Lecture 2

Topics:

- Sublinear time approximation of average degree
- Estimate number of connected components

Last time

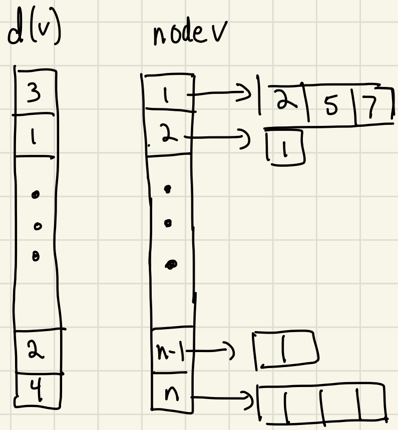
Estimating the average degree of a graph

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 $\leftarrow$ lets assume $\delta = 1/4$

Output $\tilde{d}$ s.t. $\Pr\left[|\tilde{d} - \bar{d}| \leq \varepsilon \bar{d} \right] \geq 1 - \delta$

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

Last time we saw that "naive sampling" ie.

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

output ave degree of sample:

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

Does not work so well, although we did prove that if <u>all</u> $\deg(v)$ are in $[\Delta, 10\Delta]$ then constantly many samples are sufficient. for the naive sampling algorithm

In general, we saw a handwavy argument that

- $\Omega(n)$ time is need to give a multiplicative estimate for average degree (this used "ultrasparse" graphs)
- $\Omega(\sqrt{n})$ time is needed for estimating average degree, even when the average degree is >1.

Today we will see the general case, & a different algorithm

General Case: "Order" edges to control outdegree

Our plan:

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

assume distinct ID's

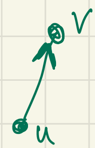
def. $u \prec v$ if

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

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"

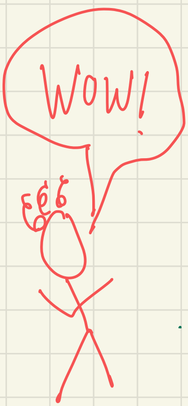
Observation $\sum_{u \in V} \deg^+(u) = m = \frac{n}{2} \cdot \bar{d}$

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

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

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

benefit:



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

Proof

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

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

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, $\deg(v) > \sqrt{2m}$ ← 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 $= 2\cdot m$

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

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 $v_i \prec u_i$ 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

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

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

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

$$\times E[X_i \mid u \text{ chosen in } (2) \text{ \& } v \text{ chosen in } (1)]$$

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

if $v \prec u$ then $X_i = 2\deg(v)$ else $X_i = 0$

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

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

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

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

Pf $Var[X_i] \equiv E[X_i^2] - E[X_i]^2 \le E[X_i^2]$ as above

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

$$= \frac{4}{n} \sum_{v \in V} \underbrace{deg^+(v)}_{\le \sqrt{2m}} \cdot deg(v) \longleftarrow \text{key insight}$$

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

$$\le 4 \cdot \sqrt{2m} \cdot \bar{d} \qquad \blacksquare$$

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 ← averaging more!

  important but pairwise independence is good enough

- $\boxed{\text{Chebyshev's } \neq \text{: } Pr[\,|X - E[X]| \ge b\,] \le \frac{Var[X]}{b^2}}$

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

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

Pf

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

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

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

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

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

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

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

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!