DATA STREAMING : FLOW OF MASSIVE SEQUENCES OF DATA

- lots of data — n elements
- cannot store it — polylog space in n

MOTIVATIONS

- IP traffic in a router
- telephone calls
- query logs

CHALLENGE

- simple problems (statistics) become difficult

example EASY : find MAX | DIFFICULT : find QUANTILES (most frequent top k, etc)

Problems cannot be solved deterministically

FEW EXCEPTIONS: Find the missing element in a permutation of 1. 2. ... n (just n-1 elements arrive)

- sum $2 \log n$ bit
- XOR $\log n$ bit

COUNT-MIN SKETCH (Cormode - Muthu)

n items numbered $1 \ldots n$

F frequency array $F[i] = \#$ times i appears in the stream

Operations: $F[i]{+}{+}$, $F[i]{-}{-}$, invariant: $F[i] \geq 0$

▷ cannot store F entirely, have only $O(\log n)$ bits

▷ find approximation $\tilde{F}$ s.t.

$$\forall i: \quad \underbrace{F[i] \leq \tilde{F}[i] \leq F[i] + \varepsilon \|F\|}_{\text{with probability } 1-\delta}$$

$$\|F\| = \sum_{j=1}^{n} F[j]$$

Need a couple of notions

- k-wise independence
  - random variables
  - hash functions
- Markov's inequality

K-wise limited independence

$X_1, X_2, \ldots, X_n$ random variables with support $S_1, S_2 \ldots S_n$

K-wise independent if

$\forall$ choice $i_1 < i_2 < \cdots < i_K \in [n]$ and $a_{i_j} \in S_{i_j}$

$$\Pr[X_{i_1} = a_{i_1} \wedge X_{i_2} = a_{i_2} \wedge \ldots X_{i_K} = a_{i_K}] = \Pr[X_{i_1} = a_{i_1}] \times \Pr[X_{i_2} = a_{i_2}] \times \cdots \times \Pr[X_{i_K} = a_{i_K}]$$

Implication: $\mathbb{E}[X_{i_1} X_{i_2} \cdots X_{i_K}] = \mathbb{E}[X_{i_1}] \times \mathbb{E}[X_{i_2}] \times \cdots \times \mathbb{E}[X_{i_K}]$

$\{h\}_{h \in H}$ family of hash functions $[n] \to [b]$

$\Pr[h \in H] = \frac{1}{|H|}$ uniform distributions

Family H is K-wise independent if

$\forall x_1, x_2, \ldots, x_K \in [n]$ $b_1, b_2 \ldots, b_K \in [b]$

$$\Pr_{h \in H}[h(x_1) = b_1 \wedge h(x_2) = b_2 \wedge \ldots \wedge h(x_K) = b_K] = \Pr_{h \in H}[h(x_1) = b_1] \times \Pr_{h \in H}[h(x_2) = b_2] \times \cdots \times \Pr_{h \in H}[h(x_K) = b_K]$$

$[X_{i_j} = h(x_j)]$

$[a_j = b_j]$

Example

$$h(x) = a_0 + a_1 x + a_2 x^2 + \cdots + a_{K-1} x^{K-1}$$

$a_i \in$ finite field F

- $|H| = |F|^K$
- there $|F|^{K-1}$ solutions to $h(x_i) = b_i$

$a_0 = b_i - a_1 x_i - a_2 x_i^2 - \cdots - a_K x_i^K$

$\Rightarrow \underbrace{\Pr[h(x_1)=b_1 \wedge h(x_2)=b_2 \wedge \cdots \wedge h(x_k)=b_k]}_{\frac{1}{|F|^k}}$

$\underbrace{\Pr[h(x_1)=b_1]}_{\frac{1}{F}} \wedge \Pr[h(x_2)=b_2] \wedge \cdots \wedge \Pr[h(x_k)=b_k]$

$\frac{1}{F}$ $\frac{1}{F}$ $\frac{1}{F}$

$= \frac{1}{F^k}$

QED

Space to store $h(x)$ is that of $a_0 a_1 \ldots a_{n-1}$

$O(k \lg |F|)$ bits

Ex. Take prime $p \in [n-1..2n]$: show $h'(x) = (h(x) \bmod p) \bmod b$ is approximately k-wise indep.

Markov's Inequality

X = random variable ≥ 0, real $a > 0$

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

Proof.

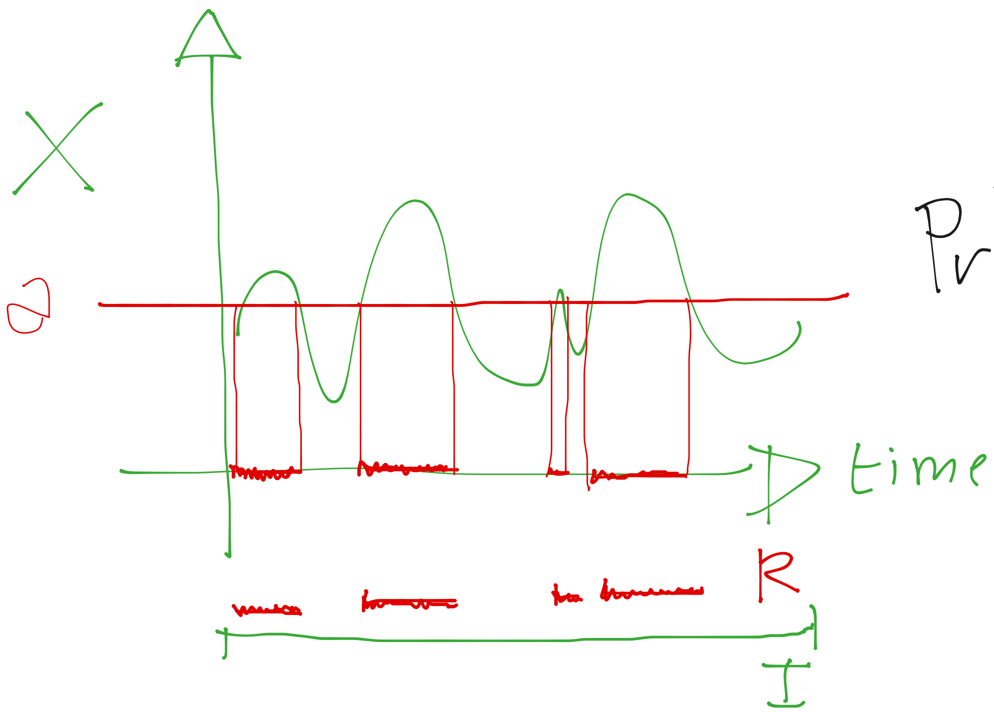
I = indicator variable = $\begin{cases} 0 & \text{if } X < a \\ 1 & \text{if } X \geq a \end{cases}$

fact $\mathbb{E}[I] = 0 \cdot \Pr[I=0] \times 1 \cdot \Pr[I=1]$
$= \Pr[I=1] = \Pr[X \geq a]$

$\Rightarrow a I \leq X$

- $X < a \Rightarrow I = 0$
- $X \geq a \Rightarrow a I = a \leq X$

$$\Rightarrow \mathbb{E}[aI] \leq \mathbb{E}[X]$$
$$\quad\;\; \| $$
$$a \cdot \mathbb{E}[I] = a \cdot \Pr[X \geq a] \quad \text{(see fact)}$$



$\Pr[X \geq a] = \frac{R}{I}$

COUNT-MIN SKETCH (Cormode - Muthu)

n items numbered $1 \ldots n$

F frequency array $F[i] = \#$ times i appears in the stream

Operations: $F[i]{+}{+}$, $F[i]{-}{-}$, invariant: $F[i] \geq 0$

▷ cannot store F entirely, have only $O(\log n)$ bits

▷ find approximation $\tilde{F}$ s.t.

$$\forall i: \quad \underbrace{F[i] \leq \tilde{F}[i] \leq F[i] + \varepsilon \, \|F\|}_{\text{with probability } 1-\delta}$$

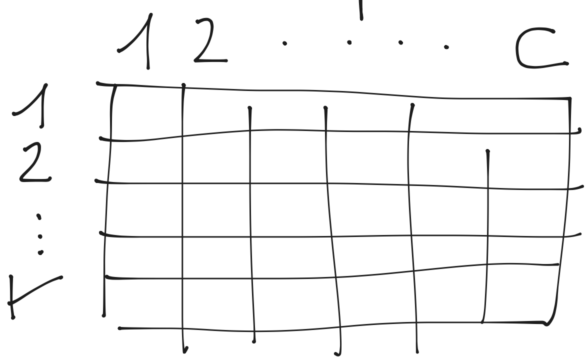
$$\|F\| = \sum_{j=1}^{n} F[j]$$

Algorithm

1. let $r = \log_2 \frac{1}{\delta}$ and $c = \frac{e}{\varepsilon}$

$e = 2.71828$ Euler's constant

2. let T be a table of $r \times c$ counters initially set to zero



3. Take a family $\mathcal{H}$ of 2-wise independent hash functions

eg $h \in \mathcal{H}$ iff $h(x) = [((ax+b) \bmod p) \bmod c] + 1$

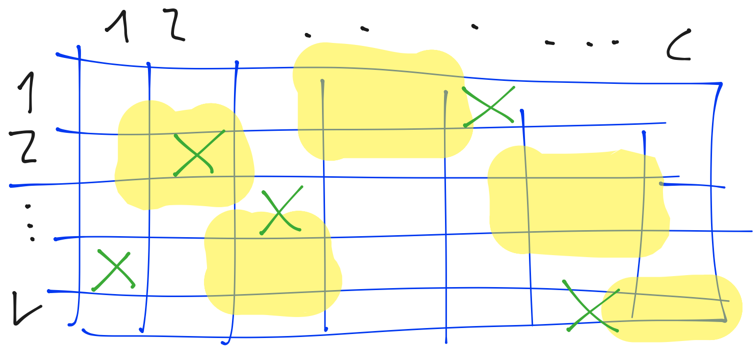
4. Choose $h_1, h_2, \ldots, h_r$ uniformly and independently from $\mathcal{H}$

eg r pairs $(a_1,b_1) \ldots (a_r,b_r)$, $a_i \neq 0$

5. given element i, associate r cells (one per row)

$T(1,h_1(i)), T(2,h_2(i)), \ldots, T(r,h_r(i))$

$i \neq i'$ might have collisions



6. OPS

$F[i]{+}{+} \Rightarrow$ increment by 1 the r cells for i

$F[i]{-}{-} \Rightarrow$ decrement, as above

7. $\tilde{F}[i] = \min\limits_{1 \le j \le r} T(j, h_j(i))$

Fact 1

Storage is $O(r \cdot c) = O(\varepsilon^{-1} \lg \delta^{-1})$ words of memory, where each word can store $\|F\|$.

proof

Each entry stores an integer in $[\|F\|]$

Fact 2 $F[i] \leq \tilde{F}[i]$

proof Let $T(j, h_j(i)) = \tilde{F}[i]$

By construction, $\exists\ i_1, i_2, \ldots, i_p$ s.t. one of them is i and

$$T(j, h_j(i)) = \sum_{\ell=1}^{p} F[i_\ell] = F[i] + X_{ji}$$

where $X_{ji} \geq 0$ is the excess (as $F[i_\ell] \geq 0$)

QED

obs To model the excess (due to hash collisions) use indicator variable

$$I_{jik} = \begin{cases} 1 & \text{if } k \neq i \text{ and } h_j(i) = h_j(k) \\ 0 & \text{otherwise} \end{cases}$$

obs $X_{ji} = \sum_{k=1}^{n} I_{jik} \cdot F[k]$

follows from the fact that $I_{jik} = 1$ iff there is a collision in cell $T(j, h_j(i))$

obs $\mathbb{E}[I_{jik}] = \frac{\varepsilon}{e}$

proof $\mathbb{E}[I_{jik}] = 0 \cdot \Pr[I_{jik} = 0] + 1 \cdot \Pr[I_{jik} = 1]$

$$= \Pr[I_{jik} = 1]$$

$$= \Pr[\exists a \in [c] : k \neq i \wedge h_j(i) = a \wedge h_j(k) = a]$$

$$= \sum_{a \in [c]} \Pr[k \neq i \wedge h_j(i) = a \wedge h_j(k) = a]$$

use the fact that h_j is 2-wise independent

$$= \sum_{a \in [c]} \underbrace{\Pr[h_j(i) = a]}_{\frac{1}{c}} \times \underbrace{\Pr[k \neq i \wedge h_j(k) = a]}_{\text{"almost" } \frac{1}{c}}$$

since one element in the domain is missing

$$= \sum_{a \in [c]} \frac{1}{c^2} = \frac{1}{c} = \frac{\varepsilon}{e}$$

Fact 3 $\quad \tilde{F}[i] \leq F[i] + \varepsilon \|F\|$ with probability $\geq 1-\delta$

proof

Let $\tilde{F}[i] = F[i] + X_{ji}$, where $X_{ji} \geq 0$ is the excess

$$\triangleright\ \mathbb{E}[X_{ji}] = \mathbb{E}\left[\sum_k I_{jik} \cdot F[k]\right] = \sum_k \mathbb{E}[I_{jik} \cdot F[k]] \leq \sum_k \left(F[k] \cdot \mathbb{E}[I_{jik}]\right) = \sum_k \left(F[k] \cdot \frac{\varepsilon}{e}\right) = \frac{\varepsilon}{e}\|F\|$$

$$\mathbb{E}[X_j] = \frac{\varepsilon}{e}\|F\| \Leftrightarrow \varepsilon\|F\| = e\,\mathbb{E}[X_{ji}]$$

$$\triangleright\ \Pr[\tilde{F}[i] > F[i] + \varepsilon\|F\|] =$$

$$\Pr[\forall_j\ \cancel{F[i]} + X_{ji} > \cancel{F[i]} + \varepsilon\|F\|] =$$

$$\Pr[\forall_j\ X_{ji} > \underbrace{\varepsilon\|F\|}_{e\,\mathbb{E}[X_{ji}]}] =$$

$$\Pr[\forall_j\ X_{ji} > e\,\mathbb{E}[X_{ji}]] =$$

$h_1, h_2, \dots, h_r$ chosen uniformly and independently

$$\underbrace{\Pr[X_{j1} > e\,\mathbb{E}[X_{j1}]]}_{} \times \Pr[X_{j2} > e\,\mathbb{E}[X_{j2}]] \times \cdots \times \Pr[X_{jr} > e\,\mathbb{E}[X_{jr}]]$$

$\hookrightarrow$ Markov's inequality $\leq \dfrac{\mathbb{E}[X_{ji}]}{e\,\mathbb{E}[X_{ji}]} = \dfrac{1}{e} < \dfrac{1}{2} \quad 1 \leq i \leq r$

$$< \left(\frac{1}{2}\right)^r = \left(\frac{1}{2}\right)^{\log_2 \frac{1}{\delta}} = \delta$$

Q.E.D

Roberto Grossi '12