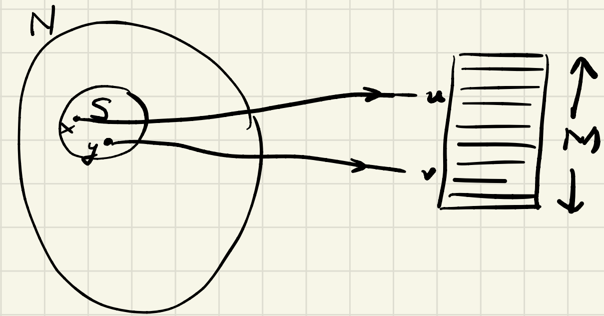
Hashing & Streaming Algorithms

$\underline{\underline{h}} : [N] \longrightarrow [M]$

$[N] = \{0, 1, \ldots, N-1\}$

$N >> M$



$h \in_R \mathcal{H}$

$\mathcal{H}$ Universal family of hash functions:

$$\forall x \neq y \qquad \underset{h \in_R \mathcal{H}}{P}\left[h(x) = h(y) \right] = \underline{\underline{\frac{1}{M}}}$$

$\underline{\text{e.g.}}$ $\quad x = x_1 \; x_2 \; x_3 \; x_4$ $\qquad x_i \in [256]$

Instead $p = 257$

Pick $a = (a_1, a_2, a_3, a_4)$ random mod 257

$$\left.\begin{array}{l} \underline{\underline{h_a}}(x) = a_1 x_1 + a_2 x_2 + a_3 x_3 + a_4 x_4 \pmod p \\ |\mathcal{H}| = 257^4 \end{array}\right\} \text{ pairwise indep?? No.}$$

Pairwise independent family of hash functions:

$$\mathcal{H} \qquad \forall x \neq y,\ u, v \quad P\left[h(x) = u \ \& \ h(y) = v \right] = \frac{1}{M^2}$$

pick $a = a_1 \, a_2 \, a_3 \, a_4 \quad b \pmod p$

$$\mathcal{H}' \qquad h'_{a,b}(x) = a_1 x_1 + a_2 x_2 + a_3 x_3 + a_4 x_4 + b \pmod p$$

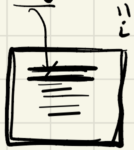
$|\mathcal{H}'| = 257^5$ $\qquad$ pairwise indep $\Rightarrow$ universal

Streaming algorithms

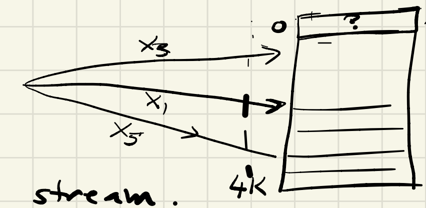
52 34 1 10 34 27 34

$x_1, x_2, x_3, x_4, \ldots, x_-, \ldots, x_N, \ldots$

$x_4 = i$, $x_- = i$, $x_N = i$



$\underline{\log N}$ bits — $\underline{\text{length of stream}}$.



Frequencies: i — m_i times

$m_{34} = 3$

Frequency moments: $F_K = \sum_i m_i^K$

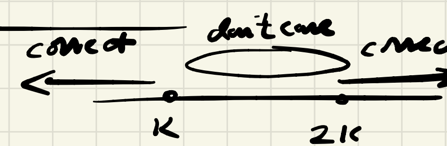
$F_0 = \sum_i m_i^0$ = # distinct elements in stream.

$F_1 = \sum_i m_i$ = length of stream

$F_2 = \sum_i m_i^2 \approx$ "variance"

---

distinct elt within factor of 2.



K: # dist elts $\leq K$ — small

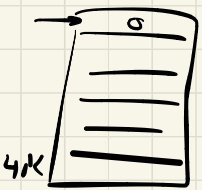
dist elts $\geq 2K$ — large

$K <$ # dist elts $< 2K$ — arbitrary answer.

Solution: Pick h random pairwise indep hash fn:

Check $h(x_i) = 0$? $\qquad$ $h'(x_i) = 0$?

Assume $\geq 2K$ distinct elts $\qquad$ $P\left[\begin{matrix}\text{one of them}\\ \text{hashes to } 0\end{matrix}\right] \geq 3/8.$



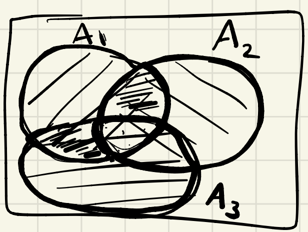
$i = 1, \dots, 2K$ $\qquad$ A_i = event i hashes to 0.

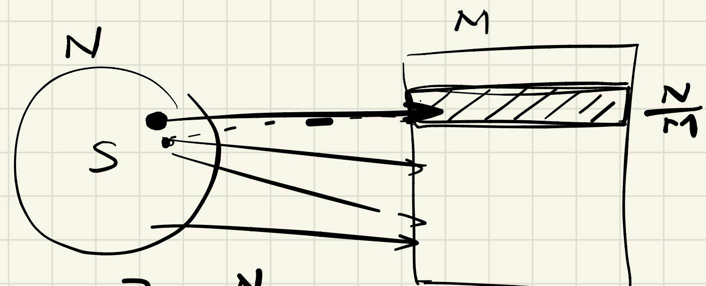
$$P[\cup A_i] \geq \sum_i P[A_i] - \sum_{i,j} P[A_i \cap A_j]$$

$$= 2K \times \frac{1}{4K} - \binom{2K}{2} \frac{1}{(4K)^2}$$

$$= \frac{1}{2} - \frac{\cancel{(2K)^2}}{2} \frac{1}{\cancel{(4K)^2}_{4}}$$

$$= \frac{1}{2} - \frac{1}{8} = \frac{3}{8}.$$





Variance calculation is valid for pairwise independence $\longrightarrow$

$$\left.\begin{aligned} E[\#\text{ in bin } j] &= \frac{N}{M} \\ Var(\#\text{ in bin } j) &= \left(\frac{1}{M}\right)\left(1 - \frac{1}{M}\right)N \end{aligned}\right\} \text{Chebyshev}$$