DATA STREAMING : FLOW OF MASSIVE SEQUENCES OF DATA

- lots of data
- cannot store it

n elements

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 $1, 2, \ldots n$ (just $n-1$ elements arrive)

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

COUNT-MIN SKETCH (Cormode - Muthu)

n items numbered $1 \dots n$

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.

$$\forall i: \quad F[i] \leq \tilde{F}[i] \leq F[i] + \varepsilon \|F\|$$

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 $a_{i_j} \in S_{i_j}$

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

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

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

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

Family H is **k-wise independent** if

$\forall x_1, x_2, \ldots, x_k \in [n]$ $b_1, b_2 \ldots b_k \in [b]$

$$\Pr_{h \in H}\left[h(x_1) = b_1 \wedge h(x_2) = b_2 \wedge \ldots \wedge h(x_k) = b_k\right] = \Pr_{h \in H}\left[h(x_1) = b_1\right] \times \Pr_{h \in H}\left[h(x_2) = b_2\right] \times \cdots \times \Pr_{h \in H}\left[h(x_k) = b_k\right]$$

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

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 $h(x_i) = b_i$

$a_0 = b_i - a_1 x_i - a_2 x_i^2 - \cdots - a_k x_i^k$

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

$\Pr[h(x_1)=b_1] \wedge \Pr[h(x_2)=b_2] \wedge \cdots \wedge \Pr[h(x_k)=b_k]$

$\frac{1}{|F|} \quad \frac{1}{|F|} \quad \frac{1}{|F|}$

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

Space to store $h(X)$ is that of $a_0, a_1, \ldots a_{k-1}$

$O(k \lg |F|)$ 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 $\mathbb{E}[I] = 0 \cdot Pr[I=0] \times 1 \cdot Pr[I=1]$

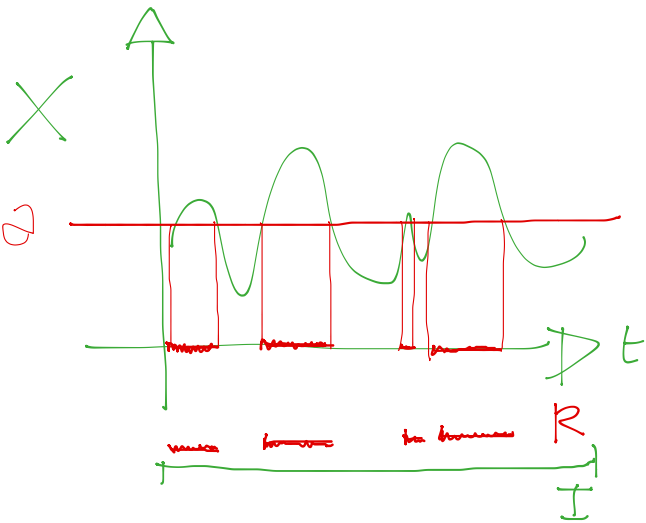
$= Pr[I=1] = Pr[X \geq a]$

$\triangleright$ $a\,I \leq X$

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

$\Rightarrow \mathbb{E}[a\,I] \leq \mathbb{E}[X]$

$a \cdot \mathbb{E}[I] = a \cdot Pr[X \geq a]$ (see fact)



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

COUNT-MIN SKETCH (Cormode - Muthu)

n items numbered $1 \dots n$

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 $O(\lg n)$ bits

▷ find approximation $\tilde{F}$ s.t.

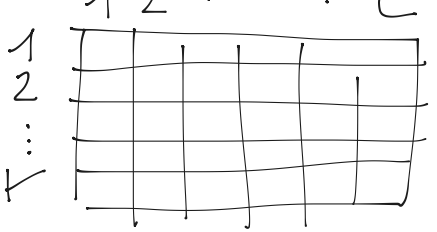
$$\forall i: \quad F[i] \leq \underbrace{\tilde{F}[i] \leq F[i] + \varepsilon \|F\|}_{\text{with probability } 1-\delta}$$

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

Algorithm

1. Let $r = \lg_2 \frac{1}{\delta}$ 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



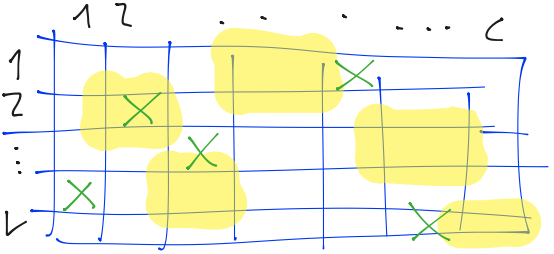
3. Take a family $\mathcal{H}$ of 2-wise independent hash functions
eg $h \in \mathcal{H}$ iff $h(x) = [(ax+b) \bmod p) \bmod c] + 1$

4. Choose $h_1, h_2, \ldots, h_r$ uniformly and independently from $\mathcal{H}$ eg r pairs $(a_1,b_1) \cdots (a_r,b_r)$ $a_i \neq 0$

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

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

$F[i]--$ ⇒ decrement, as above

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

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 $F[i] \le \widetilde{F}[i]$

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

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

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

where $X_{ji} \ge 0$ is the excess ($as\ F[i_k] \ge 0$)

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 $X_{j,i} = \sum_{k=1}^{n} I_{j,i,k} \cdot T[k]$

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

obs $\mathbb{E}[I_{j,i,k}] = \frac{\varepsilon}{c}$

proof $\mathbb{E}[I_{j,i,k}] = 0 \cdot Pr[I_{j,i,k}=0] + 1 \cdot Pr[I_{j,i,k}=1]$

$= Pr[I_{j,i,k}=1]$

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

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

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

use the fact that h_j is 2-wise independent

$Pr[h_j(i)=a] = \frac{1}{c}$

$Pr[k \neq i \wedge h_j(k)=a]$ "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

$$\widetilde{F}[i] \leq F[i] + \varepsilon \|F\|$$

with probability $\geq 1 - \delta$

proof

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

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

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

$$\Pr\left[\widetilde{F}[i] > F[i] + \varepsilon\|F\|\right] =$$

$$\Pr\left[\forall j \; F[i] + X_{ji} > F[i] + \varepsilon\|F\|\right] =$$

$$\Pr\left[\forall j \; X_{ji} > \underbrace{\varepsilon\|F\|}_{e\,\mathbb{E}[X_{ji}]}\right] =$$

$$\Pr\left[\forall j \; X_{ji} > e\,\mathbb{E}[X_{ji}]\right] =$$

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

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

$$\Pr\left[X_{ji} > e\,\mathbb{E}[X_{ji}]\right] \overset{\text{Markov's inequality}}{\leq} \frac{\mathbb{E}[X_{ji}]}{e\,\mathbb{E}[X_{ji}]} = \frac{1}{e} < \frac{1}{2} \qquad 1 \leq i \leq r$$

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

Q.E.D

Roberto Grossi '12