Approximate Average Degree

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

\[
\Pr\left[ |\tilde{d} - \bar{d}| \leq \epsilon \bar{d} \right] \geq 1 - \delta.
\]

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 \( \bar{d} \geq 1 \).
- We are given access to the following two queries:
  1. “degree queries”: Given \( v \in V \), output \( \deg(v) \)
  2. “neighbor queries”: Given \( (v,j) \in V \times N \), 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

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

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

1: $k \leftarrow \frac{20}{\varepsilon^2} \log(2/\delta)$
2: for $i = 1, \ldots, k$ do
3: \hspace{1em} Pick $v_i \in u$ \\
4: \hspace{1em} $X_i \leftarrow \deg(v_i)$
5: \hspace{1em} end for
6: return $\tilde{d} \leftarrow \frac{1}{k} \sum_{i=1}^{k} X_i$

Proof.

\[
E[\tilde{d}] = \frac{1}{k} \sum_{i=1}^{k} E[X_i] \quad \text{(linearity of expectation)}
\]
\[
= E[X_i] \quad \text{(i.i.d)}
\]
\[
= \sum_v \Pr[v_i \text{ is picked}] \cdot \deg(v_i)
\]
\[
= \frac{1}{n} \cdot \sum_v \deg(v_i)
\]
\[
= \frac{2m}{n} = \bar{d}
\]

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

Before we proceed, let’s introduce today’s Chernoff bound:

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

\[
\Pr[|Y - E[Y]| > b] \leq 2 \cdot \exp(-2b^2/k).
\]

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 $\bar{d} = \frac{10\Delta}{k} Z$.

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

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

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

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

where second last step follows by $\bar{d}^2/\Delta^2 \geq 1$ by assumption, and the last step follows by choice of $k$. \qed
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 $\text{deg}^+(\cdot)$.

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

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

Then, we define $\text{deg}^+(u)$ 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 $\text{deg}^+(\cdot)$ count the “out-edges”. Then, this directly implies

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

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 $v \in V$, $\text{deg}^+(v) \leq \sqrt{2m}$.

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

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

Now, we present our algorithm for the general case.

**Algorithm 2** Approximating Degree

1: $k \leftarrow \frac{16}{\epsilon^2} \sqrt{n}$
2: for $i = 1, \ldots, k$ do
3: Pick $v_i \in u V$ \hspace{1cm} \triangleright\text{Step 1}
4: Pick $u_i \in u N(v_i)$ \hspace{1cm} \triangleright\text{Step 2}
5: Let $X_i = \begin{cases} 2 \text{deg}(v_i) & \text{if } v_i \prec u_i \\ 0 & \text{otherwise} \end{cases}$
6: end for
7: return $\tilde{d} \leftarrow \frac{1}{k} \sum_{i=1}^{k} X_i$

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

3
Proof.

\[ E[X_i] = \sum_{v \in V} \Pr[v \text{ picked in } 1] \cdot E[X_i \mid v \text{ picked in } 1] \]

\[ = \frac{1}{n} \sum_{v \in V} \sum_{u \in N(v)} \Pr[u \text{ picked in } 2] \cdot E[X_i \mid v \text{ picked in } 1 \text{ and } u \text{ picked in } 2] \]

\[ = \frac{1}{n} \sum_{v \in V} \sum_{u \in N(v), v < u} \frac{1}{\deg(v)} \cdot 2 \deg(v) \]

\[ = \frac{2}{n} \sum_{v \in V} \deg^+(v) \]

\[ = d. \]

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

\[ \square \]

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