STREAMING ALGORITHMS

Algorithm Design II
Lecture notes by Rasmus Pagh

McG: BOOK CHAPTER BY ANDREW McGREGOR

PLAN:
- STREAMING MODEL [McG SECT. 0]
- HEAVY HITTERS
  - RESERVOIR SAMPLING [McG 1.2] + CHERNOFF BOUND
  - MISRA-GRIES [McG 1.1]
  - COUNT-MIN [McG 3.1], PAIRWISE INDEPENDENCE.
- DISTINCT ELEMENTS
  - SAMPLING DOES NOT WORK
  - k-MIN SUMMARIES
  - (- SKETCHES [McG 2.1])
  - (- APPLICATION: 4 DEGREES OF SEPARATION)

COLLECT "PAGE STREAM"

STREAMING MODEL:

SIMPLE FORM: SEQUENCE OF NUMBERS $x_1, x_2, \ldots$ e.g.

(GENERAL FORM: SEQUENCE (i_t, Δ_t) OF UPDATES TO ENTRY i_t, ADDING Δ_t FOR $t=1,2,\ldots$)

GOAL: ANSWER QUESTIONS ABOUT THE MULTISET (OR VECTOR) USING SMALL SPACE, S.

TODAY:
- WHAT ARE THE HEAVIEST ITEMS (HEAVY HITTERS)
- HOW MANY DISTINCT ITEMS ARE THERE

RESERVOIR SAMPLING

SAMPLE OF $X_1, \dots, X_s$ IS SIMPLY $\{X_1, \dots, X_s\}$.

AFTER RECEIVING X_{t+1}, SAMPLE IT WITH PROBABILITY $\frac{s}{t+1}$, REPLACING A RANDOM ELEMENT IN THE SAMPLE.

CLAIM: MAINTAINS A RANDOM SUBSET OF SIZE s FROM $X_1, \dots, X_t$.

Q: WHAT IS THE PROBABILITY THAT WE SAMPLE AN ITEM x WITH FREQUENCY f? LET $Z_i = \begin{cases} 1 & \text{IF } X_i \text{ IS SAMPLED AND } X_i = x \\ 0 & \text{OTHERWISE} \end{cases}$.

NUMBER OF SAMPLES IS $Z = \sum_{i=1}^{t} Z_i$.

$$E[Z] = \sum_{i, X_i = x} E[Z_i] = ft \frac{s}{t} = fs.$$

$$\mathrm{Var}(Z) = E[(Z - E[Z])^2] = E[Z^2] - E[Z]^2$$

$$= \sum_{i=1}^{t} E[Z_i^2] + 2 \sum_{i<j} E[Z_i Z_j] - (fs)^2$$

$$= fs + 2 \frac{s}{t} \frac{(s-1)}{t} \cdot \binom{ft}{2} - (fs)^2 \leq fs.$$

CHEBYSHEV'S INEQUALITY: $\Pr[|Z - E[Z]| > t] < \frac{\mathrm{Var}(Z)}{t^2}$

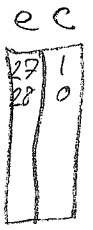
A: $\Pr[Z = 0] < \frac{fs}{(fs)^2} = 1/fs$.

MISRA-GRIES

STORE ELEMENT-COUNTER PAIRS $(e_1, c_1), \ldots, (e_s, c_s)$.

WHEN SEING x_i:

- IF $x_i = e_j$, INCR. c_j
- ELSE IF $c_j = 0$, SET $c_j = 1$, $e_j = x_i$
- ELSE DECREMENT ALL c_j BY 1.

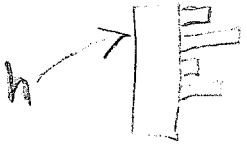


LEMMA: $\hat{x}_i = \begin{cases} c_j & \text{if } e_j = x_i \\ 0 & \text{otherwise} \end{cases}$ DIFFERS AT MOST $\frac{t-1}{s-1}$ FROM THE FREQUENCY f_t OF x_i.

PROOF: CASE 3 CAN HAPPEN AT MOST $\frac{t}{s-1}$ TIMES, x_i NOT ADDED

COUNT-MIN

IDEA IS A "RANDOMIZED HISTOGRAM", WHERE A HASH FUNCTION h PARTITIONS THE ITEMS. HOW MUCH DO WE OVERCOUNT?



MARKOV'S INEQUALITY:

$E[\text{ITEMS IN BUCKET } h(x)] = \frac{t}{s}$

$\Rightarrow \Pr[\text{ITEMS IN BUCKET } h(x) > \frac{2t}{s}] < \frac{1}{2}$

ADVANTAGE:
- DELETIONS
- WEIGHTS

TO REDUCE ERROR PROB., REPEAT AND TAKE MIN.

OBS: ENOUGH TO HAVE PAIRWISE INDEP: $\forall x, y$, $\Pr[h(x) = i \mid h(y) = j] = 1/s^2$.

E.G. THINK OF X AS BINARY VECTOR, TAKE RANDOM MATRIX $A \in \mathbb{R}^{\log s \times r}$ AND COMPUTE $h(y) = Ax$.

DISTINCT ELEMENTS

SAMPLING DOES NOT WORK WELL. TO DISTINGUISH BETWEEN $\sqrt{n}$ ELEMENTS OCCURRING ONCE, AND OCCURRING TWICE, NEED $\Omega(\sqrt{n})$ SAMPLES.

MIN-HASH SUMMARY

TAKE PAIRWISE INDEP. HASH FUNCTION $h: N \to [0, 1]$.

STORE s ITEMS HAVING THE SMALLEST HASH VALUES.

IF v IS THE sTH SMALLEST ESTIMATE # DISTINCT ITEMS AS $\frac{k}{v}$.

INTUITION:

ANALYSIS: $X_i = \begin{cases} 1 & \text{if } h(k_i) < \frac{k}{2d} \\ 0 & \text{otherwise} \end{cases}$ $X = \sum X_i$, $E[X] = \frac{k}{2}$

$$\Pr[\text{ESTIMATE 2x TOO LARGE}] \leq \Pr[X \geq k] < \frac{\text{Var}(z)}{(k/2)^2} < \frac{2}{k}.$$

SIMILARLY FOR EST 2x TOO SMALL.

COLLATZ 81

4-DEGREES OF SEPARATION