Lecture 1 :- Saving memory.

Heavy hitters:- "Detect abnormally high frequency of 'requests' from a small number of sources."

Example :- - Too many requests to UPI from one phone number
- Too many requests to a website from one IP address.

Possible solution :- 'Maintain a frequency table'
- Keys : Idenfiers : phone numbers / IP addresses
- Value : How many requests with this identifier have been received recently?

Q: What is the signature of heavy hitting with such a table?

"Few entries account for most of the total frequency".

One possible way of detecting this :-

Comapare $\sum_{i=1}^{M} f_i =: F_1$ against $\sum_{i=1}^{M} f_i^2 := F_2$

where f_i is the frequency of the ith identifier,
M : number of distinct identifiers.

Possible formalization of heavy-hitting: $F_2 \gg F_1$.

Note :- F_1 is easy to compute: just need a single counter: $O(\log F_1)$ (Will write $N = F_1$ sometimes).

What about F_2?

- If the frequency-table is stored, can compute F_2 in $O(M\,\mathrm{poly}(\log(N)))$ time.
(or iteratively in $O(\log N)$ time per update).

Storage for the frequency table could be $\Omega(M \log N)$ if we use the standard hash table / Balanced binary map style storage.

Eg. with IPv4 : $M \approx 4\times10^9$
Suppose each count is stored as a standard 4 byte integer.
Total storage = 16 GB.

$\approx$ Order of memory available on a single node.

[Q] Can we reduce the storage for F_2-computation perhaps at the cost of an approximation factor?

Comments:-

[The solution we will discuss is quite beautiful, and is from Alon, Mattias and Szegedy '96.]

[However, as far as I know, engineering solutions typically use diffrent heuristics.

↳ Because they want very fast updates, low memory footprint, and they are happy to use other heuristic signatures of heavy hitting]

↳ (*) POSSIBLE PROJECT IDEA.

AMS algorithm :- Randomized estimator for F_2.

Suppose $\varepsilon_i \in_{uar} \{-1, +1\}$, $1 \le i \le M$, ε_i are independent. are given. Then one needs just one more counter to get such an estimator.

$$Z = \sum_{i=1}^{M} \varepsilon_i f_i$$!! One counter !!

$$Y = Z^2$$

$$E[Z] = E\left[\sum_{i=1}^{M} \varepsilon_i f_i\right]$$

$$= \sum_{i=1}^{M} f_i \, E[\varepsilon_i]$$ (Linearity of expectation)

$$= 0.$$

$$E[Y] = E\left[\left(\sum_{i=1}^{M} f_i \varepsilon_i\right)^2\right]$$

$$= E\left[\sum_{i=1}^{M} f_i^2 \varepsilon_i^2 + 2\sum_{i=1}^{M}\sum_{j=1}^{i-1} f_i f_j \varepsilon_i \varepsilon_j\right]$$

$$= \sum_{i=1}^{M} f_i^2 + 2\sum_{i=1}^{M}\sum_{j=1}^{i-1} f_i f_j \, E[\varepsilon_i \varepsilon_j]$$

(Linearity of expectation)

Assuming $(\varepsilon_i)_{i=1}^{M}$ are pairwise-independent.

$$\left(E[\alpha(\varepsilon_i)\beta(\varepsilon_j)] = E[\alpha(\varepsilon_i)]\, E[\beta(\varepsilon_j)] \quad \forall\, i \ne j,\ \alpha, \beta \text{ 'reasonable' fns.} \right)$$

we have $E[\varepsilon_i \varepsilon_j] = E[\varepsilon_i] E[\varepsilon_j] = 0$
$\forall i \neq j$

Therefore :-

$$E[Y] = \sum_{i=1}^{M} f_i^2 = F_2$$

But this is only in expectation.

$$\left[T = \begin{cases} -10^9 & \text{w.p. } 1/2 \\ +10^9 & \text{w.p. } 1/2 \end{cases} \quad E[T] = 0 \text{, but } T \text{ is never close to zero!!} \right]$$

We want to say that Y (or some modification of Y) does not have such a pathology. We will check the variance of Y

$$\left[\mathrm{Var}(T) = 10^{18} \right]$$

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

For example if $\Pr\left[|Z - E[Z]| > 10 |E[Z]| \right] > \frac{1}{2}$

then $\mathrm{Var}(Z) \geqslant 50\, E[Z]^2$.

"Large deviations from mean with non-negligible probability"
$\Longrightarrow$ "Large" variance.

(Note

$$\begin{aligned} \mathrm{Var}[Z] &= E\left[(Z - E[Z])^2 \right] \\ &= E\left[Z^2 - 2 \cdot Z \cdot E[Z] + E[Z]^2 \right] \\ &= E[Z^2] - E[Z]^2 \end{aligned}$$

)

Let's check the variance of Y.

$$\mathrm{Var}(Y) = E[Y^2] - E[Y]^2$$
$$= E[Y^2] - F_2.$$

$$E[Y^2] = E[Z^4]$$
$$= E\left[\left(\sum_{i=1}^{M} f_i \varepsilon_i\right)^4\right]$$
$$= E\Bigg[\sum_{i=1}^{M} f_i^4 \varepsilon_i^4 + 4 \sum_{i=1}^{M} \sum_{\substack{j=1 \\ j\neq i}}^{M} f_i^3 f_j \varepsilon_i^3 \varepsilon_j$$
$$+ 6 \sum_{i=1}^{M} \sum_{j=1}^{i-1} f_i^2 f_j^2 \varepsilon_i^2 \varepsilon_j^2$$
$$+ 12 \sum_{i=1}^{M} \sum_{j=1, j\neq i}^{M} \sum_{k=1}^{j-1} f_i^2 f_j f_k \varepsilon_i^2 \varepsilon_j \varepsilon_k$$
$$+ 24 \sum_{i=1}^{M} \sum_{j=1}^{i-1} \sum_{k=1}^{j-1} \sum_{l=1}^{k-1} f_i f_j f_k f_l \varepsilon_i \varepsilon_j \varepsilon_k \varepsilon_l\Bigg]$$

$$= \sum_{i=1}^{M} f_i^4 + 6 \sum_{i=1}^{M} \sum_{j=1}^{i-1} f_i^2 f_j^2$$
$$+ E[\text{Blue} + \text{Orange} + \text{Pink}]$$

Blue terms: 0 expectation when $(\varepsilon_i)_{i=1}^{M}$ are pairwise independent.

Orange terms: 0 expectation when $(\varepsilon_i)_{i=1}^{M}$ are 3-wise independent.

Pink terms: 0 expectation when $(\varepsilon_i)_{i=1}^{M}$ are 4-wise independent.

k-wise independence of random variables $X_1, X_2, \dots X_n$:
$\forall$ distinct $i_1, i_2, \dots i_k \in [n]$

$$E\left[\left(\prod_{j=1}^{k} \alpha_j(X_{i_j})\right)\right] = \prod_{j=1}^{k} E[\alpha_j(X_{i_j})]$$

$\alpha_1 \dots \alpha_k$ are 'reasonable' fns.

Therefore :- if $(\varepsilon_i)_{i=1}^{M}$ are 4-wise independent, then,

$$E[Y] = F_2 = \sum_{i=1}^{M} f_i^2$$

$$E[Y^2] = \sum_{i=1}^{M} f_i^4 + 6\sum_{i=1}^{M}\sum_{j=1}^{i-1} f_i^2 f_j^2 .$$

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

$$= \sum_{i=1}^{n} f_i^4 + 6\sum_{i=1}^{M}\sum_{j=1}^{i-1} f_i^2 f_j^2 - \sum_{i=1}^{M} f_i^4 - 2\sum_{i=1}^{M}\sum_{j=1}^{i-1} f_i^2 f_j^2$$

$$= 4\sum_{i=1}^{M}\sum_{j=1}^{i-1} f_i^2 f_j^2$$

$$= 2\sum_{i=1}^{M}\sum_{\substack{j=1;\\ j\neq i}}^{M} f_i^2 f_j^2$$

$$\leq 2\sum_{i=1}^{M}\sum_{j=1}^{M} f_i^2 f_j^2 = 2\left(\sum_{i=1}^{M} f_i^2\right)\left(\sum_{j=1}^{M} f_j^2\right)$$

$$= 2F_2^2$$

$$\mathrm{Var}(Y) \leq 2F_2^2 .$$

Markov / Chebyshev argument: Let S be a non-negative random variable. Let $\lambda > 0$.

$$E[S] = E[S(I[S>\lambda] + I[S\le\lambda])]$$

$$= E[S \cdot I[S>\lambda]] + \underbrace{E[S\cdot I[S\le\lambda]]}_{\ge 0.\ (\because S \ge 0.)}$$

$$\ge E[S\cdot I[S>\lambda]]$$

Now, $S\cdot I[S>\lambda] \ge \lambda I[S>\lambda]$

$$\ge \lambda\, E[I[S>\lambda]]$$

$$= \lambda \Pr[S>\lambda].$$

$\Leftrightarrow$ $\Pr[S>\lambda] \le \frac{1}{\lambda} E[S]$ for every $\lambda > 0$ and S a non-negative random variable.

Let's apply this to $S = (Z - E[Z])^2$. Then for any $\lambda > 0$, this gives

$$\Pr[(Z-E[Z])^2 > \lambda^2] \le \frac{1}{\lambda^2} E[(Z-E[Z])^2]$$

$$\Leftrightarrow \Pr[|Z - E[Z]| > \lambda] \le \frac{\mathrm{Var}(Z)}{\lambda^2}.$$

(Chebyshev inequality)

So for Y, $\left[E[Y] = F_2 \quad Var(Y) \leq 2F_2^2 \right.$

$$Pr\left[|Y - F_2| > \alpha F_2 \right] \leq \frac{Var(Y)}{\alpha^2 F_2^2}$$

$$\leq \frac{2}{\alpha^2}.$$

So we at least get

$$Pr\left[Y \geq 3F_2 \right] \leq \frac{1}{2}.$$

But we only get anything interesting when $\alpha > 1$.

But then, Y could be zero!!

Want to reduce variance :-

Keep $Y_1, Y_2, \ldots, Y_s$ independent copies of Y. and let the final estimator G be

$$G = \frac{1}{s} \sum_{i=1}^{s} Y_i$$

Then

$$E[G] = E[Y_1] = F_2$$

$$Var[G] = \frac{1}{s} Var[Y_1] \leq \frac{2F_2^2}{s}.$$

So, for $\varepsilon > 0$, by choosing $s \geq \left\lceil \frac{16}{\varepsilon^2} \right\rceil$, we get

$$Pr\left[|G - F_2| > \varepsilon F_2 \right] \leq \frac{1}{8}.$$