

L14: Distinct Counting + Mergeable Summaries

Data Mining
Oct 20, 2022
Ieff M. Phillips



Streaming

Input

$$A = \langle a_1, a_2, \dots, a_m \rangle$$

$a_i \in [u]$

observe in order, in one pass

small space

$$O(\text{poly}(\log(m), \log(u)))$$

counter label

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

frequency

$$F_0 = \# \text{ distinct items } j \in [u]$$
$$\# \{f_j > 0\}$$

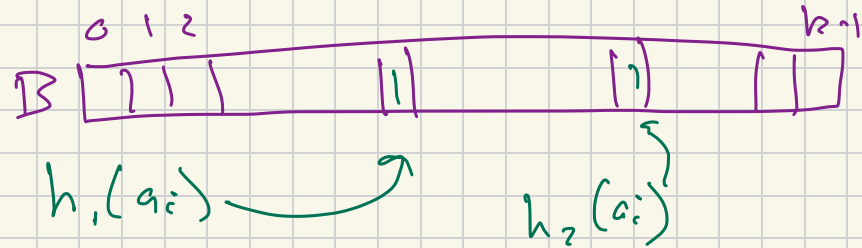
→ Bloom Filter

→ Hyper Log Log

Leverage hash function

$$h: [v] \rightarrow [k]$$

(padding needs)



Bloom Filter

Maintain which items

- no false negatives

- allow false positives

set
 S in A

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

t hash functions $h_1, h_2, \dots, h_t$

Array B of k bits (init all $B[s] = 0$)

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

Query $g \in [v]$? is in S ?
check all $h_j(g) = 1$?
if all = 1 $\Rightarrow$ True.

hash function

$$h: [U] \rightarrow [0, 1) \subset \mathbb{R}$$

continuous

how do we keep track of count?



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

$$\hat{F}_0 = \frac{1}{\gamma} - 1$$

$$E[\hat{F}_0] = F_0 = \# \text{ distinct items}$$

unbiased estimator of F_0

Amplification

t hash functions

$$h_1, h_2, \dots, h_t \sim \mathcal{H}$$

$$\bar{y} = \frac{1}{t} \sum_{j=1}^t y_{h_j}$$

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

$$\epsilon = \frac{1}{10}$$

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

$$\bar{F}_0 = \frac{1}{\bar{y}} - 1$$
$$\left(\frac{\text{actuals}}{\bar{F}_0} = \frac{1}{t} \sum_{j=1}^t F_{0,j} \right)$$

with prob $1 - \delta$

$$|\bar{F}_0 - F_0| \leq \epsilon F_0$$

↑
error
tolerance

$\pm 10\%$ true
count.

Issue #1: high
prob failure

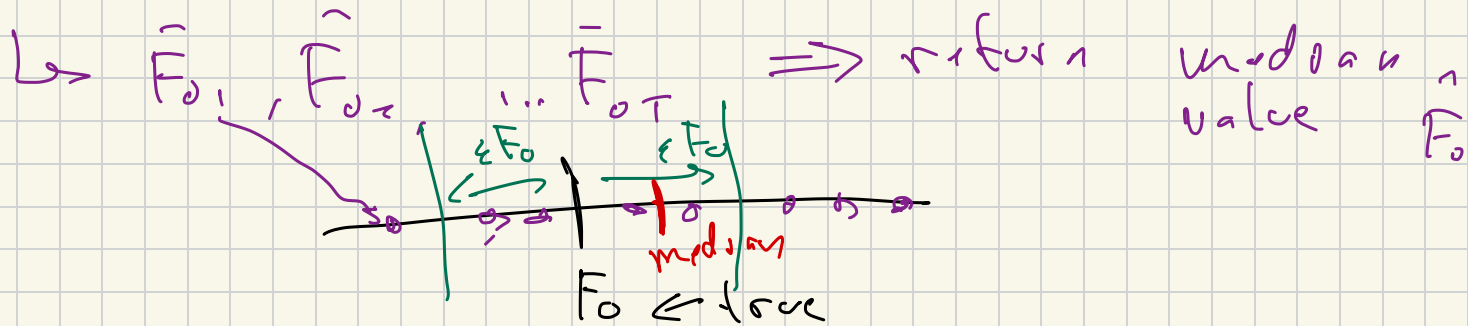
Issue #2: continuous
vs small values
 y_{h_j}

Median Trick

Run w/ ϵ , $\delta = 1/2$
total size: $O(1/\epsilon^2)$

Repeat $\Rightarrow$

$$T = 2 \log_2 \frac{1}{\delta} \text{ times} / \text{copies}$$



HyperLogLog

Flojolet - Merton

From continuous $x \rightarrow$ bids

$$h_j(a_i) \Rightarrow [0, 1) \subset \mathbb{R}$$

$$h(a_1) \rightarrow 0 \text{ } \boxed{1} \text{ } 0 \text{ } \boxed{1} \text{ } \boxed{1} \text{ } 0 \text{ } 0 \dots \rightarrow \frac{1}{4} + \frac{1}{16} + \frac{1}{32}$$

$$h(a_2) \rightarrow 0 \text{ } 0 \text{ } 0 \text{ } \boxed{1} \text{ } 0 \text{ } 0 \text{ } 0 \text{ } \dots \rightarrow \frac{1}{16}$$

$\frac{1}{2} \quad \frac{1}{4} \quad \frac{1}{8} \quad \frac{1}{16} \quad \frac{1}{32}$

$$h(a_3) \rightarrow 0 \text{ } 1 \text{ } 0 \text{ } 1 \text{ } 1 \text{ } 0 \text{ } 0$$

$$h(a_4) \rightarrow 1 \text{ } 0 \text{ } 0 \text{ } 1 \text{ } 1 \text{ } 0 \text{ } 0$$

$$a_3 = a_1$$

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

where is first 1 bid.
 $\text{storage} = \# \text{ } 0\text{'s} \Rightarrow \log \log(F_0) \text{ 2-approx bits}$

Mergeable Summaries

Summary as Approx Data Structures

fixed space

$S(A)$

• insert 1 item

• return query (approx) ϵ, δ

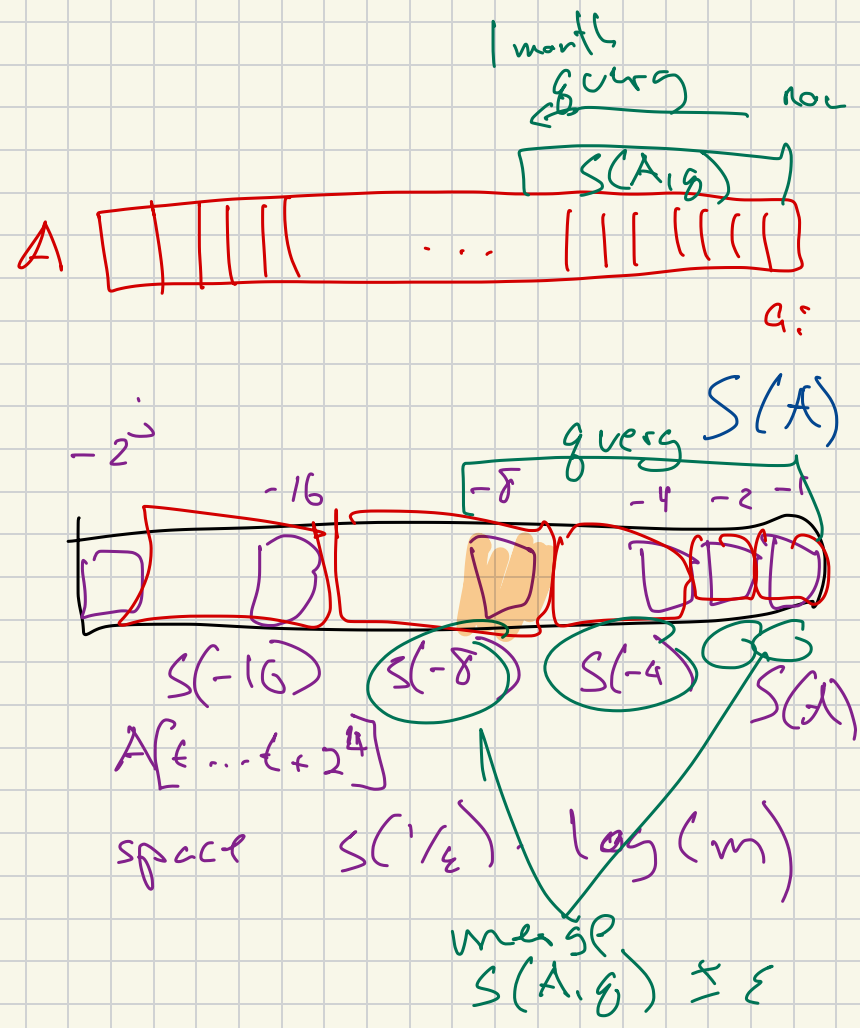
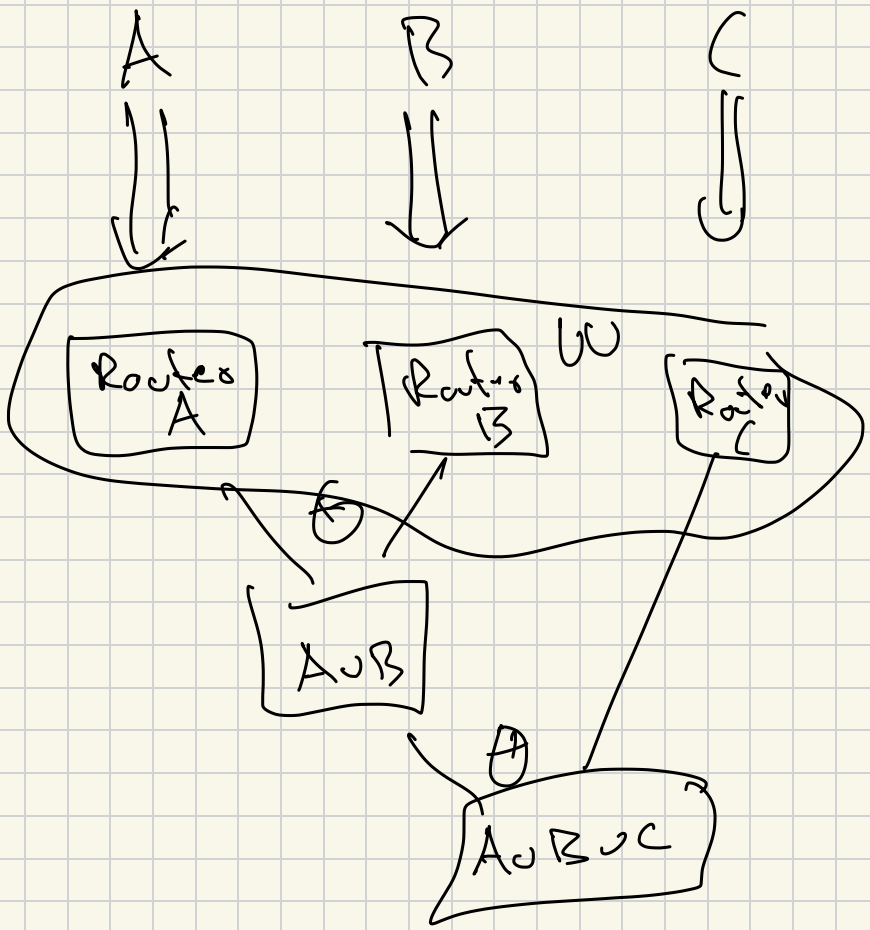
• merge $(S_A, S_B) \rightarrow S_{A \cup B}$

S_A (stream S_B another stream

space $S(1/\epsilon)$

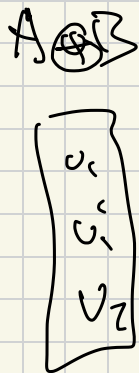
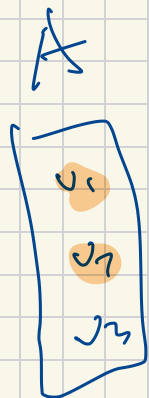
most $S_{A \cup B}$ space $S(1/\epsilon)$ some

MG, CMS $S(1/\epsilon) = \frac{1}{\epsilon}$
Sample, Fo $S(1/\epsilon) = \frac{1}{\epsilon^2}$



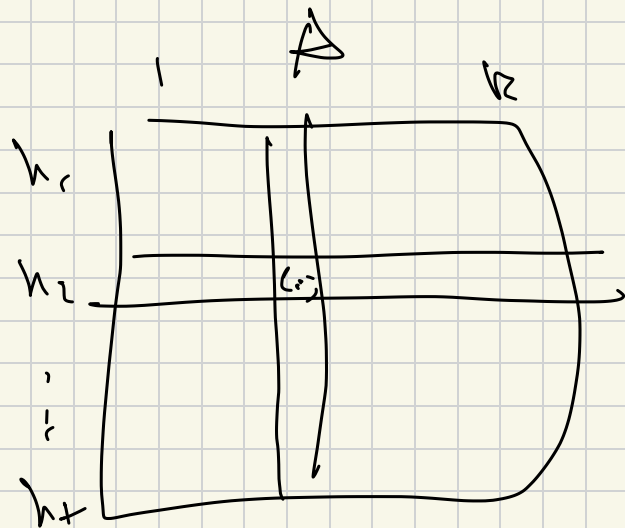
Merge Sample

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



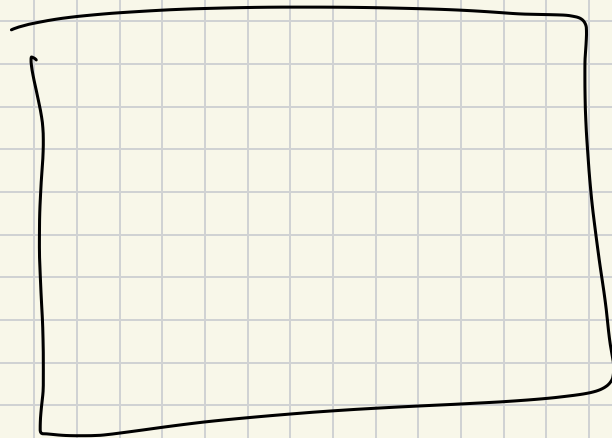
maintain min v_i
 a_i associated
min k

Geo Matrix



CMS

B



$A \cup B$

$$c_{ij} = c_{ij}^A + c_{ij}^B$$