Course Notes

Part I

Probabilistic Combinatorics

and

Algorithms

J. A. Verstraete
Department of Mathematics
University of California San Diego
9500 Gilman Drive
La Jolla California 92037-0112

jacques@ucsd.edu

1 The Regularity Lemma

Szemerédi's Regularity Lemma tells us that every graph can be partitioned into a constant number of sets of vertices in such a way that for most of the pairs of sets in the partition, the bipartite graph of edges between them has many of the properties that one would expect in a random bipartite graph with appropriate edge probability. This is in contrast to the statements of Ramsey theory which ensure that logarithmic sized subsets of vertices need not exhibit any randomness at all – at least one will be empty or complete by Ramsey's Theorem. Szemerédi originally used his lemma to prove his celebrated theorem that sets of integers of positive density contain arbitrarily long arithmetic progressions. Since the first version of the regularity lemma in 1975, the lemma has been used to a great extent in many different areas of mathematics, in particular, in Green and Tao's proof that the primes contain arbitrarily long arithmetic progressions. The lemma has also been generalized to hypergraphs, where the lemma (or rather, theorem) lies in finding a general framework for a number of different problems in combinatorial number theory.

1.1 Regular pairs

A crucial property of random graphs we shall use is that if X and Y are disjoint subsets in $G_{n,p}$, and X and Y are not too small, then the number of edges between X and Y is highly concentrated around $p|X||Y|$, for fixed p.

Let X and Y be parts of a bipartite graph. We define the density of the pair (X, Y) to be

$$d(X, Y) = e(X, Y)/|X||Y|.$$

A pair (X, Y) is said to be ε-regular if, for all $X^* \subset X$ and $Y^* \subset Y$ with $|X^*| > \varepsilon|X|$ and $|Y^*| > \varepsilon|Y|$,

$$|d(X, Y) - d(X^*, Y^*)| \leq \varepsilon.$$

As an exercise in the use of the Chernoff Bound, one should show that if we consider a random bipartite graph $G(X, Y, p)$ with parts X and Y and edge probability p, then the pair (X, Y) is almost surely ε-regular of density $d \in [p - \delta, p + \delta]$, for any ε and $\delta > 0$. A partition $(V_0, V_1, V_2, \ldots, V_k)$ of a graph G is called ε-regular if $|V_0| \leq \varepsilon n$, $|V_i| = |V_j| = \varepsilon n$ for all $i, j \geq 1$, and (V_i, V_j) is ε-regular for all but $\varepsilon\binom{k}{2}$ pairs $\{i, j\} \subset \{1, 2, \ldots, k\}$. The Regularity Lemma may be stated as follows:

Lemma 1 *Let m be an integer and let $\varepsilon > 0$. Then there exists an integer $M(m, \varepsilon)$ such that if G is a graph of order at least $M(m, \varepsilon)$, then G has an ε-regular partition into $m \leq k \leq M(m, \varepsilon)$ sets.*

In other words, every graph has a partition into equicardinal sets in which most pairs of sets are ε-regular. Note that we may redistribute the vertices of V_0 into the sets $V_1, V_2, \ldots, V_k$ without damaging the regularity, and therefore we can assume if we wish that V_0 is empty and

$$|V_1| \leq |V_2| \leq \cdots \leq |V_k| \leq |V_1| + 1.$$

For convenience, we refer to this, too, as an ε-regular partition.

1.2 Applications of the Regularity Lemma

Before proving the Regularity Lemma, let us give a number of well-known applications of it. Let us begin with some elementary properties of regular pairs.

Lemma 2 *Let (X, Y) be an ε-regular pair of density d. If $Y^* \subset Y$ and $|Y^*| > \varepsilon|Y|$, then*

$$|\{x \in X : |N(x) \cap Y^*| < (d - \varepsilon)|Y^*|\}| < \varepsilon|X|.$$

Proof ▷ Let $X^* = \{x \in X : |\Gamma(x) \cap Y^*| < (d - \varepsilon)|Y^*|\}$. Then $d(X^*, Y^*) < (d - \varepsilon)$. By ε-regularity, therefore, $|X^*| < \varepsilon|X|$, as required. ∎

A well-known application of the regularity lemma is to the embedding of graphs into dense graphs. We will now state a key "embedding lemma" which, in conjunction with the regularity lemma, can be used to great effect.

1.2.1 Counting Embeddings of Graphs

An embedding of a graph G into a graph H is a graph isomorphism from H to a subgraph of G isomorphic to H. In general, we write $H \to G$ for an embedding of H into G. We are interested in determining $\|H \to G\|$, the number of isomorphic copies of H in G. We need the following notation. Let $G(t)$ denote the inflation of t: i.e. replace vertices with independent sets of t vertices, called clusters and a complete bipartite graph between t-sets whenever the corresponding vertices were joined in G.

Lemma 3 *Let R be a graph on $\{1, 2, \ldots, k\}$ and let H be a subgraph of $R(t)$ of maximum degree Δ. Then, for every pair of positive real numbers d and ε and each integer ℓ satisfying*

$$\begin{aligned} \varepsilon < \eta &= \frac{(d-\varepsilon)^\Delta}{\Delta + 2} \\ (t-1) &\leq \eta\ell, \end{aligned}$$

the graph G obtained by replacing $V(R)$ with disjoint sets $V_1, \ldots, V_k$ of size ℓ, and an ε-regular pair (V_i, V_j) of density d for all edges $ij \in R$, contains $(\eta\ell)^{|H|}$ isomorphic copies of H.

Proof ▷ For convenience, let $V(H) = \{x_1, x_2, \ldots, x_h\}$, let the clusters of $R(t)$ be $U_1, U_2, \ldots, U_k$, and $\delta = (d - \varepsilon)$. Let $\psi : \{1, 2, \ldots, h\} \to \{1, 2, \ldots, k\}$ be defined by $\psi(j) = i$ iff $x_j \in U_i$. Initially, let $C_{1,j} = V_{\psi(j)}$ for $j = 1, 2, \ldots, k$. Suppose we have defined $C_{i,j}$ for $j \geq i \geq 1$. Define $C_{i+1,i+1}$ as follows: choose $v_i \in C_{i,i}$ with

$$d(v_i, C_{i,j}) > \delta|C_{i,j}|$$

for every $j > i$ such that $x_i x_j \in E(H)$. Let $C_{i+1,j} = C_{i,j} \cap N(v_i)$ if $x_i x_j \in E(H)$ and $C_{i+1,j} = C_{i,j}$ otherwise. Then $|C_{i,j}|$ is reduced by a factor δ at most Δ times, since H has degree at most Δ. If $|C_{i,i}| \geq \varepsilon\ell$ for all i, then Lemma 2 shows that for all i there is a set of at least $(1-\varepsilon)^{\Delta}\ell \geq \ell - \Delta\varepsilon\ell$ choices of $v_i \in V_{\psi(i)}$. The number of these choices which are in $C_{i,i}$ is at least $|C_{i,i}| - \Delta\varepsilon\ell - (t-1)$. Since $|C_{i,i}| \geq \delta^{\Delta}$, it is enough that

$$\delta^{\Delta} - \Delta\varepsilon\ell - (t-1) \geq \eta\ell.$$

This is satisfied under the conditions $t - 1 < \eta\ell$ and $\varepsilon < \eta$. ∎

The graph G in the key lemma is of density roughly d. In the application of the lemma, we usually start with the graph G, find an ε-regular partition using the regularity lemma, form the graph $R(t)$, and then lift back the structure of $R(t)$ to G using the key lemma. In the statement of the regularity lemma, almost all the pairs were ε-regular and it is not hard to see that many of the ε-regular pairs have density close to the density of the whole graph. The cluster graph $R(t)$ formed from our host graph will therefore have density at least about d. The following important theorem was proved by Erdős and Stone (1966):

Theorem 4 *Let $t_{r-1}(n)$ be the number of edges in an $(r-1)$-partite n-vertex graph with as many edges as possible (i.e. all the parts have size $\lfloor n/(r-1) \rfloor$ or $\lceil n/(r-1) \rceil$. Let G be a graph with $t_{r-1}(n) + \gamma n^2$ edges. Then there exists an $n(r,t)$ such that if $n > n(r,t)$, then G contains a $K_r(t)$.*

Proof ▷ Let G be a graph with $t_{r-1}(n) + \gamma\binom{n}{2}$ edges. In order to apply the embedding lemma, we must fix some parameters. We will apply the embedding lemma with parameters $\Delta = rt$ and $d = \gamma$ and $m > 1/\gamma$. A posteriori, we will also require

$$\gamma - \frac{1}{m} - \varepsilon^2 - 4\varepsilon > 0.$$

Note that this is possible since $\gamma - \frac{1}{m} > 0$. Let us assume $n > N(\varepsilon, m)$ and $n > M(\varepsilon, m)t/\varepsilon(1-\varepsilon)$. By the Regularity Lemma, G has an ε-regular partition $(V_0, V_1, \ldots, V_k)$, where $m \leq k \leq M(\varepsilon, m)$. Form a new graph R with $V(R) = \{v_1, v_2, \ldots, v_k\}$ in which we join v_i to v_j whenever (V_i, V_j) is an ε-regular pair of density γ in G. Since

$$\varepsilon\ell = \varepsilon\frac{n - |V_0|}{k} \geq \varepsilon(1-\varepsilon)\frac{n}{M} > t,$$

and $\varepsilon < \eta$, we need only show that $K_r \subset R$ to finish the proof. Now the number of edges in V_0 is at most $\binom{\varepsilon n}{2}$, and the number of edges in each V_i is at most $\binom{\ell}{2}$. The number of edges not in regular pairs is at most

$$f(\varepsilon, k, n) = \binom{\varepsilon n}{2} + \varepsilon(1-\varepsilon)n^2 + k\binom{\ell}{2}.$$

The number of edges in pairs of density less than γ is at most $\binom{k}{2}\gamma\ell^2$. Therefore

$$t_{r-1}(n) + \gamma n^2 \leq e(G) \leq f(\varepsilon, k, n) + \binom{k}{2}\gamma\ell^2 + e(R)\ell^2.$$

It follows that

$$e(R) \geq \binom{k}{2} \left(2t_{r-1}(n)/n^2 + \gamma - 1/m - \varepsilon^2 - 4\varepsilon\right).$$

By our choice of parameters, we have $e(R) > t_{r-1}(k)$, and therefore $K_r \subset R$, as required. ∎

Let $\mathcal{F}$ be a family of graphs, and let $\mathrm{ex}(n, \mathcal{F})$ denote the maximum number of edges in a graph on n vertices which does not contain any graph in H as a subgraph. In other words, we this is the maximum number of edges in an $\mathcal{F}$-free graph on n vertices. This number is called the Turán number of $\mathcal{F}$, or the extremal number of $\mathcal{F}$. The Erdős-Stone-Simonovits Theorem determines asymptotically the Turán numbers of all graphs of chromatic number at least three. This theorem is a short step from the Erdős-Stone Theorem:

Theorem 5 *Let $\mathcal{F}$ be a non-empty family of graphs of chromatic number $r \geq 3$. Then $ex(n, \mathcal{F}) \sim ex(n, K_r)$.*

The case $r = 2$, bipartite graphs, is a much harder case to deal with. We shall return to this later.

1.2.2 Bounded degree Ramsey numbers

For any graph G, the Ramsey number $R(G, G)$ is the maximum number n such that there exists an edge-colouring of the complete graph K_n in which no subgraph $G \subset K_n$ is monochromatic. For example, if $G = K_k$, then

$$k2^{k/2} \leq R(K_k, K_k) \leq 4^k/\sqrt{\pi k}$$

and it is a notorious open problem to determine $\lim_{k\to\infty} R(K_k, K_k)^{1/k}$. In this section, we present a result of Chvátal, Rödl, Szemerédi and Trotter (1981), which states that bounded degree graphs have linear Ramsey numbers:

Theorem 6 *Let $d > 0$ and let G be a graph of maximum degree at most d. Then there exists a constant $c = c(d)$ such that $R(G, G) \leq c(d)|G|$.*

Proof ▷ Colour K_n red and blue, and let H denote the red graph. Take an ε-regular partition $(V_0, V_1, \dots, V_k)$, and let R be the graph on $\{v_1, v_2, \dots, v_k\}$ formed by joining v_i to v_j if (V_i, V_j) is ε-regular. Since the partition was ε-regular, R has at least $(1-\varepsilon)\binom{k}{2}$ edges and, if $\varepsilon < 1/r-1$, contains K_r, by Turán's Theorem. We will choose $r = R(K_{d+1}, K_{d+1})$, the Ramsey number for K_{d+1}. We now recolour the edges of $K_r \subset R$ green or yellow, according as the density of the corresponding pair (V_i, V_j) is at least or more than $1/2$. By the choice of $r = R(d+1, d+1)$, there is a monochromatic K_{d+1} or monochromatic K_{d+1} in this coloured $K_r \subset R$. By the embedding lemma, applied with $d = 1/2$ and $t = 1$, this means that either the blue graph or the red graph contains a K_r. ∎

1.2.3 Arithmetic Progressions

The upper density $\overline{d}(A)$ of a set $A \subset \mathbb{Z}$ is

$$\overline{d}(A) = \limsup_{n\to\infty} \frac{|A \cap [n]|}{n}$$

and the lower density, $\underline{d}(A)$ is defined similarly. We say that a set A has density d if $\overline{d}(A) = \underline{d}(A) = d$, and this is denoted $d(A)$. In the section on analytic number theory, we gave motivation for this definition of the density of a set. A natural question is, what does positive density imply about a set of integers? For example, what substructures must it contain?

van der Waerden proved that whenever the natural numbers are partitioned into finitely many sets, one of the sets must contain a k-term arithmetic progression. Erdős and Turán (1936) conjectured that any set of positive density contains a k-term arithmetic progression. A famous theorem of Roth (1953) states that every subset of $\{1, 2, \ldots, n\}$ of positive density contains a three-term arithmetic progression, thus answering Erdős and Turán's question in the case $k = 3$. Szemerédi (1975) and Fürstenburg (1977) independently gave the final answer. Szemerédi made use of the regularity lemma to prove his theorem. We show how to get the case $k = 3$ from the regularity lemma, as done by Ruzsa and Szemerédi (1978). We also give a more recent proof of Gowers (2004). The following two lemmas are needed:

Lemma 7 *For all $\varepsilon > 0$, there exists $\delta > 0$ such that whenever G is an n-vertex graph with at most $\delta\binom{n}{3}$ triangles, there exists a set of $\varepsilon\binom{n}{2}$ edges of G whose removal gives a triangle-free graph.*

Lemma 8 *For all $\varepsilon > 0$, there exists $\delta > 0$ such that if G is an n-vertex graph of average degree $\varepsilon(n-1)$ which is properly edge-coloured with at most n colours, then G contains at least $\delta\binom{n}{2}$ paths of three edges whose first and last edges have the same colour.*

Proof ▷ Let $\eta > 0$ and let $(V_1, V_2, \ldots, V_k)$ be an η-regular partition of G where $|V_i| \leq |V_{i+1}|$ for i and η is sufficiently small relative to ε. Let R be the graph of edges $\{i, j\}$ if (V_i, V_j) is η-regular and has density at least 2η, and delete all edges of G which are not in some (V_i, V_j) with $\{i, j\} \in E(R)$ to get a graph H. Then

$$|E(H)| \geq (\varepsilon - c\eta)\binom{n}{2}$$

for an appropriate constant $c > 0$. Delete all vertices of R of degree at most $\frac{1}{4}(\varepsilon - c\eta)k$ to get a graph R^* and all corresponding sets V_i from H to get a graph H^* such that

$$|E(H^*)| \geq \frac{1}{2}(\varepsilon - c\eta)\binom{n}{2}.$$

If M_a is the set of edges of colour a in H^*, let $V_i(a) = V(M_a) \cap V_i$. Then

$$\sum_{a\in[n]} \sum_{\{i,j\}\in R^*} \min\{|V_i(a)|, |V_j(a)|\} \geq \frac{1}{4}(\varepsilon - c\eta)k|E(H^*)|.$$

Using the bound on $|E(H^*)|$, we find an $a \in [n]$ and a pair $\{i,j\} \in R^*$ such that

$$\sum_{a\in[n]} \sum_{\{i,j\}\in R^*} \min\{|V_i(a)|,|V_j(a)|\} \geq \frac{1}{8}(\varepsilon - c\eta)^2 \frac{n}{k}.$$

Provided $16\eta < (\varepsilon - c\eta)^2$, we get some $a \in [n]$ and $\{i,j\} \in R^*$ such that

$$\min\{|V_i(a)|,|V_j(a)|\} \geq 2\eta\frac{n}{k}.$$

Let $X \subset V_i(a)$ and $Y \subset V_j(a)$ be sets of size more than $\eta\frac{n}{k}$ which have no edge of M_a between them in (V_i, V_j). Since (V_i, V_j) is η-regular, of density at least 2η, there must be at least $\eta|X||Y|$ edges of H^* between X and Y. Then any one of these $\eta|X||Y|$ edges is incident with two edges of M_a, by definition of $V_i(a)$ and $V_j(a)$. So we recover $\eta|X||Y| \sim \eta n^2/k^2$ paths of length three as required. ∎

Theorem 9 *Every set of positive density contains a positive relative density of all three-term arithmetic progressions.*

Proof ▷ (First Proof) Let A be a set of positive density. Choose n such that $|A\cap\{1,2,\dots,n\}| \geq \varepsilon n$. Form a bipartite $2n$ by $3n$ graph by joining $x+a \in \{1,2,\dots,2n\}$ to $x+2a \in \{1,2,\dots,3n\}$ for all $x \in \{1,2,\dots,n\}$ and for all $a \in A$, and coloured the edge $\{x+a, x+2a\}$ with colour a. Note that this is a proper edge-colouring: if $\{x+a,x+2a\}$ and $\{y+b,y+2b\}$ are two distinct edges of the same colour, then $a = b$. This means that $y + b \neq x + a$ and $y + 2b \neq x + 2b$, so the edges do not meet. By the preceding result, there are at least $\delta\binom{n}{2}$ paths of length three in G whose first and last edges have the some colour. We show that each corresponds to an arithmetic progression in A. Fix a path with edges $\{x + c, x + 2c\}$, $\{y + b, y + 2b\}$ and $\{x + a, x + 2a\}$. By definition, we must have $x + 2c = y + 2b$ and $y + b = x + a$. Therefore $2c - 2b = a - b$ and $b + a = 2c$. Now $\{a, b, c\}$ forms an arithmetic progression of length three in A, as required.

(2nd Proof) In the first proof, we actually showed that in the H graph on $[n] \times [2n] \times [3n]$ consisting of the union of all triangles with vertex set $\{x, x+a, x+2a\}$, any triangle must be of the form $\{x, x+a, x+2a\}$, and all of these triangles are edge-disjoint. The number of triangles is $|A|n$, and therefore we have to delete at least $|A|n$ edges to destroy all the triangles. On the other hand, for any $\varepsilon > 0$, if n is large enough, then we can delete $\varepsilon\binom{n}{2}$ edges of H to obtain a triangle-free graph. Choosing ε so that $|A|n > \varepsilon\binom{n}{2}$, this is a contradiction. ∎

We can even allow the density of our set A to be zero, provided $|A\cap\{1,2,\dots,n\}| \gg n/\log^* n$ for all n. In fact, Roth's original proof uses the Hardy-Littlewood method, which is effectively better: every set A with $|A\cap\{1,2,\dots,n\}| \gg n/(\log\log n)^{500}$ contains a three-term arithmetic progression. A much more involved proof of Bourgain shows that we can actually take $|A \cap \{1,2,\dots,n\}| \gg n/(\log n)^{1/2-\eta}$ for any positive η. It is believed by some that the right answer should be of order $n/\exp^{c\sqrt{\log n}}$ for some $c > 0$, but this seems out of reach. The following construction of Behrend shows that this order of magnitude would be close to best possible.

Construction. Let $\mathbb{S}_r$ denote the sphere of radius r in $\mathbb{R}^d$. The number of vectors $(x_1, x_2, \ldots, x_d) \in \mathbb{R}^d$ with integer co-ordinates at least $r/4$ and at most $r/2$ is at least $(r/4-1)^d$. So there exists an integer $R : r^2d/16 \leq R \leq r^2d/4$ for which the number of points in the positive orthant of the sphere $\mathbb{S}_R$ is at least

$$\frac{1}{r^2 d}\left(\frac{r}{4}-1\right)^d > \frac{1}{d}\left(\frac{r}{8}\right)^{d-2}.$$

Let V be the set of such points in the positive orthant of $\mathbb{S}_r$. Then the equation $x + y = 2z$ has no solution when x and y and z are in V, since

$$\begin{aligned} x + y = 2z \qquad &\Rightarrow \|x+y\|^2 = 4\|z\|^2 \\ &\Rightarrow \|x\|^2 + \|y\|^2 + 2\|x\|\|y\|\cos\theta = 4\|z\|^2 \\ &\Rightarrow 2 + 2\cos\theta = 4 \end{aligned}$$

which is impossible since $0 < \theta \leq \pi/2$. Define the map $\psi(x) = \sum x_i r^i$. Thus x gives the r-adic representation of $\psi(x)$. Let

$$A = \{\psi(x) : x \in V\}.$$

Then ψ restricted to V is an injection, and more importantly, if $a + b = 2c$ in A then $\psi^{-1}(a)+\psi^{-1}(b) = 2\psi^{-1}(c)$ in V, which is not possible. So A contains no three-term arithmetic progression. Furthermore,

$$|A| = |V| > \frac{1}{8}\left(\frac{r}{8}\right)^d.$$

Finally $A \subset [n]$ where $n = r^{d+1}$ and therefore

$$|A| > \frac{1}{d8^d} n^{1-2/d}.$$

Selecting $d = \sqrt{\log n}$ gives a set $A \subset [n]$ of size $n/\exp(c\sqrt{\log n})$ for some $c > 0$, and A contains no arithmetic progressions.

1.3 The Proof of the regularity lemma

The proof, although quite simple in its approach, is fairly technical. Before proceeding, we need the following definition: for a pair of disjoint sets of vertices $X, Y \subset V$, we define

$$q(X,Y) = \frac{e(X,Y)^2}{|X||Y|n^2} = \frac{|X||Y|d(X,Y)}{n^2}.$$

For a partition $\mathcal{P} = (X_1, X_2, \ldots, X_k)$ of V, we write

$$q(\mathcal{P}) = \sum_{1\leq i<j\leq k} q(X_i, X_j)$$

for the index of the partition $\mathcal{P}$. Given partitions $\mathcal{X}$ of X and $\mathcal{Y}$ of Y, we define

$$q(\mathcal{X},\mathcal{Y}) = \sum_{\substack{X'\in\mathcal{X}\\ Y'\in\mathcal{Y}}} q(X',Y').$$

Lemma 10 *If $\mathcal{Q}$ is a refinement of $\mathcal{P}$, then $q(\mathcal{Q}) \geq q(\mathcal{P})$.*

Proof $\triangleright$ We use the Cauchy-Schwartz inequality in the form

$$\sum \frac{a_i^2}{b_i} \geq \frac{(\sum a_i)^2}{\sum b_i}.$$

This follows from $\sum x_i^2 \sum y_i^2 \geq \sum x_i^2 y_i^2$ by setting $x_i = a_i/\sqrt{b_i}$ and $y_i = \sqrt{b_i}$. Now if $\mathcal{X}$ and $\mathcal{Y}$ are partitions of $X, Y \in \mathcal{P}$ then, by Cauchy-Schwarz,

$$q(\mathcal{X}, \mathcal{Y}) \geq \sum_{\substack{X' \in \mathcal{X} \\ Y' \in \mathcal{Y}}} \frac{e(X', Y')^2}{|X'||Y'|n^2} \geq \frac{1}{n^2} q(X, Y).$$

It follows immediately by summing over all pairs $(\mathcal{X}, \mathcal{Y})$ such that X and Y are distinct sets in $\mathcal{P}$ and $\mathcal{X}$ and $\mathcal{Y}$ are partitions of $\mathcal{X}$ and $\mathcal{Y}$ contained in $\mathcal{Q}$, that $q(\mathcal{Q}) \geq q(\mathcal{P})$. ∎

Proof of The Regularity Lemma. Let $\mathcal{P} = (X_0, X_1, \ldots, X_m)$ be a partition of G with $|X_1| = |X_2| = \cdots = |X_m|$ and exceptional set X_0 and $|X_0| \leq \varepsilon n - n/2^m$. If $\mathcal{P}$ is ε-regular, we are done. Suppose not, and let us now find a partition $\mathcal{P}'$ as mentioned above.

We first concentrate on each irregular pair: for each non-ε-regular pair (X, Y), there exist X_1, Y_1 with $|X_1| > \varepsilon |X|$ and $|Y_1| > \varepsilon |Y|$ and $|d(X_1, Y_1) - d(X, Y)| > \varepsilon$. Let $\mathcal{X} = (X_1, X_2)$ and $\mathcal{Y} = (Y_1, Y_2)$ where $X_2 = X \backslash X_1$ and similarly for Y_2. We claim that

$$q(\mathcal{X}, \mathcal{Y}) \geq q(X, Y) + \varepsilon^4 \frac{|X||Y|}{n^2}.$$

For convenience, let $e = e(X, Y), e_{X_i Y_j} = e_{ij}$, and $x = |X|$, $y = |Y|$, $x_i = |X_i|, y_i = |Y_i|$ and $\delta = d(X, Y) - d(X_1, Y_1)$. Then

$$\begin{aligned} n^2 q(\mathcal{X}, \mathcal{Y}) &= \sum_{1 \leq i,j \leq 2} \frac{e_{ij}^2}{x_i y_j n^2} \\ &\geq \frac{e_{11}^2}{x_1 y_1} + \frac{(e - e_{11})^2}{xy - x_1 y_1} \\ &\geq \frac{e^2}{xy} + \delta^2 x_1 y_1 \\ &\geq \frac{e^2}{xy} + \varepsilon^4 xy. \end{aligned}$$

Note that in the second line we used $(*)$, and the third line requires a little computation using the definition of δ. The fourth line came from $|\delta| > \varepsilon$. This computation shows that if we are given an irregular pair, we may refine the partition on this pair to get a partition with a slightly better index. We will next take this approach for every irregular pair in the partition.

If (X_i, X_j) is not ε-regular, we will write $\mathcal{X}_{ij}$ and $\mathcal{X}_{ji}$ instead of $\mathcal{X}$ and $\mathcal{Y}$. If (X_i, X_j) is ε-regular, let $\mathcal{X}_{ij} = \{X_i\}, \mathcal{X}_{ji} = \{X_j\}$, otherwise let $\mathcal{X}_{ij}$ and $\mathcal{X}_{ji}$ be defined as above. For each i, let $\mathcal{X}_i$ be the unique partition of X_i that refines every $\mathcal{X}_{ij}$. Take $\mathcal{Q}$ to be the partition of V consisting of all elements of $\mathcal{X}_i$ for all i, together with $\{X_0\}$. Clearly $|\mathcal{X}_i| \leq 2^{m-1}$, and so $m \leq |\mathcal{Q}| \leq m2^m$. We will get $\mathcal{P}'$ from $\mathcal{Q}$ by adding vertices to the exceptional set in $\mathcal{Q}$, thereby making all sets in $\mathcal{Q}$ the same size.

Since $\mathcal{P}$ is not ε-regular, for more that εm^2 pairs (X_i, X_j), the partition $\mathcal{X}_{ij}$ is non-trivial. Let us assume there are no edges inside $\mathcal{X}_i$ for all $i \geq 0$. Then, with $\mathcal{X}_0 = \{\{x\} : x \in X_0\}$ and $x = |X_i|$:

$$\begin{aligned} q(\mathcal{Q}) &= \sum_{i<j} q(\mathcal{X}_i, \mathcal{X}_j) + \sum_{i \geq 1} q(\mathcal{X}_0, \mathcal{X}_i) \\ &\geq \sum q(X_i, X_j) + \varepsilon m^2 \frac{\varepsilon^4 x^2}{n^2} + \sum_{i \geq 1} q(\mathcal{X}_0, \{X_i\}) \\ &> q(\mathcal{P}) + \varepsilon^5. \end{aligned}$$

At the second line, we recalled the fact that $q(\mathcal{X}, \mathcal{Y}) \geq q(X, Y)$ if $\mathcal{X}$ and $\mathcal{Y}$ partition X and Y. We now take a maximal collection $X_1', X_2', \ldots, X_\ell'$ of disjoint sets of size $\lfloor x/4^m \rfloor$ so that each X_i' is contained in some non-exceptional member of $\mathcal{Q}$, and let $X_0' = V \backslash \bigcup X_i'$. Then

$$q(\mathcal{P}') \geq q(\mathcal{Q}) \geq q(\mathcal{P}) + \varepsilon^5,$$

since $\mathcal{P}'$ is a refinement of $\mathcal{Q}$. Now no more than 4^m sets X_i' can lie inside X_i, and therefore $|\mathcal{P}'| \leq m4^m$. The sets X_i' use all but $\lfloor x/2^m \rfloor$ vertices of each X_i, since there were at most 2^m parts in $\mathcal{X}_i$, so clearly

$$|X_0'| \leq |X_0| + n/2^m \leq \varepsilon n.$$

So we have found $\mathcal{P}'$ with $q(\mathcal{P}') > q(\mathcal{P}) + \varepsilon^5$, $m \leq |\mathcal{P}'| \leq m4^m$, and exceptional set $|X_0'|$. This completes the proof. ▮

1.4 Bipartite Extremal Problems

In the last section, we determined the asymptotic behaviour of $\text{ex}(n, F)$ when F is a graph of chromatic number r. In the case $r = 2$, the situation is less satisfactory, since all we know from the Erdős-Stone-Simonovits theorem is that $\text{ex}(n, H) = o(n^2)$ when H is bipartite. It seems to be a difficult problem to determine $\text{ex}(n, H)$ when H is bipartite, and even an asymptotic determination is out of reach. The proof of the following theorem of Kövari, Sós and Turán is left as an exercise:

Theorem 11 *For any integer $r \geq 2$,*

$$ex(n, K_{r,r}) \leq \frac{1}{2}(r-1)^{1/r} n^{2-1/r} + \frac{1}{2}(r-1)n.$$

One may ask what graph invariant controls the extremal number $\mathrm{ex}(n, F)$ when F is a bipartite graph. We shall find a tight upper bound for $\mathrm{ex}(n, F)$ when F has maximum degree at most r: it turns out that $\mathrm{ex}(n, F) \ll \mathrm{ex}(n, K_{r,r})$ in this case. We need the following simple lemma, whose proof is an exercise.

Lemma 12 *Let $\mathcal{S}$ be a set system of r-element subsets of an n-element set X. If*

$$\binom{n}{s}\left[\binom{s}{r} - 1\right] < |\mathcal{S}|\binom{n}{s-r}$$

then $\mathcal{S}$ contains all r-element subsets of some s-element set S.

The main theorem we prove is:

Theorem 13 *Let H be an s by t bipartite graph of maximum degree at most r in the part of size t. Then $ex(n, H) < s(t-1)^{1/r}n^{2-1/r} + rn$.*

Proof ▷ Let $f(n, r, t)$ be the expression in the upper bound. First we recall that every graph on $2n$ vertices contains an n by n bipartite graph with more than half its edges. So if $\mathrm{ex}(n, n, H)$ denotes the maximum number of edges in an n by n bipartite graph which does not contain H, then

$$\mathrm{ex}(2n, H) < 2\mathrm{ex}(n, n, H).$$

So we only need to show $\mathrm{ex}(n, n, H) \leq \frac{1}{2}f(n, r, t)$. Let $G(A, B)$ be an n by n bipartite graph with parts A and B and more than $f(n, r, t)$ edges. We aim to find $H \subset G$. The way we do it is to find a set $S \subset B$ such that every r-element subset of S has at least t common neighbours in A.

Select, uniformly and randomly, a vertex of A. Let X denote the random set of neighbours of this vertex. For a set $R \subset B$ of r distinct vertices, let A_R denote the event that R has fewer than t common neighbours in B and let Y be the number of A_R which occur. Then

$$\mathbb{P}[A_R] \leq \frac{t-1}{n}$$

so the expected number of $R \subset B$ for which A_R occurs is at most $\frac{t-1}{n}\binom{n}{r}$. Let Z be the random variable $\frac{1}{s^r}\binom{|X|}{r} - |Y|$. Using Jensen's inequality,

$$\begin{aligned}\mathbb{E}[Z] &\geq \frac{1}{s^r}\binom{\mathbb{E}[|X|]}{r} - \frac{t-1}{n}\binom{n}{r} \\ &\geq \frac{1}{s^r}\binom{f(n, r, t)/n}{r} - \frac{t-1}{n}\binom{n}{r} \\ &\geq 0.\end{aligned}$$

Therefore there is a choice of a vertex of A for which $Z \geq 0$. In other words, we can find a vertex of A whose neighbourhood, X, contains at least $(1 - s^{-r})\binom{|X|}{r}$ sets of size r, each of which has at least t common neighbours in A. By the preceding lemma, this means that we can find a subset S of X of size s all of whose r-element subsets have at least t common neighbours in A.

Let $S_1, S_2, \ldots, S_m$ be the common neighbourhoods of the r-sets $T_1, T_2, \ldots, T_m$ in S, where $m = \binom{s}{r}$. We embed vertices and their neighbourhoods, one by one. If the vertices of H in the part of size t are $v_1, v_2, \ldots, v_t$, put v_1 into S_1 and the neighbourhood of v_1 into the set of common neighbours of v_1. Suppose we have embedded $v_1, v_2, \ldots, v_i$ and $\Gamma(v_i) \subset T_{\psi(i)}$, where $\psi : \{1, 2, \ldots, t\} \to \{1, 2, \ldots, m\}$. To embed v_{i+1}, note that v_{i+1} has at most r neighbours amonsts $v_1, v_2, \ldots, v_i$, so we choose a set $T_{\psi(i+1)}$ which contains all these neighbours. Then v_{i+1} is put into $S_{\psi(i+1)} \backslash \{v_1, v_2, \ldots, v_i\}$. This set is always non-empty. ∎

It is likely that the order of magnitude $n^{2-1/r}$ is best possible, since it is widely believed that $\text{ex}(n, K_{r,t})$ has that order of magnitude for any $t \geq r$. The recent construction of norm graphs, by Kollár, Ronyai and Szábo (2000), which use algebraic geometry, show that this is indeed true if $t > (r-1)!$.

Recall that a graph is r-degenerate if every subgraph of it has a vertex of degree at most r. In other words, the graph can be constructed by recursively adding vertices of degree at most r. Erdős and Simonovits (1968) conjectured that for any r-degenerate bipartite graph H, $\text{ex}(n, H) \leq cn^{2-1/r}$ where c depends only on H. The following theorem can be used to show, at least, that $\text{ex}(n, H) \leq cn^{2-1/4r}$:

Theorem 14 *Let G be a graph on n vertices with $cn^{2-1/4r}$ edges. Then there exists two sets X and Y in G of size at least c each such that every r-element subset of X has at least c neighbours in Y and every r-element subset of Y has at least c neighbours in X.*

It is left as an exercise to prove that if H is r-degenerate, then $\text{ex}(n, H) \ll n^{2-1/4r}$.

1.5 Cycles

While it is known that $\text{ex}(n, C_{2k+1}) = \lfloor n^2/4 \rfloor$ for any $k \in \mathbb{N}$ and large enough n, the situation for even cycles is much more difficult. The case $r = 2$ of the last section shows that $\text{ex}(n, C_4) \lesssim n^{3/2}$. It turns out that this is best possible: take the graph $\mathbb{G}_q$ whose vertices set is $\mathbb{F}_q \times \mathbb{F}_q$ and in which (x, y) is joined to (w, z) if $x + w = yz$. It is a simple exercise to show that this graph has no quadrilaterals and asymptotically $\frac{1}{2}q^3$ edges as $q \to \infty$. The asymptotic value of $\text{ex}(n, C_{2k})$ is not know for any other value of $k > 2$. For $k \in \{3, 5\}$ it is known that $\text{ex}(n, C_{2k}) \asymp n^{1+1/k}$ but even the order of magnitude is not known for any other k. Constructions of graphs without short cycles are very important in coding and information theory, and are central to the study of graph spectra and graph embeddings and drawing in various metric spaces. We will return to the spectral side of things later; for now we show that $\mathrm{ex}(n, C_{2k}) \leq n^{1+1/k}$ for all $k \in \mathbb{N}$. The following simple colouring lemma is needed.

Lemma 15 *Suppose H is a graph consisting of a cycle together with a chord joining two vertices of the cycle. Let ϕ be a bicolouring of $V(H)$. Then unless ϕ is a proper colouring of H or H is monochromatic, there exists for every positive integer $\ell < |V(H)|$ a path of ℓ-edges whose endpoints have different colours.*

We leave the proof as an exercise. This lemma allows us to prove results on cycles of length zero modulo k:

Theorem 16 *Let G be a connected bipartite graph of average degree at least $4(k-1)$ and radius r. Then G contains cycles of k consecutive even lengths, the shortest of which has length at most $2r$.*

Proof ▷ Let T be a breadth first search tree in G, so that T has height r. By the pigeonhole principle, there exist two consecutive levels X_i and X_{i+1} of the tree such that the bipartite graph B_i with parts X_i and X_{i+1} has average degree at least $2(k-1)$. Pass to a subgraph of B_i of minimum degree at least $k-1$. It is not hard to see that this subgraph contains a graph H as in the preceding lemma and such that $|V(H)| \geq 2k$. Let U be the minimal subtree of T whose leaves consist of all vertices of $V(H) \cap X_i$. Then the root of U, which is the vertex closest in T to the root of T, has degree at least two in U. This implies $U = U_1 \cup U_2$ where U_1 and U_2 are trees sharing only the root of U. Let $\phi(x) = 1$ if $x \in V(U_1)$ and $\phi(x) = 2$ otherwise. Then ϕ is not a proper colouring of H, and H is not monochromatic. So for every $\ell < 2k$, there exist a path of length ℓ with ends of different colours. In particular, if a path has even length and starts at a leaf of U_1, then it must end with a leaf of U_2. So we get paths $P_1, P_2, \dots, P_k \subset G$ of even lengths in $\{2, 4, \dots, 2k-2\}$ each starting with a leaf of U_1 and ending with a leaf of U_2. Now for each P_i, there exists in U a unique path Q_i such that $P_i \cup Q_i$ is a cycle, and the length of this cycle is $|P_i| + 2h$ where h is the height of U. It follows that G contains cycles of lengths in $\{2h+2, 2h+4, \dots, 2h+2k\}$ where $h < r$. ▮

This shows that every graph of average degree at least $8(k-1)$ contains a cycle of length zero modulo k. It seems likely that the densest bipartite graphs with no cycles of length zero mod k are the complete bipartite graphs $K_{k-1,n-k+1}$ if k is even and $K_{2k-1,n-2k+1}$ when k is odd, but this is an open question. We are almost ready to prove $\mathrm{ex}(n, C_{2k}) \ll n^{1+1/k}$, except for the following lemma:

Lemma 17 *Let G be an n-vertex graph with $cn^{1+1/k}$ edges. Then G contains a subgraph of radius at most k and average degree at least c.*

Theorem 18 *For any integer $k \geq 2$,*

$$ex(n, C_{2k}) < 8(k-1)n^{1+1/k}.$$

Proof ▷ Take a maxcut of G and then a densest component W of this maxcut. Then $e(W) \geq 4(k-1)m^{1+1/k}$ where $m = |V(W)|$. By the last lemma there is a subgraph F of W of average degree at least $4(k-1)$ and radius $r \leq k$. By the last theorem, F contains cycles $C_{2h+2}, C_{2h+4}, \ldots, C_{2h+2k}$ for some $h < r$. Since $r = k$, some cycle C_{2h+2i} has length $2k$, as required. ∎

Finally we turn to the constructions for $k = 3, 5$. Let $\mathbb{G}_{q,k}$ be the bipartite graph with parts $\mathbb{F}_q^k \times \mathbb{F}_q^k$ and where $x = (x_1, x_2, \ldots, x_k)$ and $y = (y_1, y_2, \ldots, y_k)$ are joined if

$$y_i = x_i + x_1 y_1^{i-1} \text{ for all } i \in \{2, 3, \ldots, k\}.$$

Clearly this graph has q^{k+1} edges. If $\mathbb{G}_{q,k}$ contains a cycle of length $2k$, then a certain list of equations must be satisfied in a non-trivial way. For convenience, suppose we have a cycle $(x_1, y_1, x_2, y_2, \ldots, x_k, y_k)$ where $x_i = (x_{i1}, x_{i2}, \ldots, x_{ik})$ and y_i is written similarly.

Theorem 19 *For $k \in \{2, 3, 5\}$, the graph $\mathbb{G}_{q,k}$ has no cycle of length $2k$.*

Proof ▷ Using the notation above, we get the following matrix equation over $\mathbb{F}_q$:

$$\begin{pmatrix} 1 & 1 & \cdots & 1 & 1 \\ y_{11} & y_{21} & \cdots & y_{k-1,1} & y_{k,1} \\ y_{11}^2 & y_{21}^2 & \cdots & y_{k-1,1}^2 & y_{k,1}^2 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ y_{11}^{k-1} & y_{21}^{k-1} & \cdots & y_{k-1,1}^{k-1} & y_{k,1}^{k-1} \end{pmatrix} \cdot \begin{pmatrix} x_{11} - x_{12} \\ x_{12} - x_{13} \\ \vdots \\ x_{k-1,1} - x_{k,1} \\ x_{k1} - x_{11} \end{pmatrix} = 0.$$

It is not hard to check that $x_{i,1} - x_{i+1,1} \neq 0$. Since the coefficient matrix is a Vandermonde matrix, two of its columns must be equal. Reducing the equations by removing equal columns, we obtain another Vandermonde matrix equation. Once more, any differences $x_{i,1} - x_{i+1,0}$ are not allowed to be zero so two columns of the Vandermonde matrix are again equal. This continues until no entries of the coefficient vector have the form $x_{i,1} - x_{i+1,1}$. This is the same as saying that no cycle of length $2k$ will exist in $\mathbb{G}_{q,k}$ if the cycle C_k has a proper vertex colouring such that each colour appears at least twice. For $k \in \{2, 3, 5\}$, this is the case, so this proves the theorem. ∎

A very active research area is to construct k-regular graphs on n vertices with no cycles of length at most $c \log_{k-1} n$ for as large as possible a constant c. A simple theoretical bound gives $c \leq 2$ (this is known as the Moore Bound). The largest c achieved thus far is $c = \frac{4}{3}$, by Lubotsky, Phillips and Sarnak, and independently by Lazebnik, Ustimenko and Woldar. The graphs of Lubotsky, Phillips and Sarnak have many additional beautiful properties, for example, they are Ramanujan graphs (graphs with close to optimal spectral gap), and they are explicity examples of expanders and pseudorandom graphs. We shall next discuss properties of graphs which allow us to call them pseudorandom, and expander graphs.