Lecture 15

History:

- 83 FM. very complicated
- 96 AMS - sketches

Alon, Matias, Szegedy
- F_0 # of distinct values
- General F_k alg for $k > 1$
- Special F_2 alg

A, M, S - Gibbons 96
- Joins (size of join)

Frequency moments

Estimate:

$F_k = \sum_i f_i^k$ with f_i frequency of elements appearing in a stream

F_0 - # of distinct values

F_1 - total # of elements

$F_2 = \sum_i f_i^2$ - measure of uniformity

$f_i = \bar{f}$

$F_1 = \bar{f} n$) uniform distrib

$F_2 = \bar{f}^2 n = \frac{F_1^2}{n}$

$$\boxed{\begin{aligned} V_q(\text{data}) &= \sum_i (\bar{f} - f_i)^2 \\ &= \underbrace{\sum_i f_i^2}_{F_2} - \bar{f} \underbrace{\sum_i f_i}_{\bar{f} n} \\ &= F_2 - \frac{F_1^2}{n} \end{aligned}}$$

$f_1 = N$, $f_{2 \ldots n} = 0$) $F_1 = N$, $F_2 = F_1^2$) Always $F_1^2 \geq F_2 \geq \frac{F_1^2}{n}$

General F_k alg.

- random counts
  - decide randomly when to start watching the stream
    - almost

$\tilde{f}_a$ - est of freq of a

1 word

(f_a is number of ... stream)

$X = \tilde{f}_a^k - \tilde{f}_a^{k-1}$

$E[X] = F_k$

1 2 7 1 3 2 1 4 2

AMS-sketches

$F_2 = \sum_i f_i^2$

Introduce for every $i \in I$ a random variable ξ_i

I	1	2	3	4	5
ξ_i	-1	1	1	-1	1

$\xi_i \in \{-1, 1\}$

$E[\xi_i] = 0$

ξ_i r.v. 4-wise indep

$X = \sum_i f_i \xi_i = \sum_{i \in \text{stream}} \xi_i$ - 1 word

1	2	2	3	1	1	5

$X = \xi_1 + \dots$

$f_1 \xi_1 + f_2 \xi_2 \dots$

$\sum f_i \xi_i$

Claim

$Z = X^2$ approximates F_2

$z = x^2$

$x = \sum_i f_i \xi_i$

$E\{z\} = E\{x^2\} = E\left\{\sum_i f_i \xi_i \sum_{i'} f_{i'} \xi_{i'}\right\}$

$= \sum_i \sum_{i'} E\{\xi_i \xi_{i'}\} f_i f_{i'}$

key

$$E\{\xi_i \xi_{i'}\} = \begin{cases} 1 & i = i' \\ 0 & i \neq i' \end{cases} = \delta_{ii'}$$

pairwise indep

$E\{z\} = \sum_i \sum_{i'} \delta_{ii'} f_i f_{i'}$

$= \sum_i f_i^2 = \Gamma_2$ ✓

$V_a(z)$

$E\{z^2\} = E\{x^4\}$

$= \sum_{i_1} \sum_{i_2} \sum_{i_3} \sum_{i_4} E\{\xi_{i_1} \xi_{i_2} \xi_{i_3} \xi_{i_4}\} f_{i_1} f_{i_2} f_{i_3} f_{i_4}$

4-wise indep is necessary to make $V_a(x)$ small

$$E\{\xi_{i_1} \xi_{i_2} \xi_{i_3} \xi_{i_4}\} = \begin{cases} \xi_{i_1}^2 \xi_{i_3}^2 & i_1 = i_2 \wedge i_3 = i_4 \\ & i_1 = i_3 \wedge i_2 = i_4 \\ & i_1 = i_4 \wedge i_2 = i_3 \\ 0 & \text{oth} \end{cases}$$

$E[\delta_{i_1}\delta_{i_2}\delta_{i_3}\delta_{i_4}] =$

$\delta_{i_1 i_2} \cdot \delta_{i_3 i_4} +$

$\delta_{i_1, i_3} \cdot \delta_{i_2, i_4} +$

$\delta_{i_1, i_4} \cdot \delta_{i_2, i_3} -$

$-2\,\delta_{i_1, i_2}\,\delta_{i_2, i_3}\,\delta_{i_3, i_4}$

$E[Z^4] \leq \sum f_{i_1}^2 f_{i_3}^2 - \sum\sum f_i^2 f_i^2 \ldots$

$-2\sum_i f_i^4 = 3F_2^2 - 2\sum_i f_i^4 \leq 3F_2^2$

$V_Z(Z) \leq 2F_2^2$

$STA(Z) \leq \sqrt{2}\cdot F_2$

$Z_1 \ldots Z_{1000}$

$\bar{Z} = \frac{\sum Z_i}{1000}$

$i_1 = i_2 = i_3 = i_4$