Universitat Politècnica de Catalunya
Facultat d'Informàtica de Barcelona

Degree: Master **Academic year:** Q1 2022–2023 (Final Exam)
Course: Randomized Algorithms (RA-MIRI) **Date:** January 19th, 2023
Time: 2h 30min

1. **(2.5 points)** You have a finite set U of n elements and draw r independent random subsets A_1, A_2, ..., A_r (there are 2^n subsets of U, hence we select a any given subset with probability $1/2^n$). Compute in a simple closed form the probability that the r subsets are pairwise disjoint (that is, the probability that $A_i \cap A_j = \emptyset$ for all subsets A_i and A_j with $i \neq j$).

---

2. **(2.5 points)** Suppose we insert N different keys in a hash table with separate chaining with M slots (linked lists of synonims). Assume that the hash function that we use has been drawn from a strongly universal class, thus $\mathbb{P}[h(x) = i] = 1/M$ for any key x and any i, $0 \leq i < M$. Compute the expected number of empty lists. Give an asymptotic approximation for that expectation, assuming $N = \alpha M$ for some $\alpha \in (0, 1)$.

---

3. **(2.5 points)** Consider the following algorithm (in pseudo-code):

```
// B: Bloom filter; Z = z1, z2, ...: data stream of length N
B:= empty filter; count:= 0
while (there are elements in Z) do
   z:= next item in Z
   if (z not in B)
       add z to B; count:= count + 1
   endif
endwhile
```

   (a) What can we say about `count`? What is its relation to the number n of **distinct** elements in Z?

   Assume that the parameters M (size of the bitvector) and k (number of hash functions) of the Bloom filter have been choosen to guarantee that the rate of false positives is ≤ 0.01. Give bounds relating n and `count`.

(b) Assuming we can approximate $\mathbb{E}[f(X)]$ by $f(\mathbb{E}[X])$ (which is wrong in general), and the result from the previous question, propose how to estimate n in terms of the bitvector stored in the Bloom filter, in particular, of the number of zeros in the bitvector.

---

4. **(2.5 points)** In a very simplified model of the weather of Sunny Valley, we will consider three possible states: sunny (S), cloudy (C) and rainy (R). The analysis of historical data reveals that the probability that after a sunny day the next day is also sunny is $1/2$, the probability that is cloudy is $1/3$ and the probability that is rainy is $1/6$. If the day is cloudy then the following day will be sunny with probability $1/4$, cloudy with probability $1/2$ and rainy with probability $1/4$. Finally, if a day is rainy then the next day will be rainy with probability $1/2$, cloudy with probability $1/3$ and sunny with probability $1/6$.

   (a) Draw the directed graph corresponding to the Markov chain.

   (b) Prove that the Markov chain is irreducible and aperiodic.

   (c) If day 1 is sunny, what is the probability that day 3 is rainy?

   (d) Compute the stationary distribution π and $P^* = \lim_{t\to\infty} P^t$.
