0368.416701 Sublinear Time Algorithms	December 21, 2015

Lecture 10

Lecturer: Ronitt Rubinfeld *Scribe: S. Gershtein, A. Frumkin, M. Shagam, M. Sulamy*

Last Week

We saw that certain properties of dense graphs can be tested in time independent of the size of the adjacency matrix (or the graph). We presented a property tester for $\triangle$-freeness whose running time dependence on $\frac{1}{\epsilon}$ is terrible - a tower of two's of size $\frac{1}{\epsilon}$.

Today

We will see a lower bound for Δ-free testing in the dense graph model. The bound has a super-polynomial dependence on $\frac{1}{\epsilon}$.

1 Main Theorem

1.1 Testing H-freeness

For any constant size graph H, for example a 6×6 complete bipartite graph, when we are testing G for H-freeness, that is we want to check whether G contains H as a subgraph, there is a powerful result that explicitly distinguishes the cases where the complexity is $poly(1/\epsilon)$.

Theorem 1 *If H is bipartite then there exists a 1-sided tester that performs $poly(\frac{1}{\epsilon})$ queries, whereas if H is not bipartite then no $poly(\frac{1}{\epsilon})$ number of queries suffices.*

This is a concise and surprising method of characterizing bipartiteness.
We will only prove the case where $H = \triangle$. Our tester will operate in the following manner:

- If G is $\triangle$-free output PASS.
- If G is ϵ-far from $\triangle$-free output FAIL with probability $\geq \frac{3}{4}$.

Theorem 2 *Given graph G in adjacency matrix representation, there exists a constant c, such that any 1-sided error tester for whether G is $\triangle$-free needs to perform at least $(\frac{c}{\epsilon})^{c\log(\frac{c}{\epsilon})}$ queries.*

Notice that since the algorithm has 1-sided error, it must find a triangle in order to assert that the graph is not $\triangle$-free.

1.2 Tools

Before proving the theorem we will present several useful results.

1.2.1 Goldreich-Trevisan Theorem

Given an adjacency matrix with a tester T for property P, which performs $q(n,\epsilon)$ queries. There exists a "natural tester" T' that uses $O\left(q^2\left(n,\epsilon\right)\right)$ queries:

- pick $q(n,\epsilon)$ nodes uniformly randomly
- query only the sub-matrix induced by these nodes
- decide according to acquired information

Note that the reduction preserves 1-sidedness, meaning that if T has a 1-sided error then T' also has a 1-sided error and vice versa.

Corollary: Lower bound of $\Omega(q')$ queries for natural tester gives a lower bound of $\Omega(\sqrt{q'})$ for any tester.

1.2.2 Additive Number Theory Lemma

Theorem 3 *$\forall m, \exists X \subseteq M = \{1,..,m\}$ of size at least $\frac{m}{e^{10\sqrt{\log m}}}$ with no nontrivial[1] solution to $x_1 + x_2 = 2x_3$ where $x_1, x_2, x_3 \in X$ (denoted as the "sum-free" property of X).*

Note that it is not trivial to construct such an X, since it must be both sum-free and large enough.

- Bad examples of X (X is not sum-free): $\{1, 2, 3\}$, $\{5, 9, 13\}$
- Bad example of X (X is not big enough): $\{1, 2, 4, 8, 16, ...\}$
- Good example of X: $\{1, 2, 4, 5, 10, ...\}$

Proof Let d be integer (equal to $e^{10\sqrt{\log m}}$), and $k = \left\lfloor \frac{\log m}{\log d} \right\rfloor - 1$, $\left(k \approx \frac{\log m}{10\sqrt{\log m}} \approx \frac{\log m}{10}\right)$.

Define $X_B = \left\{\sum_{i=0}^{k} x_i d^i | x_i < \frac{d}{2} \text{ for } 0 \leq i \leq k,\ \sum_{i=0}^{k} x_i^2 = B\right\}$. View $x \in M$ as represented in base d, $X = (x_0, ..., x_k)$, $x_i < d$.

Note: x_i is small, therefore summing pairs of elements in X_B does not generate a carry.

Bound on largest number in any X_B:

$$< \left(\frac{d}{2}\right) d^k + \left(\frac{d}{2}\right) d^{k-1} + ... < d^{k+1} < d^{\frac{\log m}{\log d}} = m \Rightarrow X_B \subseteq M$$

Claim 4 *X_B has the "sum-free" property, i.e., $\forall x, y, z \in X_B$ such that $x + y = 2z$ it must be that $x = y = z$.*

Proof of claim: For $\forall x, y, z \in X_B$:

$$x + y = 2z \Leftrightarrow \sum_{i=0}^{k} x_i d^i + \sum_{i=0}^{k} y_i d^i = 2\sum_{i=0}^{k} z_i d^i$$

$$\Leftrightarrow x_0 + y_0 = 2z_0$$
$$x_1 + y_1 = 2z_1$$
$$...$$
$$x_k + y_k = 2z_k$$

Subclaim If it holds that $\forall i, x_i + y_i = 2z_i$ then $\forall i, x_i^2 + y_i^2 \geq 2z_i^2$ with equality only if $x_i = y_i = z_i$.

Proof of Subclaim: $f(a) = a^2$ is strictly convex. Using Jensen's inequality:

$$\frac{\sum_{i=1}^{n} f(a_i)}{n} \geq f\left(\frac{\sum a_i}{n}\right), \text{ with equality only if } a_1 = a_2 = ... = a_n$$

$$\Rightarrow \frac{x_i^2 + y_i^2}{2} \geq \left(\frac{x_i + y_i}{2}\right)^2 = z_i^2 \Rightarrow x_i^2 + y_i^2 \geq 2z_i^2, \text{ with equality only if } x_i = y_i = z_i$$

■(subclaim)

Assume that $(x = y = z)$ does not hold, i.e. $\exists i$ such that not $(x_i = y_i = z_i)$. From the subclaim we get that:

[1] A trivial solution is defined as $x_1 = x_2 = x_3$

- For this i, $x_i^2 + y_i^2 > 2z_i^2$
- For all other i's, $x_i^2 + y_i^2 \geq 2z_i^2$

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

Recall from the definition of X_B that $B = \sum_{i=0}^{k} x_i = \sum_{i=0}^{k} y_i = \sum_{i=0}^{k} z_i$, hence we get that $B + B > 2B$ which is a contradiction! ■(claim 4)

Claim 5 *X_B can be selected such that $|X_B| \geq \frac{m}{e^{10\sqrt{\log m}}}$*

Proof of claim: Pick B to maximize $|X_B|$. How big is X_B?

$$B \leq (k+1)\left(\frac{d}{2}\right)^2 < k \cdot d^2$$

$$\sum |X_B| = \left|\bigcup X_B\right| = \left(\frac{d}{2}\right)^{k+1} > \left(\frac{d}{2}\right)^k$$

$$\Rightarrow \exists B, \text{ such that } |X_B| \geq \frac{\left(\frac{d}{2}\right)^k}{k \cdot d^2} \geq \frac{m}{e^{10\sqrt{\log m}}} \text{ (by choice of } d, k)$$

■(claim 5)

The number theory theorem immediately follows from Claims 4 and 5 ■

1.3 Proof of Main Theorem

1.3.1 Proof Outline

We will present a graph, whose construction is based on the theorem we have just proved, for which any canonical △-free tester must use $q > \left(\frac{c^*}{\epsilon}\right)^{c^* \log \frac{c^*}{\epsilon}}$ queries. From the Goldreich-Trevisan theorem follows a lower bound of $(\frac{c}{\epsilon})^{c \log(\frac{c}{\epsilon})}$ queries for any triangle-free tester. There are three constants - represented by c which is taken as the maximum of all three.

1.3.2 Initial Graph

Given a sum-free $X \subseteq \{1..m\}$ of size $\geq \frac{m}{e^{10\sqrt{\log m}}}$, we create a tri-partite graph G:

- The nodes of the graph are divided into 3 sets - $V_1 = \{1..m\}$, $V_2 = \{1..2m\}$, $V_3 = \{1..3m\}$
- Nodes from V_1 are connected to nodes from V_2 using edges $(x, x+\ell)$, $\forall x \in \{1..m\}, \ell \in X$
- Nodes from V_1 are connected to nodes from V_3 using edges $(x, x+2\ell)$, $\forall x \in \{1..m\}, \ell \in X$
- Nodes from V_2 are connected to nodes from V_3 using edges $(x+\ell, x+2\ell)$, $\forall x \in \{1..m\}, \ell \in X$
- There are no edges between two nodes of the same set

Definition 1 *Intended Triangle: Triangles created by $x, x+l, x+2l$ ($l \in X$) are defined as "intended triangles".*

Claim 6 *The total number of triangles in G is equal to $m|X|$, moreover all triangles in G are "intended triangles".*

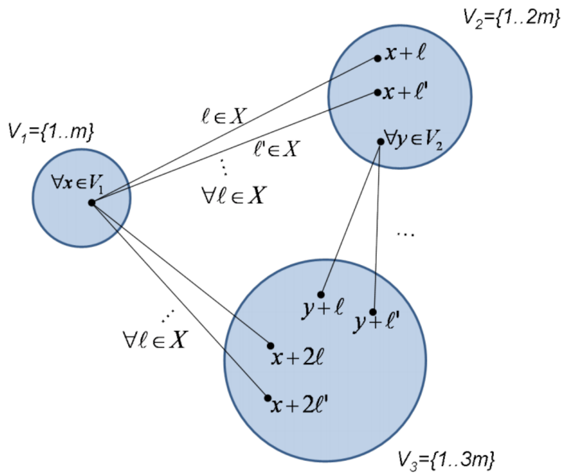

Figure 1: ***Graph G built using Theorem 3.*** *Each node in the set V_1 is connected to $|X|$ nodes in V_2 and $|X|$ nodes in V_3. Each node in the set V_2 is also connected to additional $|X|$ nodes in V_3*

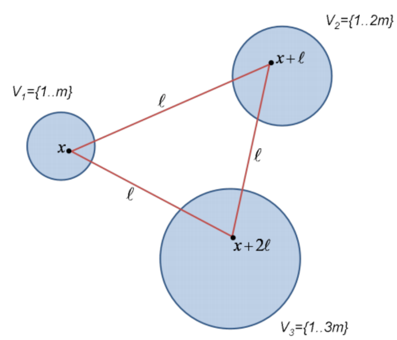

Figure 2: ***Example of an intended triangle.***

Proof of claim: Let $\triangle(u, v, w)$ be a triangle in G, connected by the edges l_1, l_2, l_3. There are no internal edges $\Rightarrow$ without loss of generality. $u \in V_1$, $v \in V_2$, $w \in V_3$
By definition of G: $u+l_1 = v$, $v+l_2 = w$, $u+2l_3 = w = v+l_2 = u+l_1+l_2 \Rightarrow 2l_3 = l_1+l_2 \Rightarrow l_1 = l_2 = l_3$ and we get that $\triangle(u, v, w)$ is an intended triangle.

- The total number of nodes in G $= 6m$
- The total number number of edges in G $=\Theta(m|x|) = \Theta\left(\frac{n^2}{e^{10\sqrt{1 \log n}}}\right)$

So $m|X|$ such intended triangles exist. ■ (claim 6)

Claim 7 *Intended triangles are edge disjoint (i.e. there are no two intended triangles with the same edge)*

Proof Idea Assume $u, v \in V_1$ are on both nodes in an intended triangle with a shared edge ℓ, as in figure 1.3.2. The triangle share two nodes so we get that $u+\ell = v+\ell$ and $u+2\ell = v+2\ell$ therefore $u = v$. Similarly we can show that any other edge in the triangle can't be shared ■

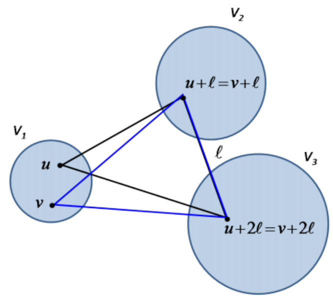

Figure 3: ***Intended triangles disjoint proof***.

Corollary 8 *Since triangles are edge disjoint and must remove ≥ 1 edges from each triangle to make graph G $\triangle$-free, then the absolute distance to $\triangle$-free is $m|X|(= \epsilon \cdot n^2)$. In other words, the distance to $\triangle$-free $= \frac{m|X|}{(6m)^2} = \Theta\left(\frac{|X|}{m}\right) = \Theta\left(\frac{1}{e^{10\sqrt{\log m}}}\right)$.*

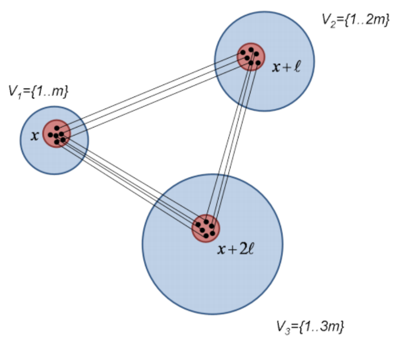

Figure 4: ***Graph Blow Up*** $G^{(s)}$.

1.3.3 Graph Blow Up

We now show a graph that is $\epsilon - far$ from being $\triangle$-free, yet any canonical triangle-free tester must use $q > \left(\frac{c^*}{\epsilon}\right)^{c^* \log \frac{c^*}{\epsilon}}$ queries in order to find a triangle with high probability for some chosen ϵ.

Let $G^{(s)}$ be a blown up version of the initial graph G shown above:

- Each vertex in G is blown up to be an independent set of size s in $G^{(s)}$.
- Each edge in G is a complete bipartite graph in $G^{(s)}$.
- We get that for any triangle in G we get s^3 triangles in $G^{(s)}$, and there are no new $\triangle$'s in $G^{(s)}$.

The parameters of $G^{(s)}$:

- number of nodes: $m \cdot s$
- number of edges: $m|x| \cdot s^2$

- number of $\triangle$: $m|x| \cdot s^3$

Claim 9 *The number of edges that need to be removed from $G^{(s)}$ to make it $\triangle$-free is at least the number of edge-disjoint $\triangle$'s $\geq m|x| \cdot s^2$.*

Proof By removing a single edge from each $\triangle$ we can remove s overlapping $\triangle$'s. The number of $\triangle$'s is $m|x|s^3$, so we are left with $m|x|s^2$ non-overlapping $\triangle$'s. ■

Claim 10 *Given ϵ, there exists a graph $G^{(s)}$ such that for any canonical tester T, Pr[T sees any triangle] $\ll 1$ unless the # of queries in $T > \left(\frac{c^*}{\epsilon}\right)^{c^* \log \frac{c^*}{\epsilon}}$ for some constant c^*.*

Proof Given ϵ, pick m to be the largest integer satisfying $\epsilon \leq \frac{1}{e^{10\sqrt{\log m}}}$. This m satisfies $m \geq \left(\frac{c}{\epsilon}\right)^{c \log \frac{c}{\epsilon}}$ for some c. Pick $s = \frac{n}{6m} \approx n \cdot \left(\frac{\epsilon}{c'}\right)^{c \log \frac{c'}{\epsilon}}$, so

$$\#edges \approx distance(absolute) \approx m|X| \cdot s^2 \approx \frac{m \cdot m}{e^{10\sqrt{\log m}}} \cdot \frac{n^2}{(6m)^2} = \epsilon n^2$$

$$\#triangles \approx \left(\frac{\epsilon}{c''}\right)^{c'' \log \frac{c''}{\epsilon}}$$

Finally, if we take a sample of size q:

$$E\left[\text{Number of } \triangle\text{'s in sample }\right] < \frac{\binom{q}{3}\left(\frac{\epsilon}{c''}\right)^{c'' \log \frac{c''}{\epsilon}} \cdot n^3}{\binom{n}{3}} \ll 1$$

unless $q > \left(\frac{c^*}{\epsilon}\right)^{c^* \log \frac{c^*}{\epsilon}}$. By Markov's inequality, Pr[see any triangle] $\ll 1$. ■

A 1-sided error sampling algorithm must see a triangle to fail, and by the last claim and the Goldreich-Trevisan theorem we get that any tester must use at least $(\frac{c}{\epsilon})^{c\log(c/\epsilon)}$ queries to see a triangle in $G^{(s)}$ with high probability.