A first Example: Diameter of a point set

note input size n is m^2

Input m points described by distance matrix D

s.t. D_{ij} is distance from i,j

+ D is: 1) symmetric

2) satisfies $\Delta \neq$

ie. $D_{ij} \leq D_{ik} + D_{kj}$ $\forall i,j,k$

Output

let i,j be s.t. D_{ij} is max $\Leftarrow$ D_{ij} is "diameter"

Output k,l s.t. $D_{kl} \geq \frac{D_{ij}}{2}$ $\Leftarrow$ 2-approximation

Algorithm

- Pick k arbitrarily
- Pick l to maximize D_{kl}
- Output k,l

runtime: $O(m) = O(\sqrt{n})$

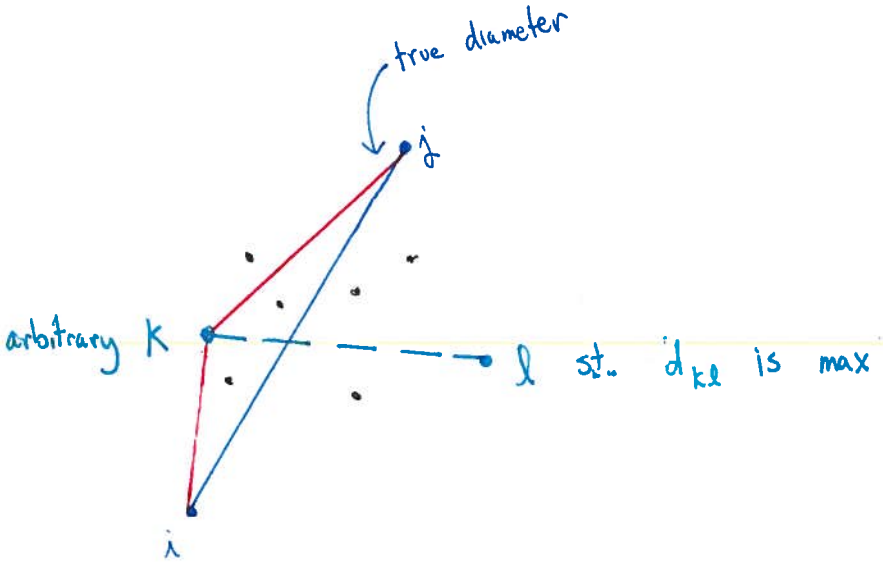
Why does it work?

$$D_{ij} \leq D_{ik} + D_{kj} \qquad \Delta\neq$$

$$= D_{ki} + D_{kj} \qquad \text{symmetry}$$

$$\leq D_{kl} + D_{kl} \qquad \text{choice of } l$$

$$\leq 2 D_{kl}$$



A second example:

How many connected components in G?

← NOTE: can solve via DFS/BFS in linear time

Input $G = (V, E)$, ε

max degree d

adjacency list representation

$|V| = n$

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

Output y s.t. if $C = \#$ conn comp

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

← "additive approx to w/in εn"

A different characterization of # conn components:

notation: $\forall v$ let $n_v \equiv \#$ nodes in v's conn comp.

observation: $\forall$ connected component $A \subseteq V$

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

So new characterization of # conn comp:

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

Why is this better?

computing $\frac{1}{n_u}$ needs $O(n)$ time? ↔ will estimate!

sum $O(n)$ terms?

Estimating $C = \sum_{u \in V} \frac{1}{n_u}$:

Two ideas:

1) estimate $\frac{1}{n_u}$ quickly ← additive estimate

2) estimate $\sum_u \frac{1}{n_u}$ via sampling bounds ← additive estimate but will use to get multiplicative error bound

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

idea n_u could be really big; so hard to compute exactly

but if n_u really big then $\frac{1}{n_u}$ is really small so can approx $\frac{1}{n_u}$ by 0 or $\frac{\varepsilon}{2}$

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

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

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

Corr $|C - \hat{C}| \leq \frac{\varepsilon n}{2}$ ← so if can compute $\hat{C}$ faster, it is useful!

Pf of lemma

if $n_u \leq 2/\varepsilon$ then $\hat{n}_u = n_u$ so $\left|\frac{1}{n_u} - \frac{1}{\hat{n}_u}\right| = 0$

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

$\Rightarrow \quad 0 \overset{\uparrow}{\leq} \frac{1}{n_u} < \frac{1}{\hat{n}_u} = \frac{\varepsilon}{2}$

↑ since $n_u > 0$

$\Rightarrow \quad \left|\frac{1}{\hat{n}_u} - \frac{1}{n_u}\right| < \varepsilon/2$ ∎

How long to compute $\hat{n}_u$?

Algorithm compute $\hat{n}_u$

Do BFS starting from u until
- visit whole component of u
- or visit $2/\varepsilon$ distinct nodes

Output # visited nodes

runtime

$O(d \cdot 1/\varepsilon)$

↑ time per step of BFS

How to estimate $\sum_u \frac{1}{\hat{n}_u}$?

Algorithm estimate $\hat{c}$

$r \leftarrow b/\varepsilon^3$

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

$\forall\, u \in U$
compute $\hat{n}_u$ via above algorithm

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

use average value estimate to estimate sum

runtime $O((d/\varepsilon) \cdot 1/\varepsilon^3) = O(d/\varepsilon^4)$

Why is it good?

Thm $\Pr\left[\,|\tilde{c}-\hat{c}| \le \varepsilon n/2\,\right] \ge 3/4$

Pf

Chernoff Bnd: $X_1 \cdots X_r$ iid $X_i \in [0,1]$

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

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

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

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

$$\delta = \frac{\varepsilon}{2}$$

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

so should pick r to be $\sim \frac{1}{p\delta^2}$ ($\Omega(\frac{1}{\varepsilon})$, $\Theta(\frac{1}{\varepsilon^2})$)

another reason why we need $\frac{1}{\hat{n}_u} \ge \frac{\varepsilon}{2}$

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

$$\le e^{-\left(\frac{b}{\varepsilon^2}\cdot\frac{\varepsilon}{n}\cdot\frac{\varepsilon^2}{4}\right)}$$ } want this to be ≥ 2 so that probability $\le e^{-2} \le \frac{1}{4}$

Since $\frac{\varepsilon}{2} \le \frac{1}{\hat{n}_u} \le 1$

summing over u: $\frac{\varepsilon n}{2} \le \hat{c} \le n$

so $\frac{\varepsilon}{2} \le \frac{\hat{c}}{n} \le 1$

$$\le e^{-\left(\frac{b}{\varepsilon}\cdot\frac{\varepsilon}{2}\cdot\frac{1}{4}\right)}$$

pick $b \ge 16$

Now we are done:

Corr. $\Pr[|c - \tilde{c}| \leq \varepsilon n] \geq 3/4$

Pf. $|c - \tilde{c}| \leq |c - \hat{c}| + |\hat{c} - \tilde{c}|$ by $\Delta \neq$

- $|c - \hat{c}|$: always $\leq \frac{\varepsilon n}{2}$ by corr
- $|\hat{c} - \tilde{c}|$: $\leq \frac{\varepsilon n}{2}$ by thm with prob $\geq 3/4$