Szemerédi's regularity lemma

notes by

Marcel K. Goh

12 January 2020

1. Definitions and notation.

In these notes, we consider undirected graphs $G = (V, E)$, where V is an arbitrary finite set and E consists of 2-element subsets of V. We sometimes write $V(G)$ for V and $E(G)$ for E, and for convenience, we will write uv instead of $\{u, v\}$ when the vertices u and v are adjacent. Fix a graph $G = (V, E)$ and let X and Y be subsets of V (not necessarily disjoint). Let $e(X, Y) = \{xy \in E : x \in X,\, y \in Y\}$ denote the number of edges that have an endpoint in each of X and Y. We define the *edge density* between X and Y to be the ratio

$$d(X, Y) = \frac{e(X, Y)}{|X||Y|}. \tag{1}$$

If X and Y are disjoint, then this is the fraction of all possible edges between X and Y that are actually present in the graph (and in the case that they are not disjoint, it isn't awfully far off anyway).

We say that a pair of vertex subsets (X, Y) is *ϵ-regular* if for all subsets $A \subseteq X$ and $B \subseteq Y$ with $|A| \geq \epsilon|X|$ and $|B| \geq \epsilon|Y|$, we have $|d(A, B) - d(X, Y)| \leq \epsilon$. This means that if we zoom in to look at the edges between a subset of X and a subset of Y, we find that the picture is sort of a "scale-model" of the whole of X and the whole of Y in the sense that the number of edges that we see is proportional to the sizes of the subsets, unless the subsets are taken to be very small. If the pair (X, Y) is *not* ϵ-regular, then there must be some $A \subseteq X$ and $B \subseteq Y$, with $|A| \geq \epsilon|X|$ and $|B| \geq \epsilon|Y|$, such that $|d(A, B) - d(X, Y)| > \epsilon$. The pair (A, B) is said to *witness* the irregularity.

A *partition* $\mathcal{P}$ is a collection $\{V_1, \ldots, V_k\}$ of disjoint subsets of V whose union is all of V. We will say that a partition is *equitable* if the sizes of any two parts do not differ by more than 1. A partition is said to be *ϵ-regular* if the sum of $|V_i||V_j|$, taken over all pairs (V_i, V_j) that are not ϵ-regular, is less than $\epsilon|V|^2$. If the partition is equitable, then this is equivalent to saying that at most ϵk^2 of the pairs (V_i, V_j) are not ϵ-regular. Szemerédi's regularity lemma says that every graph admits an ϵ-regular equitable partition into a number of parts depending only on ϵ (and not the size of the graph).

A function $f(n)$ is said to be $O\big(g(n)\big)$ if there exists a constant C such that $f(n) \leq Cg(n)$ for all integers $n \geq 1$. A function $f(n)$ is $o\big(g(n)\big)$ if $lim_{n\to\infty} f(n)/g(n) = 0$. The notation $[a, b]$ will always denote the *discrete interval* $\{n \in \mathbf{Z} : a \leq n \leq b\}$.

References for the original instances of the proofs are listed at the bottom of this document; the particular presentation of these notes is heavily informed by lectures given by Yufei Zhao in 2019.

2. The regularity lemma

In this section we will prove Szemerédi's regularity lemma via a sequence of auxiliary ones. The idea runs as follows. We begin with a partition of $G = (V, E)$ that is given to us (e.g., the trivial partition $\mathcal{P} = \{V\}$) and while the partition is not ϵ-regular, we iteratively refine it by subdividing each element of the partition into further parts. Using an "energy increment argument", we show that this process terminates after a bounded number of steps, and therefore the number of parts in the final partition is bounded. For vertex sets U and W, we define the *energy* of (U, W) to be the quantity

$$q(U, W) = \frac{|U||W|}{n^2} d(U, W)^2, \tag{2}$$

where $n = |V|$, and for partitions $\mathcal{P}_U = \{U_1, U_2, \ldots, U_k\}$ and $\mathcal{P}_W\{W_1, W_2, \ldots, W_l\}$ of U and W respectively, we will define the *energy* of the two partitions to be the sum

$$q(\mathcal{P}_U, \mathcal{P}_W) = \sum_{i=1}^{k} \sum_{j=1}^{l} q(U_i, W_j). \tag{3}$$

We will write $q(\mathcal{P}) = q(\mathcal{P}, \mathcal{P})$ when the two partitions are equal. Note that if $\mathcal{P}$ is a partition of V into k parts, we have

$$q(\mathcal{P}) = \sum_{i=k}^{k} \sum_{j=1}^{k} \frac{|V_i||V_j|}{n^2} d(V_i, V_j)^2 \leq 1, \tag{4}$$

since edge density is at most 1. The first lemma states that the energy of a pair of refined partitions is at least the energy of the original pair.

Lemma A. *Let $G = (V, E)$ be a graph, let $U, W \subseteq V$ and suppose that $\mathcal{P}_U = \{U_1, U_2, \ldots, U_k\}$ and $\mathcal{P}_W\{W_1, W_2, \ldots, W_l\}$ are partitions of U and W respectively. Then $q(\mathcal{P}_U, \mathcal{P}_W) \geq q(U, W)$.*

Proof. We will define a random variable Z as follows. We select vertices $u \in U$ and $w \in W$ uniformly at random; suppose that $U_i \in \mathcal{P}_U$ contains u and $W_j \in \mathcal{P}_W$ contains w. We let $Z = d(U_i, W_j)$. We compute the first moment

$$\mathbf{E}\{Z\} = \sum_{i=1}^{k} \frac{|U_i|}{|U|} \sum_{j=1}^{l} \frac{|W_j|}{|W|} d(U_i, W_j) = \frac{e(U, W)}{|U||W|} = d(U, W) \tag{5}$$

and the second moment

$$\mathbf{E}\{Z^2\} = \sum_{i=1}^{k} \frac{|U_i|}{|U|} \sum_{j=1}^{l} \frac{|W_j|}{|W|} d(U_i, W_j)^2 = \frac{n^2}{|U||W|} q(\mathcal{P}_U, \mathcal{P}_W) \tag{6}$$

of Z, where n denotes the size of V. By Jensen's inequality, we have $\mathbf{E}\{Z^2\} \geq \mathbf{E}\{Z\}^2$ and therefore

$$q(\mathcal{P}_U, \mathcal{P}_W) \geq \frac{|U||W|}{n^2} d(U, W) = q(U, W). \quad \blacksquare \tag{7}$$

In particular, if $\mathcal{P}$ is a partition of V and $\mathcal{P}'$ refines $\mathcal{P}$, then we can apply Lemma A to every pair (V_i, V_j) of sets in $\mathcal{P}$ to conclude that $q(\mathcal{P}') \geq q(\mathcal{P})$. The next lemma shows that the inequality in Lemma A is sometimes strict, a fact we will need for the energy increment argument.

Lemma B. *With the same definitions as in Lemma A, suppose furthermore that for some $\epsilon > 0$, the pair (U, W) is not ϵ-regular and the irregularity is witnessed by $U_1 \subseteq U$ and $W_1 \subseteq W$. Then*

$$q(\{U_1, U \setminus U_1\}, \{W_1, W \setminus W_1\}) \geq q(U, W) + \epsilon^4 \frac{|U||W|}{n^2}, \tag{8}$$

where $n = |V|$.

Proof. Define the random variable Z as in the proof of Lemma *A*. Note that the variance of Z is

$$\begin{aligned} \mathbf{V}\{Z\} &= \mathbf{E}\{Z^2\} - \mathbf{E}\{Z\}^2 \\ &= \frac{n^2}{|U||W|} q(\{U_1, U \setminus U_1\}, \{W_1, W \setminus W_1\}) - d(U, W)^2 \\ &= \frac{n^2}{|U||W|} \big(q(\{U_1, U \setminus U_1\}, \{W_1, W \setminus W_1\}) - q(U, W)\big). \end{aligned} \tag{9}$$

But we also have the formula

$$\begin{aligned} \mathbf{V}\{Z\} &= \mathbf{E}\left\{(Z - \mathbf{E}\{Z\})^2\right\} \\ &= \frac{|U_1||W_1|}{|U||W|} \big(d(U_1, W_1) - d(U, W)\big)^2 + \frac{|U_1||W \setminus W_1|}{|U||W|} \big(d(U_1, W \setminus W_1) - d(U, W)\big)^2 \\ &\quad + \frac{|U \setminus U_1||W_1|}{|U||W|} \big(d(U \setminus U_1, W_1) - d(U, W)\big)^2 + \frac{|U \setminus U_1||W \setminus W_1|}{|U||W|} \big(d(U \setminus U_1, W \setminus W_1) - d(U, W)\big)^2 \\ &\geq \frac{|U_1|}{|U|} \cdot \frac{|W_1|}{|W|} \cdot \big(d(U_1, W_1) - d(U, W)\big)^2 \\ &\geq \epsilon \cdot \epsilon \cdot \epsilon^2, \end{aligned} \tag{10}$$

where the final inequality follows from the fact that (U_1, W_1) was the witness for the non-ϵ-regularity of (U, W). Combining both calculations for the variance proves the inequality we need. $\blacksquare$

We are now able to formulate the step in the inner loop of our regularisation procedure.

Lemma C. *Let $G = (V, E)$ be a graph, let $\mathcal{P} = \{V_1, V_2 \ldots, V_k\}$ be a partition of V, and let $\epsilon > 0$. If the partition $\mathcal{P}$ is not ϵ-regular, then there exists a refinement $\mathcal{Q}$ of $\mathcal{P}$ in which every V_i is partitioned into at most 2^k parts and such that*

$$q(\mathcal{Q}) \geq q(\mathcal{P}) + \epsilon^5. \tag{11}$$

Proof. If $\mathcal{P}_1$ and $\mathcal{P}_2$ are refinements of $\mathcal{P}$ that subdivide V_i into $V_{i1} \cup V'_{i1}$ and $V_{i2} \cup V'_{i2}$ respectively, then the *common refinement* of $\mathcal{P}_1$ and $\mathcal{P}_2$ divides V_i into the union

$$V_i = (V_{i1} \cap V_{i2}) \cup (V'_{i1} \cap V_{i2}) \cup (V_{i1} \cap V'_{i2}) \cup (V'_{i1} \cap V'_{i2}); \tag{12}$$

by induction, we can similarly define the common refinement of any finite number of partitions that refine $\mathcal{P}$. For every pair (i, j) for which (V_i, V_j) is not ϵ-regular, we can find $A_{ij} \subseteq V_i$ and $A_{ji} \subseteq V_j$ that witnesses the irregularity. Lemma B will produce a refinement of $\mathcal{P}$ for each (i, j) that divides V_i and V_j each into two new parts, and we can let $\mathcal{Q}$ be the common refinement of these partitions, as defined above. Note that we have constructed $\mathcal{Q}$ such that it does not have more than 2^k parts.

Let $\mathcal{R}$ be the set of all $(i, j) \in [1, k]^2$ such that (V_i, V_j) is ϵ-regular. For each i, let $\mathcal{Q}_{V_i}$ denote the subdivision of V_i given by $\mathcal{Q}$. By Lemma A, we have

$$\begin{aligned} q(\mathcal{Q}) &= \sum_{i=1}^{k} \sum_{j=1}^{k} q(\mathcal{Q}_{V_i}, \mathcal{Q}_{V_j}) \\ &\geq \sum_{(i,j) \in \mathcal{R}} q(V_i, V_j) + \sum_{(i,j) \notin \mathcal{R}} q(\{A_{ij}, V_i \setminus A_{ij}\}, \{A_{ji}, V_j \setminus A_{ji}\}). \end{aligned} \tag{13}$$

Applying Lemma B, we find that

$$\begin{aligned} q(\mathcal{Q}) &\geq \sum_{(i,j) \in \mathcal{R}} q(V_i, V_j) + \sum_{(i,j) \notin \mathcal{R}} q(V_i, V_j) + \epsilon^4 \frac{|V_i||V_j|}{n^2} \\ &\geq q(\mathcal{P}) + \epsilon^5, \end{aligned} \tag{14}$$

as desired. ∎

With Lemma C in hand, we can state and prove Szemerédi's regularity lemma without too much further effort.

Theorem R (*Szemerédi*, 1978). *For all $\epsilon > 0$ there exists an M with such that every graph admits an ϵ-regular partition of its vertices into no more than M parts. The constant M depends only on ϵ (not on the size of the graph) and we have the upper bound*

$$M \leq 4^{4^{\cdot^{\cdot^{\cdot^4}}}} \tag{15}$$

where the tower of 4s consists of ϵ^{-5} repeated exponents.

Proof. We start with the trivial partition $\mathcal{P} = \{V\}$ and apply Lemma C while the current partition is not ϵ-regular. Since $0 \leq q(\mathcal{P}) \leq 1$, and the energy of the partition increases by at least ϵ^5 with each iteration, the algorithm terminates after at most ϵ^{-5} steps. At any particular step, if $\mathcal{P}$ has k parts, Lemma C outputs a partition with $k2^k \leq 4^k$ parts, and this observation yields the upper bound in the theorem statement. ∎

The bound we stated above is not tight, but it turns out that the tower of exponents is inescapable. It was shown by Gowers (1997) that there exists a $c > 0$ such that for all $\epsilon > 0$ small enough, one can construct a graph whose ϵ-regular partition requires more than M parts, where M is a exponential tower of 2s that is ϵ^{-c} high.

Equitable partitions. In Lemma C, one can require that the resulting partition $\mathcal{Q}$ be equitable (no two parts differ by more than one element), and the inequality would still hold, although with an increment that is possibly less than ϵ^5. The difference is negligible, in the sense that the final bound obtained for M will

still be of the same order. It is important to note that it is *not* possible to obtain an ϵ-regular partition by proving the regularity lemma with Lemma C as stated above, and then further subdividing and merging the partitions afterwards, because

3. Graph counting and removal

A *graph homomorphism* $f : H \to G$ is a function from $V(H)$ to $V(G)$ such that for any edge $uv \in E(H)$, $f(u)f(v)$ is an edge in $E(G)$. We begin by counting the number of homomorphic copies of a fixed graph H in a larger graph G, provided some regularity conditions on the vertex set of G hold.

Lemma C (*Graph counting lemma*). *Let H be a graph with vertex set $[1,k]$. Let $\epsilon > 0$ and let G be another graph with vertex subsets $V_1, V_2, \ldots, V_k \subseteq V(G)$ such that (V_i, V_j) are ϵ-regular for all $ij \in E(H)$. If $m = \big|E(H)\big|$ and n is the number of k-tuples $(v_1, v_2, \ldots, v_k) \in V_1 \times V_2 \times \cdots \times V_k$ such that $v_i v_j \in E(G)$ whenever $ij \in E(H)$, then we have*

$$\Big| n - \prod_{ij \in E(H)} d(V_i, V_j) \prod_{i=1}^{k} |V_i| \Big| \leq m\epsilon \prod_{i=1}^{k} |V_i|. \tag{16}$$

Proof. To avoid trivialities, we assume $k \geq 3$. If we sample ξ_1 uniformly from V_1, ξ_2 from V_2, and so on, and let B be the event that $\xi_i \xi_j \in E(G)$ for all $ij \in E(H)$, then the above inequality is equivalent to the statement

$$\Big| \mathbf{P}\{B\} - \prod_{ij \in E(H)} d(V_i, V_j) \Big| \leq m\epsilon. \tag{17}$$

We will prove this statement by induction on m. If $m = 0$, both sides equal 0, since the empty product equals 1. Now assume that $m \geq 1$ and by relabelling the vertices of H, we can assume that $12 \in E(H)$. Let B' be the event that $\xi_i \xi_j \in E(G)$ for all $ij \in E(H) \setminus \{12\}$. By the induction hypothesis, we have

$$\left| \mathbf{P}\{B'\} - \frac{1}{d(V_1, V_2)} \prod_{ij \in E(H)} d(V_i, V_j) \right| \leq (m-1)\epsilon, \tag{18}$$

and since edge density is bounded above by 1, we can rewrite this as

$$\Big| d(V_1, V_2)\, \mathbf{P}\{B'\} - \prod_{ij \in E(H)} d(V_i, V_j) \Big| \leq (m-1)\epsilon. \tag{19}$$

Now let $(v_3, v_4, \ldots, v_k)$ in $V_3 \times V_4 \times \cdots \times V_k$ be arbitrary and let A_1 be the set of all $v_1 \in V_1$ such that $v_1 v_i \in E(G)$ for all $1i \in E(H)$, $i \neq 2$. Likewise, let A_2 be the set of all $v_2 \in V_1$ such that $v_2 v_i \in E(G)$ for all $2i \in E(H)$, $i \neq 1$. If $|A| \geq \epsilon |V_1|$ and $|A_2| \geq \epsilon |V_2|$, then since (V_1, V_2) is ϵ-regular, the inequality $\big|d(A_1, A_2) - d(V_1, V_2)\big| \leq \epsilon$ implies that

$$\left| \frac{e(A_1, A_2)}{|V_1| \cdot |V_2|} - d(V_1, V_2) \frac{|A_1| \cdot |A_2|}{|V_1| \cdot |V_2|} \right| \leq \epsilon \frac{|A_1| \cdot |A_2|}{|V_1| \cdot |V_2|} \leq \epsilon; \tag{20}$$

this inequality also holds if $|A_1| < \epsilon |V_1$ or $|A_2| < \epsilon |V_2|$, since in this case both terms are less than ϵ. The first term is the probability that the choices for ξ_1 and ξ_2 have the correct neighbours in the set $\{v_3, v_4, \ldots, v_k\}$ *and* are connected, while the second term does not require that ξ_1 and ξ_2 be connected. In other words,

$$\frac{e(A_1, A_2)}{|V_1| \cdot |V_2|} = \mathbf{P}\{B \mid \xi_3 = v_3,\, \xi_4 = v_4,\, \ldots,\, \xi_k = v_k\} \quad \text{and} \quad \frac{|A_1| \cdot |A_2|}{|V_1| \cdot |V_2|} = \{B' \mid \xi_3 = v_3,\, \xi_4 = v_4,\, \ldots,\, \xi_k = v_k\}. \tag{21}$$

Substituting this back into (20), we have

$$\big| \mathbf{P}\{B \mid \xi_3 = v_3,\, \xi_4 = v_4,\, \ldots,\, \xi_k = v_k\} - d(V_1, V_2)\, \mathbf{P}\{B' \mid \xi_3 = v_3,\, \xi_4 = v_4,\, \ldots,\, \xi_k = v_k\} \big| \leq \epsilon. \tag{22}$$

Since this holds for *any* choice of $(v_3, v_4, \ldots, v_k)$, we have *a fortiori*

$$\left|\mathbf{P}\{B\} - d(V_1, V_2)\,\mathbf{P}\{B'\}\right| \leq \epsilon, \tag{23}$$

which results by letting $\xi_3, \xi_4, \ldots, \xi_k$ once again be chosen uniformly from their respective sets. Adding the induction hypothesis (19) and applying the triangle inequality, we have

$$\left|\mathbf{P}\{B\} - \prod_{ij\in E(H)} d(V_i, V_j)\right| \leq \left|\mathbf{P}\{B\} - d(V_1, V_2)\,\mathbf{P}\{B'\}\right| + \left|d(V_1, V_2)\,\mathbf{P}\{B'\} - \prod_{ij\in E(H)} d(V_i, V_j)\right| \leq m\epsilon, \tag{24}$$

which was what we had to show. ∎

We say that a graph G is *H-free* if it does not admit a graph homomorphism $f : H \to G$. We can now formulate the graph removal lemma, which states that if a graph G does not contain too many homomorphic copies of a graph H, then one does not need to remove that many edges to make it H-free. The following theorem was first proven in 1978 by I. Ruzsa and E. Szemerédi for the case $H = K_3$, and extended to arbitrary H in 1986 By P. Erdős, P. Frankl, and V. Rödl.

Lemma R (*Graph removal lemma*). *Fix a graph H with $k \geq 3$ vertices and m edges. For any $\epsilon > 0$ there exists $\delta > 0$ depending on H and ϵ such that every n-vertex graph G with at most δn^k homomorphic copies of H can be made H-free by removing $\leq \epsilon n^2$ edges.*

Proof. Let G be a graph on n vertices and let $\mathcal{P} = \{V_i\}$ be an $(\epsilon/4)$-regular partition of $V(G)$ given by the regularity lemma (thus it has no more than M parts, where M depends only on ϵ). We will remove all edges between parts V_i and V_j if

i) (V_i, V_j) is not $(\epsilon/4)$-regular;
ii) $d(V_i, V_j) \leq \epsilon/2$; or
iii) either V_i or V_j has $\leq (\epsilon/4M)n$ vertices.

The total number of edges deleted in case (i) is $\epsilon n^2/4$, by the definition of an $(\epsilon/4)$-regular partition. In case (ii), we remove no more than $\epsilon n^2/2$ edges and case (iii) happens no more than $M \cdot n$ times, causing no more than $\epsilon n^2/4$ edges to be deleted. Thus we have removed at most ϵn^2 edges in total. If there exists a copy of H in this post-surgery graph, the vertices $1, 2, \ldots, k$ of H must lie in parts (not necessarily distinct) $V_1, V_2, \ldots, V_k \in \mathcal{P}$, each with greater than $(\epsilon/4M)n$ vertices, such that for all $(i, j) \in [1, k]^2$, (V_i, V_j) is $(\epsilon/4)$-regular and $d(V_i, V_j) > \epsilon/2$. By the graph counting lemma, the number C of homomorphic copies of H in G satisfies

$$\left|C - \prod_{ij\in E(H)} d(V_i, V_j) \prod_{i=1}^{k} |V_i|\right| \leq m\frac{\epsilon}{4} \prod_{i=1}^{k} |V_i|. \tag{25}$$

This means that there are at least

$$\frac{1}{k!}\left(\frac{\epsilon}{2}\right)^m \left(\frac{\epsilon}{4M}\right)^k n^k - m\frac{\epsilon}{4}\left(\frac{\epsilon}{4M}\right)^k n^k \tag{26}$$

homomorphic copies of H in G, where we have divided by $k!$ to handle the overcounting that occurs if some (or indeed all) of the V_i are actually equal. So we may set

$$\delta < \left(\frac{\epsilon}{2}\right)^m \left(1 - m\frac{\epsilon}{4}\right)\left(\frac{\epsilon}{4M}\right)^k \tag{27}$$

and by contraposition, if G did not contain more than δn^k copies of H to begin with, then the procedure above removed all copies of H in G. ∎

The triangle counting lemma can be stated in a more succinct way that avoids ϵ-δ terminology: *If H has k vertices, then any graph with $o(n^k)$ copies of H can be made H-free by deleting $o(n^2)$ edges.* Of course, given a single graph on n vertices, one cannot tell whether it has $o(n^k)$ copies of H, so more precisely,

we're saying that given any $f(n) = o(n^k)$, there exists $g(n) = o(n^2)$ such that any graph on n vertices with $\leq f(n)$ copies of H can be made H-free by deleting at most $g(n)$ edges. Finally, although we have counted homomorphisms that are not necessarily injective, as the size of G gets large relative to a fixed graph H, the proportion of homomorphisms that are non-injective quickly becomes negligible, so the number of subgraphs is roughly the number of homomorphisms.

4. Roth's theorem

By letting H be the triangle graph K_3, we can prove a special case of Szemerédi's theorem, namely Roth's theorem on arithmetic progressions of length 3, which was first proved in 1953. Since any homomorphism from K_3 to a graph G must be injective, the graph counting lemma counts *bona fide* subgraphs of K_3 in larger graphs G. We begin with the following corollary of the removal lemma.

Lemma T. *If G is an n-vertex graph and every edge lies in exactly one triangle, then the number of edges of G is $o(n^2)$.*

Proof. Suppose there are m edges in G. Then the number of triangles is $m' = m/3$, but since $m \leq \binom{n}{2}$, this is $o(n^3)$. Applying the triangle removal lemma, we only need to remove $o(n^2)$ edges to make the graph triangle-free. But then again, we needed to remove at least $1/3$ of the edges to make G triangle-free, so the total number of edges is $o(n^2)$ as well. ∎

Roth's theorem states that for any $\delta > 0$, there exists N such that for any $n \geq N$, every subset of $[1, n]$ with size at least δn has an arithmetic progression of length 3. We restate this slightly differently using the asymptotic notation described above and then prove it using the triangle removal lemma.

Theorem R (*Roth's theorem*). *If $A \subseteq [1, n]$ does not have any 3-term arithmetic progressions, then $|A| = o(n)$.*

Proof. Let $A \subseteq [1, n]$ be free of 3-term arithmetic progressions. Let $m = 2n + 1$ and note that if A is taken to be a subset of the cyclic group $\mathbf{Z}/m\mathbf{Z}$, it also does not have any 3-term arithmetic progressions in this group. Construct G such that $V(G) = X \cup Y \cup Z$, where $X = Y = Z = \mathbf{Z}/m\mathbf{Z}$. We define the edge set of G as follows:

i) Connect $x \in X$ and $y \in Y$ whenever $y - x \in A$.

ii) Connect $x \in X$ and $z \in Z$ whenever $(z - x)/2 \in A$. We are allowed to divide by 2 because the multiplicative inverse of 2 in $\mathbf{Z}/m\mathbf{Z}$ is $n + 1$.

iii) Connect $y \in Y$ and $z \in Z$ whenever $z - y \in A$.

Since we are doing addition in $\mathbf{Z}/m\mathbf{Z}$, each element of A contributes m edges to $E(G)$ in each case. If $x \in X$, $y \in Y$, and $z \in Z$ are the vertices of a triangle in G, then $y - x$, $(z - x)/2$, and $z - y$ all belong to A. But A does not contain any proper arithmetic progressions of length 3, so these elements must all be the same, that is,

$$y - x = \frac{z - x}{2} = z - y. \tag{28}$$

Given any two of x, y, z, we can determine the third from this equation, so every edge belongs to exactly one triangle. Thus the number of edges is $o(m^2)$. On the other hand, the number of edges is exactly $3m|A|$, so $3m|A| = o(m^2)$ and $|A| = o(m) = o(n)$. ∎

Szemerédi's theorem has the same statement as Roth's theorem, with 3 replaced by an arbitrary integer k. Note that we cannot extend this result to longer arithmetic progressions with the graph regularity lemma alone. However, it was first shown in a 2007 paper by W. T. Gowers, that one can extend the regularity lemma to hypergraphs, and by encoding the faces of a $(k-1)$-simplex as a $(k-1)$-uniform hypergraph as above, one can prove Szemerédi's theorem directly from the hypergraph removal lemma.

5. Property testing

Fix a graph H on k vertices. Suppose we have a graph G with a large number n of vertices and we know, for some $\epsilon > 0$, that it is either H-free or "ϵ-far" from H-free, in that it cannot be made H-free by removing

less than ϵn^2 edges. Consider the following algorithm to test whether G is H-free. We sample k random vertices $v_1, v_2, \ldots, v_k \in V(G)$ (not necessarily distinct) and check if $v_i v_j \in E(G)$ whenever $ij \in E(H)$. If a copy of H is found, we report that G is ϵ-far from H-free. If C samples are performed and no copy of H is found, we report that G is H-free. Of course, there is a probability that this algorithm outputs an incorrect answer. The following theorem bounds this probability.

Theorem H (*Find homomorphism*). *There exists a constant C depending only on ϵ such that the algorithm above succeeds with probability at most $1 - 1/e$.*

Proof. If G is H-free, the algorithm always succeeds. So we are only concerned with the case where G is ϵ-far from H-free. The graph removal lemma says that in this case, there must exist at least δn^k copies of H in G for some δ depending on ϵ. Setting $C = 1/\delta$ and letting F be the event that the algorithm fails, we have

$$\mathbf{P}\{F\} \leq \left(1 - \frac{\delta n^k}{n^k}\right)^C \leq (1-\delta)^{1/\delta} \leq \frac{1}{e}.$$

Thus the algorithm succeeds with probability at most $1 - 1/e$. ∎

So one can test whether graphs G are H-free with arbitrary precision in $O(1)$ time, i.e., not depending on the size of G. For example, if we want a probability of error $< 1\%$, we only need to perform $5C$ tests no matter how large G is. Of course, since δ was given by the regularity lemma, the constant C is astronomically large depending on ϵ.

References

Paul Erdős, Péter Frankl, and Vojtěch Rödl, "The asymptotic number of graphs not containing a fixed subgraph and a problem for hypergraphs having no exponent," *Graphs and Combinatorics* **2** (1986), 113–121.

William Timothy Gowers, "Lower bounds of tower type for Szemerédi's uniformity lemma," *Geometric and Functional Analysis* **7** (1997), 322–397.

William Timothy Gowers, "Hypergraph regularity and the multidimensional Szemerédi theorem," *Annals of Mathematics* **166** (2007), 897–946.

Klaus Friedrich Roth, "On certain sets of integers," *Journal of the London Mathematical Society* **28** (1953), 104–109.

Imre Ruzsa and Endre Szemerédi, "Triple systems with no six points carrying three triangles," *Colloquia Mathematica Societatis János Bolyai* **18** (1978), 939–945.

Endre Szemerédi, "On sets of integers containing no k elements in arithmetic progression," *Acta Arithmetica* **27** (1975), 199–245.

Endre Szemerédi, "Regular partitions of graphs," *Colloques Internationaux C.N.R.S. n° 260—Problèmes Combinatoires et Théorie des Graphes, Orsay* (1978), 399–401.

Terence Tao and Van Ha Vu, *Additive Combinatorics* (Cambridge: Cambridge University Press, 2006).