University of Washington
Department of Computer Science and Engineering
CSE 522, Spring 2014

May 3, 2014

Problem Set 1, Due Wednesday, April 30, 2014

Problem #1:

Generalize the lower bound argument for the complexity of property testing for element distinctness to show that any algorithm that always accepts distinct inputs, but with probability at least 1/2 rejects all inputs with $< (1-\varepsilon)n$ distinct values requires $\Omega(\sqrt{n/\varepsilon})$ samples.

Problem #2:

Consider the following deterministic algorithm which, unlike the Misra-Gries algorithm, computes an over-estimate, rather than an under-estimate of the heavy hitter frequencies.

Space-Saving Algorithm

Initialize: $k \leftarrow \lceil 1/\varepsilon \rceil$
$A \leftarrow \emptyset$, $\quad$ A is a set of up to k pairs $(j, \tilde{f}_j)$.
Process:
for each i **do**
    if $x_i \in A$ **then**
        $\tilde{f}_{x_i} \leftarrow \tilde{f}_{x_i} + 1$
    else if $|A| < k$ **then**
        Add x_i to A
        $\tilde{f}_{x_i} \leftarrow 1$
    else
        $j' \leftarrow \mathrm{argmin}_{j \in A} \tilde{f}_j$
        $\tilde{f}_{x_i} \leftarrow \tilde{f}_{j'} + 1$
        Replace $(j', \tilde{f}_{j'})$ in A with $(x_i, \tilde{f}_{x_i})$
    end if
end for
Output: $\tilde{f} \leftarrow A$
$\tilde{f}_j$ is as given for $j \in A$, $\tilde{f}_j = 0$ if $j \notin A$.

Show that:

(a) For every $j \in A$, $\tilde{f}_j \geq f_j$.

(b) For every $j \in [M]$, $\tilde{f}_j \leq f_j + \tilde{f}_{min}$ where $\tilde{f}_{min} = \{f_j \ : \ j \in A\}$.

(c) $\sum_{j \in A} \tilde{f}_j = n$

(d) $\tilde{f}_{min} \leq \lfloor n/k \rfloor$ and hence $f_j \leq \tilde{f}_j \leq f_j + \lfloor n/k \rfloor$ for every $j \in A$.

(e) For $i \leq k$, the i-th largest $\tilde{f}_j$ value is an upper bound on the i-th largest f_j value (even though they might be for different values of j).

Problem #3:

This problem shows a tight relationship between the approximations of the Space-Saving algorithm above and the Misra-Gries algorithm when run using the same value of k:

Let A^{MG} be the set of size at most $k-1$ maintained during the execution of the Misra-Gries algorithm and $\tilde{f}^{MG}$ be the frequency values maintained by that algorithm, where $\tilde{f}_j^{MG} = 0$ for $j \notin A^{MG}$.

Similarly, define A^{SS}, the set of size up to k maintained by the Space-Saving algorithm. Let $\min^{SS}$ be the k-th largest frequency in A^{SS} where $\min^{SS} = 0$ if $|A^{SS}| < k$. Let $\tilde{f}^{SS}$ be the vector of frequency estimates maintained by the Space-Saving algorithm during its execution, where we consider $\tilde{f}_j^{SS} = \min^{SS}$ if $j \notin A^{SS}$.

Finally, let sum^{MG} be $\sum_{j=1}^{M} \tilde{f}_j^{MG}$.

Prove by induction on n that after processing the same sequence of n inputs,

$$\min{}^{SS} = (n - \text{sum}^{MG})/k$$

and for every $j \in [M]$,

$$\tilde{f}_j^{SS} = \tilde{f}_j^{MG} + \min{}^{SS}.$$

Problem #4:

Given two relations R and S with a common attribute, for query optimization it is useful to estimate the size of the join $R \bowtie S$ without actually executing it. If the frequency vectors for the attribute in the two relations are $f = (f_1, \ldots, f_M)$ and $g = (g_1, \ldots, g_M)$ then the number of tuples in that join is precisely $\sum_{j=1}^{M} f_j g_j = \langle f, g \rangle$. Design and analyze an algorithm based on the Tug-of-War Sketch for F_2 that maintains sketches for both f and g that provides a $1 \pm \varepsilon$ factor approximation of the join size with probability at least $1 - \delta$.

Hint: Replace the use of y^2 in the Tug-of-War Sketch with a product of two different values.