Lecture 1

Topics:

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

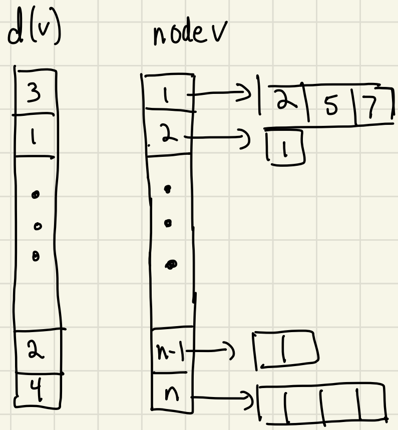
Estimating the average degree of a graph

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

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 $v_1 .. v_s$

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:

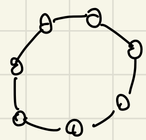
0 edges vs. 1 edge

need $\Omega(n)$ queries to distinguish

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

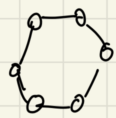
ave deg ≥ 2:

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



vs.

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

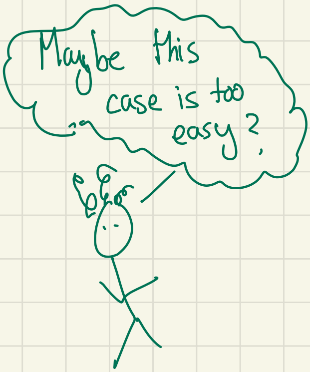


need $\Omega(n^{1/2})$ queries to find clique node

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: $O\left(\frac{1}{\varepsilon^2} \ln \frac{1}{\delta}\right)$ ← 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

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

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

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

Claim $\Pr\left[\,|\bar{d} - \tilde{d}| \le \varepsilon \bar{d}\,\right] \ge 1 - \delta$

Pf

Will use following version of Chernoff Bnd:

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

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

so can't use Chernoff

Note: X_i's are not 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 $\geq \Delta$

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

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

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

<u>Moral</u>: 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.