Sublinear Algorithms	January 31, 2012
Lecture 7	
Lecturer: Sofya Raskhodnikova	*Scribe(s): Ye Zhang*

1 Introduction

In today's lecture, we will talk about a sublinear time algorithm to test whether a dense graph is bipartite or ϵ-far from it. The content of this lecture is based on Goldreich, Goldwasser and Ron's paper [GGR98], where they propose a $\tilde{O}(\frac{1}{\epsilon^4})$ algorithm.

The better result can be found in Alon and Krivelevich's paper [AK02]. We can also prove that the lower bound for this problem is $\tilde{O}(\frac{1}{\epsilon^{1.5}})$, which only depends on the parameter ϵ.

1.1 Model

As we consider dense graphs, we choose adjacency matrix to represent it. Recall that in the previous lectures, we defined the distance between two graph G_1, G_2 (which both represented by adjacency matrix):

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

where we assume that G_1 and G_2 have n vertices.

Definition 1. *A graph $G = (V, E)$ is said to be bipartite if there exists $V_1, V_2 \subset V$ such that $V_1 \cup V_2 = V$, $V_1 \cap V_2 = \emptyset$ and for any $(u, v) \in E$, $u \in V_1$ and $v \in V_2$.*

An edge $(u, v) \in E$ is a *violating edge* if either $u, v \in V_1$ or $u, v \in V_2$. An example is shown in Figure 2, where we assume $V = \{u_1, u_2, u_3\}$. The partition is $V_1 = \{u_1, u_3\}$ and $V_2 = \{u_2\}$. Then, (u_1, u_3) is a violating edge.

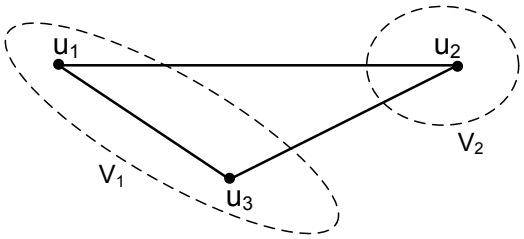

Figure 1: An example of violating edges

Corollary 2. *A graph $G = (V, E)$ (where $n = |V|$) is ϵ-far from bipartite, if for any partition V_1, V_2 (where $V_1 \cup V_2 = V$ and $V_1 \cap V_2 = \emptyset$), there exist at least ϵn^2 violating edges in E w.r.t. V_1 and V_2.*

For example, a clique (that every two vertices are connected by an edge) is not bipartite. Specifically, it is $\frac{\binom{n}{2} - n^2/4}{n^2} \approx \frac{1}{4}$-far from bipartite (i.e., compared with the complete bipartite graph).

2 First Attempt

Given a graph $G = (V, E)$ is ϵ-far from bipartite, where $n = |V|$, we observe that there are 2^n possible partitions. Suppose we sample m pairs of node, then what is the probability that we do not capture a violating edge?

Since G is ϵ-far, by Cor 2, for any fixed partition V_1, V_2, we have Pr[a random pair (u, v) is a non-violating edge] $\leq 1 - \frac{\epsilon n^2}{n^2} \leq 1 - \epsilon$. So for the fixed partition V_1, V_2, the probability $\Pr[X\&V_1]$ that m selections of random pairs are all non-violating is

$$\begin{aligned}\Pr[X\&V_1] &\leq (1-\epsilon)^m \\ &\leq [(1-\epsilon)^{1/\epsilon}]^{\epsilon m} \leq e^{-\epsilon m} \\ &\leq c2^{-n} \text{ (for } m \geq \tfrac{n \ln 2}{\epsilon})\end{aligned}$$

By a union bound, we have the probability that $\Pr[X] \leq \sum_{V_1} \Pr[X\&V_1] \leq 2^n c 2^{-n} \leq c < 1$. So by the Witness Lemma, we can repeat the above tests for $\frac{2}{c}$ time to complete the algorithm, which shown as follows. The running time and query complexity are both $O(\frac{2n}{c\epsilon}) = O(\frac{n}{\epsilon})$.

Algorithm 1: The First Attempt

Input: $G = (V, E)$ where $|V| = n$
while *Repeat the following test for* $\frac{2}{c}$ *times* **do**
  Randomly partition V to two set V_1, V_2 ;
  Uniformly choose $m = \frac{n}{\epsilon} + 1$ pairs of nodes, say (u_1, v_1) to (u_m, v_m) ;
  if (*there exists* i *such that* (u_i, v_i) *is an violating edge*) **then**
    return *reject* ;
  end
end
return *accept* ;

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

Algorithm 2: [GGR98]'s Algorithm

Input: $G = (V, E)$
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**
  return *accept* ;
end
else
  return *reject* ;
end

Note that we can test bipartitness of G' by using Breadth-First Search (BFS). Therefore, the time and query complexities can be both $O(\binom{S}{2}) = O(\frac{\log^2 \frac{1}{\epsilon}}{\epsilon^4}) = \tilde{O}(\frac{1}{\epsilon^4})$. Now we analyze the correctness.

3.1 Correctness Analysis

There are two ideas behind the analysis: (a) If G is bipartite, then its subgraph is also bipartite.
(b) Break the samples into two sets: (1) the learning set L with the size $|L| = \Theta(\frac{1}{\epsilon}\log\frac{1}{\epsilon})$ and (2) the testing set T with the 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 test in the testing set T whether there is a violating edge according to the partition.

Definition 3. *A node v is said to be influential, if its degree $deg(v) \geq \frac{\epsilon n}{4}$ where $|V| = n$.*

As $|V| = n$, there are at most n non-influential nodes, where each of the non-influential node has $< \frac{\epsilon n}{4}$ edges incident on it (because its degree $< \frac{\epsilon n}{4}$). On the other hand, there are ϵn^2 violating edges, and therefore there are at least $\frac{3\epsilon n^2}{4}$ violating edges between influential nodes.

Definition 4. *$v \in V$ is said to be 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.*

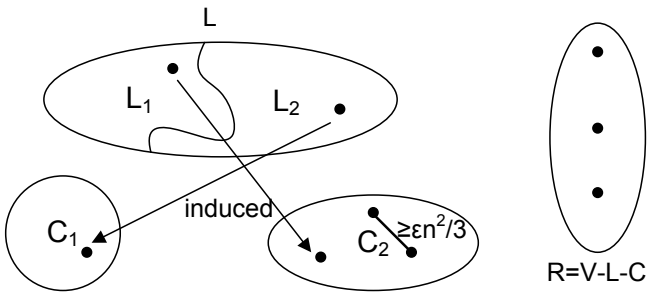

Figure 2: Sample partition. L is partitioned into L_1 and L_2. C_1 is the set covered by L_2 and C_2 is the set covered by L_1. Let $C = C_1 \cup C_2$, then there are at least $\frac{\epsilon n^2}{3}$ violating edges inside C. Denote $R = V - L - C$ be the set containing all remaining vertices.

Claim 5. Pr[*more than $\frac{\epsilon n}{4}$ influential nodes are not covered by L*] $\leq \frac{1}{6}$.

Proof. For any influential node $v \in V$, let's define an indicator $X_v = \begin{cases} 1 & \text{if } v \text{ is not covered by } L \\ 0 & \text{otherwise.} \end{cases}$
Then, let $X = \sum X_v$ and therefore, we want to bound $\Pr[X \geq \frac{\epsilon n}{4}]$.

On the other hand, we see $\Pr[X_v = 1] \leq (1 - \frac{\epsilon}{4})^{|L|}$ where $|L|$ is the set size of L. This bound is because any influential nodes has degree at least $\frac{\epsilon n}{4}$ and therefore each node of L is chosen not from those (at least) adjacent $\frac{\epsilon n}{4}$ nodes of v. Specifically,

$$\begin{aligned} \Pr[X_v = 1] &\leq (1 - \frac{\epsilon}{4})^{|L|} \\ &\leq e^{\frac{-\epsilon|L|}{4}} \\ &\leq \frac{\epsilon}{24}. \end{aligned}$$

Hence, $E(X) = \sum \Pr(X_v = 1) \leq \frac{\epsilon n}{24}$. By Markov inequality, $\Pr[X \geq \frac{\epsilon n}{4}] \leq \frac{E(X)}{\epsilon n/4} \leq \frac{1}{6}$. □

Let's call the event that "more than $\frac{\epsilon n}{4}$ influential nodes are not covered by L" as "BAD1" event.

Claim 6. *Assuming "BAD1" event does not occur, then every partition of L includes $\frac{\epsilon n^2}{3}$ violating edges in C.*

Violating edges incident to	number of vertices	degree	number of violating edges
Influential nodes in R	$\frac{\epsilon n}{4}$ $(\overline{BAD1})$	n	$\frac{\epsilon n^2}{4}$
Non-influential nodes in R	$\leq n$	$\frac{\epsilon n}{4}$ (non-influential)	$\leq \frac{\epsilon n^2}{4}$
Nodes in L	$\Theta(\frac{\log 1/\epsilon}{\epsilon})$	$\leq n$	$O(n^2) = \frac{\epsilon n^2}{6}$

Proof. First, for every partition, there are at least ϵn^2 violating edges. Therefore, there are $\geq (\epsilon n^2 - \frac{\epsilon n^2}{4} - \frac{\epsilon n^2}{4} - \frac{\epsilon n^2}{6}) \geq \frac{\epsilon n^2}{3}$ violating edges 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

$$\begin{aligned}\Pr[\text{ no pairs } (v_{2i-1}, v_{2i}) \ (i = 1, ..., \tfrac{|T|}{2}) \text{ are violating edges in } C] &\leq (1 - \frac{\epsilon}{3})^{|T|/2} \text{as Claim 6.} \\ &\leq e^{-\frac{\epsilon|T|}{6}} \\ &\leq \frac{2^{-|L|}}{6}.\end{aligned}$$

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

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

Let this event as "BAD2".

Now, we complete the correctness analysis, by:

$$\begin{aligned}\Pr[\epsilon\text{-far G is accepted}] &\leq \Pr[BAD1] + \Pr[\overline{BAD1}]\Pr[BAD2|\overline{BAD1}] \\ &\leq \frac{1}{6} + \frac{1}{6} \text{ from Claim 5 \& Claim 6.} \\ &\leq \frac{1}{3}.\end{aligned}$$

References

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

[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.