CS5120: Probability & Computing **Jan-May 2020**

Lecture 15
Heavy hitters and the Misra-Gries algorithm

Lecturer: N.R.Aravind *Scribe: N.R.Aravind*

Problem 1 *Given a stream $x_1, \ldots, x_n$ with each $x_i \in \{1, 2, \ldots, m\}$, and a positive integer k, find all elements in the stream that occurred at least $n/(k+1)$ times.*

The Misra-Gries algorithm for the above problem maintains k counters, and along with each counter, an element is stored, whose approximate count is being stored in that counter. Whenever an element is seen, the algorithm attempts to increment its counter, failing which it decrements all existing counters.

Misra-Gries MG(k):

1. Initialize elements $E_1, \ldots, E_k$ to be null, and counters $C_1, \ldots, C_k$ to zero.
2. For $i = 1$ to n do:
3. If x_i is equal to E_j for some j, then increment C_j.
4. Else if there is an E_j which is null (with corresponding counter $C_j = 0$), then set $E_j = x_i$ and $C_j = 1$.
5. Else decrement all C_j. If any C_j becomes zero, mark the corresponding E_j as null.
6. End For

The following claim guarantees that all elements with frequency at least $n/(k+1)$ will have a non-zero counter at the end.

Claim 1 *If $E_j = x$, and $f(x)$ is the frequency of x in the stream, then* $f(x) - \dfrac{n}{k+1} \leq C_j \leq f(x)$.

We now give a proof of the claim. Let $C(x)$ denote the counter of x, $a(x)$ be the number of times that x was seen in the stream without incrementing $C(x)$,

and $b(x)$ be the number of times that $C(x)$ was decremented. Then we have: $C(x) = f(x) - a(x) - b(x)$. Thus, it suffices to show that $a(x) + b(x) \leq \frac{n}{k+1}$.

To see this, consider the following version of Misra-Gries instead, which maintains a muti-set of the elements (that is, allowing multiple copies of the same element).

1. Initialize S to be the empty-set.
2. For $i = 1$ to n do:
3. If $x_i \in S$, then add (a copy of) x_i to S.
4. Else if there are less than k distinct elements in S, then add x_i to S.
5. Else add x_i, then delete it, as well as one copy of each of the k elements already present in S.
6. End For

The above algorithm can be seen as identical to the first version (with the number of copies same as the value of the counters), except for the space used (the former is succinct). But from the second algorithm, it is clear that elements are deleted in groups of size $k+1$, and thus each element can be deleted at most $n/(k+1)$ times. The number of deletions of x corresponds to the number $a(x) + b(x)$ of the first algorithm, which proves Claim 1. ■

We remark that we can rephrase the guarantee of the Misra-Gries algorithm as follows: The Misra-Gries algorithm uses $O(\frac{1}{\varepsilon} \log m)$ space to find a set of size at most $\lceil \frac{1}{\varepsilon} \rceil$, that contains every element of frequency at least εn.