CS5234

Week 1
August 16

Part One: Sublinear time

Ex Graph is 1 TB
Disk scan: 200 Mb/s ⇒ 83 min
Disk seek: 1 Mb/s ⇒ 11.5 days

BFS takes > 1 week!!!

What can we do without scanning the whole graph?

Problem	Alg
Is G connected? # connected components?	BFS $O(n+m)$
↳ $O(1/\varepsilon^2 d)$ / $O(d/\varepsilon^3)$	
Weight of MST?	Prim's $O(m \log n)$
↳ $O(dW^4 \log(W)/\varepsilon^3)$	

Average degree? | Scan $O(n)$

$\Rightarrow O(\sqrt{n} \cdot \varepsilon^{-9/2})$

Diameter? | BFS $O(n(m+n))$

$\Rightarrow O(1/\varepsilon^3)$

Max matching | Edmunds Blossom Alg

$\Rightarrow O(d^4/\varepsilon^2)$

Trade-off: approximate solution

Eg:

$MST(G)(1-\varepsilon) \leq ALG(G) \leq MST(G)(1+\varepsilon)$

or

$\begin{cases} \text{if } G \text{ is connected} \Rightarrow \text{true} \\ \text{if } G \text{ is "}\varepsilon\text{-far" from connected} \Rightarrow \text{false} \end{cases}$

Warm up:

n element array

0	0	0	0	0	0	1	0	0	0	0

0 = good
1 = error

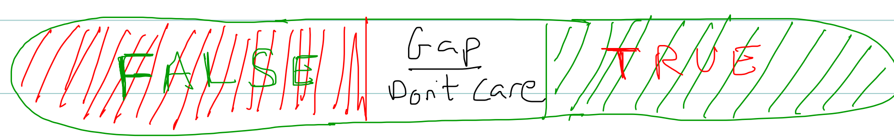
Is array all 0's?

Alg: check all cells
Time: $O(n)$
Impossible to do better: $\Omega(n)$

Is array mostly 0's??

If all 0's: return true
If $> \varepsilon n$ 1's: return false
Else: don't care

Gap:



$\boxed{\begin{array}{l} \text{All-Zeros}(A, \varepsilon) \\ \quad \text{Repeat } s \text{ times:} \\ \quad\quad \text{Choose random } i \in [1, n] \\ \quad\quad \text{if } A[i] = 1 \text{ then return false} \\ \quad \text{Return true} \end{array}}$

Claim: If A is all 0, then alg returns true.

$\boxed{\text{Fix } s = 2/\varepsilon}$ time: $O(1/\varepsilon)$

Claim: If A has $\geq \varepsilon n$ 1's, then alg returns false.

Proof: For a sample i, $\Pr(A[i]=1) \geq \frac{\varepsilon n}{n} \geq \varepsilon$

$$\Pr(\text{all samples are } 0) \leq (1-\varepsilon)^s$$
$$\leq (1-\varepsilon)^{2/\varepsilon}$$
$$\leq e^{-2}$$
$$\leq 1/3$$

Don't forget:

$\boxed{\begin{array}{l} e^{-x} \geq 1-x \\ (1-\frac{1}{x})^x \leq e^{-1} \end{array}}$

what if you want error $\leq \delta$?

With Probability $\geq 2/3$, correct

Set $s = \frac{1}{\varepsilon} \ln \frac{1}{\delta}$

Warm up #2:

n element array

0	0	0	1	1	0	1	0	0	0	1

0 = good
1 = error

What fraction of the array is 1's?

e.g., $4/11$

Tolerate ε error

ε

e.g., 0.4 ± 0.05

Probability of error $\leq 1/3$

```
PercentZeros(A, ε)
    sum = 0
    Repeat S times:
        Choose random i ∈ [1,n]
        sum = sum + A[i]
    Return (sum/S)
```

Will set $S = 1/\varepsilon^2 \Rightarrow O(1/\varepsilon^2)$ time

Hoeffding Bound

$Y_1, Y_2, \ldots, Y_s$ = iid random variables

$Y_i \in [0,1]$

$Y = \sum Y_i$

$$\Pr[|Y - E[Y]| \geq \delta] \leq 2e^{-2\delta^2/s}$$

Set $Y_i = 1$ if i^{th} sample is 1
$\qquad\quad 0$ if i^{th} sample is 0

$Y = \sum Y_i$

Alg returns (Y/s)

Let f = fraction 1's

$\Pr(Y_i = 1) = f$, $\quad E[Y_i] = 1 \cdot \Pr[Y_i = 1] + 0 \cdot \Pr[Y_i = 0]$
$\qquad\qquad\qquad\qquad\qquad\quad = f$

$E[Y] = sf$

$E[Y/s] = f$

$$\Pr[|Y/s - f| \geq \varepsilon] = \Pr[|Y - fs| \geq \varepsilon s]$$
$$= \Pr[|Y - E[Y]| \geq \varepsilon s]$$

Set $s = 1/\varepsilon^2$

$$\leq 2e^{-2(\varepsilon s)^2/s}$$
$$\leq 2e^{-2\varepsilon^2 \cdot 1/\varepsilon^2} \leq 1/3$$

Graph Connectivity

Graph $G = (V, E)$

n nodes $\rightarrow d \geq 3$

max degree d $[m \leq dn]$

Adjacency list

query: $nbr(u, i)$ returns i^{th} neighbor of u

Question: is G disconnected?

Requires $\Omega(m+n)$ time

Question 2: is G "very" disconnected?

Defn: Graph G is "ε-far" from connected if you need to modify $\varepsilon n d$ entries in adjacency list to connect it.

Note: adding an edge requires modifying 2 entries: $e = (u,v) \Rightarrow$ update u and v

$\Rightarrow$ can delete/add $\varepsilon n d / 2$ edges

Goal: if G is connected $\longrightarrow$ TRUE
if G is ε-far from connected $\to$ FALSE
otherwise $\to$ don't care

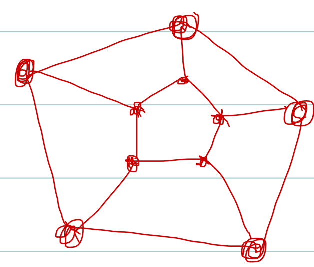
$\Pr(\text{error}) \leq 1/3$

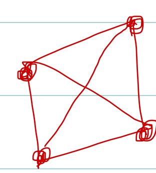
Preliminary claims

Lemma 1: if G is ε-far from connected then it has $\geq \varepsilon dn/4$ connected components.

Proof: Assume G has $< \varepsilon dn/4$ C.C.
Then can connect G by adding $< \varepsilon dn/4 - 1$ edges $\overset{?}{\Longrightarrow}$ Contradiction

$d=3$





Cannot add 1 edge to connect since $d=3$

Does not have 2 nodes with free edge slot

Full CC with ℓ nodes has $\geq \frac{(\ell-1)d}{2}$ edges
Spanning tree has $\ell-1 < \frac{(\ell-1)d}{2}$ edges
$\Rightarrow$ First delete one edge from full CC.

Lemma 1 (cont.)

Total changes: $< \varepsilon dn/4$ edges deleted
(to make room)
$< \varepsilon dn/4 - 1$ edges added
(to connect)

$\Rightarrow \leq \varepsilon dn$ changes to adjacency list to connect G
$\Rightarrow G$ <u>not</u> ε-far from connected
$\Rightarrow$ contradiction

Lemma 2: if G is ε-far from connected then: $\geq \varepsilon dn/8$ connected components in G are of size $\leq 8/\varepsilon d$

<u>Proof:</u> Counting

If $\varepsilon dn/8$ have $> 8/\varepsilon d \Rightarrow > n$ nodes
$\Rightarrow < \varepsilon dn/8$ have $> 8/\varepsilon d$
Lemma 1 says $\geq \varepsilon dn/4$ in total
$\Rightarrow \geq \varepsilon dn/4 - \varepsilon dn/8 = \varepsilon dn/8$ have $\leq 8/\varepsilon d$

Connected(G, n, d, ε)

Repeat $16/\varepsilon d$ times
  Choose u at random
  Do a BFS from u, stopping when $8/\varepsilon d$ nodes are found
  If $|CC(u)| \le 8/\varepsilon d$, return FALSE
Return TRUE

Claim: time is $O(1/\varepsilon^2 d)$

Proof: Each BFS takes time $(8/\varepsilon d)\, d$.
Repeat $16/\varepsilon d$.
$\Rightarrow (16/\varepsilon d)(8/\varepsilon d)\, d = O(1/\varepsilon^2 d)$

Claim: if G is connected, returns TRUE

Lemma 3: if G is ε-far from connected then returns FALSE

Proof: By Lemma 2, $\exists\ \varepsilon dn/8$ CC of size $\le 8/\varepsilon d$.
Each has ≥ 1 node

$$\Pr[u \text{ is in CC of size} \le 8/\varepsilon d] \ge \frac{(\varepsilon dn/8)}{n} \ge \varepsilon d/8$$

Lemma 3 (continued):

$$\Pr(\text{return TRUE}) \leq \left(1 - \varepsilon d/8\right)^{16/\varepsilon d}$$

error!

$$\leq e^{-(\varepsilon d/8)(16/\varepsilon d)} \leq e^{-2}$$

$$\leq 1/3$$

$\Rightarrow$ alg is correct w.p. $\geq 2/3$