Big Data Computing Homework 1

Fabio Fraschetti 1834942

October 2023

[Fabio Fraschetti]

1 Exercise

1.1

From a graph with n nodes we can choose k nodes in $\binom{n}{k}$ different modes. Once we have this combination we can define a Random Variable Z_j that indicate wheter the j subset is a clique or not. So the expected number of k-cliques in a graph of n nodes is:

$$E[Z_1, Z_2, \dots, Z_{\binom{n}{k}}] = E\Big[\sum_{k=1}^{\binom{n}{k}} Z_j\Big]= \binom{n}{k} E[Z_1]$$

We can do the last step because the random variables have all the same expectations. Then we know that $E[Z_1] = Pr(Z_1 = 1)$ so:

$$\binom{n}{k} E[Z_1] = \binom{n}{k} Pr(Z_1 = 1) = \binom{n}{k} p^{\binom{k}{2}}$$

where $\binom{k}{2}$ means all possible combinations of 2 nodes from the k subsets.
Now we can compute the upper and lower bounds as:

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

Where the upper bound is $\left(\frac{en}{k}\right)^k p^{\binom{k}{2}}$ and the lower bound is $\left(\frac{n}{k}\right)^k p^{\binom{k}{2}}$

1.2

The upper bound of the probability that a clique of size at least k exist is:

$$P(Z \geq 1) \leq E[Z] = \binom{n}{k} p^{\binom{k}{2}} \leq \left(\frac{en}{k}\right)^k p^{\binom{k}{2}}$$

Here I'm applying the Markov's inequality and take the upper bound of the exercise above.

$$\left(\frac{en}{k}\right)^k p^{\frac{k(k-1)}{2}} \leq \left(\frac{en}{k}\right)^k p^k = \left(\frac{epn}{k}\right)^k$$

This inequality is true for k¿2. Eventually replacing $k = \frac{epn}{1-\epsilon}$ in the last formula, we obtain our upper bound:

$$(1-\epsilon)^{\frac{epn}{1-\epsilon}}$$

2 Exercise

2.1

Here we provided an algorithm that recall the function $Sample()$ m times and store all the results in j, which in the end contain the sum of all the 1's returned by $Sample()$. Then we compute p that is the estimated probability over m samples and then we can compute $\hat{X}$ by doing p times A, that is the total area. I choose to do a mean because it is an unbiased estimator.

```
function EstimatorX
    $j = 0$
    for $i \leftarrow 0$ to $m$ do
        j=j+Sample()
        i=i+1
    $p = j/m$
    $\hat{X} = p * A$
    return $\hat{X}$
```

2.2

We need that the probability of the module $|\hat{X} - X| \leq \epsilon X$ must satisfies this requirement:

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

Now we can change first part of the inequality by modifying the $\leq$ and split the module:

$$1 - P(|\hat{X} - X| < \epsilon X) = 1 - P(\hat{X} \geq (1+\epsilon)X) - P(\hat{X} \leq (1-\epsilon)X)$$

If the sample has dimension m we apply Chernoff bound to the probabilities and a reduction:

$$1 - \Big(e^{-\frac{\epsilon^2}{3}mp} + e^{-\frac{\epsilon^2}{2}mp}\Big) \geq 1 - 2e^{-\frac{\epsilon^2}{3}mp} \geq \delta$$

From this inequality we want to find the minimum m that satisfy it. So we do the following calculus:

$$-2e^{-\frac{\epsilon^2}{3}mp} \geq \delta{-}1 \Longrightarrow e^{-\frac{\epsilon^2}{3}mp} < \frac{1-\delta}{2} \Longrightarrow \frac{\epsilon^2 mp}{3} < ln\Big(\frac{1-\delta}{2}\Big) \Longrightarrow m \geq -ln\Big(\frac{1-\delta}{2}\Big)\frac{3}{\epsilon^2 p} = -ln\Big(\frac{1-\delta}{2}\Big)\frac{3}{\epsilon^2}\frac{A}{X}$$

We obtained that the bound for m is:

$$m \geq -ln\Big(\frac{1-\delta}{2}\Big)\frac{3}{\epsilon^2}\frac{A}{X}$$

3 Exercise

3.1

Hypothesis H0: The results examined in the graph G, do not denote any evident social structure. (Professor Knowitbetter)
Hypothesis H1: The results examined in the graph G, denote an evident social structure. (Professor Knowitall)

3.2

We need to find the probability that $X_i \geq 600$ where X_i is a random variable that define the degree of the i-th node:

$$P(X_i \geq 600) = P\Big(X_i \geq (1{+}\delta)\frac{m}{n}\Big) = P\Big(X_i \geq 3{\cdot}200\Big) \leq \Bigg(\frac{e^\delta}{(1+\delta)^{1+\delta}}\Bigg)^\mu = \Bigg(\frac{e^2}{(1+2)^{1+2}}\Bigg)^{200} \sim 0$$

Here we used a Chernoff bound with $\delta > 0$ in particular equal to 2 and $\mu = \frac{1000000}{5000} = 200$. We define a level of significance of the p-value to 0.01 and in this case the H_0 is invalidate by p-value, because the result obtained is smaller then the p-value. So I am in favor of the Professor Knowitall's thesis.