ONLINE MASTERS IN **DATA SCIENCE**

DSC 257R - UNSUPERVISED LEARNING

HEAVY HITTERS

SANJOY DASGUPTA, PROFESSOR

UC San Diego

COMPUTER SCIENCE & ENGINEERING
HALICIOĞLU DATA SCIENCE INSTITUTE

The Heavy Hitters Problem

- Stream of data $x_1, x_2, \dots, x_m \in \mathcal{X}$ (where m is unknown)
- Parameter $\epsilon \in (0,1)$

Keep track of elements whose frequency is $\geq \epsilon m$,
as well as their frequencies.

How much space is needed to write down this information, if $n = |\mathcal{X}|$?

Can we design an online algorithm that uses only this much space?

Misra-Gries Algorithm

Data Structure:

- Hash table T of the size $k = 1/\epsilon$
- Each element $x \in T$ has an associated value $V[x] \in \{1, 2, \dots\}$

Algorithm:

- Table T is initially empty
- For $t = 1, 2, \dots$:
  - Get x_t
  - If $x_t \in T$: increment $V[x_t]$
  - Else: If $|T| < k$: Add x_t to T, with $V[x_t] = 1$
  - Else: for each $x \in T$:
    - Decrement $V[x]$
    - If $V[x] = 0$, remove x from T

Algorithmic Guarantees

Suppose that the number of times x appears in $x_1, \ldots, x_t$ is $\mathrm{freq_t(x)}$.

Claim. The following is true at all times t, for every $x \in \mathcal{X}$:

$$\mathrm{freq}_t(x) - t/(k+1) \leq V[x] \leq \mathrm{freq}_t(x).$$

(Take $V[x] = 0$ for any $x \notin T$.)

Key idea:

- Think of as $V[x]$ holding the number of occurrences of each item x
- Once in a while, $k+1$ of these values are decremented
- By time t, the maximum number of such decrement-steps is $t/(k\ +\ 1)$