Lecture 19

Lecturer: Ronitt Rubinfeld *Scribe: Sahil Kuchlous*

1 Introduction

In the last lecture we saw a property testing algorithm for testing triangle-freeness in dense graphs using Szemeredi's regularity lemma that depended only on ε. However, the dependence on ε was significantly worse than even exponential. In this lecture we will partially justify this by showing super-polynomial lower bounds on ε for property testing triangle-freeness in dense graphs based on results from additive combinatorics.

Theorem 1 *There exists a constant c such that any 1-sided error tester for triangle-freeness in dense graphs requires $\Omega((c/\varepsilon)^{c\log(1/\varepsilon)})$ queries.*

Note that $\Omega((c/\varepsilon)^{c\log(1/\varepsilon)})$ is worse than any poly(ε). To prove this theorem, we will need an important tool from additive combinatorics.

2 Sum-Free Sets

The goal of this section is to prove the existence of dense subsets of integers that do not contain 3-arithmetic progressions, which we will use to construct graphs that are far from triangle free but in which it takes many queries to detect a triangle.

Definition 2 (Sum-free) *A subset $X \subseteq \mathbb{Z}$ of integers is* sum-free *if there is no triple of distinct elements $x_1, x_2, x_3 \in X$ such that $x_1 + x_3 = 2x_2$.*

Lemma 3 *For all m, there exists $X \subset [m]$ such that $|X| \geq m/e^{10\sqrt{\log(m)}}$ and X is sum-free.*

To see why constructing such an X may be difficult, let's consider an example. If we try to greedily include numbers starting from 1, we get the set $\{1, 2, 4, 5, 10, \ldots\}$. However, we can not include 9, for example, because $(1, 5, 9)$ would be a bad triple. Thus, it is not obvious how dense a sum-free subset can get.

Proof [Lemma 3]

We will define parameters $d = e^{10\log m}$ and $k = \left\lfloor \frac{\log m}{\log d} \right\rfloor - 1$. Note that $k \approx \frac{\log m}{10\sqrt{\log m}} \approx \frac{\sqrt{\log m}}{10}$. Consider the set

$$X_B = \left\{ \sum_{i=0}^{k} X_i d^i \mid X_i < \frac{d}{2}, \sum_{i=0}^{k} X_i^2 = B \right\}.$$

We can think of the elements of X_B as integers $(X_k, \ldots, X_0)$ in base d, where every digit is smaller than $d/2$. This will be useful because adding values will not result in any carries. Moreover, note that the sets X_B partition all such values based on the sum of squares of their digits. Finally, note that X_i^2 can be replaced by any convex function on X_i to get a similar result. We will use these properties to show that each set X_B is sum-free.

Let us start by showing that $X_B \subset [m]$. This is because the largest value in X_B is at most

$$d^{k+1} = d^{\left\lfloor \frac{\log m}{\log d} \right\rfloor} \leq d^{\log_d m} = m^{\log_d d} = m.$$

Next, we will pick the B that maximizes $|X_B|$. Let us show that X_B must be large. Note that $|\bigcup X_B| = \sum_B |X_B| = (d/2)^{k+1}$ and the number of sets X_B is at most $(k+1)(d/2)^2 < k \cdot d^2$. Thus, the average size of X_B is at least $\frac{(d/2)^{k+1}}{(k+1)(d/2)^2}$. Substituting the values of k and d, we see that this simplifies to $m/e^{10\log m}$. Thus, for the B that maximizes $|X_B|$, we see that $|X_B| \geq m/e^{10\log m}$.

Finally, let us show that all X_B are sum-free. Consider an arbitrary triple $x, y, z \in X_B$. If $x+y = 2z$, we know that

$$\sum_{i=0}^{k} x_i d^i + \sum_{i=0}^{k} y_i d^i = 2\sum_{i=0}^{k} z_i d^i.$$

However, note that we ensured sums of values in X_B would result in no carries. Thus, this is only possible if $x_i + y_i = 2z_i$ for all $i \in [k]$. Let us show this contradicts the constraint that $\sum_{i=0}^{k} X_i^2 = B$. We know that the function $f(x) = x^2$ is convex, so by Jensen's inequality we see that

$$\frac{1}{2}(x_i^2 + y_i^2) \geq z_i^2,$$

where equality holds if and only if $x_i = y_i = z_i$. Note that Jensen's inequality holds for any convex function. We know that $x \neq y \neq z$, so there must be some $i \in [k]$ such that $\frac{1}{2}(x_i^2 + y_i^2) > z_i^2$, and for all $j \neq i$ we know that $\frac{1}{2}(x_i^2 + y_i^2) \geq z_i^2$. However, this implies that

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

so x, y and z can not be in the same set X_B, leading to a contradiction. Thus, every X_B is sum-free, so the largest X_B satisfies the lemma. ■

3 Lower Bound

To show the lower bound, we need a second tool that we will not prove.

Theorem 4 (Goldreich-Trevisan) *In the adjacency matrix model, if there exists a property tester T that makes $q(n, \varepsilon)$ (possibly adaptive) queries, then there exists a 'natural' tester T' that picks $q(n, \varepsilon)$ nodes and makes $O(q^2)$ non-adaptive queries.*

Thus, an $\Omega(q)$ lower bound for a natural tester implies an $\Omega(\sqrt{q})$ lower bound for any tester. Our goal is to find a class of graphs that is far from triangle free, but in which a natural tester can not find a triangle in $(1/\varepsilon)^{\log 1/\varepsilon}$ queries. Unfortunately, it is not true that the distance of a graph from triangle-free is equal to the number of triangles in it. For example, it is possible that a graph has nearly n triangles that all share a single edge, making its distance from triangle-free 1. Thus, we will need a more careful construction.

Let us start with a sum-free subset $X \subseteq [m]$. We will define a tripartite graph G on $V_1 = [m]$, $V_2 = [2m]$ and $V_3 = [3m]$. For every $v \in V_1$ and $x \in X$, we will add an edge from v to $v + x \in V_2$ and $v + 2x \in V_3$. Additionally, for every $v \in V_2$ and $x \in X$, we will add an edge from v to $v + x \in V_3$.

Let us analyze the properties of G. The number of vertices is $6m$ and the number of edges is $\Theta(m \cdot |X|) = \Theta(n^2/e^{10\sqrt{\log n}})$. Next, let's count the number of triangles. By our construction, it is clear that there are $m \cdot |X| = O(n^2/e^{10\sqrt{\log n}})$ triangles of the form $(v, v + x, v + 2x)$, since such a triangle exists for every $v \in V_1$ and $x \in X$. However, we can also show that these are the only triangles in G. Since G is tripartite, every triangle must contain a vertex $v_1 \in V_1$, $v_2 \in V_2$ and $v_3 \in V_3$. Let x_1 be the edge from v_1 to v_2, x_2 be the edge from v_2 to v_3 and x_3 be the edge from v_1 to v_3. Following the edge from v_1 to v_3 and the path from v_1 to v_3 via v_2, we see that $v_1 + x_1 + x_2 = v_1 + 2x_3$. This implies tat $x_1 + x_2 = 2x_3$, but since X is sum-free we know that $x_1 = x_2 = x_3$. Thus, this must be one of the triangles we identified, so G contains exactly $O(n^2/e^{10\sqrt{\log n}})$ triangles.

Next, note that the distance of G from triangle-free is at least the number of edge-disjoint triangles in G, since we must remove at least one edge from every disjoint triangle. However, we know that every triangle in G is disjoint, since if two triangles share an edge then this fixes the value of v and x, which also uniquely determines the third vertex of the triangle. Thus, the distance of G from triangle-free is $\Theta(n^2/e^{10\sqrt{\log n}})$.

Unfortunately, this distance is not sufficient; we want to find a graph that is ε-far from triangle free, but G is only $(1/e^{10\sqrt{\log n}})$-far. To fix this, we will define a new graph $G^{(s)}$ based on G as follows. Every vertex v of G will correspond to an independent set $v^{(s)}$ of s vertices in $G^{(s)}$. An edge (u, v) in G will correspond to a complete bipartite graph between $u^(s)$ and $v^{(s)}$ in $G^{(s)}$. Note that the number of vertices in $G^{(s)}$ is $\Theta(m \cdot s)$, the number of edges is $\Theta(m \cdot |X| \cdot s^2)$. While $G^{(s)}$ has a large number of triangles, these triangles are no longer disjoint. However, we can show that $G^{(s)}$ has at least $m \cdot |X| \cdot s^2$ disjoint triangles, which implies that $G^{(s)}$ is at least $\Omega(|X|/m)$-far from triangle-free. By taking $s = \Theta(n/m)$ and $m \geq (c/\varepsilon)^{c \log(1/\varepsilon)}$, we see that $G^{(s)}$ is at least ε-far from triangle-free.

Finally, we can also show that the number of triangles in $G^{(s)}$ is $\Theta(m \cdot |X| \cdot s^3) = \Theta((\varepsilon/c')^{c' \log(c'/\varepsilon)} \cdot n^3)$. This implies that if we run a natural tester on $q \leq (c''/\varepsilon)^{c'' \log(c''/\varepsilon)}$ nodes, the expected number of triangles is approximately $q^3 \cdot (\varepsilon/c')^{c' \log(c'/\varepsilon)} << 1$ (note that we are being a bit sloppy with constants here). Thus, by Markov's inequality, the probability of seeing a triangle must also be very small, so a natural algorithm cannot tell if $G^{(s)}$ is triangle-free. As mentioned earlier, combined with the Goldreich-Trevisan theorem, this completes the lower bound.