Streaming:

Models. — Turnstile vs Cash Register
— Streaming vs Semi-streaming
— One pass vs. Multipass.

- Counter
- $F_0, F_2 \dots$
- heavy hitters
- median.

AMS: F_p: $\boxed{F_p := \sum_i' (x_i(t))^p.}$ $\boxed{x_i(t) = \text{frequency vector @ time } t.}$

$C = \sum_{i=1}^{|U|} x_i h_i$ output C^2. h_i: 4-universal $U \to \{-1, 1\}$.

Note: $EC^2 = \sum_i' x_i^2$. by unbiasedness & 2-indep.

$VC^2 = E(C^2)^2 - (EC^2)^2$

$EC^4 = \sum_i' x_i^4 + 6\sum_{i \neq j}' x_i^2 x_j^2$ $\Rightarrow$ $VC^2 = 4\sum_{i<j}' x_i^2 x_j^2 \le 2(EC^2)$

variance same order as expectation.

good news.

Chebyshev:

$Pr\left[\,|C^2 - EC^2| \geq \varepsilon EC^2\right] \le \frac{Var(C^2)}{\varepsilon^2 EC^2} \le 2/\varepsilon^2$

but repeat. $Var(Est) \le k$. orig variance | or average (k time)
k times indep. Expect $\le k$. orig expect. | $Var \le \frac{1}{k}$ orig var
Exp = same.

$\Downarrow$

tail prob $\le \frac{2k}{\varepsilon^2 k^2}$ $\Rightarrow$ set $k = \frac{2}{\varepsilon^2} \cdot \frac{1}{\delta}$. ✓

Same as $\overset{\leftarrow u \rightarrow}{\boxed{A}}\ (x)$ with A having k rows, where rows A drawn from 4-wise indep family $\in \{\pm \frac{1}{\sqrt{k}}\}$

output $(Ax)^2$

And $\frac{2}{\varepsilon^2} \cdot \frac{1}{\delta}$ rows suffice. to show that $\|Ax\|^2 \approx \|x\|^2$. (JL!).

Tug of War sketch: get $E\langle Sx, Sy\rangle = \langle x, y\rangle$
$\forall x,y$ $Var(\langle Sx, Sy\rangle) \le \frac{2}{k}\|x\|^2\|y\|^2$.

$AB \approx \underbrace{AS^T}\,\underbrace{SB}$.

E ij^{th} entry $= \langle Sa_i, Sb_j\rangle = \langle a_i, b_j\rangle = (AB)_{ij}$ so its like ToW sketch.

F_0: exact requires $\Omega(\min(N, U))$ space.

apx: several ways.

Use the Flajolet Martin Idea:- smallest elt $\cong \frac{1}{d+1}$ $\Rightarrow$ output $\frac{1}{\text{smallest}} - 1$.

right expectation
high variance.

So instead:- keep track of s smallest values seen

(say hash fn goes from U to $[0,1]$). $\leftarrow$ $\boxed{\text{actually to } [0, M]}$ ~~but that's another detail for later.~~

Intuition: sth smallest element @ $\cong \frac{sM}{d}$. $= L \in [0, M]$.
(if d distinct elts)

$\Rightarrow$ output $\frac{Ms}{L}$. (slightly off — but almost OK).

What about error prob?

Claim: $\Pr$(est off by factor of $(1+\varepsilon)$) $\leq O(1/s)$.

Pf: for estimator too small.

So say $\underbrace{\frac{Ms}{L}}$ $< d/2$ $\Rightarrow$ position of sth smallest hash $> \frac{2Ms}{d}$

estimator, where $L \in [0, M]$ is position of sth smallest hash value.

$\Rightarrow$ # elts smaller than this $\leq s$

but expected # $= 2s$.

$\Rightarrow$ we fell below expectation by factor of 2.

Calculations in old Lecture notes seem incorrect. Fix!

Heavy hitters: Count Min sketch.

- Misra Gries better analysis. Sps keep more counters. (Want k heavy hitters. Keep ℓ regs).

Output $\hat{f}_1, \hat{f}_2 \ldots \hat{f}_k$. (Say n arrivals overall.)

Note: if d decrements, sum of counters $= n - d\cdot\ell$

$$= \sum_{i=1}^{\ell} \hat{f}_i \geq \sum_{i=1}^{\ell} (f_i - d) \geq \sum_{i=1}^{k} (f_i - d)$$

non neg of counters.

$\Rightarrow d = \frac{F_{[k+1..n]}}{\ell - k}$ $\Rightarrow$ our loss $\leq F_{[k+1..n]} + d\cdot k = (1 + \frac{k}{\ell - k})F_{k+1..n}$. $= F_{[1..k]} - kd$.