L14: Distinct Item Counting + Mergeable Summaries

Oct 13, 2025
Data Mining
Jeff M. Phillips

Streaming

Input $A = \langle a_1, a_2, \ldots a_m \rangle$

observe in order, one pass

small space $O(\text{poly}(\log m, \log n))$

$a_i \in [n]$

$\log m$: counter; $\log n$: label.

$f_j = |\{a_i \in A \mid a_i = j\}|$ frequency

$F_0 =$ # distinct items $j \in [n]$

$\#\ f_j > 0$

↳ need compact way to maintain unique representatives

Leverage hash function

$h : [n] \to [k]$

$\uparrow$ space of A

Bloom Filters

$t \approx \frac{k}{|S|}$

Maintain which items S in A

- no false negatives
- allow false positives

t hash functions $h_1, h_2, \ldots h_t$ iid $\mathcal{H}$

have array $B[0, \ldots, k-1]$ of bits, init 0.

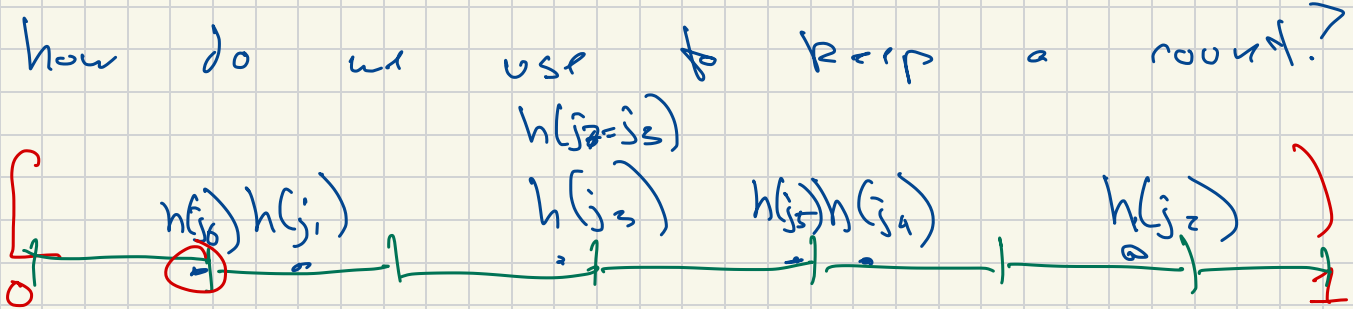
For $a_i \in A$
  for $j = 1$ to t
    set $B[h_j(a_i)] = 1$

is $q \in S$?
check if $h_j(q) = 1$ $\forall j \in 1 \ldots t$.

hash function $h: [n] \to [0,1) \subset \mathbb{R}$

continuous

how do we use to keep a count?



- each $h(j)$ random $\text{Unif}([0,1))$

- number instances = # distinct item F_0

$$\gamma = \min_{a_i \in A} h(a_i)$$

$$\hat{F}_0 = \frac{1}{\gamma} - 1 \qquad E[\hat{F}] = F_0$$

Amplification

t hash functions

$h_1, h_2 \ldots h_t$ $\overline{\text{iid}}$

$$\bar{\gamma} = \frac{1}{t} \sum_{i=1}^{t} \gamma_{h_i} \qquad \bar{F}_0 = \frac{1}{\bar{\gamma}} - 1$$

$$t = \frac{1}{\varepsilon^2} \frac{1}{\delta}$$

w.p. $1-\delta$

$$\left| \bar{F}_0 - F_0 \right| \leq \varepsilon F_0$$

$\varepsilon = \frac{1}{10}$ $\delta = \frac{1}{10}$

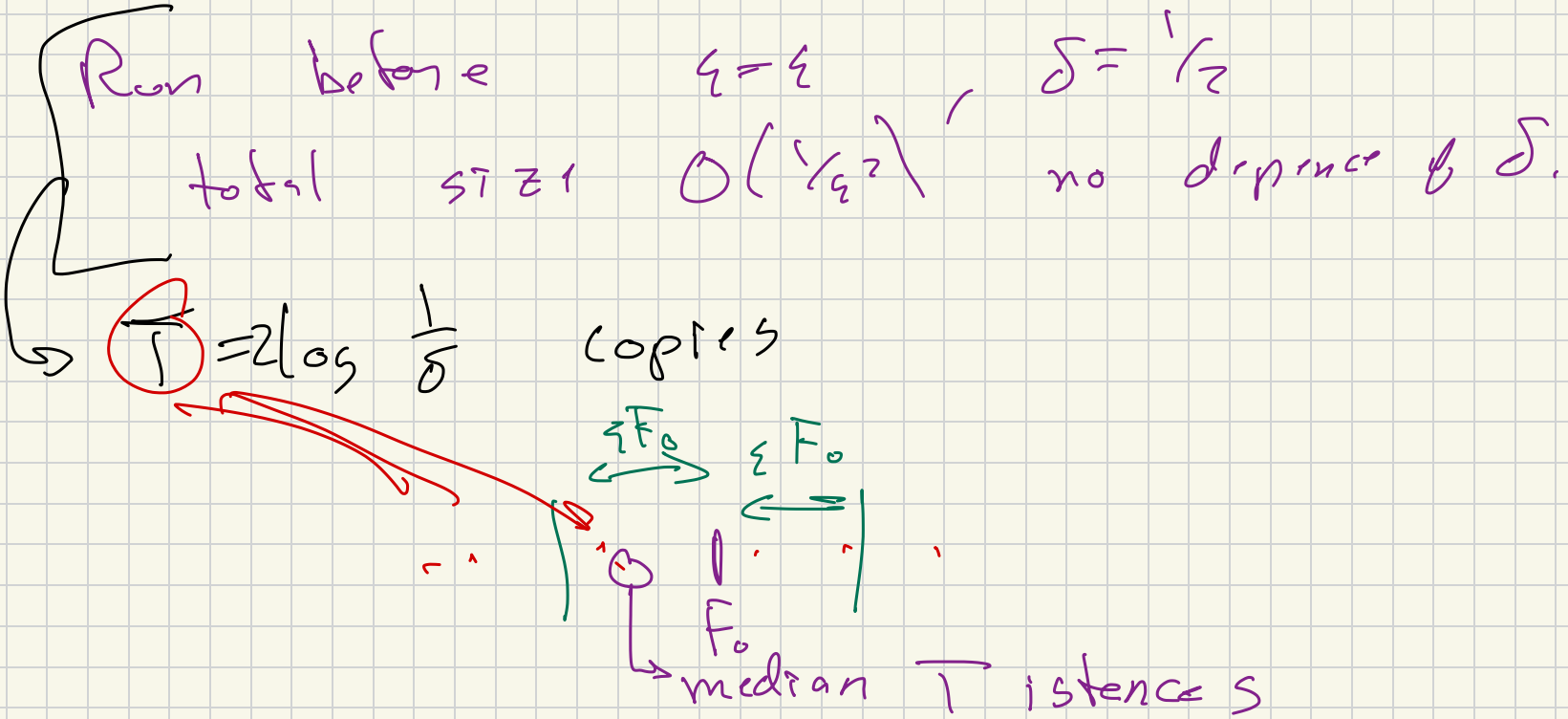
$\bar{F} = F_0 \pm 10\% F_0$

Issue #1: Failure prob high.

Issue #2: continuous, very small value γ.

Amplify prob failure δ.

Median Trick



HyperLogLog

Flajolet - Martin

From continuous $\gamma \to$ bits.

$h(a_1) \to 0\,1\,0\,1\,1\,0\,1 \quad \in [0, 1)$

$h(a_2) \to 0\,0\,0\,1\,0\,0\,1 \quad = \frac{1}{16} + \frac{1}{128}$

$\frac{1}{2}\ \frac{1}{4}\ \frac{1}{8}\ \frac{1}{16}\ \frac{1}{32}\ \frac{1}{64}\ \frac{1}{128}$

need $\log_2(F_0)$ bits.

$h(a_3) \to 0\,1\,1\,0\,1\,1\,1$

$h(a_4) \to 1\,0\,\ldots\ldots$

just first 1 bit $\Rightarrow$ 2-approx

storing the first 1 bit, requires $\log\log(F_0)$ bits

Mergeable Summaries

Summary as an Approx Data Structure

S

- insert 1 item
- return (approx) query
- merge $(S_A, S_B) \Rightarrow S_{A \cup B}$

key no increase in space as error ε.

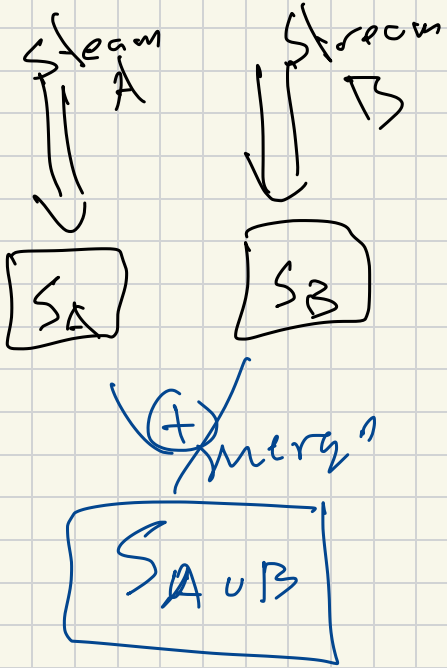
A 1 stream B 1 disjoint stream

space $s(1/\varepsilon)$ e.g. $s(1/\varepsilon) = \frac{1}{\varepsilon}$ MG, CS

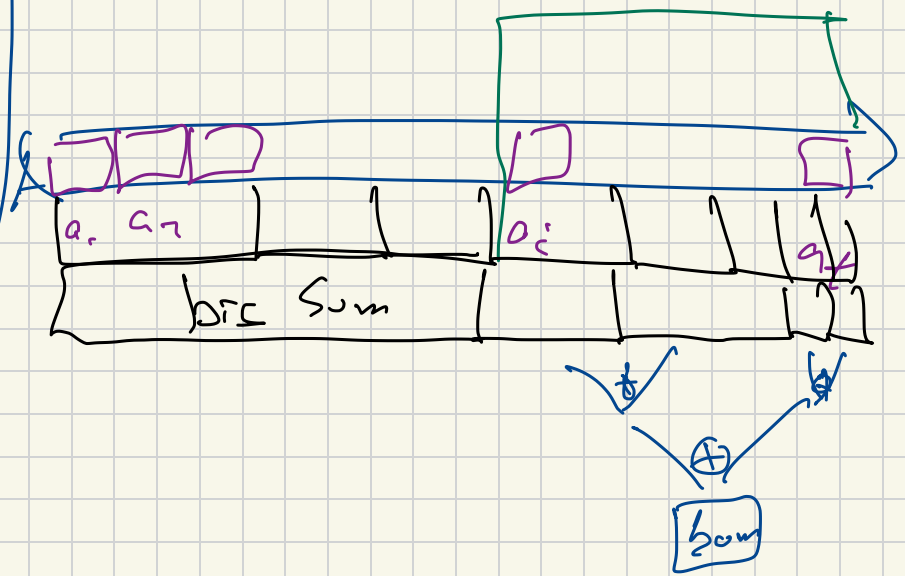
$s(1/\varepsilon) = 1/\varepsilon^2$ sample, F0

want $S_{A \cup B}$ same $s(1/\varepsilon)$!

with Merging



in parallel



Mergeable

Sampling

each $a_i \rightarrow v_i \sim \text{Unif}[0,1)$

Maintain $\min v_i \Rightarrow a_i$

$$A = \begin{bmatrix} v_1 & v_2 & v_3 \\ a_1 & a_2 & a_3 \end{bmatrix}$$

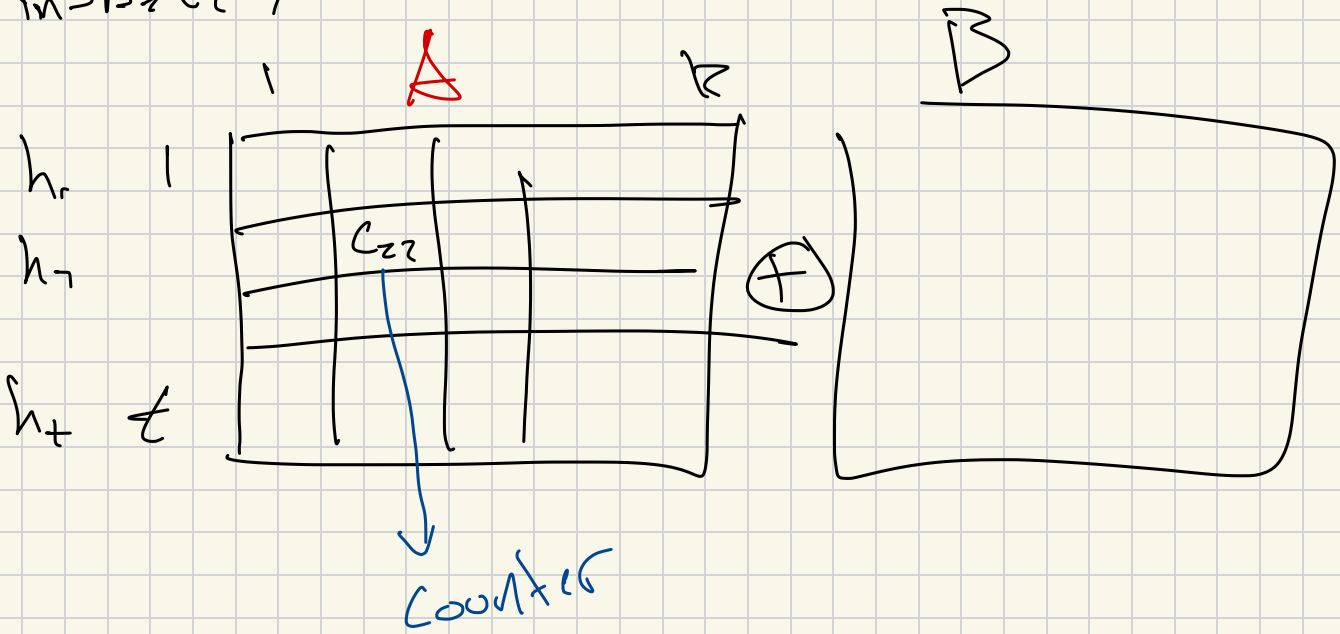
$$B = \begin{bmatrix} v_1' & v_2' & v_3' \\ b_1 & b_2 & b_3 \end{bmatrix}$$

$\Rightarrow A \cup B$

3 smallest

$$\begin{bmatrix} v_1' & v_1 & v_2' \\ b_1 & a_1 & b_2 \end{bmatrix}$$

Count MinSketch



add all counters

F_0 - sketch   Flajolet   HyperLogLog

$$\varepsilon \cdot T = \frac{1}{\varepsilon^2} \log \frac{1}{\delta}$$ of [max first 1 bit ...]

merge → "bitwise max"