Streaming Algorithms

Stream: $A = \langle a_1, a_2, \ldots, a_m \rangle$
- $a_i$ in $[n]$ size $\log n$
Compute $f(A)$ in $\text{poly}(\log m, \log n)$ space

Let $f_j = |\{a_i \in A \mid a_i = j\}|$

$F_1 = \sum_j f_j = m$ == total count
$F_2 = \sqrt{\sum_j f_j^2}$ == RMS count

Goal:

deps-FREQUENCY-ESTIMATION: Build data structure $S$.
For any $j$ in $[n]$, $\hat{f}_j = S(j)$ s.t.
- $f_j - \text{eps} \cdot F_1 \leq \hat{f}_j \leq f_j$       (MG)
- $f_j \leq \hat{f}_j \leq f_j + \text{eps} \cdot F_1$     (CMS (today))
- $|f_j - \hat{f}_j| \leq \text{eps} \cdot F_2$        (CS (maybe))

aka deps-approximate phi-HEAVY-HITTERS:
- Return all $f_j$ s.t. $f_j > \phi$
- Return no $f_j$ s.t. $f_j < \phi - \text{eps} \cdot m$

Count-Min Sketch [Cormode + Muthukrishnan '05]

t independent hash functions $\{h_1, \ldots, h_t\}$
each $h_i : [n] \rightarrow [k]$
2-d array of counters:
\[ h_1 \rightarrow [C_{1,1}] [C_{1,2}] \ldots [C_{1,k}] \]
\[ h_2 \rightarrow [C_{2,1}] [C_{2,2}] \ldots [C_{2,k}] \]
\[ \ldots \quad \ldots \quad \ldots \]
\[ h_t \rightarrow [C_{t,1}] [C_{t,2}] \ldots [C_{t,k}] \]

for each \( a \in A \rightarrow \) increment \( C_{i,h_i(a)} \) for \( i \in [t] \).

\[ \hat{f}_a = \min_{i \in [t]} C_{i,h_i(a)} \]

Set \( t = \log(1/\delta) \)
Set \( k = 2/\varepsilon \)

-----------------------------

Clearly \( f_a \leq \hat{f}_a \)

\( \hat{f}_a \leq f_a + W. \) What is \( W? \)

One hash function \( h_i \).
Adds to \( W \) when there is a collision \( h_i(a) = h_i(j) \). wp 1/k

random variable \( Y_{i,j} \)
\( Y_{i,j} = \{f_j \text{ wp } 1/k, 0 \text{ wp } 1-1/k\} \)
\( E[Y_{i,j}] = f_j/k \)

random variable \( X_i = \sum_{j \in [n], j \neq a} Y_{i,j} \)
\( E[X_i] = E[\sum_j Y_{i,j}] = \sum_j f_j/k = F_1/k = \varepsilon * F_1/2 \)

+++++++++++++++++++++++++++++++++
Markov Inequality
X a rv and a>0
Pr[|X| >= a] <= E[|X|]/a

+++++++++++++++++++++++

X_i > 0 so |X_i| = X_i
setting a = eps F_1 then
E[|X|]/a = (eps*F_1 /2)/(eps F_1) = 1/2
Pr[X_i >= eps F_1] <= 1/2

-----

Now for t *independent* hash functions:

Pr[hat{f}_a - f_a >= eps F_1]
= Pr[min_i X_i >= eps F_1]
= Pr[forall_{i in [t]} (X_i >= eps F_1)]
= Prod_{i in [t]} Pr[X_i >= eps F_1]]
<= 1/2^t
= delta  (since t = log(1/delta) )

Hence:
\[ f_a <= hat{f}_a <= f_a + eps F_1 \]
- first inequality always holds
- second inequality holds wp > 1-delta

-----
Space:
each of k*t counters requires log m space
O(k*t*log m)
Store t hash functions:  log n each
O((k log m + log n)*t) = O((1/eps) log m + log n) log (1/ delta))

-------------------------------
turnstile model: add or subtract (as long as is there)

Count Sketch:

t independent hash functions \{h_1, ..., h_t\}
each \(h_i : [n] \rightarrow [k]\)

t independent secondary hash functions \{g_1, ..., g_t\}
each \(g_i : [n] \rightarrow \{-1, +1\}\)

2-d array of counters:
\(h_1 \rightarrow [C_{1,1}], [C_{1,2}], \ldots, [C_{1,k}]\)
\(h_2 \rightarrow [C_{2,1}], [C_{2,2}], \ldots, [C_{2,k}]\)
\[ \ldots \]
\(h_t \rightarrow [C_{t,1}], [C_{t,2}], \ldots, [C_{t,k}]\)

for each \(a \in A\) -> adds \(g_i(a)\) to \(C_{i,h_i(a)}\) for \(i\) in \([t]\).

\(\hat{f}_a = \text{median}_{i \in [t]} C_{i,h_i(a)}\)

Set \(t = 2 \cdot \log(1/\delta)\)
Set \(k = 4/\epsilon^2\)

One hash function pair \(h_i, g_i\).
\(E[\hat{f}_a] = g_i(a) f_a\)
random variable: \( Y_{i,j} \) expected error caused by \( f_j \) on \( \hat{f}_a \)
\[ Y_{i,j} = \{ f_j \text{ wp } 1/2k, -f_j \text{ wp } 1/2k, 0 \text{ wp } 1-1/k \} \]

random variable: \( X_i \) expected error of \( \hat{f}_a \)
\[ X_i = \sum_j Y_{i,j} \]
\[ E[X_i] = 0 \]

\( Y_{i,j} \) pairwise independent, so
\[ \text{Var}[X] = \sum_j \text{Var}[Y_{i,j}] \]
\[ \text{Var}[Y_{i,j}] = E[Y_{i,j}^2] - E[Y_{i,j}]^2 \]
\[ = E[Y_{i,j}^2] \]
\[ = f_j^2 / k \]

\[ \text{Var}[X_i] = \sum_j f_j^2/k \leq F_2^2/k. \]

Chebyshev's Inequality:

\[ X \text{ a rv and } b>0 \]
\[ \Pr[|X-E[X]| \geq b] \leq \text{Var}(X)/b^2 \]

using \( b = \varepsilon F_2 \)
\[ \Pr[|X_i| \geq \varepsilon F_2] \leq (F_2^2/k) / (\varepsilon F_2)^2 \]
\[ = 1/(k \times \varepsilon^2) \leq 1/4 \]

since \( k = 4/\varepsilon^2 \)

-------------------------------------------

\( t \) *independent* hash function pairs:

Recall: \( \hat{f}_a = \text{median}_i \{(f_a + X_i)/g_i(a)\} \)
\[ \Pr[|f_a - \hat{f}_a| < \epsilon F_2] = \Pr[\text{median}_i X_i > \epsilon F_2] \leq 2 \cdot \Pr[\{i \in [t]\} (X_i \geq \epsilon F_2)] \leq 2 \cdot \prod_{i \in [t/2]} \Pr[X_i \geq \epsilon F_2] \leq 2 \cdot \frac{1}{4^{t/2}} \leq \delta \quad \text{(since } t = 2\log(1/\delta) \text{ )} \]

-----

Space:
each of \(k \cdot t\) counters requires \(\log m\) space
\(O(k \cdot t \cdot \log m)\)

Store \(t\) hash function pairs: \(\log n\) each
\(O((k \cdot \log m + \log n) \cdot t)\)
\(= O((1/\epsilon^2) \log m + \log n) \log (1/\delta))\)

--------------------------
--------------------------

CMS: \(\epsilon F_1\) error
space \(O((1/\epsilon) \log m + \log n) \log (1/\delta))\)

CS: \(\epsilon F_2\) error
space \(O((1/\epsilon^2) \log m + \log n) \log (1/\delta))\)

\(F_2 < F_1\) (generally), but \(1/\epsilon \ll 1/\epsilon^2\)
CMS very practical because of only \((1/\epsilon)\) term.