0368.416701 Sublinear Time Algorithms November 30, 2015

Lecture 7

Lecturer: Ronitt Rubinfeld *Scribe: A. Boag, T. Haimovich, E. Kuperwasser, A. Rosin*

1 Introduction

We would like to estimate the average degree of graph $G = (V, E)$, defined as $\bar{d} = \frac{\sum_{u \in V} d(u)}{n}$. Our estimation $\tilde{d}$ will hold $\bar{d} \leq \tilde{d} \leq (1+\epsilon)\bar{d}$.

Assumptions

1. G is a simple graph, no self-loops.
2. G isn't super-sparse, i.e. $|E| = m = \Omega(n)$.
3. $\forall v \in V$ we have two sorts of queries:
   - degree query: getting v's degree $d(v)$
   - neighbour query: getting v's i^{th} neighbour (from the neighbour array)

Reminder In the previous lesson we had two observations

- The plug-in estimation using Chernoff fails because the variance is too large
- We have a lower bound of $\Omega(\sqrt{n})$ queries

To counter the variance issue we'll use *Bucketing*

Buckets

Set $\beta = \frac{\epsilon}{c}$ and $t = O(\frac{\log n}{\epsilon})$, where t is the number of buckets we'll use.

We define a set of buckets: $B_i = \{v \mid (1+\beta)^{i-1} < d(v) \leq (1+\beta)^i\}$ for $i \in \{0, ..., t-1\}$. The buckets therefore contain vertices with similar degrees. Notice that vertices of degree 0 are not counted as the final algorithm does not use them.

The total degree of nodes in B_i is

$$(1+\beta)^{i-1} \mid B_i \mid \leq d_{B_i} \leq (1+\beta)^i \mid B_i \mid \tag{1}$$

and the total degree of the graph is

$$\sum_i (1+\beta)^{i-1} \mid B_i \mid \leq d_{total} \leq \sum_i (1+\beta)^i \mid B_i \mid \tag{2}$$

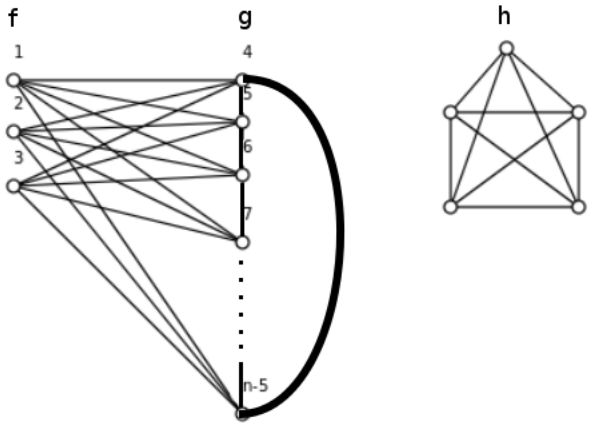
Notice that we estimated d_{total} to a multiplicative factor of $1+\beta$.

Our plan is therefore to estimate $\sum_i (1+\beta)^{i-1} \mid B_i \mid$ and divide by n to get $\bar{d}$.

1.1 First idea

1. Sample vertices into S (size to be determined)
2. Set $\forall i, S_i \leftarrow S \cap B_i$
3. Estimate $\frac{|B_i|}{n}$ for all i by $\rho_i \leftarrow \frac{|s_i|}{|S|}$
4. Output $\sum_i \rho_i (1+\beta)^{i-1}$

Example



The vertices in the above graph can be divided to 3 groups: f, g and h. The algorithm will place these vertices in three buckets: B_f, B_g, B_h.

bucket	f	g	h
degree	n-8	5	4
size	3	n-8	5

The graph, therefore, has an average degree $\bar{d} = \frac{1}{n} \cdot [(n-8) \cdot 3 + 5 \cdot (n-8) + 4 \cdot 5] \approx 8$.

However, the algorithm presented above will (with high probability) only sample vertices from g. This means we'll have an estimation of 5 instead.

1.2 Second idea

1. Take sample S, $|S| = \Theta(\sqrt{n} \cdot polylog(n) poly(\frac{1}{\epsilon}))$
2. $\forall i$, $S_i \leftarrow S \cap B_i$
3. Estimate $\frac{|B_i|}{n}$:
   (a) if $|S_i| \geq \sqrt{\frac{\epsilon}{n}} \frac{|S|}{c \cdot t}$, use $\rho_i = \frac{|S_i|}{|S|}$
   (b) else, $\rho_i \leftarrow 0$
4. Output $\sum \rho_i (1+\beta)^{i-1}$

Remark $|S| \geq \sqrt{\frac{n}{\epsilon}} \frac{log(n)}{\epsilon}$, otherwise the condition is trivially false

Definition i is *heavy* if $|S_i| \geq \sqrt{\frac{\epsilon}{n}} \frac{|S|}{c \cdot t}$, otherwise i is *light*

Analysis

Claim (with high probability) Output is not too large
(ideal case) Suppose $\forall i$, $\rho_i = \frac{|B_i|}{n}$. then $\sum \rho_i (1+\beta)^{i-1} = \sum \frac{|B_i|}{n}(1+\beta)^{i-1} \leq \bar{d}$
(reality) Suppose $\forall i$, $\rho_i \leq \frac{|B_i|}{n}(1+\gamma)$. then $\sum \rho_i (1+\beta)^{i-1} \leq \bar{d}(1+\gamma)$

- if i is light: $\rho_i = 0 \leq \frac{|B_i|}{n}$
- if i is heavy: $\rho_i \leq \frac{|B_i|}{n}(1+\gamma)$ (with high probability) by Chernoff/Hoeffding

So our assumption holds, and so does our claim.

Can the output be too small?

for heavy i: $\rho_i \geq \frac{|B_i|}{n}(1-\gamma)$

for light i: good question!

Definition We classify the types of edges in G:

1. *heavy-heavy*: both u,v in heavy bucket. those edges are counted twice.
2. *heavy-light*: one of u,v in heavy bucket, one in light bucket. those edges are counted once.
3. *light-light*: both u,v in light bucket. those edges are never counted.

Note we should count every edge twice in order to get the total degree

How many light-light edges do we have?

With high probability, $\forall$ light $i, |B_i| < \sqrt{\frac{\epsilon}{n}}\frac{2n}{c \cdot t}$

This is true since $E[S_i] = |S|\frac{|B_i|}{n}$, so if $|B_i| > \sqrt{\frac{\epsilon}{n}}\frac{2n}{c \cdot t}$ then $E[S_i] > \sqrt{\frac{\epsilon}{n}}\frac{2|S|}{c \cdot t}$, which means that with high probability i is heavy.

As a result, even if all the light nodes were connected in one clique, the total number of edges is still bounded by:

light-light edges $\leq (\frac{2\sqrt{\epsilon n}}{c \cdot t} \cdot t)^2 = O(\epsilon n)$ since we have t buckets.
So ignoring light-light edges affects our approximation of $\bar{d}$ by an additive error of ϵn at most

Remark We assumed G is not super-sparse, and thus an additive error of ϵ will become a multiplicative error of $1+\epsilon$, so in total we have a multiplicative error of $2+\epsilon$! (since the heavy-light edges are underestimated by at most factor $\frac{1}{2}$)

1.3 Third (and final) Idea

We'll fix our estimation by considering the light-heavy edges.

1. Take sample S, $|S| = \Theta(\sqrt{n} \cdot polylog(n) poly(\frac{1}{\epsilon}))$
2. $\forall i$, $S_i \leftarrow S \cap B_i$
3. Estimate $\frac{|B_i|}{n}$:
   (a) if $|S_i| \geq \sqrt{\frac{\epsilon}{n}}\frac{|S|}{c \cdot t}$, use $\rho_i = \frac{|S_i|}{|S|}$. also - set $\alpha_i \leftarrow Fix(S_i)$

(b) else, $\rho_i \leftarrow 0$

4. Output $\sum_{heavy} \rho_i (1+\beta)^{i-1}(1+\alpha_i)$

the $1+\alpha_i$ factor will function as the light-heavy correction

What's "Fix"?

We still need to show how Fix works. First, we explain how to choose a random neighbour of a given vertex.

random neighbour query(v):

- find $d(v)$ using a degree query on v
- pick $i \leftarrow random(1, ...d(v))$
- return v's i^{th} neighbour (using a neighbour query)

Now we can define $Fix(S_i)$:

- $\forall u \in S_i$:
  - $v \leftarrow$ random neighbour of u
  - set a_u as the indicator of whether (u, v) is a heavy-light edge
- Output $\leftarrow \alpha_i = \frac{\sum a_u}{|S_i|}$

Analysis

Let HL_i be the number of heavy-light edges in B_i. Consider S_i as a set of nodes in B_i chosen uniformly at random.

Observation $E[a_u] = Pr(a_u = 1)$ is the probability that an edge originating in u is heavy-light. Thus, $E[\alpha_i] = \sum_{u \in S_i} \frac{Pr(a_u=1)}{|S_i|}$ is the probability of choosing a heavy-light edge (first, choose a vertex $u \in B_i$ at random then choose a neighbour at random).

Proof The philosophy behind this proof is that since the nodes are all in the same bucket, their degrees are pretty close.

The easy case All nodes have the same degree d.

Denote by p the probability that a specific heavy-light edge is chosen. In this case: $p = \frac{1}{d|B_i|}$

As a result, $E[\alpha_i] = Pr(\text{any heavy-light edge is chosen}) = \frac{HL_i}{d|B_i|}$

The general case All nodes are in the same bucket, so all the degrees are in $((1+\beta)^{i-1}, (1+\beta)^i]$. In this case the probability of choosing a specific heavy-light edge is:

$$\frac{1}{|B_i|(1+\beta)^i} \leq p \leq \frac{1}{|B_i|(1+\beta)^{i-1}} \tag{3}$$

therefore

$$\frac{HL_i}{|B_i|(1+\beta)^i} \leq E[\alpha_i] \leq \frac{HL_i}{|B_i|(1+\beta)^{i-1}} \tag{4}$$

which means

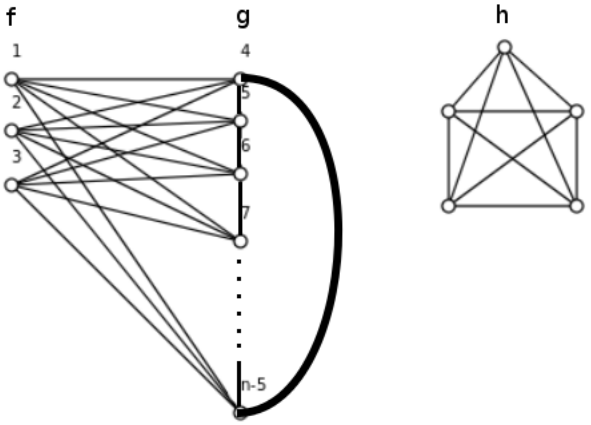
$$E[\alpha_i]|B_i|(1+\beta)^{i-1} \leq HL_i \leq E[\alpha_i]|B_i|(1+\beta)^i \tag{5}$$

Using Chernoff we'll have a $(1+\epsilon)$-approximation of $E[\alpha_i]$ by α_i , so in total we have a $(1+\epsilon)(1+\beta)$-approximation of HL_i, using $\alpha_i|B_i|(1+\beta)^{i-1}$.

All in all, $\alpha_i\rho_i(1+\beta)^{i-1}$ is a $(1+\epsilon)(1+\beta)$-approximation of $\frac{HL_i}{n}$.

Result Our final approximation of $\sum_{heavy}\rho_i(1+\beta)^{i-1}(1+\alpha_i)$ counts the heavy-heavy edges twice and the heavy-light edges once, then adds a correction by counting the heavy-light edges once more.

Back to the Example



Applying the final algorithm to our previous example, we have $\rho_f = \rho_h = 0$ and $\rho_g = 1$ - since all our samples (with high probability) end up in bucket g.

As a result, our previous approximation was 5. However, this time we compute $\alpha_g \approx \frac{3}{5}$ (since 3 out of 5 neighbours of g are heavy-light) and multiply by the correction factor. The answer: $5(1+\frac{3}{5}) = 8$.