Homework – Sketching and Streaming Algorithms

Olivier Beaumont

Deadline: November 2nd

Estimation of the number of unique visitors

In the course, we have studied an algorithm to estimate the number of unique visitors, which was based on an idealized version of hash functions, since we neglected the storage costs associated to hash functions. In the remainder of this homework, we will remove this assumption by relying on a "true" family $\mathcal{H}$ of hash functions from $[1, p]$ to $[1, p]$, such as those defined in the lectures.

This algorithm uses 2 steps that are actually performed simultaneously in practice. We have seen the second one in the lectures, and this homework describes the first one. In the first step, the goal is toi find an approximation algorithm, to a constant factor C, of the number of unique visitors t by $\tilde{t}$ in the form

$$\frac{t}{C} \leq \tilde{t} \leq Ct.$$

We will start by choosing h in a $\mathcal{H}$ pair wise independent family of functions from $[1, n = 2^k]$ in $[1, n = 2^k]$. We will define X as

$$X = \max_{i \in \text{visitors}} lsb(h(i)),$$

where $lsb(m)$ denotes the least significant bit of m, i.e. the position of the rightmost "1" in the binary form of m. Our goal is to prove that $\tilde{t} = 2^X$ is a (reasonably) good approximation of t.

Question 1: Write the procedure described above in the form of an algorithm like those seen in the lectures for the number of visits and the number of visitors (idealized case). Discuss its complexity.

Question 2: Let $m \in [1, n]$ be a random number and $j \in [0, \log n]$, what is the probability that $lsb(m) = j$?

Question 3: If t denotes the number of unique visitors, what is the expectation of the number of unique visitors such that $lsb(h(i)) = j$?

Question 4: How to use the previous property to get an "idea" of t?

Let us set j. We will denote $Z_j = \#\{i \in \text{visitors}, \ lsb(h(i)) = j\}$ and $Z_{>j} = \#\{i \in \text{visitors}, \ lsb(h(i)) > j\}$, where $\#(S)$ denotes the cardinal of the set S. Let us also set $Y_i = 1$ if $lsb(h(i)) = j$ and 0 otherwise.

Question 5: Prove that $E(Z_j) = \frac{t}{2^{j+1}}$ and that $E(Z_{>j}) < \frac{t}{2^{j+1}}$.

Question 6: Evaluate $V(Z_j)$. To do so, you can prove that $E(Y_i Y_k) = E(Y_i)E(Y_k)$ with the above assumptions.

Let us now set $j* = \lceil \log t + 5 \rceil$ and $\hat{j} = \lceil \log t - 5 \rceil$.

Question 7: Compute $E(Z_{j*})$? Give an upper bound for $Pr(Z_{j*} = 0)$ (based on Markov)?

Question 8: Compute $E(Z_{\hat{j}})$? Give an upper bound for $Pr(Z_{\hat{j}}) \geq 1)$ (based on Chebychev)?

Question 9: Find an approximation algorithm for t (computing $\tilde{t}$). What is the approximation ratio C and what is the probability that

$$\frac{t}{C} \leq \tilde{t} \leq Ct?$$

Another option for Questions 8 and 9 is the following:

Question 8 bis: Let l be the largest value in $\{0, \ldots, \log n\}$ such that $\forall k \geq l,\ Z_k = 0$. Prove that $P(l > L + 4) < 1/8$ (using Markov)

Question 9 bis:Prove $P(Z_{\hat{j}} = 0) < 1/8$ (using Chebychev)