1 Introduction

One of the most important results in extremal graph theory is the Kővári–Sós–Turán theorem. There are a number of different ways of stating it; the following way will be the most useful for us in the future.

Theorem 1 (Kővári–Sós–Turán)**.** *Let G be an N-vertex graph with at least γN^2 edges. Then G contains a copy of $K_{k,k}$, where*

$$k \geqslant c\frac{\log N}{\log \frac{1}{\gamma}}$$

and $c > 0$ is an absolute constant.

Moreover, this is tight up to the value of c: a random N-vertex graph with γN^2 edges contains, with high probability, no copy of $K_{k,k}$ for $k = C\frac{\log N}{\log \frac{1}{\gamma}}$, for some absolute constant $C > 0$.

One of the remarkable things about the Kővári–Sós–Turán theorem is that holds for all γ and N. In particular, it is meaningful even when γ is a negative power of N, or equivalently when k is fixed. For example, if we plug in $\gamma = N^{-c/k}$, we find that $K_{k,k} \subseteq G$ whenever G is an N-vertex graph with at least $N^{2-c/k}$ edges. In this formulation, we see that it is of great interest to close the constant-factor gap in Theorem 1, since doing so would determine the correct exponent for the extremal number of $K_{k,k}$.

Another natural way of viewing the Kővári–Sós–Turán theorem is through the lens of graph blowups. Given a graph H and an integer k, the *k-blowup* $H[k]$ is obtained from H by replacing every vertex by an independent set of order k, and every edge by a copy of $K_{k,k}$. In particular, $K_{k,k}$ is simply the k-blowup of K_2. Thus, the Kővári–Sós–Turán theorem states that if G contains at least γN^2 copies of K_2, then it contains a copy of $K_2[k]$ for a "large" value of k, namely $k = \Omega(\frac{\log N}{\log \frac{1}{\gamma}})$.

Therefore, it is natural to ask what happens if we assume that G contains many copies of H; in particular, can we guarantee a large blowup of H? The first result in this direction was due to Erdős, who proved a hypergraph analogue of the Kővári–Sós–Turán theorem, which, in particular, implies the following.

Theorem 2 (Erdős)**.** *Let H be an h-vertex graph. Let G be an N-vertex graph with at least γN^h copies of H. Then G contains a copy of $H[k]$, where*

$$k \geqslant c_H \frac{(\log N)^{\frac{1}{h-1}}}{\log \frac{1}{\gamma}}.$$

For example, for $H = K_3$, this states that a graph with γN^3 triangles contains a complete tripartite graph $K_{k,k,k} = K_3[k]$ for $k = \Omega_\gamma(\sqrt{\log N})$. While the underlying result about hypergraphs is again tight up to a constant factor (as witnessed by a random hypergraph), Theorem 2 is *not* tight. Indeed, a random graph with edge density $\gamma^{1/e(H)}$ has $\asymp \gamma N^h$ copies of H, but still has a blowup $H[k]$ with $k = \Theta_\gamma(\log N)$.

Since the random construction was tight for Theorem 1, it is natural to expect that Theorem 2 can be strengthened to yield a logarithmic bound. And indeed, this was proved in groundbreaking work of Nikiforov.

Theorem 3 (Nikiforov)**.** *Fix $\gamma > 0$ and an h-vertex graph H. If N is sufficiently large, and G is an N-vertex graph with at least γN^h copies of H, then G contains a copy of $H[k]$, where*

$$k \geqslant c_H \gamma^h \log N,$$

and c_H is a constant depending only on H.

Thus, Nikiforov's theorem states that if G contains $\Omega(N^h)$ copies of H, then it contains an H-blowup of size $k = \Theta(\log N)$. This dependence on N *is* tight, as witnessed by a random graph. In the case $H = K_3$, we find a complete tripartite graph of size $\Theta(\log N)$, rather than the $\Omega(\sqrt{\log N})$ size guaranteed by Theorem 2. This result is one of a number of fundamental theorems—also including the Ruzsa–Szemerédi $(6, 3)$ theorem—which demonstrate that the structure of triangles in a graph is rather different from the structure of a generic 3-uniform hypergraph.

Nonetheless, there are two major downsides to Theorem 3. The first is that the quantitative dependence on γ is not necessarily optimal. A random graph with γN^h copies of H has a copy of $H[k]$ only for $k = C_H \frac{\log N}{\log \frac{1}{\gamma}}$, which, as a function of γ, is much larger than the $c_H \gamma^h \log N$ guaranteed by Theorem 3. The second downside of Theorem 3 is the assumption that N is sufficiently large with respect to γ. As discussed above, Theorem 1 holds for all γ and N, and thus yields informative results even when we are looking for a constant-sized blowup. Thus, for example, Theorem 1 yields information on how many edges in a graph guarantee a copy of $C_4 = K_2[2]$, but Theorem 3 gives no information on how many triangles in a graph guarantee a copy of the octahedron graph, $K_3[2]$.

Addressing the first downside, there have been a number of quantitative improvements to Theorem 3. Rödl and Schacht proved that, in Theorem 3, we can take $k \geqslant \gamma^{1+o(1)} \log N$, where the $o(1)$ tends to 0 as $\gamma \to 0$, thus improving the exponent from h to a constant independent of H. Later, Fox, Luo, and I improved this result and showed that we can take $k \geqslant \gamma^{1-1/e(H)+o(1)} \log N$, thus obtaining an exponent strictly better than 1 for any fixed H. However, all of these results still yield a lower bound that is polynomial in γ, whereas the best known upper bound, coming from random graphs, is logarithmic in γ. Additionally, all of these results share the second downside of Theorem 3, namely the require N to be sufficiently large in terms of γ.

However, given our experience with Theorem 1, it is natural to expect that the random graph is essentially optimal, and thus that one can overcome both downsides. Formally, the following conjecture seems natural.

Conjecture 4. *Let H be an h-vertex graph. If G is an N-vertex graph with at least γN^h copies of H, then G contains a copy of $H[k]$, where*

$$k \geqslant c_H \frac{\log N}{\log \frac{1}{\gamma}},$$

and $c_H > 0$ is a constant depending only on H.

Again, Conjecture 4 would be tight up to the value of c_H if it's true, as witnessed by a random graph. Additionally, as it makes no assumption on the relation between N and γ, it would yield results in the constant k regime, as well as in intermediate regimes. This would be very interesting, since, for example, Rödl and Schacht proved that knowing Conjecture 4 for $H = K_3$ and γ on the order of $e^{-\sqrt{\log N}}$ would have implications for such questions in hypergraphs.

As far as I know, Conjecture 4 has never been formally written down, but its statement has been floating around for a long time. I am also not sure whether experts really believe it's true: several people have told me that they expect that for $H = K_3$, there should be a construction doing much better than random.

Theorem 1 implies that Conjecture 4 is true for all bipartite H, which is a somewhat degenerate case of the problem. The main theorem I want to discuss today confirms Conjecture 4 for a large family of H, and in particular is the first instance where Conjecture 4 is known for *any* non-bipartite H.

Theorem 5 (Girão–Hunter–W.)**.** *Let H be a triangle-free h-vertex graph. If G is an N-vertex graph with at least γN^h copies of H, then G contains a copy of $H[k]$, where*

$$k \geqslant c_H \frac{\log N}{\log \frac{1}{\gamma}},$$

and $c_H > 0$ is a constant depending only on H.

2 An application to Ramsey theory

Before discussing some ideas of the proof of Theorem 5, let me mention one cool application it has in Ramsey theory. Recall that given a graph H and an integer q, the *Ramsey number* $r(H; q)$ is defined to be the least N such that every q-coloring of $E(K_N)$ contains a monochromatic copy of H. A basic result of Chung and Graham states that

$$q^{ck} \leqslant r(K_{k,k}; q) \leqslant q^{Ck} \tag{1}$$

where $C > c > 0$ are absolute constants. That is, $r(K_{k,k}; q)$ grows polynomially in q, and exponentially in k.

However, if H is not bipartite, then it is easy to see that such polynomial behavior in q is impossible. Indeed, the standard "hypercube coloring" shows that $r(H; q) > 2^q$ for every non-bipartite H. In fact, for complete graphs, it is known that

$$2^{cqk} \leqslant r(K_k; q) \leqslant q^{qk},$$

i.e. the growth of $r(K_k; q)$ is again exponential in k, and between exponential and (barely) super-exponential in q.

However, one way of viewing the Chung–Graham result (1) is as stating that $r(H;q)$ grows polynomially in q if H is a blowup of the fixed graph K_2. Another result we proved is that something similar happens for blowups of any fixed graph, so long as Conjecture 4 holds.

Proposition 6 (Girão–Hunter–W.)**.** *Let H be a graph, and suppose that Conjecture 4 holds for H. If q is fixed and k is sufficiently large, then*

$$q^{c_H k} \leqslant r(H[k];q) \leqslant q^{C_H k},$$

where $C_H > c_H > 0$ are constants depending only on H.

Since we proved that Conjecture 4 holds for any triangle-free H, this implies that Ramsey numbers of large blowups of fixed triangle-free graphs exhibit polynomial dependence on q. Again, we stress that the assumption that k is sufficiently large with respect to q is necessary, since otherwise the bound $r(H;q) > 2^q$, holding for non-bipartite H, shows that a polynomial dependence on q is impossible.

Let me remark that Proposition 6 is not very hard to prove, but that the first thing you might try (or at least that I might try) does not work. Indeed, given that our goal is to use Conjecture 4, it is natural to start with a q-coloring of $E(K_N)$, and then to identify a color with many monochromatic copies of H. If we can do this, we can apply Conjecture 4 to the graph of edges in this color, and find a large monochromatic blowup of H. Indeed, using standard techniques of *Ramsey multiplicity*, one can find a color with many monochromatic H. However, the bounds given by this approach are too weak, and even the best possible bounds one could hope for are still too weak.

To circumvent this issue, we first "zoom in" to a small portion of the coloring, say to a specially chosen set of $\sqrt{N}$ vertices. By choosing this set appropriately, we can ensure that it contains *very* many monochromatic H—so many that we can plug into Conjecture 4. The key point is that, since Conjecture 4 yields bounds that are logarithmic in N, dropping the number of vertices from N to $\sqrt{N}$ involves essentially no loss in the bounds. The difficulty is *finding* which set of vertices to zoom into, and there are a few ways of doing this; it can be done using Szemerédi's regularity lemma, but we take an alternative density-increment approach in order to optimize some quantitative aspects of the proof.

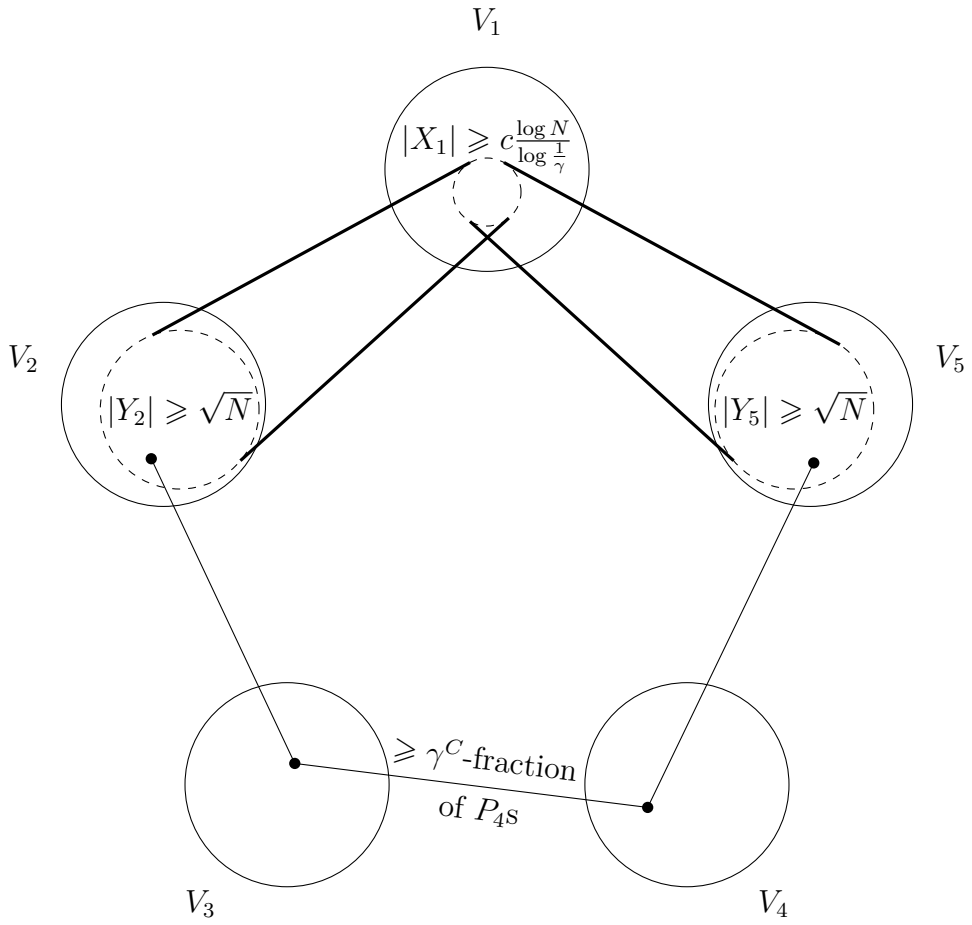
3 Proof sketch

To conclude, I want to discuss a few of the ideas that go into the proof of Theorem 5, and in particular, why the assumption that H is triangle-free arises. For concreteness, let's work with the special case $H = C_5$.

We are given an N-vertex graph G with γN^5 copies of C_5. We can actually assume (by slightly changing the parameters) that G is a 5-partite graph with parts $V_1, \dots, V_5$, where $|V_i| = N$ for all i, and with at least γN^5 *canonical* copies of C_5, that is, copies of C_5 with one vertex in each V_i and with the vertices of the C_5 going between the parts in cyclic order.

Our (first) goal is to find subsets $X_1 \subseteq V_1, Y_2 \subseteq V_2, Y_5 \subseteq V_5$ with the following properties:

- We have $|X_1| \geqslant c\frac{\log N}{\log \frac{1}{\gamma}}$, for some constant c;
- For $i \in \{2, 5\}$, we have that $|Y_i| \geqslant \sqrt{N}$;
- Every vertex in X_1 is adjacent to all vertices in Y_2 and Y_5;
- The number of canonical copies of $P_4 = C_5 \setminus \{v_1\}$ among Y_2, V_3, V_4, Y_5 is at least $\gamma^C |Y_2||V_3||V_4||Y_5|$.



Note that if we can find this structure, we can apply induction, running the same argument with $H = P_4$ among the sets Y_2, V_3, V_4, Y_5. Continuing in this fashion, we will eventually find a blowup of C_5. The key point is that, although we are paying a lot at every step—shrinking the size of some sets by a square root, and decreasing the number of copies from a γ-fraction to a γ^C-fraction—all of these losses are irrelevant once we take logarithms, since

$$\frac{\log \sqrt{N}}{\log \frac{1}{\gamma^C}} = \frac{\frac{1}{2}\log N}{C \log \frac{1}{\gamma}} = \Omega\left(\frac{\log N}{\log \frac{1}{\gamma}}\right).$$

Thus, at the end of this process, we will still only pick up a constant-factor loss (depending on H). Because of this, it suffices to do one step of the process, as described above.

In order to construct these sets, we apply a variant of the dependent random choice method, as follows. Set $s = c\frac{\log N}{\log \frac{1}{\gamma}}$ for an appropriate constant c. We pick random vertices $a_1, \dots, a_s \in V_1$, and random edges $b_1c_1, \dots, b_sc_s \in E(V_3 \times V_4)$. We define $X_1 = \{a_1, \dots, a_s\}$, and set $\mathcal{Y} \subseteq V_2 \times V_5$ to be the set of pairs $(y_2, y_5) \in V_2 \times V_5$ with the property that $a_iy_2b_ic_iy_5$ forms a copy of C_5 for all $1 \leqslant i \leqslant s$.

Note that by construction, we have that $|X_1| = s = c\frac{\log N}{\log \frac{1}{\gamma}}$. Also, we have by construction that every vertex in X_1 is adjacent to all vertices appearing in $\mathcal{Y}$, since these vertices were chosen to form copies of C_5 with the vertices in X_1. Additionally, it is not hard to show that, in expectation, we have many copies of P_4 with starting and ending vertices in $\mathcal{Y}$ and central vertices in $V_3 \times V_4$, and that, in expectation, $|\mathcal{Y}| \geqslant |V_2||V_5|/\sqrt{N}$. In fact, by appropriate computations, one can ensure that both of these events happen simultaneously with positive probability.

The final, and key, step is to observe that—deterministically—$\mathcal{Y}$ is not an arbitrary subset of $V_2 \times V_5$. Instead, $\mathcal{Y}$ is of the form $Y_2 \times Y_5$ for some $Y_2 \subseteq V_2, Y_5 \subseteq V_5$. The reason for this is as follows: for a pair (y_2, y_5) to form a C_5 with all tripes a_i, b_i, c_i, the condition we are imposing is just that y_2 is adjacent to all a_i, b_i, and that y_5 is adjacent to all a_i, c_i. These conditions have nothing to do with each other, so $\mathcal{Y}$ is just a product set. In particular, this plus the lower bound on $\mathcal{Y}$ yields all the properties we needed on the sets X_1, Y_2, Y_5.

This is the only place where we use that H (in this case C_5) is triangle-free. This argument only works because in C_5, we have no edge between vertices 2 and 5. In general, we are using that in H, the neighborhood of any vertex is an independent set. Without this assumption, the set $\mathcal{Y}$ would not necessarily be a product set, and we would lose all control. Because of this, it seems very difficult to push our technique and obtain any results even for K_3; making progess on Conjecture 4 for $H = K_3$ would be extremely interesting.