Szemerédi's regularity lemma revisited

Lewis Memorial Lecture
March 14, 2008

Terence Tao (UCLA)

Finding models of large dense graphs

Suppose we are given a large dense graph $G = (V, E)$, where V is a set of $n = |V|$ vertices (with n large), and E is a set of edges, which for us will usually be dense in the sense that $|E| \geq cn^2$ for some constant $c > 0$.

- **Example 1** G could be an Erdős-Rényi random graph $G(V, 1/2)$, with any two vertices connected by an edge with probability $1/2$.

- **Example 2** G could be a random complete bipartite graph $G(V_1, V_2, 1)$, with the two vertex classes V_1, V_2 chosen at random.

It is natural to try to classify all large dense graphs into various "types".

- Q1. Is there some "reasonable" classification of all possible "types" of such graphs?

- Q2. Can one determine what "type" a graph is, just by looking at a small (random) piece of it?

- Q3. If one knows what type of graph one has, what questions can one then answer about the graph?

Of course, these questions are related to each other, and depend on the definition of "type".

When n is large, the number of possible graphs on n vertices is enormous (of size $2^{\binom{n}{2}}$). This is still true even if we identify pairs of graphs which are isomorphic (note the number of isomorphisms is only $n! = 2^{O(n \log n)}$). So a complete description of all such graphs is infeasible in the limit $n \to \infty$.

But suppose we only want to classify all the large dense graphs approximately, by choosing some error tolerance $\varepsilon > 0$ and asking how many different "types" of large dense graph there are "up to error ε". Does the situation improve?

4

The answer is yes - if we define "type" and "error" properly. The correct viewpoint is to view large dense graphs as being built up of special components, namely the ε-regular graphs.

> **Definition** A bipartite graph $G = (V, W, E)$ is said to be ε-regular with density $0 \leq d \leq 1$ if for any $S \subset V$ and $T \subset W$ one has $|e_G(S, T) - d|S||T|| \leq \varepsilon |V||W|$, where $e_G(S, T)$ is the number of edges between S and T.

Example If V, W are sufficiently large depending on ε, then a random bipartite graph $G(V, W, d)$ between V and W, in which every pair forms an edge with independent probability d, is very likely to be ε-regular.

A more informal definition: a bipartite graph $G = (V, W, E)$ is ε-regular with density $0 \le d \le 1$ and some small ε if the probability distribution of small randomly chosen subgraphs of G is nearly indistinguishable (up to error $O(\varepsilon)$) from that of the random graph $G(V, W, d)$ in the previous example.

This allows us to compute any "local" statistic of an ε-regular graph. For instance, the number of (labeled) 4-cycles in such a graph is $(d^4 + O(\varepsilon))|V|^2|W|^2$.

The remarkable fact is that every large dense graph is essentially made up of a bounded number of regular graphs:

> **Szemerédi regularity lemma** (Szemerédi, 1975) Let $G = (V, E)$ be a graph, and let $\varepsilon > 0$. Then there exists a vertex partition $V = V_1 \cup \ldots \cup V_m$ into $m = O_\varepsilon(1)$ pieces of comparable size, such that the portion of G connecting V_i and V_j is ε-regular (with some density d_{ij}) for $1-\varepsilon$ of all pairs $1 \le i < j \le m$.

Thus, in principle, every graph is described "up to error ε" (and up to isomorphism) by a collection of numbers d_{ij} for $1 \le i < j \le m$, where m is bounded by ε but not depending on n.

The dependence of m on ε is, however, rather poor - m can be as large as a tower of exponentials of height $\varepsilon^{-O(1)}$ (Gowers, 1997). Nevertheless, this lemma is of fundamental importance in the asymptotic regime $n \to \infty$.

The regularity lemma has countless applications. Here is a typical one:

> **Triangle removal lemma**(Ruzsa-Szemerédi, 1978) Suppose a graph G on n vertices contains only $o(n^3)$ triangles. Then these triangles can be deleted by removing at most $o(n^2)$ edges.

Sketch of proof: Apply the regularity lemma. If there exist cells V_i, V_j, V_k with $d_{ij}d_{jk}d_{ki}$ large (and the pairs ij, jk, ki all regular) then this would force G to have more than $o(n^3)$ triangles.

Thus, for every i, j, k, one of the pairs ij, jk, ki is either irregular or has small density. Deleting the $o(n^2)$ edges associated to such pairs one obtains the claim.

All known proofs of the triangle removal lemma use some version of the regularity lemma (or something equivalent to that lemma)!

The same general strategy allows one to prove various property testing results for graphs using the regularity lemma; see e.g. Alon-Shapira (2005), Lovasz-Szegedy (2005), ...

The graph theoretic applications of the regularity lemma can lead to applications in other areas too. Here is a typical example:

> **Roth-Varnavides theorem**(Roth 1956; Varnavides, 1959) A subset of $\mathbf{Z}/n\mathbf{Z}$ of cardinality at least δn will contain at least $c(\delta)n^2$ arithmetic progressions of length 3 if n is large enough compared to δ, where $c(\delta) > 0$.

Sketch of proof: Suppose for contradiction that we could find a set $A \subset \mathbf{Z}/n\mathbf{Z}$ of cardinality at least δn which only had $o(n^2)$ arithmetic progressions. In particular, there

are at most $o(n^3)$ solutions to the system of constraints

$$x, y, z \in \mathbf{Z}/n\mathbf{Z}; 2x + y \in A; x - z \in A; -y - 2z \in A(*).$$

This is asserting that a certain tripartite Cayley graph on $O(n)$ vertices has only $o(n^3)$ triangles, which can then be deleted by removing only $o(n^2)$ edges. But every $a \in A$ and $r \in \mathbf{Z}/n\mathbf{Z}$ generates a solution $(x, y, z) = (r, a - 2r, r - a)$ to (*), and removing an edge deletes at most one of these δn^2 solutions, giving the required contradiction.

Thanks to recent work by Nagle-Rödl-Schacht-Skokan, Gowers, and later authors on extending the regularity lemma to hypergraphs, the above methods have been extended to give a new and fairly direct proof of

> **Szemerédi's theorem**(Szemerédi, 1975) Every set of integers of positive upper density contains arbitrarily long arithmetic progressions.

Another application of the regularity lemma is to compactify the space of all large dense graphs. Let us say that a sequence $G_i = (V_i, E_i)$ of increasingly large graphs is convergent if for any finite k, the probability distribution of a random k-element subgraph of G_i converges to a limit as $i \to \infty$. (For instance, for fixed p, the Erdős-Rényi graphs $G(n, p)$ are almost surely convergent as $n \to \infty$.)

> **Theorem** (Lovasz-Szegedy, 2004) Every sequence of increasingly large graphs has a convergent subsequence.

15

Actually, one can be more precise. Define a graphon to be a symmetric measurable function $p : [0, 1] \times [0, 1] \to [0, 1]$. One can then define a generalised Erdős-Rényi graph $G(n, p)$ on n vertices by giving each vertex v a colour $x_v \in [0, 1]$ uniformly at random, and then connecting v to w with probability $p(x_v, x_w)$ independently at random. Let $G(\infty, p)$ be the (formal) graph limit of the $G(n, p)$.

Theorem (Lovasz-Szegedy, 2004) Every sequence of increasingly large graphs has a subsequence converging to $G(\infty, p)$ for some graphon p.

Example The complete random bipartite graphs $G(V_1, V_2, 1)$ on n vertices almost surely converge to the graphon limit $G(\infty, p)$, where p is 1 on $[0, 1/2] \times (1/2, 1] \cup (1/2, 1] \cup [0, 1/2]$ and zero elsewhere.

This theorem is largely equivalent to the regularity lemma (or to a "weak" form of that lemma due to Frieze and Kannan). In particular, many consequences of the regularity lemma (e.g. the triangle removal lemma) can also be proven using graph limits. The role of the regularity lemma is then played by such classical analysis results as the Lebesgue differentiation theorem (!).

One can also view the theory of graph limits in terms of the theory of exchangeable measures from probability theory.

Given a finite deterministic graph $G = (V, E)$, one can create an infinite random graph $\tilde{G} = (\mathbf{Z}, \tilde{E})$ by randomly assigning a vertex $v_n \in V$ to each integer n and then "pulling back" the graph G by declaring n, m to be connected in $\tilde{E}$ if v_n, v_m are connected in E. This random graph is exchangeable, which means that permutations of the vertex set $\mathbf{Z}$ do not affect the distribution of the graph.

The existence of graph limits is then equivalent to the sequential compactness of exchangeable random graphs.

As pointed out recently by Austin and Diaconis-Janson, the identification of graph limits with graphons is equivalent to a classification of exchangeable random graphs (or exchangeable arrays) due to Aldous (1981) and Kallenberg (1992), which in turn generalises a classical theorem of de Finetti.

By using the theory of exchangeable measures and a compactness argument, one can show that one can regularise a graph $G = (V, E)$ by using random neighbourhoods. More precisely, if one picks m vertices $v_1, \ldots, v_m$ at random, and then partitions V into the 2^m cells formed by the neighbourhoods of these vertices, then the probability that these cells form an ε-regular partition of the graph G (after accounting for the unequal sizes of the cells) approaches 1 for a sequence of m going to infinity, where the decay rate and the sparsity of the sequence is uniformly controlled in n. (This fact was also observed by Ishigami, and may also be folklore.)

Intuitively, as m gets large, the random sample of m vertices is increasingly likely to ferret out any irregularities or biases in the distribution of edges in the graph, so that the only thing left after taking the connectivity of these m vertices into account is random noise.

This theory (and its extension to hypergraphs) was recently used by Austin and myself to establish various property testing results for graphs and hypergraphs.

Regularity in additive combinatorics and number theory

The philosophy of the regularity lemma - that a large general object can be modeled by objects of much lower complexity - is not restricted to graph theory. For instance, such lemmas have become useful in additive combinatorics and in number theory.

The arithmetic analogue of regularity is that of Gowers uniformity. Roughly speaking, two sets A, B are close in the Gowers uniformity norm of order 2 if the number of solutions to problems such as

$$a, a + r, a + 2r \in A$$

or

$$a, a + h_1, a + h_2, a + h_1 + h_2 \in A$$

are close to the number of solutions to the same problems in B.

We say that A is *ε-uniform of order* 2 if it is within ε of a random set of the same density in the Gowers uniformity norm of order 2.

More generally, A, B are close in the Gowers uniformity norm of order k if the number of solutions to problems such as

$$a, a + r, a + 2r, \ldots, a + kr \in A$$

or

$$a + \sum_{i \in I} h_i \in A \text{ for all } I \subset \{1, \ldots, k\}$$

are close to the number of solutions to the same problems in B.

We say that A is *ε-uniform of order k* if it is within ε of a random set of the same density in the Gowers uniformity norm of order k.

Here is a typical statement:

> **Arithmetic regularity lemma of order** 2 (Green, 2005) Let G be a vector space over $\mathbb{F}_2$, let $A \subset G$ be a set, and let $\varepsilon > 0$. Then one can partition G into $O_\varepsilon(1)$ affine subspaces, such that A is ε-uniform of order 2 on $1 - \varepsilon$.

This result is proven by exploiting a close connection between uniformity of order 2 of a set A, and the Fourier coefficients of that set A.

Similar (but more complicated) statements exist for more general abelian groups G, and can be used for instance to give new proofs of the Roth-Varnavides theorem (as well as various refinements thereof).

Analogues of this lemma for order 3 are known (Green-T. 2006, Gowers-Wolf 2008), and are conjectured for higher order.

A cousin of this lemma was also decisive in establishing that the primes contain arbitrarily long arithmetic progressions. Roughly speaking, the key proposition is

Sparse regularity lemma (Green-T. 2004) Let $k \geq 2$ be fixed and N be large. Let A be the (weighted) set of primes from 1 to N, with each prime p given a weight of $\log p$. Then there exists a dense (weighted) subset B of $\{1, \ldots, N\}$, with weights $O(1)$, such that A and B are close in the Gowers uniformity norm of order k.

[Actually, for technical reasons one does not work with the primes themselves, but with an arithmetic progression $\{n : Wn + b \text{ prime}\}$ of primes, for certain W, b.]

This lemma, combined with Szemerédi's theorem on arithmetic progressions, implies that the primes contain infinitely many progressions of length $k + 1$.

The proof of the lemma is related to that of certain variants of the Szemerédi regularity lemma for sparse graphs.