Lecture 1

Topics:

- Intro to course ← see slides
- Approximating diameter of point set
- Sublinear time approximation of average degree

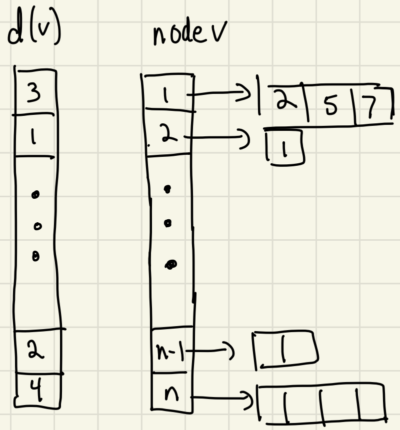
Estimating the average degree of a graph

def Average degree $\bar{d} = \frac{\sum_{u \in V} \deg(u)}{n} = \frac{2m}{n}$

Assume: G simple (no parallel edges, self-loops)

$\Omega(n)$ edges (not "ultra-sparse")

Representation via adj list + degrees:



- degree queries: on v return $\deg(v)$
- neighbor queries: on (v, j) return j^{th} nbr of v

Estimating Average Degree

Given $G = (V, E)$

$\varepsilon \in (0,1)$ approximation parameter

$\delta \in (0,1)$ confidence ← lets assume $\delta = 1/4$

Output $\tilde{d}$ s.t. $\Pr[|\tilde{d} - \bar{d}| \le \varepsilon \bar{d}] \ge 1 - \delta$

where $\bar{d} = \frac{m}{n}$ (average degree)

Naive sampling:

Pick $O(??)$ sample nodes $v_1 \dots v_s$

output ave degree of sample:

$$\frac{1}{s} \sum_i \deg(v_i)$$

Straightforward Chernoff/Hoeffding needs $\Omega(n)$ samples

lower bound?

deg(1)	deg(2)	...				deg(n)
0	0	0	n-1	0	0	0

need $\Omega(n)$ samples to find "needle in haystack"

not a possible degree sequence!!

n-1	1	1	1	1	1	1

<u>is</u> possible

Some lower bounds:

"ultrasparse" case:

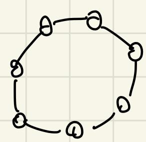
0 edges vs. 1 edge

need $\Omega(n)$ queries to distinguish

$\Rightarrow$ multiplicative approx needs $\Omega(n)$

ave deg ≥ 2:

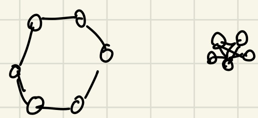
n - cycle $\bar{d} = 2$



vs.

$n - c\cdot\sqrt{n}$ cycle $\bar{d} \approx 2 + c^2$

$+ \; c\sqrt{n}$ - clique

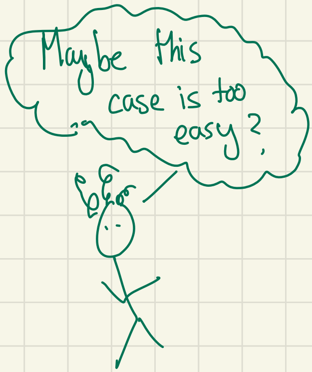


need $\Omega(n^{1/2})$ queries to find clique node

Warm up: regular graphs

Assume each node has degree Δ

Algorithm: output Δ



Better warmup: almost regular graphs

Assume each node has degree in $[\Delta, 10\Delta]$ (so $\Delta \le \bar{d} \le 10\Delta$)

Algorithm:

$K \leftarrow \frac{50}{\varepsilon^2} \ln(2/\delta)$

For $i \leftarrow 1$ to K do

- pick $v_i \in_u V$
- $X_i \leftarrow \deg(v_i)$

Output $\tilde{d} \leftarrow \frac{1}{K} \sum_{i=1}^{K} X_i$

notation: $x \in_u D$ means pick x uniformly from set D

Runtime: $O(\frac{1}{\varepsilon^2} \ln \frac{1}{\delta})$ ← Wow!! no dependence Δ or n

Behavior: What is the expected value of $\tilde{d}$?

Claim $E[\tilde{d}] = \bar{d}$ (True even without assumption)

Pf

$$E[\tilde{d}] \underset{\text{lin of exp}}{=} \frac{1}{k} \sum_{i=1}^{k} E[X_i] \underset{\text{iid}}{=} E[X_1]$$

$$= \sum_{v \in V} \frac{1}{n} \deg(v) = \frac{\sum \deg(v)}{n} = \bar{d} \quad \blacksquare$$

How likely is $\tilde{d}$ (actual output) to be far from its expected value?

Claim $\Pr[\,|\bar{d} - \tilde{d}| \leq \varepsilon \bar{d}\,] \geq 1 - \delta$

Pf

Will use following version of Chernoff Bnd:

Thm let $Y_1 \ldots Y_k$ be independent random variables s.t. $Y_i \in [0,1]$ & $Y = \sum_{i=1}^{k} Y_i$. For $b \geq 1$

$$\Pr[\,|Y - E[Y]| > b\,] \leq 2 \cdot \exp(-2b^2/k)$$

so can't use Chernoff

Note: X_i's are <u>not</u> in $[0,1]$ but are in $[\Delta, 10\Delta]$

Normalize! let $Z_i \leftarrow \frac{X_i}{10\Delta}$ then $Z_i \in [0,1]$ ✱✱✱

let $Z \leftarrow \sum_{i=1}^{k} Z_i$ so $\tilde{d} = \frac{1}{k}\sum X_i = \frac{10\Delta}{k}\cdot Z$

$+ \ \bar{d} = E[\tilde{d}] = \frac{10\Delta}{k} E[Z]$

We have: $|\tilde{d} - \underset{\substack{\uparrow \\ E[\tilde{d}]}}{\bar{d}}| \geq \varepsilon \bar{d} \iff \left|\frac{10\Delta}{k} Z - \frac{10\Delta}{k} E[Z]\right| \geq \varepsilon \bar{d}$

$\iff \underbrace{|Z - E[Z]| \geq \frac{k}{10\cdot\Delta}\cdot \varepsilon \bar{d}}_{\text{we want to make sure this isn't likely}}$

Use Chernoff on Z's

with $b = \frac{k}{10\Delta}\varepsilon \bar{d}$

$$\Pr\left[|Z - E[Z]| \geq \frac{k}{10\Delta}\varepsilon \bar{d}\right] \leq 2\cdot e^{-\left(\frac{2k^2\varepsilon^2\bar{d}^2}{100\Delta^2\cdot k}\right)}$$

$$= 2e^{-\frac{1}{50}\cdot\frac{k\varepsilon^2\bar{d}^2}{\Delta^2}}$$ ✱✱✱

$\bar{d} \geq \Delta$ by assumption on all degrees $\geq \Delta$

$$\leq 2e^{-\frac{k\varepsilon^2}{50}}$$

$k = \frac{50}{\varepsilon^2}\ln(2/\delta)$

$$= 2e^{-\frac{(\cancel{50}/\cancel{\varepsilon^2})(\ln 2/\delta)\cdot\cancel{\varepsilon^2}}{\cancel{50}}} = \delta \quad \blacksquare$$

Moral: it is a lot easier to estimate averages when the variables are all within a constant factor of each other.

★★★ marks where this assumption was used.

General Case: "Order" edges to control outdegree

Our plan:

define total order "$\prec$" on nodes:

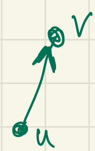
$\nwarrow$ assume distinct ID's

def. $u \prec v$ if

- $\deg(u) < \deg(v)$

or
- $\deg(u) = \deg(v)$
  $\&$ $ID(u) < ID(v)$

$\deg^+(u)$ = # nbrs of u s.t. $u \prec v$



Orient edges from small to large, $\deg^+(u)$ counts "out-edges"

Observation $\sum_{u \in V} \deg^+(u) = m = \frac{n}{2} \cdot \bar{d}$

(since each edge only counted once instead of twice as in $\sum_u \deg(u)$)

idea estimate $\operatorname{average}_u \left(\deg^+(u) \right) \equiv \frac{1}{2} \cdot \bar{d}$

problem? we can query $\deg(u)$ not $\deg^+(u)$

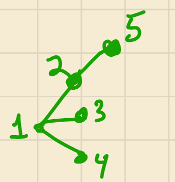
benefit:

Lemma $\forall v \in V \quad \deg^+(v) \le \sqrt{2m}$

Proof

Consider order of v_i's by $\prec$:

example:



3, 4, 5, 2, 1

low degree

high degree

v_i's ordered by α : $\overbrace{v_1 \quad v_2 \quad \cdots \quad v_k \quad \cdots}^{V \setminus H} \quad \overbrace{v_{n-\sqrt{2m}+1} \quad \cdots \quad v_n}^{H}$

edges can only point to right

define $H \subseteq V$ to be $\sqrt{2m}$ nodes with highest rank (degree) w.r.t. α

heavy nodes:

$\forall v \in H$, $\deg^+(v) \leq \sqrt{2m}$ since edges "leaving" v go to bigger nodes (which must also be in H)

light nodes:

$\forall v \in V \setminus H$, $\deg^+(v) \leq \deg(v) \leq \sqrt{2m}$:

Why?

if not, $\deg(v) > \sqrt{2m}$ $\leftarrow$ assume for contradiction

but all w in H have

$$\deg(w) \geq \deg(v) > \sqrt{2m}$$

so total degree $\sum_v d(v)$

$$> \underbrace{|H| \cdot \sqrt{2m}}_{\text{contribution from } H} + \underbrace{\text{something positive}}_{\text{contribution from } V \setminus H}$$

$$> \sqrt{2m} \cdot \sqrt{2m} = 2 \cdot m$$

but sum of degrees $= 2 \cdot m$

$$\underbrace{\rightarrow\leftarrow}_{\text{symbol for "contradiction"}} \qquad \blacksquare$$

New Algorithm:

$K \leftarrow \frac{16}{\varepsilon^2}\sqrt{n}$

for $i = 1$ to K

neighbor of v

pick $v_i \in_r V$ (1)

pick $u_i \in_r N(v_i)$ (2)

if $v_i < u_i$ then $X_i \leftarrow 2\deg(v_i)$

else $X_i \leftarrow 0$

return $\tilde{d} = \frac{1}{K}\sum_{i=1}^{K} X_i$

Question to think about: why the "2"?

Claim $E[X_i] = \bar{d}$

Pf

$$E[X_i] = \sum_{v \in V} \Pr[v \text{ chosen in } (1)] \cdot E[X_i \mid v \text{ chosen in } (1)]$$

$$= \sum_{v \in V} \frac{1}{n} \cdot E[X_i \mid v \text{ chosen in } (1)]$$

$$= \frac{1}{n} \sum_{v \in V} \sum_{u \in N(v)} \Pr[u \text{ chosen in } (2) \mid v \text{ chosen in } (1)]$$

$$\times E[X_i \mid u \text{ chosen in } (2) \text{ \& } \underbrace{v \text{ chosen in } (1)]}_{\substack{\text{if } v \propto u \\ \text{then} \\ X_i = 2\deg(v) \\ \text{else } X_i = 0}}$$

$$= \frac{1}{n} \cdot \sum_{v \in V} \sum_{\substack{u \in N(v) \\ \text{\& } v \propto u}} \frac{1}{\deg(v)} \cdot 2 \cdot \deg(v)$$

$$= \frac{2}{n} \sum_{v \in V} \sum_{\substack{u \in N(v) \\ \text{\& } v \propto u}} 1$$

$$= \frac{2}{n} \cdot \sum_{v \in V} \deg^+(v) = \frac{2m}{n} = \bar{d}$$

∎

But how many samples do we need to assure that we are close to expectation? Here is where we use the lemma!

$\boxed{\text{def. } Var[X] \equiv E[X^2] - E[X]^2}$

Claim $Var[X_i] \leq 4\sqrt{2m}\,\bar{d}$

Pf $Var[X_i] \equiv E[X_i^2] - E[X_i]^2 \leq E[X_i^2]$

as above

$$= \frac{1}{n} \sum_{v \in V} \sum_{\substack{u \in N(v) \\ v \prec u}} \frac{1}{deg(v)} \underbrace{(2\, deg(v))^2}_{X_i^2}$$

$$= \frac{4}{n} \sum_{v \in V} \underbrace{deg^+(v)}_{\leq \sqrt{2m}} \cdot deg(v)$$

key insight

$$\leq \frac{4}{n} \cdot \sqrt{2m} \sum_{v \in V} deg(v)$$

$$\leq 4 \cdot \sqrt{2m} \cdot \bar{d}$$

■

2 useful facts about variance:

- $\boxed{\text{Lemma let } Y = \frac{1}{k}\sum_{i=1}^{k} X_i \text{ where } X_i\text{'s are iid then } Var[Y] = \frac{1}{k} Var[X]}$

  so can reduce variance by sampling more! averaging

  important but pairwise independence is good enough

- $\boxed{\text{Chebyshev's } \ddagger\text{: } Pr[\,|X - E[X]| \geq b\,] \leq \frac{Var[X]}{b^2}}$

$\tilde{d}$ = output
$\bar{d}$ = average

Lemma $\Pr\left[\,|\tilde{d}-\bar{d}| \le \varepsilon \bar{d}\,\right] \ge 3/4$

Pf

$E[\tilde{d}] = \bar{d}$ by lin of expectation

$\mathrm{Var}[\tilde{d}] \le \frac{4\cdot\sqrt{2m}}{k}\cdot\bar{d}$

since $\bar{d} = E[\tilde{d}]$

$$\Pr\left[\,|\tilde{d}-\bar{d}| \ge \varepsilon\bar{d}\,\right] = \Pr\left[\,|\tilde{d}-E[\tilde{d}]| \ge \varepsilon\bar{d}\,\right]$$

$$\le \frac{\mathrm{Var}[\tilde{d}]}{(\varepsilon\bar{d})^2}$$

$$\le \frac{\frac{4\sqrt{2m}}{k}\cdot\bar{d}}{\varepsilon^2\bar{d}^2} = \frac{4\sqrt{2m}}{\varepsilon^2\cdot\underbrace{\bar{d}}_{=\frac{2m}{n}}\cdot k}$$

$$= \frac{4\sqrt{2m}\cdot n}{\varepsilon^2\cdot 2m\cdot k} = \frac{4n}{\varepsilon^2\cdot\sqrt{2m}\cdot k}$$

pick $k = \frac{16}{\varepsilon^2}\sqrt{n}$

$$= \frac{\sqrt{n}}{4\cdot\sqrt{2m}}$$

$$\le \frac{1}{4}$$

since $\sqrt{\frac{n}{2m}} = \sqrt{\frac{1}{\bar{d}}} \le 1$ since we assumed $\bar{d} \ge 1$

$\Rightarrow$ good estimate with prob $\ge 3/4$

How do we improve probability of success?

See HW 0!