Data arrives in a stream:

$S = s_1, s_2, \ldots s_T$

Ex: Twitter tweets → on avg, # hashtags / tweet?

→ # users tweet / day?

→ avg # tweets / user / day?

Facebook friend changes

→ CC in FB graph?

→ K-connected?

→ # triangles

Sensor data → avg temperature

New babies → most common names

Stock market → most traded stocks

3 Examples

1) Network router → count # flows

2) Twitter → count users

3) Directed Graph: count # reachable vertices for each (size transitive closure)

Counting distinct elements

(eg. # users who tweet each day)

$S = A, B, A, A, C, B, A, D, D, B, C, D$

$\text{count}(S) = 4$

Notation: $S = s_1, s_2, \dots, s_T$

$t = \text{count}(S)$

Solution 1: $A[1, \dots, t]$ store counts

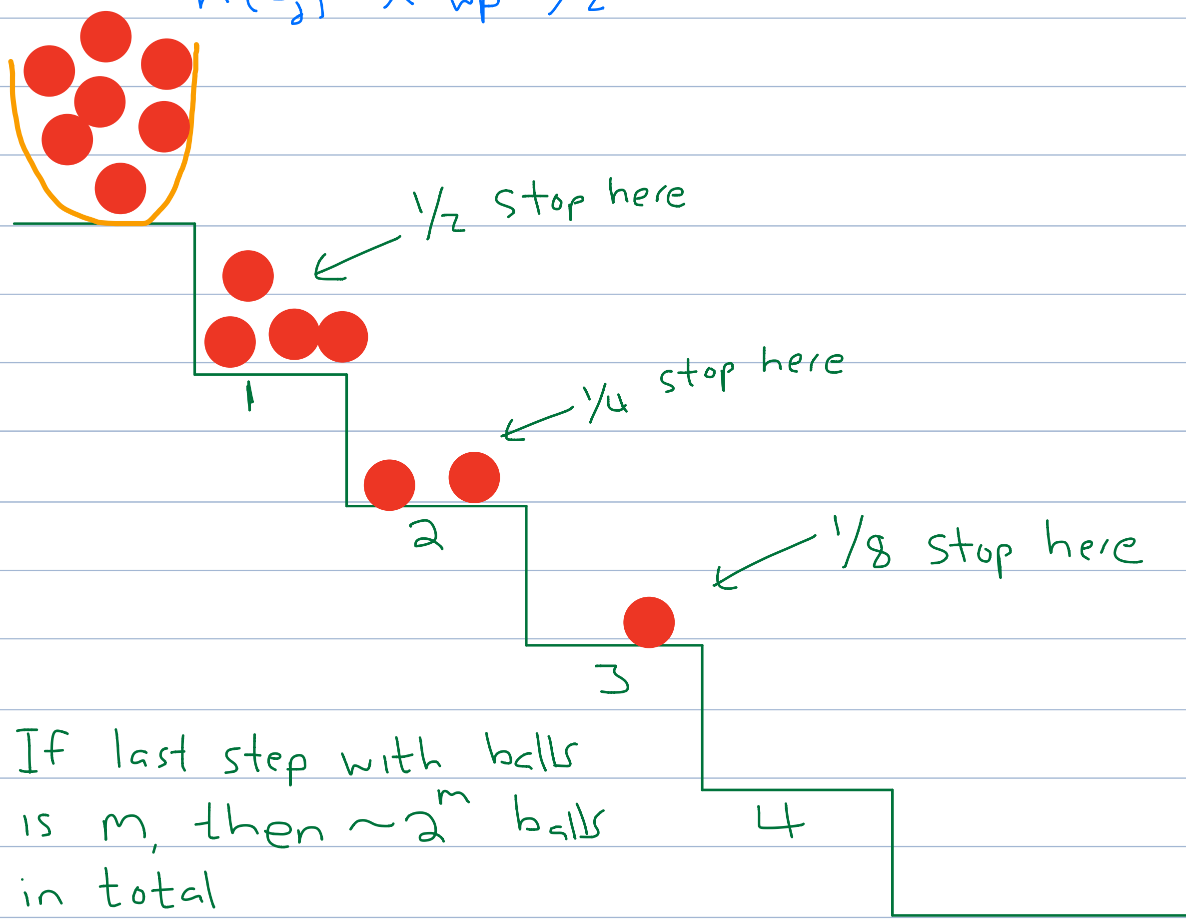
or: hash table

or: Bloom Filter

$\Rightarrow$ space $O(t)$

Idea: map each item to exponential variable

$h(s_j) = X$ wp $1/2^X$



If last step with balls is m, then $\sim 2^m$ balls in total

Flajolet-Martin (FM)

① Let $h: S_j \to [0,1]$ be uniform random hash

② Let $X = \min_{j \in [1,\dots,t]} h(S_j)$

③ Return $(1/X) - 1$

Why? $\Pr(h(S_j) < 1/2) = 1/2 \leftarrow$ Step 1

$\Pr(h(S_j) < 1/4) = 1/4 \leftarrow$ Step 2

$\vdots$

$\Pr(h(S_j) < 1/2^K) = 1/2^K \leftarrow$ Step K

If any reach step K, return $2^K \to 1/\min$

Analysis:

$$E[X] = \int_0^1 \Pr[X > \lambda]\, d\lambda$$

$$= \int_0^1 \Pr[\forall j, h(S_j) > \lambda]\, d\lambda$$

$$= \int_0^1 (1-\lambda)^t\, d\lambda$$

$$= \left. \frac{-(1-\lambda)^{t+1}}{t+1} \right|^1$$

'0

$$= \boxed{1/t+1}$$

$$\cancel{\boxed{E[FM] = E[1/X - 1] = 1/E[X] - 1 = t}}$$

cannot invert $E[1/X]$

$$Var[X] = E[X^2] - E[X]^2$$

$$E[X^2] = \int_0^1 Pr[X^2 > \lambda]\, d\lambda$$

$$= \int_0^1 Pr[X > \sqrt{\lambda}]\, d\lambda$$

$$= \int_0^1 (1-\sqrt{\lambda})^t\, d\lambda \leftarrow u = 1-\sqrt{\lambda}$$

$$\lambda = (1-u)^2$$

$$= \int_1^0 u^t(-2(1-u))\, du \qquad d\lambda = -2(1-u)\,du$$

$$= 2\int_0^1 u^t\, du - 2\int_0^1 u^{t+1}\, du$$

$$= \frac{2}{t+1} - \frac{2}{t+2}$$

$$= 2/(t+1)(t+2)$$

$$\mathrm{Var}[X] = \frac{2}{(t+1)(t+2)} - \frac{1}{(t+1)^2} = \frac{t}{(t+1)^2(t+2)}$$

$$\leq 1/(t+1)^2$$

$$\boxed{\begin{array}{l} \text{FM+} \\ \quad \text{1) Run } a \text{ copies of FM: } X_1, X_2, \dots, X_a \\ \quad \text{2) Return avg: } Z = (1/a) \sum X_j \\ \qquad\qquad\qquad\qquad\qquad \text{Return } (1/z) - 1 \end{array}}$$

$$E[Z] = (1/a) \cdot a \cdot (1/t+1) = 1/t+1$$

$$\mathrm{Var}[Z] = \frac{\sum \mathrm{Var}(X_j)}{a^2} = \frac{a\,t}{a^2(t+1)^2(t+2)} \leq \frac{1}{a(t+1)^2}$$

Chebychev

Let Y be r.v. where $E[Y]$

$\mathrm{Var}[Y]$

$$\Pr[|Y - E[Y]| \geq t] \leq \frac{\mathrm{Var}[Y]}{t^2}$$

FM+ analysis:

$$\Pr\left[\left|z - \tfrac{1}{t+1}\right| \ge \varepsilon\left(\tfrac{1}{t+1}\right)\right] \le \frac{(t+1)^2}{\varepsilon^2} \cdot \frac{1}{a(t+1)^2}$$

$$\le \; 1/\varepsilon^2 \cdot a$$

Choose $a = 4/\varepsilon^2 \Rightarrow$ error $\le 1/4$

W.P. $\ge 3/4$: $z \ge (1-\varepsilon)/(t+1)$
$z \le (1+\varepsilon)/(t+1)$

Recall: $(1+\varepsilon)^{-1} = 1 - \varepsilon + \varepsilon^2/2 - \cdots \le 1 - 2\varepsilon, \ \ge 1-\varepsilon$
$(1-\varepsilon)^{-1} = 1 + \varepsilon + \varepsilon^2/2 + \cdots \ge 1 + \varepsilon, \ \le 1 + 2\varepsilon$

$$\Rightarrow \frac{1}{z} - 1 \ge \frac{t+1}{1+\varepsilon} - 1 \ge (1-\varepsilon)t - \varepsilon$$

$$\ge (1-2\varepsilon)t$$

$$\frac{1}{z} - 1 \le \frac{t+1}{1-\varepsilon} - 1 \le (1+2\varepsilon)t + 2\varepsilon$$

$$\le (1+4\varepsilon)t$$

W.p. $\ge 3/4$, ans $= t(1 \pm 4\varepsilon)$

Better: take median of FM+

FM++

Let $Y_1, \dots, Y_b$ be b ind. FM+

Return median$[Y_j]$

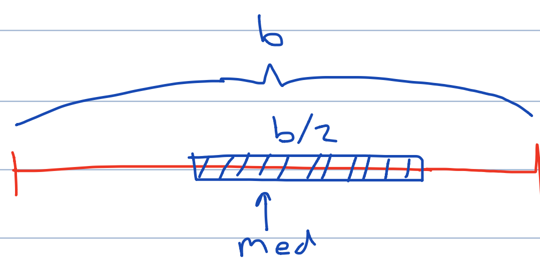
Define $X_j = 1$ if $|Y_j - t| \le 4\varepsilon t$

$\quad\;\; = 0$ otherwise

$X = \sum X_j$

$E[X_j] = 3/4$

$E[X] = 3b/4$



Key observation: median is good if $X \ge b/2$

$$\Pr[X < b/2] = \Pr\left[|X - E[X]| \ge (1/3)E[X]\right]$$

$$\le 2e^{-E[X](1/3)^2(1/3)} \le 2e^{-(3b/4)(1/27)}$$

$$\le 2e^{-b/36}$$

$b = 36 \ln(2/\delta)$

$$\le \delta$$

Conclusion: $ans = t(1 \pm 4\varepsilon)$ w.p. $1-\delta$

$O(ab)$ space $= O\left(\dfrac{\log(1/\delta)}{\varepsilon^2} \cdot \log(n) \right)$

bits of largest S_j