Today

Streaming.

Frequency Moments.

Streaming.

Input:

$x_1, x_2, x_3, \ldots, x_n$.

One at a time.

Pikachu, Squirtle, Mew, Pikachu, ...

Got to get 'em all!

Actually, no. $O(\log^c n)$ space.

Model LARGE data $_{\text{small}}$ space.

Extreme mismatch.

What to compute.

Data.

Moments!

$F_k = \sum_i m_i^k$

m_i - number of items of type i.

E.g., number of Pikachus, Squirtles, ...

F_0: Number of distinct elements.

How to compute?

F_1: Length of stream.

Easy to compute!

F_2: How to compute?

Number Distinct Elements

Claim: takes $\Omega(n)$ space for exact number of distinct items!

Pikachu, Squirtle, Mew, Squirtle, Pikachu, Squirtle

How many distinct elements?

Answer: 3.

See! $\Omega(n)$ time.

Algorithm A takes stream S
maintains number of distinct elements.
Is $x \in S$?
Add x, see if number of distinct elements change.
Must know subset of $[n]$
(at most n types)

2^n possibilities $\rightarrow$ requires $\Omega(n)$ bits!

Toy problem

Alg: Number of distinct elements
$\leq k$ Output: "no"
$\geq 2k$ Output: "yes"

Don't care if in between.

Randomized Algorithm:
(1) Choose random hash function.
$h : [n] \rightarrow [B]$, where $B = k$.
(2) If any $h(x_i) = 0$, say "yes", else "no".

$$\Pr[A(x) = No \mid N \leq k] = \left(1 - \frac{1}{B}\right)^N \geq \left(1 - \frac{1}{B}\right)^k$$

$$\Pr[A(x) = No \mid N > 2k] = \left(1 - \frac{1}{B}\right)^N \leq \left(1 - \frac{1}{B}\right)^{2k}$$

Constant gap (roughly $1/e - 1/e^2$).
Many trials, in parallel gives good result. ..more later.

Number of bits for random hash function?
k^n hash functions. $n \log k$ bits to specify!

2-wise independent hash functions

The family $\mathscr{H} : [n] \rightarrow [p]$
$h_{a,b}(x) = ax + b \mod p$, prime $p \geq n$, $a, b \in \{0, \ldots, p-1\}$
is 2-wise independent:

$$\Pr_{a,b}[h(x) = c \wedge h(y) = d] = \frac{1}{p^2} \qquad \forall x \neq y$$

Proof: If $h(x) = c$ and $h(y) = d$ then

$$ax + b = c \pmod{p} \qquad ay + b = d \pmod{p}$$

has unique solution for a, b since $(x - y) \neq 0$.
$\rightarrow$ One $h_{a,b}$ out of p^2 functions has $h(x) = c$ and $h(y) = d$. $\square$

Nonprime $|B| < p$.

$\mathscr{H} : [n] \rightarrow |B|$, $h_{a,b} = (ax + b) \pmod{p} \pmod{|B|}$

Approximately 2-wise indendependent.
Pr[collision at c and d] $\approx \frac{1}{|B|^2}(1 \pm \frac{k}{p})^2$ Assume $p >> 1$, so basically assume perfectly independent.

(k-wise independent hash family. degree k polynomials.)

Distinct elements with 2-wise hash functions.

N distinct items.

Toy Alg:

(1) Random hash h from $\mathscr{H}: [n] \to [4k]$.

(2) If $h(x_i) = 0$, say "yes", else say "no"

Union Bound: $Pr[A \cup B] \leq Pr[A] + Pr[B]$

$Pr[A_1 \cup A_2 \cup \cdots \cup A_N] \leq \sum_i Pr[A_i]$

$Pr[\text{"yes"} | N < k] \leq \sum_j Pr[\text{h(j)= 0}] \leq k(\frac{1}{4k}) \leq \frac{1}{4}$

Inclusion/Exclusion: $Pr[A \cup B] \geq Pr[A] + Pr[B] - Pr[A \cap B]$

$Pr[\cup A_i] \geq \sum_i Pr[A_i] - \sum_{i,j} Pr[A_i \cap A_j]$

$$\Pr[\text{"yes"} \mid N \geq 2k] \geq \frac{2k}{B} - \frac{2k.(2k-1)}{2B} \geq \frac{2k}{B}(1 - \frac{k}{B}) = (\frac{3}{8})$$

See this as one of two coins.

Either heads with prob $\leq \frac{1}{4}$

Either heads with prob $\leq \frac{3}{8}$

Gap of $\frac{1}{8}$.

Flip coin (in parallel) to pump up the ~~volume!~~ probability!

Core Alg: analysis cont.

$E[Z^2] = F_2$.

$Var(Z^2) = E[Z^4] - E[Z^2]^2 = 2\sum m_i^2 m_j^2 \leq 2F_2^2$

Close to expectation? $|Z^2 - \mu| \leq \varepsilon F_2$?

Chebyshev: $Pr[|X - \mu| > \Delta] \leq \frac{Var(X)}{\Delta^2}$

For Z^2, $Pr[|Z^2 - \mu| > \varepsilon F_2] \leq \frac{2F_2^2}{\varepsilon^2 F_2^2} = \frac{2}{\varepsilon^2}$

Uh oh. Bigger than one for $\varepsilon \leq 2$!

It gets better.

Simpl. Chernoff: Number of heads $\widehat{b}$ in $k = O(\frac{\log(1/\delta)}{\varepsilon^2})$ flips of bias b coin satisfies $bk(1-\varepsilon) \leq \widehat{b} \leq bk(1+\varepsilon)$ with probability $1-\delta$.

Alg:

"yes" with probability at most $1/4$ when $N < k$.

"yes" with probability at least $3/8$ when $N > 2k$.

Run $\Theta(\log \frac{1}{\delta})$ independent copies of Alg.

Output "yes" if more than $\frac{5}{16}$ yes's.

Use claim with $\varepsilon = \frac{1}{3}$.

$\to$ Correct with probability $\geq 1 - \delta$.

Run $\log n$ times to get within factor of two.

Factor of $(1+\varepsilon)$? Choose $|B| = \theta(\frac{k}{\varepsilon})$ in Alg.

"yes" with probability at most τ when $N < k$.

"yes" with probability at least $(1+\varepsilon)\tau$ when $N > (1+\varepsilon)k$.

Run $\frac{\log \frac{1}{\delta}}{\varepsilon^2}$ times to pump up the probability.

Run $\log_{1+\varepsilon} n$ times to get within factor of $1+\varepsilon$.

$O(\log n \log_{1+\varepsilon} n \frac{\log \frac{1}{\delta}}{\varepsilon^2})$ space, $(1 \pm \varepsilon)$ estimate, w/prob $1-\delta$.

Independent trials.

Run Core Alg k times. $Z_1, \ldots, Z_k$.

$(E[Z_i^2] = F_2 \; Var(Z_i^2) \leq 2F_2^2.)$

Output average. $Y = \frac{1}{k}\sum_i Z_i^2$

$E[Y] = \frac{1}{k}\sum E[Z_i^2] = F_2$

$Var(cX) = c^2 Var(X)$

$Var(X+Y) = Var(X) + Var(Y)$; independent X and Y

$Var(Y) = \frac{1}{k^2}\sum_i Var(Z_i^2) = \frac{2F_2^2}{k}$

$k = \frac{2}{\delta\varepsilon^2}$ and Chebyshev

$\Pr[|Y - \mu| \geq \varepsilon F_2] \leq \delta$

Space: $O(\frac{\log n}{\varepsilon^2 \delta})$.

Could get $O(\frac{\log n \log \frac{1}{\delta}}{\varepsilon^2})$ using a Central Limit Theorem.

Estimating F_2

Second Moment: $F_2 = \sum_j m_j^2$.

Core Alg:

(1) Random h from 4-wise ind. family $\mathscr{H}: [n] \to \pm 1$.

(2) Output $Z^2 = (\sum_i h(x_i))^2$

Show $E[Z^2] = F_2$.

$h(j) = Y_j$

$Z = \sum_{i \in [m]} h(x_i) = \sum_{j \in S} Y_j m_j$

$E[Z^2] = \sum_j E[Y_j^2] m_j^2 + \sum_{i,j} E[Y_i] E[Y_j] m_i m_j = \sum_i m_i^2 = F_2$

Show good probability of success? Calculate variance.

$Var(X) = E[X^2] - (E[X])^2$

$E[Z^4] = \sum_i E[Y_i^4 m_i^4] + 3\sum_{i,j} E[Y_i^2 Y_j^2 m_i^2 m_j^2] = \sum_i m_i^4 + 6\sum_{i,j} m_i^2 m_j^2$

$Var(Z^2) = E[Z^4] - E[Z^2]^2 = 2\sum m_i^2 m_j^2 \leq 2F_2^2$

See you on Tuesday.