7.2 Szemerédi's regularity lemma

More than 20 years ago, in the course of the proof of a major result on the Ramsey properties of arithmetic progressions, Szemerédi developed a graph theoretical tool whose fundamental importance has been realized more and more in recent years: his so-called *regularity* or *uniformity lemma*. Very roughly, the lemma says that all graphs can be approximated by random graphs in the following sense: every graph can be partitioned, into a bounded number of equal parts, so that most of its edges run between different parts and the edges between any two parts are distributed fairly uniformly—just as we would expect it if they had been generated at random.

In order to state the regularity lemma precisely, we need some definitions. Let $G = (V, E)$ be a graph, and let $X, Y \subseteq V$ be disjoint. Then we denote by $\|X, Y\|$ the number of X–Y edges of G, and call $\|X, Y\|$

$$d(X, Y) := \frac{\|X, Y\|}{|X|\,|Y|}$$ $d(X, Y)$

the *density* of the pair (X, Y). (This is a real number between 0 and 1.) *density* Given some $\epsilon > 0$, we call a pair (A, B) of disjoint sets $A, B \subseteq V$ *ϵ-regular* if all $X \subseteq A$ and $Y \subseteq B$ with *ϵ-regular pair*

$$|X| \geqslant \epsilon\,|A| \quad \text{and} \quad |Y| \geqslant \epsilon\,|B|$$

satisfy

$$\big|d(X, Y) - d(A, B)\big| \leqslant \epsilon\,.$$

The edges in an ϵ-regular pair are thus distributed fairly uniformly: the smaller ϵ, the more uniform their distribution.

Consider a partition $\{ V_0, V_1, \ldots, V_k \}$ of V in which one set V_0 has been singled out as an *exceptional set*. (This exceptional set V_0 may be empty.[3]) We call such a partition an *ϵ-regular* partition of G if it satisfies the following three conditions: *exceptional set*

ϵ-regular partition

(i) $|V_0| \leqslant \epsilon\,|V|$;

(ii) $|V_1| = \ldots = |V_k|$;

(iii) all but at most ϵk^2 of the pairs (V_i, V_j) with $1 \leqslant i < j \leqslant k$ are ϵ-regular.

The role of the exceptional set V_0 is one of pure convenience: it makes it possible to require that all the other partition sets have exactly the same size. Since condition (iii) affects only the sets $V_1, \ldots, V_k$, we

[3] So V_0 may be an exception also to our terminological rule that partition sets are not normally empty.

may think of V_0 as a kind of bin: its vertices are disregarded when the uniformity of the partition is assessed, but there are only few such vertices.

Lemma 7.2.1. (Regularity Lemma)
[9.2.2] *For every $\epsilon > 0$ and every integer $m \geqslant 1$ there exists an integer M such that every graph of order at least m admits an ϵ-regular partition $\{V_0, V_1, \ldots, V_k\}$ with $m \leqslant k \leqslant M$.*

The regularity lemma thus says that, given any $\epsilon > 0$, every graph has an ϵ-regular partition into a bounded number of sets. The upper bound M on the number of partition sets ensures that for large graphs the partition sets are large too; note that ϵ-regularity is trivial when the partition sets are singletons, and a powerful property when they are large. In addition, the lemma allows us to specify a lower bound m on the number of partition sets; by choosing m large, we may increase the proportion of edges running between different partition sets (rather than inside one), i.e. the proportion of edges that are subject to the regularity assertion.

Note that the regularity lemma is designed for use with dense graphs:[4] for sparse graphs it becomes trivial, because all densities of pairs—and hence their differences—tend to zero (Exercise 22).

The remainder of this section is devoted to the proof of the regularity lemma. Although the proof is not difficult, a reader meeting the regularity lemma here for the first time is likely to draw more insight from seeing how the lemma is typically applied than from studying the technicalities of its proof. Any such reader is encouraged to skip to the start of Section 7.3 now and come back to the proof at his or her leisure.

We shall need the following inequality for reals $\mu_1, \ldots, \mu_k > 0$ and $e_1, \ldots, e_k \geqslant 0$:

$$\sum \frac{e_i^2}{\mu_i} \geqslant \frac{\left(\sum e_i\right)^2}{\sum \mu_i}\,. \tag{1}$$

This follows from the Cauchy-Schwarz inequality $\sum a_i^2 \sum b_i^2 \geqslant \left(\sum a_i b_i\right)^2$ by taking $a_i := \sqrt{\mu_i}$ and $b_i := e_i/\sqrt{\mu_i}$.

$G = (V,E)$ Let $G = (V,E)$ be a graph and $n := |V|$. For disjoint sets $A, B \subseteq V$
n we define

$q(A,B)$
$$q(A,B) := \frac{|A|\,|B|}{n^2}\, d^2(A,B) = \frac{\|A,B\|^2}{|A|\,|B|\,n^2}\,.$$

For partitions $\mathcal{A}$ of A and $\mathcal{B}$ of B we set

$q(\mathcal{A},\mathcal{B})$
$$q(\mathcal{A},\mathcal{B}) := \sum_{A' \in \mathcal{A};\ B' \in \mathcal{B}} q(A',B')\,,$$

[4] Sparse versions do exist, though; see the notes.

and for a partition $\mathcal{P} = \{ C_1, \dots, C_k \}$ of V we let

$$q(\mathcal{P}) := \sum_{i<j} q(C_i, C_j)\,.$$ $q(\mathcal{P})$

However, if $\mathcal{P} = \{ C_0, C_1, \dots, C_k \}$ is a partition of V *with exceptional set* C_0, we treat C_0 as a set of singletons and define

$$q(\mathcal{P}) := q(\tilde{\mathcal{P}})\,,$$

where $\tilde{\mathcal{P}} := \big\{C_1, \dots, C_k\big\} \cup \big\{\{ v \} : v \in C_0\big\}$. $\tilde{\mathcal{P}}$

The function $q(\mathcal{P})$ plays a pivotal role in the proof of the regularity lemma. On the one hand, it measures the uniformity of the partition $\mathcal{P}$: if $\mathcal{P}$ has too many irregular pairs (A, B), we may take the pairs (X, Y) of subsets violating the regularity of the pairs (A, B) and make those sets X and Y into partition sets of their own; as we shall prove, this refines $\mathcal{P}$ into a partition for which q is substantially greater than for $\mathcal{P}$. Here, 'substantial' means that the increase of $q(\mathcal{P})$ is bounded below by some constant depending only on ϵ. On the other hand,

$$\begin{aligned} q(\mathcal{P}) &= \sum_{i<j} q(C_i, C_j) \\ &= \sum_{i<j} \frac{|C_i|\,|C_j|}{n^2}\, d^2(C_i, C_j) \\ &\leqslant \frac{1}{n^2} \sum_{i<j} |C_i|\,|C_j| \\ &\leqslant 1\,. \end{aligned}$$

The number of times that $q(\mathcal{P})$ can be increased by a constant is thus also bounded by a constant—in other words, after some bounded number of refinements our partition will be ϵ-regular! To complete the proof of the regularity lemma, all we have to do then is to note how many sets that last partition can possibly have if we start with a partition into m sets, and to choose this number as our desired bound M.

Let us make all this precise. We begin by showing that, when we refine a partition, the value of q will not decrease:

Lemma 7.2.2.

(i) *Let $C, D \subseteq V$ be disjoint. If $\mathcal{C}$ is a partition of C and $\mathcal{D}$ is a partition of D, then $q(\mathcal{C}, \mathcal{D}) \geqslant q(C, D)$.*

(ii) *If $\mathcal{P}, \mathcal{P}'$ are partitions of V and $\mathcal{P}'$ refines $\mathcal{P}$, then $q(\mathcal{P}') \geqslant q(\mathcal{P})$.*

Proof. (i) Let $\mathcal{C} =: \{ C_1, \dots, C_k \}$ and $\mathcal{D} =: \{ D_1, \dots, D_\ell \}$. Then

$$\begin{aligned} q(\mathcal{C}, \mathcal{D}) &= \sum_{i,j} q(C_i, D_j) \\ &= \frac{1}{n^2} \sum_{i,j} \frac{\|C_i, D_j\|^2}{|C_i|\,|D_j|} \\ &\underset{(1)}{\geqslant} \frac{1}{n^2} \frac{\big(\sum_{i,j} \|C_i, D_j\|\big)^2}{\sum_{i,j} |C_i|\,|D_j|} \\ &= \frac{1}{n^2} \frac{\|C, D\|^2}{\big(\sum_i |C_i|\big)\big(\sum_j |D_j|\big)} \\ &= q(C, D)\,. \end{aligned}$$

(ii) Let $\mathcal{P} =: \{ C_1, \dots, C_k \}$, and for $i = 1, \dots, k$ let $\mathcal{C}_i$ be the partition of C_i induced by $\mathcal{P}'$. Then

$$\begin{aligned} q(\mathcal{P}) &= \sum_{i<j} q(C_i, C_j) \\ &\underset{\text{(i)}}{\leqslant} \sum_{i<j} q(\mathcal{C}_i, \mathcal{C}_j) \\ &\leqslant q(\mathcal{P}')\,, \end{aligned}$$

since $q(\mathcal{P}') = \sum_i q(\mathcal{C}_i) + \sum_{i<j} q(\mathcal{C}_i, \mathcal{C}_j)$. □

Next, we show that refining a partition by subpartitioning an irregular pair of partition sets increases the value of q a little; since we are dealing here with a single pair only, the amount of this increase will still be less than any constant.

Lemma 7.2.3. *Let $\epsilon > 0$, and let $C, D \subseteq V$ be disjoint. If (C, D) is not ϵ-regular, then there are partitions $\mathcal{C} = (C_1, C_2)$ of C and $\mathcal{D} = (D_1, D_2)$ of D such that*

$$q(\mathcal{C}, \mathcal{D}) \geqslant q(C, D) + \epsilon^4 \frac{|C|\,|D|}{n^2}\,.$$

Proof. Suppose (C, D) is not ϵ-regular. Then there are sets $C_1 \subseteq C$ and $D_1 \subseteq D$ with $|C_1| > \epsilon\,|C|$ and $|D_1| > \epsilon\,|D|$ such that

$$|\eta| > \epsilon \tag{2}$$

η for $\eta := d(C_1, D_1) - d(C, D)$. Let $\mathcal{C} := \{ C_1, C_2 \}$ and $\mathcal{D} := \{ D_1, D_2 \}$, where $C_2 := C \setminus C_1$ and $D_2 := D \setminus D_1$.

Let us show that $\mathcal{C}$ and $\mathcal{D}$ satisfy the conclusion of the lemma. We shall write $c_i := |C_i|$, $d_i := |D_i|$, $e_{ij} := \|C_i, D_j\|$, $c := |C|$, $d := |D|$ and $e := \|C, D\|$. As in the proof of Lemma 7.2.2, c_i, d_i, e_{ij} c, d, e

$$\begin{aligned} q(\mathcal{C},\mathcal{D}) &= \frac{1}{n^2}\sum_{i,j}\frac{e_{ij}^2}{c_i d_j} \\ &= \frac{1}{n^2}\left(\frac{e_{11}^2}{c_1 d_1} + \sum_{i+j>2}\frac{e_{ij}^2}{c_i d_j}\right) \\ &\underset{(1)}{\geqslant} \frac{1}{n^2}\left(\frac{e_{11}^2}{c_1 d_1} + \frac{(e-e_{11})^2}{cd - c_1 d_1}\right). \end{aligned}$$

By definition of η, we have $e_{11} = c_1 d_1 e/cd + \eta c_1 d_1$, so

$$\begin{aligned} n^2\, q(\mathcal{C},\mathcal{D}) \geqslant\ & \frac{1}{c_1 d_1}\left(\frac{c_1 d_1 e}{cd} + \eta c_1 d_1\right)^2 \\ & + \frac{1}{cd - c_1 d_1}\left(\frac{cd - c_1 d_1}{cd} e - \eta c_1 d_1\right)^2 \\ =\ & \frac{c_1 d_1 e^2}{c^2 d^2} + \frac{2e\eta c_1 d_1}{cd} + \eta^2 c_1 d_1 \\ & + \frac{cd - c_1 d_1}{c^2 d^2} e^2 - \frac{2e\eta c_1 d_1}{cd} + \frac{\eta^2 c_1^2 d_1^2}{cd - c_1 d_1} \\ \geqslant\ & \frac{e^2}{cd} + \eta^2 c_1 d_1 \\ \underset{(2)}{\geqslant}\ & \frac{e^2}{cd} + \epsilon^4 cd \end{aligned}$$

since $c_1 \geqslant \epsilon c$ and $d_1 \geqslant \epsilon d$ by the choice of C_1 and D_1. □

Finally, we show that if a partition has enough irregular pairs of partition sets to fall short of the definition of an ϵ-regular partition, then subpartitioning all those pairs at once results in an increase of q by a constant:

Lemma 7.2.4. *Let $0 < \epsilon \leqslant 1/4$, and let $\mathcal{P} = \{ C_0, C_1, \ldots, C_k \}$ be a partition of V, with exceptional set C_0 of size $|C_0| \leqslant \epsilon n$ and $|C_1| = \ldots = |C_k| =: c$. If $\mathcal{P}$ is not ϵ-regular, then there is a partition $\mathcal{P}' = \{ C_0', C_1', \ldots, C_\ell' \}$ of V with exceptional set C_0', where $k \leqslant \ell \leqslant k4^k$, such that $|C_0'| \leqslant |C_0| + n/2^k$, all other sets C_i' have equal size, and* c

$$q(\mathcal{P}') \geqslant q(\mathcal{P}) + \epsilon^5/2\,.$$

$\mathcal{C}_{ij}$ *Proof.* For all $1 \leqslant i < j \leqslant k$, let us define a partition $\mathcal{C}_{ij}$ of C_i and a partition $\mathcal{C}_{ji}$ of C_j, as follows. If the pair (C_i, C_j) is ϵ-regular, we let $\mathcal{C}_{ij} := \{ C_i \}$ and $\mathcal{C}_{ji} := \{ C_j \}$. If not, then by Lemma 7.2.3 there are partitions $\mathcal{C}_{ij}$ of C_i and $\mathcal{C}_{ji}$ of C_j with $|\mathcal{C}_{ij}| = |\mathcal{C}_{ji}| = 2$ and

$$q(\mathcal{C}_{ij}, \mathcal{C}_{ji}) \geqslant q(C_i, C_j) + \epsilon^4 \frac{|C_i|\,|C_j|}{n^2} = q(C_i, C_j) + \frac{\epsilon^4 c^2}{n^2}\,. \tag{3}$$

$\mathcal{C}_i$ For each $i = 1, \dots, k$, let $\mathcal{C}_i$ be the unique minimal partition of C_i that refines every partition $\mathcal{C}_{ij}$ with $j \neq i$. (In other words, if we consider two elements of C_i as equivalent whenever they lie in the same partition set of $\mathcal{C}_{ij}$ for every $j \neq i$, then $\mathcal{C}_i$ is the set of equivalence classes.) Thus, $|\mathcal{C}_i| \leqslant 2^{k-1}$. Now consider the partition

$\mathcal{C}$
$$\mathcal{C} := \{ C_0 \} \cup \bigcup_{i=1}^{k} \mathcal{C}_i$$

of V, with C_0 as exceptional set. Then $\mathcal{C}$ refines $\mathcal{P}$, and

$$k \leqslant |\mathcal{C}| \leqslant k 2^k. \tag{4}$$

$\mathcal{C}_0$ Let $\mathcal{C}_0 := \big\{ \{ v \} : v \in C_0 \big\}$. Now if $\mathcal{P}$ is not ϵ-regular, then for more than ϵk^2 of the pairs (C_i, C_j) with $1 \leqslant i < j \leqslant k$ the partition $\mathcal{C}_{ij}$ is non-trivial. Hence, by our definition of q for partitions with exceptional set, and Lemma 7.2.2 (i),

$$\begin{aligned}
q(\mathcal{C}) &= \sum_{1 \leqslant i < j} q(\mathcal{C}_i, \mathcal{C}_j) + \sum_{1 \leqslant i} q(\mathcal{C}_0, \mathcal{C}_i) + \sum_{0 \leqslant i} q(\mathcal{C}_i) \\
&\geqslant \sum_{1 \leqslant i < j} q(\mathcal{C}_{ij}, \mathcal{C}_{ji}) + \sum_{1 \leqslant i} q\big(\mathcal{C}_0, \{ C_i \}\big) + q(\mathcal{C}_0) \\
&\underset{(3)}{\geqslant} \sum_{1 \leqslant i < j} q(C_i, C_j) + \epsilon k^2 \frac{\epsilon^4 c^2}{n^2} + \sum_{1 \leqslant i} q\big(\mathcal{C}_0, \{ C_i \}\big) + q(\mathcal{C}_0) \\
&= q(\mathcal{P}) + \epsilon^5 \left(\frac{kc}{n} \right)^2 \\
&\geqslant q(\mathcal{P}) + \epsilon^5/2\,.
\end{aligned}$$

(For the last inequality, recall that $|C_0| \leqslant \epsilon n \leqslant \frac{1}{4} n$, so $kc \geqslant \frac{3}{4} n$.)

In order to turn $\mathcal{C}$ into our desired partition $\mathcal{P}'$, all that remains to do is to cut its sets up into pieces of some common size, small enough that all remaining vertices can be collected into the exceptional set without making this too large. Let $C'_1, \dots, C'_\ell$ be a maximal collection of dis-
d joint sets of size $d := \lfloor c/4^k \rfloor$ such that each C'_i is contained in some

$C \in \mathcal{C} \setminus \{ C_0 \}$, and put $C_0' := V \setminus \bigcup C_i'$. Then $\mathcal{P}' = \{ C_0', C_1', \dots, C_\ell' \}$ is indeed a partition of V. Moreover, $\tilde{\mathcal{P}}'$ refines $\tilde{\mathcal{C}}$, so $\mathcal{P}'$

$$q(\mathcal{P}') \geqslant q(\mathcal{C}) \geqslant q(\mathcal{P}) + \epsilon^5/2$$

by Lemma 7.2.2 (ii). Since each set $C_i' \neq C_0'$ is also contained in one of the sets $C_1, \dots, C_k$, but no more than 4^k sets C_i' can lie inside the same C_j (by the choice of d), we also have $k \leqslant \ell \leqslant k4^k$ as required. Finally, the sets $C_1', \dots, C_\ell'$ use all but at most d vertices from each set $C \neq C_0$ of $\mathcal{C}$. Hence,

$$\begin{aligned} |C_0'| &\leqslant |C_0| + d\,|\mathcal{C}| \\ &\underset{(4)}{\leqslant} |C_0| + \frac{c}{4^k} k 2^k \\ &= |C_0| + ck/2^k \\ &\leqslant |C_0| + n/2^k. \end{aligned}$$

□

The proof of the regularity lemma now follows easily by repeated application of Lemma 7.2.4:

Proof of Lemma 7.2.1. Let $\epsilon > 0$ and $m \geqslant 1$ be given; without loss of generality, $\epsilon \leqslant 1/4$. Let $s := 2/\epsilon^5$. This number s is an upper bound on the number of iterations of Lemma 7.2.4 that can be applied to a partition of a graph before it becomes ϵ-regular; recall that $q(\mathcal{P}) \leqslant 1$ for all partitions $\mathcal{P}$. ϵ, m s

There is one formal requirement which a partition $\{ C_0, C_1, \dots, C_k \}$ with $|C_1| = \dots = |C_k|$ has to satisfy before Lemma 7.2.4 can be (re-) applied: the size $|C_0|$ of its exceptional set must not exceed ϵn. With each iteration of the lemma, however, the size of the exceptional set can grow by up to $n/2^k$. (More precisely, by up to $n/2^\ell$, where ℓ is the number of other sets in the current partition; but $\ell \geqslant k$ by the lemma, so $n/2^k$ is certainly an upper bound for the increase.) We thus want to choose k large enough that even s increments of $n/2^k$ add up to at most $\frac{1}{2}\epsilon n$, and n large enough that, for any initial value of $|C_0| < k$, we have $|C_0| \leqslant \frac{1}{2}\epsilon n$. (If we give our starting partition k non-exceptional sets $C_1, \dots, C_k$, we should allow an initial size of up to k for C_0, to be able to achieve $|C_1| = \dots = |C_k|$.)

So let $k \geqslant m$ be large enough that $2^{k-1} \geqslant s/\epsilon$. Then $s/2^k \leqslant \epsilon/2$, and hence k

$$k + \frac{s}{2^k} n \leqslant \epsilon n \qquad (5)$$

whenever $k/n \leqslant \epsilon/2$, i.e. for all $n \geqslant 2k/\epsilon$.

Let us now choose M. This should be an upper bound on the number of (non-exceptional) sets in our partition after up to s iterations

of Lemma 7.2.4, where in each iteration this number may grow from its current value r to at most $r4^r$. So let f be the function $x \mapsto x4^x$, and
M take $M := \max\{\, f^s(k), 2k/\epsilon \,\}$; the second term in the maximum ensures that any $n \geqslant M$ is large enough to satisfy (5).

We finally have to show that every graph $G = (V, E)$ of order at least m has an ϵ-regular partition $\{\, V_0, V_1, \dots, V_k \,\}$ with $m \leqslant k \leqslant M$. So
n let G be given, and let $n := |G|$. If $n \leqslant M$, we partition G into $k := n$ singletons, choosing $V_0 := \emptyset$ and $|V_1| = \ldots = |V_k| = 1$. This partition of G is clearly ϵ-regular. Suppose now that $n > M$. Let $C_0 \subseteq V$ be minimal such that k divides $|V \setminus C_0|$, and let $\{\, C_1, \dots, C_k \,\}$ be any partition of $V \setminus C_0$ into sets of equal size. Then $|C_0| < k$, and hence $|C_0| \leqslant \epsilon n$ by (5). Starting with $\{\, C_0, C_1, \dots, C_k \,\}$ we apply Lemma 7.2.4 again and again, until the partition of G obtained is ϵ-regular; this will happen after at most s iterations, since by (5) the size of the exceptional set in the partitions stays below ϵn, so the lemma could indeed be reapplied up to the theoretical maximum of s times. □