Big Data Computing **A.Y. 2023/2024**
Homework 1 **Deadline: October 22nd**

Instructions. You should hand in your homeworks within the due date and time. Late deliveries will be penalized, please check homework delivery policies in the exam info.

Handing in. You should submit your work as *one pdf* containing your answers and solutions to the theoretical assignments. Answers should be either typed up or hand written **very clearly** and scanned. The first option is preferred.

Please deliver your homework by attaching the above zip file to an email addressed to me as follows:
To: `becchetti@diag.uniroma1.it`
Subject: `HW1 <Your last name> <Your first name> <Your Sapienza ID number>`

Typesetting in Latex. Latex is very handy for typesetting (especially math), but you need to install it. If you do not want to install Latex, you can go for Overleaf, providing an integrated, Web interface, accessible for free in its basic version (which is enough for your needs). It allows you to both type in Latex using a Web interface, and compiling your code to produce a pdf document. Overleaf's documentation also contains a tutorial on Latex essentials.

Important note. Grading of your answers to the theoretical assignments will depend on i) correctness and solidity of the arguments, ii) clarity of the presentation and iii) mathematical rigour. E.g., ill-defined quantities, missing assumptions, undefined symbols etc. are going to penalize you. Rather than writing a lot, try to write what is needed to answer and write it well.

Assignment 1.

A (undirected) random graph with n vertices is a probabily distribution over the set of all possible graphs $G = (V, E)$ with n vertices. In particular, a popular model is the following: i) without loss of generality, the vertex set is $V = \{1, .., n\}$; ii) the edge set E is obtained as follows: for every $u, v \in V$ (with $u \neq v$) edge (u, v) exists with probability p *independently* of other edges. Here, $p \in [0, 1]$ is a parameter, which (together with n) completely defines the probability distribution. We are interested in the probability that G contains at least one *clique* of size at least k, for some integer $k \leq n$.[1] More precisely, answer the following questions:

(a) Denote by Z the number of cliques of size exactly k that are present in G. Compute both upper and lower bounds on $\mathbb{E}[Z]$.

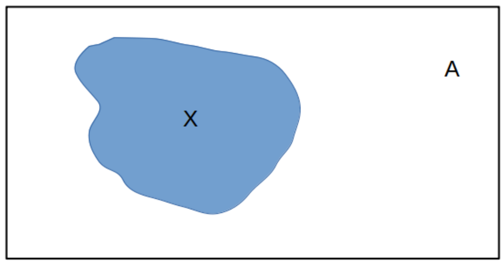
[1] You are reminded that, given an undirected graph $G = (V, E)$, a clique is a subgraph $G' = (V', E')$ of G that is fully connected. I.e., $V' \subseteq V$ and the subset $E' \subseteq E$ of the edges induced by V' contains all possible edges connecting vertices in V' (thus, exactly $\binom{|V'|}{2}$ edges).

(b) For $0 < \varepsilon < 1$, give an upper bound to the probability that a clique of size *at least* $k = \frac{epn}{1-\varepsilon}$ exists.[2]

Hints. For both questions (a) and (b), note that for integers n and k with $n \leq k$ we have $\left(\frac{n}{k}\right)^k \leq \binom{n}{k} \leq \left(\frac{en}{k}\right)^k$. For question (b), observe that if G contains a clique of size at least k then it must contain a clique of size exactly k.

Assignment 2.

Assume we want to estimate the area (in pixels) of a shape contained in a rectangular screen canvas, for example as in the following picture:



In the picture, A is the overall area (number of pixels) of the canvas, while X is the area of the shape you are interested in. All pixels in the canvas are black or white, depending on whether they belong to the shape of interest or to the rest of the canvas. In order to provide an estimate of X, you can use sample points uniformly at random from the canvas. In particular, each call to `sample()` independenly samples a pixel uniformly at random from the canvas, returning 1 if the sampled point is black (i.e., it belongs to the shape), while it returns 0 otherwise. Answer the following questions:

(a) Design an algorithm that, only using the `sample()` primitive and the values it returns, computes an unbiased estimator of X.

(b) Give a bound on the number m of calls of the `sample()` primitive that allow to provide an estimate $\hat{X}$ of X, such that $|\hat{X} - X| \leq \varepsilon X$ with probability at least $1 - \delta$. Note that this lower bound should depend on A, X, ε and δ.

In part (b) $0 < \varepsilon < 1$ measures a desired degree of accuracy, while $0 < \delta < 1$ measures a desired degree of confidence. For example, if $\varepsilon = 0.05$ and $\delta = 0.1$, it means that we want our estimate to be within 5% of the true value with 90% probability.

Hints. To answer part (b) you will need a concentration bound, the stronger the better. Observe that all samples you take are assumed to be independent.

Assignment 3.

Professor Knowitall and Professor Knowitbetter argue about the statistical significance of a phenomenon observed in a sample of a (undirected) social network graph $G = (V, E)$. In particular, G has $n = 5000$ vertices and $m = 10^6$ edges. Moreover, (at least) one of the vertices has degree 600. Professor Knowitall maintains that the presence of

[2]Note that e is the base of the natural logarithm.

even one vertex of such high degree is indicative of the presence of social structure, in particular *hub* nodes of high centrality. Professor Knowitbetter instead argues that this phenomenon is not supported by empirical evidence. In particular, the presence of a vertex of degree 600 might simply arise by chance if one considers a random graph as defined in Assignment 1, with the same number of vertices and expected degree equal to the average degree of G. Whose argument do you support and why? In more detail:

(a) Formulate the problem as one of hypothesis testing, clearly defining the null hypothesis H_0;

(b) Provide quantitative and mathematically rigorous arguments in favour of Professor Knowitall's or Professor Knowitbetter's thesis.

Hints. You might need to use a Chernoff bound at some point. Standard forms are given in the slides or in even on Wikipedia. Be sure you apply the right one and motivate its application, by clearly identifying the variables you are considering and explaining why they are independent.