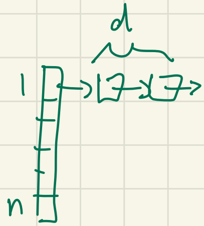
6.889 Sublinear time algorithms

Prof. Ronitt Rubinfeld

Lecture 1 Topics:

- Overview
- Diameter of a pointset
- # Connected Components

of Connected Components in graph:



Given: Graph G, max degree d ⟸ Adjacency list representation

parameter ε

$|V| = n$

$|E| = m \leq d \cdot n$

Output: let C = # conn comp in G

output y s.t.

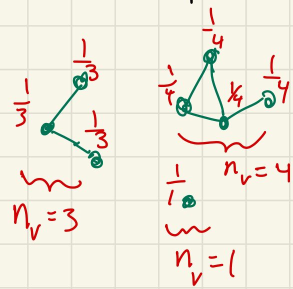
$$C - \varepsilon n \leq y \leq C + \varepsilon n$$

Main Insight:

new characterization of # conn components notation: $\forall v$ let n_v = # nodes in v's conn comp

observe: $\forall$ connected comp $A \subseteq V$

$$\sum_{u \in A} \frac{1}{n_u} = \sum_{u \in A} \frac{1}{|A|} = 1$$



conn comp $c = \sum_{u \in V} \frac{1}{n_u}$

better? compute n_u need $O(n)$ time?

compute sum of n things is $O(n)$?

$O(n^2)$ time?

Estimating $\frac{1}{n_u}$:

C = # conn comp

n_v = # nodes in v's comp

Fact $\quad C = \sum_{u \in V} \frac{1}{n_u}$

def $\quad \hat{n}_u = \min\{n_u, \underbrace{2/\varepsilon}\}$

$\hat{C} = \sum_{u \in V} \frac{1}{\hat{n}_u}$

why $2/\varepsilon$?

Lemma $\quad \forall u \quad \left| \frac{1}{n_u} - \frac{1}{\hat{n}_u} \right| \le \varepsilon/2$

Corr $\quad |C - \hat{C}| \le \frac{\varepsilon n}{2}$

Pf of lemma $\quad$ idea:

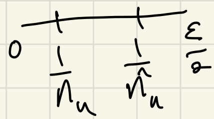
case 1 $\quad n_u$ small $(\le 2/\varepsilon) \Rightarrow n_u = \hat{n}_u$

$$\left| \frac{1}{n_u} - \frac{1}{\hat{n}_u} \right| = 0$$

Case 2 $\quad n_u$ "big" $(> 2/\varepsilon)$

$$\hat{n}_u = 2/\varepsilon < n_u$$

$$0 \le \frac{1}{n_u} < \frac{1}{\hat{n}_u} = \frac{\varepsilon}{2}$$



$$\left| \frac{1}{n_u} - \frac{1}{\hat{n}_u} \right| \le \varepsilon/2 \quad \blacksquare$$

Algorithm to compute $\hat{n}_u$:

Do BFS starting from u

until:

- visit whole component of u

OR

- visit $2/\varepsilon$ distinct nodes

Output # visited nodes

$C =$ # conn comp

$n_v =$ # nodes in v's comp

Fact $\quad C = \sum_{u \in V} \frac{1}{n_u}$

def $\quad \hat{n}_u \equiv \min\{n_u, 2/\varepsilon\}$

$\hat{C} \equiv \sum_{u \in V} \frac{1}{\hat{n}_u}$

Lemma $\quad \forall u \quad \left|\frac{1}{\hat{n}_u} - \frac{1}{n_u}\right| \le \varepsilon/2$

Corr $\quad |C - \hat{C}| \le \frac{\varepsilon n}{2}$

Runtime:

$$O(\underbrace{d}_{\text{visit new node in BFS}} \cdot 2/\varepsilon) = O(d/\varepsilon)$$

Estimate sum $\sum_u \frac{1}{\hat{n}_u}$?

Algorithm estimate $\hat{C}$

$r \leftarrow b/\varepsilon^3$

Choose $U = \{u_1 \dots u_r\}$ random nodes

$\forall\ u_i \in U$

compute $\hat{n}_u$ via above

Output $\tilde{C} = n \cdot \underbrace{\frac{1}{r} \cdot \sum_{u \in U}}_{\text{estimate of ave value of } \frac{1}{\hat{n}_u}} \frac{1}{\hat{n}_u}$

$C =$ # conn comp

$n_v =$ # nodes in v's comp

Fact $\quad C = \sum_{u \in V} \frac{1}{n_u}$

def $\quad \hat{n}_u \equiv \min\{n_u, 2/\varepsilon\}$

$\hat{C} \equiv \sum_{u \in V} \frac{1}{\hat{n}_u}$

Lemma $\quad \forall u \quad \left|\frac{1}{\hat{n}_u} - \frac{1}{n_u}\right| \le \varepsilon/2$

Corr $\quad |C - \hat{C}| \le \frac{\varepsilon n}{2}$

$\hat{C} = n \cdot$ ave value of $\frac{1}{\hat{n}_u}$

runtime: $\quad O\left(\frac{1}{\varepsilon^3} \cdot \underbrace{\frac{d}{\varepsilon}}_{\text{calls to above alg}}\right) = O\left(\frac{d}{\varepsilon^4}\right)$

C = # conn comp

n_v = # nodes in v's comp

<u>Fact</u> $\quad C = \sum_{u \in V} \frac{1}{n_u}$

<u>def</u> $\quad \hat{n}_u \equiv \min\{n_u, 2/\varepsilon\}$

$$\hat{C} \equiv \sum_{u \in V} \frac{1}{\hat{n}_u}$$

<u>Lemma</u> $\quad \forall u \quad \left|\frac{1}{\hat{n}_u} - \frac{1}{n_u}\right| \le \varepsilon/2$

<u>Corr</u> $\quad |C - \hat{C}| \le \frac{\varepsilon n}{2}$

$$\tilde{c} = n \cdot \frac{1}{r} \cdot \sum_{u \in U} \frac{1}{\hat{n}_u}$$

Chernoff Bound:

Given $X_1 \cdots X_r$ iid $X_i \in [0,1]$

Let $S = \sum_{i=1}^{r} X_i$ $\quad p = E[X_i] = E[S]/r$

Then: $\Pr\left[\left|\frac{S}{r} - p\right| \geq \delta p\right] \leq e^{-\Omega(rp\delta^2)}$

__Thm__ $\Pr\left[|\tilde{c} - \hat{c}| \leq \frac{\varepsilon n}{2}\right] \geq 3/4$

__Corr__ $\Pr\left[|\tilde{c} - c| \leq \varepsilon n\right] \geq 3/4$

$$\left(\text{since } \begin{array}{l} |\hat{c} - c| \leq \frac{\varepsilon n}{2} \\ |\tilde{c} - \hat{c}| \leq \frac{\varepsilon n}{2} \end{array} \overset{\Delta \neq}{\Longrightarrow} |\tilde{c} - c| \leq \varepsilon n \right)$$

__Pf of thm__

$$X_i = \frac{1}{\hat{n}_{u_i}}$$

$$p = E\left[\frac{1}{\hat{n}_{u_i}}\right] = \frac{1}{n} \sum_{u \in V} \frac{1}{\hat{n}_{u_i}} = \frac{\hat{c}}{n}$$

bound $\frac{\hat{c}}{n}$:

$$\frac{\varepsilon}{2} \leq \frac{1}{\hat{n}_u} \leq 1$$

$$\frac{\varepsilon n}{2} \leq \hat{c} \leq n$$

$$\text{So } \frac{\varepsilon}{2} \leq \frac{\hat{c}}{n} \leq 1$$

$$\delta = \varepsilon/2$$

$$\frac{S}{r} = \frac{1}{r} \cdot \sum_{i=1}^{r} \frac{1}{\hat{n}_{u_i}}$$

So
$$\Pr\left[\left|\frac{\tilde{c}}{n} - \frac{\hat{c}}{n}\right| \geq \frac{\varepsilon}{2} \cdot \frac{\hat{c}}{n}\right] = \Pr\left[|\tilde{c} - \hat{c}| \geq \frac{\varepsilon}{2} \cdot \hat{c}\right]$$

$$\leq e^{-\left(\frac{b}{\cancel{\varepsilon^2}} \cdot \frac{\hat{c}}{n} \cdot \cancel{\varepsilon^2}/4\right)} = e^{-\frac{b}{\cancel{\varepsilon}} \cdot \frac{\cancel{\varepsilon}}{2} \cdot \frac{1}{4}} \leq 1/4$$

pick b big enough so that

∎