BIG DATA COMPUTING

Homework 1
Concentration of measures and statistical significance

Leonardo Di Nino : 1919479

October 21, 2023

Assignment 1

Question (a)

Given such a random graph model $G(V, E)$ where $|V| = n$ and $\mathbb{P}((u, v) \in E) = p$ for each $u, v \in V$, the total number of possible cliques of size k is given by $C = \mathcal{C}_{n,k} = \binom{n}{k}$, i.e. the number of subset of K nodes. Each of these cliques is associated to a different subset of nodes $\{V_i'\}_{i=1}^{C}$ such that the clique exists only if the induced subgraph $G(V_i')$ has a number of edges equal to $\binom{k}{2}$. Stating that each edge is a random variable independent from all the others, we can then associate to each of those cliques a binary random variable $Z_i \sim Ber(q)$: $Z_i = 1$ if the cliques exists, with a probability equal to $q = p^{\binom{k}{2}} = \frac{k(k-1)}{2}$ since the clique needs such a number of edges in order to exists. So we have the succession of random variables $\{Z_i\}_{i=1}^{C}$ and $Z = \sum_{i=1}^{C} Z_i$ is the total number of cliques of size k. Finally we have the following result just leveraging the linearity of expected value:

$$\mathbb{E}[Z] = \mathbb{E}[\sum_{i=1}^{C} Z_i] = \sum_{i=1}^{C} \mathbb{E}[Z_i] = Cq = \binom{n}{k} p^{\frac{k(k-1)}{2}} \tag{1}$$

This is a closed form solution for $\mathbb{E}[Z]$, but we can easily bound it since we have the following inequality for binomial coefficients:

$$(\frac{n}{k})^k \leq \binom{n}{k} \leq (\frac{en}{k})^k \tag{2}$$

So just applying eq.(2) to eq.(1) we get:

$$(\frac{n}{k})^k p^{\frac{k(k-1)}{2}} \leq \mathbb{E}[Z] \leq (\frac{en}{k})^k p^{\frac{k(k-1)}{2}} \tag{3}$$

Question (b)

We now want to provide an upper bound for the probability for the existence of a clique of size at least equal to $\frac{epn}{1-\epsilon}$ for $0 < \epsilon < 1$.

Before moving forward let's consider the feasibility of the question also to figure out in which dynamic k ranges. In order to have consistency we have to impose $k \leq n$, i.e. $\frac{epn}{1-\epsilon} \leq n$, so that $0 \leq \epsilon \leq 1 - pe$. Still we want $1 - pe \geq 0$ so that also $p \geq \frac{1}{e}$. In this setting we are considering a high value for k: we are likely to consider $\epsilon \approx 0$ and $p \approx \frac{1}{e}$ to retrieve non trivial bounds, as we can see.

Now moving on with probabilistic modeling we can introduce the event I_u as the binary random variable whose realization is 1 if the clique of size u does exist. It is interesting to show that $I_{u+1} \subset I_u$ since $I_{u+1} = I_u \cap N_{u+1}$ being N_u the event "There is a $(u+1)$-th node that is connected to all the u nodes forming a u sized clique". We have that the following holds:

$$\begin{gathered} I_k \supset I_{k+1} \supset I_{k+2} \supset ... \supset I_n \\ \mathbb{P}(I_k) > \mathbb{P}(I_{k+1}) > \mathbb{P}(I_{k+2}) > ... > \mathbb{P}(I_n) \end{gathered} \tag{4}$$

This few considerations in modeling are actually very important, since the event of interest is that the graph has at least a clique of at least size k:

$$\mathbb{P}(\bigcup_{u=k}^{n}(I_u = 1)) = \mathbb{P}((I_k = 1)) \tag{5}$$

since we have $I_k \supset I_{k+1} \supset I_{k+2} \supset ... \supset I_n$ we have that $\bigcup_{u=k}^{n}(I_u = 1) = I_k$.

Finally we can just consider to bound $\mathbb{P}(I_k = 1)$. Firstly stating that Z still is the total number of cliques of size k and we are considering that a clique must exist we can leverage the fact that all Z_i are binary random variables as defined in question (a), so that:

$$\mathbb{P}(I_k = 1) = \mathbb{P}(Z \geq 1) = \mathbb{P}((\sum_{i=1}^{C} Z_i) \geq 1) = \mathbb{P}(\bigcup_{i=1}^{C}(Z_i = 1)) \tag{6}$$

Now we can apply Boole's inequality to eq.(6):

$$\mathbb{P}(\bigcup_{i=1}^{C}(Z_i = 1)) \leq \sum_{i=1}^{C} \mathbb{P}(Z_i = 1) \tag{7}$$

from this last naturally follows the following, since $\sum_{i=1}^{C} \mathbb{P}(Z_i = 1) = \mathbb{E}[Z]$:

$$\mathbb{P}(\bigcup_{i=1}^{C}(Z_i = 1)) \leq \mathbb{E}[Z] \tag{8}$$

and now we can use the upper bound we derived in eq.(3) making some further assumptions leveraging $p \in [0, 1]$, that implies that exponentiation is decreasing when using p as a base:

$$\mathbb{P}(\bigcup_{i=1}^{C}(Z_i = 1)) \leq \mathbb{E}[Z] \leq (\frac{en}{k})^k p^{\frac{k(k-1)}{2}} \leq (\frac{en}{k})^k p^{k^2} \leq (\frac{en}{k})^k p^k \bigg|_{k=\frac{epn}{1-\epsilon}} \tag{9}$$

Chaining everything we have the following:

$$\mathbb{P}(Z \geq 1) \leq (\frac{epn}{k})^k \tag{10}$$

We can now evaluate the term bounding on the right in $k = \frac{epn}{1-\epsilon}$:

$$(\frac{epn}{k})^k \bigg|_{k=\frac{epn}{1-\epsilon}} = (\frac{epn}{\frac{epn}{1-\epsilon}})^{\frac{epn}{1-\epsilon}} = (1-\epsilon)^{\frac{epn}{1-\epsilon}} \tag{11}$$

Finally we have our upper bound:

$$\mathbb{P}(Z \geq 1) \leq (1-\epsilon)^{\frac{epn}{1-\epsilon}} \tag{12}$$

Assignment 2

Question (a)

Given the assumptions we can configure a proper probabilistic model. In particular being A a measurable set, we can define a uniform probability distribution over it such that for each $a \in A$ we have the following probability measure: $f(a) = \frac{1}{\mu(a)}\mathbf{1}_A(a)$, being $\mu(A)$ the Lebesgue integral over A, i.e. its very area: this models the fact that the pixels are uniformly distributed in A.

We then can consider the following probability, given that $X \subseteq A$ and basic measure theory results:

$$\mathbb{P}(a \in X | X \subseteq A) = \int_X \frac{d\mu(X)}{\mu(a)} = \frac{\mu(X)}{\mu(A)} \tag{13}$$

From now on to lighten notation we'll consider $\mu(A) = A$, $p = \mathbb{P}(a \in X|X \subseteq A)$ and $\mu(X) = X$, so that $\mathbb{P}(a \in X|X \subseteq A) = p = \frac{X}{A}$, and naturally $X = Ap$.

We want to provide an unbiased estimator of X using repeated IID sampling leveraging these uniformity assumptions. So we can define for each pixel sampled $Z \sim Ber(p)$ as the binary random variable assuming value 1 if $Z \in X$ so that this distribution is parametrized exactly by $p = \mathbb{P}(a \in X|X \subseteq A)$: this is realized by the routine `sample()`.

Given this sampling scheme and an actual sample $\{Z_i\}_{i=1}^m$, we know that the maximum likelihood estimator for p is the sample mean $\overline{Z}_m$ and it's trivial to show that it is also an unbiased estimator for p:

$$\mathbb{E}[\overline{Z}_m] = \mathbb{E}[\frac{1}{m}\sum_{i=1}^m Z_i] = \frac{1}{m}\sum_{i=1}^m \mathbb{E}[Z_i] = \frac{1}{m}mp = p \tag{14}$$

Now we can leverage that $X = Ap$ is a function $\phi(p)$ and the fact that maximum likelihood estimation is invariant under continuous mapping: $\hat{\phi}_{MLE}(p) = \phi(\hat{p}_{MLE})$. This finally gives us the estimator for X that is $\hat{X} = A\overline{Z}_m$. Given the primitive `sample()` we have the following procedure:

Algorithm 1 Unbiased estimator of X

Require: $A : float, m : int$
  $S \leftarrow \mathbf{0}_n$ ▷ Pre-allocating a vector for the samples
  for $i = 1, ..., m$ **do**
    $S[i] \leftarrow sample()$
  end for
  return $\frac{A}{n}\sum_{i=1}^m S[i]$

Question(b)

Given a precision ϵ and a degree of confidence δ we want to bound m such that

$$\mathbb{P}(|\hat{X} - X| \leq \epsilon X) \geq 1 - \delta$$

We can play a bit with this expression:

$$\begin{gathered}\mathbb{P}(|\hat{X} - X| \leq \epsilon X) \geq 1 - \delta \\ \mathbb{P}(|A\hat{p} - Ap| \leq \epsilon X) \geq 1 - \delta \\ \mathbb{P}(A|\hat{p} - p| \leq \epsilon X) \geq 1 - \delta \\ \mathbb{P}(|\overline{Z}_m - \mathbb{E}[Z]| \leq \frac{\epsilon X}{A}) \geq 1 - \delta\end{gathered} \tag{15}$$

Hoeffding's inequality is a good choice in our case, since it is an exponentially decreasing bound. Hoeffding's inequality is usually in the form

$$\mathbb{P}(|\overline{X}_m - \mathbb{E}[X]| \leq \kappa) \geq 1 - 2e^{\frac{-2m\kappa^2}{(b-a)^2}}$$

under the hypotesis that X is a bounded random variable, which is true for a Bernoulli being $b = 1$ and $a = 0$. So finally we have the following:

$$\begin{aligned} \mathbb{P}(|\overline{Z}_m - \mathbb{E}[Z]| \leq \frac{\epsilon X}{A}) \geq 1 - \delta \\ \mathbb{P}(|\overline{Z}_m - \mathbb{E}[Z]| \leq \frac{\epsilon X}{A}) \geq 1 - 2e^{-2m(\frac{\epsilon X}{A})^2} \end{aligned} \tag{16}$$

Comparing directly δ and the term in Hoeffding's inequality we get:

$$\begin{aligned} \delta &= 2e^{-2m(\frac{\epsilon X}{A})^2} \\ log(\frac{\delta}{2}) &= -2m(\frac{\epsilon X}{A})^2 \\ m(\frac{\epsilon X}{A})^2 &= \frac{1}{2}log(\frac{2}{\delta}) \\ m &= \frac{A^2 log(\frac{2}{\delta})}{2\epsilon^2 X^2} \end{aligned} \tag{17}$$

The last expression is the lower bound on m respecting the initial request.

Assignment 3

Question (a)

We can model the situation as a hypotesis testing problem with the following hypotesis set:

$$\begin{cases} \mathcal{H}_0 : \text{There is at least one node whose degree is 600 as effect of randomness;} \\ \mathcal{H}_1 : \text{At least one node whose degree is 600 is a prove of existence of strong social structures} \end{cases}$$

We are going to find a formal acceptance or rejection of $\mathcal{H}_0$ through a bound on the p-value, i.e. the probability for the observed data under the null hypotesis.

Question (b)

We can be a little bit more formal in defining what we are dealing with. First of all we can retrieve the average degree from $|V| = 5000, |E| = 10^6$: $\hat{d} = \frac{2|E|}{|V|} = 400$.

Now we'll focus on a single node of interest u and we'll define D_u as the random variable that associate to u its degree. From the assumptions $\mathbb{E}[D_u] = \hat{d} = 400$, and given this we can also retrieve the following estimate for the parameter p characterizing the underlying Erdős–Rényi model in the null hypotesis: $\hat{p} \approx \frac{\hat{d}}{|V|} = 0.08$.

We then can model the degree of a fixed node through a random variable D that under $\mathcal{H}_0$ is equal to the sum of IID bernoulli random variables; in particular $D_u = \sum_{i=1}^{|V|-1} e_i$, being in this case $e_i \sim Ber(\hat{p})$ the bernoulli for the existence of the edge between the fixed node in question and all the other nodes. We have from the assumptions that $\mathbb{E}[D] = 400$: all these can give us the following upper tail Chernoff bound over the probability that $D \geq 600$:

$$\mathbb{P}(D \geq (1+\delta)\mathbb{E}[D]) \leq \exp\{-\frac{\delta^2}{2+\delta}\mathbb{E}[D]\} \tag{18}$$

being $(1+\delta)\mathbb{E}[D] = 400(1+\delta) = 600$. So we have $\delta = \frac{1}{2}$ and the following bound holds:

$$\begin{gathered} \mathbb{P}(D \geq 600) \leq \exp\{-\frac{\delta^2}{2+\delta}\mathbb{E}[D]\}\Big|_{\delta=\frac{1}{2}} \\ \mathbb{P}(D \geq 600) \leq \exp -\frac{\frac{1}{4}}{\frac{5}{2}} * 400 \\ \mathbb{P}(D \geq 600) \leq \exp -40 \\ \mathbb{P}(D \geq 600) \leq 4.24 * 10^{-18} \end{gathered} \tag{19}$$

What we are interested in is the probability that at least one vertex has such a degree: so we consider now the random variables $\{D_u\}_{u=1}^{|V|}$ and the following holds through Boole's inequality:

$$\mathbb{P}(\bigcup_{u=1}^{|V|}(D_u \geq 600)) \leq \sum_{u=1}^{|V|}\mathbb{P}(D_u \geq 600) \tag{20}$$

Since for each term in the right side of this last inequality eq.(19) holds, we can bound each term in the sum and get the following:

$$\mathbb{P}(\bigcup_{u=1}^{|V|}(D_u \geq 600)) \leq \sum_{u=1}^{|V|}\mathbb{P}(D_u \geq 600) \leq \sum_{u=1}^{|V|} 4.24 * 10^{-18} \tag{21}$$

Finally we have

$$\mathbb{P}(\bigcup_{u=1}^{|V|}(D_u \geq 600)) \leq 2.12 * 10^{-14} \tag{22}$$

Now consider a standard level for hypotesis testing $\alpha = 0.05$; due the Bonferroni correction we have that the actual level for our test is $\alpha' = \frac{\alpha}{|V|} = 10^{-5}$. The strong upper bound we have on the p-value implies the rejection of the null hypotesis for every level of confidence higher than $2.12 * 10^{-14}$, as α' effectively is. In the end, in support against the thesis of prof. Knowitbetter, we could not but reject $\mathcal{H}_0$.