Goal: Estimate # of edges

+ Sample edges almost u.a.r

in $\tilde{O}\left(\frac{n}{\sqrt{m}}\right)$ queries

n - # of vertices, m - # of edges

Goal: Estimate # of edges

+ Sample edges almost u.a.r

in $\tilde{O}\left(\frac{n}{\sqrt{m}}\right)$ queries

n - # of vertices, m - # of edges

(You have already seen an algorithm for approximating m up to multiplicative factor in $\tilde{O}\left(\frac{n}{\sqrt{m}}\right)$ queries by Goldreich & Ron)

Query Model: Adjacency list :

Query Model: Adjacency list:

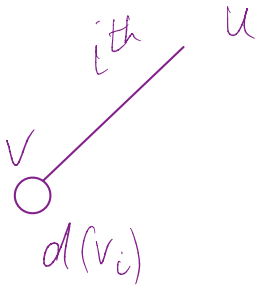
- The vertices are labeled arbitrarily $1..n$, and Alg knows n.

Query Model: Adjacency list:

- The vertices are labeled arbitrarily $1..n$, and Alg knows n.
- Degree queries: $\deg(v)$ returns d_v.

Query Model: Adjacency list:

- The vertices are labeled arbitrarily $1..n$, and Alg knows n.
- Degree queries: $deg(v)$ returns d_v.
- Neighbor queries: $nbr(v,i)$ returns the i^{th} neighbor of v. if one exists. O.w. returns $\perp$.



General approach in sublinear estimation :

General approach in sublinear estimation:

Define some unbiased estimator, s.t.:

General approach in sublinear estimation:

Define some unbiased estimator, s.t.:

1. $EX[X] = f(m)$ where from $f(m)$ we can deduce m.

   E.g.: $EX[X] = m$, or $EX[X] = \frac{1}{n} \cdot m$

   $\in \left[\frac{m}{2n}, \frac{m}{n}\right]$

2. The (1) maximum value (weaker)
   or (2) variance of X (stronger) are bounded

   $10 \cdot \frac{(H-m)}{n}$

General approach in sublinear estimation:

Define some unbiased estimator, s.t.:

1. $EX[X] = f(m)$ where from $f(m)$ we can deduce m.

   E.g.: $EX[X] = m$, or $EX[X] = \frac{1}{n} \cdot m$

2. The (1) maximum value (weaker)
   or (2) variance of X (stronger) are bounded

Then, by Chernoff inequality (for (1)) and Chebyshev's inequality (for (2))
we can bound the number of sufficient samples. Proof follows.

Chernoff inequality:

Let $X_1, \ldots, X_k$ be independent random variables, s.t. $\forall i$ X_i in $[0, B]$ and let $X = \sum_{i=1}^{k} X_i$. Then:

$$\Pr\left[|X - EX[X]| > \varepsilon \cdot EX[X]\right] < 2 \cdot \exp\left(-\frac{\varepsilon^2 \cdot EX[X]}{3 \cdot B}\right)$$

$\Rightarrow$ to get $\Pr[\text{deviation}] < \delta$, should set

k to $\frac{3B \cdot \ln(2/\delta)}{\varepsilon^2 \cdot EX[X_i]}$ $\Leftrightarrow$ $\frac{\text{Max}}{EX[X_i]}$

Chebyshev's inequality:

- $Pr[\,|X - EX[X]| > \varepsilon \cdot EX[X]\,] < \frac{Var[X]}{\varepsilon^2 \cdot EX[X]^2}$
- For $X = \sum_{i=1}^{k} X_i$, s.t. $\forall i,j \quad Var[X_i] = Var[X_j]$,

$$\Rightarrow Pr[\dots] < \frac{\sum Var[X_i]}{\varepsilon^2 \cdot k^2 \cdot EX[X_i]^2} = \frac{k \cdot Var[X_i]}{\varepsilon^2 \cdot k^2 \cdot EX[X_i]^2} = \frac{Var[X_i]}{\varepsilon^2 \cdot k \cdot EX[X_i]^2}$$

So to get $Pr[\dots] < \delta$, we need to set

$$k = \frac{Var[X_i]}{\varepsilon^2 \cdot EX[X_i]^2 \cdot \delta}$$

Chebyshev's inequality:

- $Pr[\,|X - EX[X]| > \varepsilon \cdot EX[X]\,] < \frac{Var[X]}{\varepsilon^2 \cdot EX[X]^2}$
- For $X = \sum_{i=1}^{k} X_i$, s.t. $\forall i,j \quad Var[X_i] = Var[X_j]$,

$$\Rightarrow Pr[\dots] < \frac{\sum Var[X_i]}{\varepsilon^2 \cdot k^2 \cdot EX[X_i]} = \frac{k \cdot Var[X_i]}{\varepsilon^2 \cdot k^2 \cdot EX[X_i]} = \frac{Var[X_i]}{\varepsilon^2 \cdot k \cdot EX[X_i]}$$

So to get $Pr[\dots] < \delta$, we need to set

$$k = \frac{Var[X_i]}{\varepsilon^2 \, EX[X_i]^2 \, \delta}$$

Back to estimating # edges:

First, naive, attempt:

For every $v \in V$, let

$$x_v = d(v)$$

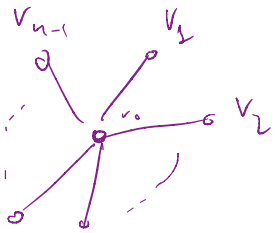
Then, $\mathbb{E}_{v \in V}[x_v] = \sum_{v \in V} \frac{1}{n} \cdot d(v) = \frac{2m}{n}$

$\Rightarrow$ unbiased estimator!

But, variance might be too high.

But, variance might be too high.

$\exists$ graphs s.t. $\mathrm{Var}_v[X_v] = n \cdot EX[X_v]$

E.g.: for 

$EX_v[X_v] = \frac{1}{n}\sum d(v) = \Theta(1)$

$\mathrm{Var}[X_v] = \frac{1}{n}\sum X_v^2 = \frac{1}{n}\sum d_v^2 = \frac{1}{n}\cdot(n-1)^2 \frac{1}{n} = \Omega(n)$

So, by Chebyshev's will have to set k

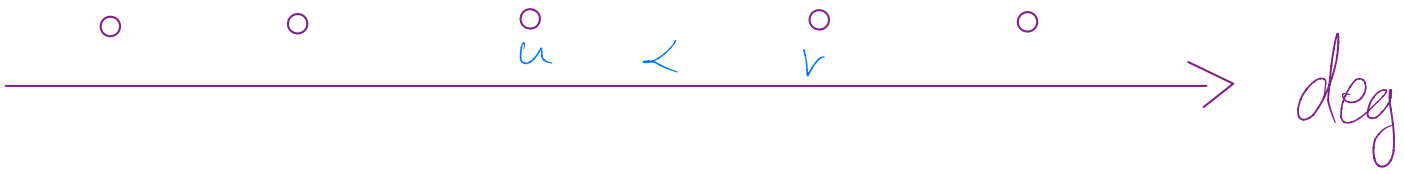
$k = \frac{\mathrm{Var}[X_v]}{\varepsilon^2 \cdot EX[X_v]^2} = \Omega\left(\frac{n}{\varepsilon^2 \cdot 1}\right) = \Omega(n)$ ☹

Second attempt:

1. Order all vertices according to degree

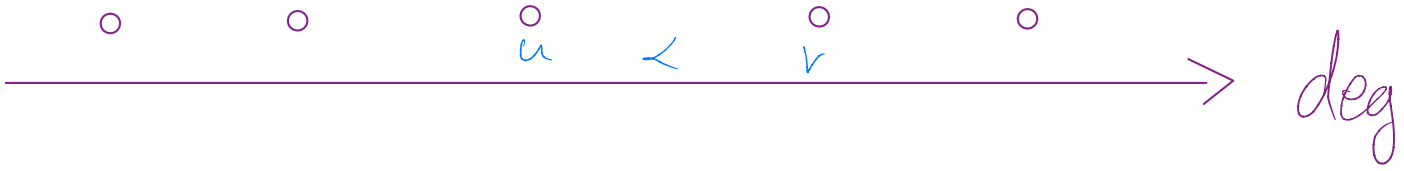
Second attempt:

1. Order all vertices according to degree



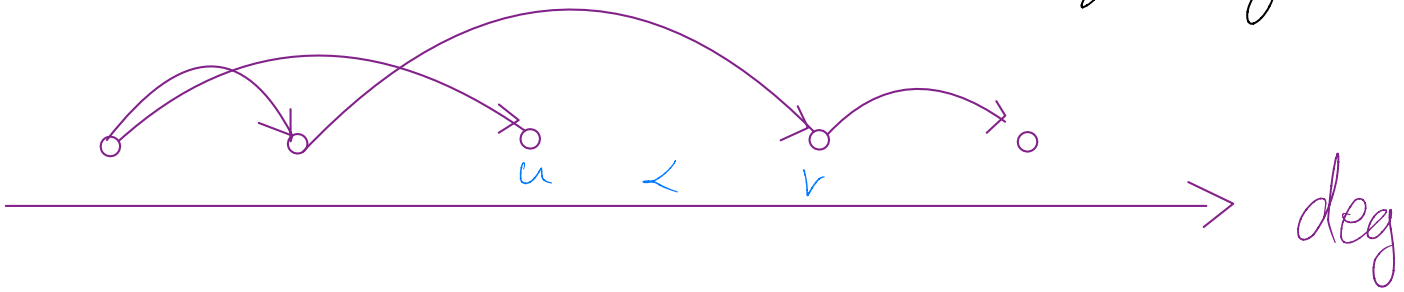
Second attempt:

1. Order all vertices according to degree
2. Orient edges from low deg to high deg vertices



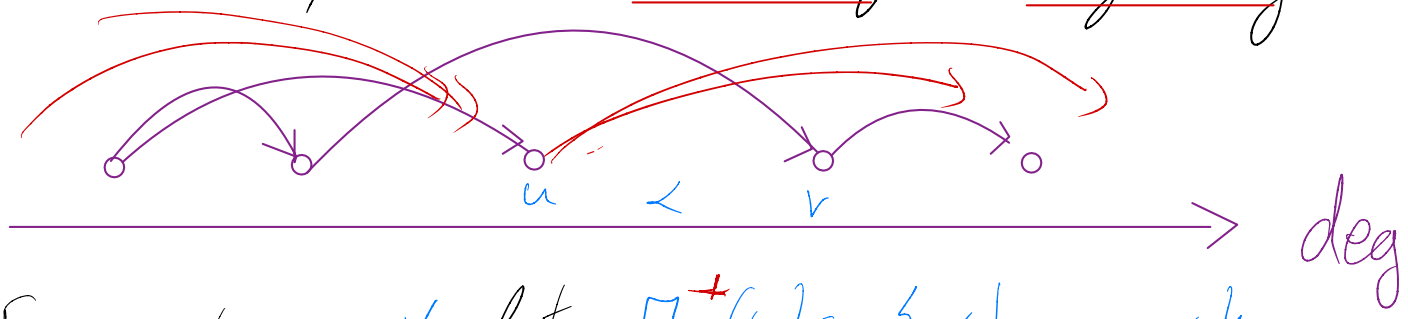
Second attempt:

1. Order all vertices according to degree
2. Orient edges from low deg to high deg vertices



Second attempt:

1. Order all vertices according to degree
2. Orient edges from low deg to high deg vertices



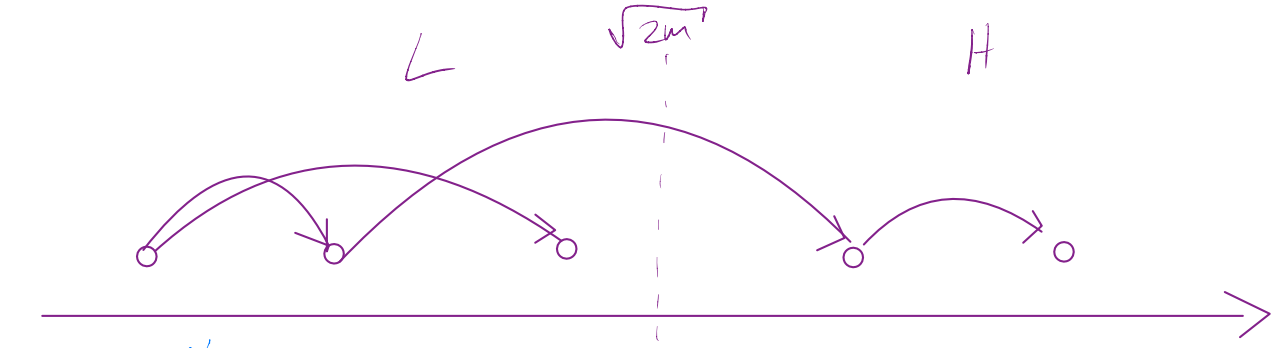
3. For each $u \in V$ let $\Gamma^+(u) = \{v \mid u < v\}$

and let $d^+(u) = |\Gamma^+(u)|$.

Observe that $\sum_{v \in V} d^+(v) = m$

(since every edge is counted from exactly one endpoint)

Claim: $\forall v,\ d^+(v) \le \sqrt{2m}$



$\forall v \in L,\ d^+(v) \le d(v) \le \sqrt{2m}$

$\forall v \in H,\ d^+(v) \le |\Gamma(v) \cap H| \le |H| \le \frac{2m}{\sqrt{2m}} = \sqrt{2m}$

2nd (wishful) attempt:

$\sum_{v \in V} d^+(v) = m$

Let $\chi_v = d^+(v)$

Then $\mathbb{E}_{v \in V}[\chi_v] = \sum_{v \in V} \frac{1}{n} d^+(v) = \frac{m}{n}$

$\Rightarrow$ unbiased estimator!

Also, by previous analysis,

$\max_{v \in V}\{\chi_v\} \leq \sqrt{2m}$ $\Rightarrow$ bounded max. value!

Hence, by Chebyshev's / Chernoff,

Sufficient to take $k = \frac{\max}{\mathbb{E}X} \cdot \frac{3\log(2/\delta)}{\varepsilon^2} = \tilde{O}\left(\frac{\sqrt{m}}{m/n}\right) = \tilde{O}\left(\frac{n}{\sqrt{m}}\right)$

Problem: We don't have access to $d^+(v)$ queries.

Solution: We'll set a random variable so that in Expectation, $E[X_u] = d^+(u)$

Problem: We don't have access to $d^+(v)$ queries.

Solution: We'll set a random variable so that in Expectation, $E[X_u] = d^+(u)$

Given $u \in V$:

1. Query $d(u)$ ✓
2. Query a uniform nbr of u, denoted v $d(v)$
3. Let

$$X_u = \begin{cases} d(u) & \text{if } u \prec v \\ 0 & \text{o.w} \end{cases}$$

$$E_v[X_u | u] = \frac{d^-(u)}{d(u)} \cdot d(u) = d^-(u)$$

Problem: We don't have access to $d^+(v)$ queries.

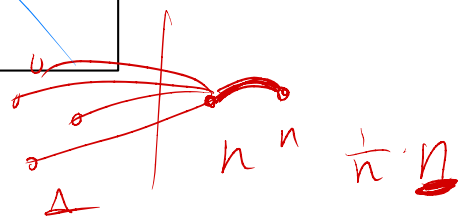
Solution: We'll set a random variable so that in Expectation, $E[X_u] = d^+(u)$

Given $u \in V$:

1. Query $d(u)$ & a uniform nbr of u, denoted v
2. Query $d(v)$
3. Let

$$X_u = \begin{cases} d(u) & \text{if } u < v \\ 0 & \text{o.w.} \end{cases}$$

$$E_x[X_u] = \frac{d^+(u)}{d(u)} \cdot d(u) = d^+(u)$$



Problem: We don't have access to $d^+(v)$ queries.

Solution: We'll set a random variable so that in Expectation, $E[X_u] = d^+(u)$

Given $u \in V$:

1. Query $d(u)$ & a uniform nbr of u, denoted v
2. Query $d(v)$
3. Let

$$X_u = \begin{cases} d(u) & \text{if } u < v \\ 0 & \text{o w} \end{cases}$$

$$Ex[X_u] = \frac{d^+(u)}{d(u)} \cdot d(u) = d^+(u)$$

Problem: max value $d(u)$ can be as high as n.

Problem: We don't have access to $d^+(v)$ queries.

Solution: We'll set a random variable so that in Expectation, $E[X_u] = d^+(u)$

Given $u \in V$:

1. Query $d(u)$ & a uniform nbr of u, denoted v
2. Query $d(v)$
3. Let

$$X_u = \begin{cases} d(u) & \text{if } u \prec v \text{ \& } d(u) \leq \sqrt{2m/\varepsilon} \\ 0 & \text{o.w.} \end{cases}$$

$$E_x[X_u] = \frac{d^+(u)}{d(u)} \cdot d(u) = d^+(u)$$

~~Problem: max value $d(u)$ can be as high as n.~~

Analysis:

$$x_u = \begin{cases} d(u) & \text{if } u < v \text{ and } d(u) \le \sqrt{2m/\varepsilon} \\ 0 & \text{o.w} \end{cases}$$

Analysis:

$X_u = \begin{cases} d(u) & \text{if } u < v \text{ and } d(u) \le \sqrt{2m/\varepsilon} \\ 0 & \text{o.w.} \end{cases}$

Observe that there are two sources of randomness:

The choice of u and the choice of u's nbr, v.

Analysis:

$X_u = \begin{cases} d(u) & \text{if } u \prec v \text{ and } d(u) \le \sqrt{2m/\varepsilon} \\ 0 & \text{o.w.} \end{cases}$

Observe that there are two sources of randomness:

The choice of u & the choice of u's nbr, v.

Fix some $u \in V$.

If $d(u) \le \sqrt{2m/\varepsilon}$, $\mathbb{E}_{v \in \Gamma(u)}[X_u \mid u] = d^+(u)$

$x_u = \begin{cases} d(u) & \text{if } u < v \text{ and } d(u) \le \sqrt{2m/\varepsilon} \\ 0 & \text{o.w.} \end{cases}$

Analysis:

Observe that there are two sources of randomness:

The choice of u and the choice of u's nbr, v.

Fix some $u \in V$.

If $d(u) \le \sqrt{2m/\varepsilon}$, $\mathbb{E}_{v \in \Gamma(u)}[x_u \mid u] = d^+(u)$

If $d(u) > \sqrt{2m/\varepsilon}$ then $\mathbb{E}[x_u \mid u] = 0$

Analysis:

$$X_u = \begin{cases} d(u) & \text{if } d(u) \le \sqrt{2m/\varepsilon} \text{ and } u < v \\ 0 & \text{o.w} \end{cases}$$

Observe that there are two sources of randomness:

The choice of u & the choice of u's nbr, v.

Fix some $u \in V$.

If $d(u) \le \sqrt{2m/\varepsilon}$, $\mathbb{E}_{v \in \Gamma(u)}[X_u \mid u] = d^+(u)$

If $d(u) > \sqrt{2m/\varepsilon}$ then $\mathbb{E}[X_u \mid u] = 0$

Therefore, $\mathbb{E}_{u,v}[X_u] = \frac{1}{n} \sum_{\substack{u \in V, \\ d(u) \le \sqrt{2m/\varepsilon}}} d^+(u)$

$$E_{X_{u,v}}[X_u] = \frac{1}{n} \sum_{\substack{u \in V \\ d(u) \le \sqrt{2m/\varepsilon}}} d^+(u)$$

$$X_u = \begin{cases} d(u) & \text{if } u \prec v \text{ and } d(u) \le \sqrt{2m/\varepsilon} \\ 0 & \text{o.w} \end{cases}$$

$$E_{X_{u,v}}[X_u] = \frac{1}{n} \sum_{\substack{u \in V \\ d(u) \le \sqrt{2m/\varepsilon}}} d^+(u)$$

$$X_u = \begin{cases} d(u) & \text{if } u < v \text{ and } d(u) \le \sqrt{2m/\varepsilon} \\ 0 & \text{o.w} \end{cases}$$

Claim: $\sum_{u \in V,\, d(u) \le \sqrt{2m/\varepsilon}} d^+(u) \in [(1-\varepsilon)m,\ m]$

Proof: There are at most $\frac{2m}{\sqrt{2m/\varepsilon}} = \sqrt{\varepsilon 2m}$ vertices whose deg. $> \sqrt{2m/\varepsilon}$. Denote their set by H

$$\sum_{v \in H} d^+(v) \le \binom{H}{2} \le \varepsilon m.$$

$$\sum_{u \in V,\, d(u) \le \sqrt{2m/\varepsilon}} d^+(u) = \sum_{u \in V} d^+(u) - \sum_{u \in H} d^+(u) \ge (1-\varepsilon)m$$

and also $\sum_{u \in V,\, d(u) \le \sqrt{2m/\varepsilon}} d^+(u) \le \sum_{u \in V} d^+(u) \le m$ ∎

$$X_u = \begin{cases} d(u) & \text{if } u < v \text{ and } d(u) \le \sqrt{2m/\varepsilon} \\ 0 & \text{o.w} \end{cases}$$

Hence, we have:

$$E_{X_{u,v}}[X_u] \in \frac{1}{n} \cdot [(1-\varepsilon)m, m]$$

and $\max_{u,v}\{X_u\} \le \sqrt{2m/\varepsilon}$

$\Rightarrow$ By Chernoff, sufficient to set

$$k = \frac{\max\{X_u\}}{E[X_u]} \cdot \frac{3\log(1/\delta)}{\varepsilon^2} \le \frac{n}{\sqrt{m}} \cdot \frac{6\log(2/\delta)}{\varepsilon^{2.5}}$$

Hence, we have:

$$E_{X_{u,v}}[X_u] \in \frac{1}{n} \cdot [(1-\varepsilon)m, m]$$

& $\max_{u,v}\{X_u\} \leq \sqrt{2m/\varepsilon}$

$\Rightarrow$ By Chernoff, sufficient to set

$$k = \frac{\max\{X_u\}}{E[X_u]} \cdot \frac{3\log(1/\delta)}{\varepsilon^2} \leq \boxed{\frac{n}{\sqrt{m}}} \cdot \frac{6\log(2/\delta)}{\varepsilon^{2.5}}$$

Final alg:

Input: $\tilde{m} \in [m, 2m]$ const. approx. of m.
ε - approximation parameter
δ - confidence parameter

1. Repeat $k = \frac{n}{\sqrt{m}} \cdot \frac{\log(1/\delta)}{\varepsilon^{2.5}}$ times:
   a. Sample a vertex $u_i \in V$ u.a.r. & query $d(u_i)$
   b. Sample a uniform nbr of u_i, denote it v_i and query $d(v_i)$
   c. Let $\chi_{u_i} = \begin{cases} d(u_i) & \text{if } u_i \prec v_i \text{ \& } d(u_i) \le \sqrt{2m}/\varepsilon \\ 0 & \text{o.w.} \end{cases}$
2. Let $\chi = \frac{1}{k} \sum_{i=1}^{k} \chi_{u_i}$
3. Return $\tilde{m} = n \cdot \chi$ as the estimate.