Efficient Removal without Efficient Regularity

Lior Gishboliner * Asaf Shapira †

May 28, 2018

Abstract

Obtaining an efficient bound for the triangle removal lemma is one of the most outstanding open problems of extremal combinatorics. Perhaps the main bottleneck for achieving this goal is that triangle-free graphs can be highly unstructured. For example, triangle-free graphs might have only regular partitions (in the sense of Szemerédi) of tower-type size. And indeed, essentially all the graph properties $\mathcal{P}$ for which removal lemmas with reasonable bounds were obtained, are such that every graph satisfying $\mathcal{P}$ has a small regular partition. So in some sense, a barrier for obtaining an efficient removal lemma for property $\mathcal{P}$ was having an efficient regularity lemma for graphs satisfying $\mathcal{P}$.

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 the induced C_4 removal lemma. We thus obtain the first efficient removal lemma that does not rely on an efficient version of the regularity lemma. This is the first substantial progress on a problem raised by Alon in 2001, and more recently by Alon, Conlon and Fox.

1 Introduction

An n-vertex graph is ε-far from satisfying a property $\mathcal{P}$ if one should add/delete at least εn^2 edges in order to turn G into a graph satisfying $\mathcal{P}$. The so called *induced removal lemma* of Alon, Fischer, Krivelevich and Szegedy [2] states that for every fixed graph H, if an n-vertex graph G is ε-far from being induced H-free, then G contains at least $n^h/\text{Rem}_H(\varepsilon)$ induced copies of H, where $h = |V(H)|$ and $\text{Rem}_H(\varepsilon)$ depends only on ε. The proof of this lemma in [2] supplied extremely weak bounds for $\text{Rem}_H(\varepsilon)$, which were later improved by Conlon and Fox [9]. However, even these improved bounds are of tower-type[1].

Alon [1] asked for which graphs H we have $\text{Rem}_H(\varepsilon) = \text{poly}(1/\varepsilon)$, that is, for which graphs H can we obtain polynomial bounds for the induced removal lemma. This question was addressed by Alon and the second author [5] who resolved this problem for all graphs H except for P_3 (the path

*School of Mathematical Sciences, Tel Aviv University, Tel Aviv, 69978, Israel.

†School of Mathematical Sciences, Tel Aviv University, Tel Aviv 69978, Israel. Email: asafico@tau.ac.il. Supported in part by ISF Grant 1028/16 and ERC Starting Grant 633509.

[1] We use tower(x) for a tower of exponents of height x, so $\text{tower}(3) = 2^{2^2}$. The original proof of the induced removal lemma in [2] gave only wowzer-type bounds, where wowzer is the iterated-tower function.

on 4 vertices) and C_4 (the 4-cycle). The former case was recently solved by Alon and Fox [4], who proved that $\mathrm{Rem}_{P_3}(\varepsilon) = \mathrm{poly}(1/\varepsilon)$. They further asked to determine if $\mathrm{Rem}_{C_4}(\varepsilon) = \mathrm{poly}(1/\varepsilon)$. This problem was also later raised by Conlon and Fox [10].

Prior to this work the best bound for $\mathrm{Rem}_{C_4}(\varepsilon)$ was the same tower-type bound that holds for all graphs H. As we explain in the next subsection, the reason is that this problem seemed to lie just outside the realm of the known techniques for proving efficient bounds for graph removal lemmas. Our main result in this paper makes the first substantial progress on this problem, by improving the tower-type bound into an exponential one.

Theorem 1.1. *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.*

We conjecture that the exponential bound in Theorem 1.1 can be further improved to a polynomial one.

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 $H \in \mathcal{F}$. Observe that for infinite families $\mathcal{F}$ it is not a priori clear that a graph which is ε-far from being induced $\mathcal{F}$-free should contain *any* constant size (that might depend on ε) subgraph that is not induced $\mathcal{F}$-free. Such a result was obtained by Alon and the second author [6], who extended the result of [2] by showing that for every family of graphs $\mathcal{F}$, there is a function $\mathrm{Rem}_{\mathcal{F}}(\varepsilon)$, so that if G is ε-far from being induced $\mathcal{F}$-free, then a random subset of $\mathrm{Rem}_{\mathcal{F}}(\varepsilon)$ vertices from $V(G)$ is not induced $\mathcal{F}$-free with probability at least (say[2]) 2/3. Needless to say that as in [2], the bounds for $\mathrm{Rem}_{\mathcal{F}}(\varepsilon)$ given by [6] are also (at least) of tower-type.

It is natural to ask if Theorem 1.1 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 this problem was the tower-type bound which follows from the general result of [6]. Here we obtain the following improved bound.

Theorem 1.2. *There is an absolute constant c, such that for every $\varepsilon \in (0,1)$, for every integer n and for every n-vertex graph G which is ε-far from being chordal, there is $4 \le \ell \le O(\varepsilon^{-18})$ such that G contains at least $n^\ell/2^{(1/\varepsilon)^c}$ induced copies of C_ℓ.*

While Theorem 1.2 asserts that if G is ε-far from being chordal it must contain an induced cycle of length $\mathrm{poly}(1/\varepsilon)$, it only implies that a sample of vertices of size $2^{(1/\varepsilon)^c}$ contains an induced cycle with probability at least 2/3. We do believe, however, that this exponential bound can be further improved to a polynomial one.

It is now natural to ask if Theorem 1.2 can be further extended to an arbitrary family of graphs $\mathcal{F}$, one of which is C_4. As our final theorem shows, this is not the case in a very strong sense.

Theorem 1.3. *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 $\mathrm{Rem}_{\mathcal{F}}(\varepsilon) \ge g(\varepsilon)$.*

In fact, for every (small enough) $\varepsilon > 0$ and every $n \ge n_0(\varepsilon)$, 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.*

[2]It is easy to see that if $\mathcal{F} = \{H\}$, this way of defining $\mathrm{Rem}_{\mathcal{F}}(\varepsilon)$ is equivalent (up to polynomial factors) to the induced removal lemma of [2], as we stated it above.

1.1 Relation to prior works

In this subsection we would like to explain why in Theorem 1.1 we managed to overcome for the first time a natural barrier, which was the main reason why one could not derive Theorem 1.1 via techniques that were previously used for proving graph removal lemmas. For simplicity we will start by discussing the triangle removal lemma, that is, the special case of the induced removal lemma when $H = K_3$. The original proof of the triangle removal lemma [25] relied on the famous regularity lemma of Szemerédi [26], which is one of the most powerful tools for tackling problems in extremal graph theory. It states that for every $\varepsilon > 0$ there is an $M = M(\varepsilon)$ so that every graph has an ε-regular partition of order at most M (see [24] for the precise definitions related to graph regularity). Since Szemerédi's proof only established that $M(\varepsilon) \leq \text{tower}(1/\varepsilon^5)$, this approach for proving the triangle removal lemma only gave the very weak bound $\text{Rem}_{K_3}(\varepsilon) \leq \text{tower}(\text{poly}(1/\varepsilon))$. Gowers' celebrated result [18], which states that $M(\varepsilon) \geq \text{tower}(\text{poly}(1/\varepsilon))$, implies that one cannot get a better bound for $\text{Rem}_{K_3}(\varepsilon)$ via the regularity lemma. In a major breakthrough, Fox [13] managed to prove the triangle removal lemma while avoiding Szemerédi's version of the regularity lemma, thus showing that $\text{Rem}_{K_3}(\varepsilon) \leq \text{tower}(O(\log 1/\varepsilon))$. 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. Hence this approach does not seem to allow one to prove (say) exponential bounds for the triangle removal lemma.

Although the best known bounds for the triangle removal are of tower-type, there are families of graphs $\mathcal{F}$ for which one *can* prove much better (non-tower-type) bounds for $\text{Rem}_{\mathcal{F}}(\varepsilon)$, that is, for the removal lemma of induced $\mathcal{F}$-freeness. One example is the result of Alon and Fox [4] mentioned above regarding induced P_3-freeness. The main point we would like to make is that all these improved bounds (save for one case discussed below) were not obtained by avoiding the regularity lemma. Instead, they still (implicitly or explicitly) used the regularity lemma, but relied on the fact that induced $\mathcal{F}$-free graphs have much smaller ε-regular partitions. For example, the result of Alon and Fox [4] regarding induced P_3-freeness can be derived from the fact that every induced P_3-free graph has an ε-regular partition of size $\text{poly}(1/\varepsilon)$. See [14] for a proof of this and other related results.

It is now natural to ask if one can use the above approach in order to obtain better bounds for the triangle removal lemma. Unfortunately, there are bipartite versions of Gowers' [18] lower bound for the regularity lemma, as well as for the variant of the regularity lemma introduced in [21]. Therefore, a graph can be triangle-free but still only have regular partitions of tower-type size. This means that any proof of the triangle removal lemma that relies on (one of the above versions of) the regularity lemma is bound to produce tower-type bounds.

With regard to induced C_4-freeness, it is easy to see 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 inducing a complete graph and the other an independent set. This means that if we take a bipartite version of Gowers' lower bound [18] (or of the one from [21]), 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 removal lemmas cannot give better-than-tower-type bounds for this problem.

Summarizing the above discussion, Theorem 1.1 is the first example showing that one *can* obtain an efficient removal lemma for a property $\mathcal{P}$, even though graphs satisfying $\mathcal{P}$ might have only regular partitions of tower-type size. To do this, our proof is the first removal lemma that avoids using the regularity lemma or one of its variants (save for the example discussed below). We are hopeful that bounds similar to those obtained in Theorem 1.1 can be obtained for removal lemmas of other properties for which the best known bounds are of tower-type, most notably for triangle freeness.

Let us end the discussion by describing the only previous example of a removal lemma that was obtained while avoiding a regularity lemma, and how it differs from Theorem 1.1. In 1984 Erdős [12] (implicitly) conjectured that k-colorability has a removal lemma, that is, that if G is ε-far from being k-colorable then a sample of $C_k(\varepsilon)$ vertices from $V(G)$ spans a non-k-colorable subgraph with probability at least $2/3$. This was first verified by Rödl and Duke [23] who used the regularity lemma in order to obtain a tower-type bound for $C_k(\varepsilon)$. This tower-type bound was dramatically improved by Goldreich, Goldwasser and Ron [16], who obtained a new proof of this result (as well as for similar *partition problems*) that avoided the regularity lemma and thus gave a polynomial bound for $C_k(\varepsilon)$. Let us try[3] to explain why k-colorability differs from triangle-freeness or induced C_4-freeness. First, as opposed to these two properties which are *local*, 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[4]. The second difference, which is more important for our quantitative investigation here, is that k-colorability is defined using global edge counts (i.e. having no edges inside a vertex partition into k sets). This can explain (at least in hindsight), why one does not need any structure theorem in order to handle this property. Instead one can 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!). It appears that arguments of this sort cannot be used to prove removal lemmas for local properties such as triangle freeness or induced C_4-freeness.

1.2 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[5] 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[6]. 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 3.6. 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 give the proofs of Theorems 1.1 and 1.2. 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 4.1. Finally, in Section 5, we give the proof of Theorem 1.3. We will make no effort to optimize the constant c appearing in Theorems 1.1 and 1.2.

[3]See also Subsection 8.3.2 of Goldreich's upcoming book [15] for a similar attempt.

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

[5]It is known [22] that *most* induced C_4-free graphs are split graphs. We stress that in our setting we have to deal with *every* induced C_4-free graph, not just typical ones!

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

2 Forbidding an induced 2-matching

Our goal in this section is to introduce several definitions and prove Lemma 2.4 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 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 2.1. *(X, Y) is induced M_2-free*[7] *if and only if there is an enumeration $x_1, \dots, x_m$ of the elements of X such that $N_Y(x_i) \subseteq N_Y(x_j)$ for every $1 \le i < j \le m$.*

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 linear ordering. Enumerate the elements of X from minimal to maximal to get the required enumeration. $\square$

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 the paper, 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 2.2. *If (X, Y) is induced M_2-free then for every integer $r \ge 1$ there are partitions $X = X_1 \cup \dots \cup X_r$ and $Y = Y_1 \cup \dots \cup Y_{r+1}$ such that $|X_i| = \frac{|X|}{r}$ for every $1 \le i \le r$, and (X_i, Y_j) is homogeneous for every $1 \le i \le r$ and $1 \le j \le r+1$ satisfying $i \ne j$.*

Proof. Let $x_1, \dots, x_m$ be the enumeration of the elements of X from Claim 2.1. For $1 \le i \le r$ define $X_i = \{x_j : \frac{(i-1)m}{r} < j \le \frac{im}{r}\}$. Let now $y_1, \dots, y_n$ 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 $N_Y(x) = \{y_1, \dots, y_k\}$ for some $0 \le k \le n$. Define $Y_1 = N_Y(x_{m/r})$, $Y_i = N_Y(x_{im/r}) \setminus N_Y(x_{(i-1)m/r})$ for $i = 2, \dots, r$ and $Y_{r+1} = Y \setminus N_Y(x_m)$.

It remains to show that (X_i, Y_j) is homogeneous for every $i \ne j$. 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. $\square$

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

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

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 $\{X_1, \dots, X_k\}$ such that each X_i is either a clique or an independent set, then every refinement of $\mathcal{P}$ is also δ-homogeneous.

Lemma 2.3. *Let $\delta \in (0, 1/2)$, let G be an n-vertex graph and let $V(G) = X_1 \cup \dots \cup X_k$ be a partition such that $X_1, \dots, X_k$ are cliques and (X_i, X_j) is induced M_2-free for every $1 \le i < j \le k$. Then there is a δ-homogeneous partition which refines $\{X_1, \dots, X_k\}$ and has at most $k\,(2/\delta)^k$ parts.*

Proof. For every $1 \le i < j \le k$, we apply Lemma 2.2 to (X_i, X_j) with parameter $r = \left\lceil \frac{1}{\delta} \right\rceil$ to get partitions $\mathcal{P}_{i,j}$ of X_i and $\mathcal{P}_{j,i}$ of X_j, $\mathcal{P}_{i,j} = \{X^1_{i,j}, \dots, X^r_{i,j}\}$, $\mathcal{P}_{j,i} = \{X^1_{j,i}, \dots, X^{r+1}_{j,i}\}$, such that $|X^p_{i,j}| = \frac{|X_i|}{r}$ for every $1 \le p \le r$, and $(X^p_{i,j}, X^q_{j,i})$ is homogeneous for every $p \neq q$. Note that

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

For every $i = 1, \dots, k$, define $\mathcal{P}_i$ to be the common refinement of the partitions $(\mathcal{P}_{i,j})_{1 \le j \le k,\, j \neq i}$. We have $|\mathcal{P}_i| \le (r+1)^{k-1} \le (\frac{1}{\delta} + 2)^{k-1} \le (2/\delta)^k$. The partition $\mathcal{P} := \bigcup_{i=1}^k \mathcal{P}_i$ refines $\{X_1, \dots, X_k\}$ 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 $1 \le i < j \le k$ and $1 \le p \le r$ such that $U \subseteq X^p_{i,j}$ and $V \subseteq X^p_{j,i}$. This follows from the fact that $X_1, \dots, X_k$ are cliques and the property of the partitions $(\mathcal{P}_{i,j})_{1 \le i \neq j \le k}$. By (1), we have

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

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

We are ready to prove the main lemma of this section. One important idea we will employ here is that of obtaining a partition $\mathcal{P}$ and then a refinement $\mathcal{Q}$ of $\mathcal{P}$ with somewhat better parameters, and then using some parts of $\mathcal{Q}$ in order to approximate $\mathcal{P}$. This idea goes back to the original proof of the induced removal lemma in [2].

Lemma 2.4. *Let $\delta \in (0, 1/2)$, let G be an n-vertex graph and let $V(G) = X_1 \cup \dots \cup X_k$ be a partition such that $X_1, \dots, X_k$ 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 $V(G) \setminus Z = Q_1 \cup \dots \cup Q_q$ which refines $\{X_1 \setminus Z, \dots, X_k \setminus Z\}$ 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. *$|W_i| \ge (\delta/2k)^{10k^2} n$ 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 is trivial for $k = 1$ so we assume that $k \ge 2$. Apply Lemma 2.3 to G with parameter δ to obtain a δ-homogeneous partition $\mathcal{P}$ which refines $\{X_1, \dots, X_k\}$. Let us define $\mathcal{Q} = \{U \in \mathcal{P} : |U| \ge \frac{\delta n}{|\mathcal{P}|}\}$ and write $\mathcal{Q} = \{Q_1, \dots, Q_q\}$. 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 $\{X_1 \setminus Z, \dots, X_k \setminus Z\}$ and that $|Z| < |\mathcal{P}| \cdot \frac{\delta n}{|\mathcal{P}|} = \delta n$. Apply Lemma 2.3 to $G[V(G) \setminus Z]$, with the partition $\{X_1 \setminus Z, \dots, X_k \setminus Z\}$ (in place of $\{X_1, \dots, X_k\}$), and with parameter $\delta' := \frac{\delta^2}{4|\mathcal{P}|^4}$. The lemma gives a δ'-homogeneous partition $\mathcal{V}$ with at most $k(8|\mathcal{P}|^4/\delta^2)^k$ 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

$$|\mathcal{W}| \leq |\mathcal{Q}| \cdot |\mathcal{V}| \leq |\mathcal{P}| \cdot k(8|\mathcal{P}|^4/\delta^2)^k. \tag{2}$$

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 $W_1, \ldots, W_q$ 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 $\frac{|\mathcal{W}| \cdot \frac{|Q_i|}{2q|\mathcal{W}|}}{|Q_i|} = \frac{1}{2q}$, as there are at most $|\mathcal{W}_i| \leq |\mathcal{W}|$ sets $W \in \mathcal{W}_i$ of size less than $\frac{|Q_i|}{2q|\mathcal{W}|}$. By the union bound, with probability larger than $\frac{1}{2}$, every $1 \leq i \leq q$ satisfies

$$|W_i| \geq \frac{|Q_i|}{2q|\mathcal{W}|} \geq \frac{\left(\frac{\delta^2}{8|\mathcal{P}|^4}\right)^k \delta n}{2k|\mathcal{P}|^3} = \frac{\delta^{2k+1} n}{2^{3k+1} k |\mathcal{P}|^{4k+3}} \geq \frac{\delta^{2k+1} n}{2^{3k+1} k^{4k+4} (2/\delta)^{4k^2+3k}} \geq \left(\frac{\delta}{2k}\right)^{10k^2} n \ ,$$

where in the second inequality we used $|Q_i| \geq \frac{\delta n}{|\mathcal{P}|}$, $q \leq |\mathcal{P}|$ and (2), in the third inequality we used the bound on $|\mathcal{P}|$ given by Lemma 2.3, and in the last inequality we used the assumption $k \geq 2$. For $1 \leq i < j \leq q$, the probability that the pair (W_i, W_j) is not homogeneous is

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

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

It is worth mentioning that the bounds in the above lemma are the sole reason why our bound in Theorem 1.1 is exponential rather than polynomial.

3 A partial structure theorem for C_4-free graphs

Our main goal in this section is to prove Lemma 3.6 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 maximum size of a clique in G.

Theorem 3.1 ([16], Theorem 7.1)**.** *For every $\varepsilon \in (0,1)$ there is $q_{3.1}(\varepsilon) = O(\varepsilon^{-5})$ 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 satisfying $\omega(G) \geq \rho n$. Suppose $q \geq q_{3.1}(\varepsilon)$ 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 ([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 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 3.3. *Let $\alpha \in (0, 1/2)$ 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 3.3 we need the following simple fact, which is proved by a standard application of the second moment method (see e.g. [7]). We first prove Claim 3.4 and then present the proof of Lemma 3.3.

Claim 3.4. *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. [7]), we have

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

where in the last inequality we used our choice of r. Finally, notice that if $|e(R) - \mathbb{E}[e(R)]| \leq \frac{\eta}{4} r^2$ 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. $\square$

Proof of Lemma 3.3. Set $\rho = 0.1\alpha^2$, $\varepsilon = \frac{\rho^2 \beta}{4} = \frac{\alpha^4 \beta}{400}$ and $r = \max\{q_{3.1}(\varepsilon), \frac{240}{\alpha}\}$. By Theorem 3.1 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 $|X| \ge \rho n$. Since $d(X) < 1-\beta$, we have $\binom{|X|}{2} - e(X) > \beta\binom{|X|}{2} \ge \beta\frac{|X|^2}{4} \ge \frac{\rho^2\beta}{4}n^2 = \varepsilon n^2$. This shows that G is ε-far from containing a clique of size ρn or larger. By our choice of r via Theorem 3.1, 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 3.4, 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. 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. $\square$

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 3.5 ([3])**.** *There is an absolute constant d 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.*

We can clearly assume that $d \ge 1$ in the above lemma, and will do so in what follows. The following is the key lemma of this section. Note that it gives us a lot of information about $G[Y]$ and $G[X_1 \cup \dots \cup X_k]$ but no information about the bipartite graph connecting $X_1 \cup \dots \cup X_k$ and Y.

Lemma 3.6. *There is an absolute constant $c_0 > 0$, such that for every $\alpha, \gamma \in (0, 1/2)$, every n-vertex graph G either contains $\Omega(\alpha^{c_0}\gamma^{c_0}n^4)$ induced copies of C_4, or admits a vertex partition $V(G) = X_1 \cup \dots \cup X_k \cup Y$ with the following properties.*

1. *$e(Y) < \alpha n^2$.*
2. *$|X_i| \ge 0.1\alpha^3 n$ and $d(X_i) \ge 1 - \gamma$ for every $1 \le i \le k$.*
3. *For every $1 \le i < j \le k$, the pair (X_i, X_j) is γ-close (i.e. not γ-far) to being induced M_2-free.*

Proof. We prove the lemma with $c_0 = \max(84, 20d)$, where d is the constant from Lemma 3.5. We inductively define two sequences of sets, $(V_i)_{i\ge 0}$ and $(X_i)_{i\ge 1}$. 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 $|V_i| \ge \alpha^{1/2}n > \alpha n$. If $e(V_i) \ge \alpha n^2$ then by Lemma 3.3, applied to $G[V_i]$ with parameters α and $\beta = 0.25\gamma^d$, either $G[V_i]$ contains $\Omega(\alpha^{80}\gamma^{20d}|V_i|^4) \ge \Omega(\alpha^{84}\gamma^{20d}n^4)$ induced copies of C_4 or there is $X_{i+1} \subseteq V_i$ with $|X_{i+1}| \ge 0.1\alpha^2|V_i| \ge 0.1\alpha^3 n$ and $d(X_i) \ge 1 - 0.25\gamma^d$. 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 $k \ge 0$. Set $Y = V_k$. We clearly have $V(G) = X_1 \cup \dots \cup X_k \cup Y$. For every $1 \le i \le k$ we have $|X_i| \ge 0.1\alpha^3 n$ and $d(X_i) \ge 1 - 0.25\gamma^d \ge 1 - \gamma$. 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 $1 \le i < j \le k$, the pair (X_i, X_j) is γ-far from being induced M_2-free. By Lemma 3.5, (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 $d(X_i), d(X_j) \ge 1 - 0.25\gamma^d$, 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. $\square$

We finish this section with the following corollary of the above structure theorem, which will be more convenient to use when proving Theorems 1.1 and 1.2 in the next section.

Lemma 3.7. *There is an absolute constant $c_0 > 0$ such that for every $\alpha, \gamma \in (0, 1/2)$, every n-vertex graph G either contains $\Omega(\alpha^{c_0}\gamma^{c_0}n^4)$ 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 $0 \leq k \leq 10\alpha^{-3}$, 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:*

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_0 in this lemma is the same as in Lemma 3.6. Apply Lemma 3.6 to G with the given α and γ. If G contains $\Omega\left(\alpha^{c_0}\gamma^{c_0}n^4\right)$ 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 3.6. Note that $k \leq 10\alpha^{-3}$ since $|X_i| \geq 0.1\alpha^3 n$ 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 3.6 we have $|E(G''[Y]) \triangle E(G[Y])| < \alpha n^2$ and $|E(G''[X]) \triangle E(G[X])| < \gamma \sum_{i=1}^{k} \binom{|X_i|}{2} + \gamma \sum_{i<j} |X_i||X_j| < \gamma n^2$. We now apply Lemma 2.4 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 $X \setminus Z = Q_1 \cup \cdots \cup Q_q$ which refines $\{X_1 \setminus Z, \ldots, X_k \setminus Z\}$, and subsets $W_i \subseteq Q_i$ such that $|W_i| \geq (\alpha/2k)^{10k^2}|X| \geq (\alpha^4/20)^{1000\alpha^{-6}}|X| \geq (\alpha/20)^{4000\alpha^{-6}}|X|$ 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 and 4, this follows from our choice of $\mathcal{Q} = \{Q_1, \ldots, Q_q\}$ and $W_1, \ldots, W_q$ via Lemma 2.4 (with parameter $\delta = \alpha$). $\square$

4 Proofs of main results

In this section we prove Theorems 1.1 and 1.2. The last ingredient we need is the following key lemma.

Lemma 4.1. *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

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

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 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) 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

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

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

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 $\sum_{y \in Y} d_2(y)$. $\square$

We are now ready to prove Theorems 1.1 and 1.2.

Proof of Theorem 1.1. Set

$$
\alpha = \frac{\varepsilon^6}{2^{13}}, \quad \gamma = \frac{1}{2}(\alpha/20)^{16000\alpha^{-6}}(\varepsilon/2)^4,
$$

and notice that $\gamma \geq 2^{-(1/\varepsilon)^{c'}}$ for some absolute constant c'. Let G be an n-vertex graph which is ε-far from being induced C_4-free. We apply Lemma 3.7 to G with the α and γ defined above. If G contains $\Omega\left(\alpha^{c_0}\gamma^{c_0}n^4\right)$ induced copies of C_4 (where c_0 is from Lemma 3.7) 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 3.7. 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 3.7, one of these options holds. By Item 3 in Lemma 3.7, the number of changes made is at most αn^2. By Item 5 in Lemma 3.7 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 (and hence 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'. By Item 4 in Lemma 3.7, 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 3.7, $G[X \setminus Z]$ 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 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. 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 4.1 hold (with respect to the family $\mathcal{F} = \{C_4\}$) since $G''[Y] = G'[Y]$ is an independent set (by Item 1 in Lemma 3.7) and $G''[X]$ is induced C_4-free by assumption. By Lemma 4.1, 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. $\square$

Proof of Theorem 1.2. Set

$$\alpha = \frac{\varepsilon^6}{2^{13}}, \quad \gamma = \frac{1}{2}(\alpha/20)^{10^5\alpha^{-9}}(\varepsilon/2)^{20\alpha^{-3}},$$

and notice that $\gamma \geq 2^{-(1/\varepsilon)^{c'}}$ for some absolute constant c'. Let G be an n-vertex graph which is ε-far from being chordal. As in the proof of Theorem 1.1, we apply Lemma 3.7 to G with the α and γ defined above. If G contains $\Omega\left(\alpha^{c_0}\gamma^{c_0}n^4\right)$ 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 3.7.

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 make (Q_i, Q_j) a complete (resp. empty) bipartite graph. As in the proof of Theorem 1.1, 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 $C = v_1 \ldots v_\ell$ of length $\ell \geq 4$. By Item 1 in Lemma 3.7, $G''[X_i \setminus Z] = G'[X_i \setminus Z]$ 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 3.7. 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 $(w_1, \ldots, w_\ell) \in W_{i_1} \times \cdots \times W_{i_\ell}$ spans an induced ℓ-cycle *in the graph* G'. By Item 4 in Lemma 3.7, 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 3.7, $G[X]$ 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 $|X||Y| \geq \frac{\varepsilon}{2}n^2$. 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 4.1. Observe that Lemma 4.1 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 3.7), and $G''[X]$ is induced $\mathcal{F}$-free (i.e. chordal) by assumption. By Lemma 4.1, 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 $|E(G'')\triangle E(G)| < 4\alpha n^2$, at least $4\alpha n^4 = \frac{\varepsilon^6}{2^{11}}n^4$ of these copies are also present in G. So in this case the assertion of the theorem holds with $\ell = 4$. This completes the proof. $\square$

5 An impossibility result

In this section we prove Theorem 1.3. It will in fact be more convenient to prove the following equivalent statement.

Theorem 5.1. *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 [11].

Theorem 5.2. *For every integer f there is $n_{5.2} = n_{5.2}(k, f)$ such that every k-uniform hypegraph with $n \geq n_{5.2}$ 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 5.3. *For every pair of integers $k \geq 3$ and $f \geq 1$ there is $n_{5.3} = n_{5.3}(k, f)$ such that for every $n \geq n_{5.3}$, the graph $B_{k,n/k}$ is $\frac{1}{2k^2}$-far from being induced $\{C_4, B_{k,f}\}$-free.*

Proof. Let $V_1, \dots, V_k$ 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 $(v_1, \dots, v_k) \in V_1 \times \dots \times V_k$ 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 $n \geq n_{5.2}(k, f)$. Thus, by Theorem 5.2, 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 $U_1 \cup \dots \cup U_k$ 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. $\square$

Proof of Theorem 5.1. 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 $n \geq n_{5.3}(k, f_k)$ and set $G = B_{k,n/k}$. By Lemma 5.3, 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, 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. $\square$

Acknowledgement: We would like to thank the referees for many valuable comments, which significantly improved the presentation of this paper.

References

[1] N. Alon, Testing subgraphs in large graphs, Random Structures and Algorithms 21 (2002), 359-370. 1

[2] N. Alon, E. Fischer, M. Krivelevich and M. Szegedy, Efficient testing of large graphs, Combinatorica 20 (2000), 451-476. 1, 1, 1, 2, 2

[3] N. Alon, E. Fischer, and I. Newman, Testing of bipartite graph properties, SIAM Journal on Computing 37 (2007), 959-976. 3, 3.5

[4] N. Alon and J. Fox, Easily testable graph properties, Combin. Probab. Comput., 24 (2015), 646-657. 1, 1.1

[5] N. Alon and A. Shapira, A characterization of easily testable induced subgraphs, Combin. Probab.Comput. 15 (2006), 791-805. 1

[6] 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. 1

[7] N. Alon and J. H. Spencer, **The Probabilistic Method**, 3rd ed., Wiley, 2008. 3, 3

[8] L. Avigad and O. Goldreich, Testing graph blow-up, Proc. of APPROX-RANDOM 2011, 389-399.

[9] D. Conlon and J. Fox, Bounds for graph regularity and removal lemmas, GAFA **22** (2012), 1191–1256. 1

[10] D. Conlon and J. Fox, Graph removal lemmas, Surveys in Combinatorics, Cambridge university press, 2013, 1-50. 1, 1.1

[11] P. Erdős, On extremal problems of graphs and generalized graphs, Israel J. Math. 2, 1964, 183-190. 5

[12] 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. 1.1

[13] J. Fox, A new proof of the graph removal lemma, Ann. of Math. **174** (2011), 561–579. 1.1

[14] L. Gishboliner and A. Shapira, Removal lemmas with polynomial bounds, Proc. of STOC 2017. 1.1

[15] O. Goldreich, **Introduction to Property Testing**, Forthcoming book, 2017. 3

[16] O. Goldreich, S. Goldwasser, and D. Ron, Property testing and its connection to learning and approximation, J. ACM 45 (1998), 653-750. 1.1, 1.2, 3, 3.1

[17] O. Goldreich and D. Ron, On proximity-oblivious testing, SIAM J. on Computing 40 (2011), 534-566.

[18] T. Gowers, Lower bounds of tower type for Szemerédi's uniformity lemma, Geom. Funct. Anal. **7** (1997), 322–337. 1.1

[19] A. Gyárfás, A. Hubenko and J. Solymosi, Large cliques in C_4-free graphs, Combinatorica, 22 (2002), 269-274. 3, 3.2

[20] L. Lovász, **Large networks and graph limits** (Vol. 60), Providence: American Mathematical Society (2012). 4

[21] G. Moshkovitz and A. Shapira, A sparse regular aproximation lemma, Transactions of the AMS, to appear. 1.1

[22] H. Prömel and A. Steger, Excluding induced subgraphs: quadrilaterals, Random Structures and Algorithms 2 (1991), 55-71. 5

[23] V. Rödl and R. Duke, On graphs with small subgraphs of large chromatic number, Graphs and Combinatorics 1 (1985), 91-96. 1.1

[24] V. Rödl and M. Schacht, Regularity lemmas for graphs, Fete of Combinatorics and Computer Science, Bolyai Soc. Math. Stud., 20 (2010), 287–325. 1.1

[25] I.Z. Ruzsa and E. Szemerédi, Triple systems with no six points carrying three triangles, in Combinatorics (Keszthely, 1976), Coll. Math. Soc. J. Bolyai 18, Volume II, 939-945. 1.1

[26] E. Szemerédi, Regular partitions of graphs, In: *Proc. Colloque Inter. CNRS*, 1978, 399–401. 1.1