Efficient Testing without Efficient Regularity

Lior Gishboliner[1] **and Asaf Shapira**[*2]

1 **School of Mathematical Sciences, Tel Aviv University, Tel Aviv, 69978, Israel**
liorgis1@post.tau.ac.il.
2 **School of Mathematical Sciences, Tel Aviv University, Tel Aviv 69978, Israel**
asafico@tau.ac.il.

Abstract

The regularity lemma of Szemerédi turned out to be the most powerful tool for studying the testability of graph properties in the dense graph model. In fact, as we argue in this paper, this lemma can be used in order to prove (essentially) all the previous results in this area. More precisely, a barrier for obtaining an efficient testing algorithm for a graph property $\mathcal{P}$ was having an efficient regularity lemma for graphs satisfying $\mathcal{P}$. The problem is that for many natural graph properties (e.g. triangle freeness) it is known that a graph can satisfy $\mathcal{P}$ and still only have regular partitions of tower-type size. This means that there was no viable path for obtaining reasonable bounds on the query complexity of testing such properties.

In this paper we consider the property of being induced C_4-free, which also suffers from the fact that a graph might satisfy this property but still have only regular partitions of tower-type size. By developing a new approach for this problem we manage to overcome this barrier and thus obtain a merely exponential bound for testing this property. This is the first substantial progress on a problem raised by Alon in 2001, and more recently by Alon, Conlon and Fox. We thus obtain the first example of an efficient testing algorithm that cannot be derived from an efficient version of the regularity lemma.

1998 ACM Subject Classification G.2.2 Graph Theory

Keywords and phrases Property testing, induced C_4-freeness

Digital Object Identifier 10.4230/LIPIcs.ITCS.2018.54

1 Introduction

The area of *property testing* was introduced in the seminal papers of Rubinfeld and Sudan [23] and Goldreich, Goldwasser and Ron [16]. As opposed to classical decision problems, where one is asked to decide if an input satisfies a predetermined property $\mathcal{P}$ or not, in property testing one is only asked to decide if the input satisfies $\mathcal{P}$ or is *far* from satisfying it. By now, problems of this type have been studied in so many areas that it will be impossible to survey them here. We thus refer the reader to the upcoming book of Goldreich [15] for more background and references on the subject.

Our focus in this paper will be testing graph properties in the *dense graph model*, introduced in the aforementioned [16], which was the first model in which property testing problems have been systematically studied. In this model, the input graph G is given via its $n \times n$ adjacency matrix, and we assume that there is an oracle that can answer queries of the form: is (i, j) an edge of G? We say that an n-vertex graph G is ε-far from satisfying property $\mathcal{P}$ if one should add/remove at least εn^2 edges in order to turn G into a graph

* Supported in part by ISF Grant 1028/16 and ERC Starting Grant 633509.

satisfying $\mathcal{P}$. An ε-tester for $\mathcal{P}$ is an algorithm that can distinguish with high probability (say, 2/3) between the case that G satisfies $\mathcal{P}$ and the case that G is ε-far from satisfying it. The maximum number of queries made by an ε-tester on n-vertex graphs is called its query complexity, and is denoted by $q(\varepsilon, n)$. We say that $\mathcal{P}$ is *testable* if it has an ε-tester which makes only $q(\varepsilon)$ queries, that is, whose query complexity depends only on ε and not on the size of the input. We say that $\mathcal{P}$ is easily testable if $q(\varepsilon) = \text{poly}(1/\varepsilon)$.

In Subsection 1.1 we (tersely) describe the main results of this paper. We elaborate on the relevant background, motivation, implications and significance of our results in Subsection 1.2.

1.1 The short story

A graph is induced H-free if it does not contain an induced copy of H. Alon, Fischer, Krivelevich and Szegedy [2] proved that for every fixed graph H, the property of being induced H-free is testable. Equivalently[1], this can be stated as saying that if an n-vertex graph G is ε-far from being induced H-free then G contains at least $n^h/q_H(\varepsilon)$ induced copies of H, where $h = |V(H)|$ and $q_H(\varepsilon)$ depends only on ε. The proof in [2] relied on the regularity lemma, and thus supplied very poor tower-type[2] bounds for $q_H(\varepsilon)$.

Alon [1] asked for which graphs H we have $q_H(\varepsilon) = \text{poly}(1/\varepsilon)$, that is, for which graphs H the property of being induced H-free is easily testable. This question was addressed by Alon and the second author [6] who resolved this problem for all graphs H save for P_3 (the path on 4 vertices) and C_4 (the 4-cycle). The former case was recently solved by Alon and Fox [5], who proved that $q_{P_3}(\varepsilon) = \text{poly}(1/\varepsilon)$. They further asked to determine if $q_{C_4}(\varepsilon) = \text{poly}(1/\varepsilon)$. This problem was also later raised by Conlon and Fox [10].

Prior to this work the best bound for $q_{C_4}(\varepsilon)$ was the same tower-type bound that holds for all graphs H. Our main result in this paper makes the first substantial progress on this problem.

▶ **Theorem 1. [Main Result]** *If an n-vertex graph G is ε-far from being induced C_4-free then G contains at least $n^4/2^{(1/\varepsilon)^c}$ induced copies of C_4, where c is an absolute constant. In particular, induced C_4-freeness is testable with query complexity $2^{(1/\varepsilon)^c}$.*

We strongly believe that the exponential bound in Theorem 1 can be further improved to a polynomial one, which would thus show that induced C_4-freeness is easily testable.

Given a (possibly infinite) family of graphs $\mathcal{F}$, we say that a graph is induced $\mathcal{F}$-free if it is induced H-free for every $H \in \mathcal{F}$. The result of [2] was extended by Alon and the second author [7] who showed that for every family of graphs $\mathcal{F}$, the property of being induced $\mathcal{F}$-free is testable. Needless to say that as in [2], the bounds involved were also of tower-type. It is natural to ask if Theorem 1 can be extended to properties defined by forbidding a family of graphs $\mathcal{F}$, one of which is C_4. The most notable and natural example is the property of being *chordal*, which is the property of not containing an induced cycle of length at least 4. Previously, the best bound for testing this property was the tower-type bound which follows from the general result of [7]. Here we obtain the following improved bound.

---

[1] This statement is usually referred to as a *removal lemma*, after the triangle removal lemma of Ruzsa and Szemerédi [24] from 1976. So in some sense, the result of [24] was the first statement in graph property testing. Also, see [1] for the short argument showing the equivalence between these two formulations.

[2] tower(x) is a tower of exponents of height x, so tower(3) $= 2^{2^2}$. In fact, the proof in [2] gave wowzer-type bounds which were later improved in [9].

▶ **Theorem 2.** *If an n-vertex graph G is ε-far from being chordal then for some $4 \leq \ell \leq O(\varepsilon^{-18})$, G contains at least $n^\ell/2^{(1/\varepsilon)^c}$ induced copies of C_ℓ, where c is an absolute constant. In particular, chordality is testable with query complexity $2^{(1/\varepsilon)^c}$.*

It is now natural to ask if Theorem 2 can be further extended to an arbitrary family $\mathcal{F}$, one of which is C_4. As our final theorem shows, this is not the case in a very strong sense.

▶ **Theorem 3.** *For every (monotone increasing) function $g : (0, 1/2) \to \mathbb{N}$ there is a family of graphs $\mathcal{F} = \mathcal{F}(g)$ so that $C_4 \in \mathcal{F}$ and the following holds. 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.*

1.2 The long story

In this subsection we would like to describe the main significance of the results stated in the previous subsection. The famous regularity lemma of Szemerédi [25] is one of the most powerful tools for tackling problems in extremal graph theory. Roughly speaking, the lemma supplies a short description of a graph via a highly structured *regular partition* of its vertices. Given the nature of the problems studied in the area of property testing, it is no surprise that this lemma has also turned out to be a powerful tool in this area. In fact, it was shown in [4] that a property can be tested if and only if it is (more or less) equivalent to the property of having certain regular partitions. In other words, the regularity lemma gives a *qualitative* explanation as to which properties are testable.

Prior to this work, the relation between the regularity lemma and graph property testing was not only qualitative but also *quantitative*. In other words, the bounds one could obtain for the regularity lemma in graphs satisfying $\mathcal{P}$ determined the bounds one could obtain for testing $\mathcal{P}$ (with one important exception discussed below). Thanks to the work of Gowers [18], we know that in the *worst case*, a graph can have only regular partitions of tower-type size. However, when designing a property testing algorithm for a property $\mathcal{P}$, one can try to prove that graphs satisfying $\mathcal{P}$ must possess much smaller regular partitions. And indeed, as we have recently shown [14], almost all the known results giving non-tower-type bounds for testing graph properties $\mathcal{P}$ in the dense model, stem from the fact that graphs satisfying $\mathcal{P}$ have small regular partitions. For example, the result of [5] showing that induced P_3-freeness is easily testable can be derived from the fact that an induced P_3-free graph has a regular partition of polynomial size. Same goes for the polynomial testability results of [1, 6, 8, 17].

We now describe the only exception to the above quantitative relation between regularity and testing. In 1984 Erdős [12] (implicitly) conjectured that k-colorability is testable. This was verified by Rödl and Duke [22] who used the regularity lemma in order to show that k-colorability is testable with a tower-type bound. This tower-type bound was dramatically improved by Goldreich, Goldwasser and Ron [16], who showed that various *partition properties*, such as Max-Cut and k-colorability are easily testable, while *not* relying on the regularity lemma. Let us try to explain the reason for this exception: first, as opposed to triangle-freeness or induced C_4-freeness which are *local* properties, the partition properties of [16] are *global*. Perhaps the best way to see this is from the perspective of graph homomorphisms: triangle-freeness means that there is no edge preserving mapping from the vertices of the triangle to the vertices of G, while 3-colorability means that there is such a mapping from the vertices of G to the vertices of the triangle[3]. The second difference, which is more important

[3] In the language of graph limits, this is the distinction between left and right homomorphisms, see [20].

for our quantitative investigation here, is that at their core, these partition properties are "edge density" properties. This can explain (at least in hindsight), why one does not need any structure theorem in order to handle these problems, and can instead rely on sampling arguments that boil down to estimating various edge densities (this is not to say that devising such proofs is an easy task!).

Given the above discussion, one can ask why then one cannot get better bounds for testing induced H-freeness for every H. It is not hard to see that there are bipartite versions of Gowers's [18] example. Therefore, even for simple properties $\mathcal{P}$ such as triangle-freeness, a graph can satisfy $\mathcal{P}$ but still only have regular partitions of tower-type size. This means that any algorithm for testing triangle-freeness that relies on the regularity lemma is bound to produce tower-type bounds. In a major breakthrough, Fox [13] managed to prove the testability of triangle-freeness while avoiding Szemerédi's version of the regularity lemma, obtaining bounds that are still of tower-type, but only of height $O(\log 1/\varepsilon)$ instead of $\text{poly}(1/\varepsilon)$. A different formulation of his proof was later given in [10] and [21]. The latter proof shows that Fox's result can be derived from a variant of the regularity lemma. Unfortunately, it was shown in [21] that this variant of the regularity lemma must also produce partitions of tower-type size. Recapping, there is currently no viable approach for getting non-tower-type bounds, even for testing triangle-freeness.

With regards to induced C_4-freeness, it is not hard to check that every split graph is induced C_4-free, where a split graph is a graph whose vertex set can be partitioned into two sets, one spanning an independent set and the other spanning a complete graph. This means that if we take a bipartite version of Gowers' lower bound [18], and put a complete graph on one of the vertex sets, we get an induced C_4-free graph that has only regular partitions of tower-type size. In particular, arguments similar to those that were previously used in order to devise efficient testing algorithms cannot give better-than-tower-type bounds for this problem.

Summarizing the above discussion, Theorem 1 is the first example showing that one *can* obtain an efficient testing algorithm for a property $\mathcal{P}$ (or equivalently, an efficient removal lemma for $\mathcal{P}$) even though graphs satisfying $\mathcal{P}$ might have only regular partitions of tower-type size. In particular, Theorem 1 exhibits the strongest separation between bounds for testing a property $\mathcal{P}$ and bounds for the regularity lemma on graphs satisfying $\mathcal{P}$. We are hopeful that bounds similar to those obtained in Theorem 1, can be obtained for other properties for which the best known bounds are of tower-type, most notably triangle freeness.

1.3 Paper overview

The main idea of the proof is to show that (very roughly speaking) every induced C_4-free graph is a split graph. To be more precise, every induced C_4-free graph is close to being a union of an independent set and few cliques, so that the bipartite graphs between these cliques are highly structured. Note that we have no guarantee on the structure of the bipartite graph connecting the independent set and the cliques[4]. Towards this goal, in Section 2 we describe some preliminary lemmas, mostly regarding the structure of bipartite graphs that do not contain an induced matching of size 2. In Section 3 we give the main partial structure theorem, stated as Lemma 13. In the course of the proof we will make a surprising application of one of the main results of Goldreich, Goldwasser and Ron [16]. In Section 4 we

[4] This unstructured part is unavoidable due to the example we mentioned earlier of putting Gowers' construction between a clique and an independent set

give the proof of Theorems 1 and 2. We will make use of the structure theorem from Section 3 but will also have to deal with the (unavoidable) *unstructured* part of the graph. This will be done in Lemma 15. Finally, in Section 5, we give the proof of Theorem 3. Throughout the paper, we make no effort to optimize the constant c appearing in Theorems 1 and 2.

2 Forbidding an induced 2-matching

Our goal in this section is to introduce several definitions and prove Lemma 7 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 as above. 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 4.** *(X,Y) is induced M_2-free 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 $y \in N_Y(x) \setminus N_Y(x')$ 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 $N_Y(x') \subseteq N_Y(x)$. Consider the poset on X in which x precedes x' if and only if $N_Y(x) \subseteq N_Y(x')$. This poset is a total order by the above. Enumerate the elements of X from minimal to maximal to get the required enumeration. ◀

We say that (X,Y) is *homogeneous* if the bipartite graph between X and Y is either complete or empty. We say that a partition $\mathcal{P} = \{P_1, \ldots, P_r\}$ of a set V is an *equipartition* if $||P_i| - |P_j|| \leq 1$ for every $1 \leq i, j \leq r$.

▶ **Lemma 5.** *If (X,Y) is induced M_2-free then for every integer $r \geq 1$ there is an equipartition $X = X_1 \cup \cdots \cup X_r$ and a partition $Y = Y_1 \cup \cdots \cup Y_{r+1}$ such that (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 4. For $1 \leq i \leq r$ define $X_i = \{x_j : \frac{(i-1)m}{r} < j \leq \frac{im}{r}\}$. Here we assume, for simplicity of presentation, that $|X|$ is divisible by r; if that is not the case then we partition X into "consecutive intervals" of sizes $\lfloor \frac{|X|}{r} \rfloor$ and $\lceil \frac{|X|}{r} \rceil$. Let now $y_1, ..., y_n$ 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. It is easy to see that a refinement of a δ-homogeneous partition is itself δ-homogeneous.

▶ **Lemma 6.** *Let $\delta > 0$, 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 $k\,(2/\delta)^k$ parts.*

Proof. For every $1 \leq i < j \leq k$, we apply Lemma 5 to (X_i, X_j) with parameter $r = \frac{1}{\delta}$ to get partitions $\mathcal{P}_{i,j}$ of X_i and $\mathcal{P}_{j,i}$ of X_j, $\mathcal{P}_{i,j} = \{X_{i,j}^1, ..., X_{i,j}^r\}$, $\mathcal{P}_{j,i} = \{X_{j,i}^1, ..., X_{j,i}^{r+1}\}$, such that $\mathcal{P}_{i,j}$ is an equipartition and $(X_{i,j}^p, X_{j,i}^q)$ is homogeneous for every $p \neq q$. Note that

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

For every $i = 1, ..., k$, define $\mathcal{P}_i$ to be the common refinement of the partitions $(\mathcal{P}_{i,j})_{1 \leq j \leq k,\, j \neq i}$. We have $|\mathcal{P}_i| \leq (r+1)^{k-1} \leq (2/\delta)^k$. The partition $\mathcal{P} := \bigcup_{i=1}^{k} \mathcal{P}_i$ refines $\{X_1, \ldots, X_k\}$ and has at most $k\,(2/\delta)^k$ parts. For every $U, V \in \mathcal{P}$, if (U, V) is not homogeneous, then there are $1 \leq i < j \leq k$ and $1 \leq p \leq r$ 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 $(\mathcal{P}_{i,j})_{1 \leq i \neq j \leq k}$. By (1), we have

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

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

▶ **Lemma 7.** *Let $\delta > 0$, 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 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), $1 \leq i < j \leq q$, is at most δn^2.*
2. *$|W_i| \geq (\delta/2k)^{10k^2} n$ for every $1 \leq i \leq q$ and (W_i, W_j) is homogeneous for every pair $1 \leq i < j \leq q$.*

Proof. Apply Lemma 6 to G with parameter δ to obtain a δ-homogeneous partition $\mathcal{P}$ which refines $\{X_1, \ldots, X_k\}$. Define $\mathcal{Q} = \{U \in \mathcal{P} : |U| \geq \frac{\delta n}{|\mathcal{P}|}\}$ 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 $|Z| < |\mathcal{P}| \cdot \frac{\delta n}{|\mathcal{P}|} = \delta n$. Apply Lemma 6 to G again (with respect to the same partition $\{X_1, \ldots, X_k\}$), now with parameter $\delta' := \frac{\delta^2}{8|\mathcal{P}|^4}$, to get a δ'-homogeneous partition $\mathcal{V}$ with at most $k(16|\mathcal{P}|^4/\delta^2)^k$ parts. Let $\mathcal{W}$ be the common refinement of $\mathcal{P}$ and $\mathcal{V}$ and note that $\mathcal{W}$ is δ'-homogeneous since it is a refinement of $\mathcal{V}$. Moreover,

$$|\mathcal{W}| \leq |\mathcal{P}| \cdot |\mathcal{V}| \leq |\mathcal{P}| \cdot k(16|\mathcal{P}|^4/\delta^2)^k. \tag{2}$$

For each $1 \leq i \leq q$, 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 $W_1, ..., W_q$ satisfy the statement in Item 2. For $1 \leq i \leq q$, the probability that

$|W_i| < \frac{|Q_i|}{2q|\mathcal{W}|}$ is smaller than $\frac{|\mathcal{W}|\cdot\frac{|Q_i|}{2q|\mathcal{W}|}}{|Q_i|} = \frac{1}{2q}$. By the union bound, with probability larger than $\frac{1}{2}$, every $1 \leq i \leq q$ satisfies

$$|W_i| \geq \frac{|Q_i|}{2q|\mathcal{W}|} \geq \frac{\left(\frac{\delta^2}{16|\mathcal{P}|^4}\right)^k \delta n}{2k|\mathcal{P}|^3} \geq \frac{\delta^{3k}n}{k(2|\mathcal{P}|)^{7k}} \geq \frac{\delta^{3k}n}{k2^k(2/\delta)^{7k^2}} \geq \left(\frac{\delta}{2k}\right)^{10k^2} n \ ,$$

where in the second inequality we used $|Q_i| \geq \frac{\delta n}{|\mathcal{P}|}$, $q \leq |\mathcal{P}|$ and (2), and in the fourth inequality we used the bound on $|\mathcal{P}|$ given by Lemma 6. For $1 \leq i < j \leq q$, the probability that the pair (W_i, W_j) is not homogeneous is

$$\sum \frac{|W||W'|}{|Q_i||Q_j|} \leq \frac{4|\mathcal{P}|^2}{\delta^2 n^2} \sum |W||W'| \leq \frac{4|\mathcal{P}|^2}{\delta^2 n^2} \cdot \delta' n^2 \leq \frac{1}{2|\mathcal{P}|^2} \ ,$$

where the sums are taken over all non-homogeneous pairs $(W, W') \in \mathcal{W}_i \times \mathcal{W}_j$, the first inequality uses $|\mathcal{Q}_i|, |\mathcal{Q}_j| \geq \frac{\delta n}{2|\mathcal{P}|}$ and the second the fact that $\mathcal{W}$ is δ'-homogeneous. By the union bound, with probability at least $1 - \binom{q}{2}\frac{1}{|\mathcal{P}|} \geq 1 - \binom{|\mathcal{P}|}{2}\frac{1}{|\mathcal{P}|} > \frac{1}{2}$, 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 1 is exponential rather than polynomial.

3 A partial structure theorem for induced C_4-free graphs

Our main goal in this section is to prove Lemma 13 stated below, which gives an approximate partial structure theorem for induced C_4-free graphs. The "approximation" will be due to the fact that the graph will only be close to having a certain nice structure, while the "partial" will be since there will be a (possibly) big part of the graph about which we will have no control. As we discussed in Section 1, this partialness is unavoidable as evidenced by split graphs.

In addition to the lemmas from the previous section, we will also need the following theorems of Goldreich, Goldwasser and Ron [16] and of Gyárfás, Hubenko and Solymosi [19]. In both cases, $\omega(G)$ denotes the maximum size of a clique in G.

▶ **Theorem 8** ([16], Theorem 7.1). *For every $\varepsilon \in (0,1)$ there is $q_8(\varepsilon) = O(\varepsilon^{-5})$ 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 containing a clique with at least ρn vertices. Suppose $q \geq q_8(\varepsilon)$ and let $Q \in \binom{V(G)}{q}$ be a randomly chosen set of q vertices of G. Then with probability at least $\frac{3}{4}$ we have $\omega(G[Q]) < (\rho - \frac{\varepsilon}{2})q$.*

▶ **Theorem 9** ([19]). *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 use derive the following important corollary of the the above two theorems. For a non-empty set $X \subseteq V(G)$, define $d(X) = e(X)/\binom{|X|}{2}$, where $e(X)$ is the number of edges of G with both endpoints in X.

▶ **Lemma 10.** *Let $\alpha \in [0, \frac{1}{2})$ and let G be a graph on n vertices with at least αn^2 edges. Then for every $\beta \in (0,1)$, either G contains $\Omega(\alpha^{80}\beta^{20}n^4)$ 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 10 we need the following simple fact.

▶ **Claim 11.** *Let $\alpha \in (0,1)$ and let G be a graph with n vertices and at least αn^2 edges. Then for every $r \geq \frac{100}{\alpha^2}$, a sample of r vertices from G spans at least $\frac{\alpha}{2}r^2$ edges with probability at least $\frac{2}{3}$.*

The proof of Claim 11 is a standard application of Chebyshev's inequality, and is thus omitted.

Proof of Lemma 10. Set $\rho = 0.1\alpha^2$, $\varepsilon = \frac{\rho^2\beta}{4} = \frac{\alpha^4\beta}{400}$ and $r = \max\{q_8(\varepsilon), \frac{100}{\alpha^2}\}$. By Theorem 8 we have $r = O(\alpha^{-20}\beta^{-5})$. We 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 $\Omega(\alpha^{80}\beta^{20}n^4)$ induced copies of C_4. Let $X \subseteq V(G)$ be such that $|X| \geq \rho n$. Since $d(X) \leq 1-\beta$, we have $\binom{|X|}{2} - e(G) \geq \beta\binom{|X|}{2} \geq \beta\frac{|X|^2}{4} \geq \frac{\rho^2\beta}{4}n^2 = \varepsilon n^2$. This shows that G is ε-far from containing a clique of size ρn or larger. By our choice of r via Theorem 8, 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 11, 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 9. We conclude that G contains at least $\frac{1}{3}\binom{n}{r}/\binom{n-4}{r-4} = \frac{1}{3}\binom{n}{4}/\binom{r}{4} = \Omega(\alpha^{80}\beta^{20}n^4)$ induced copies of C_4. ◀

The last ingredient we need is the following result of Alon, Fischer and Newman [3]. 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 to make (X,Y) induced M_2-free.

▶ **Lemma 12** ([3])**.** *There is an absolute constant $d > 0$ such that the following holds. If (X,Y) is ε-far from being induced M_2-free then (X,Y) contains at least $\varepsilon^d|X|^2|Y|^2$ induced copies of M_2.*

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.

▶ **Lemma 13.** *There is an absolute constant $c > 0$, such that for every $\alpha, \gamma \in (0,1)$, every n-vertex graph G either contains $\Omega(\alpha^c\gamma^c n^4)$ 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. We prove the lemma with $c = \max(84, 20d)$, where d is the constant from Lemma 12. We inductively define two sequences of sets, $(V_i)_{i\geq 0}$ and $(X_i)_{i\geq 1}$. Set $V_0 = V(G)$. At the i'th step (starting from $i = 0$), if $e(V_i) < \alpha n^2$ then we stop. Note that if we did not stop then $|V_i| \geq \alpha n$. If $e(V_i) \geq \alpha n^2$ then by Lemma 10, applied to $G[V_i]$ with parameters α and $\beta = 0.25\gamma^d$, either $G[V_i]$ contains $\Omega(\alpha^{80}\gamma^{20d}|V_i|^4) \geq \Omega(\alpha^{84}\gamma^{20d}n^4)$ 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 $d(X_i) \geq 1 - 0.25\gamma^d$. If the former case happens then 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 k'th 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 $d(X_i) \geq 1 - 0.25\gamma^d \geq 1 - \gamma$. Since the process stopped at the k'th 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 $0.5 \cdot 10^{-4}\alpha^{12}\gamma^d n^4$ induced copies of C_4. Assume that for some $1 \leq i < j \leq k$, the

pair (X_i, X_j) is γ-far from being induced M_2-free. By Lemma 12, (X_i, X_j) contains at least $\gamma^d|X_i|^2|X_j|^2$ 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 $d(X_i), d(X_j) \geq 1 - 0.25\gamma^d$, There are at most $0.5\gamma^d|X_i|^2|X_j|^2$ 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 $0.5\gamma^d|X_i|^2|X_j|^2 \geq 0.5 \cdot 10^{-4}\alpha^{12}\gamma^d n^4$ induced copies of C_4. ◀

We finish this section with the following corollary of the above structure theorem, which will be more convenient to use when proving Theorems 1 and 2 in the next section.

▶ **Lemma 14.** *There is an absolute constant $c > 0$ such that for every $\alpha, \gamma \in (0,1)$, every n-vertex graph G either contains $\Omega(\alpha^c\gamma^c n^4)$ induced copies of C_4 or there is a graph G' on $V(G)$, a partition $V(G) = X_1 \cup \cdots \cup X_k \cup Y$, where $k \leq 10\alpha^{-3}$, 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 $W_i \subseteq Q_i$ with the following properties.*

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. *In G', the sum of $|Q_i||Q_j|$ over all non-homogeneous pairs (Q_i, Q_j), $1 \leq i < j \leq q$, is at most αn^2.*
4. *(W_i, W_j) is homogeneous in G' for every $1 \leq i < j \leq q$ and $|W_i| \geq (\alpha/20)^{4000\alpha^{-6}}|X|$ for every $1 \leq i \leq q$.*
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. The constant c in this lemma is the same as in Lemma 13. Apply Lemma 13 to G with the given α and γ. If G contains $\Omega\left(\alpha^c\gamma^c n^4\right)$ induced copies of C_4 then the assertion of the lemma holds, and otherwise let $X_1, \ldots, X_k, Y$ be as in the statement of Lemma 13. Note that $k \leq 10\alpha^{-3}$ since $|X_i| \geq 0.1\alpha^3$ for every $1 \leq i \leq k$. 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 13 we have $|E(G''[Y])\triangle E(G[Y])| < \alpha n^2$ and $|E(G''[X])\triangle E(G[X])| < \gamma \sum_{i=1}^{k} \binom{|X_i|}{2} + \gamma \sum_{i<j} |X_i||X_j| < \gamma n^2$. We now apply Lemma 7 to $G''[X]$ with parameter $\delta = \alpha$ (and with respect to the partition $\{X_1, \ldots, X_k\}$) and obtain a subset $Z \subseteq X$ of size $|Z| < \alpha|X| \leq \alpha n$, a partition $X \setminus Z = Q_q \cup \cdots \cup Q_q$ which refines $\{X_1 \setminus Z, \ldots, X_k \setminus Z\}$, and subsets $W_i \subseteq Q_i$ such that $|W_i| \geq (\alpha/2k)^{10k^2}|X| \geq (\alpha^4/20)^{1000\alpha^{-6}}|X| \geq (\alpha/20)^{4000\alpha^{-6}}|X|$ for every $1 \leq i \leq q$.

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 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 3-4, this follows from our choice of $\mathcal{Q} = \{Q_1, \ldots, Q_q\}$ and $W_1, \ldots, W_q$ via Lemma 7 (with parameter $\delta = \alpha$). ◀

4 Proofs of main results

In this section we prove Theorems 1 and 2. The last ingredient we need is the following key lemma.

▶ **Lemma 15.** *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 $v \in V(F)$, the neighbourhood of v in F 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. Then, 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 $\{u,v\}$ and $\{u',v'\}$, 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 pairs of distinct vertices in $N_X(y)$ that are non-adjacent. Then the above discussion implies that every $y \in Y$ satisfies

$$d_2(y) \geq \frac{|\mathcal{M}(y)|^2}{2} \,. \tag{3}$$

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

$$|E(G') \triangle E(G)| = 2 \sum_{y \in Y} |\mathcal{M}(y)| \,. \tag{4}$$

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 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) and Jensen's inequality thus gives

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

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

$$\begin{aligned}\sum_{u,v \in X} \binom{t(u,v)}{2} &\geq \binom{|X|}{2} \cdot \binom{\sum_{u,v \in X} t(u,v) / \binom{|X|}{2}}{2} \\ &= \binom{|X|}{2} \cdot \binom{\sum_{y \in Y} d_2(y) / \binom{|X|}{2}}{2} \\ &\geq \frac{|X|^2}{4} \cdot \frac{(\varepsilon^2|Y|/4)^2}{4} = \frac{\varepsilon^4}{2^8}|X|^2|Y|^2,\end{aligned}$$

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 $\sum_{y \in Y} d_2(y)$. ◀

Proof of Theorem 1. We first observe that the "in particular" part of the statement, namely the testing algorithm, follows immediately from the first assertion of the theorem; indeed, the first assertion implies that if G is ε-far from being induced C_4-free, then sampling $2^{(1/\varepsilon)^c}$ 4-tuples of vertices will contain at least one induced 4-cycle with probability at least 2/3.

We now turn to prove the first assertion of the theorem. Set

$$\alpha = \frac{\varepsilon^6}{2^{11}}, \quad \gamma = \frac{1}{2}(\alpha/20)^{16000\alpha^{-6}}(\varepsilon/2)^4.$$

and notice that $\gamma \geq 2^{-(1/\varepsilon)^{c'}}$ for some absolute constant c'. We apply Lemma 14 to G with the α and γ defined above. If G contains $\Omega\left(\alpha^c\gamma^c n^4\right)$ induced copies of C_4 then we are done. 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 14. 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 14, one of these options holds. By Item 3 in Lemma 14, the number of changes made is at most αn^2. By Item 5 in Lemma 14 we have $|E(G'') \triangle E(G)| \leq |E(G'') \triangle E(G')| + |E(G') \triangle E(G)| < (3\alpha + \gamma)n^2 < \frac{\varepsilon}{2}n^2$, implying that G'' is $\frac{\varepsilon}{2}$-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 (and hence induced C_4-free) by deleting $|X \setminus Z| \cdot n \leq \frac{\varepsilon}{2}n^2$ 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'. By Item 4 in Lemma 14, G' contains

$$|W_{i_1}| \cdot |W_{i_2}| \cdot |W_{i_3}| \cdot |W_{i_4}| \geq (\alpha/20)^{16000\alpha^{-6}}|X|^4 \geq (\alpha/20)^{16000\alpha^{-6}}(\varepsilon/2)^4 n^4 = 2\gamma n^4$$

induced copies of C_4. By Item 5 in Lemma 14, $G[X \setminus Z]$ 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 γn^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. In particular, we have $|X||Y| \geq \varepsilon n^2$. Notice that the conditions of Lemma 15 hold (with respect to the family $\mathcal{F} = \{C_4\}$) since $G''[Y] = G'[Y]$ is an independent set (by Item 1 in Lemma 14) and $G''[X]$ is induced C_4-free by assumption. By Lemma 15, G'' contains at least $\frac{\varepsilon^4}{2^8}|X|^2|Y|^2 \geq \frac{\varepsilon^6}{2^8}n^4 = 8\alpha n^4$ induced copies of C_4. Since $|E(G'') \triangle E(G)| < (3\alpha + \gamma)n^2 < 4\alpha n^2$, at least $4\alpha n^4 = \frac{\varepsilon^6}{2^9}n^4$ of these copies are also present in G. This completes the proof of the theorem. ◀

Proof of Theorem 2. Set

$$\alpha = \frac{\varepsilon^6}{2^{11}}, \quad \gamma = \frac{1}{2}(\alpha/20)^{10^5\alpha^{-9}}(\varepsilon/2)^{20\alpha^{-3}}.$$

and notice that $\gamma \geq 2^{-(1/\varepsilon)^{c'}}$ for some absolute constant c'. As in the proof of Theorem 1, we apply Lemma 14 to G with the α and γ defined above. If G contains $\Omega\left(\alpha^c\delta^c n^4\right)$ induced copies of C_4 then we are done. 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 14.

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 make (Q_i, Q_j) a complete

(resp. empty) bipartite graph. As in the proof of Theorem 1, G'' is $\frac{\varepsilon}{2}$-far from being chordal, and we have $|X \setminus Z| \geq \frac{\varepsilon}{2}n$.

Assume first that $G''[X \setminus Z]$ is not chordal, namely that it contains an induced cycle $C = v_1 \dots v_\ell$ of length $\ell \geq 4$. By Item 1 in Lemma 14, $G''[X_i \setminus Z] = G'[X_i \setminus Z]$ is a clique for every $1 \leq i \leq k$. Since the cycle C does not contain a triangle, it can contain at most 2 vertices from each of these cliques, implying that $\ell = |C| \leq 2k \leq 20\alpha^{-3} = O(\varepsilon^{-18})$. The bound on k comes from Lemma 14. For $1 \leq s \leq \ell$, let i_s be such that $v_s \in Q_{i_s}$. It is easy to see that by the definition of G'', ℓ-tuple $(w_1, \dots, w_\ell) \in W_{i_1} \times \dots \times W_{i_\ell}$ spans an induced ℓ-cycle *in the graph* G'. By Item 4 in Lemma 14, G' contains

$$\prod_{j=1}^{\ell} |W_{i_j}| \geq (\alpha/20)^{4000\alpha^{-6}\ell} |X|^\ell \geq (\alpha/20)^{10^5\alpha^{-9}} (\varepsilon/2)^{20\alpha^{-3}} n^\ell = 2\gamma n^\ell$$

induced copies of C_ℓ. By Item 5 in Lemma 14, $G[X]$ and $G'[X]$ differ on less than γn^2 edges, each of which can participate in at most $n^{\ell-2}$ induced copies of C_ℓ. Thus, G contains at least γn^ℓ induced copies of C_ℓ, as required.

We now assume that $G''[X]$ is chordal. Since G'' is $\frac{\varepsilon}{2}$-far from being chordal, one must add/delete at least $\frac{\varepsilon}{2}n^2 \geq \varepsilon|X||Y|$ of the edges between X and Y to make G'' chordal. In particular, we have $|X||Y| \geq \varepsilon n^2$. Note that the family $\mathcal{F} = \{C_\ell : \ell \geq 4\}$, i.e. the family of forbidden induced subgraphs for chordality, satisfies Conditions 1-2 of Lemma 15. Observe that Lemma 15 is applicable to G'' (with respect to the family $\mathcal{F} = \{C_\ell : \ell \geq 4\}$), as $G''[Y] = G'[Y]$ is an independent set (by Item 1 in Lemma 14), and $G''[X]$ is induced $\mathcal{F}$-free (i.e. chordal) by assumption. By Lemma 15, G'' contains at least $\frac{\varepsilon^4}{2^8}|X|^2|Y|^2 \geq \frac{\varepsilon^6}{2^8}n^4 = 8\alpha n^4$ induced copies of C_4. Since $|E(G'') \triangle E(G)| < 4\alpha n^2$, at least $4\alpha n^4 = \frac{\varepsilon^6}{2^9}n^4$ of these copies are also present in G. ◀

5 An impossibility result

In this section we prove Theorem 3. It will in fact be more convenient to prove the following equivalent statement.

▶ **Theorem 16.** *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 [11].

▶ **Theorem 17.** *For every integer f there is $n_{17} = n_{17}(k, f)$ such that every k-uniform hypegraph with $n \geq n_{17}$ 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 18.** *For every pair of integers $k \geq 3$ and $f \geq 1$ there is $n_{18} = n_{18}(k, f)$ such that for every $n \geq n_{18}$, the graph $B_{k,n/k}$ is $\frac{1}{2k^2}$-far from being induced $\{C_4, B_{k,f}\}$-free.*

Proof. Let $V_1, \dots, V_k$ 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 $V_1, \dots, V_k$ whose edges are all k-tuples $(v_1, \dots, v_k) \in V_1 \times \cdots \times V_k$ such that $v_1 v_2 \dots v_k v_1$ 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 $n \geq n_{17}(k, f)$. Thus, by Theorem 17, 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 16. 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 $n \geq n_{18}(k, f_k)$ and set $G = B_{k,n/k}$. By Lemma 18, 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, for the sake of contradiction, that $x_1, \dots, x_\ell, x_1$ is an induced ℓ-cycle in G. Let $V_1, \dots, V_k$ be the sides of $G = B_{k,n/k}$. If $|\{x_1, \dots, x_\ell\} \cap V_i| \leq 1$ for every $1 \leq i \leq k$ then $x_1, \dots, x_\ell$ are contained in an induced path, which is impossible. So there is some $1 \leq i \leq k$ for which $|\{x_1, \dots, x_\ell\} \cap V_i| \geq 2$. Suppose without loss of generality that $x_1, x_2 \in V_1$ (recall that $V_1, \dots, V_k$ 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. ◀

References

1. N. Alon. Testing subgraphs in large graphs. *Random Struct. Alg.*, 21:359–370, 2002.
2. N. Alon, E. Fischer, M. Krivelevich, and M. Szegedy. Efficient testing of large graphs. *Combinatorica*, 20:451–476, 2000.
3. N. Alon, E. Fischer, and I. Newman. Testing of bipartite graph properties. *SIAM Journal on Computing*, 37:959–976, 2007.
4. 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*, 39:143–167, 2009.
5. N. Alon and J. Fox. Easily testable graph properties. *Combin. Probab. Comput.*, 24:646–657, 2015.
6. N. Alon and A. Shapira. A characterization of easily testable induced subgraphs. *Combin. Probab.Comput.*, 15:791–805, 2006.
7. N. Alon and A. Shapira. A characterization of the (natural) graph properties testable with one-sided error. *SIAM Journal on Computing*, 37:1703–1727, 2008.
8. L. Avigad and O. Goldreich. Testing graph blow-up. In *Proc. of APPROX-RANDOM*, pages 389–399. Springer, 2011.
9. D. Conlon and J. Fox. Bounds for graph regularity and removal lemmas. *GAFA*, 22:1191–1256, 2012.
10. D. Conlon and J. Fox. Graph removal lemmas. *Surveys in Combinatorics*, pages 1–50, 2013.

11 P. Erdős. On extremal problems of graphs and generalized graphs. *Israel J. Math*, 2:183–190, 1964.

12 P. Erdős. On some problems in graph theory, combinatorial analysis and combinatorial number theory. *Graph Theory and Combinatorics (Cambridge, 1983), Academic Press, London*, pages 1–17, 1984.

13 J. Fox. A new proof of the graph removal lemma. *Ann. of Math.*, 174:561–579, 2011.

14 L. Gishboliner and A. Shapira. Removal lemmas with polynomial bounds. *Proc. of STOC*, pages 510–522, 2017.

15 O. Goldreich. Introduction to property testing, Forthcoming book, 2017.

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

17 O. Goldreich and D. Ron. On proximity-oblivious testing. *SIAM J. on Computing*, 40:534–566, 2011.

18 T. Gowers. Lower bounds of tower type for szemerédi's uniformity lemma. *GAFA*, 7:322–337, 1997.

19 A. Gyárfás, A. Hubenko, and J. Solymosi. Large cliques in c_4-free graphs. *Combinatorica*, 22:269–274, 2002.

20 L. Lovász. *Large networks and graph limits*, volume 60. American Mathematical Society Providence, 2012.

21 G. Moshkovitz and A. Shapira. A sparse regular aproximation lemma.

22 V. Rödl and R. Duke. On graphs with small subgraphs of large chromatic number. *Graphs and Combinatorics*, 1:91–96, 1985.

23 R. Rubinfeld and M. Sudan. Robust charachterization of polynomials with applications to program testing. *SIAM Journal on Computing*, 25:252–271, 1996.

24 I.Z. Ruzsa and E. Szemerédi. Triple systems with no six points carrying three triangles. *Combinatorics (Keszthely, 1976), Coll. Math. Soc. J. Bolyai*, 18:939–945, 1976.

25 E. Szemerédi. Regular partitions of graphs. *In:* Proc. Colloque Inter. CNRS*, J. C. Fournier, M. Las Vergnas and D. Sotteau, eds.*, pages 399–401, 1978.