The Mathematics of Endre Szemerédi

W.T. Gowers

1 Introduction

Endre Szemerédi is famous for his work in combinatorics and theoretical computer science. He has published a very large number of papers, often involving extraordinarily intricate arguments, so it will not be possible in an article such as this to do justice to either the breadth or the depth of his work. Instead, therefore, I shall describe a representative sample of his best known theorems, and attempt to convey in an informal way some of the ideas that go into their proofs. The sample will consist of the following results. (The dates given below and throughout the paper are the dates of publication rather than the dates that the results were actually proved.)

- In 1975, he proved that every dense subset of the natural numbers contains arbitrarily long arithmetic progressions, solving a famous and decades-old problem of Erdős and Turán.
- As part of the proof of the Erdős-Turán conjecture he formulated and proved a lemma, now known as Szemerédi's regularity lemma, that became a central tool in extremal graph theory and an inspiration for many other results in graph theory and beyond.
- In 1978, with Imre Ruzsa, he proved, using the regularity lemma, that every graph with few triangles can be approximated by a graph with no triangles. This innocent-looking result has been at the heart of many developments in graph theory, hypergraph theory, additive combinatorics, and computer science.
- In 1980, with Miklós Ajtai and János Komlós, he showed that the Ramsey number $R(3,k)$ is at most $Ck^2/\log k$.

Electronic supplementary material Supplementary material is available in the online version of this chapter at http://dx.doi.org/10.1007/978-3-642-39449-2_25. Videos can also be accessed at http://www.springerimages.com/videos/978-3-642-39449-2.

W.T. Gowers (✉)
Royal Society 2010 Anniversary Research Professor, Centre for Mathematical Sciences, Wilberforce Road, Cambridge, CB3 0WB, UK
e-mail: wtg10@dpmms.cam.ac.uk

H. Holden, R. Piene (eds.), *The Abel Prize 2008-2012*,
DOI 10.1007/978-3-642-39449-2_25,

- In 1982, with János Komlós and János Pintz, he found a counterexample to an old conjecture of Heilbronn in combinatorial geometry.
- In 1983, with Miklós Ajtai and János Komlós, he constructed a parallel sorting network that sorts n objects in $O(\log n)$ rounds.
- Also in 1983, with William Trotter, he proved a theorem about point-line incidences that has become one of the central results in combinatorial geometry.
- In 1995, with Jeff Kahn and János Komlós, he obtained the first exponentially small upper bound for the probability that a random ± 1 matrix is singular.

2 Szemerédi's Theorem

As its name suggests, Szemerédi's most famous theorem is ... Szemerédi's theorem [26], which states the following.

Theorem 1 *For every positive integer k and every $\delta > 0$ there exists N such that every subset $A \subset \{1, \dots, N\}$ of size at least δN contains an arithmetic progression of length k.*

This result was first conjectured in 1936, by Erdős and Turán [12], so by the time Szemerédi proved it in 1975 it had been open for almost four decades. Since then, a number of other proofs have been discovered, but they have been discussed in several other places, and here it seems more appropriate to talk about Szemerédi's own proof. Unfortunately, his proof is long and extremely intricate, so it is out of the question to present it here, and difficult even to give an overview. However, it is significantly easier in the case $k = 3$, so I shall begin by describing the argument in that case.

2.0.1 Density Increment Strategies Common to many proofs of Szemerédi's theorem, and indeed of several other results in extremal combinatorics, is the so-called *density increment strategy*. If X is a combinatorial structure and A is a subset of X, then the *density* of A is $|A|/|X|$. Suppose now that X has many substructures that are very similar to X itself and we want to prove that if the density of A is at least δ then A has some property P. Suppose also that if any subset of A has property P then A itself has that property. (In particular, this is true when the property is of the form "contains a configuration of type T".) Then instead of aiming directly for the property P we can instead try to prove one of the following *two* statements.

1. A has property P.
2. There exists a substructure Y such that the density of $A \cap Y$ in Y is at least $\delta + c(\delta)$.

Here $c(\delta)$ is a positive constant that depends on δ and increases as δ increases. (One could get away with a weaker condition, but in practice this one always holds.) If the substructure Y is sufficiently similar to X, then it too will have plenty of substructures, so in the second case we can apply the argument again. The density

cannot continue increasing for ever, so eventually, as long as the initial structure X is large enough, we obtain a subset of A with property P, which shows that A has property P.

Another way of looking at the density increment strategy is that its existence allows us to add an extra assumption about the set A: that it has density δ in X, and density at most $\delta + c(\delta)$ in every substructure Y of X. (The proof: if not, then we can pass to Y and we have a bigger density to work with; this cannot go on for ever.) This is extremely useful, because it allows us to assume that A has the kind of homogeneity that is usually associated with random sets, and random sets behave very nicely. This very rough idea is at the heart of all proofs of Szemerédi's theorem, but turning it into a rigorous proof is not at all easy.

2.1 Sketch Proof of Szemerédi's Theorem when $k = 3$

Our aim in this subsection is to present Szemerédi's proof of the following result, which is the first non-trivial case of Theorem 1.

Theorem 2 *For every $\delta > 0$ there exists N such that every subset $A \subset \{1, \dots, N\}$ of cardinality at least δN contains an arithmetic progression of length* 3.

This result was first proved by Klaus Roth [20] in 1953. Roth used Fourier analysis, while Szemerédi's argument is, as we shall see, purely combinatorial.

We shall apply the density increment strategy. In the context of this problem, there is a very natural collection of substructures of $\{1, 2, \dots, N\}$, namely the set of all arithmetic progressions that are subsets of $\{1, 2, \dots, N\}$. So by the discussion above, we are free to assume that A has density δ in $\{1, 2, \dots, N\}$ and density at most $\delta + c(\delta)$ in every arithmetic progression $Y \subset \{1, \dots, N\}$ of length at least m. What matters here is that m should tend to infinity with N, since then we can ensure that m is as large as we like by choosing N sufficiently large. We shall choose $c(\delta)$ and m later in the argument.

Let us now introduce a second idea that appears in many arguments that use a density increment strategy. Suppose you want to show that A intersects a substructure Y with density at least $\delta + c(\delta)$. Often it is not obvious how to find such a substructure in one step, but it is much clearer how to show that A intersects some kind of "nice" set W with density at least $\delta + 2c(\delta)$ (say). In principle, that allows one to obtain a density increment in two stages. In the first stage, one obtains the "nice" set W such that $|A \cap W| \geq (\delta + 2c(\delta))|W|$. In the second stage, one shows that W can be partitioned into large substructures $Y_1, \dots, Y_r$. By averaging, one then deduces that there exists i such that $|A \cap Y_i| \geq (\delta + 2c(\delta))|Y_i|$ and one has a density increment.

Sometimes, asking for a partition is too much, but one can get away with a slightly weaker assertion. This is where the factor 2 comes in. It is enough if *almost* all of W can be partitioned into large substructures $Y_1, \dots, Y_r$: if $|W \setminus (Y_1 \cup$

$\cdots \cup Y_r)| \leq c(\delta)|W|$, then $|A \cap (Y_1 \cup \cdots \cup Y_r)| \geq (\delta + c(\delta))|Y_1 \cup \cdots \cup Y_r|$ and the averaging argument works again.

Another way of weakening the requirement that the substructures Y_i partition W is to ask instead that they form a *uniform covering* of W: that is, every element of W is contained in the same number of sets Y_i. It is easy to see that the averaging argument still works. And one can weaken that to an approximately uniform covering. However, approximate partitions will suffice here.

2.1.1 A Strategy for Obtaining a "Nice" Set

To see where a "nice" set W might come from in our case, we shall make a simple observation, but we need to set the scene first.

Let θ be a smallish positive absolute constant, and let us divide the interval $\{1, \ldots, N\}$ into three parts: the integers up to $(1/2 - \theta)N$, the integers between $(1/2 - \theta)N$ and $(1/2 + \theta)N$ and the integers between $(1/2 + \theta)N$ and N. That is, we split $\{1, \ldots, N\}$ up into a smallish interval around $N/2$ and the intervals on either side of it. Let us refer to these intervals as the left interval, the middle interval and the right interval, and write them as L, M and R. Let us also write A_L, A_M and A_R for $A \cap L$, $A \cap M$ and $A \cap R$.

The density-increment strategy allows us to assume that the density of A in each of L, M and R is at most $\delta + c(\delta)$, which implies, by an easy averaging argument, that it is approximately *equal* to δ in each of the three subintervals.

Now if we are given any subset B of A_M, then A has an empty intersection with $2.B - A_L$ (which is defined to be $\{2y - x : x \in A_L, y \in B\}$). That is because the numbers x and y belong to A and the triple $(x, y, 2y - x)$ is an arithmetic progression. (We are of course assuming that A does not contain an arithmetic progression of length 3.) Thus, A_R is disjoint from $2.B - A_L$.

The set $2.B - A_L$ is not just any old set: it is a sumset of two large and very homogeneous sets, and as such has a highly atypical structure. Could that structure allow us to partition its complement into long arithmetic progressions?

The answer is not immediately obvious, so let us try a simple-minded approach, picking a positive integer d and partitioning the complement of $2.B - A_L$ into maximal arithmetic progressions of common difference d. It turns out to be easy to characterize how many of these progressions there are, since if x is the minimal element of such an arithmetic progression, we know that $x \notin 2.B - A_L$ and $x - d \in 2.B - A_L$. That is, $x \in (2.B - A_L + d) \setminus (2.B - A_L)$.

This gives us a proof strategy. Suppose we can find a subset $B \subset A_M$ and a positive integer d such that $(2.B - A_L + d) \setminus (2.B - A_L)$ has cardinality $o(N)$. Then we can partition the complement of $2.B - A_L$ into $o(N)$ arithmetic progressions with common difference d. Since this complement contains A_R, which has density at least $\delta/4$, the average length of these progressions tends to infinity. By an easy averaging argument, we can throw away $o(N)$ points and partition the rest of the complement of $2.B - A_L$ into long arithmetic progressions. Since $2.B - A_L$ has density at least $\delta/4$ (because A_L does, and $2.B$ is non-empty) and A_R has density

roughly δ in R, the average density of A inside these arithmetic progressions is at least $(1 + c\delta)\delta = \delta + c\delta^2$ for some absolute constant c, and we have our density increment on a long arithmetic progression.

2.1.2 Implementing the Strategy

It remains to find a set B and a positive integer d such that $(2.B - A_L + d) \setminus (2.B - A_L)$ has cardinality $o(N)$. For this we use the simple observation that if $B \subset B'$ then $2.B - A_L \subset 2.B' - A_L$. This again leads to a proof strategy. Suppose we can find a sequence of subsets $B_0, B_1, \dots, B_k$ of A_M such that each B_i is of the form $B_{i-1} \cup (B_{i-1} + d_i)$ for some positive integer d_i. Then the B_i are nested, so the sets $2.B_i - A_L$ are nested, from which it follows that there exists i such that $(2.B_i - A_L) \setminus (2.B_{i-1} - A_L)$ has cardinality at most N/k. But $2.B_i - A_L = (2.B_{i-1} - A_L + 2d_i) \cup (2.B_{i-1} - A_L)$, so

$$(2.B_i - A_L) \setminus (2.B_{i-1} - A_L) = (2.B_{i-1} - A_L + 2d_i) \setminus (2.B_{i-1} - A_L),$$

which is a set of the form that we would like to have cardinality $o(N)$. Therefore, our proof will be complete if we can get k to tend to infinity.

We have now reduced the problem to the task of finding a large *Hilbert cube* inside A_M: that is, a set of the form $\{x + \sum_{j=1}^{k} \varepsilon_j d_j : \varepsilon_j \in \{0, 1\}\}$. If we can find that, then we can set $B_i = \{x + \sum_{j=1}^{i} \varepsilon_j d_j : \varepsilon_j \in \{0, 1\}\}$ and we will have $B_i = B_{i-1} \cup (B_{i-1} + d_i)$ as desired.

Claim *Let I be an interval of integers of cardinality m and let E be a subset of I of density η. Then E contains a Hilbert cube of dimension at least $c \log\log m$.*

The proof is well known and very simple. There are $\eta m(\eta m - 1) \approx \eta^2 m^2$ pairs (x, y) of distinct elements of E and at most $2m$ possible differences, so at least one difference d_1 occurs at least $\eta^2 m/2$ times (up to a tiny error). Let $E_1 = E \cap (E - d_1)$. Then E_1 has cardinality at least $\eta^2 m/2$ and $E_1 \cup (E_1 + d_1) \subset E$. We repeat this observation for E_1 to find a subset E_2 and a positive integer d_2 such that $E_2 \cup (E_2 + d_2) \subset E_1$, and so on. At each stage of the iteration, we square the density of the set, and the iteration continues for as long as we still have a density of at least m^{-1}. From this we obtain a cube of dimension k provided that $(1/\eta)^{2^k} \leq m$, which gives us the bound of $c \log\log m$.

With the claim established, the proof of the theorem is complete.

2.2 What Happens when the Progressions Are Longer?

The proof just sketched contains the germs of various ideas that appear in the proof of the general case of Szemerédi's theorem. However, the argument for longer progressions is *much* more difficult and complicated. There is a simple reason for this,

which can be summarized in the form of a slogan: *it is very easy to find arithmetic progressions of length 2*. In the argument above, we made use of the fact that if a set A contains no arithmetic progressions of length 3, and if B and C are subsets of A, then A is disjoint from the set $2C - B$. That is, every pair $(b, c) \in B \times C$ gives us a number $2c - b$ that is not allowed to belong to A (as long as $b \neq c$).

If we want to try to do something similar for progressions of length 4, then we will find ourselves considering *three* sets, B, C and D. But now a triple $(b, c, d) \in B \times C \times D$ does not usually yield for us an element that cannot belong to A: it does so only if b, c and d lie in an arithmetic progression (in which case the next term in that progression is not allowed to belong to A). As just one example of the kind of difficulty this can cause, suppose we tried to imitate the proof for progressions of length 3 as follows: split the interval $\{1, \dots, N\}$ into four subintervals, let $A_1, \dots, A_4$ be the intersections of A with those intervals, find structured subsets S_2 and S_3 of A_2 and A_3 (to play the role of the Hilbert cube), and then use points in A_1 together with the sets S_2 and S_3 to find many points that cannot belong to A_4. Whereas in the argument for progressions of length 3, every point in A_L ruled out many points from A_R, now a point in A_1 would not rule out any points at all from A_4 unless it belonged to the set $2.S_2 - S_3$. If S_2 and S_3 are very small sets (as the Hilbert cube was in the argument for progressions of length 3), then there is no reason to suppose even that the set $A_1 \cap (2.S_2 - S_3)$ is non-empty. So for an approach like this to have any chance of working, the construction of the sets S_2 and S_3 would have to depend in an essential way on the sets A_1 and A_4—by contrast with the construction of the Hilbert cube in the earlier argument, which depended on A_M only.

We shall have a little more to say about the proof for longer progressions at the end of the next section.

3 Szemerédi's Regularity Lemma

In 1947, Erdős gave a remarkably simple proof [10] that the Ramsey number $R(k, k)$ is at least $2^{k/2}$. The proof can be summarized in a single sentence: if you take the complete graph on $2^{k/2}$ vertices and randomly colour its edges with two colours, then the expected number of monochromatic cliques of size k is less than 1. This proof gave birth to the subject of random graphs, and to the realization that random graphs are in many ways easy to understand. In particular, if G is a graph with n vertices and each pair of vertices xy forms an edge with probability p, with all these events being independent, then two things happen with high probability.

1. For every large set X of vertices the density of the induced subgraph with vertex set X is approximately p.
2. If v is a fixed constant and H is a graph with v vertices and e edges, then the number of copies of H in G is approximately $p^e(1-p)^{\binom{v}{2}-e}n^v$.

To be clear, a *copy* of H means a function ϕ from the vertex set of H to the vertex set of G such that $\phi(x)\phi(y)$ is an edge of G if and only if xy is an edge of H.

3.1 Quasirandom Graphs and the Counting Lemma

Approximately 40 years later it was realized, as a result of work of Thomason [31] and Chung, Graham and Wilson [8], that the two properties above are *equivalent*. This leads to the extremely useful notion of a *quasirandom* graph, which is a graph with one, and hence both, of the above properties.

There are many further respects in which quasirandom graphs behave like random graphs. A particularly important one is an example of a *counting lemma*, for which we need the closely related notion of a quasirandom *bipartite* graph. A bipartite graph G of density p with vertex sets X and Y of sizes m and n is quasirandom if it has one of the following two properties, which again turn out to be equivalent.

1. For every large pair of subsets $X' \subset X$ and $Y' \subset Y$ the density of the induced bipartite subgraph with vertex sets X' and Y' is approximately p.
2. If v and w are fixed constants and H is a bipartite graph with vertex sets of sizes v and w vertices and e edges, then the number of copies of H in G with the vertex set of size v in X and the vertex set of size w in Y is approximately $p^e(1-p)^{vw-e}m^v n^w$.

The counting lemma is the following statement.

Lemma 1 *Let G be a k-partite graph with vertex sets $V_1, \dots, V_k$ such that V_i has cardinality n_i for each i and such that for each i, j the bipartite graph that joins V_i to V_j is quasirandom with density α_{ij}. Let H be a graph with vertex set $\{1, 2, \dots, k\}$. Then the number of copies $\phi : H \to G$ of H in G such that $\phi(i) \in V_i$ for each i is approximately $n_1 \dots n_k \prod_{ij \in E(H)} \alpha_{ij} \prod_{ij \notin E(H)} (1 - \alpha_{ij})$.*

The counting lemma has a very intuitive interpretation. Imagine that the edges between V_i and V_j are put in with probability α_{ij} and that we randomly pick a vertex v_i from each V_i. Then the probability that the vertices $v_1, \dots, v_k$ span a copy of H, in the sense that $v_i v_j$ is an edge of G if and only if ij is an edge of H, will be $\prod_{ij \in E(H)} \alpha_{ij} \prod_{ij \notin E(H)} (1 - \alpha_{ij})$. The counting lemma tells us that the same is true if the edges between each V_i and V_j form *quasirandom* bipartite graphs rather than random ones. Thus, when those graphs are quasirandom, the number of copies of H is "what one would expect".

3.2 Statement of the Regularity Lemma

The counting lemma concerns graphs with vertex sets that can be partitioned into a small number of sets in such a way that the edges between each pair form quasirandom bipartite graphs. This may seem like a rather artificially strong condition to impose on a graph. Remarkably, it is not strong at all: to oversimplify slightly, Szemerédi's regularity lemma tells us that *every* dense graph is of this apparently special form.

A more precise statement of the lemma is as follows. Given any two sets U, V of vertices in a graph, let $e(U, V)$ denote the number of pairs $(u, v) \in U \times V$ such that uv is an edge of G, and define the *density* $d(U, V)$ to be $e(U, V)/|U||V|$. We define a bipartite graph G with vertex sets X and Y and density p to be ε-*regular* if $|d(U, V) - p| \leq \varepsilon$ whenever $U \subset X$, $V \subset Y$, $|U| \geq \varepsilon|X|$ and $|V| \geq \varepsilon|Y|$.

Theorem 3 *For every $\varepsilon > 0$ there exists a positive integer K with the following property. For every finite graph G there is a partition of its vertex set into subsets $V_1, \ldots, V_k$ with sizes differing by at most 1, such that $k \leq K$ and such that for all but at most εk^2 pairs (i, j) the bipartite subgraph of G induced by V_i and V_j is ε-regular.*

The slight oversimplification alluded to earlier was that I implied that *all* the pairs were quasirandom, whereas the correct statement is that they are *almost* all quasirandom. However, this does not matter too much: for example, the counting lemma remains true if a few of the bipartite graphs are not quasirandom, since it is an approximate statement.

3.3 Sketch Proof of the Regularity Lemma

The regularity lemma has been intensively studied ever since it was originally formulated, and there are now several approaches to proving it. However, even Szemerédi's original approach is simple and conceptual, so that is the one I shall present here.

3.3.1 Energy Increment Strategies

The key idea is a cousin of the density increment strategy discussed earlier: it is what is nowadays often referred to as an *energy increment strategy*. Let G be a graph with vertex set V and let $V_1, \ldots, V_r$ be a partition of the vertex set of G. Let $|V_i| = \mu_i|V|$ for each i. Then the *mean square density* of the partition is $\sum_i \mu_i\mu_j d(V_i, V_j)^2$. This we can think of as a kind of "energy".

I have stated this definition without assuming that the V_i have approximately the same size. We shall need a further definition adapted to this context. Let B be the set of "bad" pairs: that is, pairs (i, j) such that the bipartite subgraph induced by V_i and V_j is not ε-regular. Let us say that the partition is ε-*regular* if $\sum\{\mu_i\mu_j : (i, j) \in B\} \leq \varepsilon$.

The energy increment strategy, like the density increment strategy, is a way of proving results without trying to do everything in one go. Here we shall prove that one of the following two statements is true.

- The partition $V_1 \cup \cdots \cup V_r$ is ε-regular.

- There is a refinement of the partition $V_1 \cup \cdots \cup V_r$ into at most $m = m(r)$ sets $W_1, \ldots, W_m$ such that the mean square density of the refined partition is greater than the mean square density of the original partition by at least $c(\varepsilon)$.

If we can prove something like that, then we are clearly done: we cannot keep increasing the mean square density for ever (as with the density increment strategy, we are assuming that $c(\varepsilon)$ is an increasing function of ε), so after a certain number of refinements we end up with an ε-regular partition. The size of this partition will be bounded above by a function obtained by iterating the function m some number of times that depends on ε only.

3.3.2 Implementing an Energy Increment Strategy

To get the energy increment strategy to work here, we need two simple lemmas.

Lemma 2 *Let G be a graph with vertex set V and let $\mathcal{P} = \{V_1, \ldots, V_k\}$ be a partition of V. For each i let $V_{i1}, \ldots, V_{ik_i}$ be a partition of V_i and let $\mathcal{Q}$ be the partition of V into the sets V_{ij}. Then the mean square density of G with respect to $\mathcal{Q}$ is at least as big as the mean square density of G with respect to $\mathcal{P}$.*

Proof This lemma can be proved by a direct calculation using the Cauchy-Schwarz inequality. To see in a more conceptual way why it is true, let H be the Hilbert space of real-valued functions defined on $V \times V$ with the norm $\|f\| = (\mathbb{E}_{x,y} f(x,y)^2)^{1/2}$. (Here, we write $\mathbb{E}_{x,y}$ as shorthand for $|V|^{-2} \sum_{x,y}$.) Let $P : H \to H$ be the averaging projection with respect to the partition of $V \times V$ into the sets $V_i \times V_j$. That is, if $(x,y) \in V_i \times V_j$, then $Pf(x,y) = \mathbb{E}_{u \in V_i, v \in V_j} f(u,v)$. Similarly, let Q be the averaging projection that averages over the sets $V_{ij} \times V_{rs}$. Then P and Q are orthogonal projections, and $PQ = P$. Also, if f is the characteristic function of the graph G, then $\|Pf\|^2$ and $\|Qf\|^2$ are the mean square densities of G with respect to the partitions $\mathcal{P}$ and $\mathcal{Q}$, respectively. But $\|Pf\|^2 = \|PQf\|^2 \leq \|Qf\|^2$, so the lemma is proved. □

For the next lemma we need to adapt the notion of mean square density to bipartite graphs. If we have a bipartite graph G with vertex sets X and Y and we have partitions $X = X_1 \cup \cdots \cup X_r$ and $Y = Y_1 \cup \cdots \cup Y_s$, we say that the mean square density with respect to the two partitions is $\sum_{i,j} \mu_i \nu_j d(X_i, Y_j)^2$, where $\mu_i = |X_i|/|X|$ and $\nu_j = |Y_j|/|Y|$. For the next lemma it will be useful to interpret the mean square density probabilistically. Let D be the random variable that takes a random $(x,y) \in X \times Y$ to the density $d(X_i, X_j)$ for the unique pair (i,j) such that $x \in X_i$ and $y \in Y_j$. Then the mean square density is simply $\mathbb{E}D^2$.

Lemma 3 *Let G be a bipartite graph with vertex sets X and Y and density p. Suppose that there are subsets $X_0 \subset X$ and $Y_0 \subset Y$ such that $|X_0| \geq \varepsilon |X|$, $|Y_0| \geq \varepsilon |Y|$ and $|d(X_0, Y_0) - p| \geq \varepsilon$. Let $X_1 = X \setminus X_0$ and $Y_1 = Y \setminus Y_0$. Then the mean square density of G with respect to the partitions $X = X_0 \cup X_1$ and $Y = Y_0 \cup Y_1$ is at least $p^2 + \varepsilon^4$.*

Proof Again this lemma can be proved by direct calculation. However, it is nicer to use a probabilistic argument. As we have just commented, the mean square density is $\mathbb{E}D^2$, where D is the random variable that tells you the density of G in the pair $X_i \times X_j$ that your point lies in. Now $\mathbb{E}D = p$, so $\mathbb{E}D^2 = p^2 + \mathrm{Var}(D)$. Since the probability that a random point $(x, y) \in X \times Y$ lies in $X_0 \times Y_0$ is at least ε^2, the hypotheses of the lemma imply that the variance of D is at least $\varepsilon^2.\varepsilon^2 = \varepsilon^4$, which proves the lemma. □

We shall now be a little more sketchy. Recall that we are supposing that we have a partition $V = V_1 \cup \cdots \cup V_r$ that is not ε-regular, and we would like to find a refinement that has a slightly larger mean square density. Again, let $\mu_i = |V_i|/|V|$ and let B be the set of all pairs (i, j) such that the pair (V_i, V_j) is not ε-regular. Let us write $G(V_i, V_j)$ for the bipartite graph we obtain if we restrict G to $V_i \times V_j$.

By Lemma 3, for each pair $(i, j) \in B$, we can find partitions $V_i = V_{ij}^0 \cup V_{ij}^1$ and $V_j = V_{ji}^0 \cup V_{ji}^1$ such that the mean square density of $G(V_i, V_j)$ with respect to these two partitions is at least $d(V_i, V_j)^2 + \varepsilon^4$.

For each i, we now pick a common refinement of all the partitions $V_i = V_{ji}^0 \cup V_{ji}^1$. We can do this with $s_i \leq 2^r$ sets $V_{i1}, \ldots, V_{is_i}$. Then by Lemma 2, for each $(i, j) \in B$, the mean square density of $G(V_i, V_j)$ with respect to the partitions $V_i = V_{i1} \cup \cdots \cup V_{is_i}$ and $V_j = V_{j1} \cup \cdots \cup V_{js_j}$ is at least $d(V_i, V_j)^2 + \varepsilon^4$. For all other pairs (i, j), Lemma 2 implies that the mean square density is at least $d(V_i, V_j)^2$. Since a random pair $(x, y) \in V^2$ has a probability at least ε of belonging to $V_i \times V_j$ for some $(i, j) \in B$, this implies that the mean square density of G with respect to the partition into the sets V_{ij} exceeds the mean square desnity of G with respect to $V_1, \ldots, V_r$ by at least ε^5.

This completes the energy increment strategy and shows that we can find an ε-regular partition into at most $K = K(\varepsilon)$ sets, where K is the function obtained by starting with 1 and iterating the function $r \mapsto r.2^r\ \varepsilon^{-5}$ times. In other words, K has a tower-type dependence on ε.

When we stated the regularity lemma earlier we included an extra condition that said that the sets V_i in the partition had roughly equal size. For most applications this is not necessary, but it can be quite convenient. To obtain it, one runs the above argument but at each iteration one approximates the partition one has obtained by one in which the sets all have roughly equal size. We omit the details.

3.4 The Regularity Lemma and Szemerédi's Theorem

For an excellent account of why the regularity lemma was useful to Szemerédi for proving his theorem on arithmetic progressions, I recommend a blog post on the topic by Terence Tao [28]. Here I shall attempt to convey the idea very briefly—the arguments are explained in more detail in the blog post.

Let A be a subset of $\{1, 2, \ldots, N\}$. Recall from the discussion of Szemerédi's theorem that the density increment strategy allows us to assume that $|A \cap P| \leq (\delta +$

$c(\delta))|P|$ whenever P is an arithmetic progression of length m, provided only that m tends to infinity with N. Therefore, by averaging it follows that $|A \cap P| \approx \delta|P|$ for almost all such progressions P. The same is true of more general sets that are made out of arithmetic progressions, such as sets of the form $P_1 + \cdots + P_k$ where each P_i is an arithmetic progression. (A set of this kind is called a *k-dimensional* arithmetic progression.) Let us refer loosely to sets for which this kind of conclusion follows as "structured sets".

If we now take a structured set P and look at a set of translates $P, P + r, P + 2r, \ldots, P + (M-1)r$, we can define a sequence of subsets $A_0, A_1, \ldots, A_{M-1}$ of P by setting $A_i = A \cap (P + ir) - ir = \{x \in P : x + ir \in A\}$. Not only do we expect almost all the A_i to have density roughly δ, but there are also interesting relationships *between* the A_i. To give a simple example, no element $x \in P$ can belong to more than $(\delta + c(\delta))M$ of the sets A_i, since otherwise more than $(\delta + c(\delta))M$ of the elements of the arithmetic progression $x, x + r, \ldots, x + (M-1)r$ would belong to A. It follows by averaging that if E is any subset of P, then there exists i such that $|A_i \cap E| \leq (\delta + c(\delta))|E|$.

With the help of van der Waerden's theorem, one can improve this result to one about several sets $E_1, \ldots, E_m$. Suppose that we cannot find i such that $|A_i \cap E_j| \leq (\delta + c(\delta))|E_j|$ for every j. Then we can m-colour the set $\{0, 1, \ldots, M-1\}$ by taking the colour of i to be some j such that $|A_i \cap E_j| > (\delta + c(\delta))|E_j|$. But then van der Waerden's theorem gives us a long arithmetic progression of numbers i for which we can take the same j, and that, by a small modification of the remarks in the previous paragraph, cannot happen.

If we apply this result not just to the E_j but also to their complements, then we may conclude further that $|A_i \cap E_j| \approx \delta|E_j|$ for every j. (By "$\approx$" here, we mean that the difference between the two sides is at most a small multiple of $|P|$, so if E_j is a very small set, then it tells us nothing.)

Unfortunately, m is so small compared with M that this observation is not very helpful on its own. It is here that the regularity lemma comes in. Let $E_1, \ldots, E_N$ be a collection of subsets of P, where N may be arbitrarily large. We shall use the regularity lemma to find i such that $|A_i \cap E_j| \approx \delta|E_j|$ for *almost* every j. Thus, we have made a small loss—having to change from "every" to "almost every"—but have also made a big gain—going from m sets, where m is much less than M, to N sets, where N can be as large as we like.

Whenever one has a collection of subsets of a finite set, one can think of it as a bipartite graph in which the subsets are neighbourhoods. Here we take the vertex sets as P and $\{1, \ldots, N\}$, joining $x \in P$ to j if and only if $x \in E_j$. (Thus, $E_j \subset P$ is the neighbourhood of j.) Let us apply the regularity lemma to this graph, obtaining partitions of P and $\{1, \ldots, N\}$ into a bounded number of sets. Let the partition of P be $P_1 \cup \cdots \cup P_m$. Then we can apply the earlier result to obtain i such that $|A_i \cap P_s| \approx \delta|P_s|$ for every s.

Now if (U, V) is a regular pair of density α and $A \subset U$, then the regularity condition implies that for almost every $v \in V$ the neighbourhood of v in U intersects A in a set of size approximately $\alpha|A|$. Since almost all pairs are regular after we have applied the regularity lemma, for most P_s we can conclude that $|A_i \cap P_s \cap E_j| \approx$

$\delta|E_j \cap P_s|$ for almost every j. Summing over all s, it follows that $|A_i \cap E_j| \approx \delta|E_j|$ for almost every j, as claimed.

Of course, it is far from obvious how these ideas lead to a proof of Szemerédi's theorem, but that is beyond the scope of this article. At least the above argument makes it plausible that the regularity lemma could be of use.

4 The Triangle Removal Lemma

In this section we present a beautiful result of Ruzsa and Szemerédi [22], which amongst other things gives us an alternative proof of Theorem 2 and makes full use of the regularity lemma.

Theorem 4 *For every $\varepsilon > 0$ there exists $\delta > 0$ such that if G is any graph with n vertices and at most δn^3 triangles, then one can remove a set of at most εn^2 edges from G and obtain a graph that is triangle free.*

In short: every graph with few triangles can be approximated by a graph with no triangles.

One might have thought that this result would either be false, or be true with a more or less trivial proof. However, it is neither: it is true with a non-trivial proof, and to determine even very roughly the correct dependence of δ on ε is still an important open problem. (The best known bound, due to Jacob Fox, is that δ can be taken to be $1/T(\log(1/\varepsilon))$, where T is a tower-type function. In the other direction, it is known that δ cannot be greater than $\exp(-\log(1/\varepsilon)^2)$, which is just a little bit worse than a power dependence.)

4.1 Sketch Proof of the Triangle Removal Lemma

The proof of the triangle removal lemma starts in a similar way to many applications of the regularity lemma. We carry out the following three steps.

1. Apply the regularity lemma to the graph G with a suitable parameter η, obtaining a partition $V(G) = V_1 \cup \cdots \cup V_k$ into sets of approximately equal size.
2. Remove from G all edges that belong to bipartite subgraphs $G(V_i, V_j)$ that are not η-regular.
3. Remove from G all edges that belong to bipartite subgraphs $G(V_i, V_j)$ of density less than θ, for some suitable parameter θ.

The result is to create a graph G' such that every bipartite subgraph $G'(V_i, V_j)$ that is non-empty has density at least θ and is η-regular. This is very useful, because it means that we can apply the counting lemma.

To see how this works for the triangle removal lemma, observe first that we have removed at most $(\eta + \theta)n^2$ edges from G. Also, since G' is a subgraph of G, we know that G' contains at most δn^3 triangles.

But when δ is sufficiently small, this second observation actually implies that G' contains *no* triangles. To see why, suppose that xyz is a triangle and let r, s and t be such that $x \in V_r$, $y \in V_s$ and $z \in V_t$. Then the three bipartite graphs $G'(V_r, V_s)$, $G'(V_s, V_t)$ and $G'(V_r, V_t)$ all contain at least one edge, from which it follows that they are all η-regular with density at least θ. But then the counting lemma implies (provided η is sufficiently small in terms of θ) that those three bipartite graphs when put together contain at least $(\theta^3/2)|V_r||V_s||V_t|$ triangles. But each of V_r, V_s and V_t has size approximately n/k, where k depends on η only. This is a contradiction when δ is small enough.

Chasing the parameters, we need $\eta + \theta \leq \varepsilon$ with η at most some power of θ. And then we can set $\delta = \theta^3/4K^3$, where $K = K(\eta)$ is the upper bound on k that comes from the regularity lemma. This gives the tower-type bound for $1/\delta$ in terms of $1/\varepsilon$.

4.2 Applications of the Triangle Removal Lemma

Ruzsa and Szemerédi noticed that the triangle removal lemma gave another proof of Roth's theorem (that is, Szemerédi's theorem for progressions of length 3). In this section we present a slight modification of their argument, observed by Jozsef Solymosi [24], that yields a stronger result.

4.2.1 The Corners Theorem

Theorem 5 *For every $\delta > 0$ there exists N such that every subset $A \subset \{1, \dots, N\}^2$ of cardinality at least δN^2 contains a triple $\{(x, y), (x + d, y), (x, y + d)\}$ with $d \neq 0$.*

Configurations of the form $\{(x, y), (x + d, y), (x, y + d)\}$ with $d \neq 0$ are sometimes called *corners*, and this result is sometimes referred to as the corners theorem.

To deduce the corners theorem from the triangle removal lemma, we need to construct a graph. This is done as follows. Let $X = Y = \{1, \dots, N\}$ and let $Z = \{1, \dots, 2N\}$. We construct a tripartite graph G with vertex sets X, Y and Z (regarding X and Y as disjoint copies of $\{1, \dots, N\}$ rather than as the same set) as follows.

1. $x \in X$ is joined to $y \in Y$ if and only if $(x, y) \in A$.
2. $x \in X$ is joined to $z \in Z$ if and only if $(x, z - x) \in A$.
3. $y \in Y$ is joined to $z \in Z$ if and only if $(z - y, y) \in A$.

If xyz forms a triangle in G, then we can set $d = z - x - y$, and A contains the three points $(x, y), (x, y + d)$ and $(x + d, y)$. Thus, triangles in G correspond to

corners in A. Or rather, they almost do, but if $x + y = z$ then they don't, since in that case $d = 0$ and the "corner" is just a single point.

We therefore can't quite deduce that G contains no triangles from the fact that A contains no corners. Surprisingly, it turns out that this is not a setback: it is of vital importance to the proof that there should be at least some triangles in G, as we shall see. What we *can* say is that there is a one-to-one correspondence between triangles in G and points in A. It follows that the number of triangles in G is at most N^2. Since the number of vertices in G is $4N$, the hypotheses of the triangle removal lemma are very strongly satisfied. It follows that we can remove $o(N^2)$ edges from G to form a triangle-free graph.

However, this is easily seen to be impossible. The triangles in G are edge disjoint, since if $x + y = z$ and $x' + y' = z'$ then any two of the equalities $x = x'$, $y = y'$ and $z = z'$ implies the third. Since there are at least δN^2 triangles, one must remove at least δN^2 edges from G to make it triangle free. This contradiction implies the corners theorem.

The above proof was not the first proof of the corners theorem: that was a result of Ajtai and Szemerédi [4] from 1975. Their proof naturally gave the slightly stronger result that we may take $d > 0$. However, as was observed by Ben Green, the two statements are equivalent, since one can begin by intersecting A with a random translate of $-A$ in order to obtain a dense subset B of A with the property that if it contains a corner with $d < 0$ then it must also contain a corner with $d > 0$. As with many of Szemerédi's proofs that are apparently superseded, the argument of Ajtai and Szemerédi has turned out to have unexpected importance, serving as a model for later arguments in situations where the regularity approach cannot easily be made to work.

4.2.2 Another Proof of Roth's Theorem

As suggested above, the corners theorem implies Roth's theorem. Here is the simple deduction. Let A be a subset of $\{1, \dots, N\}$ of density δ and let $A' \subset \{1, \dots, 2N\}^2$ consist of all points (x, y) such that $x - y \in A$. Then A' has density at least $\delta/4$, so by the corners theorem it contains a triple $\{(x, y), (x + d, y), (x, y + d)\}$. But then the three points $x - y - d, x - y$ and $x - y + d$ all lie in A and form an arithmetic progression.

4.2.3 Property Testing

There is considerable interest amongst theoretical computer scientists in algorithms that can test for properties of their input by making only a constant number of queries. Given that the input has size n, which tends to infinity, this might seem a hopeless task. However, one typically asks for approximate answers, and one wants them to be correct with high probability rather than total certainty.

A very simple example would be testing an input sequence of n 0s and 1s to see whether at least half of the bits are 1s. We cannot hope to do this with only constantly many queries, but what if we relax the requirement so that our aim is to output one of the following two statements and be *almost* certain that the statement we go for is true?

1. At least half the bits are 1s.
2. At most 51 % of the bits are 1s.

If we sample 10^8 bits at random, then the standard deviation of the number of 1s we get will be of order of magnitude 10^4, so with very high probability the proportion of 1s in our sample will differ from the true proportion by less than 0.5 %. Therefore, we can output the first statement if the proportion of 1s in our sample is at least 50.5 % and the second statement otherwise.

Now let us think about a more interesting problem. This time our input is a graph G with n vertices. What we would really like to determine is whether or not the graph contains a triangle, but we cannot hope to do that after looking at only a constant number of edges. However, what we can do is output, with confidence, one of the following two statements.

1. G contains a triangle.
2. G can be approximated by a triangle-free graph.

This follows directly from the triangle removal lemma. The algorithm is very simple indeed: we randomly sample a large but constant number of triples of vertices, seeing in each case whether we have the vertices of a triangle. If we ever do, then we output "G contains a triangle" and we are 100 % certain that is correct. If we never discover a triangle in our sample, then with high probability the proportion of triples in G that form triangles is very small. But then, by the triangle removal lemma, G can be approximated by a triangle-free graph.

This is typical of many property-testing results in that we either make one claim with complete certainty or the other one with near certainty.

A great deal is now known about properties that can be tested for in this way, and the regularity lemma is a central tool for proving such results.

5 A Sharp Upper Bound for the Ramsey Number $R(3, k)$

Ramsey's theorem states that for every pair of positive integers k and l, there exists a positive integer n such that every graph G with n vertices contains a clique of size k or an independent set of size l. (A *clique* is a set of vertices such that every pair of vertices in the set is joined by an edge. An *independent set* is the opposite: a set of vertices such that no two of them are joined by an edge.)

The Ramsey number $R(k, l)$ is the smallest n for which Ramsey's theorem is true. Unless k and l are small, it does not appear to be feasible to calculate Ramsey numbers exactly, so attention has turned to asymptotics. However, even these are

difficult to obtain with any accuracy. For example, the best known upper and lower bounds for $R(k,k)$ are roughly 2^{2k} and $2^{k/2}$, respectively, so the gap between them is exponential.

A simple argument shows that $R(3,k)$ is at most $k(k+1)/2$. Indeed, let G be a graph with n vertices. We would like to show that G contains either a triangle or an independent set of size k. Let us assume that G does not contain a triangle. This tells us that the neighbourhood of each vertex x (that is, the set of vertices joined to x) contains no edges.

We shall use this observation repeatedly to create an independent set $x_1,\dots,x_k$. Let x_1 be an arbitrary vertex of G, throw away x_1 and all its neighbours, and let V_1 be the set of all remaining vertices. Since the neighbours of x_1 form an independent set, either we are done or there are at most $k-1$ of them. In the second case, let x_2 be an arbitrary vertex in V_1, throw away x_2 and all its neighbours from V_1 and let V_2 be the set of all remaining vertices. Since the neighbours of x_2 form an independent set, which remains an independent set when x_1 is included, either we are done or there are at most $k-2$ of them. Continuing in this way, we end up finding an independent set provided that $n \geq k+(k-1)+\cdots+1 = k(k+1)/2$.

This bound was improved in 1968 by Graver and Yackel [13] to $Ck^2 \log\log k/\log k$. Then in a paper published in 1981 Ajtai, Komlós and Szemerédi [2] improved the bound to $Ck^2/\log k$. They subsequently found a simpler argument [1] that (slightly confusingly for the historian) was published in 1980. The 1981 paper remained important for two reasons: it made progress on another interesting problem, and it introduced the so-called *semirandom method* into combinatorics, which has become a major tool with many further applications. We shall say a little about semirandom methods in the next section, but here we give the simpler proof from the 1980 paper.

5.1 Choosing an Independent Set More Carefully

The basic strategy we presented above for proving the bound $R(3,k) \leq k(k+1)/2$ was to choose an independent set $\{x_1,\dots,x_k\}$ greedily, exploiting the fact that in a triangle-free graph with no independent set of size k, no vertex has degree more than $k-1$.

If we want to improve this argument, then a natural strategy is to be slightly less greedy. For example, perhaps we could try to find a vertex of degree less than $k-1$ so that we have fewer neighbours to worry about.

In general, that may not be possible, but if we look ahead a little further, then there is something else we can try to do, namely pick the vertices x_i in such a way that when we remove their neighbourhoods, we remove as many further edges as we can. That way, we can hope that as the selection proceeds, the average degree in the remaining graph goes down, which enables us to pick vertices with not too many neighbours.

The next lemma shows how to find a vertex whose removal will cause us to remove many edges. Let us define the *second degree* of a vertex x in a graph to be the

sum of the degrees of all the neighbours of x, and denote it by $d_2(x)$. Equivalently, it is the number of paths of length 2 that start at x (counting "paths" that begin and end at x). We also write $d(x)$ for the degree of x.

Lemma 4 *Let G be a graph with average degree t. Then there exists a vertex x such that $d_2(x) \geq td(x)$.*

Proof We show first that $\sum_x d_2(x) = \sum_x d(x)^2$. To see this, let A be the adjacency matrix of G. Then

$$\sum_x d_2(x) = \sum_{x,y,z} A(x,y)A(y,z) = \sum_y \sum_{x,z} A(y,x)A(y,z) = \sum_y \Big(\sum_x A(y,x)\Big)^2$$
$$= \sum_y d(y)^2$$

which is of course equal to $\sum_x d(x)^2$. (All we have done here is count the set of paths of length 2 in two different ways.)

It follows that $\mathbb{E}_x d_2(x) = \mathbb{E}_x d(x)^2$. Since the variance of the degrees is non-negative, $\mathbb{E}_x d(x)^2 \geq (\mathbb{E}_x d(x))^2 = t\mathbb{E}_x d(x)$. Therefore, there exists x such that $d_2(x) \geq td(x)$, as claimed. □

Theorem 6 *Let G be a triangle-free graph with n vertices and average degree t. Then G contains an independent set of size at least $n \log t/8t$.*

Proof By Lemma 4 we can find a vertex x such that $d_2(x) \geq td(x)$. If $d(x) > 4t$ then let us remove x from the graph, and otherwise let us remove x and all its neighbours from the graph.

In the second case, we remove $d(x) + 1$ vertices from the graph, and the sum of the degrees goes down by at least $2td(x)$. The latter bound follows from the fact that G contains no triangles, which means that no edge is joined to more than one neighbour of x.

In both cases, we can then choose the largest independent set in the remainder of the graph; in the second case we can add x to that independent set to get a larger independent set.

Now let us define a function $\phi : \mathbb{N}^2 \to \mathbb{N}$ as follows: $\phi(n,m)$ is the minimum size of the largest independent set that is contained in a triangle-free graph with n vertices and m edges. Our preliminary remarks have shown that

$$\phi(n,tn) \geq \min\Big\{\phi\big(n-1,t(n-4)\big), 1 + \min_{d \leq 4t} \phi\big(n-d-1,t(n-2d)\big)\Big\}.$$

Let us now prove by induction that $\phi(n,m) \geq n^2 m^{-1} \log(m/n)/8 = n \log(m/n)/8(m/n)$. Note that $2m/n$ is the average degree of the graph, so up to a constant a bound of $n/(m/n)$ is the simple-minded bound one would get if G was regular and one just removed an arbitrary vertex and its neighbours at each stage. The interest is in the logarithmic improvement.

A simple back-of-envelope calculation shows that

$$\frac{(n-1)^2}{t(n-4)}\log\left(t\frac{n-4}{n-1}\right)\geq\frac{n}{t}\log t$$

if $\log t\geq 6$.This proves the inductive step in the case that $\phi(n,tn)\geq\phi((n-1),t(n-4))$, provided that the average degree is not too small.

Another simple back-of-envelope calculation shows that

$$\frac{(n-d-1)^2}{8t(n-2d)}\log\left(t\frac{n-2d}{n-d-1}\right)+1\geq\frac{n}{8t}\log t$$

provided that $d\leq 4t$. Let us actually do this second calculation, since it is the important case—that is, the case that tells us what happens when we remove a vertex of roughly average degree and reasonably high second degree.

Since $(n-d-1)^2=n^2-2(d-1)n+(d-1)^2>n(n-2d)$, we can bound the first fraction on the left-hand side below by $n/8t$. Therefore, it remains to prove the inequality

$$\log t+\log\left(\frac{n-2d}{n-d-1}\right)+\frac{8t}{n}\geq\log t,$$

which is equivalent to the inequality

$$\log(1-2d/n)-\log\bigl(1-(d-1)/n\bigr)+8t/n\geq 0.$$

Using the approximation $\log(1+x)\approx x$ (and not being too careful about justifying it—let us assume that d/n is reasonably small and take on trust that the argument can be made completely rigorous) we need to show that $8t/n\geq(d+1)/n$, which is true since $d\leq 4t$. (The extra elbow room here compensates for the sloppiness before.)

We haven't quite finished, since it remains to discuss what happens if $\log t<6$. In this case, we use the fact that the logarithmic improvement is just a constant. To be efficient about it, we use Turán's theorem, which implies that the largest independent set in a graph of average degree t has size at least $1+(n-1)/t$. We need this to be at least $n\log t/8t$, which it is, since $\log t<6$. This completes the proof. □

Corollary 1 *The Ramsey number $R(3,k)$ is bounded above by $Ck^2/\log k$ for an absolute constant C.*

Proof The bound on $R(3,k)$ is equivalent to the assertion that a triangle-free graph G with n vertices contains an independent set of size at least $c\sqrt{n\log n}$ for an absolute constant c. This is certainly true if there is a vertex of degree at least $\sqrt{n\log n}$, since the neighbourhood of that vertex is an independent set. If not, then the average degree is at most $\sqrt{n\log n}$, and then Theorem 6 tells us that there is an independent set of size at least $c'n\log(\sqrt{n\log n})/\sqrt{n\log n}=c\sqrt{n\log n}$. □

In another famous result, Jeong Han Kim proved in 1995 a lower bound for $R(3,k)$ that matches the upper bound of Corollary 1 to within a constant [16]. Thus, the result of Ajtai, Komlós and Szemerédi was shown by Kim to be best possible.

6 A Counterexample to Heilbronn's Triangle Conjecture

Suppose that you take n points in a unit disc. Then any three of those points define a (possibly degenerate) triangle. How large can the area of the smallest of these triangles be? A trivial upper bound is Cn^{-1}: by the pigeonhole principle there must be three points that have x-coordinates equal to within $Cn^{-1}/2$, and the triangle defined by those three points then cannot have area greater than Cn^{-1}.

There is also a fairly simple lower bound of cn^{-2}, due to Erdős. For convenience let n be a prime p, and let X be the set of all points of the form $(x/p, y/p)$ such that $0 \leq x, y \leq p-1$ and $y \equiv x^2 \bmod p$. In other words, X is basically the graph of the function $x \mapsto x^2 \bmod p$. Now no three of these points lie in a line, since if they did, then they would also lie in a line mod p, and a quadratic function can equal a linear function in at most two places. Therefore, X contains no degenerate triangles. But the smallest possible area of a non-degenerate triangle with vertices in $\mathbb{Z}^2$ is $1/2$, so the smallest triangle with vertices in X has area at least $p^{-2}/2$.

Heilbronn's conjecture was that the lower bound of cn^{-2} was correct. The gap between n^{-2} and n^{-1} is embarrassingly large, and initial work of Roth and Schmidt brought it down only very slightly: Roth [19] obtained an upper bound of $Cn^{-1}(\log\log n)^{-1/2}$ in 1950, then Schmidt [23] reduced that to $Cn^{-1}(\log n)^{-1/2}$ in 1972. Also in 1972, Roth [21] eventually managed to obtain an improvement in the power of n, but to nowhere near n^{-2}.

In 1982 (the paper was received in 1980), Komlós, Pintz and Szemerédi disproved Heilbronn's conjecture by proving the following result [18].

Theorem 7 *It is possible to choose n points in the unit disc such that no three form a triangle of area less than* $cn^{-2}\log n$.

That is, they obtained a logarithmic improvement over Erdős's lower bound.

Of course, a logarithmic improvement is quite small, and one could respond by modifying the conjecture to say that the smallest triangle has area at most $n^{-2+\varepsilon}$. However, the proof was very interesting and influential.

One particularly interesting aspect of the argument was that it reduced a geometrical problem to a purely combinatorial one about hypergraphs. A *k-uniform hypergraph* is a set V of *vertices* and a set E of k-tuples of vertices. The k-tuples are called *hyperedges*, but they are often simply called *edges*. A 2-uniform hypergraph is just a graph in the normal sense.

An *independent set* in a k-uniform hypergraph is the obvious generalization of what it is for a graph: it is a set of vertices such that no k of them form an edge.

The basic strategy that Komlós, Pintz and Szemerédi used to obtain their lower bound for the Heilbronn problem was as follows.

- Begin by dropping $n^{1+\alpha}$ random points into the unit disc for some small constant $\alpha > 0$.
- Define a 3-uniform hypergraph H by taking the random points as its vertices and all triples of points that form triangles of area less than $cn^{-2}\log n$ as its edges.

- Show that with high probability H has certain combinatorial properties that show that it is "locally sparse".
- Deduce from these local sparseness properties that H contains an independent set of size n.

Since an independent set in H is a set of points in the unit disc such that no three form a triangle of area less than $cn^{-2}\log n$, this strategy, if it can be carried out, disproves the Heilbronn conjecture.

Before we discuss this strategy further, it is worth looking at an observation that Komlós, Pintz and Szemerédi make in their paper, which is that selecting points from a random set can be used to give a different proof of Erdős's lower bound. To see this, let us first consider the probability that three random points form a triangle of area less than a. If the distance between the first two points is r, then the third point needs to lie within a strip of width $4a/r$. The probability that the distance between the first two points lies between r and $r+\delta r$ is at most about $2\pi r\delta r$, so an upper bound for the probability that the three points form a triangle of area at most a is $\int_0^2 (4a/r)(2\pi r)\,dr = 16\pi a$.

Therefore, if we drop $2n$ random points into the unit disc, the expected number of triangles of area at most a is at most $16\pi a\binom{n}{3} \le 10an^3$. If we choose a to be $n^{-2}/10$, then this is at most n. Therefore, we can remove n points from the set and obtain a set of n points with no triangles of area less than $n^{-2}/10$.

Why should it be possible to gain anything over this simple approach if we choose, and then discard, more points? Let me quote from an article by Imre Bárány [5].

> According to his coauthors, Szemerédi's philosophy, that random subgraphs of a graph behave very regularly, and his vision that such a proof should work, proved decisive. Since then, the method has been applied several times and with great success.

Bárány was in fact referring to the first proof of Theorem 6 above. However, the results are closely connected: in both cases, there is some kind of sparseness condition that allows one to find a slightly larger independent set than one might naively think is possible. To make the connection clearer, let us look at the simple argument above in a slightly different way. Suppose we have m random points in the unit disc, forming a set S, and we want to choose as many of them as we can while avoiding a triangle of area a. The expected number of triangles of area a is, as we have already seen, at most $16\pi am^3$. Therefore, each point belongs, on average, to at most $16\pi am^2$ such triangles. So if a is small enough for $16\pi am^2$ to be substantially less than m, then we could imagine an algorithm that simply picks a random point $x \in S$, then throws away it and all other points $y \in S$ such that there is a point $z \in S$ for which the triangle xyz has area at most a. Typically, we will throw away at most $16\pi am^2$ points at each stage, so we can hope to obtain a subset of S with at least $m^{-1}a^{-1}/16\pi$ points.

Forgetting about the absolute constants, if $m^{-1}a^{-1} = n$ and $m \ge n$, then $a^{-1} \ge n^2$, and the larger m is, the worse a becomes. So at first it looks as though the strategy outlined above is doomed to fail. However, the argument we have just given

is quite clearly very inefficient: if xyz is a triangle of small area, there is no need to throw away both y and z: it is enough to throw away just one of them. To see why this is a big help when m is large, note that if x and y are very close, then the argument above requires us to throw away all points z in quite a wide strip about the line that joins x and y, when to avoid all those triangles of small area it would be enough just to throw away y.

Let us now see what the local sparseness properties are that enable one to choose a large independent set in a hypergraph. Once we have dropped $n^{1+\alpha}$ random points into the unit disc, the expected number of pairs of points within distance $n^{-2/3}$ is at most $n^{-4/3}n^{2+2\alpha} = n^{2/3+2\alpha}$. If α is small enough, this is substantially less than n, so we can discard a small fraction of the points and end up with no two of them closer than $n^{-2/3}$.

We have already seen how to estimate the number of edges in the hypergraph H. If $a = n^{-2}$, then it is at most $16\pi n^{1+3\alpha}$, so on average each vertex belongs to at most $16\pi n^{3\alpha}$ edges.

Define a 2-*cycle* in a 3-uniform hypergraph to be a pair of edges that intersect in a set of size 2, a *simple* 3-*cycle* to be a triple of edges of the form abx, bcy, acz, where all of a, b, c, x, y, z are distinct, and a *simple* 4-*cycle* to be a quadruple of edges of the form abx, bcy, cdz, adw, where again different letters stand for distinct vertices. In a similar way to the way we estimated the number of edges, one can show that if α is small enough, then the numbers of 2-cycles, simple 3-cycles and simple 4-cycles are all substantially less than n, so we can remove a small fraction of the vertices and obtain a hypergraph with no 2-cycles, simple 3-cycles or simple 4-cycles. Komlós, Pintz and Szemerédi called such a hypergraph *uncrowded* (though they say that the term was in fact invented by Joel Spencer—indeed, the phrasing in terms of hypergraphs was Spencer's idea as well).

The main result that Komlós, Pintz and Szemerédi proved was the following result about 3-uniform hypergraphs. Define the *degree* of a vertex to be the number of edges that contain that vertex.

Theorem 8 *Let G be an uncrowded 3-uniform hypergraph with n vertices and average degree d. Suppose that d is sufficiently large, and also at most $n^{1/20}$. Then G contains an independent set of size at least $c(n/d^{1/2})(\log d)^{1/2}$.*

We shall not prove this theorem here, but we can make a few remarks. First, note that there is an easy bound of $cn/d^{1/2}$, proved as follows. If you pick a triple at random, then the probability that it is an edge is proportional to d/n^2. Therefore, if you pick m vertices at random, then the expected number of edges that they span is Cm^3d/n^2. If this is less than $m/2$, then you can discard at most $m/2$ vertices and end up with an independent set. Solving for m we obtain the bound claimed. This is essentially the argument we used above to rederive the Erdős lower bound.

Easy examples show that this bound is best possible if we do not impose the uncrowdedness assumption. So what is that assumption doing for us?

To answer that question, note first that a very similar situation applied with Theorem 6. If G is a graph (that is, a 2-uniform hypergraph) with average degree t, then a random set of m vertices spans Ctm^2/n edges on average, and for this to be

less than $m/2$ we need $m = cn/t$. Theorem 6 improves this bound by a logarithmic factor under the additional assumption that G is triangle free, which is equivalent to saying that the neighbourhood of each vertex forms an independent set.

This suggests that we should look for a condition on hypergraphs that could play a similar role. The uncrowdedness assumption implies the following. Suppose we pick a vertex in an uncrowded hypergraph and throw away all vertices that belong to edges that contain x. If y and z are two such vertices, then we will lose all edges that contain either y or z. With the uncrowdedness assumption, the edges that contain y are all disjoint from the edges that contain z. That is because an edge that contains both y and z would form a 2-cycle or a simple 3-cycle, and if an edge containing y overlaps an edge containing z, then we would have either a 2-cycle or a simple 4-cycle.

There is, however, an important respect in which Theorem 8 differs from Theorem 6. In the case of graphs, each time we pick a vertex to go into our independent set, we must throw away all its neighbours. But with a 3-uniform hypergraph, if we pick a vertex x, then what we must ensure is that for every edge xyz we do not pick both of y and z. If we do this in a crude way by discarding all vertices that belong to an edge that contains x, then on average we throw away d vertices each time, and even if we can make some kind of logarithmic gain, we will end up with the wrong power of d in our final answer. In other words, a greedy algorithm, even if the hypergraph is very regular, gives a much worse bound than the simple random selection described earlier.

Very roughly, the strategy of the proof is this. Instead of choosing a single point at a time, one chooses small random sets of points to add to the independent set. If C is the set of points that have already been chosen, then it is necessary to discard every point z such that there exist $x, y \in C$ such that xyz is an edge. So each time a few more random points are added to C, one discards the points that need to be discarded and then chooses the next small random set from the points that remain.

At each stage s, if C_s is the set of points chosen so far and R_s is the set of points that remain, there is an important graph with vertex set R_s, and also an important hypergraph. The hypergraph is just the restriction of the original hypergraph G to R_s. The graph is the set of all pairs yz in R_s such that xyz is an edge of G for some $x \in C_s$. For the proof to work, it is vital that when we add some randomly chosen points from R_s to C_s to create the set C_{s+1} and pass to a new set R_{s+1}, the set R_{s+1} should resemble a randomly chosen subset of R_s, in the sense that the degrees in the graph and hypergraph should go down in roughly the expected way.

This kind of technique has become known as the *semirandom method*, and has been used to solve many problems in extremal combinatorics that had previously appeared to be hopelessly difficult.

7 An Optimal Parallel Sorting Network

A well-known mathematical problem is to minimize the number of pairwise comparisons needed to sort n objects that are linearly ordered. A simple argument shows

that $\log_2(n!) = cn \log n$ comparisons are necessary. Indeed, before we do any comparisons there are $n!$ possible orderings compatible with the information we have so far. But each time we do a comparison, there are two possible results, so in the worst case the number of compatible orderings is over half what it was before the comparison. This implies the bound stated.

It is also not very hard to match this lower bound with an upper bound of the same form, using a recursively defined algorithm known as Mergesort. Take your n objects and divide them into two groups of size $n/2$ (for convenience let us assume that n is a power of 2—it is easy to remove this condition afterwards). Apply Mergesort to each group (which we know how to do by induction). We now have two ordered groups A and B of $n/2$ points, which we "merge" into an ordering of all n points as follows. Let the elements of A be $a_1 < \cdots < a_m$ and let the elements of B be $b_1 < \cdots < b_m$. Then we compare a_1 with b_1, then b_2, and so on until we reach i such that $b_i < a_1 < b_{i+1}$. We then compare a_2 with b_{i+1}, b_{i+2} and so on until we find where we can slot in a_2. We keep going like this until the two sets have been fully merged. The number of comparisons we make when doing this process is at most $2m = n$, since the number $i + j$ increases each time we move to a new comparison between some a_i and some b_j. Therefore, if we define $f(k)$ to be the time that mergesort needs to sort 2^k objects, we have the recursion $f(k) \leq 2f(k-1) + 2^k$. We also know that $f(1) = 1$. It follows easily by induction that $f(k) \leq k.2^k$. Setting $n = 2^k$, we obtain a bound of $Cn \log n$. (If n is not a power of 2, we can add some dummy objects to bring the number up to the next power of 2.)

Given two bounds that are obtained by simple arguments and are equal up to a constant, one might think that there was little more to say. However, this is not the case. A general question of major importance in computer science is whether algorithms can be *parallelized*. That is, if you have a large number of processors (growing with the size of the problem), can you get the algorithm to run *much* faster?

Rather than discuss what parallel computation is in general, let us look at a simple model that is sufficient for understanding this problem. Imagine that we have n rocks that all look quite similar but that all have slightly different weights. Imagine also that we have a very accurate balance that will take at most one rock on each side. The sorting problem just discussed is equivalent to asking how many times we need to use the balance if we want to order the rocks by weight.

For the parallel sorting problem, we can have as many balances as we like. Let us assume that comparing two rocks takes some fixed time such as one minute. Then what we would like to minimize is the total time needed to determine the order of the rocks. Since we can do up to $n/2$ comparisons at the same time, and since $cn \log n$ comparisons are needed, we will need to take at least $c \log n$ minutes. But can we achieve a growth rate that is anything like as small as logarithmic?

Before I answer that question, I need to mention another word from the title of this section. I have been discussing the word "parallel" but have not yet paid any attention to the word "network", which is also a critical part of what Ajtai, Komlós and Szemerédi did. The idea here is that we decide in advance all the comparisons we are going to do.

More formally, a *comparator network* is a sequence of partitions of $\{1,\dots,n\}$ into $n/2$ pairs. Given a comparator network, we define a sorting algorithm as follows. At the ith stage, we use the ith partition to decide which rocks to compare: if r and s are paired, then we take the rocks in the rth and sth places, compare them, and put them back in the two places they came from, but switching them round if necessary so that the heavier rock is to the right of the lighter one. The *depth* of a comparator network is just the length of the sequence of partitions. If the network correctly sorts every permutation of the rocks, then it is a *sorting network*.

It is initially somewhat counterintuitive that efficient sorting networks exist, since the comparisons that are made do not depend at all on the results of earlier comparisions (which is quite unlike the behaviour of Mergesort). However, in 1968, Batcher [6] constructed a relatively simple sorting network of depth $C(\log n)^2$. Here, briefly, is how it works.

First, he shows inductively that merging two increasing sequences of length 2^{k-1} can be done with a comparator network of depth k. The idea is straightforward. The odd terms of the sequences form two increasing sequences of length 2^{k-2}, so by induction they can be merged with a network of depth $k-1$. In parallel, one can merge the even terms. This now gives a sequence such that the odd terms are in the right order and the even terms are as well. But it is not hard to check that because the original sequence was increasing in both halves, the only way that the final sequence can be out of order is if the terms in places $2r$ and $2r+1$ are the wrong way round. This can be cured with one final round of comparisons, which makes a depth of k.

This enables Mergesort to be carried out on 2^k objects with a sorting network of depth $k(k+1)/2$, since if the depth needed is $d(k)$, then $d(k) \le d(k-1)+k$: the $d(k-1)$ is needed to sort each half and the k is needed to do the merging. That gives the $C(\log n)^2$ claimed, with a good constant C. There are reasons to think that improving on a $(\log n)^2$ bound might be difficult, but the remarkable result of Ajtai, Komlós and Szemerédi [3] is that there is a sorting network with the trivially optimal depth of $C\log n$.

The full proof of this result is quite technical, though it has been simplified over the years. However, it is possible to give a flavour of the ideas. Let us begin with the concept of an *ε-approximate halver*. Let us say that a rock *is in the correct half* if it is one of the $n/2$ lightest rocks and is in one of the first $n/2$ places, or is one of the $n/2$ heaviest rocks and is in one of the last $n/2$ places. An ε-approximate halver is a comparator network such that for every initial permutation of the rocks, at most εn of them do not end up in the right half when we perform the corresponding sorting algorithm.

A natural way to build an ε-approximate halver of low depth is just to choose d partitions randomly. Suppose we do that and then perform the algorithm. Let us say that a place *has a rock of the right type* if the rock in that place is in the correct half. If at any stage, a place has a rock of the right type, then it will have a rock of the right type from that moment on. For instance, if one of the $n/2$ lightest rocks is in one of the first $n/2$ places, then it can only ever be replaced by a lighter rock, so the place will continue to have a rock of the right type.

But at each stage of the process, if there are θn rocks in the wrong half and hence $\theta n/2$ wrong rocks in each half, the probability that a rock in the wrong half gets compared with a wrong rock in the other half is at least θ, and if such a comparison takes place, then the two places are filled with rocks from the correct halves. So to argue very crudely, for as long as there are εn rocks in the wrong half, each place with a rock of the wrong type has a probability ε of being filled with a rock of the correct type. So after d rounds, it has a probability at least $1-(1-\varepsilon)^d \approx 1-\exp(-d\varepsilon)$ of being filled with a rock of the correct type. Therefore, we can take d to be around $\varepsilon^{-1}\log(\varepsilon^{-1})$.

If we could move *all* rocks to the correct half in a constant number of rounds, then we would almost be done: all we would have to do is repeat the procedure (in parallel) inside each half so that each rock was in the correct quarter, and so on all the way down. But it is easy to see that this is impossible. If there are fewer than $n/2$ rounds and we only ever compare rocks from different halves, then for each place in the first half there is some place in the second half that it never gets compared with. Pick an arbitrary place r in the first half and a place s in the second half that is never compared with r. Then put the lightest $n/2$ rocks in the first half and the heaviest $n/2$ rocks in the second half, except in places r and s. In those places put rocks of the wrong type. Then none of the comparisons will move any of the rocks.

That changes if one is allowed to make comparisons within each half, but even then a depth of $c\log n$ is necessary. The reason is that if the depth is d then there are at most 2^d places that can hold rocks that end up in any given place, so we can put a rock of the wrong type in the first place, say and also in all the 2^d places that can hold rocks that end up in the first place, then there will be a rock of the wrong type in the first place at the end of the process.

To get round this difficulty, Ajtai, Komlós and Szemerédi invented a complicated and extremely ingenious scheme for ensuring that rocks that get "left behind" are moved at a later stage. Thus, in a sense, their network was an approximation of the kind of network that we have just seen cannot exist.

There was one final ingredient of their argument, which turned the above ideas from a random sorting network into a deterministic one. That was to use bipartite expander graphs. A bipartite graph with vertex sets X and Y of the same size is called a (λ,α,d)-*expander* if for every subset $A\subset X$ of size at most $\alpha|X|$, the number of vertices in Y that are joined to at least one vertex in A is greater than $\lambda|A|$, and the same for subsets of Y.

It can be shown that whenever $\lambda\alpha<1$ and n is sufficiently large, there exists d depending on λ and α only, and a collection of d perfect matchings between two sets X and Y of size n, such that the union of these perfect matchings is a (λ,α,d)-expander.

To see how this might be useful, suppose we use such a collection of matchings to form a comparator network of depth d, taking $\alpha=\varepsilon$ and $\lambda=(1-\varepsilon)/\varepsilon$. It is easy to see that after applying the corresponding sorting algorithm, we cannot be left with $\varepsilon n/2$ heavy rocks in the light places and $\varepsilon n/2$ light rocks in the heavy places. To see this, suppose that we have a set A of $\varepsilon n/2$ places on the light side and a set

B of $\varepsilon n/2$ places on the heavy side. In each place on the light side, rocks only ever get lighter, and in each place on the heavy side, rocks only ever get heavier. The expansion property guarantees that there is an edge in the graph between A and B. Therefore, a comparison must have happened between a rock at place $a \in A$ and a rock at place $b \in B$, after which the rock in place a will always be lighter than the rock in place b. This shows that it cannot be the case that after the comparisons, all the rocks in A are in the heavy half and all the rocks in B are in the light half.

8 A Theorem on Point-Line Incidences

Suppose that you have n points $x_1, \dots, x_n$ and m lines $L_1, \dots, L_m$ in the plane. An *incidence* is simply a pair (i, j) such that $x_i \in L_j$. The following question sounds almost too simple to be interesting: how many incidences can there be? The answer, discovered by Szemerédi and Trotter, turned out to be very interesting indeed: their result is not simple at all, and its numerous consequences have made it a central result in combinatorial geometry.

The Szemerédi–Trotter theorem is the following statement [27].

Theorem 9 *Amongst any n points and m lines the number of incidences cannot be greater than $C(m + n + (mn)^{2/3})$.*

This bound looks a little strange at first, but a few observations make it seem more natural. To begin with, we could equally well write the bound as $C \max\{m, n, (mn)^{2/3}\}$. The form of the bound is telling us that there are essentially three competing examples, and which one is best depends on the relative sizes of m and n.

It is easy to see that we can have m incidences or n incidences: we just take m lines containing a point or n points along a line. To see how to obtain $(mn)^{2/3}$ incidences, consider the grid $\{1, 2, \dots, r\} \times \{1, 2, \dots, s\}$ in $\mathbb{Z}^2$. For each pair of points $(a, 1)$ and $(a + d, 2)$ such that $1 \leq a \leq r/2$ and $1 \leq d \leq r/2s$, the line joining $(a, 1)$ to $(b, 2)$ intersects this grid at all the s points $(a, 1), (a + d, 2), \dots, (a + (s - 1)d, s)$. There are $r^2/4s$ such lines.

Therefore, we can find a set of rs points and $r^2/4s$ lines with $r^2/4$ incidences. So for given m and n we need to solve the equations $rs = n$ and $r^2/4s = m$. This requires n to be at most m^2 (up to a constant) and m to be at most n^2 (also up to a constant). But if these inequalities do not hold, then one of m and n is bigger than $(mn)^{2/3}$.

This shows that the bound obtained by Szemerédi and Trotter is best possible.

For almost all of this article, I have focused on Szemerédi's original arguments, or slightly cleaned up versions of the arguments that have been produced since. In this case, however, there is a beautiful short proof discovered by Székely [25] that can be presented in full, and it seems a pity not to give it. As with most of

Szemerédi's proofs, however, his original proof of Theorem 9 is still interesting and important: there are certain generalizations that can be proved with his method that do not appear to be provable using the technique I am about to describe.

8.1 Székely's Proof of the Szemerédi–Trotter Theorem

The observation on which Székely's proof crucially depends is that a set of points and lines can be used to define a graph, and that graph has many vertices and edges. The graph is a very obvious one: its vertices are the points, and two vertices are joined if they appear consecutively along one of the lines.

If there are n points, m lines and t incidences, then the graph has n vertices and $t - m$ edges. The reason for the last assertion is that a line with k points on it gives rise to $k - 1$ edges. (I am assuming here, as I may, that each line contains at least one of the points.)

Something else that we know about this graph is that it can be drawn in the plane with at most $\binom{m}{2}$ *crossings*—that is, edges that are represented by intersecting curves (which happen in this case to be line segments). However, it turns out that we can also get a *lower* bound on the number of crossings, and that means that we are in business.

Lemma 5 *Let G be a graph with n vertices and m edges. Then any drawing of G in the plane (whether edges are represented by line segments or by more general curves) must have at least $m^3/72n^2$ crossings.*

Proof Euler's formula tells us that if G is a planar graph with V vertices, E edges and F faces, then $V - E + F = 2$. Since every face is bounded by at least three edges (if $V \geq 3$), and every edge is contained in at most two faces, $2E \geq 3F$, so $V - E + 2E/3 \geq 2$, which implies that $E \leq 3V - 6$.

To put this result a different way, if we have a drawing of a graph with n vertices and more than $3n - 6$ edges, then there must be at least one crossing. It follows that a drawing of a graph with n vertices and m edges must have at least $m - 3n$ crossings, since we can repeat the following process at least $m - 3n$ times (in fact, at least $m - 3n + 6$ times): find an edge involved in a crossing and remove it, thereby destroying that crossing.

Now a simple averaging argument allows us to improve this bound for large m. Let G be a graph with n vertices and m edges, and choose a random subgraph H of G by picking each vertex independently with probability p. Suppose that G has been drawn with t crossings. Then the expected number of vertices in H is pn and the expected number of crossings is $p^4 t$, since for a crossing to belong to the subgraph, all four vertices of the two crossing edges must survive.

But the expected number of edges is $p^2 m$, so the expected number of crossings is also at least $pm - 3pn$ by the bound above. It follows that $p^4 t \geq p^2 m - 3pn$. Choosing p to be $6n/m$, we find that $t \geq p^{-2} m/2 = m^3/72n^2$. □

Applying Lemma 5 to the graph described above, we deduce that $\binom{m}{2} \geq (t - m)^3/72n^2$. Therefore, if $t \geq 2m$, we may deduce that $m^2 \geq ct^3/n^2$, which gives us the upper bound $t \leq C(mn)^{2/3}$.

8.2 An Application of the Szemerédi–Trotter Theorem

A major conjecture in additive combinatorics, due to Erdős and Szemerédi [11], states that if A is a set of integers of size n, then one of $A + A = \{x + y : x, y \in A\}$ and $A.A = \{xy : x, y \in A\}$ must have size at least $n^{2-\varepsilon}$. Since the largest possible size of the sumset or product set is $n(n+1)/2$, this is saying that one or other of the two sets must have near-maximal size.

It is not easy to obtain any non-trivial lower bound, but with the help of the Szemerédi–Trotter theorem one can show that either $A + A$ or $A.A$ has size at least $n^{5/4}$. More precisely, we have the following result. The beautiful proof is due to Elekes [9].

Theorem 10 *Let A be a set of size n. Then $|A + A||A.A| \geq cn^{5/2}$.*

Proof As our set of points we take the Cartesian product $(A + A) \times (A.A)$ and suppose that this set has size t. As our lines we take every line of the form $\{(a + \lambda, \lambda b) : \lambda \in \mathbb{R}\}$ with $a, b \in A$. Each such line intersects $(A + A) \times (A.A)$ once for every $\lambda \in A$, and therefore in n points. Therefore, since there are n^2 lines, the number of incidences is n^3. By the Szemerédi–Trotter theorem it follows that $n^3 \leq C \max\{t, n^2, t^{2/3}n^{4/3}\}$. It follows that either $t \geq cn^3$, in which case we are trivially done, or $Ct^{2/3}n^{4/3} \geq n^3$, which translates into the stated bound $t \geq cn^{5/2}$. □

The Szemerédi–Trotter theorem and modifications of it have been used to obtain many partial results in combinatorial geometry. Some of these exploit the fact that we can replace the lines in the theorem by any collection of curves, provided that no two of those curves intersect in more than a bounded number of points. For example, the Erdős distance problem asks whether given any set of n points in the plane there must be at least $n^{1-\varepsilon}$ distinct distances between them. If there are very few distances, then there are many circles about points in the set that contain many other points in the set. This gives us a set of curves with many point-curve incidences. The argument is not as straightforward as that makes it sound, because two points can belong to several different circles, so the crossing lemma needs to be generalized to graphs with multiple edges and applied accordingly. But if two points x and y belong to many circles, then the centres of those circles all lie in a line. Therefore, if we have many examples of pairs of points that belong to many circles in the set, we have a system of lines that contain many points in the set and can apply the Szemerédi–Trotter theorem again.

Recently, in a major breakthrough, the Erdős distance problem was solved by Guth and Katz using different methods [14]. However, the Szemerédi–Trotter theorem continues to be a very important tool.

8.3 What Are the Extremal Sets in the Szemerédi–Trotter Theorem?

Let us end this section with a fascinating and somewhat open-ended question (which I learned from Jozsef Solymosi).

Question 1 Let P be a set of n points and let L be a set of n lines. Suppose that there are at least $10^{-10}n^{4/3}$ incidences between P and L. What can be said about the structures of P and L?

An answer to this question would fit very well a recurring theme in extremal combinatorics, which is to take an extremal result and to ask what happens in the near-extremal cases. For most such problems, we have an inequality and can say what happens when equality occurs. To give a simple example, if A is a set of n numbers, then $|A + A| \geq 2n - 1$, and equality holds if and only if A is an arithmetic progression. However, with the Szemerédi–Trotter theorem, the exact best possible bound is not known, so obtaining a structural result for *any* bound seems to be challenging. In the case of sumsets, a beautiful theorem of Freiman completely characterizes, at least qualitatively, all sets A such that $|A + A| \leq C|A|$ for some fixed constant C: each such set has to be a large subset of a generalized arithmetic progression of low dimension. Here, one might be looking for some kind of grid-like structure. This would follow from known results (one of which is Freiman's theorem itself) if one could show that there had to be cn^3 quadruples of points in P that formed the vertices of (possibly degenerate) parallelograms.

9 The Probability that a Random ±1 Matrix is Singular

Let M be a random $n \times n$ matrix where each entry has a 50 % chance of being 1 and a 50 % chance of being -1, with all choices independent. What is the probability that M is singular? Equivalently, what is the probability that if you choose n random 01-sequences of length n, then one of them will be in the linear span of the others?

This very basic question is surprisingly difficult to answer. Even to show that the probability tends to zero was a non-trivial open problem, solved by Komlós in 1967 [17]. (In this case the discrepancy between publication date and the date of the actual proof is quite large: the result was obtained in 1963.) He proved that the probability is at most $C/\sqrt{n}$.

There is a natural conjecture for the correct bound, which is $(2+o(1))\binom{n}{2}2^{-(n-1)}$. The heuristic argument for this is that by far the easiest way to obtain a linear dependence amongst the rows of a random ± 1 matrix ought to be to have two rows or two columns that are equal up to a ± 1 multiple.

The truth of this conjecture is still an open problem, and one that appears to need a major new idea. Given that situation, the next strongest aim it was reasonable to have was to prove that the probability was exponentially small. This too seemed out

of reach, so it was a big surprise when Kahn, Komlós and Szemerédi proved it in 1995 [15]. In the remainder of this section, let us look at some of the ideas that were involved in their proof.

9.1 The Need to Consider Dependences

One might attempt to prove the result in the following way, which works for many problems.

- Express the event E whose probability we are trying to estimate as a union of simple events.
- Give upper bounds for the probabilities of the simple events.
- Use the trivial "union bound" (that is, just add up the probabilities of the simple events) as an upper bound for the original event E.

In our case, E is the event that a random $n \times n$ ± 1 matrix M is singular. But M is singular if and only if $Ma = 0$ for some $a \in \mathbb{R}^n$, so an obvious candidate for the set of simple events is to take all events of the form "$Ma = 0$". Let us call this event E_a.

Obviously this won't do as it stands, since there are infinitely many possible a. However, we could try to identify a finite set A of vectors a such that if M is singular then there exists $a \in A$ such that $Ma = 0$. Such sets trivially exist: for each singular matrix M we pick a vector a such that $Ma = 0$ and then we put together these vectors to form our set A. However, they do not necessarily help us. For example, let F be the set of all ± 1 matrices that have two pairs of equal columns. For each matrix $M \in F$, let a be a vector with four non-zero coordinates that take the values $\pm\lambda$ and $\pm\mu$ in such a way that $Ma = 0$, and make sure that no one of these vectors is a multiple of another. The number of vectors we create is at least $2^{n(n-2)}$, and for each such vector a the probability that $Ma = 0$ is at least 2^{-2n}. Multiplying these numbers together we get $2^{n(n-4)}$, which is far bigger than 1 and therefore tells us nothing.

Of course, it was perverse of us to make sure that no two of the vectors were multiples of each other: if we had taken $\lambda = \mu = 1$ for every single vector, then the number of vectors would have dropped to cn^4. But the point is nevertheless made that for a union bound to work one would have to obtain a great deal of duplication of this kind, which is not obviously possible.

Kahn, Komlós and Szemerédi use a natural generalization of this approach. Instead of using the trivial fact that if $a = b$ then the events E_a and E_b are the same, so that only one of them needs to be considered in a union bound, they show that if several vectors a_i belong to a low-dimensional subspace S, then the events E_{a_i} are highly correlated, with the result that the event $\bigcup_i E_{a_i}$ has a much smaller probability than $\sum_i P(E_{a_i})$. In other words, linear dependencies lead to useful probabilistic dependencies.

9.1.1 The Probability that $Ma = 0$

Nevertheless, it is useful to think about the events E_a and in particular about their probabilities. Given a vector a, let $p(a)$ be the probability that $\sum_i \varepsilon_i a_i = 0$, where $(\varepsilon_1, \dots, \varepsilon_n)$ is a random ± 1 sequence. Then the probability that $Ma = 0$ is $p(a)^n$.

What sort of values can $p(a)$ take?

- If a is the vector $(1, 1, \dots, 1)$, then $p(a)$ is around $n^{-1/2}$.
- More generally, if a takes the value ± 1 d times and 0 otherwise, then $p(a)$ is around $d^{-1/2}$.
- The Littlewood-Offord inequality, or rather a slight improvement of it due to Erdős, implies a sort of converse to this observation: if the support of a has size d, then $p(a) \leq d^{-1/2}$.
- Sárközy and Szemerédi proved that if $a_1, \dots, a_n$ are distinct, then $p(a) \leq Cn^{-3/2}$.

Thus, for $p(a)$ to be large, we need a to have small support and many repeated entries.

9.1.2 Dealing with Vectors a for which $p(a)$ Is Very Small

A simple lemma shows that we can at least disregard all vectors a for which $p(a)$ is exponentially small.

Lemma 6 *For every $p \in [0, 1]$ the probability that there exists a such that $p(a) \leq p$ and $Ma = 0$ is at most np.*

Proof Let $E(p)$ be the event that such a vector a exists. Let us condition on the entire matrix M apart from the ith row, and bound from above the probability, given those $n(n-1)$ values, that $E(p)$ holds and the ith row is in the linear span of the other rows.

Now for $E(p)$ to hold conditional on these values, there must exist a with $p(a) \leq p$ that is orthogonal to all rows apart from the ith. Pick any such vector a. For the ith row to be a linear combination of the other rows, it is necessary that it too should be orthogonal to a, and this happens with probability at most p.

For $E(p)$ to hold in general, there must exist a row of M that is in the linear span of the other rows. The lemma follows. □

9.1.3 Applying Linear Dependence

A similar argument shows that we can improve the bound in Lemma 6 if we insist that a belongs to a k-codimensional subspace.

Lemma 7 *Let S be a k-codimensional subspace and let $0 \leq p \leq 1$. Then the probability that there exists $a \in S$ such that $p(a) \leq p$ and $Ma = 0$ is at most $\binom{n}{k+1} p^{k+1}$.*

Proof For such an a to exist, it is necessary that the kernel of M intersects S. Therefore, writing V_j for the orthogonal complement of the span of the first j rows of M, there can be at most $n-k-1$ values of j for which $V_j \cap S$ is a proper subset of $V_{j-1} \cap S$.

Now let us fix a set J of size $n-k-1$ and assume that $j \in J$ whenever $V_j \cap S \subset \neq V_{j-1} \cap S$. Let us condition on the values of M in the $n-k-1$ rows corresponding to J.

Now let $j \notin J$. By construction, $V_j \cap S = V_{j-1} \cap S$, so if $Ma = 0$ with $p(a) \leq p$, then $a \in V_{j-1}$, which implies that $a \in V_j$ and hence that a must be orthogonal to the jth row of M, which happens with probability at most p. We can apply this argument to each of the $k+1$ rows not corresponding to elements of J, and since those rows are independent, we obtain an upper bound of p^{k+1} for that choice of J. Applying the law of total probability and summing over all J gives the result. □

9.2 The Main Idea

Let us informally refer to a vector a as *bad* if $p(a)$ is large (meaning greater than $(1-\varepsilon)^n$ for some suitable ε). For Lemma 7 to be useful, we need to be able to show that we can cover the bad vectors efficiently with subspaces of fairly low dimension. To this end, Kahn, Komlós and Szemerédi prove a result that seems at first glance to be rather unlikely to be true. Let $a_1, \dots, a_n$ are integers and let $\mu > 0$ be a smallish absolute constant. Consider the following two random walks. At time t, the first walk chooses a random step of $\pm a_t$, each with probability $1/2$. The second walk chooses a random step of $\pm a_t$, each with probability μ, and a step of 0 with probability $1-2\mu$. Their result is that, no matter what the initial sequence $a_1, \dots, a_n$ was, as long as it has a reasonably large support, the probability that the first walk ends up at 0 is smaller by a factor of $O(\sqrt{\mu})$ than the probability that the second walk ends up at 0.

Because the binomial distribution is highly concentrated about its mean, the second walk is similar, but not identical, to a walk where we first randomly choose $d = 2\mu n$ of the a_i, replace all the others by 0, and then do a normal random walk with the new sequence. So it might seem that the result cannot be true if, for example, we take the sequence $1, 2, 4, \dots, 2^{n-2}, -(2^{n-1}-1)$, in which case the only way of getting back to the origin is to take all signs positive or all signs negative. However, in this case the probability of ending at 0 with the first walk is $2^{-(n-1)}$, while the probability with the second walk is $2\mu^n + (1-2\mu)^n$, which is much larger as long as μ is smaller than $1/4$. However, in this case the result is not really telling us very much: for it to be useful we need $p(a)$ not to be too small, which, roughly speaking, allows us to assume that the higher probability in the second walk arises for non-trivial reasons.

9.2.1 A Very Rough Sketch of the Main Argument

To see how all this helps, recall that our aim is to choose a collection of subspaces of not too large dimension that cover all the "bad" vectors a. A bad vector is one for which $p(a)$ is large, and the main step described above implies that if $p(a)$ is large, then for some d close to λn, the probability that a random ± 1 sum of d randomly chosen entries of a is 0 is larger than $p(a)$ by at least a factor $c\sqrt{\lambda}$.

Now let us choose a subspace S randomly as follows. Define a *d-vector* to be a vector in $\{1, -1, 0\}^n$ that takes non-zero values exactly d times. Also, given a vector a, define a *d-sum* of a to be a ± 1-sum of d terms of a. Equivalently, it is the inner product of a with a d-vector. For a suitable γ, choose $(1-\gamma)n$ d-vectors at random and let S be the orthogonal complement of the space spanned by these d-vectors. Thus, $a \in S$ if and only if every d-sum of a that corresponds to one of the d-vectors we have chosen is 0. If we know that the probability that a random d-sum of a is 0 is greater by a factor $c\sqrt{\lambda}$ than $p(a)$, then the probability that a belongs to the γn-dimensional subspace S is greater by a factor $(c\sqrt{\lambda})^{(1-\gamma)n}$ than the bound of roughly $p(a)^{(1-\gamma)n}$ that comes from Lemma 7.

We can now partition the interval $[(1-\varepsilon)^n, 1]$ into not too many subintervals of ps of approximately the same size. If we apply Lemma 7 to a particular value of p, then each subspace we apply it to contributes roughly $p^{(1-\gamma)n}$ (the binomial coefficient turns out not to make too much difference so I am ignoring it). But since the probability that a vector a with $p(a) \approx p$ belongs to a random such subspace is more like $(C\lambda^{-1/2})^n p^{(1-\gamma)n}$, the number of such subspaces that we need to cover all the a with $p(a) \approx p$ is roughly $(c\lambda^{1/2})^n p^{-(1-\gamma)n}$, and the total contribution is $(c\lambda^{1/2})^n$, which is exponentially small. Adding up the contributions of this kind, we find that they are dominated by the contribution of $n(1-\varepsilon)^n$ that came from the a for which $p(a)$ is very small, and the result is proved.

9.3 Subsequent Improvements

The bound obtained by Kahn, Komlós and Szemerédi was around $(0.999)^n$. In 2006 this was slightly improved, to $(0.953)^n$, by Tao and Vu [29]. The following year they obtained a bound of $(3/4 + o(1))^n$ using methods from additive combinatorics [30]. The current record is $(1/\sqrt{2} + o(1))^n$. This is a very recent result (it appeared in 2013) of Bourgain, Vu and Wood [7].

10 Conclusion

Szemerédi's work has several qualities that make it stand out and that make him one of the great mathematicians of the second half of the twentieth century, not to mention the beginning of the twenty-first. An obvious one is the sheer difficulty of so

many of his results. He has often solved open problems on which the mathematical community had become completely stuck, and his ingenious and delicate solutions have often left other mathematicians feeling that they were in a sense right to be stuck. Another quality that many of his results have had, and that the very best results in combinatorics have, is that the proofs have introduced techniques and ideas with applications that go far beyond the original problems that Szemerédi was solving. His influence permeates the whole of combinatorics and theoretical computer science, fully justifying the award of the Abel Prize.

References

1. Ajtai, M., Komlós, J., Szemerédi, E.: A note on Ramsey numbers. J. Comb. Theory, Ser. A **29**, 354–360 (1980)
2. Ajtai, M., Komlós, J., Szemerédi, E.: A dense infinite Sidon sequence. Eur. J. Comb. **2**, 1–11 (1981)
3. Ajtai, M., Komlós, J., Szemerédi, E.: An $O(n \log n)$ sorting network. Combinatorica **3**, 1–19 (1983)
4. Ajtai, M., Szemerédi, E.: Sets of lattice points that form no squares. Studia Sci. Math. Hung. **9**, 9–11 (1975)
5. Bárány, I.: Applications of graph and hypergraph theory in geometry. In: Goodman, J.E., Pach, J., Welzl, E. (eds.) Combinatorial and Computational Geometry. MSRI Publications, vol. 52, pp. 31–50 (2005)
6. Batcher, K.: Sorting networds and their applications. AFIPS Spring Joint Comput. Conf. **32**, 307–314 (1968)
7. Bourgain, J., Vu, V.H., Wood, P.M.: On the singularity probability of discrete random matrices. http://arxiv.org/abs/0905.0461
8. Chung, F., Graham, R.L., Wilson, R.M.: Quasi-random graphs. Combinatorica **9**, 345–362 (1989)
9. Elekes, Gy.: On the number of sums and products. Acta Arith. **81**, 365–367 (1997)
10. Erdős, P.: Some remarks on the theory of graphs. Bull. Am. Math. Soc. **53**, 292–294 (1947)
11. Erdős, P., Szemerédi, E.: On sums and products of integers. In: Erdős, P., Alpar, L., Halasz, G. (eds.) Studies in Pure Mathematics: To the Memory of Paul Turán, pp. 213–218. Birkhäuser, Basel (1983)
12. Erdős, P., Turán, P.: On some sequences of integers. J. Lond. Math. Soc. **11**, 261–264 (1936)
13. Graver, J.E., Yackel, J.: Some graph theoretic results associated with Ramsey's theorem. J. Comb. Theory **4**, 125–175 (1968)
14. Guth, L., Katz, N.H.: On the Erdős distinct distances problem in the plane. http://arxiv.org/abs/1011.4105
15. Kahn, J., Komlós, J., Szemerédi, E.: On the probability that a random ± 1-matrix is singular. J. Am. Math. Soc. **8**, 223–240 (1995)
16. Kim, J.H.: The Ramsey number $R(3, t)$ has order of magnitude $t^2/\log t$. Random Struct. Algorithms **7**, 173–207 (1995)
17. Komlós, J.: On the determinant of (0, 1) matrices. Studia Sci. Math. Hung. **2**, 387–399 (1968)
18. Komlós, J., Pintz, J., Szemerédi, E.: A lower bound for Heilbronn's problem. J. Lond. Math. Soc. **25**, 13–24 (1982)
19. Roth, K.F.: On a problem of Heilbronn. J. Lond. Math. Soc. **26**, 198–204 (1951)
20. Roth, K.F.: On certain sets of integers, I. J. Lond. Math. Soc. **28**, 104–109 (1953)
21. Roth, K.F.: On a problem of Heilbronn, II. Proc. Lond. Math. Soc. **25**, 193–212 (1972)
22. Ruzsa, I.Z., Szemerédi, E.: Triple systems with no six points carrying three triangles. In: Combinatorics, Proc. Fifth Hungarian Colloq., Keszthely, 1976. Coll. Math. Soc. J. Bolyai 18, Volume II, pp. 939–945. North Holland, Amsterdam (1978)

23. Schmidt, W.M.: On a problem of Heilbronn. J. Lond. Math. Soc. **4**, 545–550 (1972)
24. Solymosi, J.: Note on a generalization of Roth's theorem. In: Pach, J. (ed.) Discrete and Computational Geometry. Algorithms Combin., vol. 25, pp. 825–827. Springer, Berlin (2003)
25. Székely, L.A.: Crossing numbers and hard Erdős problems in discrete geometry. Comb. Probab. Comput. **6**, 353–358 (1997)
26. Szemerédi, E.: On sets of integers containing no k elements in arithmetic progression. Acta Arith. **27**, 199–245 (1975)
27. Szemerédi, E., Trotter, W.T.: Extremal problems in discrete geometry. Combinatorica **3**, 381–392 (1983)
28. Tao, T.: Some ingredients in Szemerédi's proof of Szemerédi's theorem. Blog post (2012). http://terrytao.wordpress.com/2012/03/23/some-ingredients-in-szemeredis-proof-of-szemeredis-theorem/
29. Tao, T., Vu, V.H.: On random ± 1 matrices: singularity and determinant. Random Struct. Algorithms **28**, 1–23 (2006)
30. Tao, T., Vu, V.H.: On the singularity probability of random Bernoulli matrices. J. Am. Math. Soc. **20**, 603–628 (2007)
31. Thomason, A.: Pseudo-random graphs. In: Karoński, M. (ed.) Proceedings of Random Graphs, Poznań, 1985, pp. 307–331. North Holland, Amsterdam (1987)