Sapienza University of Rome

Big Data Computing

Homework 1

Student

Clara Lecce, 1796575

Academic Year 2023/2024

Assignment 1

a)

We have that Z_k represents the number of cliques of size exactly k, and $k \leq n$, with n being the size of G.
The largest number of cliques, meaning the total number of possible combination, is given by $\binom{n}{k}$. We can define X_j as an indicator r.v.:

$$X_j = \begin{cases} 1 & \text{if is a clique} \\ 0 & \text{otherwise} \end{cases}$$

we can use these variables to define $Z_k = (X_1 + X_2 + \cdots + X_{\binom{n}{k}})$, where the X_j r.v. indicates whether the j^{th} subset is a clique or not.
Hence, we can estimate the expected number of cliques of size k as follows:

$$\mathbb{E}(Z_k) = \mathbb{E}\left(\sum_{j=1}^{\binom{n}{k}} X_j\right)$$

By the symmetry property of expectations (since X_j are *i.i.d.*):

$$\mathbb{E}\left(\sum_{j=1}^{\binom{n}{k}} X_j\right) = \binom{n}{k}\mathbb{E}(X_1)$$

Now, by definition of indicator r.v., $\mathbb{E}(X_1) = \mathbb{P}(X_1 = 1)$, so in order to calculate the expectation, we need to calculate the probability that the first subset of size k is a clique. We know that a clique has exactly $\binom{k}{2}$ edges, and an edge exists with probability p, so the probability of having a k-clique is:

$$\mathbb{P}(X_1 = 1) = p^{\binom{k}{2}} = \mathbb{E}(X_1)$$

So:

$$\mathbb{E}(Z_k) = \binom{n}{k}\mathbb{E}(X_1) = \binom{n}{k}p^{\binom{k}{2}}$$

Returning to the initial question, we want to find the lower and upper bounds of $\mathbb{E}(Z_k)$, and using the hint given:

$$\left(\frac{n}{k}\right)^k p^{\binom{k}{2}} \leq \mathbb{E}(Z_k) \leq \left(\frac{en}{k}\right)^k p^{\binom{k}{2}}$$

b)

To find the upper bound, we can follow the hint given.
Consider a clique T of size $k+1$. Then we can take a graph T', sub-graph of T, made up of k vertices. Since T is a clique and by definition is fully-connected, every sub-graph of T has this property, so T' is a k-clique.
Whenever exists a clique of size *at least* k, then there always will be a clique of size *exactly* k. So we can use the r.v. defined in the previous exercise, to formalize the probability we are looking for:

$$\{\text{there is a clique of size exactly k}\} = \{Z_k \geq 1\}$$

Using Boole's inequality, with $k = \frac{epn}{1-\epsilon}$, the final upper bound is the following:

$$\begin{aligned}\mathbb{P}\left(\bigcup_{j=1}^{\binom{n}{k}} X_j = 1\right) &\leq \sum_{j=1}^{\binom{n}{k}} \mathbb{P}(X_j = 1) = \binom{n}{k} p^{\binom{k}{2}} \\ &\leq \left(\frac{en}{k}\right)^k p^{\binom{k}{2}} = \left(\frac{1-\epsilon}{p}\right)^{\frac{epn}{1-\epsilon}} p^{\binom{\frac{epn}{1-\epsilon}}{2}}\end{aligned}$$

Assignment 2

a)

We can define an unbiased estimator for X in the following way: $\hat{X} = \frac{A}{m} \sum_i^m$ `sample()`, with m the number of calls of `sample()`, and we can say that is an unbiased estimator because its expected value is:

$$\mathbb{E}(\hat{X}) = \frac{A}{m} \sum_{i=1}^{m} \mathbb{E}(X_i) = \frac{A}{m} \cdot m \frac{X}{A} \Rightarrow \mathbb{E}(\hat{X}) = X$$

Where X_i is an indicator r.v. that models `sample()`, which follows $Ber\left(p = \frac{X}{A}\right)$.
So, the pseudo-code of the algorithm requested is the following:

```
unbiased_estimator(A, m):
    cnt <- 0

    for i to 1:m:
        cnt <- cnt + sample()
    X = (A/m) * cnt
```

b)

We want to find a bound of the following probability $\mathbb{P}(|\hat{X} - X| \leq \varepsilon X) \geq 1 - \delta$, which is equivalent to $\mathbb{P}(|\hat{X} - X| \geq \varepsilon X) \leq \delta$.
We can use Chebyshev's inequality to find the bound we are looking for, considering that $X = \mathbb{E}(\hat{X})$:

$$\mathbb{P}(|\hat{X} - \mathbb{E}(\hat{X})| \geq \varepsilon X) \leq \delta$$

$$\mathbb{P}(|\hat{X} - \mathbb{E}(\hat{X})| \geq \varepsilon X) \leq \frac{Var(\hat{X})}{(\varepsilon X)^2} \leq \delta$$

Keeping in mind that $\hat{X} = \frac{A}{m} \sum_i^m X_i$ and $X_i \sim Ber\left(\frac{X}{A}\right)$ and applying the properties of the variance of a Bernoulli distribution, we get:

$$\frac{Var(\hat{X})}{(\varepsilon X)^2} = \frac{\frac{A^2}{m^2} Var(\sum_i^m X_i)}{(\varepsilon X)^2} = \frac{\frac{A^2}{m^2} m \frac{X}{A}\left(1 - \frac{X}{A}\right)}{(\varepsilon X)^2} = \frac{A - X}{m \varepsilon^2 X} \leq \delta$$

Since we want to bound m, the final inequation is:

$$m \geq \frac{A - X}{\delta \varepsilon^2 X}$$

Assignment 3

a)

We can define the null hypothesis as the following:
H_0: Random Graph where the edges have probability p.
H_1: Graph has a social structure.

b)

To solve this point, we are going to use the Chernoff bound, this one in particular:

$$\mathbb{P}(X \geq (1+\delta)\mu) \leq \left(\frac{e^{\delta}}{(1+\delta)^{1+\delta}}\right)^{\mu}$$

But first we need to define the variables involved, so, starting from the calculation of the average degree:

$$d = \frac{2m}{n} = \frac{2 \cdot 10^6}{500} = 400$$

Also, considering the X defined as in Assignment 1, meaning an indicator r.v. $X = \sum_i^n X_i$ which is equal to 1 if the i^{th} vertex has an edge (X representing the degree of a given vertex), we want the probability that a given vertex has a degree of 600 or more, and since the average degree is equal to the expected degree, meaning that $d = \mathbb{E}(X) = 400 = \mu$, and putting $(1+\delta)\mu = 600 \Rightarrow \delta = 0.5$, we can rewrite the bound as the following:

$$\mathbb{P}(X \geq 600) \leq \left(\frac{e^{0.5}}{(1.5)^{1.5}}\right)^{400}$$

The probability we get is very small, implying that we are going to reject H_0, disagree with Professor Knowitbetter and agree with Professor Knowitall, and this means that G has a social structure.