WHICH GRAPHS CAN BE COUNTED IN C_4-FREE GRAPHS?

DAVID CONLON, JACOB FOX, BENNY SUDAKOV, AND YUFEI ZHAO

Abstract. For which graphs F is there a sparse F-counting lemma in C_4-free graphs? We are interested in identifying graphs F with the property that, roughly speaking, if G is an n-vertex C_4-free graph with on the order of $n^{3/2}$ edges, then the density of F in G, after a suitable normalization, is approximately at least the density of F in an ϵ-regular approximation of G. In recent work, motivated by applications in extremal and additive combinatorics, we showed that C_5 has this property. Here we construct a family of graphs with the property.

1. Introduction

When applying the regularity method in extremal graph theory, proofs can often be divided into two steps: first applying Szemerédi's regularity lemma to partition a large graph so that most pairs of parts are regular and then using a counting (or embedding) lemma to find copies of a particular subgraph in this regular partition. For dense graphs, these steps are generally well-behaved and essentially completely understood. For sparse graphs, however, both steps can break down without additional hypotheses. Here we will focus on the second step of finding appropriate counting lemmas in the sparse regime, since the regularity step is now reasonably well understood [14, 22] (although difficulties in maintaining the so-called no-dense-spots condition can arise even here).

Similar issues arise in the study of *quasirandom graphs*, a fundamental theme developed and popularized by Chung, Graham and Wilson [4], building on earlier work of Thomason [23]. In their work, they showed, somewhat surprisingly, that several distinct notions of quasirandomness in dense graphs are essentially equivalent. In particular, in an n-vertex graph G with edge density p, where p is a fixed constant, having C_4-density $p^4 + o(1)$ is equivalent to a certain discrepancy condition and this in turn implies that the F-density in G is $p^{|E(F)|} + o(1)$ for all fixed graphs F. However, as already observed by Chung and Graham in [5], these equivalences do not automatically carry over to graphs with $o(n^2)$ edges without additional assumptions. Indeed, even rather modest variants of the Chung–Graham–Wilson equivalences can fail to hold [20]. One viewpoint on our work here is that some aspect of the Chung–Graham–Wilson equivalences may be recovered if we assume that our graph is C_4-free.

Previous work on developing counting lemmas for sparse graphs has largely focused on controlling relatively dense subgraphs of sparse random or pseudorandom graphs. For instance, a counting lemma in sparse random graphs was proved by Conlon, Gowers, Samotij, and Schacht [6] in connection with the celebrated KŁR conjecture [15] (see also [2, 21]), while a counting lemma in sparse pseudorandom graphs was proved by Conlon, Fox, and Zhao [8] and later extended to hypergraphs [10], allowing them to simplify the proof of the Green–Tao theorem [13] (see also [9] for a detailed exposition incorporating many further simplifications of the original proof).

In recent work [7], motivated by applications in extremal and additive combinatorics, we pursued the study of sparse regularity in a very different setting, without any explicit pseudorandomness hypothesis. Instead, the only hypothesis on the host graph was that it be C_4-free. Under this

Conlon is supported by NSF Award DMS-2054452.

Fox is supported by a Packard Fellowship and by NSF Award DMS-1855635.

Sudakov is supported in part by SNSF grant 200021_196965.

Zhao is supported by NSF Award DMS-1764176, the MIT Solomon Buchsbaum Fund, and a Sloan Research Fellowship.

assumption, we proved a C_5-counting lemma, which, when combined with an appropriate sparse regularity lemma, led to various new results, including a C_5-removal lemma in C_4-free graphs. As an example of an additive combinatorics application, we showed that every Sidon subset of $[N]$ without nontrivial solutions to $w + x + y + z = 4u$ has at most $o(\sqrt{N})$ elements. Here a *Sidon set* is a set without nontrivial solutions to the equation $x + y = z + w$ and it is known that the maximum size of a Sidon subset of $[N]$ is $(1 + o(1))\sqrt{N}$. We refer the interested reader to [7] for further discussion of applications.

In this article, we continue the study of counting lemmas in C_4-free graphs, our main interest being the problem of determining which graphs F, besides C_5, satisfy an F-counting lemma in C_4-free graphs. We will make this question more precise in Definition 1.3 below, when we say formally what it means for a graph F to be *countable*.

Question 1.1 (Main question, informal)**.** For which graphs F is there an F-counting lemma in C_4-free graphs?

By extending the proof of [7, Theorem 1.1, see Section 4] (which was written for $F = C_5$, but easily extends), we can deduce an F-removal lemma in C_4-free graphs whenever F is countable.

Corollary 1.2 (Sparse removal lemma in C_4-free graphs)**.** *For any countable graph F and any $\epsilon > 0$, there exists $\delta = \delta(F, \epsilon) > 0$ such that every n-vertex C_4-free graph with at most $\delta n^{|V(F)|-|E(F)|/2}$ copies of F can be made F-homomorphism-free by removing at most $\epsilon n^{3/2}$ edges.*

Here "copies of F" refer to subgraphs isomorphic to F, whereas "F-homomorphism-free" means that there is no graph homomorphism from F into the resulting graph after edge removal. In particular, if F is bipartite and the number of copies of F in a C_4-free graph on n vertices is $o(n^{|V(F)|-|E(F)|/2})$, then G has $o(n^{3/2})$ edges.

Let us sketch the main ideas of the proof of Corollary 1.2, referring the reader to [7, Section 4] for further details. We first apply a sparse weak regularity lemma to approximate the C_4-free graph G by some "dense" graph H (allowing edge-weights in $[0, 1]$ for H). The counting lemma then implies that H has small F-homomorphism density. By the dense F-removal lemma, applied as a black box, one can therefore remove a collection of edges from H with small total weight so that the remaining graph contains no subgraphs to which F is homomorphic. Removing the corresponding edges from G then makes it F-homomorphism-free.

The notion of having an F-counting lemma is made precise in the following definition. Note that the conclusion we seek is one-sided, that is, we only ask for a lower bound. In practice, this is usually all that is needed in applications.

Definition 1.3. A graph F is *countable* if, for every $\epsilon > 0$, there exists $\delta = \delta(F, \epsilon) > 0$ such that if G is an n-vertex C_4-free graph on vertex set V and $H \in [0, 1]^{V \times V}$ is a symmetric matrix (i.e., an edge-weighted graph) satisfying

$$\left| \frac{e_G(A, B)}{n^{3/2}} - \frac{e_H(A, B)}{n^2} \right| \le \delta \qquad \text{for all } A, B \subseteq V, \tag{1.1}$$

(here $e_G(A, B) = \{(x, y) \in A \times B : xy \in E(G)\}$ and $e_H(A, B) = \sum_{x \in A, y \in B} H(x, y)$), then, for every $\boldsymbol{A} = (A_v)_{v \in V(F)}$ with $A_v \subseteq V$ for each $v \in V(F)$, one has

$$\frac{\hom_{\boldsymbol{A}}(F, G)}{n^{|V(F)|-|E(F)|/2}} \ge \frac{\hom_{\boldsymbol{A}}(F, H)}{n^{|V(F)|}} - \epsilon, \tag{1.2}$$

where $\hom_{\boldsymbol{A}}(F, G)$ is the number of homomorphisms from F to G where each $v \in V(F)$ is mapped to a vertex in A_v and $\hom_{\boldsymbol{A}}(F, H)$ is the weighted analogue defined by the formula

$$\hom_{\boldsymbol{A}}(F, H) := \sum_{x_v \in A_v \ \forall v \in V(F)} \prod_{uv \in E(F)} H(x_u, x_v).$$

The scaling in the denominators of the definition above is natural because the maximum number of edges in an n-vertex C_4-free graph is $(1/2 + o(1))n^{3/2}$ (see Remark 1.5 below). It may be instructive to consider what happens when G is the random graph $G(n, n^{-1/2})$ and H is the all-1 matrix, in which case, provided $|E(F')| < 2|V(F')|$ for all subgraphs F' of F, (1.1) and (1.2) with $\delta, \epsilon \to 0$ hold with high probability as $n \to \infty$.

Remark 1.4. In Definition 1.3, it suffices to only consider unweighted graphs H, since we can always randomly sample a weighted graph to get an unweighted graph with similar density properties. However, in applications, H is usually the normalized edge-density matrix of some (weak) regular partition of G, so it is more intuitive to allow edge-weights for H.

Remark 1.5. The *polarity graph* [3, 11, 12] is an n-vertex C_4-free graph G with $(1/2+o(1))n^{3/2}$ edges (which is essentially best possible by the Kővári–Sós–Turán theorem [16]). In addition, it has the property that every edge lies in exactly one triangle and it satisfies the discrepancy condition (1.1) with $\delta = O(n^{-1/4})$ and H being the all-1 matrix.

More specifically, let q be a prime power and let G_0 be the graph with $q^2 + q + 1$ vertices, each corresponding to a point of the projective plane over $\mathbb{F}_q$, i.e., elements of $\mathbb{F}_q^3 \setminus \{(0,0,0)\}$ where (x, y, z) is identified with $(\lambda x, \lambda y, \lambda y)$ for every nonzero $\lambda \in \mathbb{F}_q$, with an edge between (x, y, z) and (x', y', z') if and only if $xx' + yy' + zz' = 0$. This graph has exactly $q + 1$ loops. It is also $(q + 1)$-regular and has the property that each pair of distinct vertices has exactly one common neighbor, which in particular implies that G_0 is C_4-free. The square of its adjacency matrix is thus $qI + J$ (with J being the all-1 matrix) and, hence, all of its eigenvalues, besides the top eigenvalue $q + 1$, are $\pm\sqrt{q}$. The discrepancy claim in the previous paragraph then follows from the expander mixing lemma (see, e.g., [17]). In practice, we will actually use the induced subgraph G of this graph where we remove all vertices with loops. This inherits the discrepancy property from G_0, but has the additional property mentioned above that every edge is contained in a unique triangle (see [18] for a more detailed discussion of this point).

We now use the polarity graph to deduce a simple necessary condition for F to be countable.

Remark 1.6. If F is countable, then it has girth at least 5.

Indeed, suppose that F contains a 4-cycle $v_1v_2v_3v_4$. Let G be an n-vertex polarity graph and H the all-1 matrix. The discrepancy property (1.1) is satisfied for $\delta = o(1)$ by the previous remark. Set A_{v_1}, A_{v_2}, A_{v_3}, A_{v_4} to be disjoint vertex sets of $V(G)$, each of order $\lfloor n/4 \rfloor$, and $A_v = V(G)$ for all $v \in V(F) \setminus \{v_1, v_2, v_3, v_4\}$. Then $\hom_{\boldsymbol{A}}(F, G) = 0$ since G is C_4-free, but $\hom_{\boldsymbol{A}}(F, H) \gtrsim n^{|V(F)|}$, so F is not countable.

Now suppose that F contains a triangle. Consider the graph G' obtained from the polarity graph G by deleting one edge from each triangle of G chosen uniformly and independently at random (recall that G is a disjoint union of triangles). With probability $1 - o(1)$, the discrepancy property (1.1) remains valid with $\delta = o(1)$ and H the all-2/3 matrix. However, (1.2) fails when $A_v = V(G)$ for all v, since the fact that G' is triangle-free implies that $\hom(F, G') = 0$. So again F is not countable. (The same construction also appears in [1, Lemma 2.6].)

In the next section, we describe our main result, which gives a sufficient condition for countability, presented as a recursive construction.

2. Countable graphs

We begin with a simple proposition, whose proof may be found in Section 5.

Proposition 2.1. *Adding a pendant edge to a countable graph produces a countable graph.*

In particular, we have the following important corollary.

Corollary 2.2. *All trees are countable.*

It will be shown in the next section that it suffices to verify countability within n-vertex C_4-free graphs G with maximum degree at most $2\sqrt{n}$. This makes the following definition relevant.

Definition 2.3. A graph F is *tame* if there exists a constant $C = C(F)$ such that $\hom(F, G) \le Cn^{|V(F)|-|E(F)|/2}$ for every n-vertex C_4-free graph G with maximum degree at most $2\sqrt{n}$.

An edgeless graph is clearly tame. Here is a sufficient recursive condition for tameness.

Proposition 2.4. *Let F be a tame graph. Let F' be obtained from F by either*

(a) adding a pendant edge to F (creating a single new leaf vertex) or

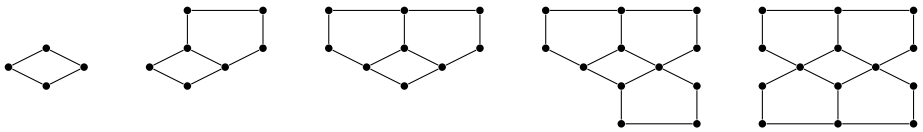
(b) joining two (not necessarily distinct) vertices of F by a 3-edge path whose two intermediate vertices are new. (If the two vertices of F are the same, then the path is a triangle.)

Then F' is tame.

Proof. Let G be an n-vertex C_4-free graph with maximum degree at most $2\sqrt{n}$. It suffices to show that $\hom(F', G) \le 4\sqrt{n}\hom(F, G)$. In case (a), this is clear, since G has maximum degree at most $2\sqrt{n}$. In case (b), we verify that the number of 3-edge walks between any pair of vertices (not necessarily distinct) in G is at most $4\sqrt{n}$. Indeed, given $x, y \in V(G)$, let w be a neighbor of x. If $w \ne y$ (at most $2\sqrt{n}$ such w), then, since G is C_4-free, there is at most one 2-edge walk from w to y. On the other hand, if $w = y$ (at most one such w), the number of 2-edge walks from $w = y$ back to itself is $\deg(y) \le 2\sqrt{n}$. □

Example 2.5. All cycles are tame, since, for each $\ell \ge 3$, one can first build an $(\ell - 3)$-edge path using (a) and then complete it to an ℓ-cycle using (b).

Example 2.6. The graphs in the sequence depicted below are also tame. To see this, observe that, at each step, we add a new path with $\ell \ge 3$ edges whose intermediate vertices are new (by again applying step (a) $\ell - 3$ times and then applying step (b) once).



Example 2.7. $K_{2,3}$ is not tame. Indeed, the n-vertex polarity graph G has $\hom(K_{2,3}, G) \ge \hom(K_{1,3}, G) = \sum_{x \in V(G)} \deg_G(x)^3 \gtrsim n^{5/2}$, which is much larger than the Cn^2 upper bound required for tameness.

Example 2.8. Let K_k' denote the 1-subdivision of K_k. Then K_k' is tame if and only if $k \le 4$. Indeed, let G be the n-vertex polarity graph. Then, since there is a homomorphism $K_k' \to K_{1,\binom{k}{2}}$ mapping all k vertices of the original K_k to the same vertex, we have that

$$\hom(K_k', G) \ge \hom(K_{1,\binom{k}{2}}, G) \gtrsim n^{1+\binom{k}{2}/2}.$$

But $1 + \binom{k}{2}/2 > k = |V(K_k')| - |E(K_k')|/2$ for $k \ge 5$, so K_k' is not tame. On the other hand, for $k \le 3$, K_k' is tame due to Proposition 2.4, while, despite the fact that Proposition 2.4 does not apply to K_4', it is still tame, as may be verified by performing a case check based on which subsets of the original four vertices of K_4 are mapped to the same vertex.

It will follow from our results below that every K_k' is countable. Therefore, K_5' (or K_k' for any $k \ge 5$) is an example of a non-tame countable graph. Moreover, since, for H the all-1 matrix, the polarity graph G satisfies the discrepancy property (1.1) with $\delta = o(1)$, we see that K_5' does not satisfy an "upper-bound counting lemma", i.e., (1.2) with $\ge \cdots - \epsilon$ replaced by $\le \cdots + \epsilon$. That is, the K_5'-counting lemma in C_4-free graphs is truly one-sided.

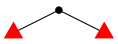
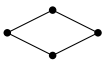
We now describe an important building block in our recursive construction of countable graphs.

Definition 2.9. Let F be a graph and $I \subseteq V(F)$ an independent set. We say that F is a *connector with ends* I (or simply that (F, I) is a connector) if

(a) F is countable and
(b) the graph $F \vee_I F$ formed by gluing two copies of F along I is tame.

Here is the simplest interesting connector.

Example 2.10. The 2-edge path $v_0v_1v_2$ is a connector with ends $\{v_0, v_2\}$. This is illustrated below, where the ends of the connector are marked by red triangles.

$F =$  $F \vee_I F =$ 

More generally, any path is a connector with ends being any independent set. However, the same statement does not extend to all trees. For instance, $K_{1,3}$ does not give rise to a triple-ended connector, since $K_{2,3}$ is not tame by Example 2.7.

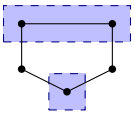
Our main result is the following recursive construction of countable graphs. It can be visualized in terms of "islands" and "bridges." We start with several disjoint tame countable components (the islands) and join them using connectors (the bridges). The theorem then says that the resulting graph is countable.

Theorem 2.11. *Let F be a graph that is an edge-disjoint union of its subgraphs $F_1, \dots, F_k, J_1, \dots, J_\ell$, satisfying all of the following conditions:*

(a) $F_1, \dots, F_k$ are countable and vertex-disjoint;
(b) $F_1, \dots, F_{k-1}$ are tame (F_k may be tame or not);
(c) for each $j \in [\ell]$, J_j is a connector with ends $I_j = V(J_j) \cap V(F_1 \cup \cdots \cup F_k)$ and I_j has at most one vertex in common with each F_i;
(d) each pair of connectors J_i and J_j share at most one vertex and the vertex they share (if any) lies in $I_i \cap I_j$.

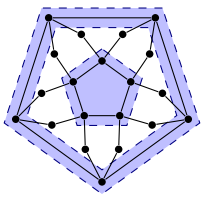
Then F is countable.

Example 2.12. The 5-cycle is countable. The "islands and bridges" decomposition is illustrated below, where each contiguous shaded region is an island. Both connectors are 2-edge-paths.

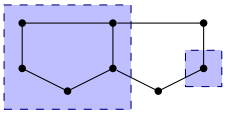


Similarly, ℓ-cycles, for $\ell \geq 5$, can be shown to be countable by starting with two islands, one an isolated vertex, as above, and the other a path of length $\ell - 4$, with 2-edge-path connectors joining the endpoints of this path to the isolated vertex. As mentioned in [7, Footnotes 1 and 3], knowing that longer cycles can be counted allows us to extend our results [7, Section 1.3] about finding solutions of translation-invariant equations in Sidon sets to equations with more than five variables.

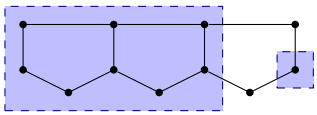
Example 2.13. Since the 5-cycle is both countable and tame, we can use it as an island to build up further countable graphs. For example, connecting a pair of 5-cycles using 2-edge-path connectors, as shown below, yields a new countable graph.



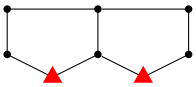
Example 2.14. Using that the 5-cycle is countable and tame, we see that the following graph is also countable, again with the islands shaded:



This graph is also tame by Proposition 2.4, so we can repeat the process to show that the following graph (and any longer chain of 5-cycles) is tame and countable.

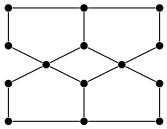


Example 2.15. The following graph is a connector (with the ends again marked by red triangles):

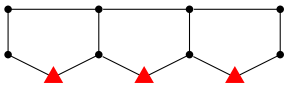


(2.1)

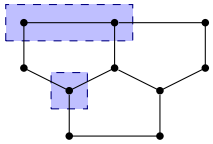
Indeed, we saw in the last example that this graph is countable, while the graph formed by gluing two copies along the ends, as shown below, is tame by Example 2.6.



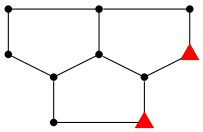
Similarly, we can check that the following graph (and any longer chain of 5-cycles) is a multi-ended connector:



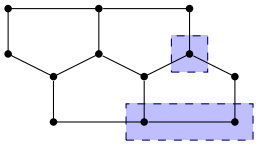
Example 2.16. The following graph is countable (one of the connectors is a 2-edge-path, while the other is (2.1)):



We can extend this example further. Since the above graph is countable, we can use Proposition 2.4 to verify that, with the ends as marked, it is also a connector:

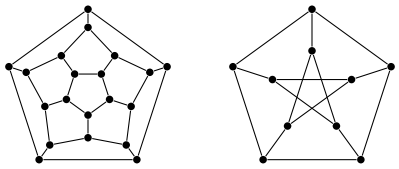


Using this connector, we deduce that the following graph is countable:



Similar inductive arguments allow us to prove the countability of many other graphs of girth at least 5. However, as we shall explain in more detail in the concluding remarks, we are far from a classification. For instance, our methods seem insufficient for showing that 3-regular graphs such as those below are countable.

Open Problem 2.17. Are the dodecahedral and Petersen graphs, shown below, countable?



In the remainder of the article, we prove Proposition 2.1 and Theorem 2.11.

3. Trimming high-degree vertices

In this brief section, we show that in the definition of countability, Definition 1.3, we can restrict to considering n-vertex C_4-free graphs G satisfying an additional maximum degree assumption, namely, that G has maximum degree at most $2\sqrt{n}$, without affecting the family of graphs which are countable.

Lemma 3.1. *Let G be a graph on a vertex set V of size n and let $H \in [0,1]^{V\times V}$ be a symmetric matrix such that*

$$\left|\frac{e_G(A,B)}{n^{3/2}} - \frac{e_H(A,B)}{n^2}\right| \le \delta \qquad \text{for all } A, B \subseteq V. \tag{3.1}$$

Let $S = \{v \in V : \deg_G(v) \le 2\sqrt{n}\}$ and let G' be the subgraph of G with the same vertex set V but only keeping edges with both endpoints in S. Then

$$\left|\frac{e_{G'}(A,B)}{n^{3/2}} - \frac{e_H(A,B)}{n^2}\right| \le 3\delta \qquad \text{for all } A, B \subseteq V.$$

Proof. Write $\overline{S} = V \setminus S$. Applying (3.1) to $(A,B) = (\overline{S}, V)$, we have

$$\delta n^2 \ge \sqrt{n} e_G(\overline{S}, V) - e_H(\overline{S}, V) \ge \sqrt{n}\cdot 2\sqrt{n}|\overline{S}| - |\overline{S}||V| = n|\overline{S}|,$$

so $|\overline{S}| \le \delta n$. For any $A, B \subseteq V$, writing $A' = A \cap S$ and $B' = B \cap S$, we have $e_{G'}(A,B) = e_G(A',B')$, so

$$\begin{aligned}\left|\sqrt{n}e_{G'}(A,B) - e_H(A,B)\right| &= \left|\sqrt{n}e_G(A',B') - e_H(A',B') + e_H(A',B') - e_H(A,B)\right| \\ &\le \left|\sqrt{n}e_G(A',B') - e_H(A',B')\right| + (|A\setminus A'| + |B\setminus B'|)n \\ &\le \delta n^2 + 2|\overline{S}|n \le 3\delta n^2.\end{aligned}$$

□

4. Notation and setup

Given a graph F, a *vertex weight function* on F (sometimes we say "on $V(F)$", as graphs and their vertex sets are interchangeable for this purpose) is a collection $\boldsymbol{\alpha} = (\alpha_v)_{v\in V(F)}$ of functions $\alpha_v \colon V \to [0,1]$ indexed by v. It will be important for our arguments that each α_v takes values in $[0,1]$ and not in some wider range.

Let $x = (x_v)_{v\in V(F)} \in V^{V(F)}$ with $x_v \in V$. For each $S \subseteq V(F)$, we write $x_S = (x_v)_{v\in S}$ for its projection onto the coordinates indexed by S. To avoid notational clutter, we will sometimes write a subgraph as the subscript rather than its vertex set. For example, if F' is a subgraph of F and $S \subseteq V(F)$, then we write $x_{F'} = x_{V(F')}$, $x_{F\setminus F'} = x_{V(F)\setminus V(F')}$, and $x_{F\setminus S} = x_{V(F)\setminus S}$.

Given a function $f \colon V^S \to \mathbb{R}$, we write

$$\int f(x_S)dx_S = |V|^{-|S|}\sum_{x_S\in V^S} f(x_S).$$

Furthermore, given a vertex weight function $\boldsymbol{\alpha} = (\alpha_v)_{v\in S}$ on S, we write

$$\int f(x_S)d^{\boldsymbol{\alpha}}x_S = \int f(x_S)\prod_{v\in S}\alpha_v(x_v)\,dx_S.$$

Given a symmetric function $g\colon V \times V \to \mathbb{R}$ and $x \in V^{V(F)}$, we define $g_F\colon V^{V(F)} \to \mathbb{R}$ by

$$g_F(x) = \prod_{uv \in F} g(x_u, x_v).$$

Given $S \subseteq V(F)$ and a vertex weight function $\boldsymbol{\alpha}$ on $F \setminus S$, we define $g_{F,S}\colon V^S \to \mathbb{R}$ by

$$g^{\boldsymbol{\alpha}}_{F,S}(x_S) = \int g_F(x_F)\, d^{\boldsymbol{\alpha}} x_{F \setminus S},$$

which (up to normalization) corresponds to counting homomorphisms $F \to G$ where the image of S is x_S and the remaining vertices of F are weighted by $\boldsymbol{\alpha}$. Such quantities also arise naturally when using flag algebras. Finally, given a vertex weight function $\boldsymbol{\alpha}$ on F, we write

$$t^{\boldsymbol{\alpha}}(F, g) = g^{\boldsymbol{\alpha}}_{F,\emptyset} = \int \prod_{uv \in F} g(x_u, x_v) \prod_{v \in V(F)} (\alpha_v(x_v) dx_v)\,,$$

which is the $\boldsymbol{\alpha}$-weighted homomorphism density of F in g.

It will also be convenient to allow our weight function notation to be a little more flexible, in the sense that we automatically ignore uninvolved vertices. For example, if $\boldsymbol{\alpha}$ is a vertex weight function on F and F' is a subgraph on a proper vertex subset, then we still write $t^{\boldsymbol{\alpha}}(F', g)$ and $d^{\boldsymbol{\alpha}} x_{F'}$ with the understanding that $\boldsymbol{\alpha}$ is now restricted to the vertex set of F'. This way we do not always have to specify the set of vertices that the weight function is defined on.

Both the discrepancy condition (1.1) and the counting lemma conclusion (1.2) can be equivalently rephrased in terms of weight functions $\boldsymbol{\alpha}$ rather than product sets $\boldsymbol{A}$. The extra flexibility allowed by considering $[0,1]$-valued weight functions will be helpful in our proofs. To see the equivalence, note that, with the function $g = \sqrt{n}G$ (here we view $G\colon V \times V \to \{0,1\}$ as the edge-indicator function of the graph G), we have

$$\frac{\hom_{\boldsymbol{A}}(F, G)}{n^{|V(F)| - |E(F)|/2}} = t^{\boldsymbol{\alpha}}(F, g)$$

for the vertex weight function $\boldsymbol{\alpha}$ on F which is equal to the indicator function of $\boldsymbol{A}$ (i.e., $\alpha_v(x) = 1$ if $x \in A_v$ and 0 otherwise). Likewise, for $h = H$,

$$\frac{\hom_{\boldsymbol{A}}(F, H)}{n^{|V(F)|}} = t^{\boldsymbol{\alpha}}(F, h).$$

Hence, the counting lemma conclusion (1.2), that

$$\frac{\hom_{\boldsymbol{A}}(F, G)}{n^{|V(F)| - |E(F)|/2}} \geq \frac{\hom_{\boldsymbol{A}}(F, H)}{n^{|V(F)|}} - \epsilon,$$

is equivalent to the statement that

$$t^{\boldsymbol{\alpha}}(F, g) \geq t^{\boldsymbol{\alpha}}(F, h) - \epsilon \tag{4.1}$$

for any $\{0,1\}$-valued vertex weight function $\boldsymbol{\alpha}$. Since $t^{\boldsymbol{\alpha}}(F, g) - t^{\boldsymbol{\alpha}}(F, h)$ is a multilinear function of the values $(\alpha_v(x))_{v \in F, x \in V}$, the extrema of the function are attained when $\alpha_v(x) \in \{0,1\}$ for all $v \in F$ and $x \in V$. This shows that the counting lemma conclusion (1.2) is equivalent to the statement that (4.1) holds for all vertex weight functions.

By the same argument, the discrepancy condition (1.1), that

$$\left| \frac{e_G(A,B)}{n^{3/2}} - \frac{e_H(A,B)}{n^2} \right| \leq \delta \qquad \text{for all } A, B \subseteq V,$$

is equivalent to

$$\left| \int (g - h)(x, y) \alpha_1(x) \alpha_2(y) dx dy \right| \leq \delta \qquad \text{for all } \alpha_1, \alpha_2\colon V \to [0,1].$$

In fact, (thanks to the trimming step in the previous section) from now on we will only need the one-sided discrepancy hypothesis

$$\int g(x,y)\alpha_1(x)\alpha_2(y)dxdy \geq \int h(x,y)\alpha_1(x)\alpha_2(y)dxdy - \delta \qquad \text{for all } \alpha_1,\alpha_2\colon V\to[0,1]. \tag{4.2}$$

Summary of what needs to be proved. To prove that F is countable, it suffices to show that there is a constant $c>0$ such that for every $\epsilon>0$ there exists $\delta>0$ satisfying the following. Let G be an n-vertex C_4-free graph on vertex set V with maximum degree at most $2\sqrt{n}$. Let $g=c\sqrt{n}G$ and let $h\colon V\times V\to[0,1]$ be a symmetric function satisfying (4.2). Then, for every vertex weight function $\boldsymbol{\alpha}$ on F, one has (4.1).

The reason that we scale by a factor of c in defining g is so that the various tameness hypotheses on subgraphs of G can be made to have the form $t(F',g)\leq 1$. Furthermore, as long as $c\leq 1/2$, the hypothesis that G has maximum degree at most $2\sqrt{n}$ implies that

$$\int g(x,y)\,dy \leq 1 \qquad \text{for all } x\in V. \tag{4.3}$$

5. Counting lemma proofs

We follow without further comment the framework discussed in the previous section.

Proof of Proposition 2.1 (adding a pendant edge preserves countability). Let F be a graph with a leaf vertex u. Let F' be F with u removed and assume that F' is countable. Suppose that

$$\int g(x,y)\alpha_1(x)\alpha_2(y)dxdy \geq \int h(x,y)\alpha_1(x)\alpha_2(y)dxdy - \epsilon \tag{5.1}$$

for all $\alpha_1,\alpha_2\colon V\to[0,1]$. Since F' is countable, we may also assume that

$$t^{\boldsymbol{\alpha}'}(F',g) \geq t^{\boldsymbol{\alpha}'}(F',h) - \epsilon \tag{5.2}$$

for every vertex weight function $\boldsymbol{\alpha}'$ on F'.

It suffices to show that these two inequalities imply that

$$t^{\boldsymbol{\alpha}}(F,g) \geq t^{\boldsymbol{\alpha}}(F,h) - 2\epsilon \tag{5.3}$$

for every vertex weight function $\boldsymbol{\alpha}$ on F. For this, define a vertex weight function $\boldsymbol{\alpha}'$ on F' by $\alpha'_v=\alpha_v$ unless v is the neighbor v of u, in which case $\alpha'_v(x_v)=\alpha_v(x_v)\int g(x_v,x_u)\alpha_u(x_u)dx_u\in[0,1]$ by (4.3). Then, by (5.2) applied with this $\boldsymbol{\alpha}'$,

$$t^{\boldsymbol{\alpha}}(F,g) = t^{\boldsymbol{\alpha}'}(F',g) \geq t^{\boldsymbol{\alpha}'}(F',h) - \epsilon.$$

Furthermore, we have

$$\begin{aligned} t^{\boldsymbol{\alpha}'}(F',h) &= \int h^{\boldsymbol{\alpha}}_{F',v}(x_v)g(x_v,x_u)\alpha_u(x_u)\alpha_v(x_v)\,dx_u dx_v \\ &\geq \int h^{\boldsymbol{\alpha}}_{F',v}(x_v)h(x_v,x_u)\alpha_u(x_u)\alpha_v(x_v)\,dx_u dx_v - \epsilon \\ &= t^{\boldsymbol{\alpha}}(F,h) - \epsilon, \end{aligned}$$

where the inequality step uses (5.1). Combining the last two displayed inequalities yields (5.3), as desired. □

Proof of Theorem 2.11 (islands and bridges). By the tameness assumptions, we can choose a sufficiently small constant $c\in(0,1]$ (depending only on F) such that, setting $g=c\sqrt{n}G\colon V\times V\to[0,\infty)$, we have

$$t(F_i,g)\leq 1 \text{ for all } i\in[k-1] \quad \text{and} \quad t(J_j\vee_{I_j} J_j,g)\leq 1 \text{ for all } j\in[\ell]. \tag{5.4}$$

Let $\epsilon \in (0,1]$ and let

$$\eta_i = \epsilon^{2^i} \text{ for each } i \in [\ell] \quad \text{and} \quad \eta = \epsilon^{2^{\ell+1}}. \tag{5.5}$$

By the countability assumption on $F_1, \dots, F_k, J_1, \dots, J_\ell$ it suffices to show that if $h\colon V \times V \to [0,1]$ satisfies

$$t^{\boldsymbol{\alpha}}(L,g) \geq t^{\boldsymbol{\alpha}}(L,h) - \eta \tag{5.6}$$

for each $L \in \{F_1, \dots, F_k, J_1, \dots, J_\ell\}$ and vertex weight function $\boldsymbol{\alpha}$ on L, then

$$t^{\boldsymbol{\alpha}}(F,g) \geq t^{\boldsymbol{\alpha}}(F,h) - (2\ell + k)\epsilon, \tag{5.7}$$

for every vertex weight function $\boldsymbol{\alpha}$ on F.

Write

$$f_{\leq t}(x) = \begin{cases} f(x) & \text{if } f(x) \leq t, \\ 0 & \text{otherwise} \end{cases} \quad \text{and} \quad f_{>t}(x) = \begin{cases} f(x) & \text{if } f(x) > t, \\ 0 & \text{otherwise.} \end{cases}$$

For each connector $(J,I) = (J_j, I_j)$ (temporarily dropping the subscript j to avoid notational clutter), writing

$$g^{\boldsymbol{\alpha}}_{J,I,>\delta^{-1}} = (g^{\boldsymbol{\alpha}}_{J,I})_{>\delta^{-1}},$$

we have, using $t(J \vee_I J, g) \leq 1$ from (5.4), that

$$\int g^{\boldsymbol{\alpha}}_{J,I,>\delta^{-1}}(x_I)\, d^{\boldsymbol{\alpha}} x_I \leq \delta \int g^2_{J,I}(x_I)\, d^{\boldsymbol{\alpha}} x_I \leq \delta t(J \vee_I J, g) \leq \delta.$$

Thus, using (5.6),

$$\int g^{\boldsymbol{\alpha}}_{J,I,\leq\delta^{-1}}(x_I)\, d^{\boldsymbol{\alpha}} x_I \geq \left(\int g^{\boldsymbol{\alpha}}_{J,I}(x_I)\, d^{\boldsymbol{\alpha}} x_I\right) - \delta \geq \left(\int h^{\boldsymbol{\alpha}}_{J,I}(x_I)\, d^{\boldsymbol{\alpha}} x_I\right) - \eta - \delta. \tag{5.8}$$

Step I. Swapping out the islands one at a time.

Write $F' = \cup_i F_i$ (islands without connectors). We have

$$\begin{aligned} t^{\boldsymbol{\alpha}}(F,g) &= \int g_F(x_F)\, d^{\boldsymbol{\alpha}} x_F \\ &= \int \prod_{i=1}^k g_{F_i}(x_{F_i}) \prod_{j=1}^\ell g^{\boldsymbol{\alpha}}_{J_j,I_j}(x_{I_j})\, d^{\boldsymbol{\alpha}} x_{F'} \\ &\geq \int \prod_{i=1}^k g_{F_i}(x_{F_i}) \prod_{j=1}^\ell g^{\boldsymbol{\alpha}}_{J_j,I_j,\leq\eta_j^{-1}}(x_{I_j})\, d^{\boldsymbol{\alpha}} x_{F'} \\ &= \int \left(\int g_{F_k}(x_{F_k}) \prod_{j=1}^\ell g^{\boldsymbol{\alpha}}_{J_j,I_j,\leq\eta_j^{-1}}(x_{I_j}) d^{\boldsymbol{\alpha}} x_{F_k}\right) \prod_{i=1}^{k-1} \left(g_{F_i}(x_{F_i}) d^{\boldsymbol{\alpha}} x_{F_i}\right). \end{aligned}$$

Now, using (5.6) for F_k and noting that the inner integral inside the parenthesis has the form $\int g_{F_k}(x_{F_k}) d^{\boldsymbol{\alpha}'} x_{F_k} \cdot \prod_{j=1}^\ell \eta_j^{-1}$ for some other vertex weight function $\boldsymbol{\alpha}'$ (absorbing the connector factors by using the fact that each connector uses at most one vertex from the island F_k), we have, continuing from above, that the last expression is

$$\geq \int \left(\int h_{F_k}(x_{F_k}) \prod_{j=1}^\ell g^{\boldsymbol{\alpha}}_{J_j,I_j,\leq\eta_j^{-1}}(x_{I_j}) d^{\boldsymbol{\alpha}} x_{F_k} - \eta \prod_{j=1}^\ell \eta_j^{-1}\right) \prod_{i=1}^{k-1} \left(g_{F_i}(x_{F_i}) d^{\boldsymbol{\alpha}} x_{F_i}\right).$$

Since $\eta\prod_{j=1}^{\ell}\eta_j^{-1}\le\epsilon$ by (5.5) and $\int g_{F_i}(x_{F_i})d^{\boldsymbol{\alpha}}x_{F_i}\le t(F_i,g)\le 1$ for each $i\in[k-1]$ by (5.4), we can continue the above as

$$\ge\int h_{F_k}(x_{F_k})\prod_{i=1}^{k-1}g_{F_i}(x_{F_i})\prod_{j=1}^{\ell}g^{\boldsymbol{\alpha}}_{J_j,I_j,\le\eta_j^{-1}}(x_{I_j})\,d^{\boldsymbol{\alpha}}x_{F'}-\epsilon.$$

We can now repeat this process to successively replace each remaining g_{F_i} factor by h_{F_i}, losing at most an additive error of ϵ at each step. (Note that even though we do not assume that $t(F_k,g)\le 1$, it is no longer needed, since what matters from now on is that $t(F_k,h)\le 1$ and this is automatically true for h, which takes values in $[0,1]$). We may therefore continue the above as

$$\ge\int\prod_{i=1}^{k}h_{F_i}(x_{F_i})\prod_{j=1}^{\ell}g^{\boldsymbol{\alpha}}_{J_j,I_j,\le\eta_j^{-1}}(x_{I_j})\,d^{\boldsymbol{\alpha}}x_{F'}-k\epsilon.$$

Step II. Swapping out the connectors one at a time.

Continuing, we have, applying (5.8) to replace $g^{\boldsymbol{\alpha}}_{J_\ell,I_\ell,\le\eta_\ell^{-1}}(x_{I_\ell})$ by $h^{\boldsymbol{\alpha}}_{J_\ell,I_\ell}(x_{J_\ell})$ (here we are applying (5.8) for each fixed $x_{F\setminus J_\ell}$ and with a different $\boldsymbol{\alpha}$ which absorbs additional factors; this step works only because each J_ℓ intersects each of $F_1,\dots,F_k$, $J_1,\dots,J_{\ell-1}$ in at most one vertex and all these intersections are contained in I_ℓ), that the last expression above is

$$\ge\int\prod_{i=1}^{k}h_{F_i}(x_{F_i})\cdot h^{\boldsymbol{\alpha}}_{J_\ell,I_\ell}(x_{I_\ell})\prod_{j=1}^{\ell-1}g^{\boldsymbol{\alpha}}_{J_j,I_j,\le\eta_j^{-1}}(x_{I_j})\,d^{\boldsymbol{\alpha}}x_{F'}-(\eta+\eta_\ell)\prod_{j=1}^{\ell-1}\eta_j^{-1}-k\epsilon.$$

We have $(\eta+\eta_\ell)\prod_{j=1}^{\ell-1}\eta_j^{-1}\le 2\epsilon$ by (5.5). Continuing, we can replace $g^{\boldsymbol{\alpha}}_{J_j,I_j,\le\eta_j^{-1}}(x_{I_j})$ by $h^{\boldsymbol{\alpha}}_{J_j,I_j}(x_{I_j})$ one at a time in *decreasing order of* j, so that the additive error at j is at most $(\eta+\eta_j)\eta_1^{-1}\cdots\eta_{j-1}^{-1}\le 2\epsilon$ (this is why we need $\eta_1,\dots,\eta_\ell$ to be rapidly decreasing). Finally, we can continue the above as

$$\begin{aligned}&\ge\int\prod_{i=1}^{k}h_{F_i}(x_{F_i})\prod_{j=1}^{\ell}h^{\boldsymbol{\alpha}}_{J_j,I_j}(x_{I_j})\,d^{\boldsymbol{\alpha}}x_{F'}-(k+2\ell)\epsilon\\&=t^{\boldsymbol{\alpha}}(F,h)-(k+2\ell)\epsilon,\end{aligned}$$

thereby proving (5.7). □

6. Concluding remarks

We conclude by exploring some of the problems that arose from our study of countability.

Classifying countable graphs. We have made partial progress on our Question 1.1 by producing a family of graphs F for which there is an F-counting lemma in C_4-free graphs. However, our results are likely far from a complete classification. We saw one necessary condition on any such F in Remark 1.6, namely, that F should have girth at least 5. It also seems necessary that the 2-density of F should be less than 2, that is, that any subgraph F' of F should satisfy $|E(F')|\le 2|V(F')|-4$. In particular, this would imply that any d-regular countable graph has $d\le 3$.

Though not a formal proof, the intuition here is that the number of copies of F' in our C_4-free graph should not be smaller than the number of edges (otherwise, we can delete all copies of F', and hence F, by removing an edge from each copy) and, for a random graph of the same density $n^{-1/2}$, the condition that the 2-density be less than 2 is necessary for this to hold. Most likely, the true conditions for countability are even more stringent than this argument suggests. Perhaps resolving the cases highlighted in Open Problem 2.17 would be a good starting point for further progress.

We remark in passing that we expect any progress on Question 1.1 to also impinge on the closely related question where we assume that there are $o(n^2)$ copies of C_4 in our n-vertex graph rather than none. Indeed, the arguments in [7] showing that C_5 is countable apply in this more general situation and the proofs here may also be adapted to this context. We suspect that the same will be true of any countable graph.

Variations on countability. There are several variants of our basic question which may be interesting. For instance, for which graphs F is there a two-sided counting lemma in C_4-free graphs? Our results are fundamentally one-sided, so new ideas are probably necessary to make progress on this question. However, we do know that for F to satisfy a two-sided counting lemma, it must, at the very least, be tame. As observed in Example 2.8, this already rules out two-sided counting for the family of subdivisions K_t' with $t \geq 5$.

Another natural variant is to ask which graphs F have an F-counting lemma in H-free graphs when H is a bipartite graph other than C_4? Our arguments apply just as well to $K_{2,t}$-free graphs as they do to C_4-free graphs, but further extensions are less obvious. We do expect our methods to extend to prove counting lemmas in C_{2k}-free graphs for any $k \geq 3$, but here the real difficulty passes back to the regularity side. Indeed, in order to apply a C_{2k+1}-counting lemma in C_{2k}-free graphs to prove a corresponding removal lemma, we also need to show that any regular partition of a C_{2k}-free graph has few edges between irregular pairs. However, we do not at present know how to do this for any $k \geq 3$. As in [7], resolving this issue would have several consequences. To give just one example, it would allow us to show that any 3-uniform hypergraph with n vertices and girth greater than $2k+1$ has $o(n^{1+1/k})$ edges, extending both the classic Ruza–Szemerédi theorem [19], which is equivalent to the case $k = 1$, and a recent result of the authors [7, Corollary 1.10] resolving the case $k = 2$.

Acknowledgments

Part of this work was completed in the summer of 2019 while Yufei Zhao was generously hosted by FIM (the Institute for Mathematical Research) during a visit to Benny Sudakov at ETH Zürich.

References

[1] Noga Alon, Béla Bollobás, Michael Krivelevich, and Benny Sudakov, *Maximum cuts and judicious partitions in graphs without short cycles*, J. Combin. Theory Ser. B **88** (2003), 329–346.
[2] József Balogh, Robert Morris, and Wojciech Samotij, *Independent sets in hypergraphs*, J. Amer. Math. Soc. **28** (2015), 669–709.
[3] W. G. Brown, *On graphs that do not contain a Thomsen graph*, Canad. Math. Bull. **9** (1966), 281–285.
[4] F. R. K. Chung, R. L. Graham, and R. M. Wilson, *Quasi-random graphs*, Combinatorica **9** (1989), 345–362.
[5] Fan Chung and Ronald Graham, *Sparse quasi-random graphs*, Combinatorica **22** (2002), 217–244.
[6] D. Conlon, W. T. Gowers, W. Samotij, and M. Schacht, *On the KLR conjecture in random graphs*, Israel J. Math. **203** (2014), 535–580.
[7] David Conlon, Jacob Fox, Benny Sudakov, and Yufei Zhao, *The regularity method for graphs with few 4-cycles*, J. Lond. Math. Soc., to appear.
[8] David Conlon, Jacob Fox, and Yufei Zhao, *Extremal results in sparse pseudorandom graphs*, Adv. Math. **256** (2014), 206–290.
[9] David Conlon, Jacob Fox, and Yufei Zhao, *The Green-Tao theorem: an exposition*, EMS Surv. Math. Sci. **1** (2014), 249–282.
[10] David Conlon, Jacob Fox, and Yufei Zhao, *A relative Szemerédi theorem*, Geom. Funct. Anal. **25** (2015), 733–762.
[11] P. Erdős and A. Rényi, *On a problem in the theory of graphs*, Magyar Tud. Akad. Mat. Kutató Int. Közl. **7** (1962), 623–641.
[12] P. Erdős, A. Rényi, and V. T. Sós, *On a problem of graph theory*, Studia Sci. Math. Hungar. **1** (1966), 215–235.
[13] Ben Green and Terence Tao, *The primes contain arbitrarily long arithmetic progressions*, Ann. of Math. (2) **167** (2008), 481–547.
[14] Y. Kohayakawa, *Szemerédi's regularity lemma for sparse graphs*, Foundations of computational mathematics (Rio de Janeiro, 1997), Springer, Berlin, 1997, pp. 216–230.

[15] Y. Kohayakawa, T. Łuczak, and V. Rödl, *On K^4-free subgraphs of random graphs*, Combinatorica **17** (1997), 173–213.
[16] T. Kövari, V. T. Sós, and P. Turán, *On a problem of K. Zarankiewicz*, Colloq. Math. **3** (1954), 50–57.
[17] M. Krivelevich and B. Sudakov, *Pseudo-random graphs*, More sets, graphs and numbers, Bolyai Soc. Math. Stud., vol. 15, Springer, Berlin, 2006, pp. 199–262.
[18] Felix Lazebnik and Jacques Verstraëte, *On hypergraphs of girth five*, Electron. J. Combin. **10** (2003), Research Paper 25, 15 pp.
[19] I. Z. Ruzsa and E. Szemerédi, *Triple systems with no six points carrying three triangles*, Combinatorics (Proc. Fifth Hungarian Colloq., Keszthely, 1976), Vol. II, Colloq. Math. Soc. János Bolyai, vol. 18, North-Holland, Amsterdam-New York, 1978, pp. 939–945.
[20] Ashwin Sah, Mehtaab Sawhney, Jonathan Tidor, and Yufei Zhao, *A counterexample to the Bollobás-Riordan conjectures on sparse graph limits*, Combin. Probab. Comput., to appear.
[21] David Saxton and Andrew Thomason, *Hypergraph containers*, Invent. Math. **201** (2015), 925–992.
[22] Alexander Scott, *Szemerédi's regularity lemma for matrices and sparse graphs*, Combin. Probab. Comput. **20** (2011), 455–466.
[23] Andrew Thomason, *Pseudorandom graphs*, Random graphs '85 (Poznań, 1985), North-Holland Math. Stud., vol. 144, North-Holland, Amsterdam, 1987, pp. 307–331.

Conlon, Department of Mathematics, California Institute of Technology, Pasadena, CA, USA
Email address: dconlon@caltech.edu

Fox, Department of Mathematics, Stanford University, Stanford, CA, USA
Email address: jacobfox@stanford.edu

Sudakov, Department of Mathematics, ETH, Zürich, 8092, Switzerland
Email address: benjamin.sudakov@math.ethz.ch

Zhao, Department of Mathematics, Massachusetts Institute of Technology, Cambridge, MA, USA
Email address: yufeiz@mit.edu