6.5240 Sublinear Time Algorithms November 12, 2026

Lecture 17

Lecturer: Ronitt Rubinfeld *Scribe: Jakob Nogler*

1 Outline

We cover the following topics:

- In Section 2, we recall the notion of γ-regularity, the Triangle Counting Lemma and Szemerédi's Regularity Lemma, as well as how the Triangle Removal Lemma gives a property tester for triangle-freeness.
- In Section 3, we prove the Triangle Removal Lemma using Szemerédi's Regularity Lemma and the Triangle Counting Lemma.
- In Section 4, in preparation for a lower bound on the query complexity of testing triangle-freeness (next lecture), we show an additive combinatorics lemma about large subsets of $\{1, \dots, m\}$ that contain no non-trivial solution to $x_1 + x_2 = 2x_3$.

2 Recap and Some Notation

Throughout, let $G = (V, E)$ be a graph. For two disjoint, non-empty sets $A, B \subseteq V$ we denote with $E(A, B)$ the set of edges $\{u, v\} \in E$ with $u \in A$ and $v \in B$, and with $e(A, B) = |E(A, B)|$ their number. We define the *density* between A and B as

$$d(A, B) = \frac{e(A, B)}{|A| \cdot |B|}.$$

Definition 1 *For $0 \leq \gamma \leq 1$, we say that a bipartite graph $G = (A \cup B, E)$ is* γ-regular *if for all subsets $A' \subseteq A$, $B' \subseteq B$ of size at least $|A'| \geq \gamma|A|$ and $|B'| \geq \gamma|B|$, we have $|d(A, B) - d(A', B')| < \gamma$.*

This last definition tries to capture a way of a bipartite graph that is *γ-regular* looks random "up to a factor of γ". That is, the density between any two large enough subsets $A' \subseteq A$ and $B' \subseteq B$ is the same as the density between A and B, up to an additive γ. For a graph $G = (V, E)$ and two disjoint sets $A, B \subseteq V$, we say that the pair (A, B) is γ-regular if the bipartite graph $(A \cup B, E(A, B))$ is γ-regular.

The next lemma shows that if three sets are pairwise regular, then the number of triangles between them behaves (up to a constant factor) as in a random graph.

Lemma 2 (Triangle Counting Lemma) *For all $0 < \eta \leq 1/2$ there are $\gamma^{\Delta}(\eta) = \eta/2$ and $\delta^{\Delta}(\eta) = (1 - \eta)\eta^3/8 \geq \eta^3/16$ such that, for any graph $G = (V, E)$ and any pairwise disjoint sets $A, B, C \subseteq V$ such that each of the pairs (A, B), (B, C) and (A, C) is $\gamma^{\Delta}(\eta)$-regular and has density more than η, the graph G contains at least $\delta^{\Delta}(\eta) \cdot |A||B||C|$ distinct triangles with one node in each of A, B and C.*

Note that in a random graph in which every pair between A, B and C is an edge with probability η we would expect $\eta^3|A||B||C|$ triangles, so we only lose a constant factor of $1/16$ compared to a random graph.

Szemerédi's Regularity Lemma, informally, tells us that every large enough graph can be partitioned into a constant number of sets such that almost all pairs of sets are regular.

Lemma 3 (Szemerédi's Regularity Lemma) *For all $m \in \mathbb{Z}_+$ and $\varepsilon' > 0$ there is a constant $T = T(m, \varepsilon')$ such that, for any graph $G = (V, E)$ with $|V| \geq T$ nodes and for any equipartition $\mathcal{A} = \{A_1, \dots, A_m\}$ of V, there exists an equipartition $\mathcal{B} = \{B_1, \dots, B_k\}$ of V that satisfies:*

(a) $\mathcal{B}$ refines $\mathcal{A}$, *i.e., for every* $i \in \{1,\dots,k\}$ *there is* $j \in \{1,\dots,m\}$ *such that* $B_i \subseteq A_j$;

(b) *the numbers of sets of* $\mathcal{B}$ *satisfies* $m \le k \le T$; *and*

(c) *at most* $\varepsilon' \cdot \binom{k}{2}$ *many pairs* $\{B_i, B_j\}$ *for* $i \neq j$ *are not* ε'*-regular, i.e., the bipartite graph* $(B_i \cup B_j, E(B_i, B_j))$ *is not* ε'*-regular.*

Unfortunately, the constant $T(m,\varepsilon')$ is huge: it is a tower of 2's (i.e., $2^{2^{2^{\cdots}}}$) of height polynomial in $1/\varepsilon'$. Last time, we stated the Triangle Removal Lemma (Lemma 5 below), which we prove in this lecture, and we saw that it directly implies that the following algorithm is a property tester for triangle-freeness:

(1) Repeat $O(1/\delta)$ times, where $\delta = \delta(\varepsilon)$ is as in Lemma 5: pick three nodes $x, y, z \in V$ uniformly at random and check whether x, y, z form a triangle.

(2) If some triangle was found, reject; otherwise accept.

Corollary 4 *The above algorithm always accepts triangle-free graphs, and it rejects graphs that are* ε*-far from being triangle free with probability at least* $2/3$. *In particular, triangle-freeness can be tested with a number of queries that only depends on* ε *(and not on* n*).*

3 Proof of Triangle Removal Lemma

Lemma 5 (Triangle Removal Lemma) *For all* $\varepsilon > 0$ *there exists a* $\delta = \delta(\varepsilon)$ *such that for all* n*-node graphs* G *that are* ε*-far from being triangle free (i.e., at least* εn^2 *edges must be removed from* G *to eliminate all triangles), we have that* G *has at least* $\delta\binom{n}{3}$ *distinct triangles.*

Proof The idea of the proof is to apply Szemerédi's Regularity Lemma to G and then remove all the edges inside the sets of the partition, between non-regular pairs and between low-density pairs. This removes fewer than εn^2 edges, so, since G is ε-far from being triangle free, some triangle survives. Its three nodes lie in three sets that are pairwise regular and dense, so the Triangle Counting Lemma gives many triangles. Finally, since the number of sets is bounded by a constant, these are a constant fraction of all $\binom{n}{3}$ triples.

Let $\varepsilon > 0$ and consider an n-node graph $G = (V, E)$ that is ε-far from being triangle free. Set $\varepsilon' = \varepsilon/10$, $T_0 = T(5/\varepsilon, \varepsilon')$ and $\delta = 6\delta^{\Delta}(\varepsilon/5)/T_0^3$. Note that $\varepsilon' = \gamma^{\Delta}(\varepsilon/5)$, i.e., ε'-regular pairs are regular enough to apply the Triangle Counting Lemma (Lemma 2) with $\eta = \varepsilon/5$.

If $n < T_0$, then we are done: since G is ε-far from being triangle free, it contains at least one triangle, and $1 > n^3/T_0^3 \ge \frac{6}{T_0^3}\binom{n}{3} \ge \delta\binom{n}{3}$, where we use that $\binom{n}{3} \le n^3/6$ and $\delta^{\Delta}(\varepsilon/5) \le 1$. Hence, from now on we assume that $n \ge T_0$.

We begin by using Szemerédi's Regularity Lemma (Lemma 3) with $m = 5/\varepsilon$, $\varepsilon' = \varepsilon/10$ and any arbitrary equipartition of V into m sets. We denote with the $V_1, \dots, V_k$ resulting equipartition that satisfies Lemma 3(a), (b) and (c). Note that, by Lemma 3(b), we have

$$5/\varepsilon \le k \le T(5/\varepsilon, \varepsilon') \qquad \Longleftrightarrow \qquad \frac{\varepsilon n}{5} \ge \frac{n}{k} \ge \frac{n}{T(5/\varepsilon, \varepsilon')}. \tag{1}$$

For the rest of the proof, for simplicity, we will assume that n/k is an integer.

Next, we define the subgraph $G' = (V, E')$ of G obtained by removing all of the following edges of E:

(i) For all $i \in \{1,\dots,k\}$ remove all the internal edges of V_i, i.e., all the edges $\{u, v\} \in E$ with $u, v \in V_i$. Note that by doing so, we delete at most

$$\sum_{i\in\{1,\dots,k\}} |E(V_i, V_i)| \le \sum_{i\in\{1,\dots,k\}} |V_i|^2 \le n/k \cdot \sum_{i\in\{1,\dots,k\}} |V_i| \le n^2/k \le \frac{\varepsilon n^2}{5},$$

where we use that each V_i has size at most n/k and (1).

(ii) Delete all edges between non-regular pairs, i.e., for all $i, j \in \{1, \ldots, k\}$ such that $i \neq j$ and $(V_i \cup V_j, E(V_i, V_j))$ is not ε'-regular, we remove all the edges $\{u, v\} \in E$ with $u \in V_i$ and $v \in V_j$. Note that by doing so, we delete at most

$$\varepsilon' \binom{k}{2} \cdot \left(\frac{n}{k}\right)^2 \leq \varepsilon' \cdot k^2 \cdot \left(\frac{n}{k}\right)^2 \leq \frac{\varepsilon n^2}{10},$$

where we use Lemma 3(c).

(iii) Delete all edges between low-density pairs, i.e., for all $i, j \in \{1, \ldots, k\}$ such that $i \neq j$ and $d(V_i, V_j) \leq \varepsilon/5$, we remove all the edges $\{u, v\} \in E$ with $u \in V_i$ and $v \in V_j$. Note that by doing so, we delete at most

$$\binom{k}{2} \cdot \frac{\varepsilon}{5} \cdot \left(\frac{n}{k}\right)^2 \leq \frac{\varepsilon}{5} \cdot \frac{k^2}{2} \cdot \left(\frac{n}{k}\right)^2 \leq \frac{\varepsilon n^2}{10},$$

where we use in the worst case there are at most $\binom{k}{2}$ low-density pairs.

Note that we removed at most $|E| - |E'| \leq \frac{\varepsilon n^2}{5} + \frac{\varepsilon n^2}{10} + \frac{\varepsilon n^2}{10} = \frac{2\varepsilon n^2}{5} < \varepsilon n^2$ edges. Now, since G is ε-far from being triangle free, we need to remove at least εn^2 edges to make G triangle free, but $|E| - |E'| < \varepsilon n^2$, we have that G' contains a triangle x, y, z as well. Moreover, note that for every pair $i \neq j$ whose edges were not removed in Step (ii) or Step (iii), the edges between V_i and V_j are exactly the same in G and in G'. Hence, such a pair is ε'-regular and has density more than $\varepsilon/5$ also in G'.

Let $a, b, c \in \{1, \ldots, k\}$ such that $x \in V_a, y \in V_b$ and $z \in V_c$. Observe that:

- since in Step (i) we removed all the edges inside each set V_i and x, y, z is a triangle we have that a, b, c are all distinct;
- by Step (ii), the sets V_a, V_b, V_c are pairwise ε'-regular, i.e., $\gamma^{\Delta}(\varepsilon/5)$-regular; and
- by Step (iii), the sets V_a, V_b, V_c have pairwise density strictly more than $\varepsilon/5$.

This allows us to use the Triangle Counting Lemma (Lemma 2) with $\eta = \varepsilon/5$ on V_a, V_b and V_c on $G'[V_a \cup V_b \cup V_c]$ to conclude that $G'[V_a \cup V_b \cup V_c]$ (and thus also G') contains a number of triangle which is at least

$$\delta^{\Delta}(\varepsilon/5) \cdot |V_a||V_b||V_c| = \delta^{\Delta}(\varepsilon/5) \cdot \left(\frac{n}{k}\right)^3 \geq \delta^{\Delta}(\varepsilon/5) \cdot \frac{n^3}{T_0^3} \geq \frac{6\delta^{\Delta}(\varepsilon/5)}{T_0^3} \binom{n}{3} = \delta \binom{n}{3},$$

where we use (1) and $\binom{n}{3} \leq n^3/6$. Since G' is a subgraph of G, every triangle of G' is also a triangle of G, and thus G contains at least $\delta\binom{n}{3}$ distinct triangles as well. Note that δ depends only on ε: it is horribly small (because of the tower-type constant T_0), but it is a constant! This concludes the proof. ■

4 A Proof of an Additive Combinatorics Lemma

Note that the hidden constant in δ of the Triangle Removal Lemma depends on the constant $T(\cdot, \cdot)$ of Szemerédi's Regularity Lemma, which, as mentioned, has a superexponential dependency on ε. An interesting question is whether such a large dependency is necessary. Alon showed that a large dependency is indeed necessary for every non-bipartite pattern (so not just for triangles), by showing that detecting any such pattern requires a number of queries that is super-polynomial in $1/\varepsilon$. Next week, we will see this proof, and in preparation for it, we now show an additive combinatorics lemma that we will use.

The setting of this lemma is the set $M = \{1, \ldots, m\}$ of all natural numbers up to some m. We are interested in how large (with respect to m) a subset $X \subseteq M$ can be without containing any distinct

$x_1, x_2, x_3 \in X$ satisfying $x_1 + x_2 = 2x_3$. Note that this last condition is equivalent to saying that x_3 is the mean of x_1 and x_2, i.e., that x_3 is equidistant from both of them.

To give a few examples, consider the following sets:

$$X = \{1, 2, 3\} \qquad X' = \{1, 2, 4, 5, 10\} \qquad X'' = \{1, 2, 4, 8\}.$$

The first set is not valid because $1 + 3 = 2 \cdot 2$, while the other two sets X' and X'' are valid. More generally, just as for X'', one can show that the set of all powers of two in M is valid, which gives a valid subset of size at least $\log m$. However, in the next lemma we show that we can do much better.

Lemma 6 (Additive Combinatorics Lemma) *For every $m \in \mathbb{Z}_+$ there exists a subset $X \subseteq M = \{1, \ldots, m\}$ of size*

$$|X| \geq \frac{m}{e^{10\sqrt{\ln m}}}$$

such that X contains no non-trivial solution to $x_1 + x_2 = 2x_3$, i.e., there are no $x_1, x_2, x_3 \in X$ with $x_1 + x_2 = 2x_3$ other than $x_1 = x_2 = x_3$.

Proof The idea of the proof is to write numbers in base d and only pick numbers with small digits whose digit vectors all lie on the same sphere. Small digits mean that adding two numbers causes no carries, so $x_1 + x_2 = 2x_3$ must hold digit by digit, i.e., the digit vector of x_3 is the midpoint of those of x_1 and x_2; but the midpoint of two distinct points on a sphere never lies on the sphere itself.

Let $L = \ln m$. If $L \leq 100$, then $10\sqrt{L} \geq L$ and thus $m/e^{10\sqrt{L}} \leq 1$, so the set $X = \{1\}$ works. Hence, from now on we assume that $L > 100$.

We set the parameters $d = \lfloor e^{\sqrt{L}} \rfloor$ and $k = \lfloor \sqrt{L} \rfloor - 1$, so that $k + 1 \leq \sqrt{L}$ and $d^{k+1} \leq e^{\sqrt{L} \cdot \sqrt{L}} = m$. Every integer $0 \leq x < d^{k+1}$ can be written uniquely in its d-ary representation $x = \sum_{i=0}^{k} x_i d^i$ with digits $x_i \in \{0, \ldots, d-1\}$, and we identify x with its vector of digits $(x_0, \ldots, x_k)$. For every integer $B \geq 0$ we define

$$X_B = \left\{ \sum_{i=0}^{k} x_i d^i \;\middle|\; x_i < \frac{d}{2} \text{ for all } i \in \{0, \ldots, k\} \text{ and } \sum_{i=0}^{k} x_i^2 = B \right\},$$

i.e., X_B contains all the numbers whose d-ary digits are all smaller than $d/2$ and whose vector of digits lies on the sphere of radius $\sqrt{B}$.

Claim 6.1 *For every $B \geq 0$, we have $X_B \subseteq \{0, \ldots, m-1\}$.*

Proof of Claim. Every element of X_B is at most $d^{k+1} - 1 < m$. □

To get a subset of $M = \{1, \ldots, m\}$, we will use $X = \{x + 1 \mid x \in X_B\}$ instead, which is fine since $x_1 + x_2 = 2x_3$ if and only if $(x_1 + 1) + (x_2 + 1) = 2(x_3 + 1)$.

Claim 6.2 *For every $B \geq 0$, the set X_B contains no non-trivial solution to $x + y = 2z$.*

Proof of Claim. Let $x, y, z \in X_B$ with $x + y = 2z$, i.e., $\sum_i x_i d^i + \sum_i y_i d^i = 2 \sum_i z_i d^i$. Since all digits are smaller than $d/2$, we have $x_i + y_i < d$ and $2z_i < d$ for all i, i.e., there are no carries on either side. By the uniqueness of the d-ary representation, this implies $x_i + y_i = 2z_i$ for all $i \in \{0, \ldots, k\}$. Next, recall that by Jensen's inequality, for the convex function $f(a) = a^2$ we have $\frac{1}{2}(f(a_1) + f(a_2)) \geq f(\frac{a_1 + a_2}{2})$, with equality if and only if $a_1 = a_2$. Hence, for all i we obtain $x_i^2 + y_i^2 \geq 2(\frac{x_i + y_i}{2})^2 = 2z_i^2$, with equality if and only if $x_i = y_i = z_i$.

Now, suppose that x, y and z are not all equal. Then there is an index i for which the inequality above is strict, while for all other indices it holds with $\geq$. Summing up over all indices, we get

$$\underbrace{\sum_{i=0}^{k} x_i^2}_{=B} + \underbrace{\sum_{i=0}^{k} y_i^2}_{=B} > 2 \underbrace{\sum_{i=0}^{k} z_i^2}_{=B},$$

i.e., $2B > 2B$, a contradiction. Hence, $x = y = z$. □

Claim 6.3 *There is a $B \geq 0$ such that $|X_B| \geq m/e^{10\sqrt{L}}$.*

Proof of Claim. We pick B such that $|X_B|$ is maximum. There are at least $(d/2)^{k+1}$ vectors of digits that are all smaller than $d/2$, and each of them belongs to some X_B. Moreover, every such vector satisfies $\sum_{i=0}^{k} x_i^2 \leq (k+1)d^2/4$, so there are at most $(k+1)d^2$ values of B for which X_B is non-empty. Thus, by the pigeonhole principle, $|X_B| \geq (d/2)^{k+1}/((k+1)d^2) = d^{k-1}/(2^{k+1}(k+1))$. Now, since $d \geq e^{\sqrt{L}}/2$ and $k - 1 \geq \sqrt{L} - 3$, we have $d^{k-1} \geq e^{\sqrt{L}(\sqrt{L}-3)}/2^{k-1} = m \cdot e^{-3\sqrt{L}}/2^{k-1}$. Plugging this in and using $k + 1 \leq \sqrt{L} \leq e^{\sqrt{L}}$ and $2^{2k} \leq 4^{\sqrt{L}} \leq e^{2\sqrt{L}}$, we get

$$|X_B| \geq \frac{m}{e^{3\sqrt{L}} \cdot 2^{2k} \cdot (k+1)} \geq \frac{m}{e^{3\sqrt{L}} \cdot e^{2\sqrt{L}} \cdot e^{\sqrt{L}}} \geq \frac{m}{e^{10\sqrt{L}}}.$$

□

Combining Claims 6.1, 6.2 and 6.3, the set $X = \{x + 1 \mid x \in X_B\}$ for the B of Claim 6.3 is the desired set. ■

Note that in the definition of X_B, the function $\sum_i x_i^2$ can be replaced by $\sum_i f(x_i)$ for any strictly convex function f, and the same proof works.