6.5240 Sublinear Time Algorithms October 5, 2022

Lecture 8

Lecturer: Shyan Akmal *Scribe: Ashley Ke*

In this lecture, we study two applications of property testing for dense graphs. We model edges of graph G with an adjacency matrix A, where

$$A_{uv} = \begin{cases} 1 & \text{if } (u,v) \text{ edge} \\ 0 & \text{otherwise.} \end{cases}$$

First, we prove a sublinear time algorithm for bipartite testing, and second, we begin to discuss a sublinear time algorithm for triangle-free testing.

1 Bipartite Testing

1.1 Setup

In this problem, we receive input graph G and adjacency matrix A. The goal is to find a sublinear time bipartite tester that will:

1. *Accept* if G is bipartite
2. *Reject* if G is ϵ-far from bipartite with probability $\geq 2/3$.

Recall that a graph G is ϵ-far from property P if $> \epsilon n^2$ entries in A must be changed to turn G into a member of P.

1.2 Algorithm

The following algorithm satisfies the desired properties of a bipartite tester. Recall that the notation $G[S]$ denotes the induced subgraph of G on S.

Algorithm 1 Bipartite Tester

Randomly sample set S of $\Theta(\frac{1}{\epsilon^2} \cdot \log(\frac{1}{\epsilon}))$ nodes.
Accept if $G[S]$ is bipartite, reject otherwise.

First, we show correctness.

Note that for any bipartite graph $G = V_1 \sqcup V_2$, any induced subgraph $G[S]$ is also bipartite because we can split S into $S \cap V_1, S \cap V_2$. Thus the algorithm always accepts.

Now, the goal is to show correctness for the ϵ-far case, which we state in the following theorem.

Theorem 1 *If G is ϵ-far from bipartite, the bipartite tester algorithm rejects with probability $\geq 2/3$.*

Before proving Theorem 1, we first discuss a failing proof idea that motivates the correct proof.

Definition 2 *We say edge $e = (u,v)$* violates *partition $V = V_1 \sqcup V_2$ if $u, v \in V_1$ or $u, v \in V_2$.*

An initial idea is to use the most basic bound on $\Pr[G[S] \text{ has no violating edge } e \text{ for } V_1 \sqcup V_2]$. Since G is ϵ-far from bipartite, there are $> \epsilon n^2$ violating edges for $V_1 \sqcup V_2$, so an arbitrary edge violates $V_1 \sqcup V_2$ with probability $\geq \epsilon$. The induced subgraph $G[S]$ has at least $|S|/2$ independent edges, thus

$$\Pr[G[S] \text{ has no violating edge } e \text{ for } V_1 \sqcup V_2] \leq (1-\epsilon)^{|S|/2}.$$

This idea fails because there are 2^n partitions $V_1 \sqcup V_2$, so union bounding gives

$$\Pr[G[S] \text{ has violating edges } e \text{ for all } V_1 \sqcup V_2] \geq 1 - 2^n(1-\epsilon)^{|S|/2}.$$

It turns out this bound is not good enough and we would need $|S| \geq \Omega(\frac{n}{\epsilon})$ for this to be sufficient. The idea for our new proof will be to bypass this 2^n factor by focusing on a smaller set of representative partitions.

Proof of Theorem 1
Goal: Show that if G is ϵ-far from bipartite, $G[S]$ is not bipartite with probability $\geq 2/3$.

We first identity the central subset of representative partitions that will bypass the 2^n factor. These will be the partitions of a set C, where C consists of the first $\Theta\left(\frac{1}{\epsilon} \cdot \log \frac{1}{\epsilon}\right)$ vertices of S. More specifically,

$$C \subseteq S,\ |C| = \Theta\left(\frac{1}{\epsilon} \cdot \log \frac{1}{\epsilon}\right).$$

Note that there are now $2^{|C|}$ partitions instead of 2^n. Looking ahead, it will be important that we chose S to have size $|S| = \Theta(\frac{1}{\epsilon}|C|)$.

Definition 3 *Edge $e = (u, v)$* ruins *partition $C = C_1 \sqcup C_2$ if $u, v \in N(C_1)$ or $u, v \in N(C_2)$.*

The key observation is the following.

Observation 4 *If edge $e \in G[S]$ ruins $C_1 \sqcup C_2$, every partition $S = S_1 \sqcup S_2$ with $C_1 \subseteq S_1, C_2 \subseteq S_2$ will have a violating edge in $G[S]$.*

Proof Without loss of generality, suppose $u, v \in N(C_1)$. Then u neighbors $u' \in C_1$ and v neighbors $v' \in C_1$. Note that since $u, v, u', v' \in S$, edges $(u, v), (u, u'), (v, v')$ are all in $G[S]$. If $(u, u'), (v, v')$ are not violating, then u, v must both be in C_2. However, this means (u, v) is violating, thus one of these three edges is always a violating edge for $S_1 \sqcup S_2$. ■

This observation is useful because it implies that $G[S]$ is not bipartite if every partition $C = C_1 \sqcup C_2$ has a ruining edge. The next claim naturally follows.

Claim 5 *With probability $\geq 2/3$, every partition $C = C_1 \sqcup C_2$ has a ruining edge $e \in G[S]$.*

To prove this claim, we will need to use the fact that G is ϵ-far from bipartite. Thus, we transition from considering partitions of C to partitions of the entire vertex set V, because each partition has $> \epsilon n^2$ violating edges.

Fix $C = C_1 \sqcup C_2$, and define $V = V_1 \sqcup V_2$ by taking $V_1 = N(C_2)$, $V_2 = V \backslash V_1$. The idea for this proof will be to use $V_1 \sqcup V_2$'s violating edges to upper bound $\Pr[C_1 \sqcup C_2$ has no ruining edge$]$. After bounding this quantity, we will then be able to union bound over these events to get an upper bound for $\Pr[\exists$ partition with no ruining edge$]$. To do this, we first need to prove a few more observations and lemmas.

Observation 6 *If edge $e = (u, v)$ violates $V_1 \sqcup V_2$ and $u, v \in N(C)$, then e is a ruining edge for $C_1 \sqcup C_2$.*

Proof We consider the two cases of whether u, v are in V_1 or V_2.

If $u, v \in V_1$, then by definition $V_1 = N(C_2)$ implies e ruins $C_1 \sqcup C_2$.

If $u, v \in V_2$, we have $u, v \in N(C)$ but $u, v \notin N(C_2)$, thus $u, v \in N(C_1)$, and e is again a ruining edge. ■

Let X be the number of edges in G with some endpoint not in $N(C)$. We will next lower bound the number of edges ruining $C_1 \sqcup C_2$ in terms of X.

We know that for graphs G ϵ-far from bipartite, any partition $V_1 \sqcup V_2$ has $> \epsilon n^2$ violating edges. Combining this with Observation 6,

$$\begin{aligned}
[\#\text{ edges ruining } C_1 \sqcup C_2] &\geq [\#e = (u,v) \text{ violating } V_1 \sqcup V_2 \text{ with } u, v \in N(C)] \\
&= [\#e \text{ violating } V_1 \sqcup V_2] - [\#e \text{ violating } V_1 \sqcup V_2 \text{ with } u \text{ or } v \text{ in } V \backslash N(C)] \\
&\geq \epsilon n^2 - [\#e \text{ with } u \text{ or } v \text{ in } V \backslash N(C)] \\
&= \epsilon n^2 - X.
\end{aligned}$$

Now we bound X. Say vertex v is *low degree* if $deg(v) < \frac{\epsilon n}{3}$, and *high degree* if $deg(v) \geq \frac{\epsilon n}{3}$. We bound the contribution to X from low and high degree vertices.

The low degree vertices contribute $\leq n \cdot \frac{\epsilon n}{3} = \frac{\epsilon n^2}{3}$ edges, because each low degree vertex has at most $\frac{\epsilon n}{3}$ edges.

The high degree vertices satisfy the following lemma.

Lemma 7 *With probability $\geq 5/6$, the number of high degree vertices $v \notin N(C)$ is $\leq \frac{\epsilon n}{3}$.*

Proof Fix a high degree vertex v. Each $u \in C$ has $\geq \frac{\epsilon}{3}$ chance of being adjacent to v, thus

$$\begin{aligned}
\Pr[v \notin N(C)] &\leq \left(1 - \frac{\epsilon}{3}\right)^{|C|} \\
&\leq e^{-\epsilon |C|/3} \\
&\leq \frac{\epsilon}{18},
\end{aligned}$$

where the last line follows from the taking constant for $|C| = \Theta(\frac{1}{\epsilon} \cdot \log \frac{1}{\epsilon})$ large enough. Thus,

$$\mathbb{E}[\#\text{ high degree vertices } v \notin N(C)] \leq \frac{\epsilon n}{18}.$$

Finally, by Markov's Inequality,

$$\begin{aligned}
\Pr\left[\#\text{ high degree vertices } v \notin N(C) \geq \frac{\epsilon n}{3}\right] &\leq \frac{\epsilon n/18}{\epsilon n/3} \\
&= 1/6,
\end{aligned}$$

as desired. ■

Therefore with probability $\geq 5/6$, there are at most $\frac{\epsilon n}{3}$ high degree vertices which each have at most n edges, for a total of $\frac{\epsilon n^2}{3}$ edges.

Thus with probability $\geq 5/6$, $X \leq \frac{\epsilon n^2}{3} + \frac{\epsilon n^2}{3} = (\frac{2\epsilon}{3})n^2$. Now we are ready to prove Claim 5.

Proof of Claim 5: Suppose the event of Lemma 7 occurs. Then, $X \leq (\frac{2\epsilon}{3})n^2$, so

$$[\#\text{ edges ruining } C_1 \sqcup C_2] \geq \epsilon n^2 - \left(\frac{2\epsilon}{3}\right) n^2 = \frac{\epsilon n^2}{3}.$$

Now, for any partition $C = C_1 \sqcup C_2$, considering $|S|/2$ pairs of independent edges gives

$$\begin{aligned}
\Pr[\text{no edge } e \in G[S] \text{ ruins } C_1 \sqcup C_2] &\leq \left(1 - \frac{\epsilon}{3}\right)^{|S|/2} && (1) \\
&\leq e^{-\epsilon |S|/6} && (2) \\
&\leq e^{-\Theta(|C|)} && (3) \\
&\leq \frac{2^{-|C|}}{6}, && (4)
\end{aligned}$$

where (2) to (3) follows from $|S| = \Theta(\frac{1}{\epsilon} \cdot |C|)$ and (3) to (4) follows from taking constant in $|S| = \Theta(|C|)$ large enough. Taking a union bound over all $2^{|C|}$ partitions $C = C_1 \sqcup C_2$,

$$\begin{aligned}\Pr[\exists\ C = C_1 \sqcup C_2 \text{ which has no ruining edge } e \in G[S]] &\leq 2^{|C|} \cdot \frac{2^{-|C|}}{6} \\ &= \frac{1}{6}.\end{aligned}$$

We use union bound once more with the probability $\leq 1/6$ that Lemma 7 fails, to get that with probability $\leq 1/3$, there exists a partition $C_1 \sqcup C_2$ with no ruining edge. Thus with probability $\geq 2/3$, every partition $C = C_1 \sqcup C_2$ has a ruining edge, completing the proof of Claim 5. ■

Now, Claim 5 and Observation 4 together imply that with probability $\geq 2/3$, $G[S]$ is not bipartite, concluding the proof of Theorem 1. ■

The runtime of this algorithm is $\Theta(|S|^2) = \Theta(\frac{1}{\epsilon^4} \log^2(\frac{1}{\epsilon}))$, which comes from querying all the edges in $G[S]$. These edges are needed for checking if $G[S]$ is bipartite, which can be done with a simple greedy algorithm over the $\Theta(|S|^2)$ edges. It turns out this runtime can be dropped to approximately $\Theta(\frac{1}{\epsilon^2})$ with a different algorithm that only requires querying edges in C.

2 Triangle-free Testing

2.1 Setup

In this problem, we explore sublinear time property testing for whether a graph is triangle-free.

Definition 8 *For graph $G = (V, E)$, a set of vertices (u, v, w) is a* triangle *if edges $(u, v), (v, w), (w, u)$ are all in E.*

We say a graph G is *triangle-free* if no three vertices form a triangle. The goal is to find a sublinear time triangle-free tester that will:

1. *Accept* if G is triangle-free
2. *Reject* if G is ϵ-far from triangle-free with probability $\geq 2/3$.

2.2 Algorithm

The following algorithm tests if a graph is triangle-free.

Algorithm 2 Triangle-Free Tester

```
t ← some sublinear parameter
for t iterations do
    Sample vertices (u, v, w) and check if it is a triangle
    Reject if yes
end for
Accept
```

Observe as a baseline that $t = O(n^3)$ would work by checking most of the $O(n^3)$ triples of vertices with high probability. What we will show is that this algorithm also works for $t = f(\epsilon)$, where f is some function of ϵ.

2.3 Analysis Overview

Our goal is to show this algorithm works for $t = f(\epsilon)$. This will follow from the next lemma.

Lemma 9 (Triangle Removal Lemma) *For all $\epsilon > 0$, there exists $\delta > 0$ such that for any graph G that is ϵ-far from triangle-free, G has at least δn^3 triangles.*

Note that given this lemma, taking $t = O(\frac{1}{\delta})$ is sufficient for Triangle-Free Tester because the probability of drawing no triangles will be $\leq (1-\delta)^{O(\frac{1}{\delta})} \leq e^{-\delta O(\frac{1}{\delta})}$, which can be made sufficiently small with large enough constant.

The remainder of this lecture discusses two important lemmas necessary for the proof of the Triangle Removal Lemma, which we will complete next class. Intuitively, the two lemmas are:

- Szemerédi Regularity Lemma (SRL): Every graph is "close" to a "random-like" graph.
- Triangle Counting Lemma: A "random-like" graph that is dense has many triangles.

To understand what a "random-like graph" is and formally state these two lemmas, we provide some definitions.

2.4 Definitions

First, we define a random graph.

Definition 10 *Let X, Y, Z be sets of nodes. A* random graph *G on $X \sqcup Y \sqcup Z$ is constructed by adding each edge*

$$(x, y) \in X \times Y \text{ with probability } p_{xy},$$
$$(y, z) \in Y \times Z \text{ with probability } p_{yz},$$
$$(z, x) \in Z \times X \text{ with probability } p_{zx},$$

where p_{xy}, p_{yz}, p_{zx} are constants and each edge is added independently.

Observe that in a random graph,

$$\Pr[(x, y, z) \in X \times Y \times Z \text{ is a triangle}] = p_{xy}p_{yz}p_{zx},$$

and the expected number of triangles is

$$\mathbb{E}[\#\text{ triangles}] = p_{xy}p_{yz}p_{zx}|X||Y||Z|.$$

Fix graph G and $X, Y \subseteq V(G)$. We also define the following terminology.

Definition 11 *Define $e(X, Y) = |\{(x, y) \in X \times Y \mid (x, y) \text{ is an edge in } E(G)\}|$.*

Definition 12 *Define the* edge density *between X and Y to be*

$$d(X, Y) = \frac{e(X, Y)}{|X||Y|}.$$

Definition 13 *The pair of vertex sets (X, Y) is called* ϵ-regular *if for all $A \subseteq X, B \subseteq Y$ with $|A| \geq \epsilon|X|$ and $|B| \geq \epsilon|Y|$,*

$$|d(A, B) - d(X, Y)| < \epsilon.$$

Definition 14 *The partition $V = V_1 \sqcup V_2 \sqcup \cdots \sqcup V_k$ is called a* ϵ-regular partition *if*

$$\sum_{\substack{i \leq i,j \leq k \\ (V_i, V_j) \text{ not } \epsilon\text{-regular}}} |V_i||V_j| \leq \epsilon n^2.$$

We are now ready to state the Szemerédi Regularity Lemma and the Triangle Counting Lemma.

2.5 Lemmas

We first state the Szemerédi Regularity Lemma, which intuitively states that every graph is "close" to a "random-like" graph.

Lemma 15 (Szemerédi Regularity Lemma) *For any $\epsilon > 0$ and positive integer m, there exists integer M such that any graph G with at least M vertices has an ϵ-regular partition on k parts, where $m \leq k \leq M$.*

The proof of this theorem is long and is left out of this class.

Next, we state and prove the Triangle Counting Lemma. Recall that the informal statement was that a "random-like" graph that is dense has many triangles. The formal definition is as follows.

Lemma 16 (Triangle Counting Lemma) *Suppose $X, Y, Z \subset V(G)$ are vertex sets that are pairwise ϵ-regular with edge densities $d_{XY}, d_{YZ}, d_{ZX} \geq 2\epsilon$. Then, the number of triangles in $X \times Y \times Z$ is at least*

$$(1-2\epsilon)(d_{XY}-\epsilon)(d_{YZ}-\epsilon)(d_{ZX}-\epsilon)|X||Y||Z|.$$

Proof Let S_Y be the set of vertices in X which have $< (d_{XY}-\epsilon)|Y|$ neighbors in Y. Observe that $|S_Y| \leq \epsilon|X|$; otherwise (S_Y, Y) contradicts (X, Y) being ϵ-regular.

Similarly, let S_Z be the set of vertices in X which have $< (d_{XZ}-\epsilon)|Z|$ neighbors in Z. By similar reasoning, $|S_Z| \leq \epsilon|X|$.

Now take $X' = X \setminus (S_Y \cup S_Z)$, X' has at least $(1-2\epsilon)|X|$ vertices. Additionally, for each $x \in X'$, $x \notin S_Y, S_Z$ implies that

$$|N_Y(x)| \geq (d_{XY}-\epsilon)|Y|$$
$$|N_Z(x)| \geq (d_{XZ}-\epsilon)|Z|.$$

Finally, (Y, Z) is ϵ-regular, implying that $d(N_Y(x), N_Z(x)) \geq d_{YZ} - \epsilon$.

Combining all these observations, for each of the at least $(1-2\epsilon)$ vertices $x \in X'$, there are at least $(d_{XY}-\epsilon)|Y|(d_{ZX}-\epsilon)|Z|$ pairs of edges to N_Y, N_Z, and $(d_{YZ}-\epsilon)$ of these are connected across Y and Z. This gives a total of

$$\geq (1-2\epsilon)|X|(d_{XY}-\epsilon)|Y|(d_{ZX}-\epsilon)|Z|(d_{YZ}-\epsilon)$$
$$= (1-2\epsilon)(d_{XY}-\epsilon)(d_{YZ}-\epsilon)(d_{ZX}-\epsilon)|X||Y||Z|$$

triangles, as desired. ■

The Szemerédi Regularity Lemma and the Triangle Counting Lemma will be used to prove the Triangle Removal Lemma next class.