1 Background

Given a graph property $\mathcal{P}$, we say that an n-vertex graph G is *ε-far from $\mathcal{P}$* if one must add or delete at least εn^2 edges to G to create a graph satisfying property $\mathcal{P}$.

A central result in extremal graph theory is the triangle removal lemma of Ruzsa and Szemerédi, which states the following.

Theorem 1 (Ruzsa–Szemerédi 1978)**.** *If an n-vertex graph G is ε-far from triangle-free, then G contains at least δn^3 triangles, where $\delta = \delta(\varepsilon) > 0$ depends only on ε.*

Despite its innocent appearance, this is an extremely deep result, with applications in theoretical computer science, number theory, and many other fields. Additionally, all known proofs of the triangle removal lemma are quite involved, and use Szemerédi's regularity lemma or ideas related to it.

It is natural to wonder whether there is a simpler proof of the triangle removal lemma, which uses only elementary counting arguments. No one has found such a proof yet, and it is likely that one does not exist. One reason to believe that any proof of the triangle removal lemma must be "complicated" is that the triangle removal lemma implies deep results in other fields, such as Roth's theorem and the Ajtai–Szemerédi corners theorem in additive combinatorics. These theorems have no known "simple" proof, so it would be surprising to find an elementary proof of the triangle removal lemma.

However, in my opinion, there is a better reason to be skeptical of the existence of a counting-based proof. This relates to the quantitative aspects of the removal lemma. In general, proofs in combinatorics which use counting arguments give polynomial dependencies between the parameters. So if there were such a proof of the triangle removal lemma, we would expect to be able to take $\delta = \mathrm{poly}(\varepsilon)$ in Theorem 1.

Theorem 2 (Ruzsa–Szemerédi 1978)**.** *One cannot take $\delta = \mathrm{poly}(\varepsilon)$ in Theorem 1.*

More precisely, for every $\varepsilon > 0$ and every sufficiently large n, there exists an n-vertex graph G which is ε-far from triangle-free, but contains fewer than $\varepsilon^{c\log(1/\varepsilon)}n^3$ triangles, for an absolute constant $c > 0$.

It's worth remarking that in the other direction, the original proof of Ruzsa and Szemerédi showed that in Theorem 1, one *can* take $1/\delta$ to be at most $2^{2^{\cdot^{\cdot^{\cdot^{2}}}}}$, where the height of the tower is $\mathrm{poly}(1/\varepsilon)$. In a major breakthrough, Fox proved that one can instead take $1/\delta$ to be "merely" a tower of twos of height $O(\log(1/\varepsilon))$. Although this is a much better bound, there remains a massive gap between it and the barely superpolynomial bound of Theorem 2, and it is a major open problem to shrink this gap.

In many applications of the triangle removal lemma, these abysmal bounds are a serious issue. So many people have asked if one can impose additional assumptions on G in order to improve the bounds on δ, and in particular to take $\delta = \mathrm{poly}(\varepsilon)$. As it turns out, one can. For example, if one imposes certain "low-complexity" assumptions on G (e.g. that G has bounded VC-dimension, or that G is a semi-algebraic graph of bounded description complexity), then one obtains polynomial bounds in the triangle removal lemma applied to

G. In a recent paper, Fox and I obtained *linear* bounds under a different sort of assumption on G: we showed that if G has minimum degree at least αn where $\alpha > \frac{1}{3}$, then one can take $\delta = O(\varepsilon)$. Moreover, the bound $\alpha > \frac{1}{3}$ is tight: for any $\alpha < \frac{1}{3}$, there exist graphs with minimum degree at least αn such that superpolynomial bounds are again necessary in the triangle removal lemma.

2 Asymmetric removal

In this talk, we will not be focusing on these quantitative questions for the triangle removal itself. Instead, we will be interested in the following *asymmetric* removal lemma.

Theorem 3. *Let H be an h-vertex graph with $\chi(H) = 3$. If an n-vertex graph G is ε-far from triangle-free, then G contains at least δn^h copies of H, where $\delta = \delta(\varepsilon, H) > 0$ depends only on ε and H.*

We call this an *asymmetric* removal lemma because we have broken the symmetry in the roles of K_3 in Theorem 1: we still assume that we are ε-far from triangle-free, but now learn something about the counts of a *different* graph H. Note that the assumption $\chi(H) = 3$ is necessary, as a balanced complete tripartite graph is $\frac{1}{9}$-far from triangle-free, yet contains no copy of any H with $\chi(H) > 3$.[1]

Theorem 3 follows from the usual proof of the triangle removal lemma, with only a little bit more work. Alternately, one can prove Theorem 3 as a direct consequence of Theorem 1, via a supersaturation argument in an auxiliary hypergraph. But both of these proofs carry with them the same terrible quantitative aspects as in the usual triangle removal lemma. So it is natural to ask: are there graphs H where one can obtain better bounds in Theorem 3?

The first person to consider this question was Csaba, who proved that in case $H = C_5$, one can do significantly better than the tower-type bounds one generally has.

Theorem 4 (Csaba 2021)**.** *One can take $\delta(\varepsilon, C_5) = 2^{-\operatorname{poly}(1/\varepsilon)}$ in Theorem 3.*

Csaba's proof also uses ideas related to the regularity method: he proves an appropriate weak regularity lemma, which only invokes exponential losses between the parameters, and which is nonetheless strong enough to count copies of C_5 in graphs which are ε-far from triangle-free.

Our first new result is an improvement of this to an essentially optimal result.

Theorem 5 (Gishboliner–Shapira–W. 2023)**.** *One can take $\delta(\varepsilon, C_5) = \operatorname{poly}(\varepsilon)$ in Theorem 3.*

More generally, we have the following result for all pairs of odd cycles. Let $1 \le k < \ell$ be integers. If G is ε-far from C_{2k+1}-free, then it contains $\Omega(\varepsilon^{4\ell+2} n^{2\ell+1})$ copies of $C_{2\ell+1}$.

[1] One might ask what happens if $\chi(H) = 2$. In this case, then question is somewhat trivial: if G is ε-far from triangle-free, then it in particular contains at least εn^2 edges, and then a classical result of Kővári–Sós–Turán implies that it contains $\operatorname{poly}(\varepsilon) n^h$ copies of any bipartite H.

So, for example, we find $\Omega(\varepsilon^{10}n^5)$ copies of C_5 in any graph which is ε-far from triangle-free. The key thing to stress about this result is that it only works because of the asymmetry (i.e. the assumption that ℓ is strictly greater than k): the argument of Ruzsa and Szemerédi shows that one cannot have polynomial bounds if one is counting copies of C_{2k+1} in graphs which are far from C_{2k+1}-free.

Note too that polynomial dependencies are the best that one could hope for, and in fact, the exponent we get is best possible up to a factor of 2. Namely, it is not hard to check that a random graph of edge density ε is $\Theta(\varepsilon)$-far from triangle-free (or more generally C_{2k+1}-free) and has $O(\varepsilon^5 n^5)$ copies of C_5 (or more generally $O(\varepsilon^{2\ell+1}n^{2\ell+1})$ copies of $C_{2\ell+1}$).

Theorem 5 is proved via elementary counting and averaging arguments, as is hopefully not too surprising given that it gives polynomial bounds. We will see the simple proof at the end of the talk.

Let us return to the statement of Theorem 3: it tells us that if G is ε-far from triangle-free, then it contains at least δn^h copies of H, for some $\delta = \delta(\varepsilon, H)$. In some cases, such as $H = K_3$, we know that δ must be superpolynomial in ε. However, in other cases, such as $H = C_5$, we find that δ can be taken to be $\text{poly}(\varepsilon)$, by Theorem 5. What happens for general H?

Definition 6. We say that a tripartite graph H is *K_3-abundant* if we have $\delta(\varepsilon, H) = \text{poly}(\varepsilon)$.

To recap what we know, Ruzsa and Szemerédi proved that K_3 itself is not K_3-abundant, and Theorem 5 implies that $C_{2\ell+1}$ is K_3-abundant for all $\ell \geq 2$. It is not too hard to show the following two facts, which give us ways of producing more K_3-abundant graphs.

- If $H_1 \subseteq H_2$ and H_2 is K_3-abundant, then so is H_1.
- If H is K_3-abundant, then so is any blowup of H.

In particular, we find that any tripartite graph containing a cycle is not K_3-abundant, whereas any graph homomorphic to C_5 is K_3-abundant.

The natural question, given all of this, is whether containing a triangle is the *only* obstruction to abundance: is every triangle-free tripartite graph K_3-abundant?

Theorem 7 (Gishboliner–Shapira–W. 2023)**.** *There exist triangle-free tripartite graphs which are not K_3-abundant (assuming Ruzsa's genus conjecture in additive combinatorics).*

3 Additive combinatorics

To understand how number theory gets involved in this graph theory problem, let's back up and see the proof of Theorem 2. Recall that we wish to prove that there exists a graph G which is ε-far from triangle-free, yet contains only $\varepsilon^{\omega(1)}n^3$ triangles, where the $\omega(1)$ tends to infinity as $\varepsilon \to 0$.

Let m be an integer and let $R \subseteq [m]$. We define the *Ruzsa–Szemerédi graph* $\mathrm{RS}(m, R)$ to be the following graph. It has three parts X, Y, Z, each of which we identify with $[3m]$. The edges are given by

$$(x, y) \in X \times Y : y - x \in R \qquad (y, z) \in Y \times Z : z - y \in R \qquad (z, x) \in Z \times X : z - x \in 2R.$$

Note that for every $x \in [m]$ and $r \in R$, we have a triangle $(x, x + r, x + 2r) \in X \times Y \times Z$. Note that these triangles are edge-disjoint, as given any two vertices in such a triangle, we can recover x and r. Therefore, in order to make $\mathrm{RS}(m, R)$ triangle-free, we need to delete at least $m|R|$ edges. In particular, if $|R| \geq 100\varepsilon m$, then $\mathrm{RS}(m, R)$ is ε-far from triangle-free.

In fact, with the same logic, we can exactly characterize the set of triangles in $\mathrm{RS}(m, R)$. Namely, suppose $(x, y, z) \in X \times Y \times Z$ forms a triangle in $\mathrm{RS}(m, R)$. Let $a = y - x, b = \frac{1}{2}(z - x), c = z - y$, and note that by definition, these are all numbers in R. Moreover, from their definitions, we see that $a + c = 2b$, i.e. that a, b, c form an arithmetic progression. Note that if $a = b = c$ (i.e. if this is a *trivial* arithmetic progression), then our triangle (x, y, z) is precisely of the form $(x, x + r, x + 2r)$ described above.

Since our goal is to produce a graph with *few* triangles, we should try to have few arithmetic progressions in R. To this end, Ruzsa and Szemerédi used the following well-known result.

Theorem 8 (Salem–Spencer 1942, Behrend 1946)**.** *There exists a set $R \subseteq [m]$ with $|R| \geq m^{1-o(1)}$ that contains no non-trivial arithmetic progression.*

To conclude the proof, we pick m to be the largest integer so that there exists $R \subseteq [m]$ with no non-trivial arithmetic progression and with $|R| \geq 100\varepsilon m$. By Theorem 8, we have that $m \geq (1/\varepsilon)^{\omega(1)}$. The graph $\mathrm{RS}(m, R)$ is then ε-far from triangle-free. Additionally, the triangles in $\mathrm{RS}(m, R)$ are fully parameterized by pairs in $[m] \times R$, as R has no non-trivial arithmetic progressions. Since $\mathrm{RS}(m, R)$ has $n = 9m$ vertices, the number of triangles in it is

$$m|R| \leq m^2 \leq \frac{1}{1000m} n^3 = \varepsilon^{\omega(1)} n^3,$$

which is what we wanted.

Let's now turn to Theorem 7. Recall what we wish to prove: there exists triangle-free tripartite graph H which is not K_3-abundant. In other words, we need three ingredients: a graph H, a graph G which is ε-far from triangle-free, and a proof that G contains few copies of H, where "few" here means $\varepsilon^{\omega(1)} n^h$.

To start with the second ingredient, we will simply use Ruzsa and Szemerédi's construction, namely the graph $\mathrm{RS}(m, R)$ defined above. We know that as long as we pick R large enough, namely $|R| \geq 100\varepsilon m$, we automatically have that this graph is ε-far from triangle-free. All that remains is defining H so that, with an appropriate choice of R, $\mathrm{RS}(m, R)$ contains few copies of H. It will be more convenient in what follows to count *homomorphisms* $H \to \mathrm{RS}(m, R)$, i.e. "copies" where we don't require the vertices of H to map into distinct vertices of $\mathrm{RS}(m, R)$.

Let us stay agnostic for the moment about what H is, and remember only that H must be tripartite and triangle-free. What do copies of H in $\mathrm{RS}(m, R)$ look like? To pick a copy

of H in $\mathrm{RS}(m, R)$, we must first decide which of the three parts X, Y, Z each vertex of H gets mapped to; this amounts to picking a proper 3-coloring of H. Having chosen this, we now have various constraints on how we choose these vertices, since edges must correspond to elements in R.

As an example, let's think of $H = C_5$, and let its vertices be $v_1, \dots, v_5$. Let's suppose that the three-coloring we've chosen is

$$v_1 \mapsto X \qquad v_2 \mapsto Y \qquad v_3 \mapsto Z \qquad v_4 \mapsto Y \qquad v_5 \mapsto Z.$$

Denote by $w_1, \dots, w_5$ the vertices of $\mathrm{RS}(m, R)$ which we embed $v_1, \dots, v_5$ to, where we think of $w_1, \dots, w_5$ as numbers in $[3m]$. Then the definition of the edges in $\mathrm{RS}(m, R)$ implies that the following are all elements of R:

$$a = w_2 - w_1 \qquad b = w_3 - w_2 \qquad c = -(w_4 - w_3) \qquad d = w_5 - w_4 \qquad e = \frac{1}{2}(w_5 - w_1).$$

From the definitions of $a, \dots, e$ we see that they satisfy the linear equation

$$a + b - c + d - 2e = 0.$$

So we've found that, just as in the Ruzsa–Szemerédi proof, a copy of $H = C_5$ in $\mathrm{RS}(m, R)$ corresponds to a solution of a certain linear equation in R.

Actually, it is pretty clear that this is a general phenomenon. Namely, let H be a tripartite graph, and fix a proper 3-coloring χ of H. Then copies of H in $\mathrm{RS}(m, R)$ which are "consistent" with χ correspond to solutions in R of a certain system of linear equations. Namely, we first create a variable x_e for every edge $e \in E(H)$. Additionally, we assign every edge of H a weight depending on the value of χ on its endpoints: an edge from color class X to Y gets weight 1, an edge from Y to Z gets weight 1, and an edge from Z to X gets weight -2. Then every cycle in H gives us an equation, where we add up the variables corresponding to the edges on the cycle, weighted by the weights above, and where we think of edges as oriented (so that if an edge in the cycle goes from Y to X, for example, it gets coefficient -1). Call the set of equations that arise in this way S_χ.

Then the upshot of the above is that every homomorphism of H to $\mathrm{RS}(m, R)$ which is consistent with χ is parameterized by a "starting vertex" $w_1 \in X$, and by a solution in R to the system of equations S_χ. In particular, if R contains no non-trivial solutions to S_χ, then there are at most $|X||R|$ such homomorphisms (as R contains $|R|$ *trivial* solutions to S_χ). If we can also ensure that $|R| \geq m^{1-o(1)}$, then we conclude that $\mathrm{RS}(m, R)$ is ε-far from triangle-free, and contains at most $\varepsilon^{\omega(1)} n^h$ copies of H consistent with χ. If we can pick R so that this works *simultaneously* for all proper colorings χ, we have proved Theorem 7.

As it turns out, dealing with all proper colorings simultaneously is not a big deal, so let's continue focusing on a single coloring χ. How do we find $R \subseteq [m]$ with $|R| \geq m^{1-o(1)}$ containing no non-trivial solution to the system of equations S_χ? Is this even possible?

Such questions are well-studied in additive combinatorics. Let E be a linear equation with integer coefficients, say the equation $\sum_{i=1}^k \alpha_i x_i = 0$. We say that E is *translation-invariant* if $\sum_{i=1}^k \alpha_i = 0$; note that all the equations in S_χ are translation-invariant, as they arise from cycles in H. The following fundamental definition is due to Ruzsa.

Definition 9. Let E be a translation-invariant linear equation with integer coefficients, say $\sum_{i=1}^{k} \alpha_i x_i = 0$. One says that E *has genus one* if for every $\varnothing \subsetneq T \subsetneq [k]$, we have that $\sum_{i \in T} \alpha_i \neq 0$.

The following theorem and conjecture, both due to Ruzsa, suggest that the property of having genus one is the fundamental one to understand the size of sets avoiding solutions to linear equations.

Theorem 10 (Ruzsa 1993)**.** *Let E be a translation-invariant linear equation that does not have genus one. If $R \subseteq [m]$ contains no non-trivial solution to E, then $|R| = O(\sqrt{m})$.*

Conjecture 11 (Ruzsa's genus conjecture [Ruzsa 1993])**.** *Let E be a translation-invariant linear equation that has genus one. Then for every integer m, there exists $R \subseteq [m]$ containing no non-trivial solution to E with $|R| \geq m^{1-o(1)}$.*

At first sight, this is bad news for us. Indeed, if H is triangle-free, then no equation in S_χ has genus one. For example, in the $H = C_5$ example we did above, the equation $a + b - c + d - 2e = 0$ we found does *not* have genus one. It is not hard to check that the only way a cycle in H can yield an equation of genus one is if the cycle has length 3. Since we are interested in triangle-free graphs, no equation in S_χ can have genus one.

However, there is a sliver of hope. We can extend the notion of genus-one equations to families of equations as follows.

Definition 12. Let S be a set of translation-invariant linear equations, where the jth equation in S is $\sum_{i=1}^{k} \alpha_i^{(j)} x_i = 0$. One says that S *has genus one* if for every $\varnothing \subsetneq T \subsetneq [k]$, there is some j with $\sum_{i \in T} \alpha_i^{(j)} \neq 0$.

Note that a set of equations can have genus one even if no particular equation in it has genus one. This is because a set T witnessing that S does not have genus one must be such a witnessing set for every equation in S simultaneously.

It is not hard to show that Conjecture 11 implies a corresponding statement for sets of genus one. Namely, Conjecture 11 implies that if S is a genus-one set of equations, then there is some $R \subseteq [m]$ with $|R| \geq m^{1-o(1)}$ such that R contains no non-trivial solutions to S. Therefore, to complete the proof of Theorem 7, it suffices to prove the following result.

Theorem 13 (Gishboliner–Shapira–W. 2023)**.** *There exists a tripartite triangle-free graph H such that for every proper coloring χ of H, the set of equations S_χ has genus one.*

Again, it's crucial to stress that any *single* equation in S_χ does not have genus one: the point is that if H is sufficiently "complicated", these equations interact with one another and ensure that S_χ has genus one.

How does one prove Theorem 13? Recall that the equations in S_χ correspond to cycles in H, and the variables in these equations correspond to edges of H. Thus, we can think of a set $\varnothing \subsetneq T \subsetneq E(H)$ as a coloring of $E(H)$ in black and white (where black edges are the ones in T), such that there is at least one edge of each color. To certify that S_χ has

genus one, we need to show that for every such coloring, there is some cycle of H such that the sum of the weights on the black edges (or, equivalently, the white edges) in the cycle is non-zero. Of course, this cannot work for an arbitrary H, so the real question is how to find an H with this property. The answer, perhaps unsurprisingly, is to use randomness.

Proposition 14 (Gishboliner–Shapira–W. 2023)**.** *Let A, B, C be disjoint sets of vertices, each of size n. Let H_0 be a random tripartite graph on $A \cup B \cup C$ where every edge is included independently with probability $p = n^{-3/4}$. Delete one edge from every triangle to form a triangle-free subgraph H. The following holds with high probability as $n \to \infty$.*

Suppose we color $E(H)$ in black and white, such that there is at least one edge of each color. Then there is a cycle of H containing either a one or two black edges; if there are two, they are consecutive and touch all three color classes (or this holds upon interchanging white and black).

Note that if a cycle has a single black edge, then the sum of the weights on the black edges is certainly non-zero. Similarly, if there are two consecutive black edges touching all three color classes, then the sum of the black weights is again non-zero. Thus, Proposition 14 completes the proof of Theorem 7.

In fact, we prove something somewhat stronger than Proposition 14. There is a short list of explicit pseudorandomness conditions on a tripartite triangle-free graph H that imply the conclusion of Proposition 14. We then verify the (simple) fact that a random tripartite graph with its triangles deleted satisfies these pseudorandomness conditions.

I will not present the proof of Proposition 14, nor state the precise pseudorandomness conditions we need. But at a very high level, the proof proceeds via Ramsey-theoretic arguments to show that if a given coloring does not have any "good" cycle, then it must have more and more structure. Eventually, this structure is enough to conclude that in this coloring, all edges must have the same color.

4 Back to triangles and pentagons

To finish this talk, I want to sketch a proof Theorem 5. I will only deal with the simplest case, of $k = 1, \ell = 2$. Namely, we assume that G is ε-far from triangle-free, and we wish to prove that G has δn^5 copies of C_5, where $\delta = \mathrm{poly}(\varepsilon)$.

So fix some n-vertex graph G which is ε-far from triangle-free. Consider a maximal collection of edge-disjoint triangles in G. If we delete all the edges from these triangles, we must destroy all triangles in G, by maximality. Therefore, there must be at least $(\varepsilon n^2)/3$ such edge-disjoint triangles. For simplicity, let's actually assume something stronger. Namely, let's assume that every vertex in G is incident to at least εn edge-disjoint triangles. There is a simple argument that shows that this assumption is essentially without loss of generality.

So now we have G, where every vertex in G is incident to at least εn edge-disjoint triangles. In particular, every vertex has degree at least εn. Fix some vertex $v \in V(G)$, and let A be the set of neighbors of v.

Every vertex $a \in A$ lies in at least εn edge-disjoint triangles, by assumption. This means that there exist vertices $b_1(a), \dots, b_{\varepsilon n}(a), c_1(a), \dots, c_{\varepsilon n}(a)$ such that for every i, the vertices $a, b_i(a), c_i(a)$ form a triangle. Since we assume these triangles are edge-disjoint, all the vertices $b_1(a), \dots, b_{\varepsilon n}(a), c_1(a), \dots, c_{\varepsilon n}(a)$ are distinct (for a fixed a).

Now, let B be the set of vertices in G that are of the form $b_i(a)$ for some index i and some $a \in A$, and similarly let C be the set of all $c_i(a)$. Then every $a \in A$ has at least εn neighbors in B. Additionally, note that an average vertex in B is of the form $b_i(a)$ for at least $\varepsilon^2 n$ choices of $a \in A$: this is because there are at least $\varepsilon^2 n^2$ choices for $(i, a) \in [\varepsilon n] \times A$, and at most n vertices in B. Therefore, an average vertex of B has at least $\varepsilon^2 n$ neighbors in C; one for each way of representing it as $b_i(a)$. Let's again assume, essentially without loss of generality, that actually *every* vertex in B has at least $\varepsilon^2 n$ neighbors in C. By the exact same argument, we see that an average vertex in C has at least $\varepsilon^2 n$ neighbors in A, and we assume that in fact, every vertex in C has at least $\varepsilon^2 n$ neighbors in A.

We now have many ways of constructing a path $(a, b, c, a') \in A \times B \times C \times A$. We first fix some $a \in A$, then have εn choices for a neighbor $b \in B$, then at least $\varepsilon^2 n$ choices for a neighbor $c \in C$, then at least $\varepsilon^2 n$ choices for a neighbor $a' \in A$. In total, there are $\text{poly}(\varepsilon) n^4$ such paths a, b, c, a'. To conlude, recall that as $a, a' \in A$, they are both neighbors of v. So v lies in $\text{poly}(\varepsilon) n^4$ copies of C_5. Repeating this argument for every $v \in V$, we find that G contains at least $\text{poly}(\varepsilon) n^5$ copies of C_5, as claimed.