CSE 4502/5717

BIG DATA ANALYTICS

LECTURE ON 9-27-22

RANDOMIZED SELECTION: FLOYD & RIVEST (1977)

(1) PICK A RANDOM SAMPLE S with $|S| = s$.

(2) PICK l_1 and l_2 FROM S s.t.:

(a) The i^{th} smallest element of $x \in [l_1, l_2]$.

(b) $|\{e \in x : l_1 \leq e \leq l_2\}|$ is "SMALL"

(3) SCAN through X and keep only

$$Y = \{ q \in X : l_1 \leq q \leq l_2 \}.$$

Make sure that Conditions 2a and 2b are MET.

Let $X_1 = \{ q \in X : q < l_1 \}$.

Let $n_1 = |X_1|$. Let $n_2 = |Y|$.

If $i > n_1$ and $i \leq (n_1 + n_2)$ then 2a is MET.

Let $\mathrm{Rank}(q, S) = j$, FOR some element $q \in S$. $\mathrm{Rank}(q, X) = r_j$

$$E[r_j] = j\frac{n}{s}.$$

LEMMA: $\mathrm{PROB}\left\{\left[r_j - j\frac{n}{s}\right] > \boxed{\sqrt{3\alpha}\,\frac{n}{\sqrt{s}}\sqrt{\log n}}\right\} \le n^{-\alpha}.$

Let $\boxed{d_1 = \sqrt{32}\,\frac{n}{\sqrt{s}}\sqrt{\log n}}$ — (1)

$E[r_j] = j\frac{n}{s}$.

$j\frac{n}{s} - d_1$ $\qquad\qquad\qquad\qquad$ $j\frac{n}{s} + d_1$

The Lemma says that r_j lies in this interval with a high prob.

OUT_OF_CORE ALGORITHMS TO BEGIN with $N = n$.

REPEAT UNTIL

1. Do ONE SCAN through the input & keep every el. in S with a prob. $\frac{M}{2N}$.

2. Let $l_1 \in S$ be such that $\text{Rank}(l_1, S) = \boxed{i\frac{\delta}{n} - \sqrt{4\alpha \log n}}$

   Let $l_2 \in S$ be such that $\text{Rank}(l_2, S) = \boxed{i\frac{\delta}{n} + \sqrt{4\alpha \log n}}$

(3) SCAN through the input to identify Y & write it in the DISK.

(4) If any of the two conditions 2a and 2b is not met, start all over;

(5) $N = |Y|$; $i = i - n_1$;

UNTIL $N \leq M$

PERFORM AN APPROPRIATE Selection on the remaining elements and output.

ANALYSIS: Use Chernoff bounds to show that $s \le \frac{3}{4}M$ w.h.p.

Step 1 takes $\frac{N}{B}$ I/O operations.

Step 2 takes no I/O.

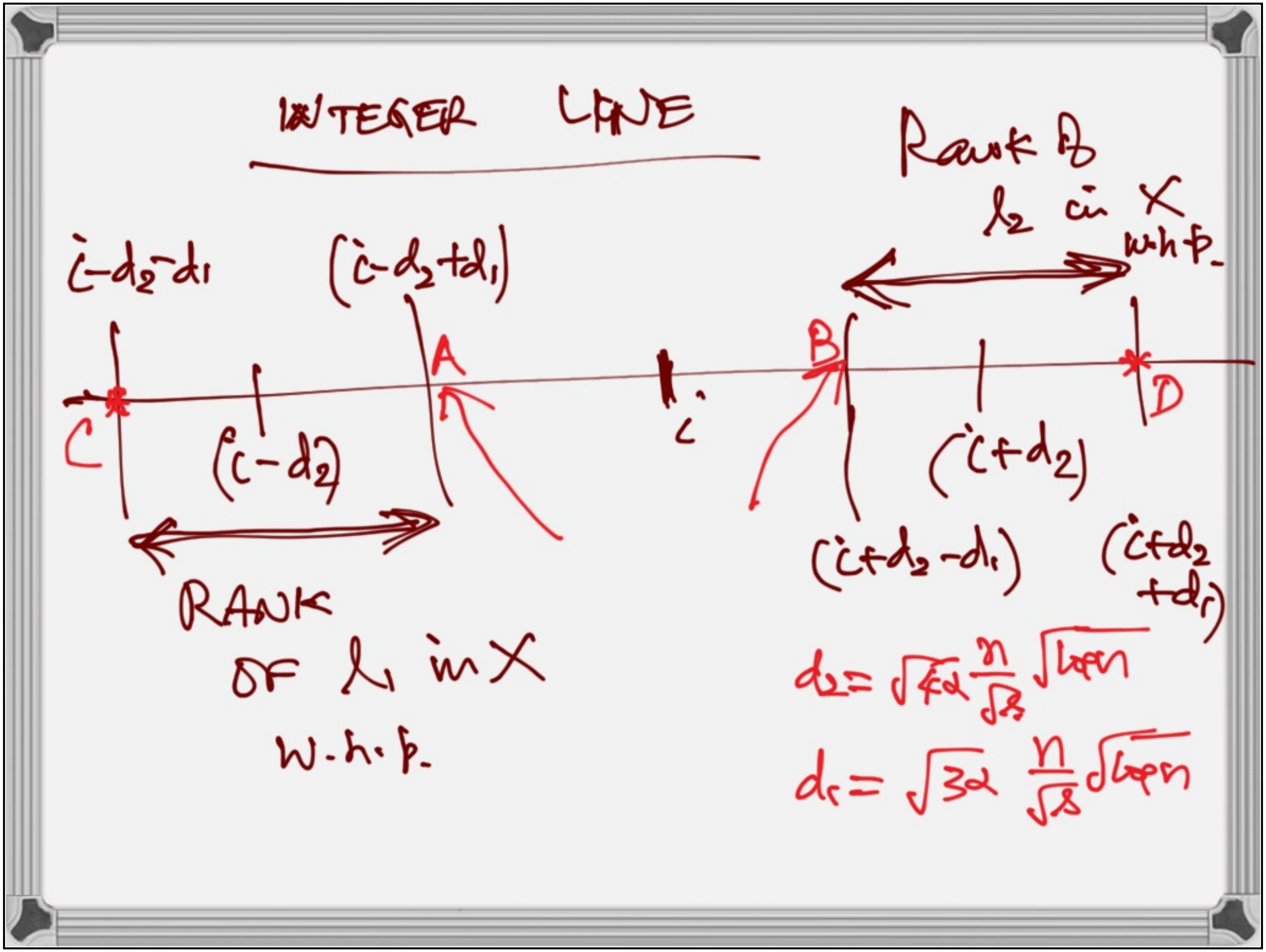
Step 3 takes $\frac{N}{B}$ I/O Operations.

$$E[\text{Rank}(l_1, x)] = i - \sqrt{4\alpha}\,\frac{N}{\sqrt{s}}\sqrt{\log n}$$

$$E[\text{Rank}(l_2, x)] = i + \sqrt{4\alpha}\,\frac{N}{\sqrt{s}}\sqrt{\log n}$$

$$d_2 = \sqrt{4\alpha}\,\frac{N}{\sqrt{s}}\sqrt{\log N} \quad \text{——— } 2$$

$$d_1 = \sqrt{3\alpha}\,\frac{N}{\sqrt{s}}\sqrt{\log N} \quad \text{——— } 1$$


INTEGER LINE
Rank of l_2 in X
w.h.p.
$i-d_2-d_1$
$(i-d_2+d_1)$
A
B
C
D
i
$(i-d_2)$
$(i+d_2)$
$(i+d_2-d_1)$
$(i+d_2+d_1)$
RANK OF l_1 in X
w.h.p.
$d_2 = \sqrt{4\alpha}\,\frac{n}{\sqrt{s}}\sqrt{\log n}$
$d_1 = \sqrt{3\alpha}\,\frac{n}{\sqrt{s}}\sqrt{\log n}$

$$|Y| \leq D - C \quad \text{w.h.p}$$

$$= (i + d_2 + d_1) - (i - d_1 - d_2)$$

$$= 2(d_1 + d_2)$$

$$= 2 \frac{N}{\sqrt{s}} \sqrt{\log N} \left(\sqrt{4\alpha} + \sqrt{3\alpha} \right).$$

$\Rightarrow$ CONDITION 2a holds w.h.p.

Also, 2b holds. $|Y| = \tilde{O}\left(\frac{N}{M^{0.4}} \right)$

Total # of I/O operations

$$= 2\frac{n}{B} + 2\frac{n}{M^{0.4}B} + 2\frac{n}{M^{0.8}B} + \cdots$$

$$= (2+\epsilon)\frac{n}{B} \text{ for some } \epsilon < 1.$$

w.h.p.

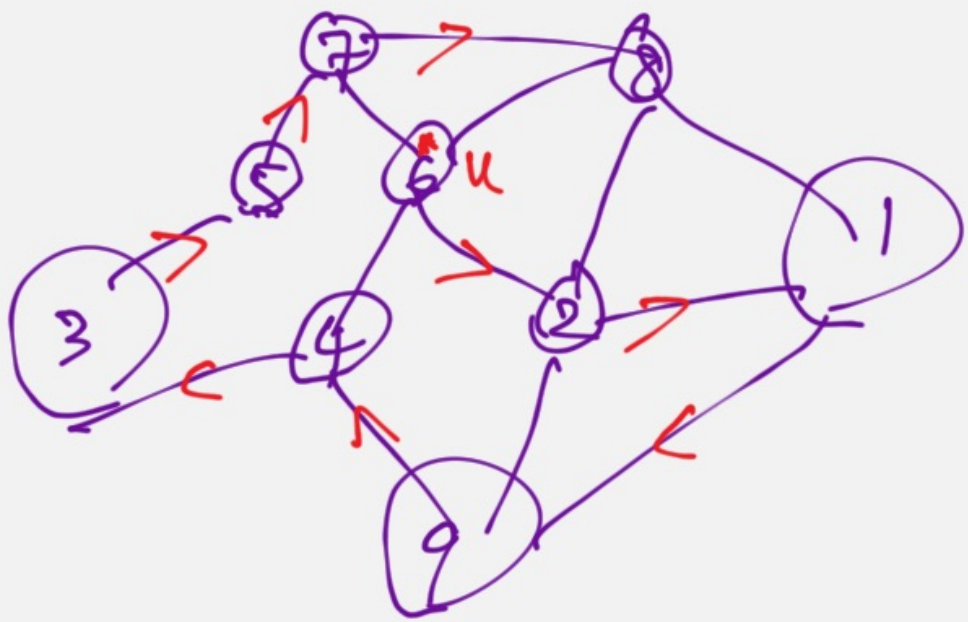
PROBLEM: GRAPH SEARCH:

INPUT: AN UNDIRECTED GRAPH $G(V,E)$.

GOAL: SEARCH the graph.

VISITED [i] = False; $1 \le i \le |V|$.

DFS (u)
    Process(u); visited [u] := true;
    For $w \in Adj(u)$ do
        IF ! visited(w) then
            DFS (w);


7
8
5
6
u
1
3
4
2
9

Out-of-Core Algorithms:

Assume that $M = \Theta(|V|)$.

Note that for every node in the graph, we have to access all of its neighbors.

Assume that we use Adjacency lists and BREADTH FIRST SEARCH.

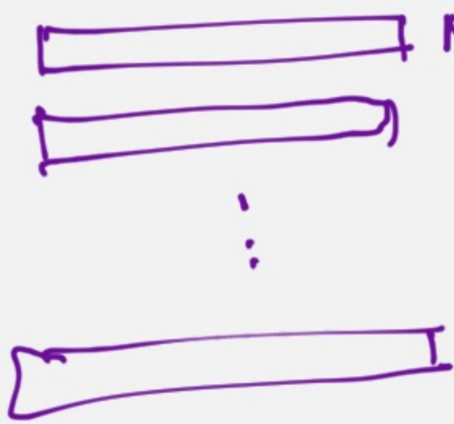
BFS (G):

Start FROM a node u.

Visit nodes at a distance 1;

Visit nodes at a distance 2;

and so on.



$n = |V|$.

I/O COMPLEXITY:

$$\sum_{u \in V} \left\lceil \frac{d_u}{B} \right\rceil ;$$

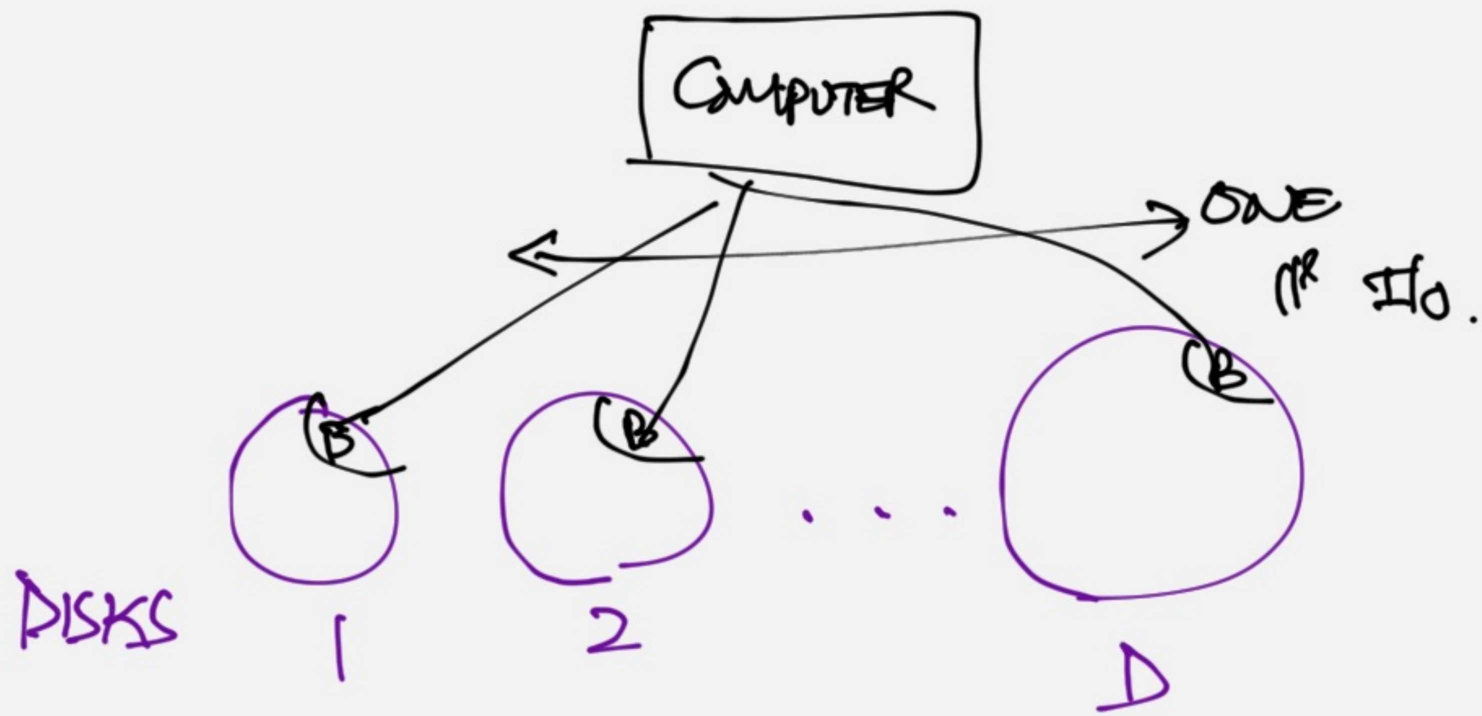
$d_u \rightarrow$ DEGREE OF u FOR ANY $u \in V$.

$$\leq \sum_{u \in V} \left(\frac{d_u}{B} + 1 \right) = \frac{2|E|}{B} + |V|$$

$$= O\left(\frac{|E|}{B} + |V| \right)$$

This is optimal if $|E| > |V| B$.

PARALLEL DISKS MODEL:



SORTING: The input will be across the disks. The output will also be across the disks.