Lecture 10

Lecturer: Sofya Raskhodnikova *Scribe(s):*

1 Introduction

Today we will cover a sublinear time algorithm for testing whether a dense graph is bipartite or ϵ-far from it. We will present a tester by Goldreich, Goldwasser and Ron [GGR98] that runs in $\tilde{O}(\frac{1}{\epsilon^4})$ time[1]. (The best tester for bipartiteness in [GGR98] runs in time $\tilde{O}(\frac{1}{\epsilon^3})$.) Alon and Krivelevich [AK02] improved the analysis of the tester we will consider to $\tilde{O}(\frac{1}{\epsilon^2})$. The tester is nonadaptive, and Bogdanov and Trevisan [BT04] showed that $\Omega(\frac{1}{\epsilon^2})$ queries are required for a nonadaptive tester of bipartiteness. They also proved a lower bound $\tilde{\Omega}(\frac{1}{\epsilon^{1.5}})$ for general (adaptive) testers for this problem.

1.1 Model

We represent a dense graph using an adjacency matrix. In one query, we can read one entry of the adjacency matrix or, in other words, find out if a pair of nodes is an edge in the graph. The distance between two graph G_1, G_2 on n vertices (each represented by an adjacency matrix) is

$$dist(G_1, G_2) = \frac{\text{the number of entries in the adjacency matrix on which } G_1 \text{ and } G_2 \text{ differ}}{n^2}.$$

If the graphs are undirected, then every edge corresponds to two entries in the adjacency matrix, and the expression above is equal to

$$dist(G_1, G_2) = \frac{\text{the number of edges present in exactly one of } G_1 \text{ and } G_2}{n^2/2}.$$

Definition 1. *A pair of sets (V_1, V_2) is a partition of the set V if V_1 and V_2 are disjoint subsets of V and $V_1 \cup V_2 = V$. An undirected graph $G = (V, E)$ is* bipartite *if there exists a partition (V_1, V_2) of V such that for every edges in G, one endpoint is in V_1 and the other is in V_2.*

Definition 2. *An edge $\{u, v\}$ of vertices is violating with respect to partition (V_1, V_2) if either $u, v \in V_1$ or $u, v \in V_2$.*

An example with $V = \{u_1, u_2, u_3\}$ is shown in Figure 2. The partition is $V_1 = \{u_1, u_3\}$ and $V_2 = \{u_2\}$. The edge $\{u_1, u_3\}$ is violating w.r.t. that partition.

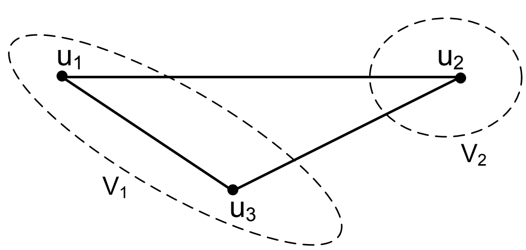

Figure 1: An example of a violating edge.

[1] $\tilde{O}$ is a variant of big-O notation that ignores logarithmic factors. Specifically, $\tilde{O}(f(n))$ is a shorthand for "there exists $k \in \mathbb{N}$ such that $O(f(n) \log^k(f(n)))$".

Observation 3. *If a graph $G = (V, E)$ on n nodes is ϵ-far from bipartite then for every partition (V_1, V_2) of V, there exist at least $\epsilon n^2/2$ violating edges w.r.t. (V_1, V_2).*

For example, consider a clique (a graph where every two vertices are connected by an edge) on n vertices. The largest number of edges a bipartite graph on n vertices can have is $n^2/4$. (This is true for a complete bipartite graph with $n/2$ vertices in each part.) A clique is $\frac{\binom{n}{2}-n^2/4}{n^2/2} \approx \frac{1}{2}$-far from bipartite.

2 First Attempt

Consider an algorithm of the following form: we will sample m pairs of nodes independently and uniformly at random and reject iff we rule out all possible partitions of V. How large does m have to be for this to be a valid tester?

In a graph $G = (V, E)$ on n nodes that is ϵ-far from bipartite, there are 2^n possible partitions of V. What is the probability that none of the m sampled pairs of nodes are violating edges? Observe that each edge corresponds to 2 ordered pairs of nodes. Since G is ϵ-far from bipartite, by Observation 3, for any fixed partition (V_1, V_2) of V,

$$\Pr[\text{a random pair } (u, v) \text{ is not violating}] \leq 1 - \frac{\epsilon n^2}{n^2} \leq 1 - \epsilon.$$

For a fixed partition (V_1, V_2), let $BAD(V_1)$ be the event that none of the m selected random pairs are violating. Let BAD be the event that there exists a partition with respect to which none of the selected pairs are violating. Then

$$\begin{aligned}
\Pr[BAD(V_1)] &\leq (1-\epsilon)^m \\
&\leq [(1-\epsilon)^{1/\epsilon}]^{\epsilon m} \leq e^{-\epsilon m} \\
&\leq \frac{1}{3} \cdot 2^{-n} \text{ (provided that } m \geq \tfrac{n \ln 2 + \ln 3}{\epsilon}).
\end{aligned}$$

By a union bound,

$$\Pr[BAD] \leq \sum_{V_1 \subseteq V} \Pr[BAD(V_1)] \leq 2^n \cdot \frac{1}{3} \cdot 2^{-n} = \frac{1}{3}.$$

Thus, an algorithm that queries $\frac{n \ln 2 + \ln 3}{\epsilon} = O(\frac{n}{\epsilon})$ independent and uniformly random pairs of nodes is an ϵ-tester for bipartiteness.

3 The $\tilde{O}(\frac{1}{\epsilon^4})$ algorithm [GGR98]

Algorithm 1: Bipartiteness Tester from [GGR98]

Input: $\epsilon \in (0, 1), n \in \mathbb{N}$ and query access to adjacency matrix of $G = (V, E)$.
Pick uniformly at random a set S of nodes where $|S| = \Theta(\frac{1}{\epsilon^2} \log \frac{1}{\epsilon})$;
Query all induced pairs of nodes (i, j) where $i, j \in S$. Let the subgraph be G' ;
if G' is bipartite **then accept**; otherwise, **reject**.

Note that we can check whether G' is bipartite by using Breadth-First Search (BFS). Therefore, both time and query complexity are $O(\binom{S}{2}) = O(\frac{\log^2 \frac{1}{\epsilon}}{\epsilon^4}) = \tilde{O}(\frac{1}{\epsilon^4})$. Now we analyze correctness.

3.1 Correctness Analysis

First observe that every subgraph of a bipartite graph is also bipartite. Thus, all bipartite graphs will be always accepted.

For the rest of the analysis assume that the input graph G on n nodes is ϵ-far from bipartite. The main idea behind the analysis is to break the samples S into two sets:

1. the learning set L of size $|L| = \Theta(\frac{1}{\epsilon}\log\frac{1}{\epsilon})$ and
2. the testing set T of size $|T| = \Theta(\frac{1}{\epsilon^2}\log\frac{1}{\epsilon})$.

The learning set L will induce a partition of V and then we can use the testing set T to check for violating pairs w.r.t. the partition.

Definition 4. *A node v of G is* influential *if its degree $deg(v) \geq \frac{\epsilon n}{8}$.*

Since $|V| = n$, there are at most n non-influential nodes. Each non-influential node has $< \frac{\epsilon n}{8}$ edges incident to it (because its degree $< \frac{\epsilon n}{8}$). Since there are ϵn^2 violating edges, there are at least $\frac{3\epsilon n^2}{8}$ violating edges between influential nodes.

Definition 5. *A node $v \in V$ is* covered *by $L \subset V$, if v has a neighbor in L. Let C be the set of all $v \in V$ covered by L.*

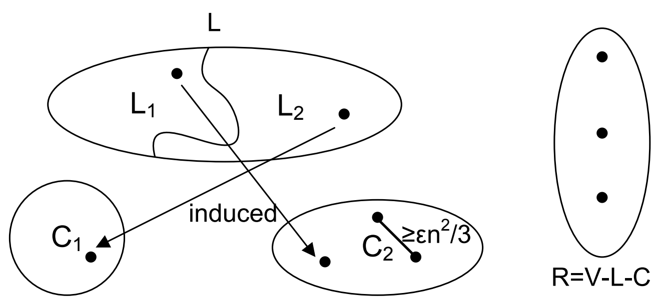

Figure 2: A partition of the learning set L into L_1 and L_2. C_1 is the set covered by L_2 and C_2 is the set covered by L_1. Let $C = C_1 \cup C_2$, then there are at least $\frac{\epsilon n^2}{8}$ violating edges between vertices in C. Let $R = V - L - C$ be the set containing all remaining vertices.

Claim 6. $\Pr[$*more than* $\frac{\epsilon n}{8}$ *influential nodes are not covered by* $L] \leq \frac{1}{6}$.

Proof. For each influential node $v \in V$, define an indicator random variable

$$X_v = \begin{cases} 1 & \text{if } v \text{ is not covered by } L; \\ 0 & \text{otherwise.} \end{cases}$$

Then, let $X = \sum X_v$ and therefore, we want to bound $\Pr[X \geq \frac{\epsilon n}{8}]$.

$\Pr[X_v = 1] \leq (1 - \frac{\epsilon}{8})^{|L|}$ where $|L|$ is the size of the set L. This bound is because an influential node has degree at least $\frac{\epsilon n}{8}$ and therefore each node of L is chosen not from those adjacent (at least) $\frac{\epsilon n}{8}$ nodes of v. Specifically,

$$\begin{aligned} \Pr[X_v = 1] &\leq (1 - \frac{\epsilon}{8})^{|L|} \\ &\leq e^{\frac{-\epsilon|L|}{8}} \\ &\leq \frac{\epsilon}{48}. \end{aligned}$$

Hence, $E(X) = \sum \Pr[X_v = 1] \leq \frac{\epsilon n}{48}$. By Markov's inequality, $\Pr[X \geq \frac{\epsilon n}{8}] \leq \frac{E(X)}{\epsilon n/8} \leq \frac{1}{6}$. □

Let BAD_1 be the event that more than $\frac{\epsilon n}{8}$ influential nodes are not covered by L.

Claim 7. *If BAD_1 does not occur, then every partition of L induces at least $\frac{\epsilon n^2}{8}$ violating edges between vertices in C.*

Proof. By Observation 3, w.r.t. every partition, there are at least $\epsilon n^2/2$ violating edges.

Violating edges incident to	Number of vertices	Degree	Number of violating edges
Influential nodes in R	$\frac{\epsilon n}{8}$ $(\overline{BAD1})$	n	$\frac{\epsilon n^2}{8}$
Non-influential nodes in R	$\leq n$	$\frac{\epsilon n}{8}$ (non-influential)	$\leq \frac{\epsilon n^2}{8}$
Nodes in L	$\Theta(\frac{\log 1/\epsilon}{\epsilon})$	$\leq n$	$o(n^2)$, i.e., at most $\frac{\epsilon n^2}{8}$

Therefore, there are at least $\epsilon n^2 - \frac{\epsilon n^2}{8} - \frac{\epsilon n^2}{8} - \frac{\epsilon n^2}{8} \geq \frac{\epsilon n^2}{8}$ violating edges between nodes in C. □

Now, fix a partition of L, which implies a partition of C. View samples from T as pairs: $(v_1, v_2), (v_3, v_4), \ldots, (v_{|T|-1}, v_{|T|})$. Then the probability

$$\begin{aligned}\Pr[\text{ no pairs } (v_{2i-1}, v_{2i})\ (i = 1, ..., \tfrac{|T|}{2}) \text{ are violating edges in } C] &\leq (1 - \frac{\epsilon}{8})^{|T|/2} \text{ by Claim 6.}\\ &\leq e^{-\frac{\epsilon|T|}{6}}\\ &\leq \frac{2^{-|L|}}{6}.\end{aligned}$$

Since there are $2^{|L|}$ partitions of L, by a union bound, the probability

$$\Pr[\text{there is a partition of } L \text{ such that no pairs } (v_{2i-1}, v_{2i})\ (i = 1, ..., \tfrac{|T|}{2}) \text{ are violating edges in } C] \leq 2^{|L|} \times \frac{2^{-|L|}}{6} \leq \frac{1}{6}.$$

Let this event be BAD_2.

Now, we complete the correctness analysis, by:

$$\begin{aligned}\Pr[\epsilon\text{-far G is accepted}] &\leq \Pr[BAD_1] + \Pr[\overline{BAD_1}]\Pr[BAD_2|\overline{BAD_1}]\\ &\leq \frac{1}{6} + \frac{1}{6} \text{ by Claims 5 and 6.}\\ &\leq \frac{1}{3}.\end{aligned}$$

References

[AK02] Noga Alon and Michael Krivelevich. Testing k-colorability. *SIAM Journal on Discrete Mathematics*, 15(2):211–227, 2002.

[BT04] Andrej Bogdanov and Luca Trevisan. Lower bounds for testing bipartiteness in dense graphs. In *19th Annual IEEE Conference on Computational Complexity (CCC 2004), 21-24 June 2004, Amherst, MA, USA*, pages 75–81, 2004.

[GGR98] O. Goldreich, S. Goldwasser, and D. Ron. Property testing and its connection to learning and approximation. *Journal of the ACM*, 45(4):653–750, 1998. Preliminary version in 37th FOCS, 1996.