6.889 Sublinear Time Algorithms	March 6, 2019

Lecture 9

Lecturer: Ronitt Rubinfeld *Scribe: Mina Dalirrooyfard*

1 Outline

Today we discuss **lower bounds for property testing**, and in particular we show the following:

Testing triangle-freeness requires super-poly dependence on ϵ.

where we want to distinguish triangle free graphs vs graphs that are ϵ-far from being triangle-free.

2 Introduction

2.1 Context

In the previous lecture, we saw a testing algorithm for triangle freeness with **constant time** in terms of n, and very bad dependence on ϵ (in the form of towers of 2).

It is natural to ask if this dependence on ϵ is actually needed. Today we answer this question for one-sided error testers. In particular, we have that:

- If H is bipartite, then poly$(1/\epsilon)$ is enough, i.e. we have a tester in poly$(1/\epsilon)$ time.
- If H is not bipartite, then poly$(1/\epsilon)$ does not suffice.

We prove the special case where H is a triangle, which is depicted in the following theorem. Note that our model is the adjacency matrix model.

Theorem 1 *There exists a constant c such that any one-sided tester for whether graph G is triangle-free needs $(\frac{c}{\epsilon})^{c\log c/\epsilon}$ queries.*

2.2 Tools

We use two main tools to prove Theorem 1. The **first tool** is the following theorem due to Goldreich-Trevisan, which converts a canonical tester to a non-canonical tester with a blow-up in the number of queries.

Theorem 2 *Assume that there exists tester T for property P in the adjacency matrix model of graphs that uses $q(n,\epsilon)$ queries where n is the number of nodes of the graph. Then the following "natural tester" T' uses $q(n,\epsilon)^2$ queries to test P: It picks $q(n,\epsilon)$ nodes, queries the submatrix under these nodes and decides for property P.*

This theorem has an important consequence: *A lower bound of $\Omega(q')$ for a natural tester results in a lower bound of $\Omega(\sqrt{q'})$ for any tester.* This is because any tester can be converted to a natural tester with a quadratic blow-up. So if we have a tester that has complexity $o(\sqrt{q'})$, then by theorem 2, there is a natural tester with query complexity $o(\sqrt{q'})$ which contradicts the assumption of having a lower bound on natural testers.

The **second tool** is the following additive number theory lemma. We use this lemma to construct graphs that are far from being triangle free and any natural tester requires $\Omega((\frac{c}{\epsilon})^{c\log c/\epsilon})$ many queries to distinguish them from triangle free graphs.

Lemma 3 *For every natural number m, there exists $X \subseteq M = \{1,2,\dots,m\}$ of size at least $m/e^{10\sqrt{\log m}}$, with no non-trivial solution to the equation $x_1 + x_2 = 2x_3$, where a trivial solution is when $x_1 = x_2 = x_3$.*

We call a set X with the property mentioned in Lemma 3 a *sum-free* set. To give some insight into sum-free sets, we provide some examples.

- Neither of the sets $\{1, 2, 3\}$ and $\{5, 9, 13\}$ are sum-free, because $1 + 3 = 2 \times 2$ and $5 + 13 = 2 \times 9$.
- One can try constructing a sum-free set by going over numbers in increasing order and selecting ones that do not contradict the sum-freeness property. This way, for $m = 10$, we get the set $\{1, 2, 4, 5, 10\}$. However, it's not clear that for each m, how big the set that results from this approach is.
- A more clear approach is to consider the powers of 2 that are less than m. But the size of this set is $\log m$ which is too small.

3 Triangle Freeness Lower Bound

In this section we first prove Lemma 3, and then using it together with Theorem 2, we prove Theorem 1.

3.1 Proof of Lemma 3.

We first fix two constants. Let $d = e^{10\sqrt{\log m}}$, and let $k = \lfloor \frac{\log m}{\log d} \rfloor - 1$. The idea is to partition a big part of the set $M = \{1, 2, \ldots, m\}$ into sum-free sets X_B for integer B, and then argue that since the number of these sets is not big, by the pigeon-hole principle one of them must be a big set itself. For an integer B, define X_B as follows.

$$X_B = \{\sum_{i=0}^{k} x_i d^i \,|\, x_i < \frac{d}{2} \text{ for } 0 \leq i \leq k \text{ and } \sum_{i=0}^{k} x_i^2 = B\}$$

Note that if we view the integers in X_B in base d, then x_is are the "digits" of these numbers. The intuition behind the first constraint for these digits, i.e. $x_i < d/2$ is that we want the sum of each two numbers in X_B be carry-free, which is used in the proof of sum-freeness of X_B. The intuition behind the second condition also appears in the proof of sum-freeness of X_B. But before showing that X_B is sum-free, we show that it is a subset of M.

Claim 4 *For any integer B, we have $X_B \subseteq M$.*

Proof Note that the largest number in X_B is less than $\sum_{i=0}^{k} d^{i+1}/2 < d^{k+1}$. Now we have $d^{k+1} \leq d^{(\lfloor \log m/\log d \rfloor - 1)+1} \leq d^{\log_d m} = m^{\log_d d} = m$. ■

Claim 5 *X_B is sum-free.*

Proof By way of contradiction, suppose that there are integers $x, y, z \in X_B$ such that $x + y = 2z$. Writing x, y and z in base d with digits x_i, y_i and z_i, respectively for $i = 0, \ldots, k$, we have that $\sum_{i=0}^{k} x_i d^i + \sum_{i=0}^{k} y_i d^i = 2\sum_{i=0}^{k} z_i d^i$. So since we have no carries, this is equivalent to having $x_i + y_i = 2z_i$ for all $i = 0, \ldots, k$. Note that since the function $f(a) = a^2$ is convex, by Jensen's inequality we have that $f(x_i) + f(y_i) \geq 2f(z_i)$, with equality if and only if $x_i = y_i = z_i$. So $x_i^2 + y_i^2 \geq 2z_i^2$, with equality if and only if $x_i = y_i = z_i$. Since x, y and z are not all equal, we have that for some i, $x_i^2 + y_i^2 > 2z_i^2$. So $\sum_{i=0}^{k} x_i^2 + \sum_{i=0}^{k} y_i^2 > 2\sum_{i=0}^{k} z_i^2$. This is a contradiction, since $\sum_{i=0}^{k} x_i^2 = \sum_{i=0}^{k} y_i^2 = \sum_{i=0}^{k} z_i^2 = B$. ■

To finish the proof of the lemma, we first see how big B can be so that X_B is non-empty, and then we derive a bound on the size of the largest X_B. Note that $B = \sum_{i=0}^{k} x_i^2 \leq (k+1)(\frac{d}{2})^2 < kd^2$. So we only consider X_B with $B < kd^2$. Now since the largest number in X_B is at most d^{k+1}, the size of the union of the sets X_B is the following: $|\cup_{B<kd^2} X_B| \geq (\frac{d}{2})^{k+1} > (\frac{d}{2})^k$. Note that $|\cup_{B<kd^2} X_B| = \sum_{B<kd^2} |X_B|$ because these sets are disjoint. So by the pigeon-hole principle, there exists $B < kd^2$ such that $|X_B| \geq (\frac{d}{2})^k / kd^2$. Plugging in the values of d and k, we see that $|X_B| \geq \frac{m}{e^{10\sqrt{\log m}}}$.

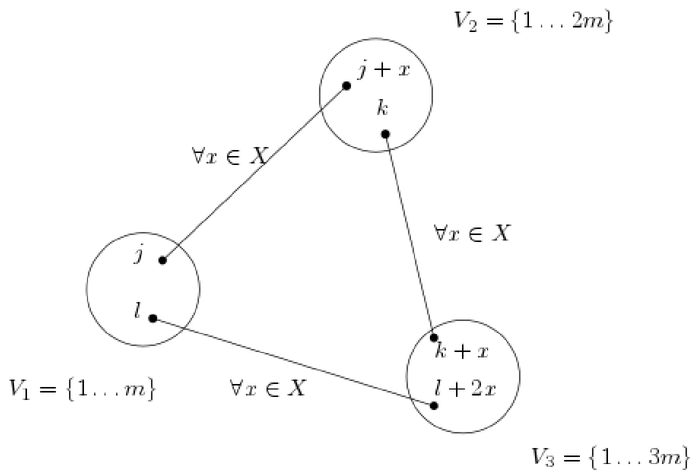

Figure 1: The graph G.

3.2 Proof of Theorem 1

Using the set $X \in \{1, \dots, m\}$ from Lemma 3, we construct a graph that is dense and far from being triangle free and we show that we need many queries to discover a triangle in it. Construct the graph G as follows: Let $V_1 = \{1, \dots, m\}$, $V_2 = \{1, \dots, 2m\}$ and $V_3 = \{1, \dots, 3m\}$ be three sets of vertices that each form an independent set. For each $x \in X$ add the following edges: Connect each $j \in V_1$ to $j + x \in V_2$. Connect each $k \in V_2$ to $k + x \in V_3$ and connect each $l \in V_1$ to $l + 2x \in V_3$. Figure 2 shows the construction.

3.2.1 G properties

The number of nodes of G is $6m$ and the number of edges is $\Theta(m|X|) = \Theta(n^2/e^{10\sqrt{logm}})$. So G is not dense enough yet. First we see how many triangles G has and how far G is from triangle freeness, and then we convert G to a dense graph.

Number of trianlges For each $j \in \{1, \dots, m\}$, there is a triangle with vertices $j, j + x, j + 2x$. we call these triangles **intended**. So the number of intended triangles is $m|X| = \Theta(n^2/e^{10\sqrt{logm}})$. We show that all the triangles in G are intended. In order to do so, first note that there are no triangles with at least two vertices in one of the sets V_1, V_2 or V_3, because there is no edge in these sets. So assume that $u \in V_1$, $v \in V_2$ and $w \in V_3$ form a triangle. Since uv is an edge, there is $x_1 \in X$ such that $v = u + x_1$. Similarly, there is x_2 and x_3, such that $w = v + x_2$ and $w = u + 2x_3$. So we have that $x_1 + x_2 = 2x_3$. Now since X is sum-free, we have that $x_1 = x_2 = x_3$, and so uvw is an intended triangle.

Number of edge-disjoint triangles We show that all intended triangles are actually edge-disjoint. Note that each intended triangle $j, j + x, j + 2x$ can be uniquely determined by the pair (j, x). Assume that the triangles $j, j + x, j + 2x$ and $j', j + x', j' + 2x'$ share an edge. No matter which edge they share, we have that $x = x'$, because the difference between endpoints of that edge in the first triangle is either x or $2x$, and in the second triangle is either x' or $2x'$. Now since they share an edge, they also share the endpoints of it, and so we see that $j = j'$.

Distance to triangle freeness In order to make G triangle free we need to remove at least one edge from each triangle. Since all triangles of G are edge-disjoint, the number of edges that we need to remove is the same as the number of trianlges which is $\Theta(n^2/e^{10\sqrt{logm}})$.

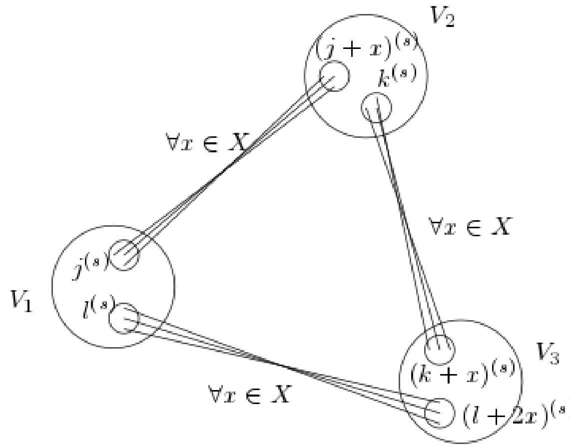

Figure 2: The graph $G^{(s)}$.

Issues with the construction First, we see that G is not $\Omega(\epsilon n^2)$-far from triangle freeness, and second, it is not dense enough. Next we fix these issues.

3.2.2 Fixed construction

Define the s**-blow-up** of G as the graph $G^{(s)}$ where each vertex u in G is replaced by an independent set $u^{(s)}$ of size s in $G^{(s)}$, and each edge uv in G is replaced by a complete bipartite graph between $u^{(s)}$ and $v^{(s)}$ in $G^{(s)}$. Note that the number of nodes in $G^{(s)}$ is $6ms$ and the number of edges is $\Theta(m|X|s^2)$. Each triangle in G is converted to s^3 trianlges in $G^{(s)}$, so there are $\Theta(m|X|s^3)$ triangles in $G^{(s)}$.

Lemma 6 *The distance of $G^{(s)}$ from triangle freeness is at least $m|X|s^2$.*

Proof We say that a triangle in $G^{(s)}$ with vertices in $u^{(s)}$, $v^{(s)}$ and $w^{(s)}$ is made from the triangle uvw in G. If two triangles in $G^{(s)}$ are made from two different triangles in G, then they are edge-disjoint, since the trianlges in G are edge-disjoint. We need to prove that we have at least $m|X|s^2$ edge-disjoint triangles in $G^{(s)}$, and in order to do so we show that each triangle in G makes s^2 edge-disjoint triangles in $G^{(s)}$. Consider the triangle uvw in G, and let $u^{(s)} = \{u_1, \ldots, u_s\}$, $v^{(s)} = \{v_1, \ldots, v_s\}$ and $w^{(s)} = \{w_1, \ldots, w_s\}$. Consider the following s^2 triangles: $T_{uvw} = \{u_i v_j w_k \mid i+j+k \equiv 0 \pmod{s}\}$. First, $|T_{uvw}| = s^2$ because i and j have s choices each and for each choice of i and j, k is uniquely determined. Moreover, suppose that $u_i v_j w_k$ and $u_{i'} v_{j'} w_{k'}$ share an edge. Then $\{i,j,k\} \cap \{i',j',k'\} \geq 2$. But since the choice of two numbers in $\{i,j,k\}$ determines the third, this means that $\{i,j,k\} = \{i',j',k'\}$, and so $u_i v_j w_k = u_{i'} v_{j'} w_{k'}$. So the triangles in T_{uvw} are edge-disjoint. ■

Finishing the proof of Theorem 1 Using the construction above, we need to set the parameters and show that this construction gives the lower bound. Given ϵ, pick m to be the largest integer satisfying $\epsilon \leq 1/e^{10\sqrt{\log m}}$. So we have $m \geq (\frac{c}{\epsilon})^{c \log c/\epsilon}$. We want the number of vertices of $G^{(s)}$ to be n, so pick $s = \lfloor \frac{n}{6m} \rfloor$ and as a result s is roughly $n(\frac{\epsilon}{c})^{c \log c/\epsilon}$ by the way we picked ϵ. To compute the number of edges, note that it is roughly $m|X|s^2$ where $|X| = \frac{m}{e^{10\sqrt{\log m}}}$. Now since $m^2 s^2 = \Theta(n)$, the number of edges is roughly $n^2/e^{10\sqrt{\log m}}$ which is ϵn^2. So the graph is **dense**. The number of triangles is $m|X|s^3$, and by plugging in the values of m, $|X|$ and s we have that it is roughly $(\frac{\epsilon}{c'})^{c' \log c'/\epsilon} n^3$ for some constant c'.

Now if we have a natural tester with sample size of $q < (\frac{c'}{\epsilon})^{c' \log c'/\epsilon}$, then we have

$$\mathbb{E}[\text{number of triangles in the sample}] < \binom{q}{3}(\frac{c'}{\epsilon})^{c' \log c'/\epsilon} << 1$$

So by Markov's inequality, the probability that we see a triangle in the sample is very small. Note that since we have one-sided error, we must find a triangle in order to output Fail. So with low probability we output fail with less than $(\frac{c'}{\epsilon})^{c' \log c'/\epsilon}$ samples, and hence we need $(\frac{c'}{\epsilon})^{c' \log c'/\epsilon}$ samples for natural testers. This gives a $(\frac{c'}{\epsilon})^{\frac{c'}{2} \log c'/\epsilon}$ lower bound for any tester by Theorem 2.