Streaming

- Model
- Reservoir Sampling
- Counting Distinct Elements
- Hash families (pairwise independence)
- Heavy-Hitters

(β) - Heavy - Hitters

INPUT: A stream $s_1 \dots s_n \in \{1 \dots N\}$

GOAL: List of all elements that appear $\geq \beta \cdot n$ times

$\beta = 1/2$

DATA STRUCTURE: "COUNT-MIN-SKETCH"

- List all heavy hitters
- Given any $a \in \{1 \dots N\}$, let f_a = frequency of 'a'

$f_a \pm \varepsilon n$

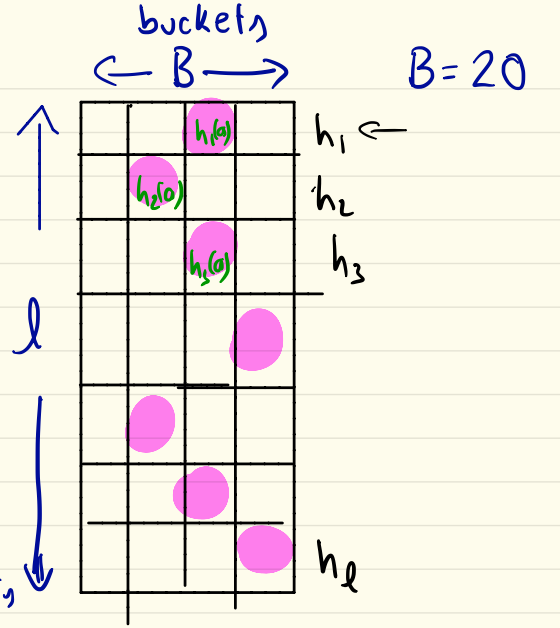
$M[i,j]$ = # of elements in the stream that map to Bucket j under h_i

INITIALIZE:

- $M[i,j] = 0$
- Hash functions

$h_1 \ldots h_\ell : \{1 \ldots N\} \longrightarrow \{1..B\}$ (buckets)

$10 \log n = \ell$



INCREMENT (element a)

```
for i = 1 to ℓ
    M[i, h_i(a)]++
```

QUERY (element a) $\approx$ approximately estimate frequency of a.

RETURN MIN $\begin{cases} M[1, h_1(a)] \\ M[2, h_2(a)] \\ \vdots \\ M[l, h_l(a)] \end{cases} \begin{matrix} \geq f_a \\ \geq f_a \\ \\ \geq f_a \end{matrix}$

Fix i, and $a \in \{1..N\}$

$$\mathop{\mathbb{E}}_{h_i} M[i, h_i(a)] = \mathop{\mathbb{E}}_{h_i}\left[f_a + \sum_{\substack{b \\ h_i(b) = h_i(a)}} f_b \right]$$

pairwise independence

$$= f_a + \left(\sum_b f_b\right)\left(\frac{1}{B}\right)$$

For a fixed $b \neq a$, $\Pr_{h_i}[h_i(b) = h_i(a)] = 1/B$

$$\leq f_a + \frac{1}{B}(n) = f_a + \frac{n}{B}$$

MARKOV'S INEQUALITY: $X \geq 0$ be a r.v

$$\Pr[X \geq t] \leq \frac{E[X]}{t}$$

$$\Pr\left[M[i, h_i(a)] - f_a \geq 2n/B \right] \leq 1/2$$

$$\Pr\left[\min_i \{M[i,h_i(a)]\} > f_a + 2n/B\right]$$

$B = 2/\varepsilon$

$\Rightarrow 2n/B = \varepsilon n$

$$= \Pr\left[\left(M[1,h_1(a)] > f_a + 2n/B\right) \text{ AND } \left(M[2,h_2(a)] \geq f_a + 2n/B\right) \text{ AND } \left(M[\ell,h_\ell(a)] \geq f_a + 2n/B\right)\right]$$

($h_1, h_2 \ldots h_\ell$ are chosen independently)

$$= \Pr\left[M(1,h_1(a)) > f_a + 2n/B\right] \cdot \Pr[\quad] \cdots \Pr\left[M(\ell,h_\ell(a)) \geq f_a + 2n/B\right]$$

$$\leq \frac{1}{2} \cdot \frac{1}{2} \cdots \frac{1}{2} = \frac{1}{2^\ell} = \frac{1}{2^{10\log n}} \approx \frac{1}{n^{10}}$$

β-Heavy Hitters: $s_1, \ldots, s_n$

for $i = 1$ to n

increment (s_i)

if $(\text{query}(s_i) \geq \beta n)$ Add s_i to list of β-heavyhitters.