Week 3: CS5234

Last Week:

1) Connected components
2) Approx-MST
3) Divisibility

Today:

1) Diameter
2) Avg. degree

Diameter

→ Graph $G=(V,E)$, undirected
n nodes, m edges

One solution: $\forall u \in V$ do: BFS(u)
⟹ $O(n(n+m))$ time

Note: different!
εn, not εnd

Defn: G is ε-far from G' if must add/delete $\lceil \varepsilon n \rceil$ edges to G to get G'.

Goal: Given G, D, decide:

1) if G has diameter $\leq D$, return TRUE
2) if G is ε-far from graph with diameter $4D+2$, return FALSE
3) otherwise, either

Surprising! Diameter is a global property

Goal: $O(1/\varepsilon^3)$ time

(n, D don't matter!!)

Trick: find a <u>local</u> property to test

<u>Defn</u>: Node $u \in V$ is $\boxed{(K, C)\text{- friendly}}$ if there are $\geq K$ nodes within distance C of u.

Key lemma: if $\geq (1-1/K)n$ nodes in G are (K, D)-friendly, then G is $2/K$-close to a graph with diameter $4D+2$.

Proof by algorithm: Label each node center
friend
fringe

Repeat:
1) Choose unlabelled node u.
2) Label u a <u>center</u>
3) Label nodes w s.t. $d(u,w) \leq D$ as <u>friend</u>.
4) Label unlabelled w st. $d(u,w) \leq 2D$ as fringe.
5) Delete u and all friends from G

Eventually, all nodes are labelled. How many centers?

For each center, 2 cases:

1) center u is not (K, D)-friendly $\Rightarrow$ $\leq n/K$ centers

2) center u is (K,D)-friendly.
a) u is not distance $2D$ from any previously chosen center (or it would have already been labelled fringe)
b) initially, u had $\geq K$ nbrs at distance $\leq D$ $\Rightarrow$ call them F (by defn of friendly)
c) Nodes in F were not previously labelled center/friend. (if $w \in F$ was friend, then $\exists$ center v st $d(w,v) \leq D$, so $d(u,v) \leq d(u,w) + d(w,v) \leq 2D$)
d) All nodes in F labelled friends, deleted
$\Rightarrow$ K nodes deleted

Conclusion: $\leq n/K$ non-friendly centers chosen

Total: $\leq 2n/K$ centers chosen

Proof (continued):

1) Choose a center u. Call it C.

2) Add $2n/K - 1$ edges connecting all centers to u.

Claim: diameter $\le 4D+2$

→ all nodes labelled ⇒ all within $2D$ of a center

→ $d(u,v) \le d(u, C_1) + d(C_1, C) + d(C, C_2) + d(C_2, v)$

ctr for u ctr for v

$\le 2D + 1 + 1 + 2D = 4D+2$

⇒ G was $2/K$ close to a graph (constructed) with diam $\le 4D+2$

Alg

Diam(G, D, ε):
 Repeat s times:
  Choose u at random.
  Do BFS to see if u is $(2/\varepsilon, D)$-friendly
  If not friendly, return FALSE.
 Return TRUE.

1) if diam $\le D$ and $n \ge 2/\varepsilon$, then all nodes friendly. ⇒ TRUE

2) if $n < 2/\varepsilon$, then can connect by $< 2/\varepsilon \le 2n/\varepsilon$ edges ⇒ special case

3) if not ε-close to $4D-2$, then $\ge \varepsilon n/2$ nodes not $(2/\varepsilon, D)$-friendly

$K = 2/\varepsilon$

$\Pr[u \text{ is not } (2/\varepsilon, D)\text{-friendly}] \ge (\varepsilon n/2)/n \ge \varepsilon/2$

$\Pr[\text{no sampled } u \text{ is not } (2/\varepsilon, D)\text{-friendly}] \le (1-\varepsilon/2)^s \le e^{-2} \le 1/3$

independent

$\boxed{\text{Choose } s = 4/\varepsilon}$

Conclusion: if ε-far from 4D+2, the FALSE w.p. $\geq 2/3$

Cost: $S = 4/\varepsilon$ iterations

Cost/iteration: BFS to find $2/\varepsilon$ nodes in distance D

$\Rightarrow$ cost $\leq (2/\varepsilon)^2 \leftarrow$ $\leq 2/\varepsilon$ nodes found

$\leq 2/\varepsilon$ nbrs of each node visited

unnecessarily

$$\boxed{\text{Total time: } (4/\varepsilon)(2/\varepsilon^2) = O(1/\varepsilon^3)}$$

Degree estimation

$G = (V, E)$, undirected, connected

Alg

```
AvgDeg(G, ε)
    min = n
    repeat K times
        total = 0
        repeat S times:
            Choose u ∈ V at random
            total = total + degree(u)
        if total/s < min then
            min = total/s
    return min
```

Sampling loop

Seems unlikely to work: $A = \{n, n, n, n, 0, 0, \dots, 0\}$

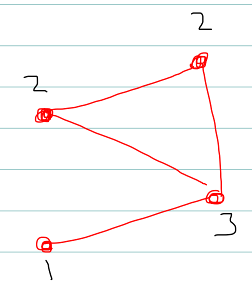
avg = 4

Sampling $\Rightarrow 0$

Sampling does not find average

Key insight: a graph is not an arbitrary array
it has structure   Cannot have degrees $[n, n, n, n, 1, 1, 1, 1, \dots, 1]$

Setup n nodes
m edges
avg degree $= \frac{1}{n} \sum_u \deg(u) = \frac{1}{n}[2m]$



avg $= 8/4 = 2$

Fix ε.

Goal: $(1/2 - \varepsilon)$-approx in
$O\left(\sqrt{n} / \varepsilon^{9/2}\right)$ time

Step 1: Show answer $\leq (1+\varepsilon) d$

Markov's Inequality!!

For one sampling loop: let $X_j = $ deg sample j

$X = \sum X_j = $ total, $\quad E[X_j] = \frac{1}{n} \sum_u \deg(u) = \frac{2m}{n} = d$

$E[X] = sd$

$$\Pr\left[X/s \geq (1+\varepsilon) d\right] = \Pr\left[X \geq (1+\varepsilon) E[X]\right] \leq \frac{E[X]}{(1+\varepsilon) E[X]}$$

$$\leq \left(\frac{1}{1+\varepsilon}\right) \leq 1 - \varepsilon/2$$

Recall: if $X < 1$: $\quad (1+X)^{-1} \geq 1 - X$
$\leq 1 - X/2$

Taylor: $(1+X)^{-1} = 1 - X - X^2/2 + \cdots$

Conclusion: $\Pr\left(X/s > (1+\varepsilon) d\right) \leq 1 - \varepsilon/2$

Too big? [cont]

$\Pr(\text{all } K \text{ sample loops} > (1+\varepsilon)d) \le (1-\varepsilon/2)^K$

↑ independent

$\boxed{\text{Choose } K = 8/\varepsilon}$

$\le (1-\varepsilon/2)^{8/\varepsilon}$

$\le e^{-4} \le 1/8$

Conclusion: w.p. $\ge 7/8$, answer is $\le (1+\varepsilon)d$

What about showing answer is not $\le (1-\varepsilon)d$??

How do you find heavy nodes?

If miss heavy nodes, then answer is too small!

Key problem: heavy (large deg) values have big impact on final result but are hard to find.

Solution: "Ignore" heavy nodes

Defn

$H = \sqrt{\varepsilon n}$ nodes with largest deg

L = all other nodes

M = max degree in L.

Question: if we only sample L, is that good enough?

(How to sample only L?? Postponed..)

Defn $\deg(H) = \sum_{u \in H} \deg(u)$

$\deg(L) = \sum_{u \in L} \deg(u)$

$$\deg(G) = \sum_{u} \deg(u) = 2m = \deg(H) + \deg(L)$$

$\boxed{\text{Claim} \quad \deg(L) \geq m(1-\varepsilon)}$

$$\deg(H) \leq 2\underbrace{\binom{\sqrt{\varepsilon n}}{2}}_{\text{edges in } H} + \underbrace{m}_{\text{edges between } L \text{ and } H} \leq 2\left[\varepsilon n/2\right] + m \leq \varepsilon n + m \overset{n \leq m}{\leq} \varepsilon m + m \leq (1+\varepsilon)m$$

$$\deg(L) = \deg(G) - \deg(H) \leq 2m - m(1+\varepsilon) = m - \varepsilon m = m(1-\varepsilon)$$

Conclusion: if we estimate avg deg L, we are within $\left(\frac{1}{2}-\varepsilon\right)$ of avg deg G $\left[\text{ie, } \frac{m}{n}\left(1-\frac{\varepsilon}{2}\right) \text{ is } \geq \frac{2m}{n}\left(\frac{1}{2}-\varepsilon\right)\right]$

How to estimate $\deg(L)$?

(Don't know which are in L.)

Trick: define new random vars.

$y_i = \min(X_i, M)$ ← max value in L

if $X_j \in H$, $y_j = M$

if $X_j \in L$, $y_j = X_j$

$X_j \geq y_j$

New goal: bound Y

$X \geq Y = \sum y_j$

$$E[y_j] = \frac{1}{n} \sum_j \min(deg(j), M)$$

$$= \frac{1}{n} \sum_{u \in H} \min(deg(u), M) + \frac{1}{n} \sum_{u \in L} \min(deg(u), M)$$

$$\geq \frac{1}{n} \sum_{u \in L} \min(deg(u), M)$$

$$\geq \frac{1}{n} \sum_{u \in L} deg(u)$$

$$\geq \frac{1}{n} deg(L)$$

$$\geq \frac{m}{n}(1-\varepsilon)$$

$$\geq (d/2)(1-\varepsilon)$$

$$E[y_j] \geq \frac{1}{n} \sum_{u \in H} \min(deg(u), M)$$

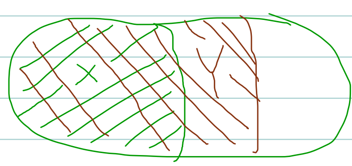
$$\geq \frac{1}{n} \cdot |H| \cdot M$$

$$\geq \frac{1}{n}(\sqrt{n\varepsilon})M \geq \frac{M}{\sqrt{n/\varepsilon}}$$

Show: $\Pr[Y \leq (1-\delta)E[Y]]$ is small, ie, $\leq 1/8K$

$$\Rightarrow \Pr[X \leq S(1-\delta)(d/2)(1-\varepsilon)] \leq \Pr[Y \leq S(1-\delta)(\tfrac{d}{2})(1-\varepsilon)]$$

$\uparrow$ $Y \leq X$



$$\leq \Pr[Y \leq (1-\delta)E[Y]] \leq 1/8K$$

$\uparrow$ $E[Y] \geq (d/2)(1-\varepsilon)S$

$$\boxed{\Pr[X/S \leq (1-\delta)(d/2)(1-\varepsilon)] \leq 1/8K}$$

Break: Chernoff Bounds

$z_1, z_2, \ldots, z_n$ are r.v. $\in [0,1]$

$Z = \sum z_j$ $\quad \boxed{0 \le \varepsilon \le 1}$

$E[Z] = M$

1) $\Pr(Z \ge (1+\varepsilon)M) \le e^{-\varepsilon^2 M/3}$

2) $\Pr(Z \le (1-\varepsilon)M) \le e^{-\varepsilon^2 M/2}$

3) $\Pr(|Z - M| \le \varepsilon M) \le 2e^{-\varepsilon^2 M/3}$

Wait!! What if $z_i \in [1, K]$?

1) $\Pr(Z \ge (1+\varepsilon)M) \le e^{-\varepsilon^2 M/3K}$

2) $\Pr(Z \le (1-\varepsilon)M) \le e^{-\varepsilon^2 M/2K}$

3) $\Pr(|Z - M| \le \varepsilon M) \le 2e^{-\varepsilon^2 M/3K}$

$$\Pr[Y \le (1-\delta)E[Y]] \le e^{-E[Y]\delta^2/2M}$$

$$\le e^{-\frac{sM}{\sqrt{n/\varepsilon}} \cdot \frac{\delta^2}{2M}} \quad \leftarrow E[Y] \ge sM/\sqrt{n/\varepsilon}$$

Choose $s = \dfrac{8\sqrt{n/\varepsilon}\,\ln K}{\delta^2}$

$$\le e^{-4\ln K} \le 1/8K$$

$\Rightarrow \Pr\left[\frac{X}{s} \le (1-\delta)(1-\varepsilon)(d/2)\right] \le 1/8K$

$\Pr[\text{all iterations good}] \geq \left(1-\frac{1}{8K}\right)^K \geq e^{-1/6} \geq 7/8$

$\Pr(\text{some iteration bad OR answer too big}) \leq 1/8 + 1/8 \leq 1/4$

$\hookrightarrow$ correct w.p. $\geq 3/4$

Set $\delta = \varepsilon$: correct = 1) answer $\leq (1+\varepsilon)d$

2) answer $\geq (1-\delta)(1-\varepsilon)(d/2)$

$\geq (1-\varepsilon)^2 (d/2)$ $\quad (1-\varepsilon)^2 = (1-2\varepsilon+\varepsilon^2) \geq 1-2\varepsilon$

$\geq (1-2\varepsilon)(d/2)$

$\geq d[1/2 - \varepsilon]$

$$\boxed{d(1/2 - \varepsilon) \leq \text{answer} \leq d(1+\varepsilon)}$$

Time: $O(KS)$

$K = \frac{8}{\varepsilon}$ $\qquad S = \frac{8\sqrt{n/\varepsilon}\,\ln(8/\varepsilon)}{\varepsilon^2}$

$O(KS) \leq \tilde{O}\left(\sqrt{n/\varepsilon} \cdot \frac{1}{\varepsilon^4}\right) = \tilde{O}\left(\sqrt{n}\,\varepsilon^{-9/2}\right)$