{2}$. Therefore, all but $2qn + \frac{(q-1)n}{2} \leq 3nq$ of the pairs of vertices are linearly independent and do not have loops, and hence have a common neighbor. We have thus proven the following.

Fact 4.3.5. *All but $3nq$ of the pairs of vertices in G have a common neighbor.*

Note that for every $t \geq 2$ and $v_1, v_{t+1} \in V(G)$, the number of paths of length t between v_1 and v_{t+1} is at most q^{t-2}. Indeed, consider a path $v_1, \dots, v_{t+1}$. Since the maximal degree in G is q, the number of choices of $v_2, \dots, v_{t-1}$ is at most q^{t-2}. Since v_t is a common neighbour of v_{t-1} and v_{t+1}, there is at most one choice for v_t given $v_2, \dots, v_{t-1}$.

A path is *good* if its endpoints have a common neighbour which is not on the path, and otherwise it is *bad*. To complete the proof, it is enough to show that for every $t \geq 3$, the number of bad paths of length t is $O(nq^{t-1})$. Indeed, we already proved that G contains at least $\left(\frac{1}{2} - o(1)\right) n^{\frac{k}{2}}$ paths of length $k-2$. Since the number of bad paths of length $k-2$ is $O(nq^{k-3}) = O\left(n^{\frac{k-1}{2}}\right)$, the number of good paths of length $k-2$ is at least $\left(\frac{1}{2} - o(1)\right) n^{\frac{k}{2}}$. A good path of length $k-2$ can be made into a k-cycle by adding the (unique) common neighbor of the endpoints of the path. Since every cycle contains k subpaths of length $k-2$, the lemma follows.

It thus remains to show that for every $t \geq 3$, the number of bad paths of length t is $O(nq^{t-1})$. There are two types of bad paths: those whose endpoints do not have a common neighbor, and those whose endpoints have a common neighbour which is on the path. First, by Fact 4.3.5, the number of pairs of vertices $u, v \in V(G)$ which do not have a common neighbor is at most $3nq$. We proved that for each such

u, v there are at most q^{t-2} paths of length t between u and v. Thus, there are at most $O(nq^{t-1})$ paths of length t whose endpoints do not have a common neighbor. Second, let $u, v \in V(G)$ be vertices having a common neighbor and let w be their unique common neighbor. The number of paths of length t from u to v in which w is at distance i from u (and hence at distance $t-i$ from v) is at most q^{t-3} if $i \in \{1, t-1\}$ and at most $q^{i-2}q^{t-i-2} = q^{t-4}$ if $2 \leq i \leq t-2$. By summing over $1 \leq i \leq t-1$ we get that the number of paths of length t from u to v which contain w is at most $2q^{t-3} + (t-3)q^{t-4} = O(q^{t-3})$. Since the number of choices for u, v is at most $\binom{n}{2}$, the total number of paths of length t that contain the common neighbor of their endpoints is $O(n^2 q^{t-3}) = O(nq^{t-1})$. In conclusion, the number of bad paths is $O(nq^{t-1})$, as required. ■

We end this section by proving the lower bound in Proposition 4.1.1.

Claim 4.3.6. *For every $\ell \geq 3$ we have $\mathrm{ex}(n, C_3, C_{2\ell}) \geq \frac{1}{2} \cdot \mathrm{ex}\left(\frac{2n}{3}, \{C_4, C_6, \ldots, C_{2\ell}\}\right)$.*

Proof. We use an argument similar to the one used in [70]. Let $G' = (A \cup B, E)$ be a maximum size $\frac{n}{3} \times \frac{n}{3}$ bipartite graph with no $C_4, C_6, \ldots, C_{2\ell}$. Let G be the graph obtained from G' by replacing every vertex of A by an edge (and replacing edges of G' by copies of $K_{2,1}$). Then G has n vertices, and one triangle per each edge of G'; so G contains $e(G') \geq \frac{1}{2} \cdot \mathrm{ex}\left(\frac{2n}{3}, \{C_4, C_6, \ldots, C_{2\ell}\}\right)$ triangles. Now assume by contradiction that C is a copy of $C_{2\ell}$ in G. By contracting the edges of C inside A, we get a closed walk C' in G' of length at most 2ℓ. For each $a \in A$, let a_1 and a_2 denote the two "copies" of a in G. If for every $a \in C' \cap A$, only one of the copies of a is in C, then $C' = C$, in contradiction to the $C_{2\ell}$-freeness of G'. So there is some $a \in A$ such that $a_1, a_2 \in C$. In the cycle C there are two paths between a_1 and a_2, and since $|C| = 2\ell \geq 6$, one of these paths must have length at least 3. Such a path has the form a_1, b_1, P, b_2, a_2 with $b_1, b_2 \in B$ distinct and with P being a path in G between b_1 and b_2 which does not go through a_1 or a_2. Contracting the edges of P inside A gives a walk in G' between b_1 and b_2 which does not go through a. It follows that G' contains a path P' between b_1 and b_2 avoiding a. Then a, b_1, P', b_2, a is a cycle in G' of length at most 2ℓ, in contradiction to the choice of G'. ■

4.4 Proof of Theorem 12: The Case $\mathrm{ex}(n, C_{2k+1}, C_{2k+3})$

In this section we give a tight upper bound for $\mathrm{ex}(n, C_{2k+1}, C_{2k+3})$ when $k \geq 2$. Let us introduce some notation that we will use throughout this chapter. For a graph G and disjoint sets $X, Y \subseteq V(G)$, we denote by $E(X, Y)$ the set of edges with one endpoint in X and one endpoint in Y, and set $e(X, Y) = |E(X, Y)|$. For $v \in V(G)$ and $X \subseteq V(G)$, denote $N_X(v) = \{x \in X : (v, x) \in E(G)\}$.

Let $U_1, \ldots, U_s$ be disjoint vertex sets in a graph. A $(U_1, \ldots, U_s)$-*path* is a path $u_1, \ldots, u_s$ with $u_i \in U_i$. Similarly, a $(U_1, \ldots, U_s)$-*cycle* is a cycle $u_1, \ldots, u_s, u_1$ with $u_i \in U_i$. Let $p(U_1, \ldots, U_s)$ denote the number of $(U_1, \ldots, U_s)$-paths and let $c(U_1, \ldots, U_s)$ denote the number of $(U_1, \ldots, U_s)$-cycles. We will frequently use a simple averaging argument, given by the following claim. The statement of Claim 4.4.1 is about unlabeled copies (as is the case throughout the chapter), but in its proof we will also consider labeled copies.

Claim 4.4.1. *Let G be a graph. If for every partition $V(G) = U_1 \cup \cdots \cup U_k$ it holds that $c(U_1, \ldots, U_k) \leq r$, then the number of copies of C_k in G is at most $\frac{1}{2}k^{k-1}r$. Similarly, if for every partition $V(G) = U_1 \cup \cdots \cup U_k$ it holds that $p(U_1, \ldots, U_k) \leq r$, then the number of copies of P_{k-1} in G is at most $\frac{1}{2}k^k r$.*

Proof. Let $V(G) = U_1 \cup \cdots \cup U_k$ be a random partition, generated according to $\mathbb{P}[v \in U_i] = \frac{1}{k}$ for each $v \in V(G)$ and $1 \leq i \leq k$, independently. Observe that a *labeled* copy $u_1, \ldots, u_k$ of P_{k-1} is a $(U_1, \ldots, U_k)$-path (namely, satisfies $u_i \in U_i$ for every $1 \leq i \leq k$) with probability k^{-k}. Similarly, a *labeled* copy of C_k is a $(U_1, \ldots, U_k)$-cycle with probability k^{-k}. Since P_{k-1} has 2 automorphisms and C_k has $2k$ automorphisms, we have $\mathbb{E}[c(U_1, \ldots, U_k)] = \#C_k(G) \cdot 2k \cdot k^{-k}$ and $\mathbb{E}[p(U_1, \ldots, U_k)] = \#P_{k-1}(G) \cdot 2 \cdot k^{-k}$, where $\#C_k(G)$ (resp. $\#P_{k-1}(G)$) denotes the number of *unlabeled* copies of C_k (resp. P_{k-1}) in G. Since these expectations are not larger than r by assumption, the claim follows. ■

In what follows, let us denote the vertices of C_{2k+1} (the $(2k+1)$-cycle) by $1, \ldots, 2k+1$, with edges $\{1,2\}, \ldots, \{2k, 2k+1\}, \{2k+1, 1\}$. For a graph G, denote by $\mathcal{I}(G)$ the set of all non-empty independent sets of G. We will need the following trivial (yet somewhat complicated to state) claim.

Claim 4.4.2. *Let J be a non-empty independent set of C_{2k+1}. Then there is $I \in \mathcal{I}(C_{2k+1})$ which contains J and satisfies the following. Let $i_1, \ldots, i_r$ be the elements of I in the order they appear when traversing the cycle $1, \ldots, 2k+1$. Then for every $1 \leq j \leq r$, i_j and i_{j+1} are at distance either 2 or 3, namely $i_{j+1} - i_j \equiv 2, 3 \pmod{2k+1}$, and if i_j and i_{j+1} are at distance 3 then $i_j \in J$ or $i_{j+1} \in J$.*

Proof. If $|J| = 1$, say without loss of generality $J = \{1\}$, then $I = \{2j-1 : 1 \leq j \leq k\}$ is easily seen to satisfy the requirements of the claim. Assume then that $|J| \geq 2$, and let $j_1, \ldots, j_r$ be the elements of J, as they appear when traversing the $(2k+1)$-cycle $1, \ldots, 2k+1$. For each $1 \leq i \leq r$, we greedily pick an independent set I_i in the path connecting j_i and j_{i+1}, which contains both j_i and j_{i+1}, as follows. In addition to j_i and j_{i+1}, we add to I_i the elements $j_i + 2, j_i + 4, \ldots$ until we reach j_{i+1} or $j_{i+1} - 1$. If we reached j_{i+1}, then the distance between every pair consecutive elements of I_i is 2, and if we reached $j_{i+1} - 1$ then this true for all pairs except for $j_{i+1} - 3, j_{i+1}$. It is now easy to see that $I = \bigcup_{i=1}^{r} I_i$ satisfies the requirements of the claim. ■

Lemma 4.4.3. *For every $k \geq 2$ it holds that $\mathrm{ex}(n, C_{2k+1}, C_{2k+3}) \leq (2k+1)^{2k} 2^{2k+1} n^k$.*

Proof. Let G be an n-vertex C_{2k+3}-free graph. By claim 4.4.1 it is sufficient to prove that for every partition $V(G) = U_1 \cup \cdots \cup U_{2k+1}$ we have $c(U_1, \ldots, U_{2k+1}) \leq 2^{2k+1} n^k$. We will actually prove that

$$c(U_1, \ldots, U_{2k+1}) \leq \sum_{I \in \mathcal{I}(C_{2k+1})} \prod_{i \in I} |U_i|. \tag{4.2}$$

This will be sufficient, as C_{2k+1} has at most 2^{2k+1} independent sets, and each of these sets contributes at most n^k to the above sum. Assume by contradiction that (4.2) is false. Let $\mathcal{C}$ denote the set of all $(U_1, \ldots, U_{2k+1})$-cycles in G. We first show that there is $C = (u_1, \ldots, u_{2k+1}) \in \mathcal{C}$ such that for every $I \in \mathcal{I}(C_{2k+1})$ there is $C' \in \mathcal{C} \setminus \{C\}$ which contains $\{u_i : i \in I\}$. We find C greedily as follows. As long as there is $C = (u_1, \ldots, u_{2k+1}) \in \mathcal{C}$ and $I \in \mathcal{I}(C_{2k+1})$ such that C is the only $(U_1, \ldots, U_{2k+1})$-cycle containing $\{u_i : i \in I\}$, we remove C from $\mathcal{C}$, and we say that C was removed due to $\{u_i : i \in I\}$. Fixing any $I \in \mathcal{I}(C_{2k+1})$ and $u_i \in U_i$ for $i \in I$, observe that at most one cycle from $\mathcal{C}$ was removed due to $\{u_i : i \in I\}$. Thus, the overall number of cycles removed is not larger than the right-hand side of (4.2). Since by our assumption (4.2) is false, there is a cycle $C = (u_1, \ldots, u_{2k+1}) \in \mathcal{C}$ which had not been removed by the end of the process. Then C satisfies our requirement. We fix such a $C = (u_1, \ldots, u_{2k+1})$ for the rest of the proof.

Let J be the set of all $1 \leq i \leq 2k+1$ such that there is $u_i' \in U_i \setminus \{u_i\}$ which is adjacent to u_{i-1} and u_{i+1}. We claim that J is a non-empty independent set (of the $(2k+1)$-cycle). To show that J is

an independent set, assume by contradiction that there is $1 \leq i \leq 2k+1$ such that $i, i+1 \in J$, and let $u'_i \in U_i \setminus \{u_i\}$ and $u'_{i+1} \in U_{i+1} \setminus \{u_{i+1}\}$ be witnesses to $i, i+1 \in J$. Then $u'_i, u_{i+1}, u_i, u'_{i+1}, u_{i+2}, \ldots, u_{i-1}, u'_i$ is a $(2k+3)$-cycle, a contradiction. We now show that $J \neq \emptyset$. Set $I' = \{2j : 2 \leq j \leq k\} \cup \{1\}$ and $I'' = \{2j : 3 \leq j \leq k\} \cup \{1, 3\}$ and note that they are both independent sets. By our choice of $C = (u_1, \ldots, u_{2k+1})$, there is $C' = (u'_1, \ldots, u'_{2k+1}) \in \mathcal{C} \setminus \{C\}$ which contains u_i for every $i \in I'$. Since $C' \neq C$, one of the following holds: either $u'_i \neq u_i$ for some $i \in \{2j+1 : 2 \leq j \leq k\}$, implying that $i \in J$ and we are done, or $(u'_2, u'_3) \neq (u_2, u_3)$. If $u'_2 = u_2$ or $u'_3 = u_3$ then $3 \in J$ or $2 \in J$, respectively, and again we are done. We deduce that $u'_2 \neq u_2$ and $u'_3 \neq u_3$. By repeating the same argument with respect to I'', we get a cycle $C'' = (u''_1, \ldots, u''_{2k+1}) \in \mathcal{C} \setminus \{C\}$ such that either $u''_i \neq u_i$ for some $i \in \{2j+1 : 3 \leq j \leq k\} \cup \{2\}$, implying that $i \in J$ and we are done, or $u''_4 \neq u_4$ and $u''_5 \neq u_5$ (here we use the assumption that $k \geq 2$ and hence $2k+1 \geq 5$). But now $u_1, u'_2, u'_3, u_4, u_3, u''_4, u''_5, u_6, \ldots, u_{2k+1}, u_1$ is a $(2k+3)$-cycle, a contradiction. See the top drawing in Figure 4.1 for an illustration.

We thus proved that J is a non-empty independent set. Apply Claim 4.4.2 to J to get $I \in \mathcal{I}(C_{2k+1})$ with the properties stated in the claim. By our choice of $C = (u_1, \ldots, u_{2k+1})$, there is $C' = (u'_1, \ldots, u'_{2k+1}) \in \mathcal{C} \setminus \{C\}$ which contains u_i for every $i \in I$. Let $i_1, \ldots, i_r$ be the elements of I in the order they appear when traversing the cycle $1, \ldots, 2k+1$. Since $C' \neq C$, there is $1 \leq j \leq r$ such that $(u'_{i_j+1}, \ldots, u'_{i_{j+1}-1}) \neq (u_{i_j+1}, \ldots, u_{i_{j+1}-1})$ [2]. Assume without loss of generality that $j = 1$ and $i_1 = 2$ (so in particular, $2 \in I$). By the guarantees of Claim 4.4.2, we have $i_2 - i_1 \equiv 2, 3 \pmod{2k+1}$, so either $i_2 = 4$ or $i_2 = 5$. Assume first that $i_2 = 4$. Then $u'_3 \neq u_3$, implying that $3 \in J$, which is impossible as $2 \in I$, $J \subseteq I$ and I is an independent set. Assume now that $i_2 = 5$. If $u'_3 = u_3$ then $u'_4 \neq u_4$ and so $4 \in J$, which is again impossible as $5 \in I$, $J \subseteq I$ and I is an independent set. So $u'_3 \neq u_3$ and similarly $u'_4 \neq u_4$. By the guarantees of Claim 4.4.2, we have that either $2 \in J$ or $5 \in J$, say without loss of generality that $2 \in J$. Then by the definition of J, there is $u''_2 \in U_2 \setminus \{u_2\}$ adjacent to u_1 and u_3. But now $u_1, u''_2, u_3, u_2, u'_3, u'_4, u_5, \ldots, u_{2k+1}, u_1$ is a $(2k+3)$-cycle, a contradiction. See the bottom drawing in Figure 4.1 for an illustration. This completes the proof. ■

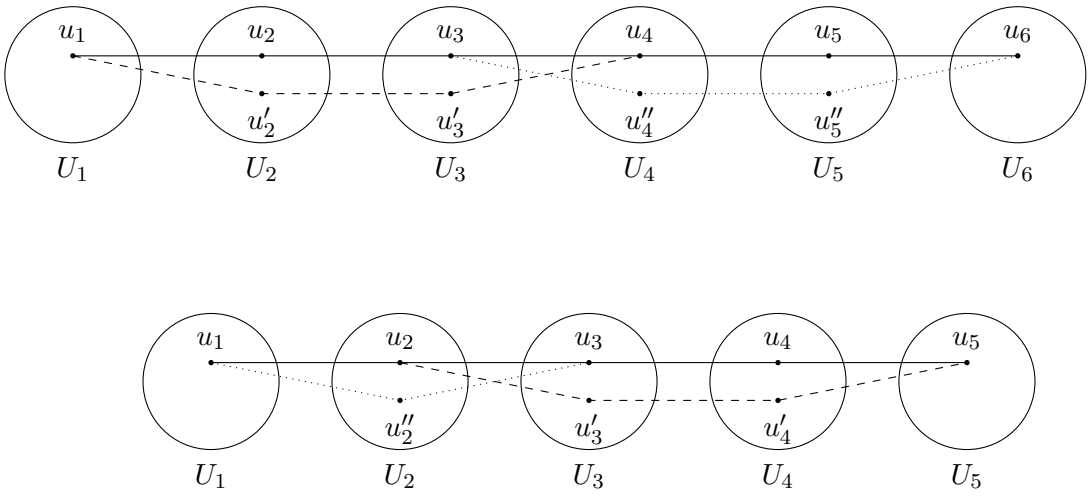

Figure 4.1: Illustrations for the proof of Lemma 4.4.3

[2] Here subscripts are taken modulo $2k+1$, while double subscripts are taken modulo r.

4.5 The Main Lemmas

In this section we establish the key lemmas to be used in the proofs of Theorems 12 and 4.1.2. We begin by proving the following lemma and some corollaries thereof. These give upper bounds on the number of paths (of a certain length) in certain types of graphs.

Lemma 4.5.1. *Let $s \geq 2$ and $\lambda \geq 1$, let G be an n-vertex graph and let $U_1, \ldots, U_s \subseteq V(G)$ be pairwise-disjoint sets such that $e(U_1, U_2) \leq \lambda(|U_1| + |U_2|)$ and $e(N_{U_{i+1}}(u_i), U_{i+2}) \leq \lambda(\left|N_{U_{i+1}}(u_i)\right| + |U_{i+2}|)$ for every $1 \leq i \leq s-2$ and $u_i \in U_i$. Then*

$$p(U_1, \ldots, U_s) \leq \begin{cases} \lambda^{(s-1)/2} n^{(s-3)/2} \left(|U_1||U_s| + \lambda n\right) & s \text{ is odd,} \\ \lambda^{s/2} n^{s/2-1} (|U_1| + |U_s|) & s \text{ is even.} \end{cases}$$

Proof. The proof is by induction on s. The base case $s = 2$ is given by our assumption that $e(U_1, U_2) \leq \lambda(|U_1| + |U_2|)$. Let then $s \geq 3$. Note that for every $u_1 \in U_1$, the sets $N_{U_2}(u_1), U_3, \ldots, U_s$ satisfy the assumptions of the lemma, so we may apply the induction hypothesis to them. Suppose first that s is odd. We have

$$\begin{aligned} p(U_1, \ldots, U_s) &= \sum_{u_1 \in U_1} p(N_{U_2}(u_1), U_3, \ldots, U_s) \leq \sum_{u_1 \in U_1} \lambda^{(s-1)/2} n^{(s-3)/2} (|N_{U_2}(u_1)| + |U_s|) \\ &= \lambda^{(s-1)/2} n^{(s-3)/2} \cdot (e(U_1, U_2) + |U_1||U_s|) \leq \lambda^{(s-1)/2} n^{(s-3)/2} \cdot (\lambda(|U_1| + |U_2|) + |U_1||U_s|) \\ &\leq \lambda^{(s-1)/2} n^{(s-3)/2} \cdot (|U_1||U_s| + \lambda n), \end{aligned}$$

where in the first inequality we used the induction hypothesis for $s-1$, and in the second inequality we used the assumption $e(U_1, U_2) \leq \lambda(|U_1| + |U_2|)$. The induction step for even s is similar. Indeed,

$$\begin{aligned} p(U_1, \ldots, U_s) &= \sum_{u_1 \in U_1} p(N_{U_2}(u_1), U_3, \ldots, U_s) \leq \sum_{u_1 \in U_1} \lambda^{(s-2)/2} n^{(s-4)/2} (|N_{U_2}(u_1)||U_s| + \lambda n) \\ &= \lambda^{(s-2)/2} n^{(s-4)/2} \cdot e(U_1, U_2) \cdot |U_s| + \lambda^{s/2} n^{s/2-1} \cdot |U_1| \\ &\leq \lambda^{(s-2)/2} n^{(s-4)/2} \cdot \lambda(|U_1| + |U_2|) \cdot |U_s| + \lambda^{s/2} n^{s/2-1} \cdot |U_1| \leq \lambda^{s/2} n^{s/2-1} \cdot (|U_1| + |U_s|), \end{aligned}$$

where in the first inequality we used the induction hypothesis for $s-1$, in the second inequality we used the assumption $e(U_1, U_2) \leq \lambda(|U_1| + |U_2|)$, and in the last inequality we used the trivial bound $|U_1| + |U_2| \leq n$. ■

We now derive two important corollaries of Lemma 4.5.1, stated as Lemmas 4.5.3 and 4.5.4. In their proof we will use the following well-known theorem of Erdős and Gallai [46].

Theorem 4.5.2 ([46])**.** *For every $t \geq 1$ we have $\mathrm{ex}(n, P_t) \leq \frac{t-1}{2} n$.*

Lemma 4.5.3. *Let $2 \leq s < t$ be integers having the same parity, let G be an n-vertex graph and let $U_1, \ldots, U_s \subseteq V(G)$ be pairwise-disjoint vertex-sets such that there is no path of length $t-1$ inside $U_1 \cup \cdots \cup U_s$ between a vertex in U_1 and a vertex in U_s. Then*[3]

$$p(U_1, \ldots, U_s) \leq \begin{cases} \left(\frac{t-s}{2}\right)^{(s-1)/2} n^{(s-3)/2} \left(|U_1||U_s| + \frac{t-s}{2} n\right) & s \text{ is odd,} \\ \left(\frac{t-s}{2}\right)^{s/2} n^{s/2-1} (|U_1| + |U_s|) & s \text{ is even.} \end{cases}$$

[3] We will only use the case that s is even.

Proof. We may and will assume that every edge in G is on some $(U_1, \ldots, U_s)$-path (as deleting all other edges does not change $p(U_1, \ldots, U_s)$). It is sufficient to show that the conditions of Lemma 4.5.1 hold for $\lambda = \frac{t-s}{2} \geq 1$. We prove the stronger statement that for every $1 \leq i \leq s-1$ and for every $U_i' \subseteq U_i$ and $U_{i+1}' \subseteq U_{i+1}$, it holds that $e(U_i', U_{i+1}') \leq \frac{t-s}{2} \left(|U_i'| + |U_{i+1}'|\right)$. If, by contradiction, this does not hold, then by Theorem 4.5.2 there is a path $P = v_1, \ldots, v_{t-s+2}$ of length $t-s+1$ in the bipartite graph (U_i', U_{i+1}'). Since $t-s+1$ is odd, we may assume without loss of generality that $v_1 \in U_i'$ and $v_{t-s+2} \in U_{i+1}'$. By our assumption, the edge (v_1, v_2) is on some $(U_1, \ldots, U_s)$-path, implying that there is a path $P' \subseteq U_1 \cup \cdots \cup U_i$ between[4] U_1 and v_1. Similarly, since the edge (v_{t-s+1}, v_{t-s+2}) is on some $(U_1, \ldots, U_s)$-path, there is a path $P'' \subseteq U_{i+1} \cup \cdots \cup U_s$ between v_{t-s+2} to U_s. Then $P'PP''$ is a path of length $t-1$ between U_1 and U_s, in contradiction to our assumption. ■

Lemma 4.5.4. *Let $s, \ell \geq 2$, let G be an n-vertex $C_{2\ell}$-free graph, let $\{u_0\}, U_1, \ldots, U_s \subseteq V(G)$ be pairwise-disjoint vertex-sets, and suppose that u_0 is adjacent to every vertex in U_1. Then*

$$p(U_1, \ldots, U_s) \leq \begin{cases} (\ell-1)^{(s-1)/2} n^{(s-3)/2} \left(|U_1||U_s| + (\ell-1)n\right) & s \text{ is odd,} \\ (\ell-1)^{s/2} n^{s/2-1} (|U_1| + |U_s|) & s \text{ is even.} \end{cases}$$

Proof. It is sufficient to show that the conditions of Lemma 4.5.1 hold with $\lambda = \ell - 1 \geq 1$. If $e(U_1, U_2) > (\ell-1)(|U_1| + |U_2|)$ then by Theorem 4.5.2, there is a path of length $2\ell-1$ in the bipartite graph (U_1, U_2). This path contains a subpath of length $2\ell-2$ with both endpoints in U_1, which closes a 2ℓ-cycle with u_0, in contradiction to the assumption of the lemma. Similarly, if $e(N_{U_{i+1}}(u_i), U_{i+2}) > (\ell-1)(\left|N_{U_{i+1}}(u_i)\right| + |U_{i+2}|)$ for some $1 \leq i \leq s-2$ and $u_i \in U_i$, then by Theorem 4.5.2 there is a path of length $2\ell-1$ in the bipartite graph with sides $N_{U_{i+1}}(u_i)$ and U_{i+2}. This path contains a subpath of length $2\ell-2$ with both endpoints in $N_{U_{i+1}}(u_i)$, which closes a 2ℓ-cycle with u_i, in contradiction to the assumption of the lemma. ■

The construction in Claim 4.3.1 shows that the bounds in the above two lemmas, as well as in Lemma 4.5.1, are tight (up to the multiplicative constants depending on the parameters λ, s, t, ℓ).

The rest of this section is devoted to proving the following key lemma, which will play a central role in the proofs of Theorems 12 and 4.1.2.

Lemma 4.5.5. *Let $\ell \geq 3$, let G be an n-vertex graph, let $X, Y, Z, W \subseteq V(G)$ be pairwise-disjoint vertex-sets and assume that the bipartite graphs (X, Y), (Y, Z) and (Z, W) are $C_{2\ell}$-free. Then there are subsets $Y' \subseteq Y$ and $Z' \subseteq Z$ such that*

1. *$e(Y', X), e(Y', Z), e(Z', Y), e(Z', W) = O(\ell n)$.*
2. *$p(X, Y \setminus Y', Z \setminus Z', W) = O(\ell^2 n^2)$.*

At the end of this section, we explain why the sets Y' and Z' in the statement of Lemma 4.5.5 are required, and why Lemma 4.5.5 is false for $\ell = 2$. The falsity of Lemma 4.5.5 for $\ell = 2$ is the reason we need a separate proof for the case $\mathrm{ex}(n, C_{2k+1}, C_{2k+3})$ (such a proof is given in Section 4.4).

For the proof of Lemma 4.5.5, we will need an upper bound on Zarankiewicz numbers for even cycles, proved by Naor and Verstraëte [87]. For integers $n, m \geq 1$ and $\ell \geq 2$, let $z(n, m, C_{2\ell})$ denote the maximal number of edges in a $C_{2\ell}$-free bipartite graph with sides of size n and m.

[4] It might be the case that $v_1 \in U_1$ (if $i = 1$), in which case P' has no edges.

Theorem 4.5.6 ([87]). *For $m \leq n$ it holds that*

$$z(n,m,C_{2\ell}) \leq \begin{cases} (2\ell-3)\left((nm)^{1/2+1/(2\ell)}+2n\right) & \ell \text{ is odd,} \\ (2\ell-3)\left(n^{1/2}m^{1/2+1/\ell}+2n\right) & \ell \text{ is even.} \end{cases}$$

The following lemma is an easy corollary of Theorem 4.5.6.

Lemma 4.5.7. *Let $\ell \geq 2$, let G be an n-vertex graph and let $X, Y \subseteq V(G)$ be disjoint sets such that the bipartite graph (X,Y) is $C_{2\ell}$-free. Let Y' be the set of all $y \in Y$ which have at least d neighbours in X. Then*

$$|Y'| \leq \begin{cases} \max\{(6\ell/d)^{2\ell/(\ell-1)}n^{(\ell+1)/(\ell-1)}, 6\ell n/d\} & \ell \text{ is odd,} \\ \max\{(6\ell/d)^{2\ell/(\ell-2)}n^{\ell/(\ell-2)}, 6\ell n/d\} & \ell \text{ is even and } \ell \geq 4, \\ 2n/(d-n^{1/2}) & \ell = 2 \text{ and } d > n^{1/2}. \end{cases}$$

Proof. Note that

$$d|Y'| \leq e(Y',X) \leq z(n,|Y'|,C_{2\ell}). \tag{4.3}$$

Suppose first that ℓ is odd. We apply Theorem 4.5.6 with parameter $m = |Y'|$. If $(|Y'|n)^{1/2+1/(2\ell)} \geq n$ then Theorem 4.5.6 gives $z(n,|Y'|,C_{2\ell}) \leq 6\ell(|Y'|n)^{1/2+1/(2\ell)}$, and if $(|Y'|n)^{1/2+1/(2\ell)} \leq n$ then Theorem 4.5.6 gives $z(n,|Y'|,C_{2\ell}) \leq 6\ell n$. By combining these inequalities with (4.3) we get that either $|Y'| \leq (6\ell/d)^{2\ell/(\ell-1)}n^{(\ell+1)/(\ell-1)}$ or $|Y'| \leq 6\ell n/d$, as required.

Suppose now that ℓ is even and $\ell \geq 4$. By Theorem 4.5.6, we have $z(n,|Y'|,C_{2\ell}) \leq 6\ell n^{1/2}|Y'|^{1/2+1/\ell}$ if $n^{1/2}|Y'|^{1/2+1/\ell} \geq n$ and $z(n,|Y'|,C_{2\ell}) \leq 6\ell n$ otherwise. By combining these inequalities with (4.3) we get that either $|Y'| \leq (6\ell/d)^{2\ell/(\ell-2)}n^{\ell/(\ell-2)}$ or $|Y'| \leq 6\ell n/d$, as required.

Finally, suppose that $\ell = 2$ and that $d > n^{1/2}$. Theorem 4.5.6 gives $z(n,|Y'|,C_4) \leq n^{1/2}|Y'| + 2n$. By combining this with (4.3) we get that $|Y'| \leq 2n/(d-n^{1/2})$, as required. ■

We are now ready to prove Lemma 4.5.5.

Proof of Lemma 4.5.5. We start by considering the case of even $\ell \geq 4$. Define the sets $Y' = \{y \in Y : |N_X(y)| \geq \ell n^{2/(\ell+2)}\}$ and $Z' = \{z \in Z : |N_W(z)| \geq \ell n^{2/(\ell+2)}\}$. Apply Lemma 4.5.7 with $d = \ell n^{2/(\ell+2)}$ to get $|Y'|, |Z'| \leq O(n^{\ell/(\ell+2)})$. By plugging these bounds into Theorem 4.5.6, one can check that $e(Y',X), e(Y',Z), e(Z',Y), e(Z',W) \leq z(n, O(n^{\ell/(\ell+2)}), C_{2\ell}) = O(\ell n)$. Next, note that by the definitions of Y' and Z' we have

$$\begin{aligned} p(X, Y \setminus Y', Z \setminus Z', W) &< e(Y \setminus Y', Z \setminus Z') \cdot \ell n^{2/(\ell+2)} \cdot \ell n^{2/(\ell+2)} \\ &\leq z(n,n,C_{2\ell}) \cdot \ell^2 n^{4/(\ell+2)} \leq O(\ell^3 n^{1+1/\ell+4/(\ell+2)}), \end{aligned}$$

where in the last inequality we used Theorem 4.5.6. So if $\ell^3 n^{1+1/\ell+4/(\ell+2)} \leq \ell^2 n^2$ then we get the required bound $p(X, Y \setminus Y', Z \setminus Z', W) = O(\ell^2 n^2)$, and the proof is complete (for even ℓ). Otherwise, we have $\ell^3 n^{1+1/\ell+4/(\ell+2)} > \ell^2 n^2$ and hence $n < \ell^{\ell(\ell+2)/(\ell^2-3\ell-2)} = \ell \cdot \ell^{(5\ell+2)/(\ell^2-3\ell-2)} \leq O(\ell)$. Since $p(X,Y,Z,W) \leq n^4$, we have $p(X,Y,Z,W) \leq n^4 = O(\ell^2 n^2)$, and again we are done.

We now consider the case of odd $\ell \geq 3$. We define $Y' = \{y \in Y : |N_X(y)| \geq \ell n^{2/(\ell+1)}\}$ and $Z' = \{z \in Z : |N_W(z)| \geq \ell n^{2/(\ell+1)}\}$. Similarly to the previous case, we apply Lemma 4.5.7 with $d = \ell n^{2/(\ell+1)}$ to obtain $|Y'|, |Z'| \leq O(n^{(\ell-1)/(\ell+1)})$. We then plug these bounds into Theorem 4.5.6

to get $e(Y',X), e(Y',Z), e(Z',Y), e(Z',W) \leq z(n, O(n^{(\ell-1)/(\ell+1)}), C_{2\ell}) = O(\ell n)$. It remains to bound $p(X, Y \setminus Y', Z \setminus Z', W)$. Assume first that $\ell \geq 5$. By the definitions of Y' and Z' we have

$$\begin{aligned} p(X, Y \setminus Y', Z \setminus Z', W) &< e(Y \setminus Y', Z \setminus Z') \cdot \ell n^{2/(\ell+1)} \cdot \ell n^{2/(\ell+1)} \\ &\leq z(n, n, C_{2\ell}) \cdot \ell^2 n^{4/(\ell+1)} \leq O(\ell^3 n^{1+1/\ell+4/(\ell+1)}), \end{aligned}$$

where in the last inequality we used Theorem 4.5.6. If $\ell^3 n^{1+1/\ell+4/(\ell+1)} \leq \ell^2 n^2$ then by the above we have $p(X, Y \setminus Y', Z \setminus Z', W) = O(\ell^2 n^2)$, as required. Otherwise, we have $\ell^3 n^{1+1/\ell+4/(\ell+1)} > \ell^2 n^2$ and hence $n < \ell^{\ell(\ell+1)/(\ell^2-4\ell-1)} = \ell \cdot \ell^{(5\ell+1)/(\ell^2-4\ell-1)} = O(\ell)$ (here we use the assumption $\ell \geq 5$). But then $p(X, Y, Z, W) \leq n^4 = O(\ell^2 n^2)$, and again we are done.

Thus, it remains to show that $p(X, Y \setminus Y', Z \setminus Z', W) = O(n^2)$ when $\ell = 3$. Recall that in this case we defined $Y' = \{y \in Y : |N_X(y)| \geq 3n^{1/2}\}$ and similarly $Z' = \{z \in Z : |N_W(z)| \geq 3n^{1/2}\}$. We need some additional definitions. Define $Y_{low} = \{y \in Y : |N_X(y)| < n^{1/3}\}$ and similarly $Z_{low} = \{z \in Z : |N_W(z)| < n^{1/3}\}$. Define $\mathcal{I} = \{i : \frac{1}{2}n^{1/3} \leq 2^i < 3n^{1/2}\}$, and for each $i \in \mathcal{I}$ set $Y_i = \{y \in Y : 2^i \leq |N_X(y)| < 2^{i+1}\}$ and $Z_i = \{z \in Z : 2^i \leq |N_W(z)| < 2^{i+1}\}$. It is immediate from these definitions that $Y \setminus Y' \subseteq Y_{low} \cup \bigcup_{i \in I} Y_i$ and similarly $Z \setminus Z' \subseteq Z_{low} \cup \bigcup_{i \in I} Z_i$. Note that

$$p(X, Y_{low}, Z_{low}, W) < e(Y_{low}, Z_{low}) \cdot n^{1/3} \cdot n^{1/3} \leq z(n, n, C_6) \cdot n^{2/3} \leq O(n^2),$$

where in the last inequality we used Theorem 4.5.6. Hence, in order to finish the proof we need to bound $p(X, \bigcup_{i \in I} Y_i, Z_{low}, W)$, $p(X, Y_{low}, \bigcup_{i \in I} Z_i, W)$ and $p(X, \bigcup_{i \in I} Y_i, \bigcup_{i \in I} Z_i, W)$. We start with the first two terms. Fix any $i \in \mathcal{I}$. By Lemma 4.5.7 with $d = 2^i$, we have $|Y_i| \leq \max\{18^3 \cdot 2^{-3i} \cdot n^2, 18 \cdot 2^{-i} \cdot n\} = O(2^{-3i} \cdot n^2)$, where we used the fact that $9 \cdot 2^{-3i} n^2 > 2^{-i} n$, which follows from $2^i < 3n^{1/2}$. So we get

$$e(Y_i, Z_{low}) \leq z(|Y_i|, n, C_6) \leq 3 \cdot \Big((|Y_i| n)^{2/3} + 2n \Big) \leq 3 \cdot \big(O(n^2 2^{-2i}) + 2n \big) \leq O(n^2 \cdot 2^{-2i}),$$

where in the second inequality we used Theorem 4.5.6, and in the last inequality we used $n^2 \cdot 2^{-2i} > n/9$ which follows from $2^i < 3n^{1/2}$. Now we have

$$\begin{aligned} p(X, \bigcup Y_i, Z_{low}, W) &= \sum_{i \in \mathcal{I}} p(X, Y_i, Z_{low}, W) < \sum_{i \in \mathcal{I}} e(Y_i, Z_{low}) \cdot 2^{i+1} \cdot n^{1/3} \leq \sum_{i \in \mathcal{I}} O(n^2 \cdot 2^{-2i}) \cdot 2^{i+1} \cdot n^{1/3} \\ &= O(n^{7/3}) \cdot \sum_{i \in \mathcal{I}} 2^{-i} \leq O(n^{7/3}) \cdot \sum_{i:\, 2^i \geq \frac{1}{2} n^{1/3}} 2^{-i} = O(n^{7/3}) \cdot O(n^{-1/3}) = O(n^2), \end{aligned}$$

where in the first inequality we used the definitions of Z_{low} and Y_i, and in the last inequality we used the definition of $\mathcal{I}$. The bound $p(X, Y_{low}, \bigcup Z_i, W) = O(n^2)$ is proved similarly.

Finally, we bound $p(X, \bigcup Y_i, \bigcup Z_i, W)$. To this end, fix any $i, j \in \mathcal{I}$. We showed above that $|Y_i| \leq O(n^2 \cdot 2^{-3i})$. By the same argument we get $|Z_j| \leq O(n^2 \cdot 2^{-3j})$. Thus we have

$$\begin{aligned} e(Y_i, Z_j) &\leq z(|Y_i|, |Z_j|, C_6) \leq 3 \cdot \Big((|Y_i||Z_j|)^{2/3} + |Y_i| + |Z_j| \Big) \\ &\leq O(n^{8/3}) \cdot 2^{-2i} \cdot 2^{-2j} + O(n^2) \cdot (2^{-3i} + 2^{-3j}) \leq O(n^{8/3}) \cdot 2^{-2i} \cdot 2^{-2j}, \end{aligned}$$

where in the second inequality we used Theorem 4.5.6, and in the last inequality we used the fact that $18 n^{8/3} \cdot 2^{-2i} \cdot 2^{-2j} \geq \max\{n^2 2^{-3i}, n^2 2^{-3j}\}$, which follows from $\frac{1}{2} n^{1/3} \leq 2^i, 2^j < 3n^{1/2}$. Now we get

$$p(X, \bigcup Y_i, \bigcup Z_i, W) = \sum_{i,j \in \mathcal{I}} p(X, Y_i, Z_j, W) \leq \sum_{i,j \in \mathcal{I}} e(Y_i, Z_j) \cdot 2^{i+1} \cdot 2^{j+1}$$

$$\leq \sum_{i,j\in\mathcal{I}} O(n^{8/3})\cdot 2^{-2i}\cdot 2^{-2j}\cdot 2^{i+1}\cdot 2^{j+1} = O(n^{8/3})\cdot \sum_{i,j\in\mathcal{I}} 2^{-i}\cdot 2^{-j}$$
$$\leq O(n^{8/3})\cdot \sum_{2^i,2^j\geq \frac{1}{2}n^{1/3}} 2^{-i}\cdot 2^{-j} = O(n^{8/3})\cdot O(n^{-2/3}) = O(n^2),$$

where in the first inequality we used the definitions of Y_i and Z_j, and in the last inequality we used the definition of $\mathcal{I}$. This completes the proof. ■

Let us explain why the sets Y' and Z' in Lemma 4.5.5 are required, (namely, that the statement $p(X,Y,Z,W) = O_\ell(n^2)$ is generally false). Note that by Theorem 4.5.6, the average degree between the four sets in Lemma 4.5.5 is $O(n^{1/3})$. One might thus guess that $p(X,Y,Z,W) = O(n\cdot(n^{1/3})^3) = O(n^2)$. To see that this is not the case, we can take Y to be a single vertex connected to all the vertices of X and Z, distribute all other vertices equally among X, Z and W, and take the bipartite graph between Z, W to be an extremal graph with no $C_{2\ell}$. Although this example satisfies $p(X,Y,Z,W) \gg n^2$, by removing the single vertex of Y we can make sure that $p(X,Y,Z,W) = O(n^2)$. This is precisely what Lemma 4.5.5 states. What we see in the proof of Theorem 4.1.2 is that if one assumes that the *entire* graph is $C_{2\ell}$-free (and not just the 3 bipartite graphs between the 4 sets) then one no longer needs to remove vertices in order to guarantee that $p(X,Y,Z,W) = O_\ell(n^2)$.

Let us note that Lemma 4.5.5 does not hold for $\ell = 2$. Indeed, in the proof of Lemma 4.3.4 we construct an n-vertex C_4-free graph, in which every vertex has degree $\Theta(n^{1/2})$ and lies on $\Theta(n^{3/2})$ paths of length 3. Taking a random vertex partition of this graph into four sets X, Y, Z, W, we see that with high probability, every vertex $y \in Y$ (resp. $z \in Z$) has $\Theta(n^{1/2})$ neighbors in X (resp. W), and every vertex in the graph lies on $\Theta(n^{3/2})$ (X,Y,Z,W)-paths. Suppose now, by contradiction, that the assertion of Lemma 4.5.5 holds for the sets X, Y, Z, W. Since every $y \in Y$ has $\Theta(n^{1/2})$ neighbors in X, and since $e(Y', X) = O(n)$, we must have $|Y'| = O(n^{1/2})$. Similarly, $|Z'| = O(n^{1/2})$. As every vertex lies on $\Theta(n^{3/2})$ (X,Y,Z,W)-paths, we have $p(X,Y,Z,W) = \Theta(n^{5/2})$ and $p(X,Y',Z,W), p(X,Y,Z',W) = O(n^2)$. But this implies that $p(X, Y\setminus Y', Z\setminus Z', W) = \Theta(n^{5/2})$, in contradiction to the statement of Lemma 4.5.5.

4.6 Proof of Theorem 4.1.2 and Proposition 4.1.1

Here we prove Theorem 4.1.2 and Proposition 4.1.1. For Theorem 4.1.2 we will need the following lemma.

Lemma 4.6.1. *Let $\ell \geq 2$ and let G be an n-vertex $C_{2\ell}$-free graph. Then every $v \in V(G)$ is the endpoint of less than $4(\ell-1)n$ paths of length 2.*

Proof. Let $v \in V(G)$ and assume, by contradiction, that v is the endpoint of $r \geq 4(\ell-1)n$ paths of length 2. Let $V(G)\setminus\{v\} = V_1 \cup V_2$ be a random partition, obtained by putting each $u \in V(G)\setminus\{v\}$ in one of the sets V_1, V_2 with probability $\frac{1}{2}$, independently. Since $\mathbb{E}[p(v,V_1,V_2)] = \frac{1}{4}r$, there is a choice of V_1, V_2 for which $e(N_{V_1}(v), V_2) = p(v,V_1,V_2) \geq \frac{1}{4}r \geq (\ell-1)n > (\ell-1)(|N_{V_1}(v)| + |V_2|)$. This stands in contradiction to Lemma 4.5.4, applied with $s = 2, u_0 = v, U_1 = N_{V_1}(v), U_2 = V_2$. ■

Next, let us recall the famous Even Cycle Theorem of Bondy and Simonovits:

Theorem 4.6.2 ([25])**.** *For every $\ell \geq 2$ we have $\mathrm{ex}(n, C_{2\ell}) \leq O(\ell n^{1+1/\ell})$.*

We note that the multiplicative constant in the above theorem was improved upon by Pikhurko [88] and Bukh and Jiang [28] (see also the survey [114]). We are now in a position to prove Theorem 4.1.2.

Proof of Theorem 4.1.2. The lower bounds were proved in Section 4.3: the lower bound for $\ell = 2$ is given by Lemma 4.3.4, and the lower bound for $\ell \geq 3$ is given by Corollary 4.3.3. Thus, it remains to prove the upper bounds. We prove both cases simultaneously by induction on k. The base cases are $k = 2, 3$. For $k = 2$, Lemma 4.6.1 implies that $\mathrm{ex}(n, P_2, C_{2\ell}) = O(\ell n^2)$, as required.

Suppose now that $k = 3$. We first handle the case $\ell \geq 3$. By Claim 4.4.1, it is enough to show that $p(X, Y, Z, W) \leq O(\ell^2 n^2)$ for every vertex-partition $X \cup Y \cup Z \cup W$ of an n-vertex $C_{2\ell}$-free graph. Let $Y' \subseteq Y$ and $Z' \subseteq Z$ be as in Lemma 4.5.5. In light of Item 2 in Lemma 4.5.5, it is enough to prove that $p(X, Y', Z, W) = O(\ell^2 n^2)$ and $p(X, Y, Z', W) = O(\ell^2 n^2)$. Fix any $y \in Y'$. By Lemma 4.6.1, we have $p(y, Z, W) = O(\ell n)$, and hence $p(X, y, Z, W) \leq O(\ell n) \cdot |N_X(y)|$. By summing over all $y \in Y'$ and using the guarantees of Lemma 4.5.5, we get

$$p(X, Y', Z, W) = \sum_{y \in Y'} p(X, y, Z, W) \leq O(\ell n) \cdot \sum_{y \in Y'} |N_X(y)| = O(\ell n) \cdot e(Y', X) \leq O(\ell^2 n^2).$$

The bound $p(X, Y, Z', W) = O(\ell^2 n^2)$ is proven similarly.

Now we handle the case $\ell = 2$. Let G be an n-vertex C_4-free graph. Observe that the number of paths of length 3 in a graph G is at most $\sum_{v \in V(G)} \#P_1(v) \cdot \#P_2(v)$, where $\#P_i(v)$ is the number of paths of length i having v as an endpoint (so $\#P_1(v)$ is just the degree of v). By combining Lemma 4.6.1 with Theorem 4.6.2, we get that $\sum_{v \in V(G)} \#P_1(v) \cdot \#P_2(v) \leq O(n) \cdot 2e(G) \leq O(n^{5/2})$, as required.

Let now $k \geq 4$. Let G be an n-vertex $C_{2\ell}$-free graph, and observe that the number of paths of length k in G is at most

$$\sum_{v \in V(G)} \#P_{k-2}(v) \cdot \#P_2(v) \leq O(\ell n) \sum_{v \in V(G)} \#P_{k-2}(v) \leq O(\ell n) \cdot \mathrm{ex}(n, P_{k-2}, C_{2\ell}),$$

where in the first inequality we used Lemma 4.6.1. Thus, $\mathrm{ex}(n, P_k, C_{2\ell}) \leq O(\ell n) \cdot \mathrm{ex}(n, P_{k-2}, C_{2\ell})$. It is now easy to see that the theorem follows by induction on k, with the base cases $k = 2, 3$. ■

Proof of Proposition 4.1.1. We start with the lower bound. For $\ell \geq 3$, this is the statement of Claim 4.3.6. For $\ell = 2$, we get it from Lemma 4.3.4 and the well-known fact that $\mathrm{ex}(n, C_4) = O(n^{3/2})$ (which follows from Theorem 4.6.2). For the upper bound, let G be an n-vertex $C_{2\ell}$-free graph, and observe that for every $v \in V(G)$, the neighborhood of v does not contain a path of length $2\ell - 2$; indeed, such a path would close a copy of $C_{2\ell}$ with v. By Theorem 4.5.2 we have $e(N(v)) \leq \frac{2\ell-3}{2} \cdot |N(v)|$. On the other hand, the number of triangles containing v is exactly $e(N(v))$, so the number of triangles in G is

$$\frac{1}{3} \sum_{v \in V(G)} e(N(v)) \leq \frac{2\ell - 3}{6} \sum_{v \in V(G)} |N(v)| = \frac{2\ell - 3}{3} \cdot e(G) \leq \frac{2\ell - 3}{3} \cdot \mathrm{ex}(n, C_{2\ell}),$$

thus completing the proof. ■

4.7 Proof of Theorem 12: All Other Cases

In this section we prove the upper bounds for all cases in Theorem 12, except for the case of two consecutive odd integers which was handled in Section 4.4. The lower bounds have already been proven in Section 4.3. We begin by deriving the following corollary of Lemmas 4.5.3 and 4.5.4.

Lemma 4.7.1. *Let $k, \ell \geq 2$, let G be an n-vertex graph and assume either that G is $C_{2\ell}$-free or that G is $C_{2\ell+1}$-free and $\ell > k$. Then for every partition $V(G) = V_1 \cup \cdots \cup V_{2k+1}$ we have*

$$c(V_1, \ldots, V_{2k+1}) \leq \ell^{k-1} n^{k-2} \cdot \left[p(V_1, V_2, V_3, V_4) + p(V_{2k+1}, V_1, V_2, V_3) \right].$$

Proof. Fix any (V_1, V_2, V_3)-path v_1, v_2, v_3. We claim that

$$p\left(N_{V_4}(v_3), V_5, \ldots, V_{2k}, N_{V_{2k+1}}(v_1)\right) \leq \ell^{k-1} n^{k-2} \cdot \left(|N_{V_4}(v_3)| + |N_{V_{2k+1}}(v_1)| \right). \tag{4.4}$$

Indeed, if G is $C_{2\ell}$-free then (4.4) follows from Lemma 4.5.4, applied with $s = 2k-2$, $u_0 = v_3$ and the sets $N_{V_4}(v_3), V_5, \ldots, V_{2k}, N_{V_{2k+1}}(v_1)$ as $U_1, \ldots, U_s$. If G is $C_{2\ell+1}$-free and $\ell > k$ then there is no path of length $2\ell - 3$ inside $V_4 \cup \cdots \cup V_{2k+1}$ between $N_{V_4}(v_3)$ and $N_{V_{2k+1}}(v_1)$, as such a path would close a $(2\ell+1)$-cycle with the path $v_1 v_2 v_3$. So (4.4) follows from Lemma 4.5.3 with parameters $s = 2k-2$ and $t = 2\ell - 2$. By summing (4.4) over all (V_1, V_2, V_3)-paths we get

$$\begin{aligned} c(V_1, \ldots, V_{2k+1}) &= \sum_{v_1, v_2, v_3} c(v_1, v_2, v_3, V_4, \ldots, V_{2k+1}) = \sum_{v_1, v_2, v_3} p(N_{V_4}(v_3), V_5, \ldots, V_{2k}, N_{V_{2k+1}}(v_1)) \\ &\leq \ell^{k-1} n^{k-2} \cdot \sum_{v_1, v_2, v_3} \left(|N_{V_4}(v_3)| + |N_{V_{2k+1}}(v_1)| \right) \\ &= \ell^{k-1} n^{k-2} \cdot \left[p(V_1, V_2, V_3, V_4) + p(V_{2k+1}, V_1, V_2, V_3) \right], \end{aligned}$$

thus completing the proof. ∎

The proof of (the upper bounds in) Theorem 12 is split into several parts: Lemma 4.7.2 handles the case that both cycle lengths are even; Lemma 4.7.3 handles the case where the forbidden cycle is even and the cycle whose number of copies is maximized is odd; finally, Lemma 4.7.4 handles the case where the cycle lengths are non-consecutive odd integers. For convenience, we rephrase each of the cases, denoting the cycle lengths by $2k$ or $2k+1$ and 2ℓ or $2\ell+1$ (rather than k and ℓ).

Lemma 4.7.2. *For every $k, \ell \geq 2$ we have $\mathrm{ex}(n, C_{2k}, C_{2\ell}) = O_k(\ell^k n^k)$.*

Proof. Let G be an n-vertex $C_{2\ell}$-free graph. By Claim 4.4.1, it is enough to prove that $c(V_1, \ldots, V_{2k}) = O(\ell^k n^k)$ for every partition $V(G) = V_1 \cup \cdots \cup V_{2k}$. Consider one such partition. Fixing $v_1 \in V_1$, we apply Lemma 4.5.4 with $s = 2k-1$, $u_0 = v_1$ and the sets $N_{V_2}(v_1), V_3, \ldots, V_{2k-1}, N_{V_{2k}}(v_1)$ as $U_1, \ldots, U_s$, to get

$$c(v_1, V_2, \ldots, V_{2k}) = p(N_{V_2}(v_1), V_3, \ldots, V_{2k-1}, N_{V_{2k}}(v_1)) \leq \ell^{k-1} n^{k-2} \cdot \left(|N_{V_2}(v_1)| \cdot |N_{V_{2k}}(v_1)| + \ell n \right).$$

By summing over all $v_1 \in V_1$, we get

$$\begin{aligned} c(V_1, \ldots, V_{2k}) &= \sum_{v_1 \in V_1} c(v_1, V_2, \ldots, V_{2k}) \leq \ell^{k-1} n^{k-2} \cdot \sum_{v_1 \in V_1} |N_{V_2}(v_1)| \cdot |N_{V_{2k}}(v_1)| + \ell^k n^{k-1} \cdot |V_1| \\ &= \ell^{k-1} n^{k-2} \cdot p(V_{2k}, V_1, V_2) + \ell^k n^{k-1} \cdot |V_1| = O(\ell^k n^k), \end{aligned}$$

where in the last inequality we used Theorem 4.1.2, which gives $p(V_{2k}, V_1, V_2) = O(\ell n^2)$. ∎

Lemma 4.7.3. *For every $k \geq 2$ we have*

$$\mathrm{ex}(n, C_{2k+1}, C_{2\ell}) \leq \begin{cases} O_k(n^{k+1/2}) & \ell = 2, \\ O_k(\ell^{k+1} n^k) & \ell \geq 3. \end{cases}$$

Proof. We start with the case $\ell \geq 3$. Let G be an n-vertex $C_{2\ell}$-free graph. By Claim 4.4.1, it is enough to prove that for every partition $V(G) = V_1 \cup \cdots \cup V_{2k+1}$ we have $c(V_1, \ldots, V_{2k+1}) = O(\ell^{k+1} n^k)$. By Theorem 4.1.2 we have $p(V_{2k+1}, V_1, V_2, V_3), p(V_1, V_2, V_3, V_4) \leq O(\ell^2 n^2)$. Plugging these estimates into Lemma 4.7.1 gives $c(V_1, \ldots, V_{2k+1}) = O(\ell^{k+1} n^k)$, as required.

The proof for the case $\ell = 2$ is similar. As in the previous case, we consider a partition $V(G) = V_1 \cup \cdots \cup V_{2k+1}$ of an n-vertex C_4-free graph. The only difference is that for $\ell = 2$, Theorem 4.1.2 gives $p(V_{2k+1}, V_1, V_2, V_3), p(V_1, V_2, V_3, V_4) = O(n^{5/2})$. Plugging this into Lemma 4.7.1 gives the required bound $c(V_1, \ldots, V_{2k+1}) = O_k(n^{k+1/2})$. ■

Lemma 4.7.4. *For every $2 \leq k < \ell - 1$ we have* $\mathrm{ex}(n, C_{2k+1}, C_{2\ell+1}) = O((2k+1)^{2k} \ell^{k+1} n^k)$.

Proof. Let G be an n-vertex $C_{2\ell+1}$-free graph. In light of Claim 4.4.1, we need to prove only that $c(V_1, \ldots, V_{2k+1}) \leq O(\ell^{k+1} n^k)$ for every partition $V(G) = V_1 \cup \cdots \cup V_{2k+1}$. Fix one such partition. We may and will assume that for every $1 \leq i \leq 2k+1$, every edge in $E(V_i, V_{i+1})$ is on some $(V_1, \ldots, V_{2k+1})$-cycle. We claim that the bipartite graph (V_i, V_{i+1}) is $C_{2\ell-2k+2}$-free for every $1 \leq i \leq 2k+1$ (with indices taken modulo $2k+1$). Assume by contradiction that there is a $(2\ell - 2k + 2)$-cycle C in the bipartite graph (V_i, V_{i+1}), and let $e \in E(V_i, V_{i+1})$ be an arbitrary edge of C. By our assumption, there is a $(V_1, \ldots, V_{2k+1})$-cycle C' containing e. But now $C \cup C' \setminus \{e\}$ is a $(2\ell+1)$-cycle, a contradiction.

In light of the above, we may apply Lemma 4.5.5 to $(V_{2k+1}, V_1, V_2, V_3)$ with $\ell - k + 1 \geq 3$ in place of ℓ and thus obtain subsets $V_1' \subseteq V_1$, $V_2' \subseteq V_2$ satisfying $e(V_1', V_{2k+1}), e(V_1', V_2), e(V_2', V_1), e(V_2', V_3) = O(\ell n)$ and $p(V_{2k+1}, V_1 \setminus V_1', V_2 \setminus V_2', V_3) = O(\ell^2 n^2)$. Similarly, applying Lemma 4.5.5 to V_1, V_2, V_3, V_4 gives subsets $V_2'' \subseteq V_2$ and $V_3'' \subseteq V_3$ such that $e(V_2'', V_1), e(V_2'', V_3), e(V_3'', V_2), e(V_3'', V_4) = O(\ell n)$ and $p(V_1, V_2 \setminus V_2'', V_3 \setminus V_3'', V_4) = O(\ell^2 n^2)$. Setting $W_1 = V_1 \setminus V_1'$, $W_2 = V_2 \setminus (V_2' \cup V_2'')$ and $W_3 = V_3 \setminus V_3''$, we see that

$$\begin{aligned} c(V_1, \ldots, V_{2k+1}) \leq\ & c(W_1, W_2, W_3, V_4, \ldots, V_{2k+1}) + c(V_1', V_2, \ldots, V_{2k+1}) + \\ & c(V_1, V_2', V_3, \ldots, V_{2k+1}) + c(V_1, V_2'', V_3, \ldots, V_{2k+1}) + c(V_1, V_2, V_3'', V_4, \ldots, V_{2k+1}). \end{aligned} \tag{4.5}$$

By our choice of V_1', V_2', V_2'', V_3'' via Lemma 4.5.5 and by the definition of the sets W_1, W_2, W_3, we have $p(V_{2k+1}, W_1, W_2, W_3) = O(\ell^2 n^2)$ and $p(W_1, W_2, W_3, V_4) = O(\ell^2 n^2)$. Plugging these bounds into Lemma 4.7.1 gives

$$c(W_1, W_2, W_3, V_4, \ldots, V_{2k+1}) \leq \ell^{k-1} n^{k-2} \cdot O(\ell^2 n^2) \leq O(\ell^{k+1} n^k).$$

It remains to bound the other four terms in (4.5). Consider the first term. Fixing any $v_1 \in V_1'$, note that there is no path of length $2\ell - 1$ in $V_2 \cup \cdots \cup V_{2k+1}$ between $N_{V_2}(v_1)$ and $N_{V_{2k+1}}(v_1)$, as such a path would close a $(2\ell+1)$-cycle with v_1. Thus, we may apply Lemma 4.5.3 with $s = 2k$, $t = 2\ell$ and $N_{V_2}(v_1), V_3, \ldots, V_{2k}, N_{V_{2k+1}}(v_1)$ as $U_1, \ldots, U_s$, to get

$$c(v_1, V_2, \ldots, V_{2k+1}) = p(N_{V_2}(v_1), V_3, \ldots, V_{2k}, N_{V_{2k+1}}(v_1)) \leq \ell^k n^{k-1} \left(|N_{V_2}(v_1)| + |N_{V_{2k+1}}(v_1)| \right).$$

By summing this over all $v_1 \in V_1'$ we obtain

$$c(V_1', V_2, \ldots, V_{2k+1}) = \sum_{v_1 \in V_1'} c(v_1, V_2, \ldots, V_{2k+1}) \leq \ell^k n^{k-1} \cdot \sum_{v_1 \in V_1'} \left(|N_{V_2}(v_1)| + |N_{V_{2k+1}}(v_1)| \right)$$

$$= \ell^k n^{k-1} \cdot \big(e(V_1', V_2) + e(V_1', V_{2k+1})\big) \leq O(\ell^{k+1} n^k),$$

where the last equality relies on the guarantees of Lemma 4.5.5. The remaining three terms in (4.5) are shown to be $O(\ell^{k+1} n^k)$ in the same manner. This completes the proof. ■

Having proven Theorem 12, we summarize our upper bounds on $\mathrm{ex}(n, C_{2k+1}, C_{2\ell+1})$ in Lemma 4.7.5 below. This lemma will then be used in Section 4.8.

Lemma 4.7.5. *There is an absolute constant c such that for every $1 \leq k < \ell$ we have the following.*

$$\mathrm{ex}(n, C_{2k+1}, C_{2\ell+1}) \leq \begin{cases} c\ell^2 n^{1+1/\ell} & k = 1, \\ c(2k+1)^{2k}(2\ell+1)^{k+1} n^k & k \geq 2. \end{cases}$$

Proof. The case $k = 1$ follows immediately by combining (4.1) with Theorem 4.6.2. As for the case $k \geq 2$, recall that by Lemma 4.7.4 we have $\mathrm{ex}(n, C_{2k+1}, C_{2\ell+1}) \leq O((2k+1)^{2k}(2\ell+1)^{k+1} n^k)$ for every $2 \leq k < \ell - 1$. In light of Lemma 4.4.3, this bound holds for $\ell = k+1$ as well (as $2\ell + 1 = 2k + 3 > 4$). ■

4.8 Proof of Theorems 10, 11 and 4.2.1

Recall that for a set L of positive integers, we denote by $\mathcal{P}_L$ the property of being *L-free*, that is, of being C_ℓ-free for every $\ell \in L$. We also put $w_L(\varepsilon) := w_{\mathcal{P}_L}(\varepsilon)$. In this section we will assume L to be an *infinite* set consisting only of *odd* integers. In what follows, $c, c', c'', c_1, c_2, \ldots$ are absolute constants which are implicitly assumed to be large enough.

The following theorem is a special case of the main result of Alon, de la Vega, Kannan and Karpinski [4]. For a graph G, denote by $\mathrm{maxcut}(G)$ the largest size of a cut in G.

Theorem 4.8.1 ([4])**.** *For every $\varepsilon \in (0,1)$, for every n-vertex graph G and for every $q \geq c\varepsilon^{-4}\log(1/\varepsilon)$, a uniformly chosen set $Q \in \binom{V(G)}{q}$ satisfies $\left|\frac{\mathrm{maxcut}(G)}{n^2} - \frac{\mathrm{maxcut}(G[Q])}{q^2}\right| < \varepsilon$ with probability at least $\frac{5}{6}$.*

We now use Theorem 4.8.1 to derive the following lemma.

Lemma 4.8.2. *For every $\varepsilon \in (0,1)$ and for every graph G which is ε-far from bipartiteness, it holds that with probability at least $\frac{2}{3}$, a random induced subgraph of G of order $c\varepsilon^{-5}$ is $\frac{\varepsilon}{2}$-far from bipartiteness.*

Proof. Let G be a graph which is ε-far from bipartiteness. Then clearly

$$\mathrm{maxcut}(G) \leq e(G) - \varepsilon n^2 = \left(\frac{e(G)}{n^2} - \varepsilon\right) n^2.$$

Set $q = c\varepsilon^{-5}$ and let $Q \in \binom{V(G)}{q}$ be chosen uniformly at random. Then with probability at least $\frac{5}{6}$ we have $\mathrm{maxcut}(G[Q]) \leq (\frac{e(G)}{n^2} - \frac{3\varepsilon}{4})q^2$, where we applied Theorem 4.8.1 with $\frac{\varepsilon}{4}$ in place of ε. By a standard second-moment-method argument one can easily show that a randomly chosen induced subgraph of order at least $c\varepsilon^{-2}$ has the same edge density as G, up to an additive error of ε. Thus, (by applying this argument with $\varepsilon/4$ in place of ε), the inequality

$$\left|\frac{e(G)}{n^2} - \frac{e(G[Q])}{q^2}\right| < \frac{\varepsilon}{4}$$

holds with probability at least $\frac{5}{6}$. Thus, with probability at least $\frac{2}{3}$ we have

$$\mathrm{maxcut}(G[Q]) \leq \left(\frac{e(G)}{n^2} - \frac{3\varepsilon}{4}\right) q^2 \leq e(G[Q]) - \frac{\varepsilon}{2} q^2,$$

which implies that $G[Q]$ is $\frac{\varepsilon}{2}$-far from bipartiteness. This completes the proof. ■

We note that Lemma 4.8.2 can also be derived from [59, Corollary 7.1.3]. Next, let us recall the following theorem of Komlós [77].

Theorem 4.8.3 ([77])**.** *For every $\varepsilon \in (0,1)$, every graph which is ε-far from bipartiteness contains an odd cycle of length at most $c\varepsilon^{-1/2}$.*

We now combine Lemma 4.8.2 with Theorem 4.8.3 to prove the following.

Lemma 4.8.4. *Let $\varepsilon \in (0,1)$, suppose that $n \geq q \geq c_1\varepsilon^{-11}$ and let G be an n-vertex graph. If G is ε-far from being bipartite then there is an odd $3 \leq s \leq c_1\varepsilon^{-1/2}$ such that with probability at least $\frac{2}{3}$, a random induced subgraph of G of order q contains at least $\left(\varepsilon^6 q/c_1\right)^s$ copies of C_s.*

Proof. By Theorem 4.8.2, a uniformly chosen $P \in \binom{V(G)}{c\varepsilon^{-5}}$ induces a graph which is $\frac{\varepsilon}{2}$-far from bipartiteness with probability at least $2/3$. By Theorem 4.8.3, such an induced subgraph contains an odd cycle of length at most $c(\varepsilon/2)^{-1/2}$. Thus, there is $3 \leq s \leq c(\varepsilon/2)^{-1/2}$ such that a random P as above contains an s-cycle with probability at least $\varepsilon^{1/2}/c'$. Set $d = 4c'\varepsilon^{-1/2}$ and let $P_1, \ldots, P_d \in \binom{V(G)}{c\varepsilon^{-5}}$ be chosen uniformly at random and independently. Setting $R = P_1 \cup \cdots \cup P_d$, we see that $G[R]$ contains an s-cycle with probability at least $1 - \left(1 - \varepsilon^{1/2}/c'\right)^{4c'\varepsilon^{-1/2}} \geq 1 - e^{-4} \geq 11/12$. Moreover, the probability that there are $1 \leq i < j \leq d$ for which $P_i \cap P_j \neq \emptyset$ is at most $\binom{d}{2} \cdot n \cdot (c\varepsilon^{-5}/n)^2 \leq c''\varepsilon^{-11}/n \leq \frac{1}{2}$, where in the last inequality we used the assumption that $n \geq c_1\varepsilon^{-11}$. Thus, setting $r = d \cdot c\varepsilon^{-5} = 4cc'\varepsilon^{-11/2}$, we see that $\mathbb{P}[|R| = r] \geq \frac{1}{2}$. Since $G[R]$ contains an s-cycle with probability at least $\frac{11}{12}$, we infer that at least a $\frac{5}{6}$-fraction of all sets $R' \in \binom{V(G)}{r}$ are such that $G[R']$ contains an s-cycle. Let $\mathcal{R}$ be the set of all $R' \in \binom{V(G)}{r}$ having this property, and note that $|\mathcal{R}| \geq \frac{5}{6}\binom{n}{r}$.

Fix any $q \geq r$. For $Q \in \binom{V(G)}{q}$, define the random variable $Z(Q) = |\binom{Q}{r} \cap \mathcal{R}|$ (namely, $Z(Q)$ is the number of sets in $\mathcal{R}$ which are contained in Q), and let $\mathcal{Q} = \left\{Q \in \binom{V(G)}{q} : Z(Q) \geq \frac{1}{2}\binom{q}{r}\right\}$. By linearity of expectation, we have $\mathbb{E}[Z] = |\mathcal{R}| \cdot \binom{q}{r}/\binom{n}{r} \geq \frac{5}{6}\binom{q}{r}$. Since $0 \leq Z \leq \binom{q}{r}$, it is now easy to deduce (by averaging) that $\mathbb{P}[Z \geq \frac{1}{2}\binom{q}{r}] \geq \frac{2}{3}$, implying that $|\mathcal{Q}| \geq \frac{2}{3}\binom{n}{q}$.

Now let $Q \in \mathcal{Q}$. By the definition of $\mathcal{Q}$, there are at least $\frac{1}{2}\binom{q}{r}$ r-sets $R \subseteq Q$ such that $G[R]$ contains a copy of C_s. On the other hand, a copy of C_s in $G[Q]$ is contained in exactly $\binom{q-s}{r-s}$ such r-sets. Thus, $G[Q]$ contains at least

$$\frac{\binom{q}{r}}{2\binom{q-s}{r-s}} = \frac{\binom{q}{s}}{2\binom{r}{s}} \geq \frac{1}{2}\left(\frac{q}{er}\right)^s \geq \left(\varepsilon^6 q/c_1\right)^s.$$

copies of C_s. This completes the proof. ■

Lemma 4.8.5, stated below, is the main lemma in this section. Its proof uses Lemmas[5] 4.7.5, 4.8.4 and 2.11.2.

[5]It is worth noting that at the cost of getting a worse upper bound in Theorem 10, one can avoid needing the full strength of Lemma 4.7.5 (which itself is essentially a restatement of Theorem 12). Indeed, weaker forms of Lemma 4.7.5 (and hence of Theorem 12) are sufficient for deriving Lemma 4.8.5, albeit with weaker bounds. This can then be used to prove (quantitatively) weaker forms of Theorem 10. However, in order to obtain the upper bound stated in Theorem 10, one needs the tight asymptotics (in the parameters n and ℓ) of Theorem 12.

Lemma 4.8.5. *There is a constant $c_2 \geq c_1$ (where c_1 is from Lemma 4.8.4) such that the following holds. Let $(\ell_i)_{i\geq 1}$ be an infinite increasing sequence of odd integers with $\ell_1 \geq 3$, and set $L = \{\ell_i : i \geq 1\}$. Then the following holds.*

1. *Let $\varepsilon \in (0,1)$ be small enough so that $c_1\varepsilon^{-1/2} \geq \ell_1$. Let ℓ_i be the maximal element of L not larger than $c_1\varepsilon^{-1/2}$, let $n \geq q \geq c_2\varepsilon^{-13} \cdot \ell_1^2 \cdot \ell_{i+1}$, and let G be an n-vertex graph which is ε-far from being bipartite. Then with probability at least $\frac{2}{3}$, a random induced subgraph of G of order q is not L-free. Thus, $w_L(\varepsilon) \leq c_2\varepsilon^{-13} \cdot \ell_1^2 \cdot \ell_{i+1}$.*
2. *For every $i \geq 1$ we have $w_L(\frac{1}{2(\ell_i+2)^2}) \geq \ell_{i+1}$.*

Proof. We start by proving the first assertion of Item 1. Let G be an n-vertex graph which is ε-far from bipartiteness. By Lemma 4.8.4, there is an odd $3 \leq s \leq c_1\varepsilon^{-1/2}$ such that for a randomly chosen $Q \in \binom{V(G)}{q}$, the graph $G[Q]$ contains at least $(\varepsilon^6 q/c_1)^s$ copies of C_s with probability at least $\frac{2}{3}$. We claim that if $G[Q]$ has this property then $G[Q]$ is not L-free. This will show that a random induced subgraph of G of order q is not L-free with probability at least $\frac{2}{3}$. This will also prove the upper bound on $w_L(\varepsilon)$ stated in Item 1, since every graph which is ε-far from being L-free is also ε-far from bipartiteness (as L contains only odd integers).

Assume first that $s = 3$. If $\ell_1 = 3$ then $G[Q]$ is clearly not L-free, as it contains at least one triangle. So we may assume that $\ell_1 = 2\ell+1 > 3$. It is easy to see that for c_2 large enough, our choice of q guarantees that

$$(\varepsilon^6 q/c_1)^3 > c\ell_1^2 q^{3/2} > c\ell^2 q^{1+1/\ell} \geq \mathrm{ex}(q, C_3, C_{2\ell+1}),$$

where in the last inequality we use Lemma 4.7.5. This means that $G[Q]$ contains more triangles than $\mathrm{ex}(q, C_3, C_{2\ell+1})$. So $G[Q]$ contains a cycle of length $\ell_1 = 2\ell + 1$ and hence is not L-free.

Assume from now on that $s > 3$. Observe that for a large enough c_2 we have

$$(\varepsilon^6 q/c_1)^s > c \cdot (c_1\varepsilon^{-1/2})^s \cdot \ell_{i+1}^{s/2} \cdot q^{s/2} \geq cs^s \ell_{i+1}^{s/2} q^{s/2} \geq cs^s \ell_{i+1}^{(s+1)/2} q^{(s-1)/2} \geq \mathrm{ex}(q, C_s, C_{\ell_{i+1}}),$$

where in the first and third inequalities we use our choice of q, in the second inequality we use $s \leq c_1\varepsilon^{-1/2}$ and in the last inequality we use Lemma 4.7.5 with $2k + 1 = s$ and $2\ell + 1 = \ell_{i+1}$, noting that $s < \ell_{i+1}$ by our choice of ℓ_i and by $s \leq c_1\varepsilon^{-1/2}$. As $G[Q]$ contains more s-cycles than $\mathrm{ex}(q, C_s, C_{\ell_{i+1}})$, it must contain a cycle of length ℓ_{i+1}. Thus, $G[Q]$ is not L-free.

We now prove the second Item. Fixing $i \geq 1$, let n be large enough so that Lemma 2.11.2 is applicable to $k = \ell_i + 2$ and $f = \ell_{i+1}$, and let G be the $\frac{n}{\ell_i+2}$-blowup of C_{ℓ_i+2}. Note that $C_{\ell_{i+1}}$ has a homomorphism into C_{ℓ_i+2}, as $\ell_{i+1} \geq \ell_i + 2$. Thus, by applying Lemma 2.11.2 with $K = C_{\ell_i+2}$ and $F = C_{\ell_{i+1}}$, we conclude that G is $\frac{1}{2(\ell_i+2)^2}$-far from being $C_{\ell_{i+1}}$-free and hence also $\frac{1}{2(\ell_i+2)^2}$-far from being L-free. On the other hand, there is no homomorphism from C_k to C_{ℓ_i+2} for any odd $k \leq \ell_i$. Thus, every subgraph of G on less than ℓ_{i+1} vertices is L-free. Item 2 of the lemma follows. ■

The proofs of Theorems 10 and 4.2.1 now follow quite easily from the above lemma.

Proof of Theorem 10. Set $\ell_1 = 3$ and $\ell_{i+1} = 2f\big(\frac{1}{2(\ell_i+2)^2}\big) + 1$. Then ℓ_i is odd for every $i \geq 1$, and $(\ell_i)_{i\geq 1}$ is increasing as f satisfies $f(x) \geq 1/x$. Setting $L = \{\ell_i : i \geq 1\}$, we will show that the property of being L-free satisfies the assertion of the theorem. More precisely, we will show that there is an absolute constant $\varepsilon_0 > 0$ such that $w_L(\varepsilon) \leq \varepsilon^{-14} f(\varepsilon/c)$ for every $\varepsilon < \varepsilon_0$, and that $w_L(\varepsilon) \geq f(\varepsilon)$ for an infinite

sequence of values of ε which tends to 0. Let $\varepsilon \in (0,1)$ be small enough so that $c_1\varepsilon^{-1/2} \geq 3 = \ell_1$, and let ℓ_i be the maximal element of L not larger than $c_1\varepsilon^{-1/2}$. Item 1 of Lemma 4.8.5 implies that

$$w_L(\varepsilon) \leq c_2\varepsilon^{-13} \cdot \ell_1^2 \cdot \ell_{i+1} = 9c_2\varepsilon^{-13} \cdot \ell_{i+1} \leq 27c_2\varepsilon^{-13} \cdot f\left(\frac{1}{2(\ell_i+2)^2}\right) \leq \varepsilon^{-14} \cdot f(\varepsilon/c),$$

where in the last inequality we used that $\ell_i \leq c_1\varepsilon^{-1/2}$, that f is decreasing, and that $1/\varepsilon > 27c_2$ (which can be guaranteed by appropriately choosing ε_0). The second part of Lemma 4.8.5 implies that for every $i \geq 1$, $w_L\left(\frac{1}{2(\ell_i+2)^2}\right) \geq \ell_{i+1} > f\left(\frac{1}{2(\ell_i+2)^2}\right)$. So there is a decreasing sequence $(\varepsilon_i)_{i\geq 1}$ with $\varepsilon_i \to 0$ (namely $\varepsilon_i = \frac{1}{2(\ell_i+2)^2}$) such that $w_L(\varepsilon_i) \geq f(\varepsilon_i)$. The theorem follows. ■

Proof of Theorem 4.2.1. The first part of Lemma 4.8.5 implies that for a sufficiently small ε we have $w_L(\varepsilon) \leq \text{poly}(1/\varepsilon) \cdot \ell_{i+1}$, where ℓ_i is the maximal element of L not larger than $c_1\varepsilon^{-1/2}$. Thus, if $\ell_{i+1} \leq \ell_i^d$ for some $d = d(L)$ and every sufficiently large i, then $w_L(\varepsilon) \leq \text{poly}(1/\varepsilon)$ for every sufficiently small ε. On the other hand, the second part of Lemma 4.8.5 implies that unless $\ell_{i+1} \leq \ell_i^d$ for some $d = d(L)$ and for every large enough i, the function $w_L(\varepsilon)$ is super-polynomial in $1/\varepsilon$ for infinitely many values of ε. We conclude that $w_L(\varepsilon) = \text{poly}(1/\varepsilon)$ if and only if $\ell_{i+1} \leq \ell_i^d$ for every large enough i, which is equivalent to having $\limsup_{j\longrightarrow\infty} \frac{\log \ell_{j+1}}{\log \ell_j} \leq d < \infty$. ■

Finally, we move on to the proof of Theorem 11. Here we will need the following lemma, which states that graphs that avoid an infinite family of odd cycles must be close to being bipartite.

Lemma 4.8.6. *Let $(\ell_i)_{i\geq 1}$ be an infinite increasing sequence of odd integers with $\ell_1 \geq 3$, and set $L = \{\ell_i : i \geq 1\}$. Then every L-free graph is $o(1)$-close to bipartiteness.*

Proof. Our goal is to show that for every sufficiently small ε there is $n_0(\varepsilon)$ such that every L-free graph on $n \geq n_0(\varepsilon)$ vertices is ε-close to being bipartite. So fix $\varepsilon > 0$ small enough to satisfy $c_1\varepsilon^{-1/2} \geq \ell_1$, and let ℓ_i be the maximal element of L not larger than $c_1\varepsilon^{-1/2}$. By (the contrapositive of) Item 1 in Lemma 4.8.5, every n-vertex L-free graph is ε-close to bipartiteness, provided that n is large enough to satisfy $n \geq c_2\varepsilon^{-13} \cdot \ell_1^2 \cdot \ell_{i+1}$. This completes the proof. ■

The quantitative version of Lemma 4.8.6 states that L-free n-vertex graphs are roughly $\Theta(\ell_i^{-2})$-close to bipartiteness, where i is the maximal integer satisfying $n \geq \text{poly}(\ell_i) \cdot \ell_{i+1} \sim \ell_{i+1}$ (here we assume that the sequence $(\ell_i)_{i\geq 1}$ grows fast enough). Let us explain why this dependence on the sequence $(\ell_i)_{i\geq 1}$ is unavoidable. For $n = \ell_{i+1}$, let G be the $\frac{n-1}{\ell_i+2}$-blowup of C_{ℓ_i+2}, plus an isolated vertex. Then G is L-free; it contains neither an odd cycle of length at most ℓ_i (as such a cycle is not homomorphic to C_{ℓ_i+2}), nor an odd cycle of length at least ℓ_{i+1} (as $\ell_{i+1} > n-1$ and G has an isolated vertex). Nonetheless, it is easy to see that G is $\Theta(\ell_i^{-2})$-far from bipartiteness. This shows that the $o(1)$-term in Lemma 4.8.6 may tend to zero arbitrarily slowly, depending on the family L. For example, if $\ell_i = \text{tower}(i)$ then $\ell_i = \log_2(\ell_{i+1})$, so every L-free n-vertex graph is roughly $\Theta(\frac{1}{\log^2 n})$-close to bipartiteness, and this is tight.

Proof of Theorem 11. By (the proof of) Theorem 10, there is an increasing sequence of odd integers $L = \{\ell_1 = 3, \ell_2, \ell_3, \dots\}$ such that $w_L(\varepsilon) \geq f(\varepsilon)$. Thus, it remains to present a 2-sided tester for L-freeness which has query complexity $\text{poly}(1/\varepsilon)$. Our ε-tester works as follows: it samples a random induced subgraph of the input of order $q = q(\varepsilon) = c\varepsilon^{-5}$ and accepts if and only if this subgraph is $\frac{\varepsilon}{2}$-close to bipartiteness. Let us prove that this algorithm is indeed a valid ε-tester for graphs of order $n \geq n_0(\varepsilon)$, where $n_0(\varepsilon)$ will be

(implicitly) chosen later. Let G be an n-vertex input graph. If G is ε-far from L-freeness then it is also ε-far from bipartiteness, so Lemma 4.8.2 implies that with probability at least $\frac{2}{3}$, G is rejected. Assume now that G is L-free. By Lemma 4.8.6, if n is large enough then G is $\frac{\varepsilon}{12}$-close to bipartiteness. Hence, there is a set $E \subseteq E(G)$ of size $|E| \leq \frac{\varepsilon}{12}n^2$ such that $G \setminus E$ (the graph obtained from G by deleting the edges in E) is bipartite. Let $Q = \{x_1, \ldots, x_q\}$ denote the vertex-set sampled by the tester. The expected number of pairs $1 \leq i < j \leq q$ for which $\{x_i, x_j\} \in E$ is $\binom{q}{2} \cdot \frac{2|E|}{n(n-1)} \leq \frac{\varepsilon}{6}q^2$. By Markov's inequality, $|E(G[Q]) \cap E| \leq \frac{\varepsilon}{2}q^2$ holds with probability at least $\frac{2}{3}$. Thus, with probability at least $\frac{2}{3}$, $G[Q]$ is $\frac{\varepsilon}{2}$-close to bipartiteness (as deleting the edges in $E(G[Q]) \cap E$ makes $G[Q]$ bipartite), and G is accepted by the tester. ■

Chapter 5

Testing Graphs against an Unknown Distribution

This chapter is devoted to proving Theorem 13 and related results (see Section 1.2.2), and is organized as follows. Section 5.1 describes the combinatorial statement at the heart of Theorem 13, as well as its connection to other related results in the area of extremal graph theory. Section 5.2 is concerned with proving *vertex-weighted* analogues of several standard regularity-method lemmas (most notably regularity and counting lemmas, and corollaries thereof). In Section 5.3 we prove the "if" direction of Theorem 13 (i.e. Theorem 5.1.1). This is by far the most challenging (and interesting) part of this chapter. The main step towards proving Theorem 13 is establishing Lemma 5.3.1, which is the key lemma of the chapter. For the reader's convenience, we give in Section 5.3.1 an overview of the key ideas used in the proof. As the proofs in Section 5.2 are somewhat routine, we encourage readers who are familiar with the regularity method to skip Section 5.2 (at least on their first read), and go directly to Section 5.3.

The "only if" direction of Theorem 13 is proved in Section 5.4. In Section 5.5 we study various variants of the VDF model, and also raise two related open problems; one is to what extent the variants of the VDF model allow for testability of non-hereditary properties, and the other asks if the sample complexity in the VDF model is the same as in the standard model (for properties that are testable in the VDF model), see Section 5.5.3. Along the way we resolve another open problem raised in [58] (see Lemma 5.5.6). Throughout this chapter, when we say that a function is increasing/decreasing we mean weakly increasing/decreasing (i.e. non-decreasing/non-increasing).

5.1 The Combinatorial Essence of Theorem 13

To prove (the "if" direction of) Theorem 13, we will actually prove the following theorem, which can be thought of as a vertex-weighted version of the infinite graph removal lemma of [10] (see Chapter 1).

Theorem 5.1.1. *For every hereditary and extendable graph property $\mathcal{P}$ there is a function $s_{\mathcal{P}} : (0,1) \to \mathbb{N}$ such that the following holds for every $\varepsilon > 0$ and for every vertex-weighted graph $(G, \mathcal{D})$ which is ε-far from $\mathcal{P}$. Let $u_1, \ldots, u_s$, $s = s_{\mathcal{P}}(\varepsilon)$, be a sequence of random vertices of G, sampled according to $\mathcal{D}$ and independently. Then $G[\{u_1, \ldots, u_s\}]$ does not satisfy $\mathcal{P}$ with probability at least $\frac{2}{3}$.*

The following similar-looking result[1] was proved by Austin and Tao [17] and Lovász and Szegedy [81].

Theorem 5.1.2 ([17, 81])**.** *For every hereditary graph property $\mathcal{P}$ there is a function $s_{\mathcal{P}} : (0,1) \to \mathbb{N}$ such that the following holds for every $\varepsilon > 0$ and for every vertex-weighted graph $(G, \mathcal{D})$ which is ε-far from $\mathcal{P}$. Let $u_1, \ldots, u_s$, $s = s_{\mathcal{P}}(\varepsilon)$, be a sequence of random vertices of G, sampled according to $\mathcal{D}$ and independently. Construct a graph S on s by letting $\{i,j\} \in E(S)$ if and only if $\{u_i, u_j\} \in E(G)$. Then S does not satisfy $\mathcal{P}$ with probability at least $\frac{2}{3}$.*

Note that Theorem 5.1.2 holds for all hereditary properties, while Theorem 5.1.1 only holds for hereditary properties which are extendable. Observe that the graph S in Theorem 5.1.2 is a blowup of the graph $G[U]$, where $U = \{u_1, \ldots, u_s\}$. Thus, the difference between Theorems 5.1.1 and 5.1.2 is that Theorem 5.1.2 only guarantees that a *blowup* of $G[U]$ does not satisfy $\mathcal{P}$ with high probability (i.e., 2/3), while Theorem 5.1.1 guarantees the stronger assertion that $G[U]$ itself does not satisfy $\mathcal{P}$ with high probability. This is an important difference: while Theorem 5.1.1 immediately implies the existence of a VDF-tester for every hereditary and extendable property $\mathcal{P}$ (see Section 5.3.3), we do not know of any way of using Theorem 5.1.2 to prove the existence of such a tester. One natural candidate for a tester derived from Theorem 5.1.2 would be the algorithm which accepts if and only if the graph S (defined in Theorem 5.1.2) does not satisfy $\mathcal{P}$. It turns out, however, that this algorithm often fails to be a valid tester[2].

It is worth noting that Theorem 5.1.2 can be deduced from the "unweighted" case, i.e. the result of [10], via a simple argument, see Lemma 5.5.6 and the discussion following it. On the other hand, the proof of Theorem 5.1.1 requires several new ideas on top of those used in [10].

5.2 Preliminary Lemmas

In this section we introduce vertex-weighted analogues of some key tools of the regularity method, most notably Szemerédi's regularity lemma [110], the strong regularity lemma [5], and the counting lemma, as well as some standard corollaries thereof. We also prove some other auxiliary lemmas needed for the proof of Theorem 13. We start with two simple lemmas regarding probability distributions on a finite set. Given a distribution $\mathcal{D}$ on a set U and a subset $W \subseteq U$, we set $\mathcal{D}(W) := \sum_{w \in W} \mathcal{D}(w)$. We denote by $\mathcal{D}_W$ the distribution $\mathcal{D}$ conditioned on W, namely $\mathcal{D}_W(w) = \frac{\mathcal{D}(w)}{\mathcal{D}(W)}$ for every $w \in W$.

Lemma 5.2.1. *For every set U, for every $\eta \in (0,1)$ and for every distribution $\mathcal{D}$ on U, there is a partition $\mathcal{P}$ of U into $\lceil 1/\eta \rceil$ parts such that $\sum_{W \in \mathcal{P}} \sum_{\{x,y\} \in \binom{W}{2}} \mathcal{D}(x)\mathcal{D}(y) \leq \eta$.*

Proof. Let $\mathcal{P}$ be a random partition of U into $k := \lceil 1/\eta \rceil$ parts, where each element is assigned to one of the parts uniformly at random and independently of all other elements. Then for every pair of distinct elements $x, y \in U$, the probability that x and y belong to the same part is exactly $\frac{1}{k}$. By linearity of

[1] We note that the results of [17] and [81] are in fact more general. The authors of [81] actually prove that the conclusion of Theorem 5.1.2 holds for all graphons. The authors of [17] prove extensions of Theorem 5.1.2 in several directions, including a version for uniform hypergraphs, and a strengthening in which the notion of testability is replaced with the stronger notion of repairability.

[2] For example, if $\mathcal{P} = C_5$-freeness then the proposed tester will reject with high probability if the input graph is a triangle with a uniform vertex-distribution (as the graph S will be a (large) blowup of a triangle, and thus contain a copy of C_5), even though this input graph clearly satisfies $\mathcal{P}$.

expectation we have

$$\mathbb{E}\left[\sum_{W\in\mathcal{P}}\sum_{\{x,y\}\in\binom{W}{2}}\mathcal{D}(x)\mathcal{D}(y)\right]=\sum_{\{x,y\}\in\binom{U}{2}}\mathcal{D}(x)\mathcal{D}(y)\cdot\frac{1}{k}<\frac{1}{2}\cdot\frac{1}{k}<\eta,$$

so there is a choice of $\mathcal{P}$ with the required property. ■

Lemma 5.2.2. *Let $a>0$ be an integer, let U be a finite set and let $\mathcal{D}$ be a distribution on U such that $\mathcal{D}(u)\leq\frac{1}{2a}$ for every $u\in U$. Then there is a partition $U=U_1\cup\cdots\cup U_a$ such that $\mathcal{D}(U_i)\geq\frac{1}{2a}$ for every $1\leq i\leq a$.*

Proof. We proof is by induction on a. The base case $a=1$ is trivial, so we assume from now on that $a\geq 2$. Let $U_1\subseteq U$ be a set of minimal size satisfying $\mathcal{D}(U_1)\geq\frac{1}{2a}$. Then $\mathcal{D}(U_1)\leq\frac{1}{a}$, because otherwise we could remove an arbitrary element of U_1 (whose weight by assumption is at most $\frac{1}{2a}$) and thus get a proper subset of U_1 having weight at least $\frac{1}{2a}$, in contradiction the minimality of U_1. Now set $U':=U\setminus U_1$, noting that $\mathcal{D}(U')\geq 1-\frac{1}{a}$. Then every $u\in U'$ satisfies

$$\mathcal{D}_{U'}(u)=\frac{\mathcal{D}(u)}{\mathcal{D}(U')}\leq\frac{\frac{1}{2a}}{1-\frac{1}{a}}=\frac{1}{2(a-1)}\ .$$

So by the induction hypothesis for $(U',\mathcal{D}_{U'})$, there is a partition $U'=U_2\cup\cdots\cup U_a$ such that

$$\mathcal{D}(U_i)=\mathcal{D}_{U'}(U_i)\cdot\mathcal{D}(U')\geq\frac{1}{2(a-1)}\cdot\mathcal{D}(U')\geq\frac{1}{2(a-1)}\cdot\left(1-\frac{1}{a}\right)=\frac{1}{2a}$$

for every $2\leq i\leq a$. This completes the proof. ■

Throughout this chapter we consider *vertex-weighted graphs*, i.e. pairs $(G,\mathcal{D})$ such that G is a graph and $\mathcal{D}$ is a distribution on $V(G)$. For a set $X\subseteq V(G)$, the *subgraph of $(G,\mathcal{D})$ induced by X* is defined to be $(G[X],\mathcal{D}_X)$, where $\mathcal{D}_X$ is the distribution $\mathcal{D}$ conditioned on X. The weight of an edge/non-edge $\{x,y\}$ (with respect to $\mathcal{D}$) is defined as $\mathcal{D}(x)\mathcal{D}(y)$. For a pair of disjoint sets $X,Y\subseteq V(G)$ with $\mathcal{D}(X),\mathcal{D}(Y)>0$, define the *density* of (X,Y), denoted $d(X,Y)$, to be $d(X,Y)=\frac{1}{\mathcal{D}(X)\mathcal{D}(Y)}\sum_{(x,y)\in E(X,Y)}\mathcal{D}(x)\mathcal{D}(y)$, where $E(X,Y)$ is the set of edges with one endpoint in X and one endpoint in Y. If $\mathcal{D}(X)=0$ or $\mathcal{D}(Y)=0$ then define $d(X,Y)=0$. A pair of disjoint vertex-sets (X,Y) is called *ε-regular* if for every $X'\subseteq X$ and $Y'\subseteq Y$ with $\mathcal{D}(X')\geq\varepsilon\mathcal{D}(X)$ and $\mathcal{D}(Y')\geq\varepsilon\mathcal{D}(Y)$, it holds that $|d(X',Y')-d(X,Y)|\leq\varepsilon$. The following lemma describes some basic properties of ε-regular pairs.

Lemma 5.2.3. *Let $(G,\mathcal{D})$ be a vertex-weighted graph, and let $X,Y\subseteq V(G)$ be disjoint vertex-sets such that $\mathcal{D}(X),\mathcal{D}(Y)>0$, and such that the pair (X,Y) is ε-regular with density d. Then the following holds.*

1. *For every $\alpha\geq\varepsilon$ and $X'\subseteq X$, $Y'\subseteq Y$ with $\mathcal{D}(X')\geq\alpha\mathcal{D}(X)$ and $\mathcal{D}(Y')\geq\alpha\mathcal{D}(Y)$, the pair (X',Y') has density at least $d-\varepsilon$ and at most $d+\varepsilon$, and is ε'-regular with $\varepsilon'=\max\{\varepsilon/\alpha,2\varepsilon\}$.*
2. *The set of vertices $x\in X$ which satisfy $|d(x,Y)-d|>\varepsilon$ has weight less than $2\varepsilon\cdot\mathcal{D}(X)$.*

Proof. Starting with Item 1, let $X' \subseteq X$ and $Y' \subseteq Y$ be such that $\mathcal{D}(X') \geq \alpha \mathcal{D}(X)$ and $\mathcal{D}(Y') \geq \alpha \mathcal{D}(Y)$. Since $\alpha \geq \varepsilon$, the ε-regularity of (X,Y) implies that $d - \varepsilon \leq d(X',Y') \leq d + \varepsilon$. Now let us show that (X',Y') is ε'-regular with $\varepsilon' = \max\{\varepsilon/\alpha, 2\varepsilon\}$. Let $X'' \subseteq X'$ and $Y'' \subseteq Y'$ be such that $\mathcal{D}(X'') \geq \varepsilon' \mathcal{D}(X')$ and $\mathcal{D}(Y'') \geq \varepsilon' \mathcal{D}(Y')$. Then $\mathcal{D}(X'') \geq \frac{\varepsilon}{\alpha} \mathcal{D}(X') \geq \varepsilon \mathcal{D}(X)$ and similarly $\mathcal{D}(Y'') \geq \varepsilon \mathcal{D}(Y)$. So by the ε-regularity of (X,Y) we have $|d(X'',Y'') - d(X,Y)| \leq \varepsilon$ and hence $|d(X'',Y'') - d(X',Y')| \leq 2\varepsilon \leq \varepsilon'$, as required.

We now prove Item 2. Let X^+ (resp. X^-) be the set of all $x \in X$ satisfying $d(x,Y) > d + \varepsilon$ (resp. $d(x,Y) < d - \varepsilon$). We have

$$\begin{aligned} d(X^+,Y) &= \frac{1}{\mathcal{D}(X^+)\mathcal{D}(Y)} \cdot \sum_{x \in X^+} \sum_{y \in N_Y(x)} \mathcal{D}(x)\mathcal{D}(y) = \frac{1}{\mathcal{D}(X^+)\mathcal{D}(Y)} \cdot \sum_{x \in X^+} \mathcal{D}(x) \cdot \mathcal{D}(Y) \cdot d(x,Y) \\ &> \frac{1}{\mathcal{D}(X^+)\mathcal{D}(Y)} \cdot \mathcal{D}(X^+)\mathcal{D}(Y) \cdot (d + \varepsilon) = d + \varepsilon. \end{aligned}$$

So unless $\mathcal{D}(X^+) < \varepsilon \mathcal{D}(X)$, we get a contradiction to the ε-regularity of (X,Y). Similarly, we must have $\mathcal{D}(X^-) < \varepsilon \mathcal{D}(X)$. The assertion follows. ■

The following is a vertex-weighted counting lemma.

Lemma 5.2.4 (Counting lemma for vertex-weighted graphs)**.** *For every integer $h \geq 2$ and $\eta \in (0,1)$ there is $\delta = \delta_{5.2.4}(h,\eta)$ such that the following holds. Let H be a graph on $[h]$ and let $U_1, \ldots, U_h$ be pairwise-disjoint vertex-sets in a vertex-weighted graph $(G, \mathcal{D})$, such that the following holds.*

1. *For every $1 \leq i < j \leq h$, if $\{i,j\} \in E(H)$ then $d(U_i,U_j) \geq \eta$ and if $\{i,j\} \notin E(H)$ then $d(U_i,U_j) \leq 1-\eta$.*
2. *For every $1 \leq i < j \leq h$, the pair (U_i,U_j) is δ-regular.*

Let $\mathcal{U}$ be the set of all $(u_1, \ldots, u_h) \in U_1 \times \cdots \times U_h$ such that $u_1, \ldots, u_h$ induce a copy of H in which u_i plays the role of i for every $1 \leq i \leq h$. Then $\sum_{(u_1,\ldots,u_h) \in \mathcal{U}} \prod_{i=1}^h \mathcal{D}(u_i) \geq \delta \prod_{i=1}^h \mathcal{D}(U_i)$.

Proof. If $\mathcal{D}(U_i) = 0$ for some $1 \leq i \leq h$ then there is nothing to prove, so suppose that $\mathcal{D}(U_i) > 0$ for every $1 \leq i \leq h$. The proof is by induction on h. The base case $h = 2$ trivially holds with $\delta = \delta(2,\eta) = \eta$. So from now on we assume that $h \geq 3$, and set

$$\delta = \delta(h,\eta) = \min\left\{ \frac{1}{4(h-1)},\ \frac{\eta}{2},\ \frac{1}{2} \cdot \left(\frac{\eta}{2}\right)^{h-1} \cdot \delta(h-1,\eta/2) \right\}.$$

For each $2 \leq i \leq h$, let W_i be the set of all vertices $u_1 \in U_1$ for which $|d(u_1,U_i) - d(U_1,U_i)| > \delta$. By Item 2 of Lemma 5.2.3, we have $\mathcal{D}(W_i) < 2\delta \cdot \mathcal{D}(U_1)$. Hence, the set $U_1' := U_1 \setminus \bigcup_{i=2}^h W_i$ satisfies $\mathcal{D}(U_1') > \mathcal{D}(U_1) - (h-1) \cdot 2\delta \cdot \mathcal{D}(U_1) \geq \frac{1}{2}\mathcal{D}(U_1)$, where in the last inequality we used our choice of δ. Now fix any $u_1 \in U_1'$. We define sets $U_2', \ldots, U_h'$ as follows: for $2 \leq i \leq h$, if $\{1,i\} \in E(H)$ then set $U_i' = N_{U_i}(u_1)$, and if $\{1,i\} \notin E(H)$ then set $U_i' = U_i \setminus N_{U_i}(u_1)$. By using Item 1 and the fact that $u_1 \in U_1'$, we get that $\mathcal{D}(U_i') \geq (\eta - \delta)\mathcal{D}(U_i) \geq \frac{\eta}{2} \cdot \mathcal{D}(U_i)$ for every $2 \leq i \leq h$. By Item 1 of Lemma 5.2.3, and by Conditions 1-2 of the current lemma, we get that for every $2 \leq i < j \leq h$, the pair (U_i',U_j') is δ'-regular with $\delta' = 2\delta/\eta \leq \delta(h-1,\eta/2)$, and that if $\{i,j\} \in E(H)$ then $d(U_i',U_j') \geq \eta - \delta \geq \eta/2$ and if $\{i,j\} \notin E(H)$ then $d(U_i',U_j') \leq 1 - \eta + \delta \leq 1 - \frac{\eta}{2}$.

We now see that the sets $U'_2, \ldots, U'_h$ satisfy the requirements of the lemma with respect to the graph $H' = H[\{2, \ldots, h\}]$ and with $\frac{\eta}{2}$ in place of η. Let $\mathcal{U}'$ be the set of all $(u_2, \ldots, u_h) \in U'_2 \times \cdots \times U'_h$ such that $u_2, \ldots, u_h$ induce a copy of H' with u_i playing the role of i for every $2 \leq i \leq h$. By the induction hypothesis, we have

$$\sum_{(u_2,\ldots,u_h)\in\mathcal{U}'} \prod_{i=2}^{h} \mathcal{D}(u_i) \geq \delta(h-1,\eta/2) \cdot \prod_{i=2}^{h} \mathcal{D}(U'_i) \geq \delta(h-1,\eta/2) \cdot (\eta/2)^{h-1} \cdot \prod_{i=2}^{h} \mathcal{D}(U_i) \geq 2\delta \prod_{i=2}^{h} \mathcal{D}(U_i).$$

For every $(u_2, \ldots, u_h) \in \mathcal{U}'$, the tuple $(u_1, \ldots, u_h)$ induces a copy of H with u_i playing the role of i for every $1 \leq i \leq h$. Hence, for every $(u_2, \ldots, u_h) \in \mathcal{U}'$ we have $(u_1, \ldots, u_h) \in \mathcal{U}$ (where $\mathcal{U}$ is defined in the statement of the lemma). Since this is true for every $u_1 \in U'_1$, we get that

$$\sum_{(u_1,\ldots,u_h)\in\mathcal{U}} \prod_{i=1}^{h} \mathcal{D}(u_i) \geq \sum_{u_1\in U'_1} \mathcal{D}(u_1) \cdot 2\delta \prod_{i=2}^{h} \mathcal{D}(U_i) = \mathcal{D}(U'_1) \cdot 2\delta \prod_{i=2}^{h} \mathcal{D}(U_i) \geq \delta \prod_{i=1}^{h} \mathcal{D}(U_i),$$

as required. ■

A partition $\mathcal{P} = \{V_1, \ldots, V_r\}$ of the vertex-set of a vertex-weighted graph $(G, \mathcal{D})$ is called ε-regular if the sum of $\mathcal{D}(V_i)\mathcal{D}(V_j)$ over all pairs $1 \leq i < j \leq r$ for which (V_i, V_j) is not an ε-regular pair, is at most ε. We now extend some basic properties of regular partitions to the vertex-weighted setting.

Lemma 5.2.5. *Let X, Y be disjoint vertex-sets in a vertex-weighted graph $(G, \mathcal{D})$, and let $\mathcal{P}_X, \mathcal{P}_Y$ be partitions of X, Y, respectively. For every $X' \in \mathcal{P}_X$ and $Y' \in \mathcal{P}_Y$, set $\varepsilon(X', Y') = d(X', Y') - d(X, Y)$. Then*

$$\sum_{X'\in\mathcal{P}_X, Y'\in\mathcal{P}_Y} \mathcal{D}(X')\mathcal{D}(Y') \cdot d(X', Y') = \mathcal{D}(X)\mathcal{D}(Y) \cdot d(X, Y),$$

and

$$\sum_{X'\in\mathcal{P}_X, Y'\in\mathcal{P}_Y} \mathcal{D}(X')\mathcal{D}(Y') \cdot d^2(X', Y') = \mathcal{D}(X)\mathcal{D}(Y) \cdot d^2(X, Y) + \sum_{X'\in\mathcal{P}_X, Y'\in\mathcal{P}_Y} \mathcal{D}(X')\mathcal{D}(Y') \cdot \varepsilon^2(X', Y').$$

As the proof of Lemma 5.2.5 is simple and routine[3], we leave it to the reader.

Let $(G, \mathcal{D})$ be a vertex-weighted graph and let $\mathcal{P} = \{P_1, \ldots, P_r\}$ be a partition of $V(G)$. The index of $\mathcal{P}$, denoted $q(\mathcal{P})$, is defined as

$$q(\mathcal{P}) = \sum_{1\leq i<j\leq r} \mathcal{D}(P_i)\mathcal{D}(P_j) \cdot d^2(P_i, P_j).$$

Lemma 5.2.6. *For every vertex-partition $\mathcal{P}$ of a vertex-weighted graph $(G, \mathcal{D})$, and for every refinement $\mathcal{P}'$ of $\mathcal{P}$, we have $q(\mathcal{P}') \geq q(\mathcal{P})$.*

Proof. Write $\mathcal{P} = \{P_1, \ldots, P_r\}$, and for each $1 \leq i \leq r$ put $\mathcal{P}'_i = \{P' \in \mathcal{P}' : P' \subseteq P_i\}$. Then

$$q(\mathcal{P}') \geq \sum_{1\leq i<j\leq r} \sum_{P'_i\in\mathcal{P}'_i, P'_j\in\mathcal{P}'_j} \mathcal{D}(P'_i)\mathcal{D}(P'_j) \cdot d^2(P'_i, P'_j) \geq \sum_{1\leq i<j\leq r} \mathcal{D}(P_i)\mathcal{D}(P_j) \cdot d^2(P_i, P_j) = q(\mathcal{P}),$$

where in the second inequality we used the second part of Lemma 5.2.5. ■

[3]Pick $X' \in \mathcal{P}_X$ and $Y' \in \mathcal{P}_Y$ randomly (and independently) with probability $\mathcal{D}(X')/\mathcal{D}(X)$ and $\mathcal{D}(Y')/\mathcal{D}(Y)$, respectively. Consider the random variable $Z := d(X', Y')$. Then the first part of Lemma 5.2.5 is simply the (easy) fact that $\mathbb{E}[Z] = d(X, Y)$, and the second part of Lemma 5.2.5 is simply the fact that $\mathbb{E}[Z^2] = \mathbb{E}[Z]^2 + \mathrm{Var}[Z]$.

Lemma 5.2.7. *Let $(G,\mathcal{D})$ be a vertex-weighted graph and let $\mathcal{P}=\{P_1,\ldots,P_r\}$ be a non-ε-regular partition of $V(G)$. Then there is a refinement $\mathcal{P}'$ of $\mathcal{P}$ such that $|\mathcal{P}'|\leq|\mathcal{P}|\cdot 2^{|\mathcal{P}|}$ and $q(\mathcal{P}')\geq q(\mathcal{P})+\varepsilon^5$.*

Proof. For each $1\leq i<j\leq r$ for which (P_i,P_j) is not ε-regular, let $P_{i,j}\subseteq P_i$, $P_{j,i}\subseteq P_j$ be such that $\mathcal{D}(P_{i,j})\geq\varepsilon\mathcal{D}(P_i),\mathcal{D}(P_{j,i})\geq\varepsilon\mathcal{D}(P_j)$, and $|d(P_{i,j},P_{j,i})-d(P_i,P_j)|>\varepsilon$. For each $1\leq i\leq r$, let $\mathcal{P}_i$ be the partition of P_i, formed by taking the common refinement of the partitions $\{P_{i,j},P_i\setminus P_{i,j}\}$, where j runs over all indices for which (P_i,P_j) is not ε-regular. Let $\mathcal{P}'=\bigcup_{i=1}^r\mathcal{P}_i$ be the resulting refinement of $\mathcal{P}$. Then clearly $|\mathcal{P}'|\leq|\mathcal{P}|\cdot 2^{|\mathcal{P}|}$. We now show that $q(\mathcal{P}')\geq q(\mathcal{P})+\varepsilon^5$. First, observe that by Lemma 5.2.5, for every $1\leq i<j\leq r$ we have $\sum_{X'\in\mathcal{P}_i,Y'\in\mathcal{P}_j}\mathcal{D}(X')\mathcal{D}(Y')\cdot d^2(X',Y')\geq\mathcal{D}(P_i)\mathcal{D}(P_j)\cdot d^2(P_i,P_j)$. Next, fix any pair $1\leq i<j\leq r$ for which (P_i,P_j) is not ε-regular. By Lemma 5.2.5 we have

$$\sum_{X'\in\mathcal{P}_i,Y'\in\mathcal{P}_j}\mathcal{D}(X')\mathcal{D}(Y')\cdot d^2(X',Y')=$$

$$\mathcal{D}(P_i)\mathcal{D}(P_j)\cdot d^2(P_i,P_j)+\sum_{X'\in\mathcal{P}_i,Y'\in\mathcal{P}_j}\mathcal{D}(X')\mathcal{D}(Y')\cdot\big(d(X',Y')-d(P_i,P_j)\big)^2\geq$$

$$\mathcal{D}(P_i)\mathcal{D}(P_j)\cdot d^2(P_i,P_j)+\sum_{X'\subseteq P_{i,j},Y'\subseteq P_{j,i}}\mathcal{D}(X')\mathcal{D}(Y')\cdot\big(d(X',Y')-d(P_i,P_j)\big)^2=$$

$$\mathcal{D}(P_i)\mathcal{D}(P_j)\cdot d^2(P_i,P_j)+\sum_{X'\subseteq P_{i,j},Y'\subseteq P_{j,i}}\mathcal{D}(X')\mathcal{D}(Y')\cdot\big[\big(d(X',Y')-d(P_{i,j},P_{j,i})\big)+(d(P_{i,j},P_{j,i})-d(P_i,P_j))\big]^2\geq$$

$$\mathcal{D}(P_i)\mathcal{D}(P_j)\cdot d^2(P_i,P_j)+\mathcal{D}(P_{i,j})\mathcal{D}(P_{j,i})\cdot(d(P_{i,j},P_{j,i})-d(P_i,P_j))^2\geq\mathcal{D}(P_i)\mathcal{D}(P_j)\cdot(d^2(P_i,P_j)+\varepsilon^4),$$

where in the penultimate inequality we used the first part of Lemma 5.2.5 to infer that

$$\sum_{X'\subseteq P_{i,j},Y'\subseteq P_{j,i}}\mathcal{D}(X')\mathcal{D}(Y')\cdot\big(d(X',Y')-d(P_{i,j},P_{j,i})\big)=0.$$

Denoting by $\mathcal{N}$ the set of pairs $1\leq i<j\leq r$ for which (P_i,P_j) is not ε-regular, we see that

$$\begin{aligned}q(\mathcal{P}')&\geq\sum_{1\leq i<j\leq r}\sum_{X'\in\mathcal{P}_i,Y'\in\mathcal{P}_j}\mathcal{D}(X')\mathcal{D}(Y')\cdot d^2(X',Y')\\&\geq\sum_{1\leq i<j\leq r}\mathcal{D}(P_i)\mathcal{D}(P_j)\cdot d^2(P_i,P_j)+\sum_{(i,j)\in\mathcal{N}}\mathcal{D}(P_i)\mathcal{D}(P_j)\cdot\varepsilon^4\\&=q(\mathcal{P})+\varepsilon^4\cdot\sum_{(i,j)\in\mathcal{N}}\mathcal{D}(P_i)\mathcal{D}(P_j)\geq q(\mathcal{P})+\varepsilon^5,\end{aligned}$$

where in the last inequality we used the assumption that $\mathcal{P}$ is not ε-regular. ■

We now prove vertex-weighted versions[4] of Szemerédi's regularity lemma [110] and of the strong regularity lemma [5].

Lemma 5.2.8 (Szemerédi's regularity lemma for vertex-weighted graphs)**.** *For every $\varepsilon\in(0,1)$ and $m\geq 0$ there is $T=T_{5.2.8}(\varepsilon,m)$ such that for every vertex-weighted graph $(G,\mathcal{D})$ and for every partition $\mathcal{P}_0$ of $V(G)$ of size not larger than m, there is an ε-regular partition $\mathcal{P}$ of $V(G)$ which has at most T parts and refines $\mathcal{P}_0$.*

[4]We note that a weighted version of Szemerédi's regularity lemma, where both vertex-weights and edge-weights are allowed, was proved in [37], but only under the assumption that all vertex-weights are $o(1)$. Hence this result is unsuitable in our setting.

Proof. For $i \geq 0$, if $\mathcal{P}_i$ is not ε-regular then we apply Lemma 5.2.7 to obtain a partition $\mathcal{P}_{i+1}$ which refines $\mathcal{P}_i$ and satisfies $|\mathcal{P}_{i+1}| \leq |\mathcal{P}_i| \cdot 2^{|\mathcal{P}_i|}$ and $q(\mathcal{P}_{i+1}) \geq q(\mathcal{P}_i) + \varepsilon^5$. Since the index of any partition is at most 1, this process must end after at most ε^{-5} steps. When the process ends, we have an ε-regular partition. Since the number of steps depends only on ε, the size of the resulting final partition can be upper-bounded by a function of ε and $|\mathcal{P}_0|$, as required. ■

Lemma 5.2.9 (Strong regularity lemma for vertex-weighted graphs)**.** *For every function $\mathcal{E} : \mathbb{N} \to (0,1)$ and for every integer m, there is $S = S_{5.2.9}(\mathcal{E}, m)$ such that for every vertex-weighted graph $(G, \mathcal{D})$ and for every partition $\mathcal{P}_0$ of $V(G)$ of size at most m, there is a refinement $\mathcal{P}$ of $\mathcal{P}_0$, and a refinement $\mathcal{Q}$ of $\mathcal{P}$, such that the following holds.*

1. *$|\mathcal{Q}| \leq S$.*
2. *The partition $\mathcal{Q}$ is $\mathcal{E}(|\mathcal{P}|)$-regular.*
3. *$\sum_{P_1,P_2 \in \mathcal{P}} \sum_{Q_1 \subseteq P_1, Q_2 \subseteq P_2} \mathcal{D}(Q_1)\mathcal{D}(Q_2) \cdot |d(Q_i,Q_j) - d(P_i,P_j)| \leq \mathcal{E}(0)$. Here the outer sum is over all unordered pairs of distinct $P_1, P_2 \in \mathcal{P}$, and the inner sum is over all $Q_1, Q_2 \in \mathcal{Q}$ such that $Q_i \subseteq P_i$ for $i = 1, 2$.*

Proof. We may assume, without loss of generality, that $\mathcal{E}$ is monotone decreasing. Let $\mathcal{P}_1$ be the partition obtained by applying Lemma 2.2.4 with parameter $\varepsilon = \mathcal{E}(0)$ and with the partition $\mathcal{P}_0$. Next, for each $i \geq 1$, apply Lemma 2.2.4 with parameter $\mathcal{E}(|\mathcal{P}_i|)$ and with the partition $\mathcal{P}_i$ to obtain a partition $\mathcal{P}_{i+1}$ which is $\mathcal{E}(|\mathcal{P}_i|)$-regular and refines $\mathcal{P}_i$. In light of Lemma 5.2.6, and as the index of any partition is at most 1, there must be some $1 \leq i \leq \frac{1}{\mathcal{E}^2(0)}$ for which $q(\mathcal{P}_{i+1}) \leq q(\mathcal{P}_i) + \mathcal{E}^2(0)$. For such an i, set $\mathcal{P} = \mathcal{P}_i$ and $\mathcal{Q} = \mathcal{P}_{i+1}$. Since $|\mathcal{P}_0| \leq m$ and the number of steps in the process is at most $\mathcal{E}^2(0)$, and since the size of the partition guaranteed by Lemma 2.2.4 can be bounded from above by a function of the parameters of this lemma (which in our case depend only on $\mathcal{E}$ and m), we see that $|\mathcal{Q}|$ too can be bounded from above by a function of $\mathcal{E}$ and m. This proves Item 1.

Item 2 is immediate from our choice of $\mathcal{Q}$. It remains to prove Item 3. By the definition of the index and by our choice of $\mathcal{P}$ and $\mathcal{Q}$, we have

$$q(\mathcal{P}) + \mathcal{E}^2(0) \geq q(\mathcal{Q}) \geq \sum_{P_1,P_2 \in \mathcal{P}} \sum_{Q_1 \subseteq P_1, Q_2 \subseteq P_2} \mathcal{D}(Q_1)\mathcal{D}(Q_2) \cdot d^2(Q_1,Q_2) =$$
$$\sum_{P_1,P_2 \in \mathcal{P}} \mathcal{D}(P_1)\mathcal{D}(P_2) \cdot d^2(P_1,P_2) + \sum_{P_1,P_2 \in \mathcal{P}} \sum_{Q_1 \subseteq P_1, Q_2 \subseteq P_2} \mathcal{D}(Q_1)\mathcal{D}(Q_2) \cdot (d(Q_1,Q_2) - d(P_1,P_2))^2 =$$
$$q(\mathcal{P}) + \sum_{P_1,P_2 \in \mathcal{P}} \sum_{Q_1 \subseteq P_1, Q_2 \subseteq P_2} \mathcal{D}(Q_1)\mathcal{D}(Q_2) \cdot (d(Q_1,Q_2) - d(P_1,P_2))^2,$$

where in the first equality we used the second part of Lemma 5.2.5. The above implies that

$$\sum_{P_1,P_2 \in \mathcal{P}} \sum_{Q_1 \subseteq P_1, Q_2 \subseteq P_2} \mathcal{D}(Q_1)\mathcal{D}(Q_2) \cdot (d(Q_1,Q_2) - d(P_1,P_2))^2 \leq \mathcal{E}^2(0),$$

and hence

$$\sum_{P_1,P_2 \in \mathcal{P}} \sum_{Q_1 \subseteq P_1, Q_2 \subseteq P_2} \mathcal{D}(Q_1)\mathcal{D}(Q_2) \cdot |d(Q_1,Q_2) - d(P_1,P_2)| \leq$$

$$\sqrt{\sum_{P_1,P_2\in\mathcal{P}}\sum_{Q_1\subseteq P_1,Q_2\subseteq P_2}\mathcal{D}(Q_1)\mathcal{D}(Q_2)\cdot\left(d(Q_1,Q_2)-d(P_1,P_2)\right)^2}\leq\mathcal{E}(0),$$

where the first inequality follows from Cauchy-Schwarz. This completes the proof. ■

Our last two lemmas are vertex-weighted analogues of well-known corollaries to Szemerédi's regularity lemma and the strong regularity lemma, respectively. The "unweighted" versions of these corollaries were used in [10] in order to prove that every hereditary property is testable in the standard model.

Lemma 5.2.10. *For every integer $t\geq 1$ and for every $\delta>0$ there is $\zeta=\zeta_{5.2.10}(t,\delta)>0$, such that the following holds. Let $(G,\mathcal{D})$ be a vertex-weighted graph such that all vertices in G have weight less than ζ. Then there are pairwise-disjoint vertex-sets $Q_1,\ldots,Q_t\subseteq V(G)$ with the following properties.*

1. *$\mathcal{D}(Q_i)\geq\zeta$ for every $1\leq i\leq t$.*
2. *(Q_i,Q_j) is δ-regular for every $1\leq i<j\leq t$.*
3. *Either all pairs (Q_i,Q_j) have density at least $\frac{1}{2}$, or all pairs (Q_i,Q_j) have density less than $\frac{1}{2}$.*

Proof. Setting $a=4^t$ and $\varepsilon=\frac{\delta}{4a^4}$, we will prove the lemma with

$$\zeta=\zeta_{5.2.10}(t,\delta)=\frac{1}{4a^2\cdot T_{2.2.4}(\varepsilon,a)}\ .$$

Let $(G,\mathcal{D})$ satisfying $\mathcal{D}(v)<\zeta$ for every $v\in V(G)$. Apply Lemma 5.2.2 with $U=V(G)$, with the distribution $\mathcal{D}$, and with a as defined above. Lemma 5.2.2 supplies a partition $V(G)=U_1\cup\cdots\cup U_a$ such that $\mathcal{D}(U_i)\geq\frac{1}{2a}$ for every $1\leq i\leq a$. Now apply Lemma 2.2.4 to $(G,\mathcal{D})$ with parameter ε and with the partition $\mathcal{P}_0:=\{U_1,\ldots,U_a\}$, to obtain an ε-regular partition $\mathcal{P}$ which refines $\mathcal{P}_0$. For each $1\leq i\leq a$, put $\mathcal{P}_i=\{P\in\mathcal{P}:P\subseteq U_i\}$, and sample $P_i\in\mathcal{P}_i$ with probability proportional to the weight of the parts, i.e. $P_i=P$ with probability $\frac{\mathcal{D}(P)}{\mathcal{D}(U_i)}$ for every $P\in\mathcal{P}_i$. We claim that with positive probability, $\mathcal{D}(P_i)\geq\zeta$ for every $1\leq i\leq a$, and all pairs (P_i,P_j) are δ-regular. For every $1\leq i\leq a$, the probability that $\mathcal{D}(P_i)<\zeta$ is less than $\frac{\zeta\cdot|\mathcal{P}|}{\mathcal{D}(U_i)}\leq\frac{\zeta\cdot T_{2.2.4}(\varepsilon,a)}{1/2a}\leq\frac{1}{2a}$, where in the first inequality we used the guarantees of Lemma 2.2.4. By the union bound, with probability at least $\frac{1}{2}$ we have $\mathcal{D}(P_i)\geq\zeta$ for every $1\leq i\leq a$. Next, observe that since $\mathcal{P}$ is ε-regular and as $\varepsilon\leq\delta$, the probability that (P_i,P_j) is not δ-regular (for some specific $1\leq i<j\leq a$) is at most $\frac{\varepsilon}{\mathcal{D}(U_i)\mathcal{D}(U_j)}\leq 4a^2\varepsilon\leq\frac{1}{a^2}$. So by taking the union bound over all pairs $1\leq i<j\leq a$, we get that with probability at least $1-\binom{a}{2}\cdot\frac{1}{a^2}>\frac{1}{2}$, all pairs (P_i,P_j) are δ-regular. This proves our assertion.

We thus showed that there is a choice of $P_1,\ldots,P_a$ such that $\mathcal{D}(P_i)\geq\zeta$ for every $1\leq i\leq a$ and such that (P_i,P_j) is δ-regular for every $1\leq i<j\leq a$. Now consider an auxiliary graph on $[a]$ in which $\{i,j\}$ is an edge if $d(P_i,P_j)\geq\frac{1}{2}$ and $\{i,j\}$ is a non-edge if $d(P_i,P_j)<\frac{1}{2}$. As $a=4^t$, a well-known bound on Ramsey numbers implies that this graph contains either a clique or an independent set $\{i_1,\ldots,i_t\}$. Then $Q_1=P_{i_1},\ldots,Q_t=P_{i_t}$ satisfy the requirements of the lemma. ■

Lemma 5.2.11. *For every function $\mathcal{E}:\mathbb{N}\to(0,1)$ and for every integer m, there is $S=S_{5.2.11}(\mathcal{E},m)>0$ such that for every vertex-weighted graph $(G,\mathcal{D})$ and for every partition $\mathcal{P}_0$ of $V(G)$ having size at most m, there is a partition $\mathcal{P}=\{P_0,P_1,\ldots,P_r\}$ of $V(G)$ and vertex-sets $Q_i\subseteq P_i$ for $1\leq i\leq r$, such that the following holds:*

1. $\mathcal{D}(P_0) < \mathcal{E}(0)$.

2. *For every* $1 \le i \le r$, P_i *is contained in some part of* $\mathcal{P}_0$.

3. $\mathcal{D}(Q_i) \ge 1/S$ *for every* $1 \le i \le r$. *In particular,* $r \le S$.

4. *For every* $1 \le i < j \le r$, *the pair* (Q_i, Q_j) *is* $\mathcal{E}(r)$*-regular.*

5. $\sum_{1 \le i < j \le r} \mathcal{D}(P_i)\mathcal{D}(P_j) \cdot |d(Q_i, Q_j) - d(P_i, P_j)| \le \mathcal{E}(0)$.

Proof. We may and will assume $\mathcal{E}$ is monotone decreasing[5]. For convenience, put $\varepsilon = \mathcal{E}(0)$. Let $\mathcal{E}' : \mathbb{N} \to (0,1)$ be the function $\mathcal{E}'(r) = \min\left\{\mathcal{E}(r), \frac{\varepsilon^2}{2r^4}, \frac{\varepsilon}{3}\right\}$. We will show that one can choose $S = S_{5.2.11}(\mathcal{E}, m) := \frac{3s^3}{\varepsilon}$, where $s := S_{5.2.9}(\mathcal{E}', m)$. Apply Lemma 5.2.9 to $(G, \mathcal{D})$ with parameter $\mathcal{E}'$ and with the given partition $\mathcal{P}_0$, to obtain partitions $\mathcal{P}'$ and $\mathcal{Q}$ such that $\mathcal{P}'$ refines $\mathcal{P}_0$, $\mathcal{Q}$ refines $\mathcal{P}'$, and Items 1-3 in Lemma 5.2.9 hold. Let P_0 be the union of all parts of $\mathcal{P}'$ of weight less than $\varepsilon/|\mathcal{P}'|$, and let $P_1, \dots, P_r$ be the parts of $\mathcal{P}'$ of weight at least $\varepsilon/|\mathcal{P}'|$. Then we have $\mathcal{D}(P_0) < |\mathcal{P}'| \cdot \varepsilon/|\mathcal{P}'| = \varepsilon$, establishing Item 1. Now set $\mathcal{P} = \{P_0, P_1, \dots, P_r\}$. It is evident that Item 2 holds.

For each $1 \le i \le r$, denote $\mathcal{Q}_i = \{Q \in \mathcal{Q} : Q \subseteq P_i\}$, and sample $Q_i \in \mathcal{Q}_i$ with probability proportional to the weight of the parts; in other words, for each $Q \in \mathcal{Q}_i$, the probability that $Q_i = Q$ is $\frac{\mathcal{D}(Q)}{\mathcal{D}(P_i)}$. We will show that with positive probability, $Q_1, \dots, Q_r$ satisfy Items 3-5. For each $1 \le i \le r$, the probability that $\mathcal{D}(Q_i) < \frac{\mathcal{D}(P_i)}{3r|\mathcal{Q}|}$ is less than $|\mathcal{Q}| \cdot \frac{1}{3r|\mathcal{Q}|} = \frac{1}{3r}$. By the union bound, the probability that there is $1 \le i \le r$ for which $\mathcal{D}(Q_i) < \frac{\mathcal{D}(P_i)}{3r|\mathcal{Q}|}$ is less than $\frac{1}{3}$. So with probability larger than $\frac{2}{3}$, for every $1 \le i \le r$ we have

$$\mathcal{D}(Q_i) \ge \frac{\mathcal{D}(P_i)}{3r|\mathcal{Q}|} \ge \frac{\varepsilon}{3|\mathcal{P}'|^2|\mathcal{Q}|} \ge \frac{\varepsilon}{3|\mathcal{Q}|^3} \ge \frac{\varepsilon}{3s^3} = \frac{1}{S},$$

where the last inequality is due to our choice of $\mathcal{Q}$ via Lemma 5.2.9.

We now prove that Item 4 holds with probability greater than $\frac{2}{3}$. Fix any $1 \le i < j \le r$. Since $\mathcal{Q}$ is ε'-regular with $\varepsilon' = \mathcal{E}'(|\mathcal{P}|') \le \min\left\{\mathcal{E}(|\mathcal{P}'|), \frac{\varepsilon^2}{2|\mathcal{P}'|^4}\right\}$, and since $\mathcal{E}(|\mathcal{P}'|) \le \mathcal{E}(r)$ (by the monotonicity of $\mathcal{E}$), the probability that the pair (Q_i, Q_j) is not $\mathcal{E}(r)$-regular is at most $\frac{\varepsilon^2/(2|\mathcal{P}'|^4)}{\mathcal{D}(P_i)\mathcal{D}(P_j)} \le \frac{1}{2}|\mathcal{P}'|^{-2} \le \frac{1}{2}r^{-2}$, where the first inequality holds because $\mathcal{D}(P_i), \mathcal{D}(P_j) \ge \varepsilon/|\mathcal{P}'|$. By the union bound over all pairs $1 \le i < j \le r$, the probability that there is $1 \le i < j \le r$ for which (Q_i, Q_j) is not $\mathcal{E}(r)$-regular is at most $\binom{r}{2} \cdot \frac{1}{2}r^{-2} < \frac{1}{3}$.

It remains to show that Item 5 holds with probability at least $\frac{2}{3}$. Observe that

$$\mathbb{E}\left[\sum_{1 \le i < j \le r} \mathcal{D}(P_i)\mathcal{D}(P_j) \cdot |d(Q_i, Q_j) - d(P_i, P_j)|\right] =$$
$$\sum_{1 \le i < j \le r} \sum_{Q_i' \in \mathcal{Q}_i, Q_j' \in \mathcal{Q}_j} \mathcal{D}(Q_i')\mathcal{D}(Q_j') \cdot \left|d(Q_i', Q_j') - d(P_i, P_j)\right| \le \frac{\varepsilon}{3},$$

where in the inequality we used Item 3 of Lemma 5.2.9, our choice of $\mathcal{E}'$, and the fact that $P_1, \dots, P_r \in \mathcal{P}'$. So by Markov's inequality, the probability that Item 5 fails is at most $\frac{1}{3}$, as required. ■

[5]Indeed, we can replace $\mathcal{E}$ with $\mathcal{E}'(r) = \min_{s \le r} \mathcal{E}(s)$, which is clearly monotone decreasing.

5.3 Proof of the "If" Part of Theorem 13

In this section we prove Theorem 5.1.1 and then use it to derive the "if" part of Theorem 13. This section is broken up into subsections, as follows. In Section 5.3.1 we give a high-level overview of the main obstacle one needs to overcome in proving Theorem 13, and the main idea behind the way we overcome it. In Section 5.3.2 we state and prove Lemma 5.3.1, which constitutes the main ingredient in the proof of Theorem 13. Finally, we prove (the "if" direction of) Theorem 13 in Section 5.3.3.

5.3.1 Proof Overview

The main difficulty: Let $\mathcal{P}$ be a hereditary and extendable graph property. We are given a graph G and a distribution $\mathcal{D}$ so that G is ε-far from $\mathcal{P}$ with respect to $\mathcal{D}$. Our goal is to show that a sample[6] of $O(1)$ vertices of G (sampled according to $\mathcal{D}$) finds with high probability (whp) an induced subgraph F of G which does not satisfy $\mathcal{P}$. There are two ways one can try to achieve this goal. First, one can take a blowup G' of G, in which a vertex is replaced by a cluster of vertices whose size is proportional to the vertex's weight under $\mathcal{D}$, and thus (try to) "reduce" the problem to the non-weighted (i.e., uniform) case. While this approach allows one to handle some properties[7], it does not work in general. The main bottleneck seems to be that a copy of F in G' does not correspond necessarily to a copy of F in G, since F might contain several of the vertices that replaced some vertex v of G. Moreover, if this vertex v has weight $\Omega(1)$ (i.e., if its weight is bounded from below by a function of ε), then even a sample whose size depends only on ε will very likely contain several of the vertices of G' that replaced v.

A second approach to test for $\mathcal{P}$ in the VDF model would be to just reprove the result of [10], while replacing the regularity lemmas used there with regularity lemmas for vertex-weighted graphs. While such vertex-weighted analogues are indeed not hard to obtain (see e.g. Lemmas 5.2.4 and 5.2.8-5.2.11), the main problem is again vertices of high weight. Now the issue is that parts of the regular partition might contain only a single vertex of high weight (and many vertices of very small, or even 0, weight), a situation in which one would not be able to embed graphs F that need to use more than one vertex from the same part.

The key new idea: The main idea is then to prove a lemma that allows one to partition G into three sets X, Y, Z with the following properties: (i) Z will have total weight at most $\varepsilon/2$; (ii) all vertices in X will have weight at least $\Omega(1)$ (i.e., weight bounded from below by a function of ε); (iii) Y will have a highly regular Szemerédi partition, that is, there will be a partition of the vertices of Y into sets $P_1, \ldots, P_r$ so that the bipartite graphs between (almost) all pairs (P_i, P_j) are pseudo-random (in an appropriate sense which takes into account the vertex-weights); (iv) each of the parts P_i will have "sufficiently many" vertices; and (v) for each $x \in X$ and $1 \leq i \leq r$, either x will be connected to all vertices of P_i or to none of them. Let us now see how a partition with the above five properties can allow one to test $\mathcal{P}$.

We first claim that $G[X \cup Y]$ (i.e. the graph induced by $X \cup Y$) is $\varepsilon/2$-far from satisfying $\mathcal{P}$. Indeed, if this is not the case, then we can first turn $G[X \cup Y]$ into a graph satisfying $\mathcal{P}$ by making changes of total weight less than $\varepsilon/2$, and then use the fact that $\mathcal{P}$ is extendable and the fact that the total weight of Z is at most $\varepsilon/2$ in order to reconnect the vertices of Z to $X \cup Y$ (and amongst themselves) so that the resulting

[6] Throughout Section 5.3.1, we use $O(1)$ to denote a positive number which is bounded from above by a function of ε (namely, independent of n and $\mathcal{D}$).

[7] Indeed, this is the approach used in [58].

graph will satisfy $\mathcal{P}$. The total weight of edges we change by doing this is less than ε, a contradiction.

We now examine the partition $P_1, \ldots, P_r$ of Y and perform a "cleaning" procedure analogous to the one performed in applications of the regularity lemma. By this we mean that we make within Y (and only within Y!) changes of total weight less than $\varepsilon/2$ so that if after these changes the set Y contains an induced copy of some graph F, then in the original graph, a sample of $O(1)$ vertices from Y finds one such copy with high probability (whp). Here we will also rely on property (iv) of the partition. The fact that $G[X \cup Y]$ is $\varepsilon/2$-far from satisfying $\mathcal{P}$ and that we made changes of total weight less than $\varepsilon/2$ when cleaning Y, means that after the cleaning, the subgraph induced by $X \cup Y$ indeed has an induced copy of a graph F that does not satisfy $\mathcal{P}$. We now claim that a sample of $O(1)$ vertices of G finds a copy of F whp. First, since the total weight of Z is small, sampling from G is (effectively) like sampling from $G[X \cup Y]$. Now let F_X (resp. F_Y) be the subgraph of F induced by X (resp. Y). By the above discussion, a sample of size $O(1)$ finds a copy of F_Y whp. Now, and this is the first crucial point, the above property (v) guarantees that the vertices of X belonging to F_X form a copy of F with *every* set of vertices in Y which forms a copy of F_Y. Moreover, and this is the second crucial point, property (ii) above guarantees that a sample of $O(1)$ vertices finds *the*[8] copy of F_X contained in X whp. Altogether, the algorithm finds an induced copy of F using $O(1)$ samples.

The new regularity lemma: As it turns out, one cannot hope to partition G as described in the first paragraph above, and instead we will have to define a partition with a much more complicated set of features, see Lemma 5.3.1 below. One of the main difficulties in proving this lemma is that on the one hand, we would like to make sure that the weight of every vertex in Y is very small compared to the total weight of each of the parts $P_1, \ldots, P_r$ (this will ensure that each of these parts has sufficiently many vertices, guaranteeing property (iv) above), while on the other hand, the number r of parts $P_1, \ldots, P_r$ needs to be very large in order to satisfy property (v) above (which forces the (average) weight of each part to be very small). The proof of Lemma 5.3.1 will use some of the lemmas of Section 5.2, most notably Lemma 5.2.11, which we will need to iterate (at least implicitly) in order to find the sought-after partition in the statement of Lemma 5.3.1.

5.3.2 The Key Lemma

Lemma 5.3.1. *For every function $\Psi : \mathbb{N} \to \mathbb{N}$ and $\varepsilon > 0$ there is $S = S_{5.3.1}(\Psi, \varepsilon) > 0$ such that for every vertex-weighted graph $(G, \mathcal{D})$ there is a partition $V(G) = X \cup Y \cup Z$, a partition $\mathcal{P} = \{P_1, \ldots, P_r\}$ of Y, vertex-sets $Q_i \subseteq P_i$, and pairwise-disjoint vertex-sets $Q_{i,1}, \ldots, Q_{i,t} \subseteq Q_i$, where $t = \Psi(|X| + r)$, such that the following holds:*

1. *$\mathcal{D}(Z) < \varepsilon$.*
2. *Every vertex in X has weight at least $1/S$.*
3. *For every $x \in X$ and for every $1 \leq i \leq r$, either x is adjacent to all vertices of P_i, or to none of the vertices of P_i.*

[8] By "the" we mean that X might contain only a single copy of F_X, but this copy has to be of weight $\Omega(1)$. This is in sharp contrast to the situation within Y, where each copy of F_Y might have very small weight, but the *total* weight of such copies must be $\Omega(1)$.

4. $\sum_{1\le i\le r}\sum_{\{x,y\}\in\binom{P_i}{2}}\mathcal{D}(x)\mathcal{D}(y)\le\varepsilon$.

5. $\sum_{1\le i<j\le r}\mathcal{D}(P_i)\mathcal{D}(P_j)\cdot|d(Q_i,Q_j)-d(P_i,P_j)|\le\varepsilon$.

6. *For every* $1\le i\le r$, *all pairs* $(Q_{i,k},Q_{i,\ell})$ *are* $\frac{1}{\Psi(|X|+r)}$*-regular, and either all pairs* $(Q_{i,k},Q_{i,\ell})$ *have density at least* $\frac{1}{2}$, *or all pairs* $(Q_{i,k},Q_{i,\ell})$ *have density less than* $\frac{1}{2}$.

7. *For every* $1\le i<j\le r$ *and* $1\le k,\ell\le t$, *the pair* $(Q_{i,k},Q_{j,\ell})$ *is* $\frac{1}{\Psi(|X|+r)}$*-regular and* $|d(Q_{i,k},Q_{j,\ell})-d(Q_i,Q_j)|\le\frac{1}{\Psi(|X|+r)}$.

8. *For every* $1\le i\le r$ *and* $1\le k\le t$, $\mathcal{D}(Q_{i,k})\ge 1/S$.

Note that Items 2 and 8 in Lemma 5.3.1 together imply that $|X|+rt\le S$. The following lemma constitutes the main part of the proof of Lemma 5.3.1. After proving Lemma 5.3.2, we deduce Lemma 5.3.1 from Lemmas 5.3.2 and 5.2.10.

Lemma 5.3.2. *For every function* $\Psi:\mathbb{N}\to\mathbb{N}$ *and* $\varepsilon>0$ *there is* $S=S_{5.3.2}(\Psi,\varepsilon)>0$ *such that for every vertex-weighted graph* $(G,\mathcal{D})$ *there is a partition* $V(G)=X\cup Y\cup Z$, *a partition* $\mathcal{P}=\{P_1,\dots,P_r\}$ *of* Y *and vertex-sets* $Q_i\subseteq P_i$ *(for* $1\le i\le r$*) such that Items 1-5 in Lemma 5.3.1 hold (with respect to* $S=S_{5.3.2}(\Psi,\varepsilon)$*), and such that the following two conditions are satisfied.*

(a) For every $1\le i<j\le r$, *the pair* (Q_i,Q_j) *is* $\frac{1}{\Psi(|X|+r)}$*-regular.*

(b) For every $1\le i\le r$ *the following holds:* $\mathcal{D}(Q_i)\ge 1/S$, *and all vertices in* Q_i *have weight less than* $\frac{1}{\Psi(|X|+r)}\cdot\mathcal{D}(Q_i)$.

Proof. We may and will assume that the function Ψ is monotone increasing[9], and that the function $S_{5.2.11}(\mathcal{E},m)$, whose existence is guaranteed by Lemma 5.2.11, is monotone decreasing in $\mathcal{E}$ and monotone increasing in m. Here, being monotone decreasing in $\mathcal{E}$ means that if a pair of functions $\mathcal{E}_1,\mathcal{E}_2:\mathbb{N}\to(0,1)$ satisfy $\mathcal{E}_1(r)\le\mathcal{E}_2(r)$ for every $r\in\mathbb{N}$, then $S_{5.2.11}(\mathcal{E}_1,m)\ge S_{5.2.11}(\mathcal{E}_2,m)$ for every m. For each $s\in\mathbb{N}$, define the function $\mathcal{E}_s:\mathbb{N}\to(0,1)$ by

$$\mathcal{E}_s(r)=\min\left\{\frac{\varepsilon}{2},\frac{1}{\Psi(s+r)}\right\}.$$

Now define the functions $S',S'':\mathbb{N}\to\mathbb{N}$ by setting:

$$S'(s)=S_{5.2.11}\left(\mathcal{E}_s,2^s\cdot\lceil 1/\varepsilon\rceil\right),\quad S''(s)=\max\left\{s,\ \frac{2S'(s)}{\varepsilon}\cdot\Psi\left(s+S'(s)\right)\right\}.$$

Note that $S''(s)\ge s$ for every $s\in\mathbb{N}$, and that S' and S'' are monotone increasing. We define a monotone increasing sequence $s_1,s_2,\dots$ as follows: $s_1=1$, and for each $i\ge 2$, $s_i=S''(s_{i-1})$. We will show that the lemma holds with

$$S=S_{5.3.2}(\Psi,\varepsilon)=s_{\lceil 2/\varepsilon\rceil}\ .$$

Let $(G,\mathcal{D})$ be a vertex-weighted graph. We iteratively define a sequence of pairwise-disjoint vertex-sets $X_1,X_2,\dots\subseteq V(G)$ as follows: let X_1 be the set of all vertices of G of weight at least $1/s_1$; for each

[9]To guarantee that Ψ is monotone increasing, we can simply replace Ψ with the function $\Psi'(s):=\max\{\Psi(0),\dots,\Psi(s)\}$.

$i \geq 2$, let X_i be the set of all vertices in $V(G) \setminus (X_1 \cup \cdots \cup X_{i-1})$ having weight at least $1/s_i$. Since $X_1, X_2, \ldots$ are pairwise-disjoint, there must be $1 \leq i \leq \lceil 2/\varepsilon \rceil$ for which $\mathcal{D}(X_i) \leq \varepsilon/2$. We now set $Z' = X_i$, $X = X_1 \cup \cdots \cup X_{i-1}$ and $Y' = V(G) \setminus (X \cup Z') = V(G) \setminus (X_1 \cup \cdots \cup X_i)$. Note that $\mathcal{D}(Z') \leq \varepsilon/2$. Setting $s := s_{i-1} \leq s_{\lceil 2/\varepsilon \rceil - 1} \leq S$, note that every vertex in X has weight at least $\frac{1}{s}$ (so in particular $|X| \leq s$), while every vertex in Y' has weight less than $\frac{1}{s_i} = \frac{1}{S''(s)}$.

If $\mathcal{D}(Y') < \frac{\varepsilon}{2}$ then $\mathcal{D}(Y' \cup Z') < \varepsilon$, so the assertion of the lemma holds for $Y = \emptyset$ and $Z = Z' \cup Y'$, and we are done. So we may and will assume from now on that $\mathcal{D}(Y') \geq \frac{\varepsilon}{2}$. Let $\mathcal{P}'_0$ be a partition of Y' into $\lceil 1/\varepsilon \rceil$ parts such that $\sum_{P \in \mathcal{P}'_0} \sum_{\{x,y\} \in \binom{P}{2}} \mathcal{D}(x)\mathcal{D}(y) \leq \varepsilon$, as guaranteed by Lemma 5.2.1. For every $x \in X$, consider the partition $\mathcal{P}_x := \{N_{Y'}(x), Y' \setminus N_{Y'}(x)\}$ of Y'. Let $\mathcal{P}_0$ be the common refinement of the partitions $\mathcal{P}'_0$ and $(\mathcal{P}_x)_{x \in X}$. Then for every $x \in X$ and $P \in \mathcal{P}_0$, either x is adjacent to every vertex of P, or x is not adjacent to any vertex of P. Moreover, we have $|\mathcal{P}_0| \leq 2^{|X|} \cdot \lceil 1/\varepsilon \rceil \leq 2^s \cdot \lceil 1/\varepsilon \rceil$.

Now apply Lemma 5.2.11 to $(G[Y'], \mathcal{D}_{Y'})$ with parameters $\mathcal{E}_s$ and $m = 2^s \cdot \lceil 1/\varepsilon \rceil$, and with the partition $\mathcal{P}_0$ (noting that $|\mathcal{P}_0| \leq m$), to obtain a partition $\mathcal{P} = \{P_0, P_1, \ldots, P_r\}$ of Y' and vertex-sets $Q_i \subseteq P_i$ (for $1 \leq i \leq r$), with the properties stated in that lemma. Note that in particular we have

$$r \leq S_{5.2.11}\left(\mathcal{E}_s, 2^s \cdot \lceil 1/\varepsilon \rceil\right) = S'(s). \tag{5.1}$$

Set $Z = Z' \cup P_0$ and $Y = Y' \setminus P_0$, noting that $\mathcal{D}(P_0) < \mathcal{E}_s(0) \leq \frac{\varepsilon}{2}$, and hence $\mathcal{D}(Z) = \mathcal{D}(Z') + \mathcal{D}(P_0) < \varepsilon$, as required by Item 1 in Lemma 5.3.1. Items 3 and 4 in Lemma 5.3.1 hold because each of the sets $P_1, \ldots, P_r$ is contained in some part of $\mathcal{P}_0$, and hence also in some part of $\mathcal{P}'_0$. Item 2 of Lemma 5.3.1 was already verified above, and Item 5 of Lemma 5.3.1 is guaranteed by Lemma 5.2.11. Item (a) holds because Lemma 5.2.11 guarantees that all pairs (Q_i, Q_j) are $\mathcal{E}_s(r)$-regular, and because $\mathcal{E}_s(r) \leq \frac{1}{\Psi(s+r)} \leq \frac{1}{\Psi(|X|+r)}$ (here we used our choice of $\mathcal{E}_s$, the fact that $|X| \leq s$, and the monotonicity of Ψ). It remains to prove Item (b). For each $1 \leq i \leq r$, we have

$$\begin{aligned} \mathcal{D}(Q_i) &= \mathcal{D}_{Y'}(Q_i) \cdot \mathcal{D}(Y') \geq \mathcal{D}_{Y'}(Q_i) \cdot \frac{\varepsilon}{2} \geq \frac{\varepsilon}{2S_{5.2.11}\left(\mathcal{E}_s, 2^s \cdot \lceil 1/\varepsilon \rceil\right)} \\ &= \frac{\varepsilon}{2S'(s)} \geq \frac{1}{S''(s)} \geq \frac{1}{S''(s_{\lceil 2/\varepsilon \rceil - 1})} = \frac{1}{s_{\lceil 2/\varepsilon \rceil}} = \frac{1}{S}\,, \end{aligned} \tag{5.2}$$

where in the second inequality we used the guarantees of Lemma 5.2.11, and later we used our choice of S' and S'', the monotonicity of S'', and the fact that $s \leq s_{\lceil 2/\varepsilon \rceil - 1}$. Next, fix $1 \leq i \leq r$ and recall that all vertices in $Q_i \subseteq Y \subseteq Y'$ have weight less than

$$\begin{aligned} \frac{1}{S''(s)} &\leq \frac{1}{\Psi\left(s + S'(s)\right)} \cdot \frac{\varepsilon}{2S'(s)} \\ &\leq \frac{1}{\Psi(s+r)} \cdot \mathcal{D}(Q_i) \leq \frac{1}{\Psi(|X|+r)} \cdot \mathcal{D}(Q_i), \end{aligned}$$

where in the first inequality we used our choice of S'', in the last two inequalities we used the monotonicity of Ψ, and in the second inequality we also used (5.1) and an intermediate step in (5.2). This shows that $\mathcal{D}(u) < \frac{1}{\Psi(|X|+r)} \cdot \mathcal{D}(Q_i)$ for every $1 \leq i \leq r$ and $u \in Q_i$, as required. ■

Proof of Lemma 5.3.1. Define the functions

$$\zeta : \mathbb{N} \to (0,1),\ \zeta(m) = \zeta_{5.2.10}\left(\Psi(m), \frac{1}{\Psi(m)}\right),$$

and

$$\Psi' : \mathbb{N} \to \mathbb{N},\ \Psi'(m) = \frac{2\Psi(m)}{\zeta(m)}.$$

We may and will assume that the function $\zeta_{5.2.10}(t,\delta)$ is monotone decreasing in t and monotone increasing in δ. This assumption implies that the function ζ defined above is monotone decreasing. We prove the lemma with

$$S = S_{5.3.1}(\Psi,\varepsilon) := \frac{S_{5.3.2}(\Psi',\varepsilon)}{\zeta(S_{5.3.2}(\Psi',\varepsilon))} \geq S_{5.3.2}(\Psi',\varepsilon)\ .$$

Let $(G,\mathcal{D})$ be a vertex-weighted graph. Apply Lemma 5.3.2 to $(G,\mathcal{D})$ with parameters Ψ' and ε, to obtain a partition $V(G) = X \cup Y \cup Z$, a partition $\mathcal{P} = \{P_1,\ldots,P_r\}$ of Y, and subsets $Q_i \subseteq P_i$ (for $1 \leq i \leq r$) such that Items 1-5 of Lemma 5.3.1 hold (with respect to $S_{5.3.2}(\Psi',\varepsilon)$), and so do Items (a) and (b) of Lemma 5.3.2.

Let us now prove that Items 6-8 (in Lemma 5.3.1) hold. It will be convenient to put $m := |X| + r$. By Item (b) in Lemma 5.3.2 and by our choice of Ψ', we have

$$\mathcal{D}(u) < \frac{1}{\Psi'(m)} \cdot \mathcal{D}(Q_i) < \frac{\zeta(m)}{\Psi(m)} \cdot \mathcal{D}(Q_i) \leq \zeta(m) \cdot \mathcal{D}(Q_i) \tag{5.3}$$

for every $1 \leq i \leq r$ and $u \in Q_i$. Recalling our choice of ζ, we see that Lemma 5.2.10 is applicable to $(G[Q_i],\mathcal{D}_{Q_i})$ with parameters $t = \Psi(m) = \Psi(|X|+r)$ and $\delta = \frac{1}{\Psi(m)} = \frac{1}{\Psi(|X|+r)}$. Applying Lemma 5.2.10 with this input, we obtain pairwise-disjoint vertex-sets $Q_{i,1},\ldots,Q_{i,t} \subseteq Q_i$ satisfying the properties stated in that lemma. The guarantees of Lemma 5.2.10 immediately establish Item 6, and also imply that for every $1 \leq k \leq t$ we have

$$\mathcal{D}(Q_{i,k}) \geq \zeta(m) \cdot \mathcal{D}(Q_i) = \zeta(|X|+r) \cdot \mathcal{D}(Q_i) \geq \zeta(|X|+r) \cdot \frac{1}{S_{5.3.2}(\Psi',\varepsilon)} \geq \frac{\zeta(S_{5.3.2}(\Psi',\varepsilon))}{S_{5.3.2}(\Psi',\varepsilon)} = \frac{1}{S},$$

where in the second and third inequalities we used the fact that $|X|+r, \frac{1}{\mathcal{D}(Q_i)} \leq S_{5.3.2}(\Psi',\varepsilon)$, as guaranteed by Item 2 of Lemma 5.3.1 and Item (b) of Lemma 5.3.2; in the third inequality we also used the monotonicity of ζ. This establishes Item 8. It remains to prove Item 7. By Item (a) of Lemma 5.3.2, the pair (Q_i,Q_j) is $\frac{1}{\Psi'(m)}$-regular for every $1 \leq i < j \leq r$. Fix any $1 \leq k,\ell \leq t$. Recalling that $\frac{1}{\Psi'(m)} = \frac{\zeta(m)}{2\Psi(m)}$ and that $\mathcal{D}(Q_{i,k}) \geq \zeta(m)\cdot\mathcal{D}(Q_i)$, $\mathcal{D}(Q_{j,\ell}) \geq \zeta(m)\cdot\mathcal{D}(Q_j)$, we apply Item 1 of Lemma 5.2.3 to $Q_i,Q_j,Q_{i,k},Q_{j,\ell}$ with parameter $\alpha = \zeta(m)$, to conclude that $|d(Q_{i,k},Q_{j,\ell}) - d(Q_i,Q_j)| \leq \frac{1}{\Psi'(m)} \leq \frac{1}{\Psi(m)} = \frac{1}{\Psi(|X|+r)}$, and that the pair $(Q_{i,k},Q_{j,\ell})$ is $\frac{1}{\Psi(|X|+r)}$-regular, as required. ■

5.3.3 The "If" Part of Theorem 13: Proof of Theorem 5.1.1

We first explain how to derive the "if" part of Theorem 13 from Theorem 5.1.1. Let $\mathcal{P}$ be a hereditary and extendable graph property. Our tester for $\mathcal{P}$ will work as follows: given an input $(G,\mathcal{D})$ and a proximity parameter ε, the tester samples a sequence of vertices $u_1,\ldots,u_s \in V(G)$ independently and with distribution $\mathcal{D}$, where $s = s_{\mathcal{P}}(\varepsilon)$ is as in Theorem 5.1.1; the tester then accepts if and only if $G[\{u_1,\ldots,u_s\}]$ satisfies $\mathcal{P}$. Since $\mathcal{P}$ is hereditary, this tester accepts with probability 1 if the input graph satisfies $\mathcal{P}$. In the other direction, Theorem 5.1.1 immediately implies that if the input $(G,\mathcal{D})$ is ε-far from $\mathcal{P}$ then the tester rejects with probability at least $\frac{2}{3}$, as required.

From now on our goal is to prove Theorem 5.1.1. We start by introducing variants of some definitions from [10]. An *embedding scheme* is a complete graph K with a vertex partition $A_K \cup B_K$, such that every

vertex in B_K is colored black or white, every edge with an endpoint in A_K is colored black or white, and every edge contained in B is colored black, white or grey. Note that one of A_k, B_k may be empty; that the vertices of A_K are not colored; and that the edges with at least one endpoint in A_K cannot be colored grey. An *embedding* from a graph F to an embedding scheme K is a map $\varphi : V(F) \to V(K)$ such that the following holds:

1. For every $a \in A_K$ we have $|\varphi^{-1}(a)| \leq 1$.
2. For every $b \in B_K$, if b is colored black then $\varphi^{-1}(b)$ induces a complete graph, and if b is colored white then $\varphi^{-1}(b)$ induces an empty graph.
3. For every $\{x, y\} \in \binom{V(K)}{2}$, if $\{x, y\}$ is colored black then the bipartite graph between $\varphi^{-1}(x)$ and $\varphi^{-1}(y)$ is complete, and if $\{x, y\}$ is colored white then the bipartite graph between $\varphi^{-1}(x)$ and $\varphi^{-1}(y)$ is empty (note that there are no restrictions in the case that $\{x, y\}$ is colored grey).

Note that Condition 3 implies that for every $a \in A_K$ and $x \in V(K) \setminus \{a\}$, the bipartite graph between $\varphi^{-1}(a)$ and $\varphi^{-1}(x)$ is either complete or empty. We use the notation $F \to K$ to mean that there is an embedding from F to K. For a graph-family $\mathcal{F}$ and an integer m, let $\mathcal{F}_m$ be the family of all embedding schemes K on at most m vertices, such that there is an embedding from some $F \in \mathcal{F}$ to K. We now introduce a variant of the function $\Psi_{\mathcal{F}}$ defined in [10].

Definition 5.3.3. *For a graph-family $\mathcal{F}$ and an integer m for which $\mathcal{F}_m \neq \emptyset$, define*

$$\Psi_{\mathcal{F}}(m) = \max_{K \in \mathcal{F}_m} \min_{F \in \mathcal{F} : F \to K} |V(F)|.$$

If $\mathcal{F}_m = \emptyset$ then define $\Psi_{\mathcal{F}}(m) = 0$.

We are now ready to prove Theorem 5.1.1.

Proof of Theorem 5.1.1. Let $\mathcal{P}$ be a hereditary and extendable graph property. Let $\mathcal{F} = \mathcal{F}(\mathcal{P})$ be the family of graphs which do not satisfy $\mathcal{P}$. Fix $\varepsilon \in (0, 1)$, and let $\Psi : \mathbb{N} \to \mathbb{N}$ be the function

$$\Psi(m) = \max\left\{\frac{8}{\varepsilon}, \Psi_{\mathcal{F}}(m), \frac{1}{\delta_{5.2.4}(\Psi_{\mathcal{F}}(m), \frac{\varepsilon}{8})}\right\},$$

where $\Psi_{\mathcal{F}}$ is defined in Definition 5.3.3. We may and will assume that the function $\delta_{5.2.4}(h, \eta)$ is monotone decreasing in h and monotone increasing in η. Set $S := S_{5.3.1}(\Psi, \frac{\varepsilon}{4})$. We prove the theorem with

$$s = s_{\mathcal{P}}(\varepsilon) := \frac{2S^{S+1}}{\delta_{5.2.4}\left(S, \frac{\varepsilon}{8}\right)}\ . \tag{5.4}$$

Let $(G, \mathcal{D})$ be a vertex-weighted graph which is ε-far from $\mathcal{P}$. Apply Lemma 5.3.1 to $(G, \mathcal{D})$ with parameter $\frac{\varepsilon}{4}$ and with Ψ as above, to obtain a partition $V(G) = X \cup Y \cup Z$, a partition $\{P_1, \ldots, P_r\}$ of Y, subsets $Q_i \subseteq P_i$ (for $1 \leq i \leq r$), and pairwise-disjoint subsets $Q_{i,1}, \ldots, Q_{i,t} \subseteq Q_i$, such that $t = \Psi(|X| + r)$ and Items 1-8 in Lemma 5.3.1 hold.

We claim that G is $\frac{3\varepsilon}{4}$-far from any graph G' on $V(G)$ which satisfies $G'[X \cup Y] \in \mathcal{P}$. So suppose by contradiction that there is a graph G' on $V(G)$ such that $G'[X \cup Y]$ satisfies $\mathcal{P}$ and such that G' is $\frac{3\varepsilon}{4}$-close

to G. Since $\mathcal{P}$ is extendable, there is a graph G'' on $V(G) = V(G')$ such that $G''[X \cup Y] = G'[X \cup Y]$ and such that G'' satisfies $\mathcal{P}$. In order to turn G' into G'', we only need to add/delete edges which are incident to vertices of Z. Therefore, the total weight of edge-changes needed to turn G' into G'' is at most $\mathcal{D}(Z) < \frac{\varepsilon}{4}$, as guaranteed by Item 1 of Lemma 5.3.1. So we see that G can be turned into G'', which satisfies $\mathcal{P}$, by adding/deleting edges whose total weight is less than $\frac{3\varepsilon}{4} + \frac{\varepsilon}{4} = \varepsilon$, in contradiction the assumption that $(G, \mathcal{D})$ is ε-far from $\mathcal{P}$.

We thus proved that G is $\frac{3\varepsilon}{4}$-far from any graph G' satisfying $G'[X \cup Y] \in \mathcal{P}$. Now, let G' be the graph obtained from G by doing the following changes:

1. For every $1 \leq i \leq r$, if $d(Q_{i,k}, Q_{i,\ell}) \geq \frac{1}{2}$ for every $1 \leq k < \ell \leq t$ then turn P_i into a clique, and if $d(Q_{i,k}, Q_{i,\ell}) < \frac{1}{2}$ for every $1 \leq k < \ell \leq t$, then turn P_i into an independent set. By Item 6 in Lemma 5.3.1, one of these options has to hold. The total weight of edge-changes needed in this item is at most $\frac{\varepsilon}{4}$ by Item 4 of Lemma 5.3.1.

2. For every $1 \leq i < j \leq r$, if $d(Q_i, Q_j) > 1 - \frac{\varepsilon}{4}$ then add all edges between P_i and P_j, and if $d(Q_i, Q_j) < \frac{\varepsilon}{4}$ then remove all edges between P_i and P_j (note that if $\frac{\varepsilon}{4} \leq d(Q_i, Q_j) \leq 1 - \frac{\varepsilon}{4}$ then no changes are made in the bipartite graph between P_i and P_j). The total weight of edge-changes needed in this item is less than $\frac{\varepsilon}{2}$ by Item 5 of Lemma 5.3.1. Indeed, observe that the total weight of changes between P_i, P_j is less than $\mathcal{D}(P_i)\mathcal{D}(P_j) \cdot \left(|d(Q_i, Q_j) - d(P_i, P_j)| + \frac{\varepsilon}{4}\right)$ by the triangle inequality. Hence, the total weight of changes is less than

$$\sum_{1 \leq i < j \leq r} \mathcal{D}(P_i)\mathcal{D}(P_j) \cdot \left(|d(Q_i, Q_j) - d(P_i, P_j)| + \frac{\varepsilon}{4}\right) \leq$$
$$\frac{\varepsilon}{4} + \sum_{1 \leq i < j \leq r} \mathcal{D}(P_i)\mathcal{D}(P_j) \cdot |d(Q_i, Q_j) - d(P_i, P_j)| \leq \frac{\varepsilon}{2}.$$

Note that no edge with an endpoint in X was added/deleted in Items 1-2, so G' and G agree on all edges that are incident to vertices of X.

We see that the total weight of edge-changes made in Items 1-2 is less than $\frac{3\varepsilon}{4}$. So $G'[X \cup Y]$ cannot satisfy $\mathcal{P}$, implying that $G'[X \cup Y] \in \mathcal{F}$. Note that by definition (see Items 1-2 above), the graph G' has the following properties:

(a) For every $1 \leq i \leq r$, P_i is either a clique or an independent set in G'. Moreover, P_i is a clique in G' then $d_G(Q_{i,k}, Q_{i,\ell}) \geq \frac{1}{2}$ for every $1 \leq k < \ell \leq t$, and if P_i is an independent set in G' then $d_G(Q_{i,k}, Q_{i,\ell}) < \frac{1}{2}$ for every $1 \leq k < \ell \leq t$.

(b) For every pair $1 \leq i < j \leq r$, if there is an edge in G' between P_i and P_j then $d_G(Q_i, Q_j) \geq \frac{\varepsilon}{4}$. Then by Item 7 of Lemma 5.3.1 we have that $d_G(Q_{i,k}, Q_{j,\ell}) \geq \frac{\varepsilon}{4} - \frac{1}{\Psi(|X|+r)} \geq \frac{\varepsilon}{8}$ for every $1 \leq k, \ell \leq t$. Analogously, if there is a non-edge in G' between P_i and P_j then $d_G(Q_i, Q_j) \leq 1 - \frac{\varepsilon}{4}$, which implies (by Item 7 of Lemma 5.3.1) that $d_G(Q_{i,k}, Q_{j,\ell}) \leq 1 - \frac{\varepsilon}{4} + \frac{1}{\Psi(|X|+r)} \leq 1 - \frac{\varepsilon}{8}$ for every $1 \leq k, \ell \leq t$.

Now let K be the following embedding scheme: $A_K = X$ and $B_K = \{b_1, \ldots, b_r\}$; for each $1 \leq i \leq r$, vertex b_i is colored black if P_i is a clique in G' and white if P_i is an independent set in G'; for each $x, x' \in X$, edge $\{x, x'\}$ is colored black if $\{x, x'\} \in E(G)$ and white if $\{x, x'\} \notin E(G)$; for each $x \in X$, $1 \leq i \leq r$, edge $\{x, b_i\}$ is colored black if the bipartite graph between x and P_i is complete and white if this bipartite

graph is empty (Item 3 in Lemma 5.3.1 implies that one of these options must hold); finally, for every $1 \leq i < j \leq r$, edge $\{b_i, b_j\}$ is colored black if the bipartite graph between P_i and P_j is complete in G', white if the bipartite graph between P_i and P_j is empty in G', and grey otherwise.

Observe that the map $\varphi : X \cup Y \to V(K)$ which maps x to itself (for every $x \in X = A_K$) and P_i to b_i (for every $1 \leq i \leq r$), is an embedding from $G'[X \cup Y]$ to K. Since $|V(K)| = |X| + r$, we have $K \in \mathcal{F}_m$ for $m := |X| + r$. By the definition of the function $\Psi_{\mathcal{F}}$ (see Definition 5.3.3), there is $F \in \mathcal{F}$ such that $F \to K$ and $|V(F)| \leq \Psi_{\mathcal{F}}(m) = \Psi_{\mathcal{F}}(|X| + r) \leq \Psi(|X| + r) = t$.

Now, fixing an embedding ρ from F to K, write $W_i := \rho^{-1}(b_i) = \{w_{i,1}, \ldots, w_{i,f_i}\}$ for $1 \leq i \leq r$. Put $W = W_1 \cup \cdots \cup W_r$ and $H = F[W]$. We claim that the sets $(Q_{i,k})_{1 \leq i \leq r, 1 \leq k \leq f_i}$ satisfy the requirements 1-2 in Lemma 5.2.4 with respect to $h = |V(F)| \leq \Psi_{\mathcal{F}}(m)$, $\eta = \frac{\varepsilon}{8}$ and H as above, *in the graph* G. In other words, we show that one can apply Lemma 5.2.4 with the sets $U_1, \ldots, U_h$ being $(Q_{i,k})_{1 \leq i \leq r, 1 \leq k \leq f_i}$, and with G as the host graph. We actually already proved that Item 1 in Lemma 5.2.4 holds; indeed, this follows from the fact that $F \to K$, the definition of the embedding scheme K, and Items (a)-(b) above. Item 2 of Lemma 5.2.4 follows from Items 6-7 of Lemma 5.3.1, which together imply that for every $1 \leq i \leq j \leq r$ and $1 \leq k \leq f_i, 1 \leq \ell \leq f_j$ (with the exception of $(i,k) = (j,\ell)$), the pair $(Q_{i,k}, Q_{j,\ell})$ is δ-regular with $\delta = \frac{1}{\Psi(m)} \leq \delta_{5.2.4}(\Psi_{\mathcal{F}}(m), \frac{\varepsilon}{8}) \leq \delta_{5.2.4}(h, \frac{\varepsilon}{8})$, as required.

We thus showed that Lemma 5.2.4 is applicable to the tuple of sets $(Q_{i,k})_{1 \leq i \leq r, 1 \leq k \leq f_i}$ and the graph $H = F[W]$ (with the parameters defined above). Let $\mathcal{U}$ be the set of all tuples $(u_{i,k})_{1 \leq i \leq r, 1 \leq k \leq f_i}$, where $u_{i,k} \in Q_{i,k}$, which induce (in G) a copy of $H = F[W]$ in which $u_{i,k}$ plays the role of $w_{i,k}$ for every $1 \leq i \leq r$ and $1 \leq k \leq f_i$. By Lemma 5.2.4, we have

$$\sum_{(u_{i,k})_{i,k} \in \mathcal{U}} \prod_{i=1}^{r} \prod_{k=1}^{f_i} \mathcal{D}(u_{i,k}) \geq \delta_{5.2.4}\left(h, \frac{\varepsilon}{8}\right) \cdot \prod_{i=1}^{r} \prod_{k=1}^{f_i} \mathcal{D}(U_{i,k}) \geq \delta_{5.2.4}\left(\Psi_{\mathcal{F}}(m), \frac{\varepsilon}{8}\right) \cdot S^{-|W|}, \tag{5.5}$$

where in the last inequality we used the guarantees of Item 8 in Lemma 5.3.1 and the monotonicity of the function $\delta_{5.2.4}$. Observe that for every $(u_{i,k})_{i,k} \in \mathcal{U}$, the subgraph of G induced by the vertex-set $X \cup \{u_{i,k} : 1 \leq i \leq r, 1 \leq k \leq f_i\}$ contains an induced copy of F. Indeed, this follows from the definition of $\mathcal{U}$, the fact that $F \to K$, and the definition of the embedding scheme K. Now sample an $(|X| + |W|)$-tuple of vertices from G according to the distribution $\mathcal{D}$ and independently. Note that if every vertex in X appears in the first $|X|$ vertices of the sample, and if the tuple of the last $|W|$ vertices of the sample belongs to $\mathcal{U}$, then the subgraph induced by the sample contains an induced copy of F and hence does not satisfy $\mathcal{P}$ (as $F \in \mathcal{F}$). The probability for this event is at least

$$\delta_{5.2.4}\left(\Psi_{\mathcal{F}}(m), \frac{\varepsilon}{8}\right) \cdot S^{-|X|-|W|} .$$

Here we used (5.5) and Item 2 in Lemma 5.3.1. Next, note that $|X| + |W| \leq |X| + rt \leq S$, where in the last inequality we used Items 2 and 8 of Lemma 5.3.1. Similarly, $\Psi_{\mathcal{F}}(m) \leq t \leq S$. So we see that a sample of S random vertices induces a graph which does not satisfy $\mathcal{P}$ with probability at least $\delta_{5.2.4}\left(S, \frac{\varepsilon}{8}\right) \cdot S^{-S}$. Therefore, a sample of $s = s_{\mathcal{P}}(\varepsilon)$ vertices (see (5.4)) induces a graph not satisfying $\mathcal{P}$ with probability at least

$$1 - \left(1 - \delta_{5.2.4}\left(S, \frac{\varepsilon}{8}\right) \cdot S^{-S}\right)^{s/S} = 1 - \left(1 - \delta_{5.2.4}\left(S, \frac{\varepsilon}{8}\right) \cdot S^{-S}\right)^{\frac{2S^S}{\delta_{5.2.4}(S, \frac{\varepsilon}{8})}} \geq 1 - e^{-2} \geq \frac{2}{3},$$

as required. This completes the proof. ■

It is natural to ask about the dependence on ε of the sample complexity of the tester supplied by Theorem 13. One answer is that one cannot prove any upper bound on the sample complexity which will hold uniformly for all properties $\mathcal{P}$, because it was shown in [12] that no such bound exists even in the standard model. Suppose then that one is interested only in "simple" properties such as induced H-freeness (for some fixed H). In this case, it is not too hard to see that although we are iterating Lemma 5.2.11, which has wowzer-type bounds[10] in this setting even for unweighted graphs (see [34, 76]), we are still getting "only" a wowzer-type bound in Theorem 13. We should also point that it might be possible to use the ideas in [34], together with those presented here, in order to get tower-type bounds on the sample complexity of testing induced H-freeness in the VDF model.

5.4 Proof of the "Only-If" Part of Theorem 13

The proof of the "only-if" part of Theorem 13 is divided between Propositions 5.4.1 and 5.4.2. As shown in [58], we can (and will) always assume that a VDF tester only queries the input graph on pairs of vertices which it has sampled.

Proposition 5.4.1. *If a graph property $\mathcal{P}$ is not extendable, then $\mathcal{P}$ is not testable in the VDF model.*

Proof. Since $\mathcal{P}$ is not extendable, there is a graph $G_1 \in \mathcal{P}$, such that no $(|V(G_1)|+1)$-vertex graph satisfying $\mathcal{P}$ contains G_1 as an induced subgraph. Let G_2 be a graph obtained from G_1 by adding a "new" vertex v (and putting an arbitrary bipartite graph between v and $V(G_1)$), let $\mathcal{D}_1$ be the uniform distribution on $V(G_1)$, and let $\mathcal{D}_2$ be the distribution on $V(G_2)$ which assigns weight $\frac{1}{|V(G_1)|}$ to each $u \in V(G_1) \subseteq V(G_2)$ and weight[11] 0 to v.

It is clear that for every integer q, a sample of q vertices from G_1 according to $\mathcal{D}_1$ is indistinguishable from a sample of q vertices from G_2 according to $\mathcal{D}_2$. Observe that G_1 satisfies $\mathcal{P}$ while $(G_2, \mathcal{D}_2)$ is $\frac{1}{|V(G_1)|^2}$-far from $\mathcal{P}$. To see that the latter statement is true, observe that by our choice of G_1, no matter how we change the bipartite graph between v and $V(G_1)$, we will always get a graph that does not satisfy $\mathcal{P}$. Hence, in order to make G_2 satisfy $\mathcal{P}$, one must change the adjacency relation between a pair of vertices from $V(G_1)$, whose weight (under $\mathcal{D}_2$) is $\frac{1}{|V(G_1)|}$. Now, the fact that $(G_1, \mathcal{D}_1)$ and $(G_2, \mathcal{D}_2)$ are indistinguishable implies that $\mathcal{P}$ is not testable[12] in the VDF model. ■

Proposition 5.4.2. *If a graph property $\mathcal{P}$ is not hereditary, then $\mathcal{P}$ is not testable in the VDF model.*

[10] To be precise, we mean here that the "standard" way of establishing Lemma 5.2.11 (which is also the way we prove this lemma here) is via the strong regularity lemma (see Lemma 5.2.9), which is known to only give wowzer-type bounds [34, 76]. In [34], (an unweighted variant of) Lemma 5.2.11 was proved without the use of the strong regularity lemma, thus giving better, tower-type, bounds. This is alluded to in the following sentence.

[11] Evidently, if one does not wish to allow vertices of weight 0, then one can instead assign to v a weight tending to 0; or, more accurately, a weight that is small enough with respect to (the inverse of) the sample complexity of an alleged tester for $\mathcal{P}$ (in a proof by contradiction that such a tester does not exist).

[12] We note that if $\mathcal{P}$ is non-extendable but hereditary, then one can easily obtain infinitely many examples showing that $\mathcal{P}$ is not testable (rather than just the one example given in the proof of Proposition 5.4.1). Indeed, instead of adding just one vertex to G_1, one can add to G_1 any number k of vertices (for a large k), and give these new vertices weight $o(1/k)$, while distributing the remaining weight uniformly among the vertices of G_1 (note that such an assignment is precisely what the NLW VDF model, defined in Section 5.5 forbids). The assumption that $\mathcal{P}$ is hereditary implies that every graph obtained in this way is $\frac{1-o(1)}{|V(G_1)|^2}$-far from satisfying $\mathcal{P}$.

Proof. Since $\mathcal{P}$ is not hereditary, there is a graph G_1 and an induced subgraph G_2 of G_1, such that G_1 satisfies $\mathcal{P}$ but G_2 does not. Let $\mathcal{D}_2$ be the uniform distribution on $V(G_2)$, and let $\mathcal{D}_1$ be the distribution on $V(G_1)$ which is supported on $V(G_2) \subseteq V(G_1)$ and uniform when conditioned on $V(G_2)$, i.e. $\mathcal{D}_1(u) = \frac{1}{|V(G_2)|}$ if $u \in V(G_2)$ and $\mathcal{D}_1(u) = 0$ if $u \in V(G_1) \setminus V(G_2)$. Clearly, for every integer q, a sample of q vertices from G_1 according to $\mathcal{D}_1$ is indistinguishable from a sample of q vertices from G_2 according to $\mathcal{D}_2$. Also, G_1 satisfies $\mathcal{P}$, whereas $(G_2, \mathcal{D}_2)$ is $\frac{1}{|V(G_2)|^2}$-far from $\mathcal{P}$ because $G_2 \notin \mathcal{P}$. Thus, $\mathcal{P}$ is not testable[13] in the VDF model. ■

5.5 On Variations of the VDF Model and Related Problems

Let us give the precise definitions of the variations of the VDF model that we consider here.

The "large inputs" model In this model, a property $\mathcal{P}$ is testable if there exists a function $M_{\mathcal{P}} : (0,1) \to \mathbb{N}$ such that for every $\varepsilon > 0$, $\mathcal{P}$ is ε-testable with sample complexity depending only on ε under the promise that inputs $(G, \mathcal{D})$ always satisfy $|V(G)| \geq M_{\mathcal{P}}(\varepsilon)$

The "size-aware" model In this model, testers are allowed to receive, as part of the input, the number of vertices of the input graph.

The "no heavy-weights" (NHW) model In this model, a property $\mathcal{P}$ is testable if there exists a function $c_{\mathcal{P}} : (0,1) \to (0,1)$ such that for every $\varepsilon > 0$, $\mathcal{P}$ is ε-testable with sample complexity depending only on ε under the promise that inputs $(G, \mathcal{D})$ always satisfy $\max_{v \in V(G)} \mathcal{D}(v) \leq c_{\mathcal{P}}(\varepsilon)$.

The "no light-weights" (NLW) model In this model, a property $\mathcal{P}$ is testable if for all $\varepsilon, \delta > 0$, $\mathcal{P}$ is ε-testable with sample complexity depending only on ε and δ under the promise that inputs $(G, \mathcal{D})$ always satisfy $\min_{v \in V(G)} \mathcal{D}(v) \geq \delta / |V(G)|$.

Note that if $\mathcal{P}$ is testable in the "large inputs" model then it is also testable in the NHW model (because by setting $c_{\mathcal{P}}(\varepsilon) := 1/M_{\mathcal{P}}(\varepsilon)$ we can make sure that the input graph has at least $M_{\mathcal{P}}(\varepsilon)$ vertices). Still, we decided to handle the NHW model separately (instead of deducing the results for this model from their counterparts for the "large inputs" model). This decision is due to two reasons: one is that in the course of studying the NHW model, we answer another question raised in [58]; and the other is that our proof that every hereditary property $\mathcal{P}$ is testable in the NHW model shows that $\mathcal{P}$ is testable in this model by a tester that accepts if and only if the subgraph induced by the sample satisfies $\mathcal{P}$. In contrast, the tester for $\mathcal{P}$ in the "large inputs" model is not always of this form.

In Sections 5.5.1 and 5.5.2 we show that every hereditary graph property is testable in each of the four models defined above.

[13] In analogy to Footnote 12, we note that if $\mathcal{P}$ is non-hereditary but extendable, then one can obtain infinitely many examples showing that $\mathcal{P}$ is not testable (rather than just the one given in the proof of Proposition 5.4.2). Indeed, the extendability of $\mathcal{P}$ implies that there are arbitrarily large graphs which satisfy $\mathcal{P}$ and contain G_1 (and hence also G_2) as an induced subgraph. Each of these graphs (together with an appropriate distribution, as in the proof of Proposition 5.4.2) is a witness to the non-testability of $\mathcal{P}$.

5.5.1 Every Hereditary Property is Testable in the "Large Inputs", "Size-Aware" and NLW Models

Let us introduce some definitions that we will use throughout Section 5.5.1. Let $\mathcal{P}$ be a hereditary graph property. A graph F is called *$\mathcal{P}$-good* if for every $r \geq |V(F)|$ there is an r-vertex graph which satisfies $\mathcal{P}$ and contains F as an induced subgraph; this in particular implies that F itself satisfies $\mathcal{P}$. If F is not $\mathcal{P}$-good then it is called *$\mathcal{P}$-bad*, and we denote by $r_{\mathcal{P}}(F)$ the minimal $r \geq |V(F)|$ such that there is no r-vertex graph which satisfies $\mathcal{P}$ and contains F as an induced subgraph. In particular, if F does not satisfy $\mathcal{P}$ then it is $\mathcal{P}$-bad and $r_{\mathcal{P}}(F) = |V(F)|$. Note that since $\mathcal{P}$ is hereditary, if F is $\mathcal{P}$-bad then there is no graph on r vertices for any $r \geq r_{\mathcal{P}}(F)$ which satisfies $\mathcal{P}$ and contains F as an induced subgraph.

Now let $\mathcal{H} = \mathcal{H}(\mathcal{P})$ be the property of being $\mathcal{P}$-good. Then $\mathcal{H} \subseteq \mathcal{P}$ and $\mathcal{H}$ is hereditary, which follows from the definition of $\mathcal{P}$-goodness and the fact that $\mathcal{P}$ is hereditary. Observe moreover that $\mathcal{H}$ is extendable (in fact, if $\mathcal{P}$ itself is extendable then $\mathcal{H} = \mathcal{P}$). For an integer $s \geq 1$, let $R_{\mathcal{P}}(s)$ be the maximum of $r_{\mathcal{P}}(F)$ over all $\mathcal{P}$-bad graphs F with at most s vertices; if no such graphs exist, we set $R_{\mathcal{P}}(s) = 0$ (this will not matter later on). We are now ready to prove that every hereditary property is testable in the "large inputs" model. This is done in the following proposition.

Proposition 5.5.1. *For every hereditary property $\mathcal{P}$ there are functions $M_{\mathcal{P}}, s_{\mathcal{P}} : (0,1) \to \mathbb{N}$ such that for every $\varepsilon > 0$, the property $\mathcal{P}$ is ε-testable with one-sided error and sample complexity $s_{\mathcal{P}}(\varepsilon)$ under the promise that inputs $(G, \mathcal{D})$ always satisfy $|V(G)| \geq M_{\mathcal{P}}(\varepsilon)$ vertices*

Proof. Consider the (hereditary and extendable) property $\mathcal{H} = \mathcal{H}(\mathcal{P})$ defined above. By Theorem 5.1.1, there is a function $s_{\mathcal{H}} : (0,1) \to \mathbb{N}$ such that for every $\varepsilon > 0$ and for every vertex-weighted graph $(G, \mathcal{D})$ which is ε-far from $\mathcal{H}$, a sample of s vertices from G (taken from $\mathcal{D}$) induces a subgraph which does not satisfy $\mathcal{H}$ with probability at least $\frac{2}{3}$.

Our ("large inputs"-model) tester for $\mathcal{P}$ samples $s_{\mathcal{H}}(\varepsilon)$ vertices, and accepts if and only if the subgraph induced by the sample satisfies $\mathcal{H}$. We prove the lemma with $M = M_{\mathcal{P}}(\varepsilon) := R_{\mathcal{P}}(s_{\mathcal{H}}(\varepsilon))$.

Let $(G, \mathcal{D})$ be a vertex-weighted graph with $|V(G)| \geq M$. Suppose first that G satisfies $\mathcal{P}$. Our goal is to show that the subgraph induced by a sample of $s_{\mathcal{H}}(\varepsilon)$ vertices, taken from $\mathcal{D}$ and independently, satisfies $\mathcal{H}$ with probability 1. So suppose by contradiction that G contains an induced subgraph F on at most $s_{\mathcal{H}}(\varepsilon)$ vertices which does not satisfy $\mathcal{H}$. In other words, F is $\mathcal{P}$-bad. By the definition of $r_{\mathcal{P}}(F)$, there is no graph on $r_{\mathcal{P}}(F)$ vertices which satisfies $\mathcal{P}$ and contains F as an induced subgraph. As $|V(G)| \geq M = R_{\mathcal{P}}(s_{\mathcal{H}}(\varepsilon)) \geq r_{\mathcal{P}}(F)$, and as $\mathcal{P}$ is hereditary, we get that G does not satisfy $\mathcal{P}$, a contradiction.

Suppose now that $(G, \mathcal{D})$ is ε-far from $\mathcal{P}$. Then $(G, \mathcal{D})$ is also ε-far from $\mathcal{H}$, as $\mathcal{H} \subseteq \mathcal{P}$. By our choice of $s_{\mathcal{H}}(\varepsilon)$ via Theorem 5.1.1, a sample of $s_{\mathcal{H}}(\varepsilon)$ vertices of G, taken from $\mathcal{D}$ and independently, does not satisfy $\mathcal{H}$ with probability at least $\frac{2}{3}$. So our tester rejects $(G, \mathcal{D})$ with probability at least $\frac{2}{3}$, as required. ■

It is natural to ask whether we can replace the function $M_{\mathcal{P}}(\varepsilon)$ in Proposition 5.5.1 by a constant depending only on $\mathcal{P}$ (and not on ε). As is shown in the following proposition, we cannot.

Proposition 5.5.2. *There is a hereditary property $\mathcal{P}$ such that for every $M > 0$, there is no tester for $\mathcal{P}$ in the VDF model even under the promise that input graphs always have at least M vertices.*

Proof. For each $k \geq 3$, let C_k^* be the graph obtained from the k-cycle C_k by adding an isolated vertex. Consider the property $\mathcal{P} = \{C_k^* : k \geq 3\}$-freeness. Let $M > 0$. Set $G = C_M$ and $G' = C_M^*$. Let $\mathcal{D}$ be

the uniform distribution on $V(G)$, and let $\mathcal{D}'$ be the distribution on $V(G')$ which assigns weight 0 to the isolated vertex in G', and is uniform on the rest of the vertices of G'. Then $G \in \mathcal{P}$ and $(G', \mathcal{D}')$ is $\frac{1}{M^2}$-far from $\mathcal{P}$, but a sample (of any number of vertices) from $(G, \mathcal{D})$ is indistinguishable from a sample of the same size from $(G', \mathcal{D}')$. This shows that $\mathcal{P}$ is not testable even under the promise that input graphs have at least M vertices. ■

We now move on to consider the "size-aware" model.

Proposition 5.5.3. *Every hereditary property is testable with one-sided error in the "size-aware" model.*

Proof. Let $\mathcal{P}$ be a hereditary property. Our goal is to design (and prove the correctness of) a one-sided-error tester for $\mathcal{P}$ in the VDF model, under the promise that $|V(G)|$ is given to the tester as part of the input. Let $M_{\mathcal{P}} : (0,1) \to \mathbb{N}$ be as in Proposition 5.5.1. On input $\varepsilon \in (0,1)$ and $(G, \mathcal{D})$, our tester for $\mathcal{P}$ in the "size-aware" model works as follows:

1. If $|V(G)| \geq M_{\mathcal{P}}(\varepsilon)$, then invoke the tester whose existence is guaranteed by Proposition 5.5.1, and accept if and only if this tester accepts.

2. Otherwise, i.e. if $|V(G)| < M_{\mathcal{P}}(\varepsilon)$, then do the following: setting $M := M_{\mathcal{P}}(\varepsilon)$ and $t := M \log(3M)/\varepsilon$, sample vertices $u_1, \ldots, u_t \in V(G)$ according to $\mathcal{D}$ and independently, and put $U := \{u_1, \ldots, u_t\}$. Accept if and only if there exists a graph on $|V(G)|$ vertices which satisfies $\mathcal{P}$ and contains $G[U]$ as an induced subgraph (in the notation introduced at the beginning of Section 5.5.1, this is the same as saying that $r_{\mathcal{P}}(G[U]) > |V(G)|$).

Let us prove the correctness of our tester. First, Proposition 5.5.1 guarantees that if $|V(G)| \geq M_{\mathcal{P}}(\varepsilon)$ then the tester works correctly; namely, it accepts with probability 1 if $G \in \mathcal{P}$, and rejects with probability at least $\frac{2}{3}$ if $(G, \mathcal{D})$ is ε-far from $\mathcal{P}$.

So from now on we may assume that $|V(G)| < M_{\mathcal{P}}(\varepsilon)$. Suppose first that $G \in \mathcal{P}$. Evidently, for every $U \subseteq V(G)$ there is a graph on $|V(G)|$ vertices which satisfies $\mathcal{P}$ and contains $G[U]$ as an induced subgraph (indeed, G is such a graph). Hence, the tester accepts G with probability 1, as required.

Suppose now that $(G, \mathcal{D})$ is ε-far from $\mathcal{P}$. Observe that for each $v \in V(G)$, the probability that $v \notin U$ is

$$(1 - \mathcal{D}(v))^t \leq e^{-\mathcal{D}(v) \cdot t} = \left(\frac{1}{3M}\right)^{-\mathcal{D}(v) \cdot M/\varepsilon}.$$

Hence, if $\mathcal{D}(v) \geq \varepsilon/M$ then the probability that $v \notin U$ is at most $\frac{1}{3M}$. By the union bound, the probability that there is $v \in V(G) \setminus U$ with $\mathcal{D}(v) \geq \varepsilon/M$ is at most $\frac{1}{3}$. Suppose then that every $v \in V(G) \setminus U$ satisfies $\mathcal{D}(v) < \varepsilon/M$ (as we just showed, this happens with probability at least $\frac{2}{3}$). Then $\mathcal{D}(V(G) \setminus U) < |V(G)| \cdot \varepsilon/M < \varepsilon$ (here we used our assumption that $|V(G)| < M$). Now, if (by contradiction) there were a graph G' on $|V(G)|$ vertices which satisfied $\mathcal{P}$ and contained $G[U]$ as an induced subgraph, then one could turn G into G' by only adding/deleting edges which are incident to vertices in $V(G) \setminus U$. Since $\mathcal{D}(V(G) \setminus U) < \varepsilon$, this would stand in contradiction to the assumption that $(G, \mathcal{D})$ is ε-far from $\mathcal{P}$. We conclude that there is no such graph G'. This implies that $(G, \mathcal{D})$ is rejected with probability at least $\frac{2}{3}$, as required. ■

Finally, we prove that every hereditary property is testable in the NLW model. This is done in the following proposition.

Proposition 5.5.4. *For every hereditary property $\mathcal{P}$ there is a function $t_{\mathcal{P}} : (0,1)^2 \to \mathbb{N}$ such that for all $\varepsilon, \delta > 0$, the property $\mathcal{P}$ is ε-testable with one-sided error and sample complexity $t_{\mathcal{P}}(\varepsilon, \delta)$ under the promise that inputs $(G, \mathcal{D})$ always satisfy $\min_{v \in V(G)} \mathcal{D}(v) \geq \delta/|V(G)|$.*

Proof. We start by specifying the function $t_{\mathcal{P}}(\varepsilon, \delta)$. Consider the (extendable and hereditary) property $\mathcal{H} = \mathcal{H}(\mathcal{P})$ defined above. By Theorem 5.1.1, there is a function $s_{\mathcal{H}} : (0,1) \to \mathbb{N}$ such that for every $\varepsilon > 0$ and for every vertex-weighted graph $(G, \mathcal{D})$ which is ε-far from $\mathcal{H}$, a sample of $s_{\mathcal{H}}(\varepsilon)$ vertices of G (taken from $\mathcal{D}$) induces a subgraph which does not satisfy $\mathcal{H}$ with probability[14] at least $\frac{5}{6}$. Now set $R := R_{\mathcal{P}}(s_{\mathcal{H}}(\varepsilon))$ and

$$t = t_{\mathcal{P}}(\varepsilon, \delta) := \max\left\{s_{\mathcal{H}}(\varepsilon),\ 2R\log(6R)/\delta\right\}.$$

Our tester for $\mathcal{P}$ in the NLW model simply samples a sequence of $t_{\mathcal{P}}(\varepsilon, \delta)$ vertices of the input and accepts if and only if the subgraph induced by the sample satisfies $\mathcal{P}$. Evidently, this tester accepts with probability 1 if the input satisfies $\mathcal{P}$. So to establish the correctness of our tester, it suffices to show that it rejects with probability at least $\frac{2}{3}$ if the input $(G, \mathcal{D})$ is ε-far from $\mathcal{P}$.

Let $\varepsilon > 0$ and let $(G, \mathcal{D})$ be a vertex-weighted graph on n vertices which is ε-far from $\mathcal{P}$, and in which all vertices have weight at least δ/n. Let $u_1, \ldots, u_t$ be a sequence of $t = t_{\mathcal{P}}(\varepsilon, \delta)$ random vertices of G, taken according to $\mathcal{D}$ and independently, and set $U = \{u_1, \ldots, u_t\}$. We need to show that with probability at least $\frac{2}{3}$, $G[U]$ does not satisfy $\mathcal{P}$. Suppose first that $n < 2R$. We claim that in this case we have $U = V(G)$ with probability at least $\frac{2}{3}$ (this is clearly sufficient because G itself does not satisfy $\mathcal{P}$). For a vertex $v \in V(G)$, the probability that $u_i \neq v$ for every $1 \leq i \leq t$ is

$$(1 - \mathcal{D}(v))^t \leq \left(1 - \frac{\delta}{n}\right)^t < \left(1 - \frac{\delta}{2R}\right)^t \leq e^{-\frac{\delta t}{2R}} \leq \frac{1}{6R}.$$

So by the union bound over all $n < 2R$ vertices of G, we see that with probability at least $\frac{2}{3}$, $U = V(G)$, as required.

Suppose now that $n \geq 2R$. Our choice of $s = s_{\mathcal{H}}(\varepsilon)$ guarantees that with probability at least $\frac{5}{6}$, the graph $F := G[\{u_1, \ldots, u_s\}]$ does not satisfy $\mathcal{H}$, meaning that it is $\mathcal{P}$-bad. We will now show that with probability at least $\frac{5}{6}$, we have $|U| \geq R$. This will imply that with probability at least $\frac{2}{3}$, $G[U]$ contains as an induced subgraph a $\mathcal{P}$-bad graph F on at most $s_{\mathcal{H}}(\varepsilon)$ vertices, and also $|U| \geq R = R_{\mathcal{P}}(s_{\mathcal{H}}(\varepsilon)) \geq r_{\mathcal{P}}(F)$. By the definition of $r_{\mathcal{P}}(F)$, this in turn will imply that $G[U]$ does not satisfy $\mathcal{P}$, as required.

So from now on, our goal is to show that $|U| \geq R$ with probability at least $\frac{5}{6}$. Fix a partition of $V(G)$ into R sets $V_1, \ldots, V_R$, each of size at least $\lfloor \frac{n}{R} \rfloor \geq \frac{n}{2R}$. For each $1 \leq i \leq R$, let A_i be the event that $U \cap V_i \neq \emptyset$. Note that if A_i occurs for every $1 \leq i \leq R$, then $|U| \geq R$. Since $\mathcal{D}(V_i) \geq |V_i| \cdot \frac{\delta}{n} \geq \frac{n}{2R} \cdot \frac{\delta}{n} = \frac{\delta}{2R}$, the probability that A_i does not occur is at most

$$(1 - \mathcal{D}(V_i))^t \leq \left(1 - \frac{\delta}{2R}\right)^t \leq e^{-\frac{\delta t}{2R}} \leq \frac{1}{6R}.$$

By the union bound, the probability that there is $1 \leq i \leq R$ for which A_i does not occur, is at most $\frac{1}{6}$, as required. This completes the proof. ■

[14] The statement of Theorem 5.1.1 only guarantees a success probability of $\frac{2}{3}$, but this can clearly be amplified to $\frac{5}{6}$ by repeating the experiment $O(1)$ times.

5.5.2 Every Hereditary Property is Testable in the NHW Model

Proposition 5.5.5. *For every hereditary property $\mathcal{P}$ there are functions $t_{\mathcal{P}} : (0,1) \to \mathbb{N}$ and $c_{\mathcal{P}} : (0,1) \to (0,1)$ such that for every $\varepsilon > 0$, the property $\mathcal{P}$ is ε-testable with one-sided error and sample complexity $t_{\mathcal{P}}(\varepsilon)$ under the promise that inputs $(G, \mathcal{D})$ always satisfy $\max_{v \in V(G)} \mathcal{D}(v) \leq c_{\mathcal{P}}(\varepsilon)$.*

The key idea in the proof of Proposition 5.5.5, which appeared in [58], is to "blow up" the vertex-weighted input graph $(G, \mathcal{D})$ by replacing each vertex v with a vertex-set whose size is proportional to $\mathcal{D}(v)$, and thus obtain an (unweighted) graph G', to which one can apply known testability results in the standard model. To this end, let us introduce some definitions. For a graph G, say on $V(G) = \{v_1, \ldots, v_n\}$, and for integers $b_1, \ldots, b_n \geq 0$, a $(b_1, \ldots, b_n)$-*blowup* of G is any graph admitting a vertex-partition $V_1 \cup \cdots \cup V_n$ such that $|V_i| = b_i$ for every $1 \leq i \leq n$, and such that the bipartite graph between V_i and V_j is complete if $\{v_i, v_j\} \in E(G)$ and empty if $\{v_i, v_j\} \notin E(G)$. The sets $V_1, \ldots, V_n$ are called the *blowup-sets*. Note that we do not pose any restrictions on the graphs induced by the sets $V_1, \ldots, V_n$; these graphs may be arbitrary.

For the rest of this section, as well as in Section 5.5.3, we will assume that all vertex-weights are rational, as this will make for cleaner results. Then, at the end of Section 5.5.3, we will detail the changes that need to be made in order to handle general (i.e., possibly irrational) weights.

Let G be a graph, let $\mathcal{D}$ be a distribution on $V(G) = \{v_1, \ldots, v_n\}$, and let $N \in \mathbb{N}$ be such that $\mathcal{D}(v_i) \cdot N$ is an integer for every $1 \leq i \leq n$; such an N is called *suitable*. A $(\mathcal{D}, N)$-*blowup* of G is a $(b_1, \ldots, b_n)$-blowup of G with $b_i = \mathcal{D}(v_i) \cdot N$ for every $1 \leq i \leq n$. Note that a blowup is always treated as "unweighted" (in other words, the distribution on its vertices is uniform). Goldreich [58] proved that for every graph F and $\varepsilon \in (0,1)$, if a vertex-weighted graph $(G, \mathcal{D})$ is ε-far from being F-free, then for every suitable N, any $(\mathcal{D}, N)$-blowup of G is $\frac{\varepsilon}{\binom{|V(F)|}{2}}$-far from being F-free. Goldreich further asked whether the $\binom{|V(F)|}{2}^{-1}$-factor can be avoided. In the following lemma we show that this is indeed the case, and moreover that an analogous statement holds for *every* hereditary property. This lemma is also the key ingredient in the proof of Proposition 5.5.5.

Lemma 5.5.6. *Let $\mathcal{P}$ be a hereditary graph property and let $(G, \mathcal{D})$ be a vertex-weighted graph which is ε-far from $\mathcal{P}$. Then for every suitable N, any $(\mathcal{D}, N)$-blowup of G is ε-far from $\mathcal{P}$.*

Proof. Fix any suitable N and let G' be a $(\mathcal{D}, N)$-blowup of G. As above, we use $v_1, \ldots, v_n$ to denote the vertices of G, and $V_1, \ldots, V_n$ to denote the corresponding blowup sets. Suppose by contradiction that there is a graph H' on $V(G')$ that satisfies $\mathcal{P}$ and is ε-close to G'. Let H be the random graph defined as follows: the vertex-set of H is $V(H) = V(G) = \{v_1, \ldots, v_n\}$. To define the edge-set of H, sample for each $1 \leq i \leq n$ a vertex $u_i \in V_i$ uniformly at random, and make $\{v_i, v_j\}$ an edge in H if and only if $\{u_i, u_j\}$ is an edge in H' (for $1 \leq i < j \leq n$). Then H satisfies $\mathcal{P}$ (with probability 1) because H is isomorphic to an induced subgraph of H' and $\mathcal{P}$ is hereditary. Let us compute the expected distance between H and G (here the distance is with respect to the distribution $\mathcal{D}$). For each $1 \leq i < j \leq n$, the probability that $\{v_i, v_j\} \in E(G) \triangle E(H)$ is precisely

$$\frac{|E_{G'}(V_i, V_j) \triangle E_{H'}(V_i, V_j)|}{|V_i||V_j|} = \frac{|E_{G'}(V_i, V_j) \triangle E_{H'}(V_i, V_j)|}{\mathcal{D}(v_i)\mathcal{D}(v_j)N^2} .$$

Hence, the expected distance between H and G is

$$\sum_{1\leq i<j\leq n} \mathcal{D}(v_i)\mathcal{D}(v_j)\cdot\frac{|E_{G'}(V_i,V_j)\triangle E_{H'}(V_i,V_j)|}{\mathcal{D}(v_i)\mathcal{D}(v_j)N^2} =$$
$$\frac{1}{N^2}\sum_{1\leq i<j\leq n}|E_{G'}(V_i,V_j)\triangle E_{H'}(V_i,V_j)| \leq \varepsilon,$$

where the last inequality uses the assumption that G' is ε-close to H'. So G is ε-close to a graph H which satisfies $\mathcal{P}$, a contradiction. ■

Discussion By combining Lemma 5.5.6 with the result of [10] that all hereditary properties are testable with one-sided error in the standard model, we obtain the following: for every hereditary property $\mathcal{P}$, for every vertex-weighted graph $(G,\mathcal{D})$ which is ε-far from $\mathcal{P}$, for every suitable N and for every $(\mathcal{D},N)$-blowup G' of G, it holds that G' is ε-far from $\mathcal{P}$ with respect to the uniform distribution, and hence a sample of some $s=s_{\mathcal{P}}(\varepsilon)$ vertices of G', taken uniformly and independently, induces a graph which w.h.p. does not satisfy $\mathcal{P}$. Observe that this induced subgraph of G' has (essentially) the same distribution as the graph S on $[s]$ obtained by sampling vertices $u_1,\ldots,u_s\in V(G)$ from $\mathcal{D}$ independently, and letting $\{i,j\}\in E(S)$ if and only if $\{u_i,u_j\}\in E(S)$. Note that S is precisely the graph defined in Theorem 5.1.2. We thus established Theorem 5.1.2, as promised in Section 5.1.

As noted in Section 5.1, the graph S defined above is a *blowup* of an induced subgraph of G, but is not necessarily a subgraph of G in itself. This is because the sequence $u_1,\ldots,u_s$ might contain several vertices which belong to the same blowup-set. In other words, it may be the case that G' contains "forbidden subgraphs" which use several vertices from one of the blowup-sets, and thus do not correspond to "forbidden subgraphs" in G. This creates an obstacle for proving Proposition 5.5.5, because in order to prove this proposition we need to know that a (suitably chosen) random induced subgraph of G (and not just the blowup thereof) does not satisfy $\mathcal{P}$ w.h.p. Avoiding this obstacle is precisely the reason for the assumption that all vertices in G have relatively small weight. This assumption guarantees that it is unlikely to sample more than once from one of the blowup-sets (or, in other words, that S is isomorphic to $G[\{u_1,\ldots,u_s\}]$.). We note that a different way of dealing with this obstacle is to restrict ourselves to properties for which we can guarantee, by appropriately choosing the graphs inside the blowup-sets, that there would not be any minimal forbidden subgraph which uses several vertices from one of the blowup-sets, see Section 5.5.3.

Proof of Proposition 5.5.5. We start by specifying the functions $t_{\mathcal{P}}$ and $c_{\mathcal{P}}$. By the main result of [10], there is a function $q_{\mathcal{P}}:(0,1)\to\mathbb{N}$ such that for every $\varepsilon>0$ and for every (unweighted) graph G which is ε-far from $\mathcal{P}$, a sample of $q_{\mathcal{P}}(\varepsilon)$ vertices from G, taken *uniformly at random* and independently, induces a graph which does not satisfy $\mathcal{P}$ with probability at least $\frac{5}{6}$. Now set $t_{\mathcal{P}}(\varepsilon):=q_{\mathcal{P}}(\varepsilon)$ and

$$c_{\mathcal{P}}(\varepsilon):=\frac{1}{3q_{\mathcal{P}}^2(\varepsilon)}.$$

Our tester for $\mathcal{P}$ in the NHW model simply samples a sequence of $t=t_{\mathcal{P}}(\varepsilon)$ vertices of the input and accepts if and only if the subgraph induced by the sample satisfies $\mathcal{P}$. Evidently, this tester accepts with probability 1 if the input satisfies $\mathcal{P}$. So to establish the correctness of our tester, it suffices to show that it rejects with probability at least $\frac{2}{3}$ if the input $(G,\mathcal{D})$ is ε-far from $\mathcal{P}$.

Let $\varepsilon > 0$ and let $(G, \mathcal{D})$ be a vertex-weighted graph on n vertices which is ε-far from $\mathcal{P}$, and in which all vertices have weight at most c, where $c = c_{\mathcal{P}}(\varepsilon)$. Write $V(G) = \{v_1, \ldots, v_n\}$ and fix a positive integer N such that $\mathcal{D}(v_i) \cdot N$ is an integer for every $1 \leq i \leq n$. Let G' be an arbitrary $(\mathcal{D}, N)$-blowup of G, and denote the blowup-sets by $V_1, \ldots, V_n$. By Lemma 5.5.6, G' is ε-far from $\mathcal{P}$. This implies that a random sequence $u_1, \ldots, u_q$ of $q = q_{\mathcal{P}}(\varepsilon)$ vertices of G', sampled uniformly and independently, induces a graph which does not satisfy $\mathcal{P}$ with probability at least $\frac{5}{6}$.

Let $\varphi : V(G') \to V(G)$ be the map which maps all elements of V_i to v_i (for every $1 \leq i \leq n$). Observe that for $u \in V(G')$ sampled uniformly, the random vertex $\varphi(u) \in V(G)$ has the distribution $\mathcal{D}$ (because $|V_i| = \mathcal{D}(v_i) \cdot N = \mathcal{D}(v_i) \cdot |V(G')|$). Furthermore, if a set $U \subseteq V(G')$ satisfies $|V_i \cap U| \leq 1$ for every $1 \leq i \leq n$, then $G[\varphi(U)]$ is isomorphic to $G'[U]$. Let $u_1, \ldots, u_q$ be a random sequence of vertices of G', sampled uniformly and independently, and set $U := \{u_1, \ldots, u_q\}$. Recall that $G'[U]$ does not satisfy $\mathcal{P}$ with probability at least $\frac{5}{6}$. Furthermore, the probability that $|V_i \cap U| \geq 2$ for some $1 \leq i \leq n$ is at most

$$\sum_{i=1}^{n} \binom{q}{2} \cdot \mathcal{D}^2(v_i) \leq \frac{q^2}{2} \cdot c \cdot \sum_{i=1}^{n} \mathcal{D}(v_i) = \frac{q^2}{2} \cdot c = \frac{1}{6}.$$

We conclude that with probability at least $\frac{2}{3}$, $G'[U]$ does not satisfy $\mathcal{P}$ and $|V_i \cap U| \leq 1$ for every $1 \leq i \leq n$, implying that $G[\varphi(U)]$ does not satisfy $\mathcal{P}$ either. This completes the proof. ■

It is natural to ask whether the function $c_{\mathcal{P}}(\varepsilon)$ appearing in Proposition 5.5.5 needs to depend on ε, namely whether the statement of this proposition holds even if $c_{\mathcal{P}}$ is a constant function (depending only on $\mathcal{P}$). It follows from Proposition 5.5.2, however, that this is not the case. In other words, allowing $c_{\mathcal{P}}(\varepsilon)$ to depend on ε is unavoidable.

5.5.3 Testing in the VDF Model vs. Testing in the Standard Model

It is natural to ask about the relation between the sample complexity for testing a property in the VDF model and the sample complexity for testing it in the standard model. More specifically, it will be interesting to resolve the following:

Problem 5.5.7. *Is it true that every hereditary and extendable property $\mathcal{P}$ can be tested in the VDF model with the same (or close to the same) sample complexity as in the (standard) dense graph model?*

While at present we cannot answer this question, we can show that many natural properties $\mathcal{P}$ can be tested in the VDF model with (exactly) the same sample complexity as that of the (optimal) tester for $\mathcal{P}$ in the standard model, which works by sampling a certain number of vertices and accepting if and only if they induce a graph which satisfies $\mathcal{P}$. This is explained in the following paragraph.

As mentioned in Section 5.5.2, the assumption made in Proposition 5.5.5 regarding the non-existence of high-weight vertices is needed in order to handle the possibility of having copies of some (forbidden) graph F in G' which do not correspond to copies of F in G. For some graph properties, however, such an assumption is not required, as we can make sure that every copy of a minimal forbidden graph in G' will correspond to such a copy in G. To make this precise, we need the following definition. A family of graphs $\mathcal{F}$ is said to be *blowup-avoidable* if for every graph G, say on $\{v_1, \ldots, v_n\}$, and for every n-tuple of integers $b_1, \ldots, b_n \geq 0$, there is a $(b_1, \ldots, b_n)$-blowup G' of G with blowup-sets $V_1, \ldots, V_n$, such that there is no induced copy of any $F \in \mathcal{F}$ in G' which intersects some V_i in at least 2 vertices; in other words, for

every $F \in \mathcal{F}$, every induced copy of F in G' corresponds to an induced copy of F in G. We say that a hereditary property $\mathcal{P}$ is *blowup-avoidable* if the family of minimal forbidden induced subgraphs for $\mathcal{P}$ is blowup-avoidable. We now prove the following proposition, which partially resolves Problem 5.5.7. The proof is similar to that of Proposition 5.5.5.

Proposition 5.5.8. *Let $\mathcal{P}$ be a hereditary property which is blowup-avoidable, and suppose that $\mathcal{P}$ admits a standard-model tester which works by sampling $q_{\mathcal{P}}(\varepsilon)$ vertices uniformly at random and independently, and accepting if and only if the subgraph induced by the sample satisfies $\mathcal{P}$. Then (assuming all vertex-weights are rational) $\mathcal{P}$ is testable in the VDF model by a tester having one-sided error and sample complexity $q_{\mathcal{P}}(\varepsilon)$.*

Proof. Given an input $(G, \mathcal{D})$, our VDF tester for $\mathcal{P}$ works by sampling (from $\mathcal{D}$) a sequence of $q_{\mathcal{P}}(\varepsilon)$ vertices, and accepting if and only if the subgraph induced by the sample satisfies $\mathcal{P}$. Since $\mathcal{P}$ is hereditary, this tester accepts with probability 1 if the input graph satisfies $\mathcal{P}$. So it remains to show that if the input $(G, \mathcal{D})$ is ε-far from $\mathcal{P}$, then with probability at least $\frac{2}{3}$, a sequence of $q_{\mathcal{P}}(\varepsilon)$ vertices of G, sampled according to $\mathcal{D}$ and independently, induces a graph which does not satisfy $\mathcal{P}$.

Let $\mathcal{F} = \mathcal{F}(\mathcal{P})$ be the family of minimal forbidden induced subgraphs for $\mathcal{P}$. Let $(G, \mathcal{D})$ be a vertex-weighted graph on n vertices, which is ε-far from $\mathcal{P}$. Write $V(G) = \{v_1, \ldots, v_n\}$ and fix a positive integer N such that $\mathcal{D}(v_i) \cdot N$ is an integer for every $1 \leq i \leq n$. As $\mathcal{P}$ is blowup-avoidable, there is a $(\mathcal{D}, N)$-blowup G' of G with blowup-sets $V_1, \ldots, V_n$, such that there is no induced copy of any $F \in \mathcal{F}$ in G' which intersects some V_i in at least 2 vertices. By Lemma 5.5.6, G' is ε-far from $\mathcal{P}$. So by our choice of $q_{\mathcal{P}}(\varepsilon)$, with probability at least $\frac{2}{3}$ it holds that a sequence of $q_{\mathcal{P}}(\varepsilon)$ vertices of G', sampled uniformly and independently, induces a graph which does not satisfy $\mathcal{P}$, and hence contains an induced copy of some $F \in \mathcal{F}$.

Let $\varphi : V(G') \to V(G)$ be the map which maps all elements of V_i to v_i (for every $1 \leq i \leq n$). Observe that for $u \in V(G')$ sampled uniformly, the random vertex $\varphi(u) \in V(G)$ has the distribution $\mathcal{D}$. Note that by our choice of G', if $u_1, \ldots, u_r \in V(G')$ span an induced copy of some $F \in \mathcal{F}$ (in the graph G'), then $\varphi|_{\{u_1,\ldots,u_r\}}$ is injective (and hence an isomorphism), which implies that $\varphi(u_1), \ldots, \varphi(u_r)$ span an induced copy of F in G. It is now easy to see that a sequence of $q_{\mathcal{P}}(\varepsilon)$ vertices of G, sampled from $\mathcal{D}$ and independently, does not satisfy $\mathcal{P}$ with probability at least $\frac{2}{3}$, as required. ■

Discussion To demonstrate the usefulness of Proposition 5.5.8, observe that induced F-freeness is blowup-avoidable for every $F \in \{P_3, P_4, C_4\}$; indeed, this is established by taking the blowup-sets (in the definition of blowup-avoidability) to be cliques. (Here P_k denotes the path with k vertices.) By combining Proposition 5.5.8 with results for the standard model, namely Theorems 4 and 7, we immediately get that induced F-freeness is testable in the VDF model with sample complexity $\mathrm{poly}(1/\varepsilon)$ if $F \in \{P_3, P_4\}$, and with sample complexity at most $2^{\mathrm{poly}(1/\varepsilon)}$ if $F = C_4$.

Let us describe another corollary of Proposition 5.5.8. We say that a graph property $\mathcal{P}$ is *closed under blowups* if for every graph G satisfying $\mathcal{P}$, every blowup of G in which the blowup-sets are independent also satisfies $\mathcal{P}$. We claim that if a hereditary property $\mathcal{P}$ is closed under blowups then it is also blowup-avoidable. To see this, let $\mathcal{F}$ be the set of minimal forbidden induced subgraphs for $\mathcal{P}$, let G be an n-vertex graph, let $b_1, \ldots, b_n \geq 0$ be integers and let G' be the $(b_1, \ldots, b_n)$-blowup of G in which the blowup-sets, $V_1, \ldots, V_n$, are independent. Let $F \in \mathcal{F}$ and suppose that G' contains an induced copy of F. If, by contradiction, this copy intersects some V_i in more than one vertex, then F is a blowup of some graph F' with $|V(F')| < |V(F)|$, where the blowup-sets are independent sets. Since $\mathcal{P}$ is closed under blowups

and $F \notin \mathcal{P}$, we must have $F' \notin \mathcal{P}$; but this contradicts the fact that F is a minimal forbidden induced subgraph for $\mathcal{P}$. So we see that the conclusion of Proposition 5.5.8 applies to hereditary properties which are closed under blowups. Some examples of such properties include K_t-freeness; the property of having a homomorphism into a fixed graph H (and in particular the property of being k-colorable); and the property of being the blowup of a fixed graph H (cf. [18]).

On the negative side, there are many natural hereditary properties which are extendable but not blowup-avoidable, such as the property of being H-free for a graph H which is neither a clique nor contains isolated vertices. It would be interesting to resolve Problem 5.5.7 for these properties.

Irrational weights Let us consider the case where we allow the distribution $\mathcal{D}$ to assign general (i.e., not necessarily rational) weights. We use a simple (but somewhat long to outline) argument of approximating irrational weights by rational ones. Let V be a finite set. Observe that for every distribution $\mathcal{D}$ on V and for every $\delta > 0$, there is a distribution $\mathcal{D}'$ on V assigning rational weights and satisfying $|\mathcal{D}'(v) - \mathcal{D}(v)| \leq \delta$ for every $v \in V$. Now, let $u_1, \ldots, u_q \in V$ (resp. $u'_1, \ldots, u'_q \in V$) be random vertices sampled independently from $\mathcal{D}$ (resp. $\mathcal{D}'$). It is easy to see that the distributions of $(u_1, \ldots, u_q)$ and $(u'_1, \ldots, u'_q)$ have total variation distance at most $q\delta$. Indeed, for every sequence of vertices $x_1, \ldots, x_q \in V$,

$$\begin{aligned}
\left|\mathbb{P}[u'_1 = x_1, \ldots, u'_q = x_q] - \mathbb{P}[u_1 = x_1, \ldots, u_q = x_q]\right| &= \left|\prod_{i=1}^{q} \mathcal{D}'(x_i) - \prod_{i=1}^{q} \mathcal{D}(x_i)\right| \\
&= \left|\sum_{j=1}^{q} \big(\mathcal{D}'(x_j) - \mathcal{D}(x_j)\big) \prod_{i=1}^{j-1} \mathcal{D}(x_i) \prod_{i=j+1}^{q} \mathcal{D}'(x_i)\right| \\
&\leq \sum_{j=1}^{q} \left|\mathcal{D}'(x_j) - \mathcal{D}(x_j)\right| \leq q\delta.
\end{aligned}$$

It follows that if G_1, G_2 are graphs on V, then $|\mathrm{dist}_{\mathcal{D}'}(G_1, G_2) - \mathrm{dist}_{\mathcal{D}}(G_1, G_2)| \leq \delta$, because $\mathrm{dist}_{\mathcal{D}}(G_1, G_2)$ equals one half the probability that random vertices $u, v \in V$ sampled (independently) according to $\mathcal{D}$ satisfy $\{u, v\} \in E(G_1) \triangle E(G_2)$, and similarly for $\mathrm{dist}_{\mathcal{D}'}(G_1, G_2)$.

Now suppose that $\mathcal{T}$ is an $\frac{\varepsilon}{2}$-tester for some property $\mathcal{P}$, which has success probability at least $\frac{5}{6}$ under the promise that all weights of the input distribution are rational. Suppose further that $\mathcal{T}$ works by sampling q vertices (according to the given distribution) and inspecting the graph which they span; recall that in our setting, all testers operate in this manner. We observe that $\mathcal{T}$ is in fact a valid ε-tester with success probability $\frac{2}{3}$ even if we allow irrational weights. Indeed, given an input $(G, \mathcal{D})$ to $\mathcal{T}$, simply consider the above distribution $\mathcal{D}'$ with $\delta := \min\{\frac{\varepsilon}{2}, \frac{1}{6q}\}$, noting that if $(G, \mathcal{D})$ is ε-far from $\mathcal{P}$, then $(G, \mathcal{D}')$ is $\frac{\varepsilon}{2}$-far from $\mathcal{P}$ (by the above discussion). Furthermore, the probability that $\mathcal{T}$ accepts $(G, \mathcal{D})$ deviates by at most $q\delta \leq \frac{1}{6}$ from the probability that $\mathcal{T}$ accepts $(G, \mathcal{D}')$, because $q\delta$ is an upper bound on the total variation distance between the distributions of q random vertices sampled from $\mathcal{D}$ and $\mathcal{D}'$, respectively. So we see that $\mathcal{T}$ is indeed a valid ε-tester (even with irrational weights allowed). It follows that Propositions 5.5.5 and 5.5.8 also holds for general weights, where in Proposition 5.5.8 we need to slightly increase the sample complexity to (say) $q_{\mathcal{P}}(\varepsilon/2)$ (in Proposition 5.5.5 this increase can be absorbed by the function t).

5.5.4 Which Properties are Testable in the Variations of the VDF Model?

It may be interesting to characterize the graph properties which are testable in each of the four variations of the VDF model (defined at the beginning of Section 5.5). We sometimes call these the *restricted models*.

Problem 5.5.9. *Which graph properties are testable in the "large inputs"/"size-aware"/NHW/NLW model?*

While at the moment we are unable to resolve Problem 5.5.9, we can rule out some initial guesses. A first guess might be that only hereditary properties are testable in these models. This, however, turns out to be false; for example, connectivity and hamiltonicity are testable in each of these models, as implied by the following proposition.

Proposition 5.5.10. *Let $\mathcal{P}$ be a graph property such that for every $\varepsilon > 0$ there is $M(\varepsilon)$ so that every vertex-weighted graph on at least $M(\varepsilon)$ vertices is ε-close to $\mathcal{P}$. Then $\mathcal{P}$ is testable in all four variations of the VDF models.*

Proof. The fact that $\mathcal{P}$ is testable in the "large inputs" (resp. NHW) model is trivial; indeed, by choosing $M_{\mathcal{P}}(\varepsilon) := M(\varepsilon)$ (resp. $c_{\mathcal{P}}(\varepsilon) := 1/M(\varepsilon)$) we can make sure that every input graph will be ε-close to $\mathcal{P}$, so a tester that always accepts without making any queries is a valid tester for $\mathcal{P}$.

Let us now consider the NLW model. Given $\varepsilon, \delta > 0$ and an input graph $(G, \mathcal{D})$ with all vertex-weights at least $\delta/|V(G)|$, our tester for $\mathcal{P}$ works as follows: setting $M := M(\varepsilon)$, the tester samples $t := M \log(3M)/\delta$ vertices according to $\mathcal{D}$ and independently; if the number of distinct vertices in the sample is at least M then the tester accepts (without making any queries), and otherwise the tester accepts if and only if the subgraph induced by the sample satisfies $\mathcal{P}$. To see that this is a valid tester, observe that if G has less than M vertices, then with probability at least $\frac{2}{3}$, the tester samples all of the vertices (this follows from our choice of t and the fact that every vertex has weight at least $\delta/|V(G)| > \delta/M$). And if G has at least M vertices then $(G, \mathcal{D})$ is ε-close to $\mathcal{P}$ by assumption.

Finally, let us prove that $\mathcal{P}$ is testable in the "size-aware" model. In this model, our tester for $\mathcal{P}$ works as follows. On input $\varepsilon > 0$ and $(G, \mathcal{D})$, the tester distinguishes between the cases $|V(G)| \geq M(\varepsilon)$ and $|V(G)| < M(\varepsilon)$. In the former case, the tester accepts without making any queries, and in the latter case, the tester samples $t := M \log(3M)/\varepsilon$ vertices $u_1, \ldots, u_t \in V(G)$ according to $\mathcal{D}$ and independently (where $M = M(\varepsilon)$), and accepts if and only if there is a graph on $|V(G)|$ vertices which satisfies $\mathcal{P}$ and contains $G[\{u_1, \ldots, u_t\}]$ as an induced subgraph. The proof of correctness for this tester is similar to the proof of Proposition 5.5.3, and we leave the details to the reader. ■

In order to apply Proposition 5.5.10 to the properties of connectivity and hamiltonicity, we observe that any vertex-weighted graph $(G, \mathcal{D})$ with $|V(G)| \geq 1/\varepsilon$ is ε-close to being hamiltonian (and hence also connected). To see that this holds, take a random (cyclic) ordering $v_1, \ldots, v_n$ of the vertices of G, and observe that for every pair of distinct $u, w \in V(G)$, the probability that there is $1 \leq i \leq n$ such that $\{u, w\} = \{v_i, v_{i+1}\}$ is $n/\binom{n}{2} = \frac{2}{n-1}$. This implies that the expected value of $\sum_{i=1}^{n} \mathcal{D}(v_i)\mathcal{D}(v_{i+1})$ is $\frac{2}{n-1} \cdot \sum_{u,w \in V(G)} \mathcal{D}(u)\mathcal{D}(w) = \frac{2}{n-1} \cdot \frac{1}{2} \cdot \left(1 - \sum_{v \in V(G)} \mathcal{D}(v)^2\right) \leq \frac{1}{n-1} \cdot \left(1 - \frac{1}{n}\right) = \frac{1}{n}$, where the last inequality follows from Cauchy-Schwarz (and the first sum is over *unordered* pairs $\{u, w\}$). Let us also note that for connectivity there is a simpler argument: if $(G, \mathcal{D})$ is a vertex-weighted graph with $|V(G)| \geq 1/\varepsilon$, then there is $v \in V(G)$ with $\mathcal{D}(v) \leq \varepsilon$, and we can make G connected by connecting v to all other vertices.

Note that in some of the restricted models (e.g. the NLW model), the tester given by (the proof of) Proposition 5.5.10 has 2-sided error. It is also not hard to see that the NLW model admits no 1-sided-error tester for, e.g., connectivity. This shows that (some of) the restricted models allow for properties which are testable with 2-sided error but not with 1-sided error (unlike the "ordinary" VDF model, where we know that every testable property can be tested with 1-sided error, as follows from Theorems 13 and 5.1.1; see also [58, Theorem 2.3]).

Another natural guess regarding the answer to Problem 5.5.9 would be that every property which is testable in the standard model is also testable in the restricted models. (See [7] for a characterization of the properties testable in the standard model.) This guess is ruled out by the following proposition, that describes a property which is testable in the standard model but not in any of the restricted models. Here we somewhat diverge from the type of properties considered so far; while up to now we only considered properties which are testable in the standard model *with 1-sided error*, the property described in Proposition 5.5.11 requires 2-sided error (for testing in the standard model).

Proposition 5.5.11. *The property $\mathcal{P}$ of having edge-density*[15] *at most $\frac{1}{4}$ is not testable in either of the four variations of the VDF model.*

Proof. Let G_1 be the n-vertex graph consisting of a clique of size $\frac{n}{2}$ and $\frac{n}{2}$ isolated vertices, and let $\mathcal{D}_1$ be the uniform distribution on $V(G_1)$. Let G_2 be the n-vertex graph consisting of a clique X of size $\frac{3n}{4}$ and $\frac{n}{4}$ isolated vertices, and let $\mathcal{D}_2$ be the distribution on $V(G_2)$ that assigns weight $\frac{2}{3n}$ to every vertex of X, and weight $\frac{2}{n}$ to every vertex of $V(G_2) \setminus X$. Note that $(G_1, \mathcal{D}_1)$ and $(G_2, \mathcal{D}_2)$ are valid inputs in each of the four variations of the VDF model (provided that n is large enough), and that G_1 satisfies $\mathcal{P}$ while $(G_2, \mathcal{D}_2)$ is $\Omega(1)$-far from $\mathcal{P}$. On the other hand, we now show that for every q, a sample of q vertices from $(G_1, \mathcal{D}_1)$ is indistinguishable from a sample of q vertices from $(G_2, \mathcal{D}_2)$ (provided that n is large enough with respect to q). To this end, let U_i be a set of q random vertices of G_i sampled according to $\mathcal{D}_i$ and independently (for $i = 1, 2$). Then for both $i = 1, 2$, the graph $G_i[U_i]$ consists of a clique and some isolated vertices. Letting X_i be the clique in $G_i[U_i]$, we have

$$\begin{aligned}\mathbb{P}[|X_1| = k] &= o(1) + \binom{q}{k} \cdot \prod_{i=0}^{k-1} \left(\frac{n}{2} - i\right) \cdot \prod_{i=0}^{q-k-1} \left(\frac{n}{2} - i\right) \cdot \left(\frac{1}{n}\right)^q \\ &= (1 + o(1))\binom{q}{k} \cdot \left(\frac{1}{2}\right)^q,\end{aligned}$$

$$\begin{aligned}&\mathbb{P}[|X_2| = k] = \\ &o(1) + \binom{q}{k} \cdot \prod_{i=0}^{k-1} \left(\frac{3n}{4} - i\right) \cdot \prod_{i=0}^{q-k-1} \left(\frac{n}{4} - i\right) \cdot \left(\frac{2}{3n}\right)^k \cdot \left(\frac{2}{n}\right)^{q-k} = \\ &(1 + o(1))\binom{q}{k} \cdot \left(\frac{1}{2}\right)^q,\end{aligned}$$

where in both cases, the additive term $o(1)$ accounts for the event that some vertex has been sampled more than once. So we see that $|\mathbb{P}[|X_1| = k] - \mathbb{P}[|X_2| = k]| = o(1)$. This implies that the total variation distance

[15] The edge-density of a (possibly vertex-weighted) graph G is defined as $2e(G)/|V(G)|^2$; in other words, the density is defined with respect to the uniform distribution on $V(G)$; it ignores the given distribution $\mathcal{D}$.

between the distribution of $G_1[U_1]$ and the distribution of $G_2[U_2]$ is $o(1)$. It follows that $\mathcal{P}$ is not testable in any of the four variations of the VDF model (note that knowing n does not help to distinguish between $(G_1, \mathcal{D}_1)$ and $(G_2, \mathcal{D}_2)$, since these graphs have the same number of vertices). ■

The proof of Proposition 5.5.11 can be adapted to show that other properties which are testable in the standard model are not testable in any of the four variations of the VDF model. These include the property of having a cut with at least αn^2 edges (for $0 < \alpha < \frac{1}{4}$) and the property of containing a clique with at least αn vertices (for $0 < \alpha < 1$). Again, all these properties require 2-sided error (in the standard model).

The proof of Proposition 5.5.11 relies on the fact that, by definition, the VDF models do not allow for uniform samples. It would be interesting to study which properties are testable in the setting where one can sample vertices both according to the uniform distribution and according to the given distribution $\mathcal{D}$; see [58, Section 4] for some results in this direction.

Chapter 6

Testing Linear Inequalities of Subgraph Statistics

In this chapter we prove Theorems 16 and 18. Recall that $\Pi_{h,w,b}$ denotes the property of all graphs G satisfying $\sum_H w_H \cdot p(H,G) \leq b$, where $p(H,G)$ is the induced density of H in G; H runs over all graphs with h vertices (for some fixed h); and b and $(w_H)_H$ are positive rationals. We remind the reader that Theorem 16 asserts that there is a choice of (h,w,b) for which $\Pi_{h,w,b}$ is not testable, while Theorem 18 states that $\Pi_{h,w,b}$ has a size-oblivious POT only if the triple of parameters (h,w,b) has the so-called removal property (see Definition 17).

6.1 Proof of Theorem 16

Let $\Pi_{h,w,b}$ be as in the statement of Theorem 16. Our goal is to show that $\Pi_{h,w,b}$ is not testable. Before delving into the details of the proof, let us give a rough outline of it. The main idea behind the proof is to show that $\Pi_{h,w,b}$ encodes the property of being *quasirandom* with edge density $\frac{1}{2}$ (for the definition of a quasirandom graph, see the paragraph below Fact 6.1.3). More precisely, we show that if a graph G satisfies $\Pi_{h,w,b}$, then its edge density must be roughly $1/2$ and its C_4 density roughly $1/16$, which is known to imply that G is quasirandom (see [33]).

Then, to show that $\Pi_{h,w,b}$ is not testable, we argue as follows. We fix an n-vertex graph G which satisfies $\Pi_{h,w,b}$, let Γ be the N/n-blowup of G (for n, N to be chosen later), and let $V_1, \ldots, V_n$ be the parts of this blowup, corresponding to the vertices of G. Now, every pair of parts V_i, V_j forms either a complete or an empty bipartite graph in Γ, which means that Γ cannot be $\frac{1}{n}$-quasirandom (with density $\frac{1}{2}$). It follows that in order to turn Γ into a quasirandom graph, one must make many changes in all bipartite graphs (V_i, V_j). Therefore, Γ is $\Omega(1)$-far from being quasirandom, and hence also $\Omega(1)$-far from $\Pi_{h,w,b}$.

Now, fix an N-vertex graph Γ^* which satisfies $\Pi_{h,w,b}$. We use the counting lemma (see Lemma 6.1.4) to argue that as G and Γ^* are quasirandom and Γ is a blowup of G, the small-subgraphs statistics of Γ and Γ^* are roughly the same, meaning that a tester which makes few samples cannot distinguish between them. Since n, N can be chosen arbitrarily large, $\Pi_{h,w,b}$ does not have a tester whose sample complexity is independent of the size of the input.

We now fill in the details of the above rough plan. As a first step towards proving the theorem, we

give a different description of $\Pi_{h,w,b}$ in terms of injective densities of edges and 4-cycles, see Lemma 6.1.2 below. First we need to introduce some notation. For a graph G, denote

$$z(G) := \sum_{H:|V(H)|=4} w_H \cdot p(H,G),$$

where the weights w_H are defined in the statement of Theorem 16. Under this notation, we have $\Pi_{h,w,b} = \{G : z(G) \leq b\}$. It will be convenient to denote $n^{\underline{h}} := n \cdot (n-1) \cdot \cdots \cdot (n-h+1)$. For a pair of graphs H and G, define

$$t_{\text{inj}}(H,G) = \frac{1}{n^{\underline{h}}} |\{\varphi \colon V(H) \to V(G) \text{ injective s.t. } uv \in E(H) \Rightarrow \varphi(u)\varphi(v) \in E(G)\}|,$$

and

$$t_{\text{ind}}(H,G) = \frac{1}{n^{\underline{h}}} |\{\varphi \colon V(H) \to V(G) \text{ injective s.t. } uv \in E(H) \Leftrightarrow \varphi(u)\varphi(v) \in E(G)\}|.$$

Note that $t_{\text{ind}}(H,G) = p(H,G) \cdot \text{aut}(H)/h!$, where $\text{aut}(H)$ is the number of automorphisms of H. Let us recall the following basic property of subgraph densities:

Fact 6.1.1. *For every pair of graphs F, G and $h \geq v(F)$, it holds that $p(F,G) = \sum_H p(F,H) \cdot p(H,G)$, where the sum is over all h-vertex graphs H.*

The following lemma gives a simpler description of the property $\Pi_{h,w,b}$.

Lemma 6.1.2. $\Pi_{h,w,b} = \{G : \phi(G) \leq 0\}$, *where* $\phi(G) = 2t_{inj}(C_4,G) - t_{inj}(K_2,G) + \frac{3}{8}$.

Proof. First, note that $t_{inj}(K_2,G) = p(K_2,G)$. Now, as C_4, D_4, K_4 are the only 4-vertex graphs containing C_4 as a subgraph, we have

$$\begin{aligned} t_{\text{inj}}(C_4,G) &= t_{\text{ind}}(C_4,G) + 2t_{\text{ind}}(D_4,G) + t_{\text{ind}}(K_4,G) \\ &= \frac{\text{aut}(C_4)}{4!} \cdot p(C_4,G) + 2 \cdot \frac{\text{aut}(D_4)}{4!} \cdot p(D_4,G) + \frac{\text{aut}(K_4)}{4!} \cdot p(K_4,G) \\ &= \frac{1}{3}p(C_4,G) + \frac{1}{3}p(D_4,G) + p(K_4,G). \end{aligned}$$

Plugging the above into the definition of $\phi(G)$, we get:

$$\begin{aligned} \phi(G) &= \frac{2}{3}p(C_4,G) + \frac{2}{3}p(D_4,G) + 2p(K_4,G) - p(K_2,G) + \frac{3}{8} \\ &= \frac{2}{3}p(C_4,G) + \frac{2}{3}p(D_4,G) + 2p(K_4,G) + p(\overline{K_2},G) - \frac{5}{8} \\ &= \frac{2}{3}p(C_4,G) + \frac{2}{3}p(D_4,G) + 2p(K_4,G) + \sum_{H:|V(H)|=4} p(\overline{K_2},H)p(H,G) - \frac{5}{8} \\ &= \sum_{H:|V(H)|=4} 2w_H \cdot p(H,G) - \frac{5}{8}. \end{aligned}$$

Here, the penultimate inequality uses Fact 6.1.1. So we see that $\phi(G) \leq 0$ holds if and only if $\sum_{H:|V(H)|=4} w_H \cdot p(H,G) \leq 5/16$, namely if and only if $G \in \Pi_{h,w,b}$, as required. ■

We will need the following well-known fact, which is closely related to the well-known Kovári-Sós-Turán theorem [78]. For a proof of this fact, see e.g. [2, Lemma 2.1].

Fact 6.1.3. *Every n-vertex graph G satisfies*[1] $t_{inj}(C_4, G) \geq t_{inj}(K_2, G)^4 - O\left(\frac{1}{n}\right)$.

We now give some background on quasirandomness. For a thorough overview of the subject, we refer the reader to [79]. In what follows, we write $x = y \pm z$ to mean that $x \in [y - z, y + z]$. An n-vertex graph G is *δ-quasirandom* (with density $\frac{1}{2}$) if for every pair of disjoint sets $U, V \subseteq V(G)$ such that $|U|, |V| \geq \delta n$, it holds that $e(U, V) = \left(\frac{1}{2} \pm \delta\right)|U||V|$.

For a family of graphs $\mathcal{F}$ and a graph G, define $p(\mathcal{F}, G) := \sum_{F \in \mathcal{F}} p(F, G)$. The well-known *counting lemma* states that a quasirandom graph has approximately the same distribution of small subgraphs as a random graph with the same density. Here we use the following variant (see e.g. [33]).

Lemma 6.1.4 (Counting lemma)**.** *For every $s \geq 2$ and $\varepsilon > 0$ there is $\delta = \delta(s, \varepsilon) > 0$ such that for every family $\mathcal{F}$ of s-vertex graphs and for every δ-quasirandom graph G, it holds that*

$$p(\mathcal{F}, G) = \sum_{F \in \mathcal{F}} 2^{-\binom{s}{2}} \frac{s!}{aut(F)} \pm \varepsilon.$$

Note that for each s-vertex graph F, the quantity $2^{-\binom{s}{2}} \frac{s!}{\mathrm{aut}(F)}$ is just the expected value of $p(F, G(n, \frac{1}{2}))$.

The following seminal result of Chung, Graham and Wilson [33] states that quasirandomness essentially boils down to having the "right" densities of edges and 4-cycles.

Theorem 6.1.5 ([33])**.** *For every $\delta \in (0, 1)$ there are $\gamma = \gamma(\delta)$ and $n_0 = n_0(\delta)$ such that if a graph G on at least n_0 vertices satisfies*

$$t_{inj}(K_2, G) = \frac{1}{2} \pm \gamma \quad \text{and} \quad t_{inj}(C_4, G) \leq \frac{1}{16} + \gamma, \tag{6.1}$$

then G is δ-quasirandom.

An important ingredient in the proof of Theorem 16 is the following lemma, which shows that graphs that satisfy $\Pi_{h,w,b}$ must be quasirandom.

Lemma 6.1.6. *For every $\delta \in (0, 1)$ there is $n_0(\delta)$ such that every graph $G \in \Pi_{h,w,b}$ on $n \geq n_0(\delta)$ vertices is δ-quasirandom.*

Proof. In light of Theorem 6.1.5, it is enough to show that for every $\gamma \in (0, 1)$ there is $n_0(\gamma)$ such that every $G \in \Pi_{h,w,b}$ on $n \geq n_0(\gamma)$ vertices satisfies (6.1). So fix any $\gamma \in (0, 1)$, and let $G \in \Pi_{h,w,b}$ be a graph on $n \geq n_0$ vertices (for n_0 to be chosen later). By Fact 6.1.3, we have

$$t_{\mathrm{inj}}(C_4, G) \geq t_{\mathrm{inj}}(K_2, G)^4 - O\left(\frac{1}{n}\right) \geq t_{\mathrm{inj}}(K_2, G)^4 - \frac{\gamma^2}{2}, \tag{6.2}$$

where the last inequality holds if n is large enough (as a function of γ). So we see that

$$2t_{\mathrm{inj}}(K_2, G)^4 - t_{\mathrm{inj}}(K_2, G) + \frac{3}{8} \leq 2t_{\mathrm{inj}}(C_4, G) + \gamma^2 - t_{\mathrm{inj}}(K_2, G) + \frac{3}{8} = \phi(G) + \gamma^2 \leq \gamma^2, \tag{6.3}$$

[1] Usually this inequality is stated in terms of the *homomorphism density*, as $\mathrm{t}(C_4, G) \geq \mathrm{t}(K_2, G)^4$ (in fact, this is the form in which Fact 6.1.3 appears in [2]). The error-term $O(\frac{1}{n})$ accounts for the difference between the homomorphism density and the injective density, see [79, Section 5.2.3].

where the first inequality follows from (6.2), and the second inequality follows from Lemma 6.1.2. Note that the function $x \mapsto 2x^4 - x + \frac{3}{8}$ is convex, and attains its minimum at $x = 1/2$. Therefore, if we had $t_{\text{inj}}(K_2, G) > \frac{1}{2} + \gamma$, then we would have

$$2t_{\text{inj}}(K_2,G)^4 - t_{\text{inj}}(K_2,G) + \frac{3}{8} > 2\left(\frac{1}{2}+\gamma\right)^4 - \left(\frac{1}{2}+\gamma\right) + \frac{3}{8} = 2\gamma^4 + 4\gamma^3 + 3\gamma^2 > \gamma^2.$$

Similarly, if we had $t_{\text{inj}}(K_2, G) < \frac{1}{2} - \gamma$, then we would have

$$2t_{\text{inj}}(K_2,G)^4 - t_{\text{inj}}(K_2,G) + \frac{3}{8} > 2\left(\frac{1}{2}-\gamma\right)^4 - \left(\frac{1}{2}-\gamma\right) + \frac{3}{8} = 2\gamma^4 - 4\gamma^3 + 3\gamma^2 > \gamma^2.$$

In any case, we see that $|t_{\text{inj}}(K_2, G) - \frac{1}{2}| > \gamma$ would stand in contradiction to (6.3). Hence, $t_{\text{inj}}(K_2, G) = \frac{1}{2} \pm \gamma$. By using again the fact that $\phi(G) \leq 0$ (see Lemma 6.1.2), we get that

$$t_{\text{inj}}(C_4,G) \leq \frac{t_{\text{inj}}(K_2,G)}{2} - \frac{3}{16} \leq \frac{1}{4} + \frac{\gamma}{2} - \frac{3}{16} < \frac{1}{16} + \gamma.$$

We have thus shown that (6.1) holds, as required. ■

By combining Lemmas 6.1.6 and 6.1.4, we immediately obtain the following corollary.

Corollary 6.1.7. *For every $s \geq 2$ and $\varepsilon \in (0,1)$ there is $n_1 = n_1(s,\varepsilon)$ such that every $G \in \Pi_{h,w,b}$ on $n \geq n_1$ vertices satisfies the following. For every family $\mathcal{F}$ of s-vertex graphs, it holds that*

$$p(\mathcal{F},G) = \sum_{F\in\mathcal{F}} 2^{-\binom{s}{2}} \frac{s!}{aut(F)} \pm \varepsilon.$$

We are now ready to prove Theorem 16.

Proof of Theorem 16. We start by showing that $\Pi_{h,w,b}$ is non-empty. More specifically, we prove that for every integer $n \geq 4$, there exists an n-vertex graph satisfying $\Pi_{h,w,b}$. Let $G \sim G(n, \frac{1}{2})$. It is easy to see that $\mathbb{E}[t_{\text{inj}}(K_2, G)] = \frac{1}{2}$ and $\mathbb{E}[t_{\text{inj}}(C_4, G)] = \frac{1}{16}$. Hence,

$$\mathbb{E}[\phi(G)] = 2\mathbb{E}[t_{\text{inj}}(C_4,G)] - \mathbb{E}[t_{\text{inj}}(K_2,G)] + \frac{3}{8} = 0.$$

It follows that there is an n-vertex graph with $\phi(G) \leq 0$, and hence $G \in \Pi_{h,w,b}$ by Lemma 6.1.2.

Now suppose by contradiction that $\Pi_{h,w,b}$ is testable, and let $\mathcal{T}$ be a tester for $\Pi_{h,w,b}$. Denote by s the sample complexity of $\mathcal{T}$ when invoked with approximation parameter $\varepsilon = 0.1$.

In what follows, it will be convenient to assume that when invoked with input G (and approximation parameter 0.1), $\mathcal{T}$ works not by sampling s vertices *independently* (as required in Definition 1), but rather that $\mathcal{T}$ samples the vertices *without repetition*; or, equivalently, that $\mathcal{T}$ samples a subset of $V(G)$ of size s (uniformly at random). This assumption regarding the operation of $\mathcal{T}$ is justified because for input graphs G with sufficiently many vertices, sampling from $V(G)$ with repetition is essentially the same as sampling without repetition. Thus, if $\Pi_{h,w,b}$ has a tester which samples with repetition (as we assume here), then it also has a tester which samples without repetition. We leave the details to the reader.

So from now on we assume that when invoked with input G (and approximation parameter 0.1), $\mathcal{T}$ works by sampling a subset of $V(G)$ of size s (uniformly at random). It follows that for every $n \geq 1$,

there is a family $\mathcal{F} = \mathcal{F}(n)$ of (rejection) graphs of order s such that when invoked on input graphs with n vertices, $\mathcal{T}$ rejects if and only if the subgraph induced by its sample belongs to $\mathcal{F}(n)$. The fact that $\mathcal{T}$ is a valid tester implies that the following holds for every n-vertex graph G.

1. $p(\mathcal{F}, G) \leq \frac{1}{3}$ if $G \in \Pi_{h,w,b}$;
2. $p(\mathcal{F}, G) \geq \frac{2}{3}$ if G is 0.1-far from $\Pi_{h,w,b}$.

Let $n = \max\{10, 5s^2, n_1(s, \frac{1}{9})\}$ and $N = \max\{n_0(\frac{1}{n}), n_1(s, \frac{1}{9})\}$, where n_1 is from Corollary 6.1.7 and n_0 is from Lemma 6.1.6. Let G be an arbitrary n-vertex graph which satisfies $\Pi_{h,w,b}$, and suppose (for convenience) that $V(G) = [n]$. Let Γ be the $\frac{N}{n}$-blow-up of G. That is, Γ is obtained from G by replacing each vertex $i \in [n] = V(G)$ with a vertex-set V_i of size N/n, and replacing edges (resp. non-edges) of G with complete (resp. empty) bipartite graphs. Note that $|V(\Gamma)| = N$.

We claim that Γ is 0.1-far from $\Pi_{h,w,b}$. Indeed, fix any $\Gamma' \in \Pi_{h,w,b}$ with N vertices. By our choice of N via Lemma 6.1.6, Γ' is $\frac{1}{n}$-quasirandom. As $|V_1| = \cdots = |V_n| = \frac{N}{n}$, quasirandomness implies that $e_{\Gamma'}(V_i, V_j) = (\frac{1}{2} \pm \frac{1}{n}) \cdot (N/n)^2$ for every pair $1 \leq i < j \leq n$. But since $e_{\Gamma}(V_i, V_j) \in \{0, (N/n)^2\}$, we must change at least $(\frac{1}{2} - \frac{1}{n})(N/n)^2 \geq 0.4(N/n)^2$ edges between V_i and V_j for every $1 \leq i < j \leq n$, in order to turn Γ into Γ'. Therefore, the distance between Γ and Γ' is at least $\binom{n}{2} \cdot 0.4(N/n)^2 \geq 0.1N^2$. This shows that Γ is indeed 0.1-far from $\Pi_{h,w,b}$, as required.

Now, let $S \in \binom{V(\Gamma)}{s}$ be chosen uniformly at random, and let $\mathcal{B}$ be the event that there exists $1 \leq i \leq n$ for which $|S \cap V_i| > 1$. Note that $\mathbb{P}(\mathcal{B}) \leq \binom{s}{2}/n < \frac{1}{9}$, where the last inequality follows from our choice of n. Observe that conditioned on $\mathcal{B}^c$, the probability that $\Gamma[S]$ is isomorphic to a given s-vertex graph F is exactly $p(F, G)$. Hence, setting $\mathcal{F} := \mathcal{F}(N)$ and $\rho := \sum_{F \in \mathcal{F}} 2^{-\binom{s}{2}} \frac{s!}{\mathrm{aut}(F)}$, we have

$$\begin{aligned} p(\mathcal{F}, \Gamma) = \mathbb{P}[\Gamma[S] \in \mathcal{F}] &\leq \mathbb{P}(\Gamma[S] \in \mathcal{F} \mid \mathcal{B}^c) + \mathbb{P}(\mathcal{B}) < \mathbb{P}(\Gamma[S] \in \mathcal{F} \mid \mathcal{B}^c) + \frac{1}{9} \\ &= p(\mathcal{F}, G) + \frac{1}{9} \leq \sum_{F \in \mathcal{F}} 2^{-\binom{s}{2}} \frac{s!}{\mathrm{aut}(F)} + \frac{1}{9} + \frac{1}{9} = \rho + \frac{2}{9}, \end{aligned} \tag{6.4}$$

where in the last inequality we used our choice of n via Corollary 6.1.7. As Γ is 0.1-far from $\Pi_{h,w,b}$, Item (b) above implies that $p(\mathcal{F}, \Gamma) \geq \frac{2}{3}$, which together with (6.4) implies that $\rho > \frac{4}{9}$.

Now fix an arbitrary N-vertex graph $\Gamma^* \in \Pi_{h,w,b}$. Then by Item (a) above, $p(\mathcal{F}, \Gamma^*) \leq \frac{1}{3}$. But our choice of N via Corollary 6.1.7 implies that

$$p(\mathcal{F}, \Gamma^*) \geq \sum_{F \in \mathcal{F}} 2^{-\binom{s}{2}} \frac{s!}{\mathrm{aut}(F)} - \frac{1}{9} = \rho - \frac{1}{9} > \frac{4}{9} - \frac{1}{9} = \frac{1}{3},$$

a contradiction. This completes the proof of the theorem. ■

A careful examination of the proof of Lemma 6.1.6 can reveal how we came up with the function $\phi(G) = 2t_{\mathrm{inj}}(C_4, G) - t_{\mathrm{inj}}(K_2, G) + \frac{3}{8}$, from which we then obtained the choice of weight function w and independent coefficient b appearing in the statement of Theorem 16. Evidently, our plan for proving Theorem 16 was to find a linear inequality involving subgraph densities, which encodes the property of being quasirandom (with density $\frac{1}{2}$). Since quasirandomness depends only on the densities of edges and 4-cycles (see Theorem 6.1.5), it is natural to look for an inequality involving only these two parameters. Since every quasirandom graph satisfies $t_{\mathrm{inj}}(C_4, G) \approx t_{\mathrm{inj}}(K_2, G)^4$, it makes sense to try the following heuristic:

start with a polynomial of the form $p(x) = x^4 + ax + b$, plug in $x = t_{\text{inj}}(K_2, G)$ and replace $x^4 = t_{\text{inj}}(K_2, G)^4$ with $t_{\text{inj}}(C_4, G)$, hoping that the resulting linear inequality $t_{\text{inj}}(C_4, G) + a \cdot t_{\text{inj}}(K_2, G) + b \leq 0$ will have the required properties. For this to work, it is necessary that the polynomial p has a global minimum at $x = \frac{1}{2}$ and that p equals 0 at this point (so as to force graphs satisfying the inequality to have density $\frac{1}{2}$). Solving the constraints $p(\frac{1}{2}) = p'(\frac{1}{2}) = 0$ for a and b, one obtains $a = -\frac{1}{2}$ and $b = \frac{3}{16}$. Multiplying the resulting p by 2, one recovers the aforementioned function $\phi(G)$.

6.2 Proof of Theorem 18

We start with the following simple proposition (which was already observed in [62]). In what follows, by "standard tester" we mean a tester as in Definition 1 (as opposed to a POT).

Proposition 6.2.1. *If a graph property Π has a POT, then it also has a standard tester. Moreover, if it has a size-oblivious POT, then it also has a size-oblivious standard tester.*

Proof. Let $\mathcal{T}$ be a POT for Π, and let $c \in (0, 1]$ and $f : (0, 1] \to (0, 1]$ be as in Definition 14. Now, let $\mathcal{A}$ be the algorithm which, given an input graph G and a proximity parameter $\varepsilon > 0$, invokes $\mathcal{T}$ (independently) $t = \Theta(1/f(\varepsilon)^2)$ times and accepts if and only if $\mathcal{T}$ accepted in at least $(c - \frac{f(\varepsilon)}{2})t$ of the tests. It is easy to show (using standard concentration inequalities) that $\mathcal{A}$ accepts with probability at least $\frac{2}{3}$ if G satisfies Π, and rejects with probability at least $\frac{2}{3}$ if G is ε-far from Π (we leave the details to the reader). Moreover, it is evident that if $\mathcal{T}$ is size-oblivious then so is $\mathcal{A}$. To complete the proof, it remains to transform $\mathcal{A}$ into an algorithm which works as described in Definition 1. This is achieved by applying the transformation of [63]. A careful examination of this transformation shows that it preserves the property of being size-oblivious. ■

Next, we prove the following auxiliary lemma.

Lemma 6.2.2. *Suppose that a graph property Π has a size-oblivious (standard) ε-tester $\mathcal{T}$ with sample complexity $s = s(\varepsilon)$. Then for every $n \geq s^4$ and for every n-vertex graph G which is ε-far from Π, the following holds: for U chosen uniformly at random from $\binom{V(G)}{s^4}$, we have $\mathbb{P}[G[U] \in \Pi] \leq e^{-\Omega(s)}$.*

Proof. We use a double-sampling trick which is implicit in [59]. Let $\mathcal{A}$ be the family of all s-vertex graphs A such that $\mathcal{T}$ accepts if it sees a subgraph isomorphic to A. For a graph G, we say that a sequence of subsets $S_1, \ldots, S_s \in \binom{V(G)}{s}$ is *good* if $G[S_i] \in \mathcal{A}$ for at least half of the values of $1 \leq i \leq s$; otherwise $S_1, \ldots, S_s$ is *bad*. For a sequence of vertices $W = (x_1, \ldots, x_{s^2})$, we say that W is good (resp. bad) if $\{x_1, \ldots, x_s\}, \{x_{s+1}, \ldots, x_{2s}\}, \ldots, \{x_{s^2-s+1}, \ldots, x_{s^2}\}$ is good (resp. bad). Note that for a random $S \in \binom{V(G)}{s}$, if $G \in \Pi$ then $\mathbb{P}[G[S] \in \mathcal{A}] \geq \frac{2}{3}$, and if G is ε-far from Π then $\mathbb{P}[G[S] \in \mathcal{A}] \leq \frac{1}{3}$. Using a standard Chernoff-type bound, one can show that the following holds for $S_1, \ldots, S_s \in \binom{V(G)}{s}$ chosen uniformly at random and independently.

1. If G satisfies Π then $S_1, \ldots, S_s$ is good with probability at least $1 - e^{-Cs}$.
2. If G is ε-far from Π then $S_1, \ldots, S_s$ is bad with probability at least $1 - e^{-Cs}$.

In both items above, $C > 0$ is an absolute constant.

The probability that there exists a pair $1 \leq i < j \leq s$ for which $S_i \cap S_j \neq \emptyset$ is at most $\binom{s}{2}\frac{s^2}{n} < \frac{1}{2}$, where the inequality follows from the assumption that $n \geq s^4$. It follows that with probability larger than $\frac{1}{2}$, the sets $S_1, \ldots, S_s$ are pairwise-disjoint. Conditioned on the event that $S_1, \ldots, S_s$ are pairwise-disjoint, the set $S := S_1 \cup \cdots \cup S_s$ has the distribution of an element of $\binom{V(G)}{s^2}$ chosen uniformly at random. Thus, a random sequence of vertices $S = (x_1, \ldots, x_{s^2})$ chosen *without repetition* from a given graph G satisfies the following.

1. If G satisfies Π then S is good with probability at least $1 - 2e^{-Cs}$.
2. If G is ε-far from Π then S is bad with probability at least $1 - 2e^{-Cs}$.

Now let G be a graph on $n \geq s^4$ vertices which is ε-far from Π. Consider a random pair (U, S), where U is chosen uniformly at random from $\binom{V(G)}{s^4}$, and $S = (x_1, \ldots, x_{s^2})$ is a sequence of vertices sampled randomly without repetition from U. Then the marginal distribution of S is that of a uniform sequence of s^2 vertices of G, sampled without repetition. Thus, by viewing S as a sample from G (and recalling that G is ε-far from Π), we see that $\mathbb{P}[S \text{ is good}] \leq 2e^{-Cs}$ (by Item 2 above). On the other hand, if $G[U] \in \Pi$, then, by viewing S as a sample from $G[U]$, we see that $\mathbb{P}[S \text{ is good} \mid U] \geq 1 - 2e^{-Cs}$ (by Item 1 above). By combining these two facts, we conclude that

$$\mathbb{P}[G[U] \in \Pi] \leq \frac{\mathbb{P}[S \text{ is good}]}{\mathbb{P}[S \text{ is good} \mid G[U] \in \Pi]} \leq \frac{2e^{-Cs}}{1 - 2e^{-Cs}} \leq 4e^{-Cs} = e^{-\Omega(s)}.$$

■

Proof of Theorem 18. Let (h, w, b) be a tuple for which $\Pi_{h,w,b}$ has a size-oblivious POT. By Proposition 6.2.1, $\Pi_{h,w,b}$ also has a size-oblivious standard tester $\mathcal{T}$, meaning that $\Pi_{h,w,b}$ satisfies the condition of Lemma 6.2.2. Denote by $s = s(\varepsilon)$ the sample complexity of $\mathcal{T}$. We may and will assume that s is large enough as a function of the parameters h and b.

Denote $z(G) := \sum_H w_H \cdot p(H, G)$. By multiplying the inequality $\sum_H w_H \cdot p(H, G) \leq b$ by an appropriate integer, we can assume without loss of generality that b and all weights $(w_H : H)$ are integers.

Let $\varepsilon \in (0, 1]$ and let G be a graph which is ε-far from $\Pi_{h,w,b}$. Our goal is to show that $z(G) \geq b + f(\varepsilon)$, for a function $f : (0, 1] \to (0, 1]$ to be chosen later. Suppose first that $n < s^4$. As G does not satisfy $\Pi_{h,w,b}$, we have $z(G) = \sum_H w_H \cdot p(H, G) > b$. Now, since b and $(w_H : H)$ are all integers, and as $p(H, G)$ is an integer multiple of $\binom{n}{h}^{-1}$ for every H, we must have

$$z(G) \geq b + \binom{n}{h}^{-1} > b + n^{-h} > b + s^{-4h},$$

implying that our assertion holds with $f(\varepsilon) = s(\varepsilon)^{-4h}$ in this case.

Suppose now that $n \geq s^4$, which is necessary in order to apply Lemma 6.2.2. That lemma implies that a randomly chosen $U \in \binom{V(G)}{s^4}$ satisfies $G[U] \notin \Pi_{h,w,b}$ with probability at least $1 - e^{-\Omega(s)}$. As before, we observe that if a k-vertex graph K does not satisfy $\Pi_{h,w,b}$, then necessarily

$$z(K) = \sum_H w_H \cdot p(H, K) \geq b + \binom{k}{h}^{-1} > b + k^{-h},$$

as b and all weights w_H are integers. Thus, if $G[U] \notin \Pi_{h,w,b}$ then

$$z(G[U]) > b + |U|^{-h} = b + s^{-4h}.$$

Observe (crucially) that $z(G)$ is the average of $z(G[U])$ over all $U \in \binom{V(G)}{s^4}$. Thus, using the guarantees of Lemma 6.2.2, we obtain

$$z(G) \geq (1 - e^{-\Omega(s)})(b + s^{-4h}) > b + \frac{1}{2}s^{-4h},$$

where the last inequality holds provided that s is large enough as a function of h and b. So we may take the function f in Definition 17 to be $f(\varepsilon) = \frac{1}{2}s(\varepsilon)^{-4h}$. This completes the proof. ■

Chapter 7

A New Bound for the Brown–Erdős–Sós Problem

This chapter is dedicated to the proof of Theorem 20. Let us mention that the Brown–Erdős–Sós conjecture, namely Conjecture 19, has a more general form (see [14, 45, 100]), which we now state. Recall that a (v, e)-*configuration* is a hypergraph with e edges and at most v vertices. Denote by $f_r(n, v, e)$ the largest number of edges in an n-vertex r-uniform hypergraph (r-graph for short) that contains no (v, e)-configuration. With this notation, Conjecture 19 can be restated as saying that $f_3(n, e + 3, e) = o(n^2)$, and Theorem 20 as saying that $f_3(n, e + 18 \log e/ \log \log e, e) = o(n^2)$ (for all $e \geq 3$). The aforementioned general form of the Brown–Erdős–Sós conjecture states that for every $2 \leq k < r$ and $e \geq 3$ it holds that $f_r(n, (r-k)e+k+1, e) = o(n^k)$. It is worth noting that this particular choice of the parameters v, e is due to the fact that $f_r(n, (r-k)e+k, e) = \Theta(n^k)$, as proved by Brown, Erdős and Sós [26]. The Brown–Erdős–Sós conjecture can then be thought of as stating that the extremal function $f_r(n, v, e)$ drops significantly if v is increased by 1 from $(r-k)e+k$ to $(r-k)e+k+1$. A related question is then whether or not the drop is actually by a polynomial factor, namely, whether it is true that $f_r(n, (r-k)e+k+1, e) \geq n^{k-o(1)}$. This lower bound is known to hold in several cases [14, 45, 98]. Lastly, it is worth mentioning that the case $k = r-1, e = r$ of the Brown–Erdős–Sós conjecture, namely, the statement that $f_r(n, 2r, r) = o(n^r)$, follows from the $(r-1)$-uniform hypergraph removal lemma (see [52, 67, 86, 89, 90]).

It is a folklore observation that the general version of the Brown–Erdős–Sós conjecture, stated above, is in fact equivalent to the special case stated as Conjecture 19 (corresponding to $k = 2$ and $r = 3$). Since this reduction does not appear in the literature, we give its proof here. We will, in fact, prove the following more general statement:

Proposition 7.0.1. *For every $2 \leq k < r$, $e \geq 3$ and $d \geq 1$,*

$$f_r(n, (r-k)e+k+d, e) \leq \binom{r}{3} e n^{k-2} \cdot f_3(n, e+2+d, e).$$

Setting $d = 1$ in the above proposition readily implies that Conjecture 19 is indeed equivalent to the general form of the Brown–Erdős–Sós conjecture stated above. The reason for stating the proposition for arbitrary d is that it allows us to infer approximate versions of the general Brown–Erdős–Sós conjecture from approximate versions of Conjecture 19. In particular, by combining Theorem 20 with Proposition 7.0.1, we immediately obtain the following corollary.

Corollary 7.0.2. *For every $2 \leq k < r$ and $e \geq 3$,*

$$f_r\left(n, (r-k)e + k - 2 + 18\log e/\log\log e, e\right) = o(n^k).$$

We note that by being more careful, one can replace the multiplicative constant 18 in Theorem 7.0.2 by $4 + o(1)$. The rest of this chapter is organized as follows. In Section 7.1, we give an overview of the main ideas which go into the proof of Theorem 20. We also state the two key lemmas of this chapter and explain how they imply Theorem 20. We then prove these two lemmas in Sections 7.2 and 7.3. Finally, in Section 7.4, we discuss an application of our results to a generalized Ramsey problem of Erdős and Gyárfás which is known to have connections to the Brown–Erdős–Sós problem. All logarithms are natural unless explicitly stated otherwise.

7.1 Outline of the Proof

Our goal in this section is fourfold. We first give an overview of the proof of Theorem 20. In doing so, we will state the two key lemmas, Lemmas 7.1.6 and 7.1.8, used in its proof. We will then proceed to show how these two lemmas can be used in order to prove Theorem 20. Finally, in Section 7.1.4, we prove Proposition 7.0.1.

7.1.1 Proof Overview and the Key Lemmas

First, let us restate Theorem 20 as follows:

Theorem 7.1.1. *For every $e \geq 3$ and $\varepsilon \in (0,1)$, there is $n_0 = n_0(e,\varepsilon)$ such that every 3-graph with $n \geq n_0$ vertices and at least εn^2 edges contains a (v,e)-configuration with $v - e \leq 18\log e/\log\log e$.*

Our first simple (yet crucial) observation towards the proof of Theorem 7.1.1 is that, in order to prove the theorem, it is enough to prove the following approximate version.

Lemma 7.1.2. *For every $e \geq 40320 = 8!$ and $\varepsilon \in (0,1)$, there is $n_0 = n_0(e,\varepsilon)$ such that every 3-graph H with $n \geq n_0$ vertices and at least εn^2 edges contains a (v',e')-configuration satisfying $e - \sqrt{e} \leq e' \leq e$ and $v' - e' \leq 8\log e/\log\log e$.*

In Section 7.1.3 we will show how to quickly derive Theorem 7.1.1 from the above lemma. So let us proceed with the overview of the proof of Lemma 7.1.2. We will heavily rely on the hypergraph removal lemma, which states the following.

Theorem 7.1.3 (Hypergraph removal lemma [67, 86, 89, 90])**.** *For every $k \geq 2$ and $\varepsilon > 0$ there exists $\gamma = \gamma(k,\varepsilon) > 0$ such that the following holds. Let $n \geq 1$ and let J be a k-uniform n-vertex hypergraph which contains a collection of at least εn^k pairwise edge-disjoint $(k+1)$-cliques. Then J contains at least γn^{k+1} $(k+1)$-cliques.*

As mentioned in Chapter 1, Sárközy and Selkow [99] have shown that $f_3(n, e+2+\lfloor\log_2 e\rfloor, e) = o(n^2)$. The proof of this result in [99] roughly proceeds as follows: suppose one has already proved that every sufficiently large n-vertex 3-graph with $\Omega(n^2)$ edges contains an $(e+k,e)$-configuration (for some values

of e and k). Using this fact, one then shows that every such 3-graph also contains a $(2e+k+2, 2e+1)$-configuration. In other words, at the price of increasing $v-e$ by 1, we multiply the number of edges by roughly 2 (and hence the term $\log_2 e$ in the aforementioned result of [99]). The proof of [99] used the graph removal lemma (at least implicitly[1]). As we mentioned before, Solymosi and Solymosi [107] improved the bound of [99] for the special case $e = 10$. The way they achieved this was by cleverly replacing the application of the graph removal lemma with an application of the 3-graph removal lemma. Roughly speaking, this allowed them to multiply a $(6,3)$-configuration by 3, instead of by 2 as in [99].

The above discussion naturally leads one to try and extend the approach of [107] by showing that after multiplying the initial configuration by 3, one can use the 4-graph removal lemma to multiply the resulting configuration by 4, etc. Performing k such steps should (roughly) give a $(k!+k, k!)$-configuration, or equivalently, a (v,e)-configuration with $v-e = O(\log e/\log\log e)$. There is one big challenge and two problems with this approach. The challenge is of course how to achieve this repeated multiplication process.[2] As to the problems, the first is that we do not know how to guarantee that one can indeed keep multiplying the size of the configurations. In other words, it is entirely possible that this process might get "stuck" along the way (this scenario is described in Item 1 of Lemma 7.1.6). More importantly, even if the process succeeds in producing a $(k!+k, k!)$-configuration for every k, it is not clear how to interpolate so as to prove Theorem 7.1.1 for values of e with $(k-1)! < e < k!$. That is, our process only guarantees the existence of suitable configurations for a very sparse set of values of e. It it tempting to guess that the resulting $(k!+k, k!)$-configurations are "degenerate", in the sense that one can repeatedly remove from them vertices of degree 1, thus maintaining the difference $v-e$. This is however false. Having said that, we will return to this degeneracy issue after the statement of Lemma 7.1.8.

In what follows, it will be convenient to use the following notation.

Definition 7.1.4. *For a 3-graph F and $U \subseteq V(F)$, the* difference *of U is defined as $\Delta(U) := |U| - e(U)$. We will write $\Delta(F)$ for $\Delta(V(F))$, i.e., $\Delta(F) := v(F) - e(F)$, and call $\Delta(F)$ the difference of F.*

Our first key lemma, Lemma 7.1.6 below, comes close to achieving what is described in the paragraph above. Given an n-vertex 3-graph H with $\Omega(n^2)$ edges, the lemma almost resolves the challenge mentioned in the previous paragraph by either showing that H contains configurations with difference k and size roughly $k!$ (this is the statement of Item 2) or getting stuck in the scenario described in Item 1. The silver lining in Item 1 is that we get an arithmetic progression of values v for which we can construct (v,e)-configurations of small difference. The problem is that the common difference of this arithmetic progression might be much larger than $\sqrt{e}$, so this lemma alone cannot be used in order to prove Lemma 7.1.2.

The key definition in Lemma 7.1.6 is the notion of a nice 3-graph, which we now define. Satisfying this definition makes a 3-graph amenable to the arguments we use in the proof of Lemma 7.1.6.

Definition 7.1.5. *Let F be a 3-graph and put $k := \Delta(F) = v(F) - e(F)$. We call F* nice *if there is an independent set $A \subseteq V(F)$ of size $k+1$ such that the following holds for every $U \subseteq V(F)$.*

1. *$\Delta(U) \geq |U \cap A| - \mathbb{1}_{A \subseteq U}$.*
2. *If $|U \cap A| \leq k-1$ and $U \setminus A \neq \emptyset$, then $\Delta(U) \geq |U \cap A| + 1$.*

[1] We will extend their approach in Lemma 7.1.8 by using the removal lemma explicitly.

[2] The special case in [107] of multiplying a $(6,3)$-configuration by 3 proceeds by case analysis which is not generalizable.

Lemma 7.1.6. *There is a sequence $(F_k)_{k\geq 3}$ of 3-graphs such that $\Delta(F_k) = v(F_k) - e(F_k) = k$, F_k is nice for each $k \geq 4$, $e(F_3) = 3$ and $e(F_k) = 5k!/12$ for each $k \geq 4$, and the following holds. For every $k \geq 4$, $r \geq 1$ and $\varepsilon \in (0,1)$, there are $\eta = \eta(k,r,\varepsilon) \in (0,1)$ and $n_0 = n_0(k,r,\varepsilon)$ such that every 3-graph H with $n \geq n_0$ vertices and at least εn^2 edges satisfies (at least) one of the following:*

1. *There are $3 \leq j \leq k-1$ and $j \leq q \leq v(F_j)-1$ such that, for every $1 \leq i \leq r$, the 3-graph H contains a (v',e')-configuration with $v' - e' \leq j$ and $v' = q + i \cdot (v(F_j) - q)$.*
2. *H contains at least ηn^k copies of F_k.*

Remark 7.1.7. *A recurring theme in our arguments is that, given some suitable 3-graph F, we will be able to show that every sufficiently large n-vertex 3-graph H with $\Omega(n^2)$ edges contains $\Omega(n^{v(F)-e(F)})$ copies of F (unless H satisfies the assertion of Theorem 7.1.1 for some other reason). This estimate for the number of copies of F is tight, since a random hypergraph with edge density $\frac{1}{n}$ has $O(n^{v(F)-e(F)})$ copies of F w.h.p.*

The proof of Lemma 7.1.6 proceeds by induction on k. Namely, assuming H contains $\Omega(n^{k-1})$ copies of F_{k-1}, we show that either H contains $\Omega(n^k)$ copies of F_k or Item 1 holds. This is done as follows. Recalling that F_{k-1} is nice (for $k \geq 5$), we fix a set $A \subseteq V(F_{k-1})$ of size $|A| = k$ which witnesses this fact (see Definition 7.1.5). For each embedding $\varphi : V(F_{k-1}) \to V(H)$ of F_{k-1} into H, we consider the set $\varphi(A) \subseteq V(H)$. By a straightforward argument (combining an application of the multicolor Ramsey theorem with a simple cleaning procedure), we can show that either there are embeddings $\varphi_1, \ldots, \varphi_r : V(F_{k-1}) \to V(H)$ and a set $U \subseteq V(F_{k-1})$ such that $|U| \geq k-1$, $|U \cap A| \geq k-2$ and $\varphi_1|_U = \cdots = \varphi_r|_U$; or there is a family $\mathcal{F}$ of $\Omega(n^{k-1})$ embeddings $\varphi : V(F_{k-1}) \to V(H)$ such that, for any two $\varphi, \varphi' \in \mathcal{F}$, the set $U = \{u \in V(F_{k-1}) : \varphi(u) = \varphi'(u)\}$ (i.e., the set of elements on which φ and φ' agree) satisfies $|U \cap A| \leq k-2$ (and $U \subseteq A$ if $|U \cap A| = k-2$). In the former case, Items 1-2 of Definition 7.1.5 imply that the union of the copies of F_k corresponding to $\varphi_1, \ldots, \varphi_r$ has difference at most $k-1$ (which is also the difference of F_{k-1}), from which it easily follows that Item 1 in Lemma 7.1.6 holds. In the latter case, we define an auxiliary $(k-1)$-uniform hypergraph by putting a $(k-1)$-uniform k-clique on the set $\varphi(A)$ for each $A \in \mathcal{F}$. The aforementioned property of $\mathcal{F}$ implies that these cliques are pairwise edge-disjoint, which allows us to apply the hypergraph removal lemma (Theorem 7.1.3) and thus infer that the number of k-cliques in our auxiliary hypergraph is at least $\Omega(n^k)$. Using again our guarantees regarding $\mathcal{F}$, we can show that most such k-cliques correspond to copies of a particular 3-graph consisting of k copies of F_{k-1} which do *not* intersect outside of the set A. This 3-graph is then chosen as F_k. One of the challenges in the proof is to then show that F_k is itself nice, thus allowing the induction to continue. The full details appear in Section 7.2.

We now move to our next key lemma, Lemma 7.1.8 below. Let us say that a 3-graph is d-degenerate if it is possible to repeatedly remove from it a set of at least d vertices which touches at most d edges. As we mentioned above, the 3-graphs F_k are not 1-degenerate, so it is not possible to take one of these 3-graphs and repeatedly remove vertices of degree at most 1 so as to obtain configurations with any desired number of edges, while not increasing the difference. One can argue, however, that since Lemma 7.1.2 only asks for e' to satisfy $e - \sqrt{e} \leq e' \leq e$, it is enough to show that the 3-graphs F_k are $\sqrt{e(F_k)}$-degenerate. Unfortunately, we cannot do even this. Instead, we will overcome the problem by using Lemma 7.1.8. This lemma states that if H contains many copies of some nice 3-graph G, then it also contains copies of 3-graphs $G_0 = G, G_1, G_2, \ldots$ which are all $e(G)$-degenerate and whose sizes increase. In fact, as in Lemma

7.1.6, we cannot always guarantee success in finding copies of $G_1, G_2, \ldots, G_\ell$ in H, since the process might get stuck in a situation analogous to the one in Lemma 7.1.6. Finally, the price we have to pay for the degeneracy guaranteed by Item 2 of Lemma 7.1.8 is that the size of the 3-graphs $G_1, G_2, \ldots, G_\ell$ only grows by a factor of roughly k at each step. Hence, just like Lemma 7.1.6, Lemma 7.1.8 also falls short of proving Lemma 7.1.2.

Lemma 7.1.8. *Let G be a nice 3-graph, put $k := \Delta(G) = v(G) - e(G)$ and assume that $k \geq 2$. Then there is a sequence of 3-graphs $(G_\ell)_{\ell \geq 0}$ having the following properties.*

1. *$G_0 = G$, $\Delta(G_\ell) = v(G_\ell) - e(G_\ell) = k + \ell$ and $e(G_\ell) = \frac{k^{\ell+1}-1}{k-1} \cdot e(G)$.*
2. *For every $\ell \geq 0$ and every $0 \leq t \leq e(G_\ell)/e(G)$, the 3-graph G_ℓ contains a (v', e')-configuration with $v' - e' \leq k + \ell$ and $e' = t \cdot e(G)$.*
3. *For every $\ell \geq 0$, $r \geq 0$ and $\varepsilon \in (0, 1)$, there are $\delta = \delta(\ell, r, \varepsilon)$ and $n_0 = n_0(\ell, r, \varepsilon)$ such that, for every 3-graph H on $n \geq n_0$ vertices, if H contains at least εn^k copies of G, then (at least) one of the following conditions is satisfied:*
   (a) *There are $0 \leq j \leq \ell - 1$ and $k + j \leq q \leq v(G_j) - 1$ such that, for every $1 \leq i \leq r$, the 3-graph H contains a (v', e')-configuration which contains a copy of G_j, where $v' - e' \leq k + j$ and $v' = q + i \cdot (v(G_j) - q)$.*
   (b) *H contains at least $\delta \cdot n^{k+\ell}$ copies of G_ℓ.*

Strictly speaking, we cannot apply Lemma 7.1.8 with G being an edge, since an edge is not a nice 3-graph (indeed, it has difference $k = 2$ but evidently contains no independent set of size $k + 1 = 3$). However, one can check that the proof also works when G is an edge and, more generally, in any case where $k := \Delta(G) = 2$ and one can choose a (not necessarily independent) $A \subseteq V(G)$ of size 3 which satisfies Items 1-2 in Definition 7.1.5. By applying Lemma 7.1.8 with G being an edge, one recovers the construction used by Sárközy and Selkow [99] to prove their result that $f_3(n, e + 2 + \lfloor \log_2 e \rfloor, e) = o(n^2)$. Generalizing this construction to other graphs G (e.g., for $k \geq 3$) presents a challenge, which we overcome by using some of the ideas from the proof of Lemma 7.1.6.

We now sketch the derivation of Lemma 7.1.2 from Lemmas 7.1.6 and 7.1.8 (the formal proof appears in Section 7.1.2). Given e, choose k so that $(2k)! \approx e$; so $k! \approx \sqrt{e}$ and $k = O(\log e / \log \log e)$. We first apply Lemma 7.1.6 with k. If we are at Item 1, then we get an arithmetic progression with difference at most $v(F_k) - k \leq k! \leq \sqrt{e}$ of values v' for which we can find (v', e')-configurations of difference at most k, thus completing the proof in this case. Suppose then that we are at Item 2, implying that H contains $\Omega(n^k)$ copies of F_k. Since F_k is nice, we can apply Lemma 7.1.8 with $G = F_k$. Since $e(F_k) \approx k!$ and $(2k)! \approx e$, choosing, say, $\ell = 3k$ guarantees that $e(G_\ell) \approx e(F_k) \cdot k^\ell > e$ (via Item 1 of Lemma 7.1.8). If the application of Lemma 7.1.8 results in Item 3(b), then we can use Item 2 of that lemma to find a (v', e')-configuration of difference $O(k + \ell) = O(k)$ with $e - \sqrt{e} \leq e - e(G) \leq e' \leq e$, thus completing the proof. Finally, suppose that we are at Item 3(a). In this case we can find a (v', e')-configuration G' of difference $O(k + \ell) = O(k)$ with $e - e(G_j) \leq e' \leq e$. With the help of a simple trick we can also find in H a copy G^* of G_j which is *edge-disjoint* from G'. As in case 3(b) above, we use Item 2 to find a sub-configuration G'' of G^* with $e - e(G') - e(G) \leq e(G'') \leq e - e(G')$. If we now take G''' to be the union of G' and G'', we infer that G''' has difference $O(k)$ and $e - \sqrt{e} \leq e - e(G) \leq e(G''') \leq e$. So again we are done.

7.1.2 Deriving Lemma 7.1.2 from Lemmas 7.1.6 and 7.1.8

The required integer $n_0 = n_0(e, \varepsilon)$ will be chosen implicitly. Let $(F_k)_{k\geq 3}$ be the nice 3-graphs whose existence is guaranteed by Lemma 7.1.6. Recall that $e(F_k) = 5k!/12$ for each $k \geq 4$ and that $e(F_3) = 3$. Let $K \geq 8$ be such that $K! \leq e < (K+1)!$ and put $k := \lfloor K/2 \rfloor \geq 4$. Note that $e(F_k) \leq k! \leq (K/2)! \leq \sqrt{K!} \leq \sqrt{e}$. It is not hard to check that $K \leq 2\log e/\log\log e$ and hence $k \leq \log e/\log\log e$. We will now apply our second construction, given by Lemma 7.1.8. Set $G := F_k$ and let $(G_\ell)_{\ell\geq 0}$ be the sequence of 3-graphs whose existence is guaranteed by Lemma 7.1.8. Let ℓ be the minimal integer satisfying $e(G_\ell) \geq e$. Then $\ell \geq 1$ (because $e(G_0) = e(G) = e(F_k) < e$). We will now bound ℓ in terms of k. For our purposes, it will be enough to show that $\ell \leq 3k$. To this end, observe that

$$e(G_{3k}) = \frac{k^{3k+1}-1}{k-1} \cdot e(G) \geq k^{3k} = \lfloor K/2 \rfloor^{3\lfloor K/2 \rfloor} \geq (K+1)! > e,$$

where the first equality follows from Item 1 of Lemma 7.1.8 and the penultimate inequality holds for every $K \geq 8$. The fact that $e(G_{3k}) > e$ now readily implies that $\ell \leq 3k$.

Let H be a 3-graph with $n \geq n_0$ vertices and at least εn^2 edges. Partition $E(H)$ into equal-sized parts $E_1, \ldots, E_{\ell+1}$ and, for each $1 \leq i \leq \ell+1$, let H_i be the hypergraph $(V(H), E_i)$. Note that $e(H_i) \geq e(H)/(\ell+1) \geq \varepsilon n^2/(\ell+1)$ for each $1 \leq i \leq \ell+1$.

Claim 7.1.9. *For each $1 \leq m \leq \ell+1$, either H_m satisfies the assertion of Lemma 7.1.2 or there exists $0 \leq j \leq \ell-1$ such that H_m contains a (v', e')-configuration which contains a copy of G_j, where $v'-e' \leq k+j$ and $e - e(G_j) \leq e' \leq e$.*

Proof. Evidently, it is enough to prove the claim for $m = 1$. We apply Lemma 7.1.6 to H_1 with parameters $r = e+k$ and $\varepsilon/(\ell+1)$. Suppose first that the assertion of Item 1 in Lemma 7.1.6 holds and let $3 \leq j \leq k-1$ and $j \leq q \leq v(F_j)-1$ be as in that item. Let i be the maximal integer satisfying $q + i \cdot (v(F_j) - q) \leq e + j$ and note that $1 \leq i \leq e+j \leq e+k$. We may thus infer from Item 1 in Lemma 7.1.6 that H_1 contains a (v', e')-configuration with

$$v' = q + i \cdot (v(F_j) - q) \leq e + j, \tag{7.1}$$

and

$$v' - e' \leq j < k \leq \log e/\log\log e. \tag{7.2}$$

Note that the maximality of i guarantees that

$$v' > e + j - (v(F_j) - q). \tag{7.3}$$

We now observe that we can assume that $e' \leq e$. Indeed, since by (7.2) we have $v' - e' \leq j$, then we can remove edges until the equality $e' = v' - j$ holds. Having done that, we are guaranteed by (7.1) that $e' \leq e$. As to the lower bound on e', by (7.3) we have $e - e' = e + j - v' < v(F_j) - q \leq v(F_j) - j$. By Lemma 7.1.6, we have $v(F_j) - j = 5j!/12$ if $j \geq 4$ and $v(F_j) - j = 3$ if $j = 3$. In either case, we get $e - e' \leq j! \leq k! \leq \sqrt{e}$. So we see that H_1 satisfies the assertion of Lemma 7.1.2, as required. This completes the proof for the case that the assertion of Item 1 in Lemma 7.1.6 holds.

Suppose from now on that the assertion of Item 2 in Lemma 7.1.6 holds, namely, that H_1 contains at least ηn^k copies of $F_k = G$. This means that we may apply Lemma 7.1.8 to H_1. By Item 3 of Lemma 7.1.8, applied with $r = e + k + \ell$ and with η in place of ε, the 3-graph H_1 satisfies (at least) one of the following:

(a) There are some $0 \leq j \leq \ell - 1$ and $k + j \leq q \leq v(G_j) - 1$ such that, for every $1 \leq i \leq e + k + \ell$, H_1 contains a (v', e')-configuration which contains a copy of G_j, where $v' - e' \leq k + j$ and $v' = q + i \cdot (v(G_j) - q)$.

(b) H_1 contains a copy of G_ℓ (in fact, at least $\delta(\ell, r, \eta) \cdot n^{k+\ell}$ such copies).

Suppose first that H_1 satisfies Item (b). Let $t \geq 0$ be the maximal integer satisfying $t \cdot e(G) \leq e$ and note that $t \leq e/e(G) \leq e(G_\ell)/e(G)$, where the second inequality uses our choice of ℓ. By Item 2 of Lemma 7.1.8, H_1 contains a (v', e')-configuration with $v' - e' \leq k + \ell \leq 4k \leq 4 \log e / \log \log e$ and $e' = t \cdot e(G) \leq e$. By our choice of t, we have $e - e' < e(G) = 5k!/12 \leq k! \leq \sqrt{e}$. So in this case the assertion of Lemma 7.1.2 indeed holds for H_1.

From now on we assume that H_1 satisfies Item (a) and let $0 \leq j \leq \ell - 1$ and $k + j \leq q \leq v(G_j) - 1$ be as in that item. Let i be the maximal integer satisfying $q + i \cdot (v(G_j) - q) \leq e + k + j$. Then $1 \leq i \leq e + k + j < e + k + \ell$. We may thus rely on (a) above to conclude that H_1 contains a (v', e')-configuration which contains a copy of G_j, where

$$v' = q + i \cdot (v(G_j) - q) \leq e + k + j, \tag{7.4}$$

and

$$v' - e' \leq k + j. \tag{7.5}$$

Note that the maximality of i guarantees that

$$v' > e + k + j - (v(G_j) - q). \tag{7.6}$$

We now observe that we can assume that $e' \leq e$. Indeed, since by (7.5) we have $v' - e' \leq k + j$ then we can remove edges until the equality $e' = v' - (k + j)$ holds. By (7.4), this would guarantee that $e' \leq e$. Note (crucially) that since $e(G_j) = v(G_j) - k - j \leq v' - k - j$, we can make sure that even after removing the required number of edges we still have a copy of G_j. As to the lower bound on e', by (7.5) and (7.6) we have $e - e' \leq e - v' + k + j < v(G_j) - q \leq v(G_j) - k - j = e(G_j)$. We conclude that H_1 indeed contains a (v', e')-configuration with the properties stated in the claim. ■

We now return to the proof of the lemma. If some H_m satisfies the assertion of Lemma 7.1.2 then we are done. Otherwise, Claim 7.1.9 implies that for each $1 \leq m \leq \ell + 1$ there is $0 \leq j_m \leq \ell - 1$ such that H_m contains a (v', e')-configuration which contains a copy of G_{j_m}, where $v' - e' \leq k + j_m$ and $e - e(G_{j_m}) \leq e' \leq e$. By the pigeonhole principle, there are two indices $1 \leq i \leq \ell + 1$ whose j_m's are equal. It follows that for some $0 \leq j \leq \ell - 1$, H contains *edge-disjoint* subgraphs G^* and G' such that G^* is isomorphic to G_j and G' satisfies $v(G') - e(G') \leq k + j$ and $e - e(G_j) \leq e(G') \leq e$. Let t be the maximal integer satisfying $t \cdot e(G) \leq e - e(G')$ and note that $0 \leq t \leq e(G_j)/e(G)$. Then, by Item 2 of Lemma 7.1.8 (with j in place of ℓ), there is a subgraph G'' of G^* such that $v(G'') - e(G'') \leq k + j$ and $e(G'') = t \cdot e(G)$. Our choice of t implies that $0 \leq e - e(G') - e(G'') < e(G) \leq k! \leq \sqrt{e}$. Now, letting G''' be the union of G' and G'', we see that $e - \sqrt{e} \leq e(G''') \leq e$ and

$$v(G''') - e(G''') \leq v(G') - e(G') + v(G'') - e(G'') \leq 2(k + j) \leq 2(k + \ell) \leq 8k \leq 8 \log e / \log \log e.$$

So we see that the assertion of the lemma holds with G''' as the required (v', e')-configuration.

7.1.3 Deriving Theorem 7.1.1 from Lemma 7.1.2

Our goal is to show that for every $e \geq 3$ and $\varepsilon \in (0,1)$, there is $n_0 = n_0(e,\varepsilon)$ such that every 3-graph with $n \geq n_0$ vertices and at least εn^2 edges contains a (v,e)-configuration with $v - e \leq 18 \log e / \log\log e$. As in the proof of Lemma 7.1.2, the required integer $n_0 = n_0(e,\varepsilon)$ will be chosen implicitly. The proof is by induction on e. Let H be a 3-graph with $n \geq n_0$ vertices and at least εn^2 edges. By the main result of [99] (mentioned above), H contains a (v,e)-configuration with $v - e \leq 2 + \lfloor \log_2 e \rfloor$. If $e \leq \exp(2^{16})$, then we have $2 + \lfloor \log_2 e \rfloor \leq 2 + 16 \log e / \log\log e \leq 18 \log e / \log\log e$ (where the second inequality holds whenever $e \geq 3$), thus completing the proof in this case. So suppose from now on that $e > \exp(2^{16}) \geq 40320$. (The inequality $e \geq 40320$ is required to apply Lemma 7.1.2.)

By Lemma 7.1.2, H contains a (v',e')-configuration F' satisfying $e - \sqrt{e} \leq e' \leq e$ and $v' - e' \leq 8 \log e / \log\log e$. Set $e'' := e - e'$, noting that $0 \leq e'' \leq \sqrt{e}$. If $e'' \leq 15$, then, by adding at most 15 edges to F', one obtains a (v,e)-configuration with $v - e \leq v' + 3e'' - (e' + e'') = v' - e' + 2e'' \leq 8 \log e / \log\log e + 30 \leq 18 \log e / \log\log e$, as required. (Here the last inequality is guaranteed by our assumption that e is large.) So suppose from now on that $e'' \geq 16$. Let H' be the 3-graph obtained from H by deleting the edges of F'. Since $e(H') \geq e(H) - e(F') \geq \varepsilon n^2 - e(F') \geq \frac{\varepsilon}{2} n^2$ (provided that n is large enough), we may apply the induction hypothesis to H', with parameter e'' in place of e, and thus obtain a (v'',e'')-configuration F'' which is edge-disjoint from F' (because it is contained in H') and satisfies

$$v'' - e'' \leq \frac{18 \log e''}{\log\log e''} \leq \frac{18 \log \sqrt{e}}{\log\log \sqrt{e}} = \frac{9 \log e}{\log\log e - \log 2}.$$

Here, in the second inequality we used the fact that the function $x \mapsto \log x / \log\log x$ is monotone increasing for $x \geq 16$. Letting F be the union of F' and F'', we see that $e(F) = e(F') + e(F'') = e$ and $v(F) \leq v(F') + v(F'')$, implying that

$$v(F) - e(F) \leq v(F') - e(F') + v(F'') - e(F'') \leq \frac{8 \log e}{\log\log e} + \frac{9 \log e}{\log\log e - \log 2} \leq \frac{18 \log e}{\log\log e},$$

where the last inequality holds whenever $e \geq \exp(2^{10})$. This completes the proof of the theorem.

7.1.4 Proof of Proposition 7.0.1

Let $2 \leq k < r$, $e \geq 3$ and $d \geq 1$. Let H be an n-vertex r-graph with

$$e(H) \geq \binom{r}{3} e n^{k-2} \cdot f_3(n, e+2+d, e).$$

Our goal is to show that H contains a (v,e)-configuration with $v \leq (r-k)e + k + d$. By averaging, there are vertices $v_1, \ldots, v_{k-2}$ such that at least $\binom{r}{3} e \cdot f_3(n, e+2+d, e)$ of the edges of H contain $v_1, \ldots, v_{k-2}$. Set $E_0 = \{X \setminus \{v_1, \ldots, v_{k-2}\} : v_1, \ldots, v_{k-2} \in X \in E(H)\}$, noting that $|E_0| \geq \binom{r}{3} e \cdot f_3(n, e+2+d, e)$ and that $|Y| = r - k + 2$ for each $Y \in E_0$. We now consider two cases. Suppose first that there is a triple $T \in \binom{V(H)}{3}$ and distinct $Y_1, \ldots, Y_e \in E_0$ such that $T \subseteq Y_i$ for each $1 \leq i \leq e$. Setting $X_i = Y_i \cup \{v_1, \ldots, v_{k-2}\}$ for each $1 \leq i \leq e$, we observe that $|X_1 \cup \cdots \cup X_e| \leq (r-k-1) \cdot e + k - 2 + 3 \leq (r-k)e + k$. It follows that H contains a (v,e)-configuration with $v \leq (r-k)e + k$, thus completing the proof in this case.

Suppose now that for each $T \in \binom{V(H)}{3}$ it holds that $\#\{Y \in E_0 : T \subseteq Y\} \leq e - 1$. Then, for each $Y \in E_0$, there are at most $\binom{r}{3}(e-1)$ sets $Y' \in E_0 \setminus \{Y\}$ such that $|Y \cap Y'| \geq 3$. This means that there

exists $E_1 \subseteq E_0$ of size

$$|E_1| \geq \frac{|E_0|}{\binom{r}{3}(e-1)+1} > f_3(n, e+2+d, e), \tag{7.7}$$

such that $|Y \cap Y'| \leq 2$ for each pair of distinct $Y, Y' \in E_1$.[3] For each $Y \in E_1$, choose arbitrarily a triple $T_Y \in \binom{Y}{3}$. Let H' be the 3-graph on $V(H)$ whose edge-set is $E(H') = \{T_Y : Y \in E_1\}$. Then $e(H') = |E_1| > f_3(n, e+2+d, e)$, where the equality holds due to our choice of E_1 and the inequality due to (7.7). It follows that H' contains an $(e+2+d, e)$-configuration F. Now observe that the edge-set $\{Y \cup \{v_1, \ldots, v_{k-2}\} : Y \in E_1 \text{ and } T_Y \in E(F)\}$ spans in H a (v, e)-configuration with $v \leq v(F) + (r-k-1)e + k - 2 \leq e + 2 + d + (r-k-1)e + k - 2 = (r-k)e + k + d$, as required.

7.2 Proof of Lemma 7.1.6

In this section we prove Lemma 7.1.6. The construction of the 3-graphs F_k appearing in the statement of the lemma, as well as the proof that these 3-graphs have the required properties, is done by induction on k. The inductive step, which constitutes the main part of the proof of Lemma 7.1.6, is given by the following lemma.

Lemma 7.2.1. *Let F be a nice 3-graph, put $k = v(F) - e(F)$ and assume that $k \geq 3$. Then there exists a nice 3-graph F' such that $v(F') - e(F') = k+1$, $e(F') = (k+1) \cdot e(F)$ and the following holds. For every $r \geq 1$ and $\varepsilon \in (0,1)$, there are $\delta = \delta(F, r, \varepsilon) \in (0,1)$ and $n_0 = n_0(F, r, \varepsilon)$ such that every 3-graph H with $n \geq n_0$ vertices and at least εn^k copies of F satisfies (at least) one of the following:*

1. *There is $k \leq q \leq v(F) - 1$ such that, for every $1 \leq i \leq r$, H contains a (v', e') configuration with $v' - e' \leq k$ and $v' = q + i \cdot (v(F) - q)$.*
2. *H contains at least δn^{k+1} copies of F'.*

Ideally, we would like to start the induction by invoking Lemma 7.2.1 with F being an edge (so $k = \Delta(F) = 2$). As is the case with Lemma 7.1.8 (see the remark following this lemma), Lemma 7.2.1 does in fact work with F being an edge, even though an edge is not nice as per Definition 7.1.5. The 3-graph F' supplied by Lemma 7.2.1 (when applied with F being an edge) is the linear 3-cycle (see Figure 7.1). In fact, applying Lemma 7.2.1 with F being an edge recovers the proof of the (6,3)-theorem, which was discussed in Section 7.1.1. Unfortunately, the linear 3-cycle is not nice (this time in a meaningful way; it really cannot be used as an input to Lemma 7.2.1), preventing us from continuing the induction. To make matters even worse, there is in fact no 3-graph F with difference $k = 3$ which is known to be a viable input to Lemma 7.2.1. Indeed, note that in order for the lemma to be useful when applied with input F, we need to know that F is nice and that it is *abundant* in every sufficiently large n-vertex 3-graph with $\Omega(n^2)$ edges (or at least in every such 3-graph that does not satisfy the conclusion of Theorem 7.1.1 for some other reason). Unfortunately, no such suitable F (of difference 3) is known. Here, we say that a 3-graph F is abundant[4] in an n-vertex 3-graph H if H contains $\Omega(n^{v(F)-e(F)})$ copies of F.

[3] To see that such an E_1 indeed exists, consider an auxiliary graph on E_0 in which Y, Y' are adjacent if and only if $|Y \cap Y'| \geq 3$ and recall the simple fact that every graph G contains an independent set of size at least $\frac{v(G)}{\Delta(G)+1}$ (where $\Delta(G)$ is the maximum degree of G). Now take E_1 to be such an independent set.

[4] In particular, the edge is trivially abundant in every hypergraph with $\Omega(n^2)$ edges and the condition (resp. conclusion) of Lemma 7.2.1 can be stated as saying that F (resp. F') is abundant in H.

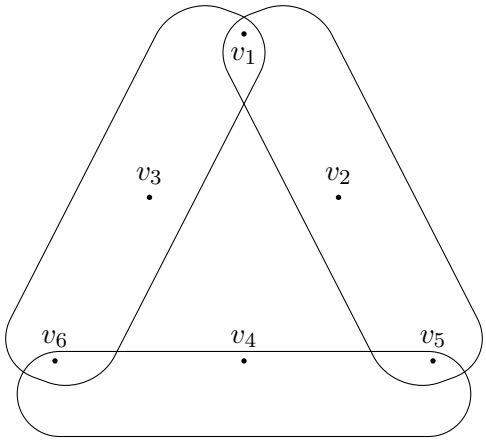

Figure 7.1: The 3-uniform linear 3-cycle

In light of this situation, the base step of our induction will have to involve a nice 3-graph F having difference at least 4. Fortunately, as stated in the following lemma, there does exist a nice F of difference 4 which can be shown to be abundant in every 3-graph H with n vertices and $\Omega(n^2)$ edges, unless H satisfies the assertion of Theorem 7.1.1 for a trivial reason.

Lemma 7.2.2. *There is a nice 3-graph F with $v(F) = 14$ and $e(F) = 10$ having the following property. For every $r \geq 1$ and $\varepsilon \in (0,1)$, there are $\delta = \delta(r,\varepsilon) \in (0,1)$ and $n_0 = n_0(r,\varepsilon)$ such that every 3-graph H with $n \geq n_0$ vertices and at least εn^2 edges satisfies (at least) one of the following:*

1. *For every $1 \leq i \leq r$, H contains a $(3i+3, 3i)$-configuration.*
2. *H contains at least δn^4 copies of F.*

We note that the 3-graph F in the above lemma played a key role in the proof in [107] that $f_3(n, 14, 10) = o(n^2)$. As such, the abundance statement regarding F was already proven in [107]. Consequently, our main task in the proof of Lemma 7.2.2 is to show that F is nice.

The rest of this section is organized as follows. In Section 7.2.1, we derive Lemma 7.1.6 from Lemmas 7.2.1 and 7.2.2. We then prove these two lemmas in Sections 7.2.2 and 7.2.3, respectively.

7.2.1 Deriving Lemma 7.1.6 from Lemmas 7.2.1 and 7.2.2

Let F_3 be the linear 3-cycle (which has 6 vertices and 3 edges). Let F_4 be the nice 3-graph whose existence is guaranteed by Lemma 7.2.2. For each $k \geq 5$, let F_k be the nice 3-graph F' obtained by applying Lemma 7.2.1 with $F := F_{k-1}$. Then it is easy to check by induction that, for every $k \geq 4$, it holds that $v(F_k) - e(F_k) = k$, $e(F_k) = 5k!/12$ and the 3-graph F_k is nice.

Let $r \geq 1$ and $\varepsilon \in (0,1)$. We define a sequence $(\delta_k)_{k \geq 4}$ as follows. Let $\delta_4 = \delta(r,\varepsilon)$ be defined via Lemma 7.2.2 and, for each $k \geq 5$, let $\delta_k = \delta\left(F_{k-1}, r, \delta_{k-1}\right)$ be given by Lemma 7.2.1. We now show by induction on $k \geq 4$ that the assertion of the lemma holds with $\eta = \eta(k,r,\varepsilon) := \delta_k$. For $k = 4$, Lemma 7.2.2 readily implies that H either satisfies the assertion of Item 2 of Lemma 7.1.6 or satisfies the assertion of Item 1 with $j = 3$ and $q = 3$. Let now $k \geq 5$. By the induction hypothesis, H satisfies the assertion of (at least)

one of the items of Lemma 7.1.6 with parameter $k-1$ (in place of k). If this is the case for Item 1, then the same item is also satisfied with parameter k and we are done. Suppose then that H satisfies the assertion of Item 2 (with parameter $k-1$), namely, that H contains at least $\delta_{k-1} \cdot n^{k-1}$ copies of F_{k-1}. Then, by Lemma 7.2.1 (with parameters $F = F_{k-1}$ and δ_{k-1} in place of ε), either H satisfies the assertion of Item 1 in Lemma 7.1.6 (with $j = k-1$) or it contains at least $\delta_k \cdot n^k = \eta(k, r, \varepsilon) \cdot n^k$ copies of F_k, as required by Item 2.

7.2.2 Proof of Lemma 7.2.1

Let $A \subseteq V(F)$ be as in Definition 7.1.5. It will be convenient to set $v := v(F)$ and to assume (without loss of generality) that $V(F) = [v]$ and $A = [k+1] \subseteq [v]$. The required nice 3-graph F' is defined as follows: take vertices $x_1, \ldots, x_{k+1}, x'_1, \ldots, x'_{k+1}$ and, for each $1 \leq i \leq k+1$, add a copy F_i of F in which x_j plays the role of $j \in V(F)$ for each $j \in [k+1] \setminus \{i\}$, x'_i plays the role of $i \in V(F)$ and all other $v(F) - k - 1$ vertices are new.

Let us calculate the number of vertices and edges in F'. First, as $A \subseteq V(F)$ is independent, the copies $F_1, \ldots, F_{k+1}$ (which comprise F') do not share edges. Hence, $e(F') = (k+1) \cdot e(F)$. Second, we have $v(F') = k + 1 + (k+1) \cdot (v(F) - k) = k + 1 + (k+1) \cdot e(F) = e(F') + k + 1$, as required.

We now show that F' is nice. We will show that F' satisfies the requirements of Definition 7.1.5 with respect to the set $A' := \{x'_1, \ldots, x'_{k+1}, x_1\}$. (We remark that in the definition of A' we could replace x_1 with any other vertex among $x_1, \ldots, x_{k+1}$.) For the rest of the proof, we set $X = \{x_1, \ldots, x_{k+1}\}$, $X' = \{x'_1, \ldots, x'_{k+1}\}$ and $A_i = (X \setminus \{x_i\}) \cup \{x'_i\}$ for each $1 \leq i \leq k+1$. Observe that for each $1 \leq i \leq k+1$, the vertices of A_i are precisely the vertices which play the roles of the vertices of $A = \{1, \ldots, k+1\} \subseteq V(F)$ in the copy F_i of F.

It is evident that $|A'| = k+2$ and easy to see that A' is independent in F'. Our goal is then to show that every $U \subseteq V(F')$ satisfies the assertion of Items 1-2 in Definition 7.1.5 (with A' in place of A). So let $U \subseteq V(F')$ and put $U_i = U \cap V(F_i)$ for each $1 \leq i \leq k+1$. Since every vertex of X belongs to exactly k of the copies $F_1, \ldots, F_{k+1}$ and every other vertex of F' belongs to exactly one of these copies, we have

$$|U| = \sum_{i=1}^{k+1} |U_i| - (k-1)|U \cap X|.$$

Since $F_1, \ldots, F_{k+1}$ are pairwise edge-disjoint, we have

$$e(U) = \sum_{i=1}^{k+1} e(U_i).$$

It follows that

$$\Delta(U) = \sum_{i=1}^{k+1} \Delta(U_i) - (k-1)|U \cap X|. \tag{7.8}$$

For each $1 \leq i \leq k+1$, it follows from the niceness of F (and the fact that A_i plays the role of A in the copy F_i of F) that

$$\Delta(U_i) \geq |U_i \cap A_i| - \mathbb{1}_{A_i \subseteq U_i}. \tag{7.9}$$

Setting $s := \#\{1 \le i \le k+1 : A_i \subseteq U_i\}$, we plug (7.9) into (7.8) to obtain

$$\begin{aligned}\Delta(U) &\ge \sum_{i=1}^{k+1} |U_i \cap A_i| - (k-1)|U \cap X| - s = |U \cap X| + |U \cap X'| - s \\ &= |U \cap A'| + |U \cap \{x_2, \ldots, x_{k+1}\}| - s.\end{aligned} \tag{7.10}$$

To see that the first equality in (7.10) holds, note that $A_1 \cup \cdots \cup A_{k+1} = X \cup X'$ and recall that every element of X (resp. X') belongs to exactly k (resp. 1) of the sets $A_1, \ldots, A_{k+1}$.

We first prove that $\Delta(U) \ge |U \cap A'| - \mathbb{1}_{A' \subseteq U}$, as required by Item 1 in Definition 7.1.5. If $s = 0$, then (7.10) readily gives $\Delta(U) \ge |U \cap A'|$. Suppose then that $s \ge 1$ and let $1 \le i \le k+1$ be such that $A_i \subseteq U_i$. Then $\{x_2, \ldots, x_{k+1}\} \setminus \{x_i\} \subseteq U$, implying that $|U \cap \{x_2, \ldots, x_{k+1}\}| \ge k-1$. Furthermore, if $s \ge 2$, then $\{x_2, \ldots, x_{k+1}\} \subseteq U$, in which case $|U \cap \{x_2, \ldots, x_{k+1}\}| = k$. Hence, it follows from (7.10) that $\Delta(U) \ge |U \cap A'| - \mathbb{1}_{s=k+1}$. We also note, for later use, that if $1 \le s \le k-1$ then $\Delta(U) \ge |U \cap A'| + 1$ (here we use the assumption that $k \ge 3$). Observe that if $s = k+1$, then $A_i \subseteq U_i$ for every $1 \le i \le k+1$, implying that $A' \subseteq X \cup X' \subseteq U$. So we indeed have $\Delta(U) \ge |U \cap A'| - \mathbb{1}_{A' \subseteq U}$, as required.

Next, we assume that $|U \cap A'| \le k$ and $U \setminus A' \ne \emptyset$ and show that in this case $\Delta(U) \ge |U \cap A'| + 1$ (as required by Item 2 in Definition 7.1.5). The assumption that $|U \cap A'| \le k$ implies that $s \le k-1$, because if $s \ge k$, then $|U \cap X'| \ge k$ and $x_1 \in U$, which means that $|U \cap A'| \ge k+1$. We already saw that $\Delta(U) \ge |U \cap A'| + 1$ if $1 \le s \le k-1$, so it remains to handle the case that $s = 0$, namely, that $A_i \not\subseteq U_i$ for each $1 \le i \le k+1$. If $U \cap \{x_2, \ldots, x_{k+1}\} \ne \emptyset$, then (7.10) readily implies that $\Delta(U) \ge |U \cap A'| + 1$ (since $s = 0$). So suppose that $U \cap \{x_2, \ldots, x_{k+1}\} = \emptyset$. Since $U \setminus A' \ne \emptyset$, there is $1 \le i \le k+1$ such that $U_i \setminus A' \ne \emptyset$. Our assumption that $U \cap \{x_2, \ldots, x_{k+1}\} = \emptyset$ implies that $|U_i \cap A_i| \le k-1$ and $U_i \setminus A_i \ne \emptyset$ (here we use the fact that $A_i \subseteq A' \cup \{x_2, \ldots, x_{k+1}\}$ and $U_i \setminus A' \ne \emptyset$). Now it follows from the niceness of F (or, more precisely, of the copy F_i of F) that $\Delta(U_i) \ge |U_i \cap A_i| + 1$. Moreover, by (7.9), we have $\Delta(U_j) \ge |U_j \cap A_j|$ for each $1 \le j \le k+1$ (this follows from our assumption that $s = 0$). By plugging all of this into (7.8), in a manner similar to the derivation of (7.10), we obtain

$$\Delta(U) \ge |U_i \cap A_i| + 1 + \sum_{j \in [k+1] \setminus \{i\}} |U_j \cap A_j| - (k-1)|U \cap X| = |U \cap X| + |U \cap X'| + 1 \ge |U \cap A'| + 1,$$

as required.

Having proven that F' is nice, we go on to show that the assertion of the lemma holds. Given $r \ge 1$ and $\varepsilon \in (0,1)$, we set

$$\delta = \delta(F, r, \varepsilon) = \frac{1}{2}\gamma\left(k,\ 2^{-v(1+2^v r)} \cdot v^{-v} \cdot \varepsilon\right)$$

and $n_0 = n_0(F, r, \varepsilon) = 1/\delta$. Here γ is from Theorem 7.1.3 and $v = v(F)$ as before.

Let H be a 3-uniform hypergraph with $n \ge n_0$ vertices and at least εn^k copies of F. Partition the vertices of H randomly into sets $C_1, \ldots, C_v$ by choosing, for each vertex $x \in V(H)$, a part C_i $(1 \le i \le v)$ uniformly at random and independently (of the choices made for all other vertices of H) and placing x in this part. A copy of F in H will be called *good* if, for each $i = 1, \ldots, v$, the vertex playing the role of i in this copy is in C_i. Since H contains at least εn^k copies of F, there are in expectation at least $v^{-v} \cdot \varepsilon n^k$ good copies of F. So fix a partition $C_1, \ldots, C_v$ with at least this number of good copies of F and denote the set of these copies by $\mathcal{F}$. It will be convenient to identify each good copy of F with the corresponding

embedding $\varphi : V(F) \to V(H)$ which maps each $i \in [v] = V(F)$ to a vertex in C_i. So we will assume that the elements of $\mathcal{F}$ are such mappings.

We now define an auxiliary graph $\mathcal{G}$ on $\mathcal{F}$ as follows: for each pair $\varphi_1, \varphi_2 \in \mathcal{F}$, we let $\{\varphi_1, \varphi_2\}$ be an edge in $\mathcal{G}$ if and only if the set $U := U(\varphi_1, \varphi_2) := \{i \in V(F) : \varphi_1(i) = \varphi_2(i)\}$ satisfies either $|U \cap A| \geq k$ or $|U \cap A| = k - 1$ and $U \setminus A \neq \emptyset$. We distinguish between two cases. Suppose first that there is $\varphi \in \mathcal{F}$ whose degree in $\mathcal{G}$ is at least

$$d := 2^{v(1+2^v r)}.$$

Let $\varphi_1, \ldots, \varphi_d$ be distinct neighbors of φ in $\mathcal{G}$. By the pigeonhole principle, there is $I_0 \subseteq [d]$ of size at least $2^{-v}d = 2^{v2^v r}$ and a set $U_0 \subseteq V(F)$ such that, for all $i \in I_0$, it holds that $U(\varphi, \varphi_i) = U_0$. Note that by the definition of $\mathcal{G}$, we have either $|U_0 \cap A| \geq k$ or $|U_0 \cap A| = k-1$ and $U_0 \setminus A \neq \emptyset$. We now consider the complete graph on I_0 and color each edge $\{i, j\} \in \binom{I_0}{2}$ of this graph with color $U(\varphi_i, \varphi_j)$. A well-known bound for multicolor Ramsey numbers (see [36]) implies that in every c-coloring of the edges of the complete graph on c^{cr} vertices, there is a monochromatic complete subgraph on r vertices. Applying this claim with $c = 2^v$, we conclude that there is $I \subseteq I_0$ of size $|I| = r$, and a set $U \subseteq V(F)$, such that $U(\varphi_i, \varphi_j) = U$ for all $\{i, j\} \in \binom{I}{2}$. Observe that for each $\{i, j\} \in \binom{I}{2}$, we have $U = U(\varphi_i, \varphi_j) \supseteq U(\varphi, \varphi_i) \cap U(\varphi, \varphi_j) = U_0$. This implies that either $|U \cap A| \geq k$ or $|U \cap A| = k - 1$ and $U \setminus A \neq \emptyset$. Our choice of A via Definition 7.1.5 implies that in both cases $\Delta(U) \geq k$. Note also that $U \neq V(F)$ because the copies of F corresponding to $(\varphi_i : i \in I)$ are distinct.

We now show that the assertion of Item 1 in the lemma holds. Suppose without loss of generality that $I = \{1, \ldots, r\}$, and write $V_i := \varphi_i(V(F) \setminus U) \subseteq V(H)$ for $1 \leq i \leq r$. Note that $V_1, \ldots, V_r$ are pairwise disjoint. We also put $W := \varphi_1(U) = \cdots = \varphi_r(U)$. Now, fix any $1 \leq i \leq r$ and set $V := V_1 \cup \cdots \cup V_i \cup W$. Then $|V| = |U| + i \cdot (v(F) - |U|) = i \cdot v(F) - (i-1) \cdot |U|$ and $e_H(V) \geq e_F(U) + i \cdot (e(F) - e_F(U)) = i \cdot e(F) - (i-1) \cdot e_F(U)$. It follows that

$$\begin{aligned} |V| - e_H(V) &\leq i \cdot (v(F) - e(F)) - (i-1)(|U| - e_F(U)) = i \cdot k - (i-1) \cdot \Delta(U) \\ &\leq i \cdot k - (i-1) \cdot k = k. \end{aligned}$$

Setting $q := |U|$, we note that $q = |U| \geq \Delta(U) \geq k$ and $q \leq v(F) - 1$ (as $U \neq V(F)$). Now we see that the assertion of Item 1 of the lemma holds with this choice of q. This completes the proof in the case that $\mathcal{G}$ has a vertex of degree at least d.

From now on we assume that the maximum degree of $\mathcal{G}$ is strictly smaller than d and prove that the assertion of Item 2 in the lemma holds. Let $\mathcal{F}^* \subseteq \mathcal{F}$ be an independent set[5] of $\mathcal{G}$ of size at least $v(\mathcal{G})/d = |\mathcal{F}|/d$. Recall that we identify $V(F)$ with $[v]$ and A with $[k+1]$. We now define an auxiliary k-uniform $(k+1)$-partite hypergraph J with parts $C_1, \ldots, C_{k+1}$, as follows. For each $\varphi \in \mathcal{F}^*$, put a k-uniform $(k+1)$-clique in J on the vertices $\varphi(1) \in C_1, \ldots, \varphi(k+1) \in C_{k+1}$. We denote this clique by K_φ. Note that by the definition of J, every edge of J is contained in a copy of F in H, which corresponds to some embedding $\varphi \in \mathcal{F}^*$.

Our first goal is to show that the cliques $(K_\varphi : \varphi \in \mathcal{F}^*)$ are pairwise edge-disjoint. So fix any distinct $\varphi_1, \varphi_2 \in \mathcal{F}^*$ and suppose, for the sake of contradiction, that the cliques $K_{\varphi_1}, K_{\varphi_2}$ share an edge. Then there is $W \subseteq A = [k+1]$ of size $|W| = k$ such that $\varphi_1(i) = \varphi_2(i)$ for every $i \in W$. It follows that

[5]Here we use the simple fact (which was already used in Section 7.1.4) that every graph G has an independent set of size at least $v(G)/(\Delta(G)+1)$, where $\Delta(G)$ is the maximum degree of G.

$W \subseteq U := U(\varphi_1, \varphi_2)$ and hence $|U \cap A| \geq |W| = k$. But this means that φ_1 and φ_2 are adjacent in $\mathcal{G}$, in contradiction to the fact that $\mathcal{F}^*$ is an independent set of $\mathcal{G}$.

We have thus shown that the cliques $(K_\varphi : \varphi \in \mathcal{F}^*)$ are pairwise edge-disjoint. It follows that J contains a collection of $|\mathcal{F}^*| \geq |\mathcal{F}|/d \geq 2^{-v(1+2^v r)} \cdot v^{-v} \cdot \varepsilon n^k$ pairwise edge-disjoint $(k+1)$-cliques. By Theorem 7.1.3 and our choice of $\delta = \delta(F, r, \varepsilon)$, the number of $(k+1)$-cliques in J is at least $2\delta n^{k+1}$.

A $(k+1)$-clique K in J is called *colorful* if it is not equal to K_φ for any $\varphi \in \mathcal{F}^*$. Note that all but at most n^k of the $(k+1)$-cliques in J are colorful (because the non-colorful cliques are pairwise edge-disjoint). It follows that J contains at least $2\delta n^{k+1} - n^k \geq \delta n^{k+1}$ colorful $(k+1)$-cliques (here we use our choice of n_0).

Fix any colorful $(k+1)$-clique $K = \{c_1, \ldots, c_{k+1}\}$, with c_i being the unique vertex in $K \cap C_i$ for each $1 \leq i \leq k+1$. By the definition of J, for each $i \in [k+1]$ there is $\varphi_i \in \mathcal{F}^*$ such that $\varphi_i(j) = c_j$ for every $j \in [k+1] \setminus \{i\}$. We claim that $\varphi_1, \ldots, \varphi_{k+1}$ are pairwise distinct. Suppose, for the sake of contradiction, that $\varphi_i = \varphi_{i'} =: \varphi$ for some $1 \leq i < i' \leq k+1$. Then, for each $1 \leq j \leq k+1$, we have $\varphi(j) = c_j$ because one of i, i' does not equal j. So we see that $K = K_\varphi$, in contradiction to the assumption that K is colorful. We conclude that $\varphi_1, \ldots, \varphi_{k+1}$ are indeed pairwise distinct. It now follows that $\varphi_i(i) \neq c_i$ for each $1 \leq i \leq k+1$. Indeed, if $\varphi_i(i) = c_i$ then, fixing any $j \in [k+1] \setminus \{i\}$, we observe that $\varphi_i(\ell) = \varphi_j(\ell)$ for each $\ell \in [k+1] \setminus \{j\}$, in contradiction to the fact that K_{φ_i} and K_{φ_j} are edge-disjoint.

Recall that F' consists of vertices $x_1, \ldots, x_{k+1}, x'_1, \ldots, x'_{k+1}$ and copies $F_1, \ldots, F_{k+1}$ of F such that the vertex playing the role of $j \in [k+1] \subseteq V(F)$ in F_i is x_j if $j \neq i$ and x'_j if $j = i$ (for every $1 \leq i, j \leq k+1$) and $F_1, \ldots, F_{k+1}$ do not intersect outside of $X = \{x_1, \ldots, x_{k+1}\}$. Now let $\varphi = \varphi_K : V(F') \to V(H)$ be the function which, for each $1 \leq i \leq k+1$, maps x_i to c_i and agrees with φ_i on the vertices of F_i (where we identify $V(F_i)$ with $V(F)$). Then $\varphi(x_i) = c_i$ and $\varphi(x'_i) = \varphi_i(i)$ for each $1 \leq i \leq k+1$. It is not hard to see that in order to show that φ is an embedding of F' into H it is enough to verify that $\mathrm{Im}(\varphi_i) \cap \mathrm{Im}(\varphi_j) = \{c_1, \ldots, c_{k+1}\} \setminus \{c_i, c_j\}$ for each $1 \leq i < j \leq k+1$. So fix any $1 \leq i < j \leq k+1$ and consider the set $U = U(\varphi_i, \varphi_j) = \{\ell \in V(F) : \varphi_i(\ell) = \varphi_j(\ell)\}$. Then $U \cap [k+1] = [k+1] \setminus \{i, j\}$ and, in particular, $|U \cap A| = k-1$. If $U = U \cap [k+1]$, then we are done (because in this case we would have $\mathrm{Im}(\varphi_i) \cap \mathrm{Im}(\varphi_j) = \{c_1, \ldots, c_{k+1}\} \setminus \{c_i, c_j\}$, as required). On the other hand, if $U \neq U \cap [k+1]$, then $U \setminus A \neq \emptyset$, which implies that φ_i and φ_j are adjacent in $\mathcal{G}$, in contradiction to the fact that $\varphi_i, \varphi_j \in \mathcal{F}^*$ and that $\mathcal{F}^*$ is an independent set of $\mathcal{G}$. We have thus shown that each colorful $(k+1)$-clique in J gives rise to a copy of F' in H. It is also easy to see that these copies are pairwise distinct. It follows that H contains at least δn^{k+1} copies of F'.

7.2.3 Proof of Lemma 7.2.2

In the proof of Lemma 7.2.2, we will need the following simple claim that can be verified by exhausting all possible cases. The proof is thus omitted.

Claim 7.2.3. *Consider the 3-uniform linear 3-cycle on vertices $v_1, \ldots, v_6$, as depicted in Figure 7.1, and let $U \subseteq \{v_1, \ldots, v_6\}$. Then $\Delta(U) \geq |U \cap \{v_1, \ldots, v_4\}| - \mathbb{1}_{\{v_1, \ldots, v_4\} \subseteq U}$. Moreover, if $U \setminus \{v_1, \ldots, v_4\} \neq \emptyset$ and either $v_1 \notin U$ or $U \cap \{v_2, v_3\} = \emptyset$, then $\Delta(U) \geq |U \cap \{v_1, \ldots, v_4\}| + 1$.*

Let F denote the 3-uniform linear 3-cycle (see Figure 7.1). Claim 7.2.3 implies that F satisfies Condition 1 in Definition 7.1.5 with respect to $A = \{v_1, \ldots, v_4\}$. However, F does *not* satisfy Condition 2 in that definition, as evidenced, e.g., by the set $U = \{v_1, v_2, v_5\}$. So the "moreover"-part of Claim 7.2.3 can be

thought of as a (non-equivalent) variant of Condition 2 in Definition 7.1.5. We also note that by going over all possible choices of A, one can easily verify that F is *not* nice.

Proof of Lemma 7.2.2. Let F be the 3-graph depicted in Figure 7.2, having vertices

$$w_1, w_2, w_3, w_4, w_1', w_2', w_3', w_4', x_5, x_6, y_5, y_6, z_5, z_6,$$

and edges

$$\{w_1, w_2, x_5\}, \{x_5, w_4', x_6\}, \{x_6, w_3, w_1\}, \{x_5, w_4, y_6\}, \{y_6, w_3', w_1\},$$
$$\{w_1, w_2', y_5\}, \{y_5, w_4, x_6\}, \{w_1', w_2, z_5\}, \{z_5, w_4, z_6\}, \{z_6, w_3, w_1'\}.$$

Then $v(F) = 14$ and $e(F) = 10$. Solymosi and Solymosi [107] (implicitly) proved that for every 3-graph H with $n \geq n_0(r, \varepsilon)$ vertices and at least εn^2 edges, either H satisfies the assertion of Item 1 in the lemma or H contains at least $\delta(r, \varepsilon) \cdot n^4$ copies of F (with $\delta(r, \varepsilon)$ being roughly $\gamma(3, \varepsilon/r)$, where γ is from Theorem 18). So, in order to complete the proof, it is enough to show that F is nice.

We prove that F satisfies the requirements of Definition 7.1.5 with $A := \{w_4, w_1', w_2', w_3', w_4'\}$. To this end, define $V_1 = \{w_1', w_2, z_5, w_4, z_6, w_3\}$, $V_2 = \{w_1, w_2', y_5, w_4, x_6, w_3\}$, $V_3 = \{w_1, w_2, x_5, w_4, y_6, w_3'\}$ and $V_4 = \{w_1, w_2, x_5, w_4', x_6, w_3\}$. Observe that $F[V_i]$ is a linear 3-cycle for every $1 \leq i \leq 4$. Furthermore, considering the vertex-labeling of the linear 3-cycle in Figure 7.1, we see that for each $1 \leq i, j \leq 4$, the role of v_j in $F[V_i]$ is played by w_j if $j \neq i$ and by w_j' if $j = i$. Now fix any $U \subseteq V(F)$ and let us show that U satisfies Items 1-2 in Definition 7.1.5. For each $1 \leq i \leq 4$, define $U_i = U \cap V_i$ and $A_i := (\{w_1, \ldots, w_4\} \setminus \{w_i\}) \cup \{w_i'\}$. Note that by Claim 7.2.3 we have $\Delta(U_i) \geq |U_i \cap A_i| - \mathbb{1}_{A_i \subseteq U_i}$.

Let us now express $\Delta(U)$ in terms of $\Delta(U_1), \ldots, \Delta(U_4)$. It is easy to check that

$$|U| = \sum_{i=1}^{4} |U_i| - 2 \cdot |U \cap \{w_1, \ldots, w_4\}| - |U \cap \{x_5, x_6\}| \tag{7.11}$$

and

$$e(U) = \sum_{i=1}^{4} e(U_i) - \mathbb{1}_{\{w_1, w_2, x_5\} \subseteq U} - \mathbb{1}_{\{w_1, w_3, x_6\} \subseteq U}. \tag{7.12}$$

Setting $r := \sum_{i=1}^{4} (\Delta(U_i) - |U_i \cap A_i|)$ and

$$t := |U \cap \{w_1, w_2, w_3\}| - |U \cap \{x_5, x_6\}| + \mathbb{1}_{\{w_1, w_2, x_5\} \subseteq U} + \mathbb{1}_{\{w_1, w_3, x_6\} \subseteq U},$$

we combine (7.11) and (7.12) to obtain

$$\begin{aligned}
\Delta(U) &= \sum_{i=1}^{4} \Delta(U_i) - 2 \cdot |U \cap \{w_1, \ldots, w_4\}| - |U \cap \{x_5, x_6\}| + \mathbb{1}_{\{w_1, w_2, x_5\} \subseteq U} + \mathbb{1}_{\{w_1, w_2, x_6\} \subseteq U} \\
&= \sum_{i=1}^{4} |U_i \cap A_i| + r - 2 \cdot |U \cap \{w_1, \ldots, w_4\}| - |U \cap \{w_1, w_2, w_3\}| + t \\
&= |U \cap A| + r + t. \qquad (7.13)
\end{aligned}$$

To complete the proof, it is enough to show that $r + t \geq -\mathbb{1}_{A \subseteq U}$ and that $r + t \geq 1$ if $|U \cap A| \leq 3$ and $U \setminus A \neq \emptyset$. In what follows we will frequently use the fact that $\Delta(U_i) \geq |U_i \cap A_i| - \mathbb{1}_{A_i \subseteq U_i}$ for each

$1 \leq i \leq 4$, as mentioned above. We consider two cases, depending on whether $w_1 \in U$ or not. Suppose first that $w_1 \notin U$. In this case we have $t = |U \cap \{w_2, w_3\}| - |U \cap \{x_5, x_6\}|$. Furthermore, $A_i \not\subseteq U_i$ for each $2 \leq i \leq 4$, which implies that $\Delta(U_i) \geq |U_i \cap A_i|$ for these values of i. Note that if $x_5 \in U$, then $U_i \setminus A_i \neq \emptyset$ for $i = 3, 4$, so, by the "moreover"-part of Claim 7.2.3 (and as $w_1 \notin U$), we have $\Delta(U_i) \geq |U_i \cap A_i| + 1$ for these values of i. Similarly, if $x_6 \in U$, then $\Delta(U_i) \geq |U_i \cap A_i| + 1$ for $i = 2, 4$. Altogether, we conclude that $r \geq |U \cap \{x_5, x_6\}| + 1 - \mathbb{1}_{U \cap \{x_5, x_6\} = \emptyset} - \mathbb{1}_{A_1 \subseteq U_1}$ and hence

$$r + t \geq |U \cap \{w_2, w_3\}| + 1 - \mathbb{1}_{U \cap \{x_5, x_6\} = \emptyset} - \mathbb{1}_{A_1 \subseteq U_1}. \tag{7.14}$$

If $A_1 \subseteq U_1$, then $\{w_2, w_3\} \subseteq U$ and hence $r + t \geq 1$. So we assume from now on that $A_1 \not\subseteq U_1$. It then easily follows from (7.14) that $r + t \geq 1$ unless $U \cap \{w_2, w_3, x_5, x_6\} = \emptyset$. Suppose then that $U \cap \{w_2, w_3, x_5, x_6\} = \emptyset$ and note that in this case $r \geq 0$ and $t = 0$, so in particular $r + t \geq 0 \geq -\mathbb{1}_{A \subseteq U}$. Furthermore, if $U \setminus A \neq \emptyset$, then $U \setminus (A_1 \cup \cdots \cup A_4) \neq \emptyset$ (because $U \cap \{w_1, w_2, w_3\} = \emptyset$), so there must be some $1 \leq i \leq 4$ such that $U_i \setminus A_i \neq \emptyset$. Now Claim 7.2.3 implies that $\Delta(U_i) \geq |U_i \cap A_i| + 1$ and hence $r \geq 1$. We conclude that if $U \setminus A \neq \emptyset$, then $r + t \geq 1$, as required.

Having handled the case that $w_1 \notin U$, we assume from now on that $w_1 \in U$. Here we consider several subcases, depending on the intersection of U with $\{w_2, w_3\}$. Suppose first that $U \cap \{w_2, w_3\} = \emptyset$. Then $A_i \not\subseteq U_i$ for each $1 \leq i \leq 4$, implying that $r \geq 0$. Furthermore, $t = 1 - |U \cap \{x_5, x_6\}|$. So if $U \cap \{x_5, x_6\} = \emptyset$, then $r + t \geq 1$ and we are done. On the other hand, if $U \cap \{x_5, x_6\} \neq \emptyset$, then $U_4 \setminus A_4 \neq \emptyset$, which implies, by Claim 7.2.3, that $\Delta(U_4) \geq |U_4 \cap A_4| + 1$. This shows that $r + t \geq 0 \geq -\mathbb{1}_{A \subseteq U}$ and in fact $r + t \geq 1$ if $|U \cap \{x_5, x_6\}| \leq 1$. So from now on we assume that $\{x_5, x_6\} \subseteq U$ and show that $r + t \geq 1$ unless $|U \cap A| \geq 4$. As $\{x_5, x_6\} \subseteq U$, we have $U_i \setminus A_i \neq \emptyset$ for $i = 2, 3$. It now follows from Claim 7.2.3 that for each $i = 2, 3$, if $w_i' \notin U$, then $\Delta(U_i) \geq |U_i \cap A_i| + 1$, which, combined with $\Delta(U_4) \geq |U_4 \cap A_4| + 1$, implies that $r \geq 2$ and hence $r + t \geq 1$. So, we are done unless $w_2', w_3' \in U$. Suppose then that $w_2', w_3' \in U$. If $w_4 \notin U$, then either $U_2 = \{w_1, w_2', x_6\}$ or $U_2 = \{w_1, w_2', y_5, x_6\}$, and in both cases $\Delta(U_2) = 3 = |U_2 \cap A_2| + 1$. But this implies that $r \geq 2$, again giving $r + t \geq 1$. Therefore, we may assume that $w_4 \in U$. Similarly, if $w_4' \notin U$, then $U_4 = \{w_1, x_5, x_6\}$, from which it follows that $\Delta(U_4) = 3 = |U_4 \cap A_4| + 2$ and hence $r \geq 2$. So again, we may assume that $w_4' \in U$. Altogether, we see that $r + t \geq 1$ unless $\{w_2', w_3', w_4, w_4'\} \subseteq U$, which only holds if $|U \cap A| \geq 4$.

Suppose now that $|U \cap \{w_2, w_3\}| = 1$. By symmetry, we may assume without loss of generality that $w_2 \in U$ and $w_3 \notin U$. Then $t = 2 - \mathbb{1}_{x_6 \in U}$ and $A_i \not\subseteq U_i$ for every $i \in \{1, 2, 4\}$. It follows that $r + t \geq 2 - \mathbb{1}_{x_6 \in U} - \mathbb{1}_{A_3 \subseteq U_3}$ and hence $r + t \geq 1$ unless $x_6 \in U$ and $A_3 \subseteq U_3$. Suppose then that $x_6 \in U$ and $\{w_3', w_4\} \subseteq A_3 \subseteq U_3 \subseteq U$. As $x_6 \in U$, we have $U_2 \setminus A_2 \neq \emptyset$. Therefore, if $w_2' \notin U$, then by Claim 7.2.3 we have $\Delta(U_2) \geq |U_2 \cap A_2| + 1$, which implies that $r \geq 0$ and hence $r + t \geq 1$. So we may assume that $w_2' \in U$. Similarly, if $w_4' \notin U$, then either $U_4 = \{w_1, w_2, x_6\}$ or $U_4 = \{w_1, w_2, x_5, x_6\}$. Since in both cases $\Delta(U_4) = |U_4 \cap A_4| + 1$, we infer that if $w_4' \notin U$, then $r \geq 0$ and hence $r + t \geq 1$. Overall, we see that $r + t \geq 1$ unless $\{w_2', w_3', w_4, w_4'\} \subseteq U$, as required.

It remains to handle the case that $\{w_2, w_3\} \subseteq U$. In this case, we have $t = 3$, so $r + t \geq 0$ unless $r = -4$. But if $r = -4$, then $A_i \subseteq U_i$ for each $1 \leq i \leq 4$, which implies that $A \subseteq U$. So we see that $r + t \geq -\mathbb{1}_{A \subseteq U}$, as required. Furthermore, if $|U \cap A| \leq 3$, then $\#\{1 \leq i \leq 4 : A_i \subseteq U_i\} \leq 2$ (indeed, if $A_i \subseteq U_i$ for at least 3 indices $1 \leq i \leq 4$, then $|U \cap \{w_1', \ldots, w_4'\}| \geq 3$ and $w_4 \in U$, implying that $|U \cap A| \geq 4$), so in fact we have $r \geq -2$ and hence $\Delta(U) \geq |U \cap A| + 1$. This completes the proof. ■

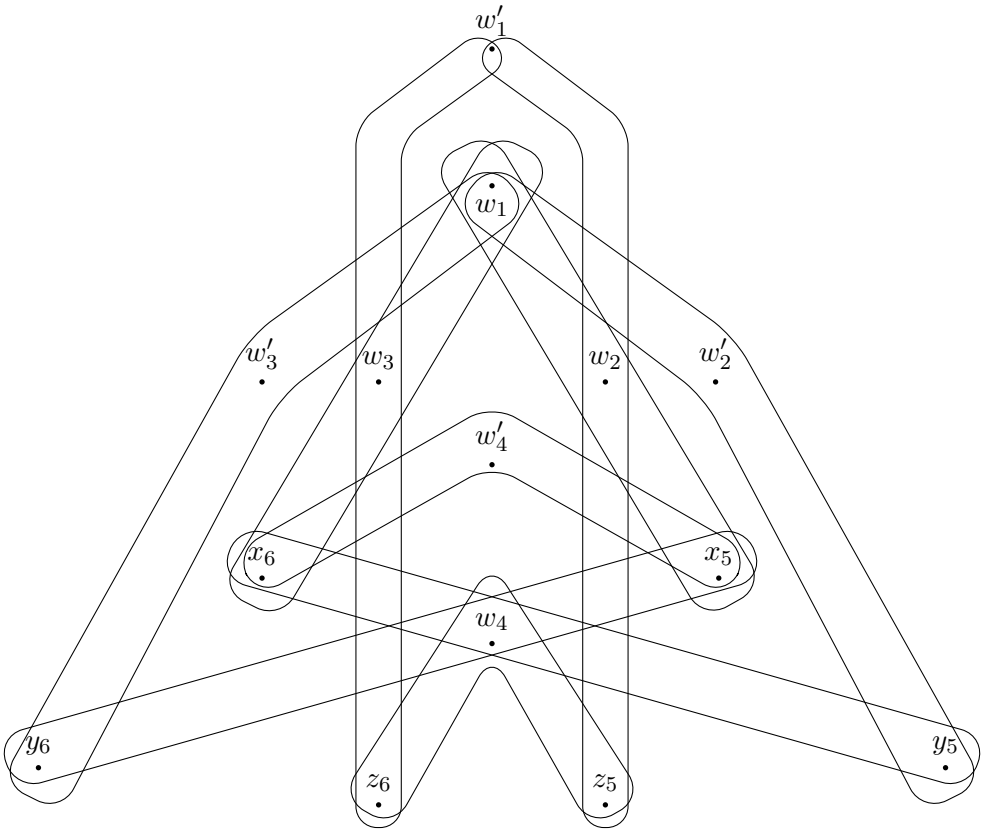

Figure 7.2: The (14,10)-configuration used in Lemma 7.2.2

7.3 Proof of Lemma 7.1.8

In this section, we prove Lemma 7.1.8 through a sequence of claims. We start by defining the 3-graphs $(G_\ell)_{\ell\geq 0}$ appearing in the statement of the lemma. Very roughly speaking, G_ℓ can be thought of as the 3-graph obtained by starting with a complete k-ary tree of height ℓ and replacing each of its vertices by a copy of G.

In each of the graphs G_ℓ we identify a special subset of vertices which will play a crucial role. More precisely, for every $\ell \geq 0$, the graph G_ℓ will contain a subset of vertices $A_\ell \subseteq V(G_\ell)$ which we will denote by $x_1, \ldots, x_k$ and $y_0, \ldots, y_\ell$. If G^* is a copy of some G_ℓ, then we use $x_i(G^*)$ and $y_i(G^*)$ to denote the vertices of G^* playing the roles of x_i and y_i in G^*. We also set $A_\ell(G^*) = \{x_1(G^*), \ldots, x_k(G^*), y_0(G^*), \ldots, y_\ell(G^*)\}$. When both G^* and the value of ℓ are clear from the context, we will simply write $A_\ell, x_1, \ldots, x_k, y_0, \ldots, y_\ell$.

Recall that G is assumed to be nice; so let $A \subseteq V(G)$ be as in Definition 7.1.5, noting that $|A| = k+1$ and that A is an independent set. Assuming the vertices of A are (arbitrarily) named $x_1, \ldots, x_k, y_0$, we now set G_0 to be G, $y_0(G_0)$ to be y_0 and $x_i(G_0)$ to be x_i for every $1 \leq i \leq k$. In particular, $A_0(G_0) = A$. Proceeding by induction, we fix $\ell \geq 1$ and assume that $G_{\ell-1}$, as well as the vertices $x_i(G_{\ell-1})$ and $y_i(G_{\ell-1})$ (and thus also the set $A_{\ell-1}(G_{\ell-1})$), have already been defined. Now G_ℓ is defined as follows. Start with a set of $k+\ell+1$ vertices $x_1, \ldots, x_k$ and $y_0, \ldots, y_\ell$. We will set $x_i(G_\ell)$ to be x_i for every $1 \leq i \leq k$ and $y_i(G_\ell)$ to be y_i for every $0 \leq i \leq \ell$. In addition to these $k+\ell+1$ vertices, we also have k additional vertices $x'_1, \ldots, x'_k$. For each $1 \leq i \leq k$, add a copy of $G_{\ell-1}$, denoted $G^i_{\ell-1}$, in which x_j plays the role of $x_j(G_{\ell-1})$ for each $j \in [k] \setminus \{i\}$, x'_i plays the role of $x_i(G_{\ell-1})$, y_j plays the role of $y_j(G_{\ell-1})$ for each $0 \leq j \leq \ell-1$ and all other $v(G_{\ell-1}) - k - \ell$ vertices are "new". As a last step, add a copy G^ℓ of G in which x_i plays the role of $x_i(G)$ for each $1 \leq i \leq k$, y_ℓ plays the role of $y_0(G)$ and all other $v(G) - k - 1$ vertices are "new". The resulting 3-graph is G_ℓ.

Claim 7.3.1. *For every $\ell \geq 0$, the set $A_\ell(G_\ell) \subseteq V(G_\ell)$ is independent and the graph G_ℓ satisfies the assertion of Item 1 of Lemma 7.1.8.*

Proof. We first prove by induction on ℓ that $A_\ell(G_\ell)$ is an independent set. For $\ell = 0$, this is guaranteed by our choice of $A_0(G_0) = A$. So fixing $\ell \geq 1$ and assuming the claim holds for $\ell - 1$, we now prove it for ℓ. By the definition of G_ℓ, each edge of G_ℓ belongs to one of the 3-graphs $G^1_{\ell-1}, \ldots, G^k_{\ell-1}, G^\ell$. Moreover, we have $V(G^i_{\ell-1}) \cap A_\ell(G_\ell) \subseteq A_{\ell-1}(G^i_{\ell-1})$ for every $1 \leq i \leq k$ and similarly $V(G^\ell) \cap A_\ell(G_\ell) = A_0(G^\ell)$. So the fact that $A_\ell(G_\ell)$ is independent follows from the induction hypothesis for $\ell - 1$ and from the case $\ell = 0$.

Since $A_\ell(G_\ell)$ is independent, the subgraphs $G^1_{\ell-1}, \ldots, G^k_{\ell-1}, G^\ell$, which comprise G_ℓ, are pairwise edge-disjoint. This implies that $e(G_\ell) = k \cdot e(G_{\ell-1}) + e(G)$. We now prove the two assertions of Item 1 of the lemma by induction on ℓ. The case $\ell = 0$ is immediate. As for the induction step, observe that for each $\ell \geq 1$, we have

$$e(G_\ell) = k \cdot e(G_{\ell-1}) + e(G) = \left(k \cdot \frac{k^\ell - 1}{k-1} + 1\right) \cdot e(G) = \frac{k^{\ell+1} - 1}{k-1} \cdot e(G),$$

where the second equality follows from the induction hypothesis for $\ell - 1$. Moreover, we have

$$\begin{aligned} v(G_\ell) &= 2k + \ell + 1 + k \cdot (v(G_{\ell-1}) - k - \ell) + v(G) - k - 1 \\ &= k + \ell + k \cdot (v(G_{\ell-1}) - k - \ell + 1) + v(G) - k \\ &= k + \ell + k \cdot e(G_{\ell-1}) + e(G) = k + \ell + e(G_\ell). \end{aligned}$$

Here we used the fact that $\Delta(G) = k$ and the induction hypothesis that $\Delta(G_{\ell-1}) = k + \ell - 1$. The above two expressions for $e(G_\ell)$ and $v(G_\ell)$ imply both assertions of Item 1. ■

Item 2 of Lemma 7.1.8 follows from the following stronger claim.

Claim 7.3.2. *Let $\ell \geq 1$ and $e(G_{\ell-1})/e(G) < t \leq e(G_\ell)/e(G)$. Then there is a subgraph G' of G_ℓ such that $v(G') - e(G') \leq k + \ell$, $e(G') = t \cdot e(G)$ and $A_\ell(G_\ell) \subseteq V(G')$.*

Before proving Claim 7.3.2, let us use this claim to establish the assertion of Item 2 of the lemma by induction on ℓ. The case $\ell = 0$ is trivial, so let $\ell \geq 1$ and $1 \leq t \leq e(G_\ell)/e(G)$. If $t > e(G_{\ell-1})/e(G)$, then the assertion of Item 2 follows from Claim 7.3.2 and if $t \leq e(G_{\ell-1})/e(G)$, then it follows from the induction hypothesis for $\ell - 1$ and the fact that G_ℓ contains a copy of $G_{\ell-1}$.

In the proof of Claim 7.3.2, we will need the following simple claim. Recall that $G^1_{\ell-1}, \ldots, G^k_{\ell-1}$ are the copies of $G_{\ell-1}$ which feature in the definition of G_ℓ.

Claim 7.3.3. *Let $0 \leq \ell' < \ell$. Then G_ℓ contains a copy G^* of $G_{\ell'}$ such that $V(G^*) \subseteq V(G^k_{\ell-1})$, $x_i(G^*) = x_i(G_\ell)$ for each $1 \leq i \leq k-1$ and $y_i(G^*) = y_i(G_\ell)$ for each $0 \leq i \leq \ell'$.*

Proof. The proof is by induction on ℓ, with the base case $\ell = 0$ holding vacuously. Let $0 \leq \ell' < \ell$. If $\ell' = \ell - 1$ then $G^* = G^k_{\ell-1}$ is easily seen to satisfy the requirements of the claim. Suppose then that $\ell' \leq \ell - 2$. By the induction hypothesis, $G_{\ell-1}$ contains a copy G^{**} of $G_{\ell'}$ such that $x_i(G^{**}) = x_i(G_{\ell-1})$ for each $1 \leq i \leq k-1$ and $y_i(G^{**}) = y_i(G_{\ell-1})$ for each $0 \leq i \leq \ell'$. Let G^* be the subgraph playing the role of G^{**} in the copy $G^k_{\ell-1}$ of $G_{\ell-1}$. Then it is evident that $V(G^*) \subseteq V(G^k_{\ell-1})$. Moreover, for each $1 \leq i \leq k-1$, we have $x_i(G^*) = x_i(G^k_{\ell-1}) = x_i(G_\ell)$, where the first equality follows from our choice of G^* and the second equality follows from the definition of G_ℓ. A similar argument shows that $y_i(G^*) = y_i(G^k_{\ell-1}) = y_i(G_\ell)$ for each $0 \leq i \leq \ell'$. ■

Proof of Claim 7.3.2. The proof is by induction on ℓ. We start with the base case $\ell = 1$. Let $1 < t \leq e(G_1)/e(G) = k + 1$. Recall that $G^1_0, \ldots, G^k_0$ and G^1 are the copies of $G_0 = G$ which feature in the definition of G_1. Let G' be the subgraph of G_1 consisting of $G^1_0, \ldots, G^{t-1}_0$ and G^1. Then $e(G') = (t-1) \cdot e(G) + e(G) = t \cdot e(G)$. Moreover, $A_1(G_1) = \{x_1(G_1), \ldots, x_k(G_1), y_0(G_1), y_1(G_1)\} \subseteq V(G')$ because $\{x_1(G_1), \ldots, x_k(G_1), y_1(G_1)\} \subseteq V(G^1) \subseteq V(G')$ and $y_0(G_1) \in V(G^1_0) \subseteq V(G')$ (here we are using the fact that $t \geq 2$). Finally, note that

$$v(G') = |A_1(G_1)| + (t-1) \cdot (v(G) - k) + (v(G) - k - 1) = k + 1 + t \cdot e(G) = e(G') + k + 1,$$

as required.

Now let $\ell \geq 2$ and let t be such that

$$(k^\ell - 1)/(k-1) = e(G_{\ell-1})/e(G) < t \leq e(G_\ell)/e(G) = (k^{\ell+1} - 1)/(k-1).$$

Here the equalities follow from Item 1 of the lemma. Let d be the unique integer satisfying

$$d \cdot (k^\ell - 1)/(k-1) + 1 \leq t < (d+1) \cdot (k^\ell - 1)/(k-1) + 1$$

and note that $1 \leq d \leq k$, where the first inequality follows from the assumption $t > (k^\ell - 1)/(k-1)$ and the second inequality follows from the assumption $t \leq (k^{\ell+1} - 1)/(k-1) = k \cdot (k^\ell - 1)/(k-1) + 1$. Set

$$t' = t - d \cdot (k^\ell - 1)/(k-1) - 1, \tag{7.15}$$

noting that $0 \leq t' < (k^\ell - 1)/(k-1)$. Observe also that if $d = k$ then $t = k \cdot (k^\ell - 1)/(k-1) + 1$ (as t is never larger than this number), in which case $t' = 0$.

Let $\ell' \geq 0$ be the minimum integer satisfying $t' \leq e(G_{\ell'})/e(G) = (k^{\ell'+1} - 1)/(k-1)$. Note that $\ell' < \ell$ because $t' < (k^\ell - 1)/(k-1)$. We now show that there is a subgraph G'' of $G^k_{\ell-1}$ such that $v(G'') - e(G'') \leq k + \ell'$, $e(G'') = t' \cdot e(G)$ and $|V(G'') \cap A_\ell(G_\ell)| \geq k + \ell'$. We begin by noting that by Claim 7.3.3, G_ℓ contains a copy G^* of $G_{\ell'}$ such that $V(G^*) \subseteq V(G^k_{\ell-1})$, $x_i(G^*) = x_i(G_\ell)$ for each $1 \leq i \leq k-1$ and $y_i(G^*) = y_i(G_\ell)$ for each $0 \leq i \leq \ell'$. In particular, $|A_{\ell'}(G^*) \cap A_\ell(G_\ell)| \geq k + \ell'$. If $t' = 0$, then take G'' to be the empty graph on the vertex set $\{x_1(G_\ell), \ldots, x_{k-1}(G_\ell), y_0(G_\ell)\} \subseteq V(G^k_{\ell-1})$, noting that this G'' satisfies all required properties. If $t' = 1$, in which case $\ell' = 0$, then take G'' to be G^* (which is isomorphic to $G_{\ell'} = G_0 = G$). Then $v(G'') - e(G'') = v(G) - e(G) = k$, $e(G'') = e(G)$ and $|V(G'') \cap A_\ell(G_\ell)| \geq |A_{\ell'}(G^*) \cap A_\ell(G_\ell)| \geq k$, as required. Next, suppose that $t' > 1$, which means that $\ell' \geq 1$. Then $e(G_{\ell'-1})/e(G) < t' \leq e(G_{\ell'})/e(G)$ by our choice of ℓ'. By the induction hypothesis for ℓ' (which we apply to the copy G^* of $G_{\ell'}$), there is a subgraph G'' of G^* such that $v(G'') - e(G'') \leq k + \ell'$, $e(G'') = t' \cdot e(G)$ and $A_{\ell'}(G^*) \subseteq V(G'')$. In particular, we have $|V(G'') \cap A_\ell(G_\ell)| \geq |A_{\ell'}(G^*) \cap A_\ell(G_\ell)| \geq k + \ell'$, as required. This shows that a subgraph G'' of $G^k_{\ell-1}$ with the desired properties indeed exists.

Now, let G' be the subgraph of G_ℓ consisting of G^ℓ, $G^1_{\ell-1}, \ldots, G^d_{\ell-1}$ and the 3-graph G'' as in the previous paragraph. Recall that G'' is a subgraph of $G^k_{\ell-1}$ and that if $d = k$ then $t' = 0$ and hence $e(G'') = 0$. It follows that G'' is edge-disjoint from $G^1_{\ell-1}, \ldots, G^d_{\ell-1}, G^\ell$, which are themselves pairwise edge-disjoint by the definition of G_ℓ. This in turn implies that

$$e(G') = d \cdot e(G_{\ell-1}) + e(G) + e(G'') = \left(d \cdot \frac{k^\ell - 1}{k-1} + 1 + t' \right) \cdot e(G) = t \cdot e(G). \tag{7.16}$$

Here, the second equality follows from Item 1 of the lemma and from our choice of G'', while the last equality uses our choice of t' in (7.15). Next, we observe that $A_\ell(G_\ell) \subseteq V(G')$. Indeed, this follows from the fact that $A_\ell(G_\ell) \setminus \{x_1(G_\ell), y_\ell(G_\ell)\} \subseteq V(G^1_{\ell-1}) \subseteq V(G')$ (recall that $d \geq 1$) and that $x_1(G_\ell), y_\ell(G_\ell) \in V(G^\ell) \subseteq V(G')$. Finally, it remains to estimate $v(G') - e(G')$. To this end, note that

$$\begin{aligned} v(G') &= |A_\ell(G_\ell)| + d \cdot (v(G_{\ell-1}) - k - \ell + 1) + (v(G) - k - 1) + |V(G'') \setminus A_\ell(G_\ell)| \\ &\leq k + \ell + d \cdot (v(G_{\ell-1}) - k - \ell + 1) + (v(G) - k) + (v(G'') - k - \ell') \\ &\leq k + \ell + d \cdot e(G_{\ell-1}) + e(G) + e(G'') = e(G') + k + \ell, \end{aligned}$$

where in the first equality we used the definition of G'; in the first inequality we used the fact that $|A_\ell(G_\ell)| = k + \ell + 1$ and $|V(G'') \cap A_\ell(G_\ell)| \geq k + \ell'$; in the second inequality we used the guarantees of Item 1 of the lemma and the fact that $v(G'') - e(G'') \leq k + \ell'$; and in the last equality we used (7.16). We have thus shown that $v(G') - e(G') \leq k + \ell$. This completes the proof of the claim. ■

The rest of this section is devoted to establishing Item 3 of Lemma 7.1.8. To this end, we first prove the following claim, which shows that the niceness of G (with respect to the set A) is carried over to some extent to all G_ℓ. From now on, we will write $A_\ell = \{x_1, \ldots, x_k, y_0, \ldots, y_\ell\}$ (omitting G_ℓ from the notation). We also set $X := \{x_1, \ldots, x_k\}$.

Claim 7.3.4. *Let $\ell \geq 0$ and let $U \subseteq V(G_\ell)$ be such that $\{y_0, \ldots, y_{\ell-1}\} \subseteq U$. Then*

1. *$\Delta(U) \geq |U \cap A_\ell| - \mathbb{1}_{\{x_1, \ldots, x_k, y_\ell\} \subseteq U}$. In particular, if $|U \cap A_\ell| \geq k + \ell$, then $\Delta(U) \geq k + \ell$.*

2. *If* $|U \cap X| \leq k-2$ *and* $U \setminus A_\ell \neq \emptyset$*, then* $\Delta(U) \geq |U \cap A_\ell| + 1$.

3. *If* $|U \cap X| \geq k-1$ *and* $U \cap V(G^\ell)$ *is not contained in* X*, then* $\Delta(U) \geq k + \ell$.

Proof. We first prove Items 1-2 by induction on ℓ and then use these items to derive Item 3. In the base case $\ell = 0$, Items 1-2 immediately follow from the fact that $G_0 = G$ is nice and from our choice of $A_0 = A$ via Definition 7.1.5. Let now $\ell \geq 1$ and let $U \subseteq V(G_\ell)$. We start with Item 1. For $1 \leq i \leq k$, put $U_i := U \cap V(G^i_{\ell-1})$. Similarly, put $U_0 := U \cap V(G^\ell)$ and note that

$$|U \cap A_\ell| = |U_0 \cap \{x_1, \ldots, x_k, y_\ell\}| + \ell \tag{7.17}$$

because $y_0, \ldots, y_{\ell-1} \in U$ by assumption. Since A_ℓ is independent (see Claim 7.3.1), we have $e(U) = \sum_{i=0}^{k} e(U_i)$. Observe also that

$$|U| = \sum_{i=0}^{k} |U_i| - (k-1) \cdot (|U \cap X| + |U \cap \{y_0, \ldots, y_{\ell-1}\}|),$$

as each element of $X \cup \{y_0, \ldots, y_{\ell-1}\}$ is contained in exactly k of the sets $V(G^1_{\ell-1}), \ldots, V(G^k_{\ell-1}), V(G^\ell)$ and each of the other vertices of G_ℓ is contained in exactly one of these sets. From the above formulas for $e(U)$ and $|U|$, it follows that

$$\Delta(U) = \sum_{i=0}^{k} \Delta(U_i) - (k-1) \cdot (|U \cap X| + \ell). \tag{7.18}$$

Here we used the fact that $\{y_0, \ldots, y_{\ell-1}\} \subseteq U$ by assumption. Recall that by the definition of G_ℓ, for each $1 \leq i \leq k$, we have

$$A_{\ell-1}(G^i_{\ell-1}) = \{x_1, \ldots, x_k, y_0, \ldots, y_{\ell-1}, x'_i\} \setminus \{x_i\}.$$

By the induction hypothesis for $\ell - 1$, applied to the copy $G^i_{\ell-1}$ of $G_{\ell-1}$, we get

$$\Delta(U_i) \geq |U_i \cap A_{\ell-1}(G^i_{\ell-1})| - \mathbb{1}_{A_{\ell-1}(G^i_{\ell-1}) \subseteq U_i} \geq |U_i \cap (A_\ell \setminus \{x_i, y_\ell\})|, \tag{7.19}$$

where the second inequality follows by considering whether $x'_i \in U_i$ or not. From (7.19), we obtain

$$\begin{aligned} \sum_{i=1}^{k} \Delta(U_i) &\geq \sum_{i=1}^{k} |U_i \cap (A_\ell \setminus \{x_i, y_\ell\})| \\ &= (k-1) \cdot |U \cap X| + k \cdot |U \cap \{y_0, \ldots, y_{\ell-1}\}| \\ &= (k-1) \cdot |U \cap X| + k\ell, \end{aligned} \tag{7.20}$$

where in the first equality we used the fact that each element of X belongs to exactly $k-1$ of the sets $A_\ell \setminus \{x_i, y_\ell\}$ (where $1 \leq i \leq k$) and each element of $\{y_0, \ldots, y_{\ell-1}\}$ belongs to all of these sets. Plugging the above into (7.18) gives

$$\Delta(U) \geq \Delta(U_0) + \ell. \tag{7.21}$$

Since G is nice and G^ℓ is a copy of G in which y_ℓ plays the role of $y_0(G)$, we have

$$\Delta(U_0) \geq |U_0 \cap \{x_1, \ldots, x_k, y_\ell\}| - \mathbb{1}_{\{x_1, \ldots, x_k, y_\ell\} \subseteq U_0}. \tag{7.22}$$

By combining (7.17), (7.21) and (7.22), we get

$$\Delta(U) \geq \Delta(U_0) + \ell \geq |U_0 \cap \{x_1, \ldots, x_k, y_\ell\}| - \mathbb{1}_{\{x_1,\ldots,x_k,y_\ell\} \subseteq U} + \ell = |U \cap A_\ell| - \mathbb{1}_{\{x_1,\ldots,x_k,y_\ell\} \subseteq U},$$

thus establishing Item 1.

Next, we prove Item 2. Suppose then that $|U \cap X| \leq k-2$ and $U \setminus A_\ell \neq \emptyset$. The inequality $|U \cap X| \leq k-2$ implies that $|U_0 \cap \{x_1, \ldots, x_k, y_\ell\}| \leq k-1$ and that $A_{\ell-1}(G^i_{\ell-1}) \not\subseteq U_i$ for each $1 \leq i \leq k$. Since $U \setminus A_\ell \neq \emptyset$, there is $0 \leq i \leq k$ such that $U_i \setminus A_\ell \neq \emptyset$. Suppose first that $i = 0$. Then $U_0 \setminus \{x_1, \ldots, x_k, y_\ell\} \neq \emptyset$, which, combined with $|U_0 \cap \{x_1, \ldots, x_k, y_\ell\}| \leq k-1$, implies that $\Delta(U_0) \geq |U_0 \cap \{x_1, \ldots, x_k, y_\ell\}| + 1$. Here we used the niceness of G (see Item 2 in Definition 7.1.5). By plugging our bound on $\Delta(U_0)$ into (7.21) and using (7.17), we get $\Delta(U) \geq \Delta(U_0) + \ell \geq |U_0 \cap \{x_1, \ldots, x_k, y_\ell\}| + 1 + \ell = |U \cap A_\ell| + 1$, as required. Now suppose that $1 \leq i \leq k$. We claim that

$$\Delta(U_i) \geq |U_i \cap (A_\ell \setminus \{x_i, y_\ell\})| + 1. \tag{7.23}$$

In other words, we show that the inequality bounding the leftmost term in (7.19) by the rightmost one is strict. If $x'_i \in U_i$, then

$$\Delta(U_i) \geq |U_i \cap A_{\ell-1}(G^i_{\ell-1})| - \mathbb{1}_{A_{\ell-1}(G^i_{\ell-1}) \subseteq U_i} = |U_i \cap A_{\ell-1}(G^i_{\ell-1})| \geq |U_i \cap (A_\ell \setminus \{x_i, y_\ell\})| + 1,$$

as required. Here, in the first inequality we used (7.19), in the equality we used the fact that $A_{\ell-1}(G^i_{\ell-1}) \not\subseteq U_i$ (as mentioned above) and in the last inequality we used the fact that $x'_i \in A_{\ell-1}(G^i_{\ell-1}) \setminus A_\ell$. So suppose now that $x'_i \notin U_i$ and note that in this case $U_i \setminus A_{\ell-1}(G^i_{\ell-1}) \neq \emptyset$ because $U_i \setminus A_\ell \neq \emptyset$ and $A_{\ell-1}(G^i_{\ell-1}) \subseteq A_\ell \cup \{x'_i\}$. Moreover, the intersection of U_i with the set $\{x_1(G^i_{\ell-1}), \ldots, x_k(G^i_{\ell-1})\} = \{x_1, \ldots, x_k, x'_i\} \setminus \{x_i\}$ is of size at most $k-2$, because $|U \cap X| \leq k-2$. So by the induction hypothesis, applied to the copy $G^i_{\ell-1}$ of $G_{\ell-1}$, we have

$$\Delta(U_i) \geq |U_i \cap A_{\ell-1}(G^i_{\ell-1})| + 1 \geq |U_i \cap (A_\ell \setminus \{x_i, y_\ell\})| + 1,$$

where the last inequality uses (7.19). We have thus proven (7.23). By repeating the calculation in (7.20) and plugging in (7.23) and (7.19) (which we use for each $j \in [k] \setminus \{i\}$), we obtain

$$\begin{aligned}
\Delta(U) &= \sum_{i=0}^{k} \Delta(U_i) - (k-1) \cdot (|U \cap X| + \ell) \geq \Delta(U_0) + \ell + 1 \\
&\geq |U_0 \cap \{x_1, \ldots, x_k, y_\ell\}| + \ell - \mathbb{1}_{\{x_1,\ldots,x_k,y_\ell\} \subseteq U_0} + 1 \\
&= |U_0 \cap \{x_1, \ldots, x_k, y_\ell\}| + \ell + 1 = |U \cap A_\ell| + 1.
\end{aligned}$$

Here, the second inequality uses (7.22) and the last equality uses (7.17). This completes the inductive proof of Items 1-2.

It remains to deduce Item 3 from Items 1-2. Suppose then that $|U \cap X| \geq k-1$ and that $U_0 \not\subseteq X$. If $X \subseteq U$ or $y_\ell \in U$, then $|U \cap A_\ell| \geq k + \ell$, in which case Item 1 implies that $\Delta(U) \geq k + \ell$, as required. So we may assume that $|U \cap X| = k-1$ and $y_\ell \notin U$. Since U_0 is not contained in X, we must have $U_0 \setminus \{x_1, \ldots, x_k, y_\ell\} \neq \emptyset$. So by the niceness of G we have $\Delta(U_0) \geq |U_0 \cap \{x_1, \ldots, x_k, y_\ell\}| + 1 = k$. Plugging this into (7.21) gives $\Delta(U) \geq k + \ell$, as required. ■

Item 3 of Lemma 7.1.8 will be derived from the following claim, in a manner similar to the derivation of Lemma 7.1.6 from Lemma 7.2.1.

Claim 7.3.5. *For every $\ell \geq 0$, $r \geq 0$ and $\varepsilon \in (0,1)$, there are $\delta = \delta(\ell, r, \varepsilon)$ and $n_0 = n_0(\ell, r, \varepsilon)$ such that, for every 3-graph H on $n \geq n_0$ vertices, if H contains at least $\varepsilon n^{k+\ell}$ copies of G_ℓ, then (at least) one of the following conditions is satisfied:*

1. *There is $k + \ell \leq q \leq v(G_\ell) - 1$ such that, for every $1 \leq i \leq r$, the 3-graph H contains a (v', e')-configuration which contains a copy of G_ℓ, where $v' - e' \leq k + \ell$ and $v' = q + i \cdot (v(G_\ell) - q)$.*
2. *H contains at least $\delta \cdot n^{k+\ell+1}$ copies of $G_{\ell+1}$.*

Proof. We proceed similarly to the proof of Lemma 7.2.1. Fixing $\ell \geq 0$, we set $v := v(G_\ell)$,

$$\zeta := 2^{-v(1+2^v r)} \cdot v^{-v} \cdot \varepsilon,$$

$\delta = \delta(\ell, r, \varepsilon) = \frac{\zeta}{4} \cdot \gamma\left(k, \frac{\zeta}{2}\right)$ and $n_0 = n_0(\ell, r, \varepsilon) = \frac{2}{\gamma(k, \frac{\zeta}{2})}$, where γ is from Theorem 7.1.3.

Let H be a 3-graph on $n \geq n_0$ vertices, which contains at least $\varepsilon n^{k+\ell}$ copies of G_ℓ. Partition the vertices of H randomly into sets $(C_z : z \in V(G_\ell))$ by choosing, for each vertex $x \in V(H)$, a vertex $z \in V(G_\ell)$ uniformly at random and independently (of the choices made for all other vertices of H) and placing x in part C_z. A copy of G_ℓ in H will be called *good* if, for each $z \in V(G_\ell)$, the vertex playing the role of z in this copy belongs to C_z. Since H contains at least $\varepsilon n^{k+\ell}$ copies of G_ℓ, there are in expectation at least $v^{-v} \cdot \varepsilon n^{k+\ell}$ good copies of G_ℓ. So fix a partition $(C_z : z \in V(G_\ell))$ with at least this number of good copies of G_ℓ and denote the set of these copies by $\mathcal{F}$. We will identify each good copy of G_ℓ with the corresponding embedding $\varphi : V(G_\ell) \to V(H)$, while noting that $\varphi(z) \in C_z$ for each $z \in V(G_\ell)$. Recall that G^ℓ is the copy of G featured in the definition of G_ℓ. Define an auxiliary graph $\mathcal{G}$ on $\mathcal{F}$ as follows. For each pair of distinct $\varphi_1, \varphi_2 \in \mathcal{F}$, we set $U(\varphi_1, \varphi_2) := \{z \in V(G_\ell) : \varphi_1(z) = \varphi_2(z)\}$ and let $\{\varphi_1, \varphi_2\}$ be an edge in $\mathcal{G}$ if and only if $U := U(\varphi_1, \varphi_2)$ satisfies $\{y_0, \ldots, y_{\ell-1}\} \subseteq U$, as well as (at least) one of the following three conditions:

(i) $|U \cap A_\ell| \geq k + \ell$.

(ii) $y_\ell \in U$ and either $|U \cap X| \geq k - 1$ or $|U \cap X| = k - 2$ and $U \setminus A_\ell \neq \emptyset$.

(iii) $|U \cap X| \geq k - 1$ and $U \cap V(G^\ell)$ is not contained in X.

Suppose first that there is $\varphi \in \mathcal{F}$ whose degree in $\mathcal{G}$ is at least

$$d := 2^{v(1+2^v r)}.$$

Let $\varphi_1, \ldots, \varphi_d$ be distinct neighbors of φ in $\mathcal{G}$. By the pigeonhole principle, there is $I' \subseteq [d]$ of size at least $2^{-v} d = 2^{v 2^v r}$ and a set $U' \subseteq V(G_\ell)$ such that, for all $i \in I'$, it holds that $U(\varphi, \varphi_i) = U'$. As in the proof of Lemma 7.2.1, we consider the coloring $\{i, j\} \mapsto U(\varphi_i, \varphi_j)$ of the pairs $\{i, j\} \in \binom{I'}{2}$ and use a bound for multicolor Ramsey numbers [36] to obtain a set $I \subseteq I'$ of size $|I| = r$ and a set $U \subseteq V(G_\ell)$ such that $U(\varphi_i, \varphi_j) = U$ for all $\{i, j\} \in \binom{I}{2}$. Observe that for each $\{i, j\} \in \binom{I}{2}$, we have $U \supseteq U(\varphi, \varphi_i) \cap U(\varphi, \varphi_j) = U'$. In particular, $\{y_0, \ldots, y_{\ell-1}\} \subseteq U' \subseteq U$ (by the definition of $\mathcal{G}$). Note also that $U \neq V(G_\ell)$ because the copies $(\varphi_i : i \in I)$ of G_ℓ are distinct.

We now use Claim 7.3.4 to prove that $\Delta(U) \geq k+\ell$. The definition of the graph $\mathcal{G}$ implies that the set U' must satisfy one of the conditions (i)-(iii) above. Note that for each of these three conditions, if it is satisfied by U', then it is also satisfied by every superset of U' and, in particular, by U. Now, if U satisfies Condition (i) (resp. (iii)), then the bound $\Delta(U) \geq k+\ell$ immediately follows from Item 1 (resp. 3) of Claim 7.3.4. Suppose then that U satisfies Condition (ii). If $|U \cap X| \geq k-1$, then $|U \cap A_\ell| \geq k+\ell$ (since Condition (ii) supposes that $y_\ell \in U$), so again we can apply Item 1 of Claim 7.3.4. Finally, if $|U \cap X| = k-2$ and $U \setminus A_\ell \neq \emptyset$, then we have $\Delta(U) \geq |U \cap A_\ell| + 1 = k+\ell$, where the inequality is given by Item 2 of Claim 7.3.4 and the equality holds because $\{y_0, \ldots, y_\ell\} \subseteq U$ and $|U \cap X| = k-2$. We have thus shown that $\Delta(U) \geq k+\ell$ in all cases.

Suppose without loss of generality that $I = [r]$. Put $W := \varphi_1(U) = \cdots = \varphi_r(U)$ and denote $V_i := \varphi_i(V(G_\ell) \setminus U) \subseteq V(H)$ for each $1 \leq i \leq r$. Note that $V_1, \ldots, V_r$ are pairwise disjoint. Now, fix any $1 \leq i \leq r$ and set $V := V_1 \cup \cdots \cup V_i \cup W$. Then

$$|V| = |U| + i \cdot (v(G_\ell) - |U|) = i \cdot v(G_\ell) - (i-1) \cdot |U|$$

and

$$e_H(V) \geq e(U) + i \cdot (e(G_\ell) - e(U)) = i \cdot e(G_\ell) - (i-1) \cdot e(U).$$

It follows that

$$\begin{aligned} |V| - e_H(V) &\leq i \cdot (v(G_\ell) - e(G_\ell)) - (i-1)(|U| - e(U)) = i \cdot (k+\ell) - (i-1) \cdot \Delta(U) \\ &\leq i \cdot (k+\ell) - (i-1) \cdot (k+\ell) = k+\ell. \end{aligned}$$

Moreover, it is evident that $H[V]$ contains a copy of G_ℓ. Finally, note that $|U| \geq \Delta(U) \geq k+\ell$ and $|U| \leq v(G_\ell) - 1$ (because $U \neq V(G_\ell)$, as mentioned above). Combining all the above, we see that the assertion of Item 1 in the claim holds with $q := |U|$. This completes the proof in the case that $\mathcal{G}$ has a vertex of degree at least d.

From now on we assume that the maximum degree of $\mathcal{G}$ is strictly smaller than d. Let $\mathcal{F}^* \subseteq \mathcal{F}$ be an independent set in $\mathcal{G}$ of size at least $v(\mathcal{G})/d = |\mathcal{F}|/d$. For each ℓ-tuple of vertices $u = (u_0, \ldots, u_{\ell-1}) \in \tilde{C} := C_{y_0} \times \cdots \times C_{y_{\ell-1}}$, we denote by $\mathcal{F}^*(u)$ the set of all $\varphi \in \mathcal{F}^*$ such that $\varphi(y_i) = u_i$ for each $0 \leq i \leq \ell-1$. Note that

$$\sum_{u \in \tilde{C}} |\mathcal{F}^*(u)| = |\mathcal{F}^*| \geq \frac{|\mathcal{F}|}{d} \geq \frac{\varepsilon n^{k+\ell}}{v^v d} = \zeta n^{k+\ell}\,. \tag{7.24}$$

We claim that $|\mathcal{F}^*(u)| \leq n^k$ for each $u \in \tilde{C}$. To see this, fix any such u and let $\varphi_1, \varphi_2 \in \mathcal{F}^*(u)$ be distinct. If $\varphi_1(x_i) = \varphi_2(x_i)$ for each $1 \leq i \leq k$, then $\{x_1, \ldots, x_k, y_0, \ldots, y_{\ell-1}\} \subseteq U(\varphi_1, \varphi_2)$. But then U satisfies Condition (i) above, implying that $\{\varphi_1, \varphi_2\} \in E(\mathcal{G})$, in contradiction to the fact that $\mathcal{F}^*$ is an independent set in $\mathcal{G}$. So we see that for each $u \in \tilde{C}$ and for each $\varphi \in \mathcal{F}^*(u)$, the values of $\varphi(x_1), \ldots, \varphi(x_k)$ determine φ uniquely. It follows that indeed $|\mathcal{F}^*(u)| \leq n^k$. Now, by using (7.24) and averaging, we get that there are at least $\frac{\zeta}{2} n^\ell$ tuples $u \in \tilde{C}$ which satisfy $|\mathcal{F}^*(u)| \geq \frac{\zeta}{2} n^k$. Let $C \subseteq \tilde{C}$ be the set of all such tuples u. We will show that for every $u = (u_0, \ldots, u_{\ell-1}) \in C$, there are at least $\frac{1}{2}\gamma(k, \frac{\zeta}{2}) \cdot n^{k+1}$ copies of $G_{\ell+1}$ in H in which u_i plays the role of $y_i(G_{\ell+1})$ for every $0 \leq i \leq \ell-1$. Combining this with the fact that $|C| \geq \frac{\zeta}{2} n^\ell$, we will conclude that H contains at least $\frac{\zeta}{2} n^\ell \cdot \frac{1}{2}\gamma(k, \frac{\zeta}{2}) \cdot n^{k+1} = \delta n^{k+\ell+1}$ copies of $G_{\ell+1}$, as required.

Fix any $u \in C$. Let us define an auxiliary k-uniform $(k+1)$-partite hypergraph $J(u)$ with parts $C_{x_1}, \ldots, C_{x_k}, C_{y_\ell}$, as follows. For each $\varphi \in \mathcal{F}^*(u)$, put a k-uniform $(k+1)$-clique in $J(u)$ on the vertices $\varphi(x_1) \in C_{x_1}, \ldots, \varphi(x_k) \in C_{x_k}$, $\varphi(y_\ell) \in C_{y_\ell}$. We denote this clique by K_φ. We claim that the cliques $(K_\varphi : \varphi \in \mathcal{F}^*(u))$ are pairwise edge-disjoint. To this end, fix any pair of distinct $\varphi_1, \varphi_2 \in \mathcal{F}^*(u)$ and suppose, for the sake of contradiction, that the cliques $K_{\varphi_1}, K_{\varphi_2}$ share an edge. Then there is $Z \subseteq \{x_1, \ldots, x_k, y_\ell\}$ of size $|Z| = k$ such that $\varphi_1(z) = \varphi_2(z)$ for every $z \in Z$. It follows that $Z \cup \{y_0, \ldots, y_{\ell-1}\} \subseteq U(\varphi_1, \varphi_2)$. Therefore, $|U(\varphi_1, \varphi_2) \cap A_\ell| \geq k + \ell$, implying that $U(\varphi_1, \varphi_2)$ satisfies Condition (i) above. This in turn implies that $\{\varphi_1, \varphi_2\} \in E(\mathcal{G})$, which contradicts the fact that $\mathcal{F}^*(u) \subseteq \mathcal{F}(u)$ is an independent set in $\mathcal{G}$. We have thus shown that the cliques $(K_\varphi : \varphi \in \mathcal{F}^*(u))$ are indeed pairwise edge-disjoint.

It follows from the previous paragraph that $J(u)$ contains a collection of $|\mathcal{F}^*(u)| \geq \frac{\zeta}{2} n^k$ pairwise edge-disjoint $(k+1)$-cliques. By Theorem 7.1.3, the number of $(k+1)$-cliques in $J(u)$ is at least $\gamma(k, \frac{\zeta}{2}) \cdot n^{k+1}$. A $(k+1)$-clique K in $J(u)$ is called *colorful* if it is not equal to K_φ for any $\varphi \in \mathcal{F}^*(u)$. Since there are at most $|\mathcal{F}^*(u)| \leq n^k$ non-colorful $(k+1)$-cliques, the number of colorful $(k+1)$-cliques in $J(u)$ is at least $\gamma(k, \frac{\zeta}{2}) \cdot n^{k+1} - n^k \geq \frac{1}{2}\gamma(k, \frac{\zeta}{2}) \cdot n^{k+1}$ (here we use our choice of n_0).

To complete the proof, it remains to show that each colorful $(k+1)$-clique in $J(u)$ corresponds to a copy of $G_{\ell+1}$ in H. Fix any colorful $(k+1)$-clique $K = \{w_1, \ldots, w_k, u_\ell\}$, where u_ℓ is the unique vertex of K contained in C_{y_ℓ} and, for each $1 \leq i \leq k$, w_i is the unique vertex of K contained in C_{x_i}. By the definition of $J(u)$, each of the $k+1$ edges of K corresponds to an embedding of G_ℓ into H. More precisely, there are $\varphi_0, \varphi_1, \ldots, \varphi_k \in \mathcal{F}^*(u)$ such that:

- For each $1 \leq i \leq k$, $\varphi_i(y_\ell) = u_\ell$ and $\varphi_i(x_j) = w_j$ for each $j \in [k] \setminus \{i\}$.
- $\varphi_0(x_i) = w_i$ for each $1 \leq i \leq k$.

We claim that $\varphi_0, \ldots, \varphi_k$ are pairwise distinct. Assume, for the sake of contradiction, that $\varphi_i = \varphi_{i'} =: \varphi$ for some $0 \leq i < i' \leq k$. Then $\varphi(x_j) = w_j$ for each $1 \leq j \leq k$. Indeed, this follows from the two items above and from the (obvious) fact that one of i, i' does not equal j. Similarly, since i, i' cannot both equal 0, the first item above implies that $\varphi(y_\ell) = u_\ell$. We now see that $K = K_\varphi$, in contradiction to the assumption that K is colorful. Hence, $\varphi_0, \ldots, \varphi_k$ are indeed pairwise distinct. Now the edge-disjointness of the cliques $K_{\varphi_0}, K_{\varphi_1}, \ldots, K_{\varphi_k}$ implies that $w_i' := \varphi_i(x_i) \neq w_i$ for each $1 \leq i \leq k$ and that $u_{\ell+1} := \varphi_0(y_\ell) \neq u_\ell$.

We now show how to construct a copy of $G_{\ell+1}$ using the copies of G_ℓ corresponding to $\varphi_1, \ldots, \varphi_k$ and the copy of G corresponding to $\varphi_0(G^\ell)$. In this copy of $G_{\ell+1}$, the role of $x_i(G_{\ell+1})$ will be played by w_i for every $1 \leq i \leq k$, the role of the vertex $x_i' \in V(G_{\ell+1})$ will be played by w_i' for every $1 \leq i \leq k$ (recall the definition of $G_{\ell+1}$) and the role of $y_i(G_{\ell+1})$ will be played by u_i for every $0 \leq i \leq \ell+1$. (Recall that the vertices $u_0, \ldots, u_{\ell-1}$ have already been fixed via the choice of u.) Note that for each $1 \leq i \leq k$, the embedding φ_i corresponds to a copy of G_ℓ in which w_j plays the role of $x_j(G_\ell)$ for every $j \in [k] \setminus \{i\}$, w_i' plays the role of $x_i(G_\ell)$ and u_j plays the role of $y_j(G_\ell)$ for every $0 \leq j \leq \ell$. This copy of G_ℓ will play the role of G_ℓ^i in our copy of $G_{\ell+1}$. Similarly, restricting φ_0 to $V(G^\ell)$ gives a copy of G in which w_i plays the role of $x_i(G)$ for each $1 \leq i \leq k$ and $u_{\ell+1}$ plays the role of $y_0(G)$ (as $y_0(G^\ell) = y_\ell(G_\ell)$ and $\varphi_0(y_\ell(G_\ell)) = u_{\ell+1}$). By the definition of $G_{\ell+1}$, in order to show that $\mathrm{Im}(\varphi_1) \cup \cdots \cup \mathrm{Im}(\varphi_k) \cup \varphi_0(V(G^\ell))$ spans a copy of $G_{\ell+1}$, it suffices to verify that the k copies of G_ℓ given by $\varphi_1, \ldots, \varphi_k$, and the copy of G given by $\varphi_0(G^\ell)$, do not intersect outside of $\{w_1, \ldots, w_k, u_0, \ldots, u_\ell\}$. Therefore, our goal is to show that $\mathrm{Im}(\varphi_i) \cap \mathrm{Im}(\varphi_j) = \{w_1, \ldots, w_k, u_0, \ldots, u_\ell\} \setminus \{w_i, w_j\}$ for each $1 \leq i < j \leq k$ and that $\mathrm{Im}(\varphi_i) \cap \varphi_0(V(G^\ell)) = \{w_1, \ldots, w_k\} \setminus \{w_i\}$ for each $1 \leq i \leq k$. We start with the former

statement. Fix any $1 \leq i < j \leq k$. Setting $U := U(\varphi_i, \varphi_j)$, note that $\mathrm{Im}(\varphi_i) \cap \mathrm{Im}(\varphi_j) = \varphi_i(U) = \varphi_j(U)$, that $y_0, \ldots, y_\ell \in U$ and that $U \cap X = X \setminus \{x_i, x_j\}$ and hence $|U \cap X| = k-2$. If we had $U \setminus A_\ell \neq \emptyset$, then U would satisfy Condition (ii) above, which in turn would imply that $\{\varphi_i, \varphi_j\} \in E(\mathcal{G})$, thus contradicting the fact that $\mathcal{F}^*(u) \subseteq \mathcal{F}^*$ is an independent set in $\mathcal{G}$. So we see that $U \subseteq A_\ell$ and therefore $U = A_\ell \setminus \{x_i, x_j\}$. This in turn is equivalent to having $\mathrm{Im}(\varphi_i) \cap \mathrm{Im}(\varphi_j) = \{w_1, \ldots, w_k, u_0, \ldots, u_\ell\} \setminus \{w_i, w_j\}$, as required.

Let us now show that $\mathrm{Im}(\varphi_i) \cap \varphi_0(V(G^\ell)) = \{w_1, \ldots, w_k\} \setminus \{w_i\}$ holds for every $1 \leq i \leq k$. Fixing $1 \leq i \leq k$, set $U := U(\varphi_i, \varphi_0)$ and note that $A_\ell \setminus \{x_i, y_\ell\} = \{x_1, \ldots, x_k, y_0, \ldots, y_{\ell-1}\} \setminus \{x_i\} \subseteq U$. Now, if $U \cap V(G^\ell)$ were not contained in X, then U would satisfy Condition (iii) above, which would imply the false statement that $\{\varphi_i, \varphi_0\} \in E(\mathcal{G})$. So we see that $U \cap V(G^\ell) \subseteq X$. Moreover, $x_i \notin U$, because otherwise the $(k+1)$-cliques corresponding to φ_i and φ_0, respectively, would not be edge-disjoint (or, alternatively, because otherwise U would satisfy Condition (i) above). So we see that $U \cap V(G^\ell) = \{x_1, \ldots, x_k\} \setminus \{x_i\}$, which implies that $\mathrm{Im}(\varphi_i) \cap \varphi_0(V(G^\ell)) = \{w_1, \ldots, w_k\} \setminus \{w_i\}$. ■

Finally, we use Claim 7.3.5 in order to establish Item 3 of the lemma by induction on ℓ. The case $\ell = 0$ is trivial. Let us now fix $\ell \geq 0$, assume the validity of Item 3 for ℓ and prove the analogous statement for $\ell + 1$. It is easy to see that if the assertion of 3(a) holds for parameter ℓ, then it also holds for parameter $\ell + 1$. So we may assume that the assertion of Item 3(b) holds, namely, that H contains at least $\varepsilon' \cdot n^{k+\ell}$ copies of G_ℓ (where $\varepsilon' := \delta(\ell, r, \varepsilon)$, as given by Item 3 in the lemma). So we may apply Claim 7.3.5 to H (with parameter ε' in place of ε). If Item 1 of Claim 7.3.5 holds, then Item 3(a) of Lemma 7.1.8 holds with $\ell + 1$ in place of ℓ (and with $j = \ell$). If instead Item 2 of Claim 7.3.5 holds, then H contains at least $\delta \cdot n^{k+\ell+1}$ copies of $G_{\ell+1}$, as required by Item 3(b) in Lemma 7.1.8. This completes the proof of the lemma.

7.4 An Improved Bound for a Generalized Ramsey Problem of Erdős and Gyárfás

The Brown–Erdős–Sós problem has a known connection to (a special case of) the following generalized Ramsey problem, introduced by Erdős and Gyárfás in [47]. Let $g(n, p, q)$ denote the minimum number of colors in a coloring of the edges of K_n in which every copy of K_p receives at least q colors. For a fixed $p \geq 4$, Erdős and Gyárfás [47] showed that $g(n, p, q)$ is quadratic in n if and only if $q \geq q_{\mathrm{quad}}(p) := \binom{p}{2} - \lfloor \frac{p}{2} \rfloor + 2$ and that $\Omega(n^2) \leq g(n, p, q_{\mathrm{quad}}(p)) \leq \binom{n}{2} - \varepsilon n^2$ for some $\varepsilon = \varepsilon(p) > 0$. They then asked for which $q_{\mathrm{quad}}(p) < q \leq \binom{p}{2}$ it holds that $g(n, p, q) = \binom{n}{2} - o(n^2)$, observing that this question is related to the Brown–Erdős–Sós problem and using this relation to prove several partial results. The relation was further exploited by Sárközy and Selkow, who combined it with their result in [99] (or, more precisely, with a 4-uniform analogue thereof) to show that $g(n, p, q) = \binom{n}{2} - o(n^2)$ whenever $q > q_{\mathrm{quad}}(p) + \lceil \frac{\log_2 p}{2} \rceil$. By using our improved bound for the Brown–Erdős–Sós problem (i.e., Corollary 7.0.2), we can improve upon the result of Sárközy and Selkow [101]. For completeness, we now sketch the proof of the reduction from the above generalized Ramsey problem to the Brown–Erdős–Sós problem. This reduction has been used implicitly in [47, 101].

Proposition 7.4.1. *Let $p \geq 4$ and $q_{quad}(p) < q \leq \binom{p}{2}$. Set $e := \binom{p}{2} - q + 1$. If $f_4(n, p, e) = o(n^2)$, then $g(n, p, q) = \binom{n}{2} - o(n^2)$.*

Proof. Assume that $f_4(n, p, e) = o(n^2)$ and suppose, for the sake of contradiction, that (for infinitely

many n) there is a coloring of the edges of K_n with $t := \binom{n}{2} - \varepsilon n^2$ colors (where $\varepsilon > 0$ is fixed) in which every copy of K_p receives at least q colors. Then at least εn^2 edges have the same color as some other edge.

Observe that each color appears fewer than $\lfloor \frac{p}{2} \rfloor$ times. Indeed, otherwise take edges $e_1, \ldots, e_{\lfloor \frac{p}{2} \rfloor}$, all having the same color, and supplement them with (a suitable number of) vertices to obtain a copy of K_p which receives at most $\binom{p}{2} - \lfloor \frac{p}{2} \rfloor + 1 < q_{\text{quad}}(p) < q$, a contradiction. It follows that at least $\varepsilon n^2 / \lfloor p/2 \rfloor \geq 2\varepsilon n^2/p$ colors appear at least twice. For each such color c, fix a pair of distinct edges (e_1^c, e_2^c) which are colored with c. We claim that there are less than $(p-1)n/2$ colors c for which e_1^c and e_2^c intersect. Indeed, assign to each such intersecting pair of edges their common vertex. If the number of intersecting pairs is at least $(p-1)n/2$, then there is a vertex u which is the common vertex for at least $\lfloor \frac{p-1}{2} \rfloor$ such edge-pairs. In other words, there are distinct vertices $(x_i, y_i : 1 \leq i \leq \lfloor \frac{p-1}{2} \rfloor)$ such that the color of $\{u, x_i\}$ is the same as that of y_i for each $1 \leq i \leq \lfloor \frac{p-1}{2} \rfloor$. As before, by adding a suitable number of vertices one obtains a copy of K_p which receives at most $\binom{p}{2} - \lfloor \frac{p-1}{2} \rfloor < q_{\text{quad}}(p) < q$ colors, in contradiction to our assumption.

It follows from the above two paragraphs that there are at least $2\varepsilon n^2/p - (p-1)n/2 \geq \varepsilon n^2/p$ colors c (appearing at least twice) for which e_1^c, e_2^c are disjoint. Define an auxiliary 4-graph H on $V(K_n)$ by putting a (4-uniform) edge on $e_1^c \cup e_2^c$ for each color c for which e_1^c, e_2^c are disjoint. Since K_4 has 3 perfect matchings, we have $e(H) \geq \frac{\varepsilon n^2}{3p}$. Observe, crucially, that H contains no (p, e)-configuration. Indeed, if H contained a (p, e)-configuration, then, by the definition of H and our choice of e, the vertex set of this configuration would correspond to a copy of K_p receiving at most $\binom{p}{2} - e = q - 1$ colors, which is impossible. We thus conclude that $e(H) \leq f_4(n, p, e)$. On the other hand, $e(H) \geq \frac{\varepsilon n^2}{3p}$, implying that $f_4(n, p, e) = \Omega(n^2)$, in contradiction to our assumption. ■

By Corollary 7.0.2, applied with parameters $r = 4$, $k = 2$ and $e = \binom{p}{2} - q + 1$, the bound $f_4(n, p, e) = o(n^2)$ holds whenever $p \geq 2e + 18 \log e / \log \log e = 2(\binom{p}{2} - q + 1) + 18 \log e / \log \log e$. By rearranging, we get the inequality $q \geq \binom{p}{2} - \frac{p}{2} + 1 + 18 \log e / \log \log e$. Recalling the value of $q_{\text{quad}}(p)$ and using the (obvious) fact that $e \leq \binom{p}{2}$, we see that this inequality holds whenever $q \geq q_{\text{quad}}(p) + C \log p / \log \log p$ for some suitable absolute constant C. By combining this with Proposition 7.4.1, we obtain the following improvement upon the aforementioned result from [101].

Theorem 7.4.2. *There is an absolute constant C such that $g(n, p, q) = \binom{n}{2} - o(n^2)$ for every $p \geq 4$ and $q \geq q_{quad}(p) + C \log p / \log \log p$.*

Bibliography

[1] N. Alon, Ramsey graphs cannot be defined by real polynomials, Journal of Graph Theory 14 (1990), 651-661.

[2] N. Alon, Testing subgraphs in large graphs, Random Structures and Algorithms 21 (2002), 359-370.

[3] N. Alon, private communication, 2013.

[4] N. Alon, W.F. de la Vega, R. Kannan and M. Karpinski, Random sampling and approximation of MAX-CSP problems, JCSS 67 (2003), 212-243.

[5] N. Alon, E. Fischer, M. Krivelevich and M. Szegedy, Efficient testing of large graphs, Combinatorica 20 (2000), 451-476.

[6] N. Alon, E. Fischer, and I. Newman. Testing of bipartite graph properties. SIAM Journal on Computing 37 (2007), 959-976.

[7] N. Alon, E. Fischer, I. Newman and A. Shapira, A combinatorial characterization of the testable graph properties: it's all about regularity. SIAM Journal on Computing, 39(1) (2009), 143–167.

[8] N. Alon and J. Fox, Easily testable graph properties, Combin. Probab. Comput. 24 (2015), 646-657.

[9] N. Alon and A. Shapira, A characterization of easily testable induced subgraphs, Combin. Probab.Comput. 15 (2006), 791-805.

[10] N. Alon and A. Shapira, A characterization of the (natural) graph properties testable with one-sided error, SIAM Journal on Computing 37 (2008), 1703-1727.

[11] N. Alon and A. Shapira, A separation theorem in property testing, Combinatorica 28 (2008), 261-281.

[12] N. Alon and A. Shapira, Every monotone graph property is testable, SIAM Journal on Computing 38 (2008), 505-522.

[13] N. Alon and A. Shapira, Testing subgraphs in directed graphs, JCSS 69 (2004), 354-382.

[14] N. Alon and A. Shapira, On an extremal hypergraph problem of Brown, Erdős and Sós, Combinatorica 26 (2006), 627–645.

[15] N. Alon and C. Shikhelman, Many T copies in H-free graphs, Journal of Combinatorial Theory, Series B, 121 (2016), 146-172.

[16] N. Alon and J. H. Spencer, **The probabilistic method**, 3rd ed., Wiley, 2008.

[17] T. Austin and T. Tao, On the testability and repair of hereditary hypergraph properties, Random Structures and Algorithms 36 (2010), 373-463.

[18] L. Avigad and O. Goldreich, Testing graph blow-up. In Studies in Complexity and Cryptography, Miscellanea on the Interplay between Randomness and Computation (2011), pp. 156–172. Springer, Berlin, Heidelberg.

[19] R. C. Baker, G. Harman and J. Pintz, The difference between consecutive primes, II. Proceedings of the London Mathematical Society, 83 (2001), 532-562.

[20] 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.

[21] M. Blum, M. Luby and R. Rubinfeld, Self-testing/correcting with applications to numerical problems, Journal of Computer and System Sciences 47 (1993), 549–595.

[22] D. Bokal, G. Fijavž, M. Juvan, P. M. Kayll and B. Mohar, The circular chromatic number of a digraph, J. Graph Theory 46 (2004), 227-240.

[23] B. Bollobás, **Extremal graph theory**, Courier Corporation, 2004.

[24] B. Bollobás and E. Győri, Pentagons vs. triangles, Discrete Math. 308 (2008), 4332-4336.

[25] J. A. Bondy and M. Simonovits, Cycles of even length in graphs, Journal of Combinatorial Theory, Series B, 16 (1974), 97-105.

[26] W. G. Brown, P. Erdős and V. T. Sós, Some extremal problems on r-graphs, in: New Directions in the Theory of Graphs, Proc. 3rd Ann Arbor Conference on Graph Theory, Academic Press, New York, 1973, 55–63.

[27] W. G. Brown, P. Erdős and V. T. Sós, On the existence of triangulated spheres in 3-graphs and related problems, Period. Math. Hungar. 3 (1973), 221–228.

[28] B. Bukh and Z. Jiang, A bound on the number of edges in graphs without an even cycle, 2017, Combinatorics, Probability and Computing, 26(1), pp.1-15.

[29] X. Chen, Z. Liu, R.A. Servedio, Y. Sheng and J. Xie, Distribution-free junta testing. In Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing (2018), 749–759.

[30] X. Chen and J. Xie, Tight bounds for the distribution-free testing of monotone conjunctions. In Proceedings of the Twenty-Seventh Annual ACM-SIAM Symposium on Discrete Algorithms, SODA (2016), 54–71.

[31] A. Chernikov and S. Starchenko, Regularity lemma for distal structures. J. Eur. Math. Soc. 20 (2018), 2437-2466.

[32] A. Chernikov and S. Starchenko, Definable regularity lemmas for NIP hypergraphs. arXiv preprint arXiv:1607.07701, 2016.

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

[34] D. Conlon and J. Fox, Bounds for graph regularity and removal lemmas, GAFA **22** (2012), 1191-1256.

[35] D. Conlon and J. Fox, Graph removal lemmas, in: Surveys in combinatorics 2013, 1–49, London Math. Soc. Lecture Note Ser., 409, Cambridge Univ. Press, Cambridge, 2013.

[36] D. Conlon, J. Fox and B. Sudakov, Recent developments in graph Ramsey theory, Surveys in combinatorics 2015, 49–118, London Math. Soc. Lecture Note Ser., 424, Cambridge Univ. Press, Cambridge, 2015.

[37] B. Csaba and A. Pluhár, A weighted regularity lemma with applications. International Journal of Combinatorics, 2014.

[38] E. Dolev and D. Ron, Distribution-free testing for monomials with a sublinear number of queries. Theory of Computing, 7(1) (2011), 155–176.

[39] P. Erdős, On extremal problems of graphs and generalized graphs, Israel J. Math. 2, 1964, 183-190.

[40] P. Erdős, On circuits and subgraphs of chromatic graphs, Mathematika 9 (1962), 170-175.

[41] P. Erdős, On the number of complete subgraphs contained in certain graphs, Magyar Tud. Akad.. Mat. Kut. Int. Kőzl, 7 (1962), 459-474.

[42] P. Erdős, On some problems in graph theory, combinatorial analysis and combinatorial number theory. In *Graph theory and combinatorics (Cambridge, 1983)*, pages 1-17. Academic Press, London, 1984.

[43] P. Erdős, Problems and results on graphs and hypergraphs: similarities and differences, Mathematics of Ramsey theory (J.. Nesetril and V. Rödl, eds.), Algorithms Combin., vol. 5, Springer, Berlin, 1990, 12-28.

[44] P. Erdős, Problems and results in combinatorial number theory, Journées arithmétiques de Bordeaux, Astérisque, 24–25 (1975), 295–310.

[45] 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.

[46] P. Erdős and T. Gallai, On maximal paths and circuits of graphs, Acta Math. Acad. Sci. Hung. 10 (1959) 337-356.

[47] P. Erdős and A. Gyárfás, A variant of the classical Ramsey problem, Combinatorica 17 (1997), 459–467.

[48] P. Erdős and A. Rényi, On a problem in the theory of graphs (in Hungarian), Publ. Math. Inst. Hungar. Acad. Sci. 7 (1962), 215-235.

[49] J. Fox, A new proof of the graph removal lemma, Ann. of Math. **174** (2011), 561–579.

[50] J. Fox, M. Gromov, V. Lafforgue, A. Naor, and J. Pach, Overlap properties of geometric expanders. Journal für die reine und angewandte Mathematik (Crelles Journal), 2012(671), 49-83, 2012.

[51] J. Fox, J. Pach and A. Suk, A polynomial regularity lemma for semi-algebraic hypergraphs and its applications in geometry and property testing. arXiv preprint arXiv:1502.01730, 2015.

[52] P. Frankl and V. Rödl, Extremal problems on set systems, Random Structures Algorithms, 20(2), 131-164, 2002.

[53] Z. Füredi and L. Özkahya, On 3-uniform hypergraphs without a cycle of a given length. Discrete Applied Mathematics, 216, pp.582–588, 2017.

[54] D. Gerbner, E. Győri, A. Methuku and M. Vizer, Generalized Turán problems for even cycles. Acta Mathematica Universitatis Comenianae, 88(3), 723-728, 2019.

[55] D. Glasner and R. A. Servedio. Distribution-free testing lower bound for basic boolean functions. Theory of Computing 5 (2009), 191–216.

[56] O. Goldreich, Contemplations on testing graph properties, ECCC Technical Report 2005. Also, Studies in Complexity and Cryptography 6650 (2011), 547-554.

[57] O. Goldreich, **Introduction to Property Testing**. Cambridge University Press, 2017.

[58] O. Goldreich, Testing Graphs in Vertex-Distribution-Free Models. Preprint available online.

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

[60] O. Goldreich and D. Ron, On proximity-oblivious testing. SIAM Journal on Computing 40 (2011), 534–566.

[61] O. Goldreich and D. Ron, Property testing in bounded degree graphs. Algorithmica 32 (2002), 302–343.

[62] O. Goldreich and I. Shinkar, Two-sided error proximity oblivious testing, Random Structures Algorithms 48 (2016), 341–383.

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

[64] M. C. Golumbic, **Algorithmic graph theory and perfect graphs** (Vol. 57), 2004, Elsevier.

[65] M. C. Golumbic, Trivially perfect graphs, Discrete Mathematics 24 (1978), 105-107.

[66] T. Gowers, Lower bounds of tower type for Szemerédi's uniformity lemma, Geom. Funct. Anal. **7** (1997), 322–337.

[67] W. T. Gowers, Hypergraph regularity and the multidimensional Szemerédi theorem, Ann. of Math. (2007), 897–946.

[68] A. Grzesik, On the maximum number of five-cycles in a triangle-free graph, J. Combin. Theory Ser. B 102 (2012), 1061-1066.

[69] A. Gyárfás, A. Hubenko and J. Solymosi, Large cliques in C_4-free graphs, Combinatorica, 22 (2002), 269-274.

[70] E. Győri and H. Li, The maximum number of triangles in C_{2k+1}-free graphs, Combinatorics, Probability and Computing 21 (2012), 187-191.

[71] S. Halevy and E. Kushilevitz, Distribution-free property testing. In Approximation, Randomization, and Combinatorial Optimization: Algorithms and Techniques (2003) (pp. 302-317). Springer, Berlin, Heidelberg.

[72] F. Harary, **Graph Theory**, Massachusetts: Addison-Wesley (1972), 71-83.

[73] H. Hatami, J. Hladký, D. Král', S. Norine and A. Razborov, On the number of pentagons in triangle-free graphs, J. Combin. Theory Ser. A 120 (2013), 722-732.

[74] P. Hell and J. Nesetril, **Graphs and homomorphisms**. Oxford University Press, 2004.

[75] R. de Joannis de Verclos, Chordal graphs are easily testable, arXiv preprint: arXiv:1902.06135, 2019.

[76] S. Kalyanasundaram and A. Shapira, A Wowzer-type lower bound for the strong regularity-lemma, Proc. of the London Math Soc 106 (2013), 621–649.

[77] J. Komlós, Covering odd cycles, Combinatorica 17 (1997), 393-400.

[78] T. Kovári, V. Sós and P. Turán, On a problem of K. Zarankiewicz. In Colloquium Mathematicum, 1954, 50–57.

[79] L. Lovász, **Large networks and graph limits**, 2012, Providence: American Mathematical Society.

[80] L. Lov´asz and B. Szegedy, Regularity partitions and the topology of graphons, in An irregular mind, 415-445, Bolyai Soc. Math. Stud. 21, János Bolyai Math. Soc., Budapest, 2010.

[81] L. Lovász and B. Szegedy, Testing properties of graphs and functions, Israel J. Math. 178 (2010), 113-156.

[82] M. Malliaris and S. Shelah, Regularity lemmas for stable graphs. Trans. Amer. Math. Soc. 366, 1551-1585, 2014.

[83] C. McDiarmid, A random bit-flipping method for seeking agreement, Random Structures and Algorithms, 1996, 121-130.

[84] J. W. Moon, **Topics on tournaments**, Holt, Rinehart and Winston, New York, 1968.

[85] G. Moshkovitz and A. Shapira, A sparse regular approximation lemma, Transactions of the AMS 371, 6779-6814, 2019.

[86] B. Nagle, V. Rödl and M. Schacht, The counting lemma for regular k-uniform hypergraphs, Random Structures Algorithms 28 (2006), 113–179.

[87] A. Naor, J. Verstraëte, A note on bipartite graphs without $2k$-cycles, Combin. Probab. Comput. 14 (2005), no. 5-6, 845-849.

[88] O. Pikhurko, A note on the Turán function of even cycles, 2012, Proceedings of the American Mathematical Society, 140(11), pp.3687-3692.

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

[90] V. Rödl and J. Skokan, Applications of the regularity lemma for uniform hypergraphs, Random Structures Algorithms 28 (2006), 180–194.

[91] V. Rödl and R. Duke, On graphs with small subgraphs of large chromatic number, Graphs and Combinatorics 1 (1985), 91-96.

[92] V. Rödl, B. Nagle, J. Skokan, M. Schacht and Y. Kohayakawa, The hypergraph regularity method and its applications. Proceedings of the National Academy of Sciences, 102(23), 8109-8113, 2005.

[93] V. Rödl and M. Schacht, Generalizations of the removal lemma, Combinatorica 29, 2009, 467-501.

[94] V. Rödl and M. Schacht, Regularity lemmas for graphs, Fete of Combinatorics and Computer Science, Bolyai Soc. Math. Stud., 20 (2010), 287–325.

[95] K. F. Roth, On certain sets of integers II, J. Lond. Math. Soc. 29 (1954), 20–26.

[96] R. Rubinfeld and M. Sudan, Robust characterizations of polynomials with applications to program testing, SIAM Journal on Computing 25 (1996), 252-271.

[97] I. Z. Ruzsa, Solving a linear equation in a set of integers I, Acta Arithmetica 65 (1993), 259-282.

[98] I. Z. Ruzsa and E. Szemerédi, Triple systems with no six points carrying three triangles, in Combinatorics (Proc. Fifth Hungarian Colloq., Keszthely, 1976), Vol. II, pp. 939–945, Colloq. Math. Soc. János Bolyai, 18, North-Holland, Amsterdam-New York, 1978.

[99] G. N. Sárközy and S. Selkow, An extension of the Ruzsa-Szemerédi theorem, Combinatorica 25 (2004), 77–84.

[100] G. N. Sárközy and S. Selkow, On a Turán-type hypergraph problem of Brown, Erdős and T. Sós, Discrete mathematics 297 (2005), 190–195.

[101] G. N. Sárközy and S. Selkow, An application of the regularity lemma in generalized Ramsey theory, J. Graph Theory 44 (2003), 39–49.

[102] A. Shapira and M. Tyomkyn, A Ramsey variant of the Brown–Erdős–Sós conjecture, preprint available at arXiv:1910.13546, 2019.

[103] P. Simon, A note on "Regularity lemma for distal structures", Proceedings of the American Mathematical Society, 144(8), 3573-3578, 2016.

[104] M. Simonovits, Extremal graph problems with symmetrical extremal graphs, additional chromatic conditions, Discrete Math. 7 (1974), 349-376.

[105] M. Simonovits, Paul Erdős' influence on extremal graph theory, in The mathematics of Paul Erdős, II, 148-192, Algorithms Combin., 14, Springer, Berlin, 1997.

[106] C. Sohler, Almost optimal canonical property testers for satisfiability, In Foundations of Computer Science (FOCS), 2012 IEEE 53rd Annual Symposium (2012), pp. 541-550, IEEE.

[107] D. Solymosi and J. Solymosi, Small cores in 3-uniform hypergraphs, J. Combin. Theory Ser. B 122 (2017), 897–910.

[108] E. Szemerédi, On sets of integers containing no four elements in arithmetic progression, Acta Math. Acad. Sci. Hungar. 20 (1969), 89–104.

[109] E. Szemerédi, On sets of integers containing no k elements in arithmetic progression, Acta Arith. 27 (1975), 199–245.

[110] E. Szemerédi, Regular partitions of graphs. In: Problèmes combinatoires et théorie des graphes (Colloq. Internat. CNRS, Univ. Orsay, Orsay, 1976), pp. 399–401, Colloq. Internat. CNRS, 260, CNRS, Paris, 1978.

[111] T. Tao, A variant of the hypergraph removal lemma, J. Combin. Theory Ser. A 113, 1257–1280, 2006.

[112] T. Tao, Szemerédi's regularity lemma revisited. Contributions to Discrete Mathematics, 1(1), 2006.

[113] P. Turán, On an extremal problem in graph theory (in Hungarian), Mat. Fiz. Lapok 48 (1941), 436-452.

[114] J. Verstraëte, Extremal problems for cycles in graphs, In Recent Trends in Combinatorics, pp. 83-116, Springer International Publishing, 2016.