6.5240 Sublinear Time Algorithms	September 9th, 2022
Lecture 3	
Lecturer: Ronitt Rubinfeld	*Scribe: Guanghao Ye*

1 Approximate Average Degree

1.1 Problem Setup

Let's first formally state the problem:

Problem 1. *Given a graph $G = (V, E)$, an approximation parameter $\epsilon \in (0, 1)$, and a confidence parameter $\delta \in (0, 1)$. The goal is to output a $\tilde{d}$ such that*

$$\Pr\left[|\tilde{d} - \bar{d}| \le \epsilon \bar{d}\right] \ge 1 - \delta.$$

where $\bar{d} = \frac{2m}{n}$ is the average degree *of the graph.*

Throughout the lecture, we will have the following assumptions:

- The average degree $\bar{d} \ge 1$.
- We are given access to the following two queries:
  1. "degree queries": Given $v \in V$, output $\deg(v)$
  2. "neighbor queries": Given $(v, j) \in V \times \mathbb{N}$, output j-th neighbor of v.

1.2 Lower bound

Recall in the last lecture, we have shown that when the average degree is very small, it requires $\Omega(n)$ many queries. For example, considering to distinguish the graph with a single edge and the graph with no edge.

Here, we (informally) show a lower bound of $\Omega(\sqrt{n})$ queries. Let's consider the following two graphs: The cycle graph with n nodes C_n has average degree $\bar{d} = 2$. We construct another graph G consists of two connected components where one is a cycle graph with $n - c\sqrt{n}$ many nodes and the other component is a clique with $c\sqrt{n}$ many nodes. Then, the average degree for this graph is

$$\bar{d} = \frac{2m}{n} = \frac{2\left(\binom{c\sqrt{n}}{2} + n - c\sqrt{n}\right)}{n} = \frac{2n + c^2 n - c\sqrt{n}}{n} = 2 + c^2 - \frac{c}{\sqrt{n}} \approx 2 + c^2.$$

However, to distinguish these two graphs, the algorithm at least needs to sample one node from the clique. This shows $\Omega(\sqrt{n})$ many queries are necessary.

In today's lecture, we will show $\tilde{O}(\sqrt{n})$ many queries suffice.

1.3 Algorithm

1.3.1 Warm-up: Almost regular graphs

Let's consider a slightly easier problem: Assume each node has degree in $[\Delta, 10\Delta]$.

It's easy to see that the algorithm above has runtime $O(\frac{1}{\epsilon^2}\log(1/\delta))$.

Now, we show $\tilde{d}$ is a good approximation for the average degree.

Claim 2. *The output $\tilde{d}$ is an unbiased estimator: $\mathbb{E}[\tilde{d}] = \bar{d}$.*

Algorithm 1 Approximating Degree for almost regular graphs

$k \leftarrow \frac{50}{\varepsilon^2}\log(2/\delta)$
for $i = 1, \ldots, k$ **do**
  Pick $v_i \in_u V$ ▷ We use $x \in_u D$ to denote x is chosen uniformly at random from set D
  $X_i \leftarrow \deg(v_i)$
end for
return $\tilde{d} \leftarrow \frac{1}{k}\sum_{i=1}^{k} X_i$

Proof.

$$\begin{aligned}\mathbb{E}[\tilde{d}] &= \frac{1}{k}\sum_{i=1}^{k} E[X_i] \qquad \text{(linearity of expectation)}\\ &= \mathbb{E}[X_i] \qquad \text{(i.i.d)}\\ &= \sum_{v} \Pr[v_i \text{ is picked}] \cdot \deg(v_i)\\ &= \frac{1}{n} \cdot \sum_{v} \deg(v_i)\\ &= \frac{2m}{n} = \bar{d}\end{aligned}$$

□

Claim 3. *The output $\tilde{d}$ satisfies the requirement of Problem 1:* $\Pr\left[|\tilde{d} - \bar{d}| \leq \epsilon \bar{d}\right] \geq 1 - \delta$.

Before we proceed, let's introduce todays's Chernoff bound:

Theorem 4 (Hoeffding's inequality)**.** *$Y_1, \ldots, Y_k$ are independent random variable's such that $Y_i \in [0, 1]$ and $Y = \sum_{i=1}^{k} Y_i$. For $b \geq 1$, we have*

$$\Pr\left[|Y - \mathbb{E}[Y]| > b\right] \leq 2 \cdot \exp(-2b^2/k).$$

Now, we are ready to prove the claim above.

Proof of Claim 3. By the assumption of almost regular graph, X_i's are in $[\Delta, 10\Delta]$. Let $Z_i \leftarrow \frac{X_i}{10\Delta}$ and $Z = \sum_i Z_i$, then we have $Z_i \in [0, 1]$ and $\tilde{d} = \frac{10\Delta}{k} Z$.

Note that $\mathbb{E}[Z] = \frac{k}{10\Delta}\mathbb{E}[\tilde{d}] = \frac{k\bar{d}}{10\Delta}$. This implies

$$\left|\tilde{d} - \bar{d}\right| \leq \varepsilon\bar{d} \iff \left|\frac{10\Delta}{k} Z - \frac{10\Delta}{k}\mathbb{E}[Z]\right| \leq \varepsilon\bar{d} \iff |Z - \mathbb{E}[Z]| \leq \varepsilon\bar{d} \cdot \frac{k}{10\Delta}.$$

Using Theorem 4 above on Z, with $b = \frac{k}{10\Delta}\varepsilon\bar{d}$, we get

$$\Pr\left[|Z - \mathbb{E}[Z]| \geq \frac{k}{10\Delta}\varepsilon\bar{d}\right] \leq 2\exp(-2\frac{\varepsilon^2\bar{d}^2k^2}{100\Delta^2 k}) \leq 2\exp(-\frac{1}{50}k\varepsilon^2) \leq \delta$$

where second last step follows by $\bar{d}^2/\Delta^2 \geq 1$ by assumption, and the last step follows by choice of k. □

1.3.2 General Case

From Markov's inequality, we know that at most a $1/C$ fraction of nodes have degree larger than $C\bar{d}$. This implies most nodes satisfy the warm up case! However, the rest of nodes can have large degrees. To cope with this, we define a new notion of degree, denoted by $\deg^+(\cdot)$.

We first assign a total order on the nodes of graph by assuming each node has a unique ID, then the order is given by the ID.

Definition 5. *Given two nodes $u, v \in V$, we say $u \prec v$ if*

- $\deg(u) < \deg(v)$
- *or* $\deg(u) = \deg(v)$ *and u has smaller ID than v.*

Then, we define $\deg^+(u)$ as the number of nodes v in u's neighborhood such that $u \prec v$.

Intuitively, if we orienting edges from small to large, the $\deg^+(\cdot)$ count the "out-edges". Then, this directly implies

$$\sum_{u \in V} \deg^+(u) = m = \frac{n\bar{d}}{2}. \tag{1}$$

The benefits of having this notion is that the newly defined degree cannot be too large for any node in the graph:

Claim 6. *For any node $v \in V$, $\deg^+(v) \leq \sqrt{2m}$.*

Proof. We define the vertex set $H \subseteq V$ to be $\sqrt{2m}$ nodes with highest rank (degree) w.r.t. $\prec$. For any $v \in H$, $\deg^+(v) \leq \sqrt{2m}$. since edge leaving v go to bigger nodes, must be also in H.

For any $v \in V \setminus H$, we will show $\deg^+(v) \leq \deg(v) \leq \sqrt{2m}$. For the sake of contradiction, we assume $\deg(v) > \sqrt{2m}$, then all $w \in H$ have $\deg(w) \geq \deg(v)$, Then, we have total degree $\geq |H| \cdot \deg(v) \geq \sqrt{2m} \cdot \sqrt{2m} = 2m$ but total degree is $2m$. This is a contradiction. □

Now, we present our algorithm for the general case.

Algorithm 2 Approximating Degree

$k \leftarrow \frac{16}{\varepsilon^2}\sqrt{n}$
for $i = 1, \ldots, k$ **do**
  Pick $v_i \in_u V$ ▷ Step ①
  Pick $u_i \in_u N(v_i)$ ▷ Step ②
  Let $X_i = \begin{cases} 2\deg(v_i) & \text{if } v_i \prec u_i \\ 0 & \end{cases}$
end for
return $\tilde{d} \leftarrow \frac{1}{k}\sum_{i=1}^{k} X_i$

Claim 7. *X_i is an unbiased estimator of $\bar{d}$: $\mathbb{E}[X_i] = \bar{d}$.*

Proof.

$$
\begin{aligned}
\mathbb{E}[X_i] &= \sum_{v \in V} \Pr[v \text{ picked in } ①] \cdot \mathbb{E}[X_i \mid v \text{ picked in } ①] \\
&= \frac{1}{n} \sum_{v \in V} \sum_{u \in N(v)} \Pr[u \text{ pickied in } ②] \cdot \mathbb{E}[X_i \mid v \text{ picked in } ① \text{ and } u \text{ picked in } ②] \\
&= \frac{1}{n} \sum_{v \in V} \sum_{u \in N(v), v \prec u} \frac{1}{\deg(v)} \cdot 2\deg(v) \\
&= \frac{2}{n} \sum_{v \in V} \deg^+(v) \\
&= \bar{d}.
\end{aligned}
$$

where the third step follows by definition of X_i given in Algorithm 2, the fourth step follows by definition of $\deg^+(\cdot)$, and the last step follows by Equation (1).

□

In the next lecture, we will show $\mathrm{Var}[X_i]$ is small by using the upper bound on $\deg^+(\cdot)$.