0368.416701 Sublinear Time Algorithms	November 16, 2009
Lecture 5	
Lecturer: Ronitt Rubinfeld	*Scribe: Hadas Birin& Michal Rosen*

1 Lecture Outline

So far, we covered algorithms on sparse graphs, where the bound on the degree d assisted us in achieving sub linear time. The input to these algorithms were graphs represented by adjacency lists.
Today we will explore property testing in dense graphs:

- Testing Bipartiteness
- A canonical tester

2 Property Testing of dense Graphs

Definition 2.1 (Adjacency Matrix Model) *Given a graph $G = (V, E)$, an algorithm in the Adjacency Matrix Model receives G as input in the form of a matrix A such that*

$$A_{ij} = \begin{cases} 1 & (i,j) \in E \\ 0 & otherwise \end{cases} .$$

For a given (i, j), querying A_{ij} is one time step.
We often refer to an entry in the adjacency matrix as an edge slot.

Definition 2.2 (Graph Property) *A graph property P is a property of graphs that depends only on the abstract structure, not on graph representations such as particular labellings or drawings of the graph. Given a property P and a domain D, let $\mathcal{P} = \{G \in D \mid G \text{ has property } P\}$.*

Definition 2.3 (ϵ-far from $\mathcal{P}$) *Given a property P and a graph G,*
let $G' \in \mathcal{P}$ be a graph with the minimal number of changes to edge slots in G's adjacency matrix.
G is ϵ-far from $\mathcal{P}$ if the number of change slots between G and G' is at least ϵn^2.

Definition 2.4 (Property Tester) *A property tester $\mathcal{T}$ for the property P is defined by*

if $x \in \mathcal{P}$, then with high probability $\mathcal{A}(x) = pass$
if x is ϵ-far from $\mathcal{P}$, then with high probability $\mathcal{A}(x) = fail$

3 Testing Bipartiteness

Definition 3.1 (Bipartite Graph) *Given a graph $G = (V, E)$, G is **bipartite** if there exists a partition of V into (V_1, V_2) such that $\forall (u, v) \in E$, $u \in V_i$ and $v \in V_j$ for $i \neq j$.*

Definition 3.2 (Violating Edge) *Given a graph $G = (V, E)$, a partition (V_1, V_2) of V and an edge $(u, v) \in E$, we say that (u, v) **violates** (V_1, V_2) if $u, v \in V_1$ or $u, v \in V2$.*
So G is bipartite if and only if there exists a partition of V with no violating edges.

Remark The property of bipartiteness is anti-monotone, i.e. to make a non-bipartite graph g be bipartite, we must remove edges. This leads to an equivalent definition for being ϵ-far from bipartiteness.

Definition 3.3 (ϵ-far from bipartite) *For two graphs G, G', $dist(G, G')$ is the fraction of locations in A that are different (i.e. all i, j such that $A_{ij}^{G} \neq A_{ij}^{G'}$).*
G is ϵ-far from property P if for all G' that have property P, $dist(G, G') > \epsilon$.

Remark

1. For sparse graphs with less than ϵn^2 edges, the definition above is not interesting as we can remove all edges to make the graph bipartite, and therefore a tester can always output pass.
2. For sparse graphs, the sample complexity for testing bipartiteness is known to have a lower bound of $\Omega(\sqrt{n})$.
3. The best lower bound known in this model, the adjacency matrix model, is $\widetilde{\Omega}(\frac{1}{\epsilon^{1.5}})$, due to: A. Bogdanov and L. Trevisan. Lower bounds for testing bipartiteness in dense graphs.
4. With methods similar to the ones we'll use today, we can test for 3-coloring in constant time!

We will now make a first attempt at testing for bipartiteness.
How about sampling $m = \theta(\frac{1}{\epsilon} \log \frac{1}{\delta})$ edges?
Assume G is ϵ-far from being bipartite, therefore it has $\geq \epsilon n^2$ violating edges. Therefore,

$$\Pr_{e \in_R E}[e \text{ is not a violating edge}] < 1 - \epsilon$$

Then for all m samples:

$$\Pr[\text{We didn't hit a violating edge in all } m \text{ samples}] < (1-\epsilon)^m$$

Therefore,

$$\Pr[\text{Hitting a violating edge in at least one of } m \text{ samples}] \geq 1-(1-\epsilon)^m = 1-(1-\epsilon)^{\frac{1}{\epsilon} ln \frac{1}{\delta}} \approx 1-e^{-ln \frac{1}{\delta}} = 1-\delta$$

The problem is that an edge is violating **with respect to a given partition**. In order to reject graphs that are ϵ-far from bipartite, we need to test whether for every partition there are at least ϵn^2 violating edges.

Lets try checking all possible partitions.

Algorithm 3.1 *TestBipartite* 0 (G)

1. *Pick $m = \theta(\frac{1}{\epsilon} \log \frac{1}{\delta})$ random edge slots (i, j) and query A_{ij}.*
2. *For every partition (V_1, V_2):*
   (a) *$violating_{(V_1,V_2)}$ = the number of violating edges in the sample with respect to (V_1, V_2)*
   (b) *If $violating_{(V_1,V_2)} = 0$ output PASS*
3. *Output FAIL*

Figure 1: TestBipartite 0.

Claim 3.4 *TestBipartite* 0 *is a tester for bipartiteness.*

Proof If G is bipartite, then there exists a partition of V into (V_1, V_2) with no violating edges. When $TestBipartite\ 0$ iterates all possible partitions, it will also check partition (V_1, V_2) and output PASS. Assume G is ϵ-far from bipartite.
For any partition of V into (V_1, V_2) there are at least ϵn^2 violating edges. The algorithm samples m independent samples and so for every partition (V_1, V_2) we have that

$Pr[violating_{V_1,V_2} > 0] \geq 1 - (1-\epsilon)^m \geq 1 - (1-\epsilon)^{\frac{1}{\epsilon} ln \frac{1}{\delta}} \approx 1 - e^{-ln \frac{1}{\delta}} = 1 - \delta$

With union bound on all partitions we get

$Pr[\ \forall V_1, V_2\ violating_{V_1,V_2} > 0] \geq (1-\delta)^{2^n}$

So if we take $\delta < \frac{1}{2^n}$ we get that $TestBipartite\ 0$ is indeed a tester for bipartiteness. ■

Observation 3.5 *By the proof above, the running time of* $TestBipartite\ 0$ *is* $\Omega(\frac{1}{\epsilon} \log \frac{1}{\frac{1}{2^n}}) = \Omega(\frac{n}{\epsilon})$*, i.e. sub-linear in the size of the input (which is* $O(n^2)$*).*

Algorithm 3.2 $TestBipartite\ 1$

1. (a) *Choose* U *nodes uniformly, s.t* $|U| = \Theta\left(\frac{1}{\epsilon} \log \frac{1}{\epsilon}\right)$
   (b) *Choose* U' *nodes uniformly, s.t* $|U'| = \Theta\left(\frac{1}{\epsilon^2} \log \frac{1}{\epsilon}\right)$. *Think of* U' *as a group of pairs:* $U' = P = \{(v_1, u_1), (v_2, u_2) \ldots\}$
2. $\forall (U_1, U_2)$ *partition of* U*:*
   (a) *Check* $\forall (u_i, v_i) \in P$*:* $X_i = DoesViolatePartition(U_1, U_2, u_i, v_i)$.
   (b) *If* $\frac{1}{|P|} \sum_{i=1}^{|P|} X_i \leq \frac{3}{4}\epsilon$ *ACCEPT and halt.*
   Else continue.
3. *FAIL*

Figure 2: TestBipartite 1

$DoesViolatePartition$ checks if the pair (u_i, v_i) violates the partition $(V_a^{U_1,U_2}, V_b^{U_1,U_2})$, which we induce from (U_1, U_2) to the rest of the graph as follows:

Algorithm 3.3 $InduceToRestOfGraph(U_1, U_2, V, E)$

1. $\forall u \in U_1$*: put* $u \in V_a^{U_1,U_2}$.
2. $\forall u \in U_2$*: put* $u \in V_b^{U_1,U_2}$.
3. $\forall u \in V \setminus (U_1 \cup U_2)$*:*
   (a) *If* u *has a neighbor in* U_1*: put* $u \in V_b^{U_1,U_2}$.
   (b) *else: put* $u \in V_a^{U_1,U_2}$.

Figure 3: InduceToRestOfGraph.

Note: We don't need to run $InduceToRestOfGraph$ in advance. We will run it only for the vertices in P to figure out in which part of the partition $\{V_a^{U_1,U_2}, V_b^{U_1,U_2}\}$ they fall. For each vertex v it will take $O\left(\frac{1}{\epsilon} \log \frac{1}{\epsilon}\right)$ (to check if $v \in U_1$, $v \in U_2$, or checking neighbors).

Theorem 3.6 $TestBipartite\ 1$ *is a property tester for bipartiteness. More precisely,*

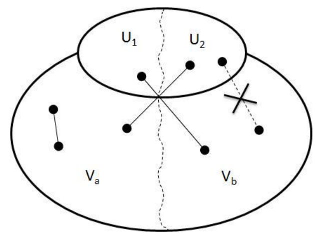

Figure 4: InduceToRestOfGraph

Algorithm 3.4 $DoesViolatePartition(U_1, U_2, u_i, v_i)$

1. *Find* $x \in \{a, b\}$ *for* u_i, *where* $u_i \in V_x^{U_1,U_2}$ *according to* $InduceToRestOfGraph$.
2. *Find* $y \in \{a, b\}$ *for* v_i, *where* $v_i \in V_y^{U_1,U_2}$ *according to* $InduceToRestOfGraph$.
3. *If* $x = y$ *return* 1.
4. *Else return 0.*

Figure 5: DoesViolatePartition.

(1) if G *is bipartite,* $TestBipartite$ 1 *PASSES with probability* $\geq \frac{3}{4}$.

(2) if G *is* $\epsilon - far$ *from bipartiteness,* $Pr[\ TestBipartite\ 1$ *outputs FAIL* $] \geq \frac{7}{8}$.

Proof of Theorem 1(1):
Assume G is bipartite.
Therefore, there exists a partition (Y_1, Y_2) of V with no violating edges.
For a sample U, let $U_1 = U \cap Y_1$ and $U_2 = U \cap Y_2$.
From $InduceToRestOfGraph$ we get $(V_a^{U_1,U_2}, V_b^{U_1,U_2})$.
How close is $(V_a^{U_1,U_2}, V_b^{U_1,U_2})$ to (Y_1, Y_2)?

Observation 3.7 *If* (Y_1, Y_2) *is a bipartition, no vertex* v *has a neighbor in both* U_1 *and* U_2 *(because* $U_1 \subseteq Y_1$ *and* $U_2 \subseteq Y_2$*). Therefore, a difference between* $(V_a^{U_1,U_2}, V_b^{U_1,U_2})$ *and* (Y_1, Y_2) *(if exists) is due to nodes that don't have neighbors in* U.

We have two kind of vertices:

1. v with small degree ($d(v) < \frac{\epsilon}{4}n$).
2. v with high degree ($d(v) \geq \frac{\epsilon}{4}n$).

Definition 3.1 *Let* $HighDeg$ *be the event where at most* $\frac{\epsilon}{4}n$ *"high degree" nodes in* V *don't have neighbors in* U.

Lemma 3.8 $Pr_U[\neg HighDeg] \le \frac{1}{8}$ *where* $|U| \ge \frac{4}{\epsilon}\log\frac{32}{\epsilon}$

Proof $\forall v \in V$ define $\sigma_v = \begin{cases} 1 & \text{if } v \text{ is a "high degree" node and v has no neighbors in } U \\ 0 & \text{otherwise} \end{cases}$.

$$E[\sigma_v] = Pr[\sigma_v = 1] \le (1 - \frac{\frac{\epsilon}{4}n}{n})^{|U|}$$

Since the number of 1s in v's row $\ge \frac{\epsilon}{4}n$, and n is the number of entries in v's row. Therefore,

$$E[\sigma_v] \le (1 - \frac{\frac{\epsilon}{4}n}{n})^{|U|} \le (1 - \frac{\epsilon}{4})^{\frac{4}{\epsilon}\log\frac{32}{\epsilon}} \le \frac{1}{e}^{\log\frac{32}{\epsilon}} = \frac{\epsilon}{32}$$

Thus, by Markov's inequality:

$$\Pr[\sum_{v \in V} \sigma_v \ge 8 \cdot \frac{\epsilon n}{32}] = \frac{\epsilon n}{4} \le \frac{1}{8}$$

■

How many violating edges are in $(V_a^{U_1,U_2}, V_b^{U_1,U_2})$?
Let N be the number of violating edges in $(V_a^{U_1,U_2}, V_b^{U_1,U_2})$ under the assumption of $HighDeg$. Then:

$$N \le \underbrace{0}_{\#\text{violating edges in } (Y_1,Y_2)} + \underbrace{\frac{\epsilon}{4}n}_{\text{Bound on the degree of small degree nodes}} \cdot \underbrace{n}_{\text{Bound on \# small degree nodes}} + \underbrace{n}_{\text{Bound on degree of high degree nodes}} \cdot \underbrace{\frac{\epsilon}{4}n}_{\text{Bound on \#high degree nodes}} \le \frac{\epsilon}{2}n^2$$

Corollary 3.9 $N \le \frac{\epsilon}{2}n^2$ *with probability* $\frac{7}{8}$.

Assuming $N \le \frac{\epsilon}{2}n^2$ we get that

$$\forall (u_i, v_i) \in P:$$
$$\Pr[(u_i, v_i) \text{ violates } (V_a^{U_1,U_2}, V_b^{U_1,U_2})] \le \frac{\epsilon}{2}$$
$$\Longrightarrow E_{(u_i,v_i)\in P}[\mathbf{1}_{(u_i,v_i) \text{ violates } (V_a^{U_1,U_2}, V_b^{U_1,U_2})}] \le \frac{\epsilon}{2}$$

Therefore we expect the fraction of violating pairs in P to be $\le \frac{\epsilon}{2}$.

Claim 3.10 *$Pr[$Fraction of violating edges in the sample $\geq \frac{3}{4}\epsilon \mid HighDeg\ ] < \frac{1}{8}$*

Proof Sample $|P| = c \cdot \frac{1}{\epsilon^2} \log \frac{1}{\epsilon}$ for some $c > 1$.
Let $X_1, X_2, \ldots X_{|P|}$ be i.i.d s.t $X_i = \begin{cases} 1 & \text{if } e_i \text{ violates } (V_a^{U_1,U_2}, V_b^{U_1,U_2}) \\ 0 & \text{otherwise} \end{cases}$.
$E[X_i | HighDeg] \leq \frac{\epsilon}{2}$.
By Chernoff:

$$\Pr[\frac{1}{|P|}\sum_{i=1}^{|P|} X_i \geq \frac{3}{4}\epsilon] \geq$$
$$Pr[\frac{1}{|P|}\sum_{i=1}^{|P|} X_i \geq (1+\frac{1}{2})E[X_i|HighDeg]\ ] \leq$$
$$e^{-(\frac{1}{2})^2 \frac{\epsilon}{2} \frac{|P|}{3}} = e^{-\frac{\epsilon}{24} c \cdot \frac{1}{\epsilon^2} \log \frac{1}{\epsilon}} = \epsilon^{\frac{c}{24\epsilon}} \quad \underbrace{< \frac{1}{8}}_{\text{choose } c \text{ s.t}}$$

■

Lemma 3.11 $\Pr[$ *TestBipartite* 1 *outputs FAIL on a bipartite graph*$] \leq \frac{1}{4}$

Proof

$$\Pr[\ TestBipartite\ 1 \text{ outputs } FAIL \text{ on a bipartite graph}]$$
$$\leq \Pr[\text{Fraction of violating edges in the sample } \geq \frac{3}{4}\epsilon | HighDeg] \cdot \Pr[HighDeg]$$
$$+ Pr[\text{Fraction of violating edges in the sample } \geq \frac{3}{4}\epsilon | \neg HighDeg] \cdot \Pr[\neg HighDeg]$$
$$\leq \frac{1}{8} \cdot 1 + 1 \cdot \frac{1}{8} = \frac{1}{4}.$$

■

That proves the first item in 3.6.
■

Proof of Theorem 1(2):
Suppose G is ϵ-far from bipartite.
Therefore, all partitions (Y_1, Y_2) have $\geq \epsilon n^2$ violating edges.
In particular, $\forall (U_1, U_2)$ partition of U, $(V_a^{U_1,U_2}, V_b^{U_1,U_2})$ has $\geq \epsilon n^2$ violating edges.

$$\Pr_{(u_i,v_i)\in P}[(u_i, v_i) \text{ violates } (V_a^{U_1,U_2}, V_b^{U_1,U_2})] \geq \frac{\epsilon n^2}{n^2} = \epsilon$$
$$\Longrightarrow E_{(u_i,v_i)\in P}[\mathbf{1}_{(u_i,v_i) \text{ violates } (V_a^{U_1,U_2}, V_b^{U_1,U_2})}] \geq \epsilon$$

■

Proof of Theorem 3.6(2):
Suppose G is ϵ-far from bipartite.
Therefore, all partitions (Y_1, Y_2) have $\geq \epsilon n^2$ violating edges.
In particular, $\forall (U_1, U_2)$ partition of U, $(V_a^{U_1,U_2}, V_b^{U_1,U_2})$ has $\geq \epsilon n^2$ violating edges.

$$\Pr_{(u_i,v_i)\in P}[(u_i, v_i) \text{ violates } (V_a^{U_1,U_2}, V_b^{U_1,U_2})] \geq \frac{\epsilon n^2}{n^2} = \epsilon$$
$$\implies E_{(u_i,v_i)\in P}[\mathbf{1}_{(u_i,v_i) \text{ violates } (V_a^{U_1,U_2}, V_b^{U_1,U_2})}] \geq \epsilon$$

Let $X_1, X_2, \ldots X_{|P|}$ be i.i.d s.t $X_i = \begin{cases} 1 & \text{if } e_i \text{ violates } (V_a^{U_1,U_2}, V_b^{U_1,U_2}) \\ 0 & \text{otherwise} \end{cases}$.

By Chernoff:

$$\begin{aligned}
&\Pr[\frac{1}{|P|}\sum_{i=1}^{|P|} X_i \leq \frac{3}{4}\epsilon] \\
&= Pr[\frac{1}{|P|}\sum_{i=1}^{|P|} X_i \leq (1-\frac{1}{4})\epsilon] \\
&< e^{-(\frac{1}{4})^2\frac{|P|}{2}\epsilon} = e^{-\frac{1}{32}\frac{1}{\epsilon^2}\log\frac{1}{\epsilon}\epsilon c} \\
&= \epsilon^{\frac{c}{32\epsilon}}
\end{aligned}$$

Therefore, by union bound:

$$\begin{aligned}
&\Pr[\text{Algorithm outputs PASS}] \\
&= \Pr[\text{There is a partition with fraction of violating pairs }] \leq \frac{3}{4}\epsilon] \\
&< 2^{|P|}\epsilon^{\frac{c}{32\epsilon}} \underbrace{=}_{\text{for some } d>0} 2^{d\frac{1}{\epsilon^2}\log\frac{1}{\epsilon}}\epsilon^{\frac{c}{32\epsilon}} \\
&= \epsilon^{-d\frac{1}{\epsilon^2}+\frac{c}{32\epsilon}} = \epsilon^{\frac{c\cdot d-32}{32\epsilon}} \underbrace{< \frac{1}{8}}_{\text{Choose } c,d \text{ s.t}}
\end{aligned}$$

■

4 A Canonical Tester

Theorem 4.1 *Let P be any graph property in the adjacency matrix model.*
Suppose T is a tester for P with query complexity $q(n, \epsilon)$.
Then, there is a tester T' with query complexity of $O(q^2)$ in the following form:

1. *Select $2q(n, \epsilon)$ nodes randomly.*
2. *Query all pairs in the sampling.*
3. *Make a decision.*

Moreover, if T has one-sided error, so does T'.