These lecture notes have not undergone rigorous peer-review. Please email quanquan.liu@yale.edu if you see any errors.

1 Introduction

We continue with our discussion of estimating the average degree today. Recall the problem I posed at the end of class from last week. The question is: given an input graph $G = (V, E)$, direct (orient) the edges of the graph from the lower degree endpoint to higher degree endpoint (breaking ties by vertex ID). What is the maximum outdegree of any vertex?

We'll now show that the maximum outdegree is bounded by $\sqrt{2m}$ where m is the number of edges in the graph. First, note that in order for an edge to be oriented from vertex u to v, the degree of v must be at least the degree of u. Hence, in order for v to have out-degree X, the edges from v must be to vertices with degree at least X. Thus, assuming that v has out-degree X which is equal to the maximum out-degree, $X^2 \leq 2m$ since each of m edges can contribute to at most 2 vertex's degree; then $X \leq \sqrt{2m}$.

We have thus proven the following lemma.

Lemma 1.1. *Given a graph $G = (V, E)$ with m edges and unique vertex IDs in $[n]$, when all edges are oriented from low-degree to high-degree vertices (breaking ties by smaller vertex ID), the maximum out-degree of any vertex is $\sqrt{2m}$.*

We now proceed with the main subject of today's class which is estimating the average degree of vertices in the adjacency-list sublinear model. Today's lecture if based off the recent paper of Eden, Ron and Seshadhri [ERS17]. In the previous lecture, we discussed an algorithm for estimating the average degree by bucketing vertices by degree and sampling from these buckets. We completely disregard buckets with too small samples and use the counts from the remaining buckets to produce our estimate. In today's class, instead of counting each edge twice as in the case of the bucketing algorithm, we instead count each edge once and *assign the responsibility* of counting that edge to the lower degree endpoint (breaking ties by smaller index).

We first start with some definitions.

Theorem 1 (Total Ordering). *A total ordering $\prec$ on vertices of $G = (V, E)$ is an ordering where for any pair of distinct vertices $u \neq v \in V$, $u \prec v$ if and only if either:*

- $\deg(u) < \deg(v)$, *or*
- $\deg(u) = \deg(v)$ *and* $ID(u) < ID(v)$.

Let $\deg^+(v)$ be the out-degree of vertex v in the ordering.

Our goal today is to estimate $\deg^+(v)$. First, note the following simple observation.

Observation 1.2. $\sum_{v \in V} \deg^+(v) = \frac{n\bar{d}}{2}$.

The observation holds since $\sum_{v \in V} \deg^+(v) = m$ and $n\bar{d} = 2m$.

2 Algorithm

Now we introduce our algorithm with pseudocode given in Algorithm 1.

We now prove the first lemma which shows that the expectation of *any* X_i variable is equal to the average degree.

Algorithm 1: $(1+\varepsilon)$-Approx Average Degree Estimation in Adjacency-List Sublinear Model

```
Function AvgDeg($\varepsilon, n$)
    $k \leftarrow \frac{16}{\varepsilon^2} \cdot \sqrt{n}$
    for $i \leftarrow 1$ to $k$ do
        Sample a vertex $v_i$ uniformly at random
        Sample a neighbor $u_i \in N(v_i)$ uniformly at random
        if $v_i \prec u_i$ then
            $X_i \leftarrow 2 \cdot \deg(v_i)$
        else
            $X_i \leftarrow 0$
        end
    end
    return $\tilde{d} \leftarrow \frac{1}{k} \cdot \sum_{i=1}^{k} X_i$
```

Lemma 2.1. $\mathbb{E}[X_i] = \overline{d}$.

Proof. We use the Law of Total Expectation to write the following expressions.

$$\begin{aligned}
\mathbb{E}[X_i] &= \sum_{v \in V} \left(\Pr[v \text{ is sampled from } V] \cdot \mathbb{E}[X_i \mid v \text{ is sampled from } V]\right) \\
&= \sum_{v \in V} \frac{1}{n} \cdot \sum_{u \in N(v)} \left(\Pr[u \text{ is sampled from } N(v) \mid v \text{ is sampled}] \cdot \mathbb{E}[X_i \mid u \text{ and } v \text{ are sampled}]\right) \\
&= \sum_{v \in V} \frac{1}{n} \cdot \sum_{u \in N(v), v \prec u} \left(\frac{1}{\deg(v)} \cdot 2\deg(v)\right) \\
&= \frac{1}{n} \cdot \sum_{v \in V} 2\deg^{+}(v) = \frac{2m}{n}.
\end{aligned}$$

The last line follows since the expression $\sum_{u \in N(v), v \prec u} 2$ means that each of v's out-degree neighbors contributes 2 to the sum. □

Now that we have shown the expectation, we will now upper bound the variance of each X_i.

Lemma 2.2. $\mathbf{Var}[X_i] \leq 4\sqrt{2m} \cdot \overline{d}$.

Proof. We know $\mathbf{Var}[X_i] = \mathbb{E}[X_i^2] - \mathbb{E}[X_i]^2 \leq \mathbb{E}[X_i^2]$. Then, using the exact same calculation that we used in the proof of Lemma 2.1 except we substitute X_i^2 for X_i, we can show

$$\begin{aligned}
\mathbb{E}[X_i^2] &= \sum_{v \in V} \big(\mathsf{Pr}[v \text{ is sampled from } V] \cdot \mathbb{E}[X_i^2 \mid v \text{ is sampled from } V]\big) \\
&= \sum_{v \in V} \frac{1}{n} \cdot \sum_{u \in N(v)} \big(\mathsf{Pr}[u \text{ is sampled from } N(v) \mid v \text{ is sampled}] \cdot \mathbb{E}[X_i^2 \mid u \text{ and } v \text{ are sampled}]\big) \\
&= \sum_{v \in V} \frac{1}{n} \cdot \sum_{u \in N(v), v \prec u} \left(\frac{1}{\deg(v)} \cdot (2\deg(v))^2\right) \\
&= \frac{1}{n} \cdot \sum_{v \in V} \deg^+(v) \cdot (4\deg(v)) \\
&\leq \frac{1}{n} \cdot \left(\sum_{v \in V} \deg^+(v)\right) \cdot \sum_{v \in V} 4\deg(v) \\
&= \frac{4\sqrt{2m}}{n} \cdot \sum_{v \in V} \deg(v) = 4\sqrt{2m} \cdot \overline{d}.
\end{aligned}$$

□

We now show our concentration bound using the median-of-means trick that we have seen before in our previous lectures. Recall that the median of means trick averages the valuesof k independent triangles to reduce the variance of by a factor of k given independent samples.

Lemma 2.3. **Var** $\left[\frac{1}{k}\sum_{i=1}^{k} X_i\right] = \frac{\mathbf{Var}[X_i]}{k}$.

Now, we can use Chebyshev's inequality to bound our probability of success by at least $3/4$; then we can amplify the probability of success using the median trick by performing $O(\log(1/\delta))$ independent trials to get a probability of success of at least $1 - \delta$. Recall Chebyshev's inequality below.

Theorem 2 (Chebyshev's Inequality)**.** *Chebyshev's Inequality states that for any (not necessarily positive) random variable X with finite expected value μ and finite non-zero variance σ^2, the probability that X is more than k standard deviations away from μ is at most $1/k^2$. Formally, it is expressed as:*

$$P(|X - \mu| \geq k) \leq \frac{\mathbf{Var}[X]}{k^2}$$

for all $k > 0$.

Theorem 2.4. *Algorithm 1 gives a $(1+\varepsilon)$-approximation of the average degree of input graph $G = (V, E)$ using $O\left(\frac{\sqrt{n}\log(1/\delta)}{\varepsilon^2}\right)$ with probability at least $1 - \delta$.*

Proof. Using Chebyshev's inequality, we get that the probability of failure is at most

$$\begin{aligned}
\mathsf{Pr}\left[|\tilde{d} - \mathbb{E}[\tilde{d}] > \varepsilon \cdot \overline{d}\right] &\leq \frac{\mathbf{Var}[\tilde{d}]}{\varepsilon^2 \overline{d}^2} \\
&= \frac{4\sqrt{2m} \cdot \overline{d}}{k\varepsilon^2 \overline{d}^2} \quad \text{where } k \text{ is defined in Algorithm 1} \\
&= \frac{4\sqrt{2mn}}{k\varepsilon^2 2m} \\
&= \frac{4n}{k\varepsilon^2 \sqrt{2m}} \\
&= \frac{\sqrt{n}}{4\sqrt{2m}} \quad \text{since } k = \frac{16}{\varepsilon^2} \cdot \sqrt{n} \\
&< \frac{1}{4} \quad \text{since } \overline{d} \geq 1.
\end{aligned}$$

Finally, we repeat the entire Algorithm 1 over $O(\log(1/\delta))$ trials to use the median trick to amplify our probability of success to $1 - \delta$. □

This lecture uses notes from [Ass20] and [Ras20].

References

[Ass20] Sepehr Assadi. CS 514: Advanced Algorithms II – Sublinear Algorithms. https://sepehr.assadi.info/courses/cs514-f21/lec2.pdf, 2020.

[ERS17] Talya Eden, Dana Ron, and C Seshadhri. Sublinear time estimation of degree distribution moments: The degeneracy connection. In *44th International Colloquium on Automata, Languages, and Programming (ICALP 2017)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2017.

[Ras20] Sofya Raskhodnikova. CS 591 Sublinear Algorithms. https://cs-people.bu.edu/sofya/sublinear-course/slides/Sublinear20-lec17.pdf, 2020.