Assignment 2

CS203B

Abhishek Pardhi | 200026 | apardhi20@iitk.ac.in

Question 1

Endsem allocation

You are allocated as the Tutor of CS203, with n students. Rajat has created 2 sets of Endsem papers to decrease cheating. He has asked you to help decide which paper should be given to whom. You scraped through the data on Hello, and found out who have been project partners in previous courses, as they will be friends now. Thus, you have found out m friendship connections among the students. You reported this to Rajat, and he said he is fine with any allocation that disrupts atleast half of the friendship connections. A friendship connection is disrupted if the students get different sets of papers.

1A

You are really busy, and just randomly allocated each student to set 1 or set 2. Show that the expected value of disrupted friendship connections is $\frac{m}{2}$.

Answer: Let $C_1, C_2, \ldots, C_m$ be the Bernoulli random variables such that $C_i = 1$ with probability p denotes that i^{th} connection got disrupted and $C_i = 1$ with probability $1 - p$ denotes that i^{th} connection got restored.
This probability $p = \frac{1}{2}$ since i^{th} connection can be disrupted/restored with equal probability. Also, let C denotes the total number of disruptions of connections:

$$\Rightarrow C = \sum_{i=1}^{i=m} C_i$$

$$\Rightarrow E[C_i] = 1 \cdot p + 0 \cdot (1 - p) = p = \frac{1}{2}$$

Now, using linearity of expectation

$$\Rightarrow E[C] = \sum_{i=1}^{m} E[C_i]$$

$$\Rightarrow E[C] = mp$$

$$\Rightarrow E[C] = \frac{m}{2}$$

Hence, the expected value of disrupted friendship connections is $\frac{m}{2}$.

1B

Getting expected value is not enough, you need to find a proper allocation. But you cannot go over all the 2^n allocations as $n \approx 150$. Using the construction for pairwise independence given in class, show that you can find an allocation with at least half of the friendship connections disrupted in poly(n)-time.

Answer: Let there exist a graph $G = (V, E)$ such that vertices V represents students having connections and edges E represents connection between students. Let's assign color B(blue) and R(red) to these vertices.
Notice that this graph coloring represents the same situation as in the question, means, an edge has its end points with same color represents that friendship between those two students got restored, other wise if they have different color then their friendship got disrupted.

Now, let there be a random coloring distribution. Now lets traverse vertices of this graph one by one.
Define an activity as flipping the color of a vertex.
Now, when we traverse each vertices of our graph G in order $O(||V|)$ time, we will perform this above activity on only those vertices which has more than half of its neighbours with different color as itself.
Why we are flipping its color?
Because on flipping, there will be an increase of atleast 1 edge with different color on its end point. Also, any finite graph has an unfriendly partition, which means we can always get a vertex v such that we can perform the above activity on it. The checking time(if we are able to perform activity on vertex v or not) will take $O(|E|)$ time.
There, our code will take $O(|V||E|)$ time. since $|E| = O(n^2)$ and $|V| = O(n)$
There, our code will take $\boxed{\text{O}(\text{n}^3)}$ time, which is in order of polynomial time.

Question 2

Estimating the number of tickets
You are given a bag full of N tickets numbered $1,\ldots,N$(N is unknown to you). You can take out tickets one at a time, note their label, and put them back in the bag. Your task is to estimate N. We will do this in the same way as we estimated π in lecture:

2A

Assume you drew out k tickets. What will be the expected value of the mean of these tickets ? Calculate N in terms of this mean, call this $\tilde{N}$.

Answer: Let $X_1, X_2, \ldots, X_k$ be random variables denoting the label obtained while drawing k tickets. Let $\bar{X}$ be the mean of these random variables. Then

$$\Rightarrow \bar{X} = \frac{\sum_{i=1}^{k} X_i}{k}$$

using linearity of expectation:

$$\Rightarrow \bar{N} = E[\bar{X}] = \frac{\sum_{i=1}^{k} E[X_i]}{k} = E[X]$$

$$\Rightarrow E[X] = \sum_{x=1}^{\infty} x \cdot \mathbb{P}(X = x)$$

$$\Rightarrow E[X] = \sum_{x=1}^{N} \frac{x}{N} = \frac{N(N+1)}{2} = \frac{N+1}{2}$$

$$\Rightarrow \bar{N} = \frac{N+1}{2}$$

$$\Rightarrow \boxed{\text{N}=2\bar{N}-1}$$

2B

Chernoff bound can be extended to work on the case when the Random Variables take values other than $\{0,1\}$. This is known as Hoeffding's inequality. Use it to find a lower bound on the probability that the error in N, using the above calculation, will be less than $\delta N(\delta < 1/2)$. (in terms of N,δ,k)

Answer: Note that the above used random variables $X_i's$ are independent and are bounded between $1 \leq X_i \leq N$, therefore we can apply Hoeffding's inequality on these random variables. Let $S_k = X_1 + \cdots + X_k$, then

$$\Rightarrow \mathbb{P}(|S_k - E[S_k]| \geq t) \leq 2e^{\left(\frac{-2t^2}{\sum_{i=1}^{k}(N-1)^2}\right)}$$

$$\Rightarrow \mathbb{P}(|k\bar{X} - k\bar{N}| \geq t) \leq 2e^{\left(\frac{-2t^2}{k(N-1)^2}\right)}$$

$$\Rightarrow \mathbb{P}(|\bar{X} - \bar{N}| \geq \frac{t}{k}) \leq 2e^{\left(\frac{-2t^2}{k(N-1)^2}\right)}$$

$$\Rightarrow \mathbb{P}(|\bar{X} - \frac{N+1}{2}| \geq \frac{t}{k}) \leq 2e^{\left(\frac{-2t^2}{k(N-1)^2}\right)}$$

$$\Rightarrow \mathbb{P}(|2\bar{X} - N - 1| \geq \frac{2t}{k}) \leq 2e^{\left(\frac{-2t^2}{k(N-1)^2}\right)}$$

$$\Rightarrow \mathbb{P}(|(2\bar{X} - 1) - N| \geq \frac{2t}{k}) \leq 2e^{\left(\frac{-2t^2}{k(N-1)^2}\right)}$$

Or,

$$\Rightarrow \mathbb{P}(|(2\bar{X} - 1) - N| \leq \frac{2t}{k}) \geq 1 - 2e^{\left(\frac{-2t^2}{k(N-1)^2}\right)}$$

Here $2\bar{X} - 1$ is the random variable denoting value of N and N is the expected value of N, therefore

$$\Rightarrow \frac{2t}{k} = \delta N$$

$$\Rightarrow t = \frac{\delta N k}{2}$$

putting this in the above inequality, we get

$$\Rightarrow \mathbb{P}(|(2\bar{X} - 1) - N| \leq \delta N) \geq 1 - 2e^{\left(\frac{-\delta^2 N^2 k}{2(N-1)^2}\right)}$$

Therefore, lower bound of the required probability is $\boxed{1\text{-}2e^{\left(\frac{-\delta^2 N^2 k}{2(N-1)^2}\right)}}$

2C

Assume k, N are odd. In calculation of part (a), instead of using the value of mean, we use the median of the labels of tickets drawn. Prove a lower bound of $1 - 2e^{-\frac{k(1+2\delta)^2}{2(3-2\delta)}}$ on the probability that the error in N using the median will be less than $\delta N(\delta < 1/2)$. (in terms of N, δ, k)

Answer: Let $X_1, X_2, \ldots, X_k$ be random variables denoting the label obtained while drawing k tickets. Let $\tilde{X}$ be the median of these random variables. Then

$$\Rightarrow \tilde{X} = X_i$$

where $i = (N+1)/2$ is the index of sequence of X when arranged in ascending order. Now

$$\Rightarrow \mathbb{P}(X_i = k) =$$

2D

Start with a random hidden value of N in range $10^4 - 10^6$. Write a function that gives k values from $[N]$ when queried with equal probability. Use these values to calculate $\tilde{N}$ as in part (a) and (c), and plot them with respect to increasing $k \leq 1000$. Repeat this estimation for a total of 3 different N, and put the plots in the main answer file. Submit the code you used to generate these plots, along with a readme on how to execute the code, zipped together with the main answer file into a single .zip file.

Answer: Provided in the zipped folder as '200026_Q2.py'. From the plots it can be seen that the required value $\tilde{N}$ is centered around its expected value $\frac{N+1}{2}$.

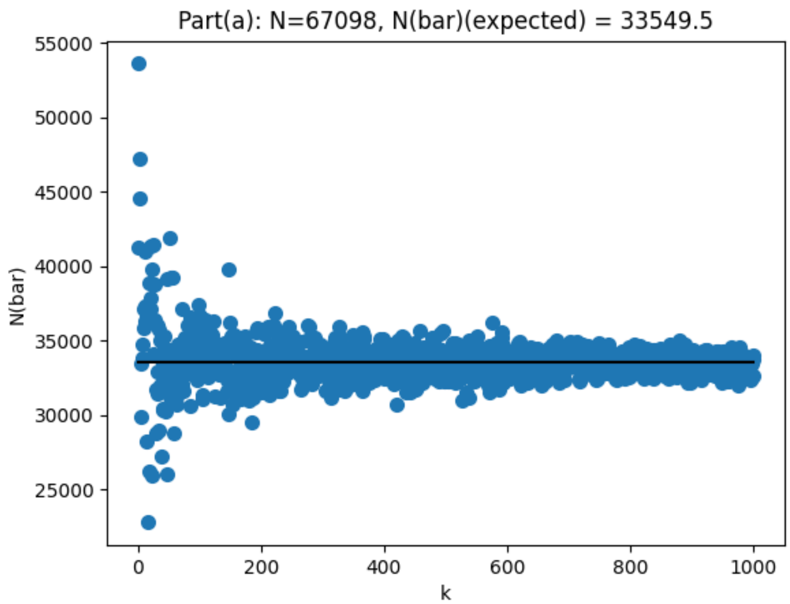

Figure 1: $\tilde{N}$ using Mean

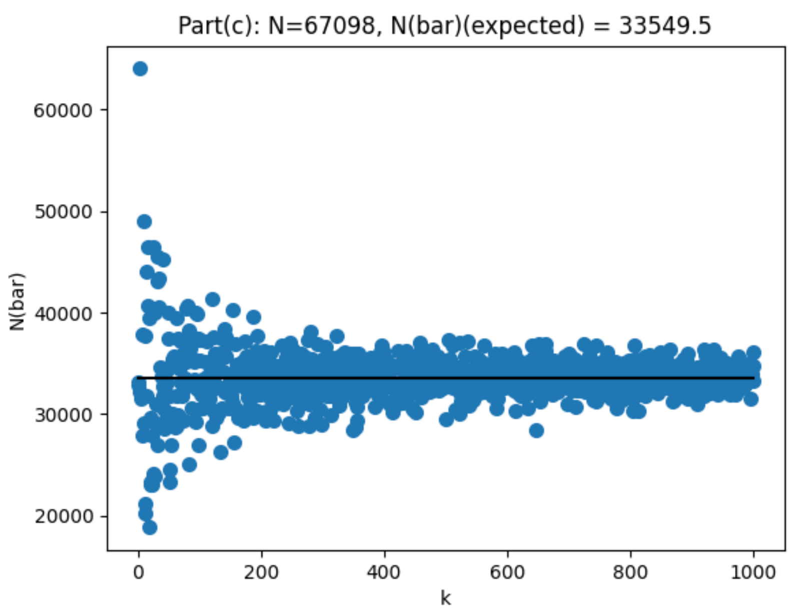

Figure 2: $\tilde{N}$ using Median

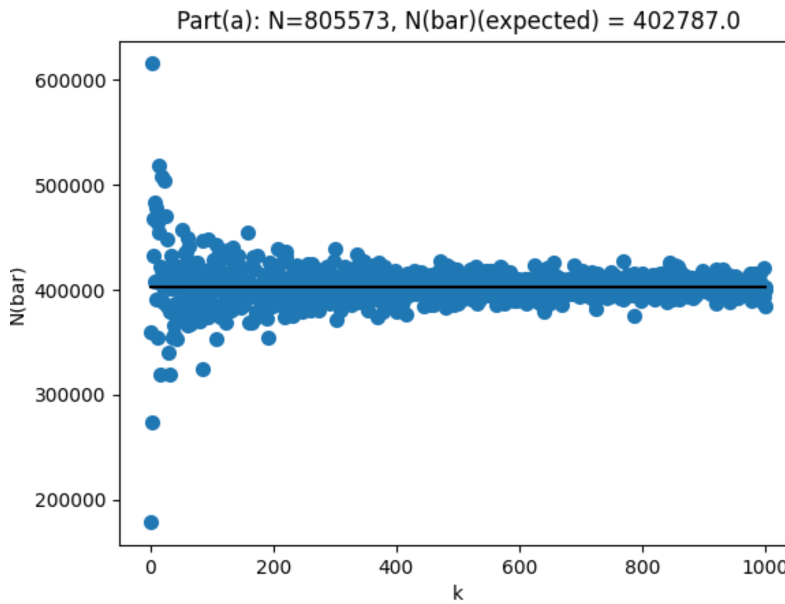

Figure 3: $\tilde{N}$ using Mean

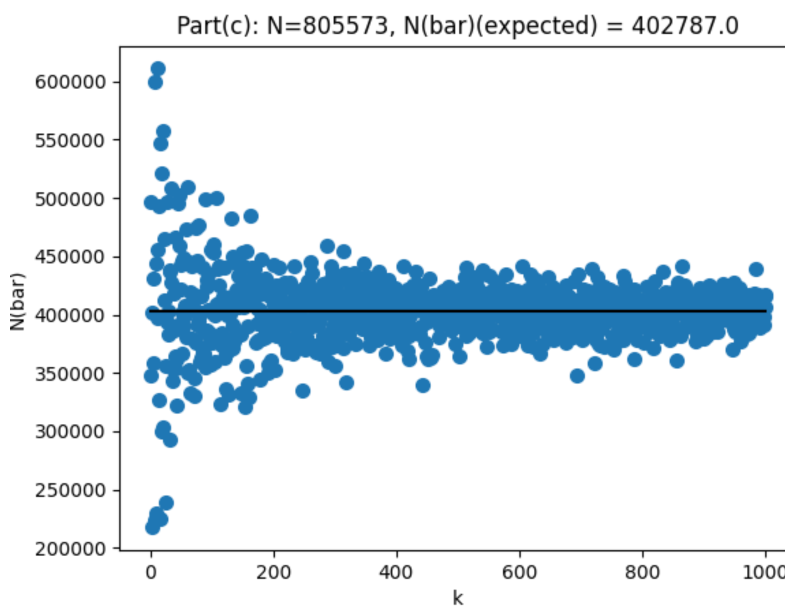

Figure 4: $\tilde{N}$ using Median

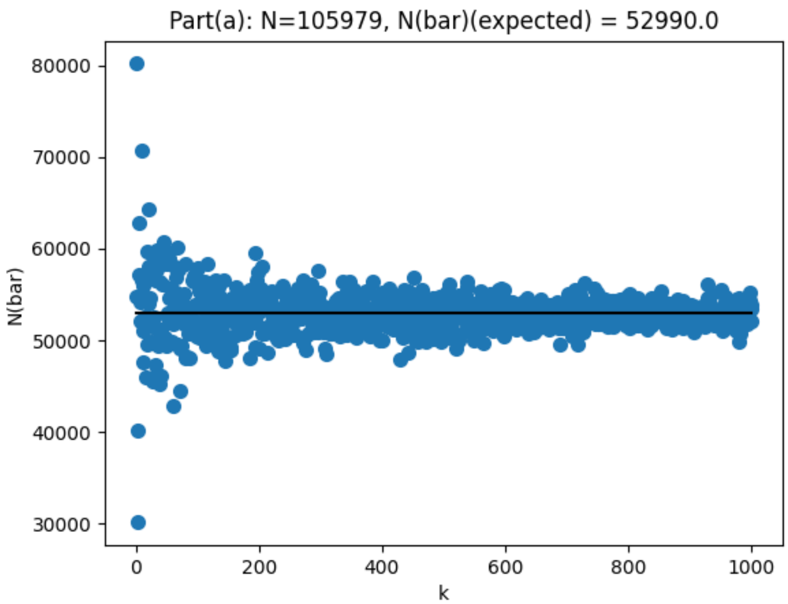

Figure 5: $\tilde{N}$ using Mean

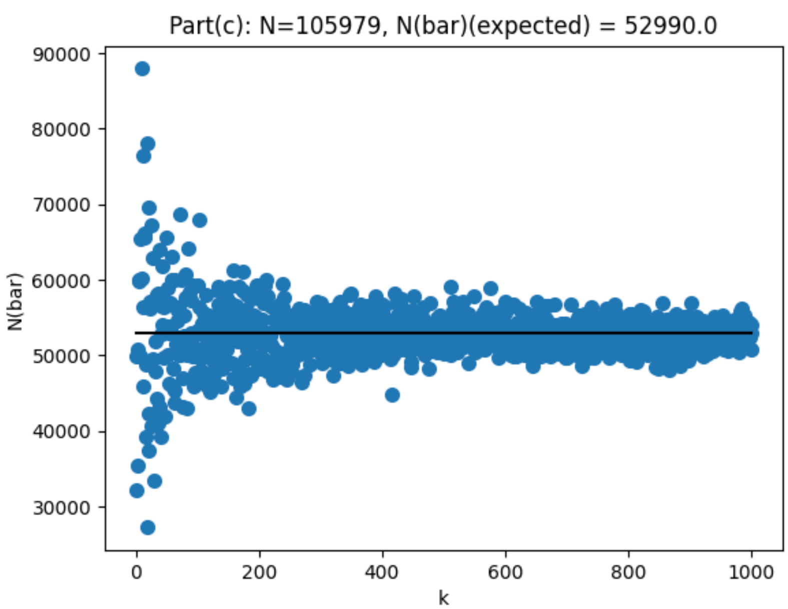

Figure 6: $\tilde{N}$ using Median

Question 3

Markov Chain
Consider a homogeneous regular Markov chain with state space S of size S, and transition matrix M. Suppose that M is symmetric and entry-wise positive.

3A

Show that all the eigenvalues of M are bounded by 1 and that the uniform distribution is the unique stationary probability distribution for M.

Answer: Sum of row terms in transition matrix M is 1 since it covers the exhaustive events of transitioning from a given state i to all possible states j.
Also, we know that there exist an eigen value = 1 for a matrix having sum of each of its rows = 1. Here's the proof of above statement:
We just need to prove $Mv = \lambda v$ such that $\lambda = 1$ for some v. Lets take v to be a unit vector:

$$\Rightarrow MI = \begin{bmatrix} \sum_{j=0}^{j=n} M_{0j} \\ \vdots \\ \sum_{j=0}^{j=n} M_{(n-1)j} \end{bmatrix} = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} = 1 \cdot I$$

Hence, 1 is an eigen value of transition matrix M.
Now, let us assume there exit an eigen value λ of M such that $\lambda > 1$

$$\Rightarrow M^n v = \lambda^n v$$

Also, we know a that $tr(A) = \sum_{i=1}^{|S|} \lambda$ for as matrix A, where $tr()$ is trace of A

$$\Rightarrow tr(M^n) = \sum_{i=1}^{|S|} \lambda_i{}^n$$

But the all elements of M^n are bounded by 1, means the $tr(M^n)$ is bounded by $|S|$.
Notice that the RHS of above equation will diverge on increasing n because λ^n will increase exponentially since $\lambda > 1$.
But the upped bound of RHS is bounded by $|S|$ (since LHS is bounded by $|S|$ and RHS=LHS).

Hence, proved by contradiction, that all the eigenvalues of M are bounded by 1.
We know that Regular Markov chain has a stationary distribution which equals to the uniform probability distribution $\frac{1}{|S|}I$.
Now, let us assume that there exit two eigen values equal to 1 for which $M^n v = 1 \cdot v$ holds for their corresponding eigen vectors say v_1 and v_2 such that $v_1 \neq v_2$. This implies that there exit two stationary distributions v_1 and v_2 such that $v_1 \neq v_2$.
But we know that stationary distribution is equal to $\frac{1}{|S|}I$ means $v_1 = v_2$.
Hence, proved by contradiction that the uniform distribution is the unique stationary probability distribution for M.

3B

Starting from the stationary distribution, express the probability of returning to the same state as the state at $t = 0$ after $n \in \mathbb{N}$ steps in terms of the eigenvalues of M. Compute the limit of the above probability as $n \to \infty$.

Answer: Let M be the transition matrix of this markov chain and μ_∞ be the stationary probability distribution. We know that probability distribution vector after n steps depends on the initial distribution as follows:

$$\Rightarrow \mu_n^T = \mu_0^T M^n$$

in our case, $\mu_0 = \mu_\infty$

$$\Rightarrow \mu_n^T = \mu_\infty^T M^n$$

and $\mu_n = \mu_\infty$ since we are returning to the initial state after taking n steps.

$$\Rightarrow \mu_\infty^T = \mu_\infty^T M^n$$

Taking transpose on both sides:

$$\Rightarrow (M^T)^n \mu_\infty = \mu_\infty$$

But since M is a symmetric matrix, $M^T = M$

$$\Rightarrow M^n \mu_\infty = \mu_\infty$$

Notice the above equation:

$$\Rightarrow M^n \mu_\infty = 1 \cdot \mu_\infty$$

Hence, μ_∞ is an eigen vector of M^n with corresponding eigen value = 1.

Now,
$\mathbb{P}$(returning to the same state)=
$\sum_{i=1}^{|S|} \mathbb{P}$(initial state = S_i)$\cdot\mathbb{P}$(final state = S_i | initial state = S_i)

Since our initial distribution is $\frac{1}{|S|}I$, therefore
$\mathbb{P}$(initial state = S_i)=$\frac{1}{|S|}$
$\Rightarrow \mathbb{P}$(returning to the same state)=$\frac{1}{|S|}\sum_{i=1}^{|S|} \mathbb{P}$(final state = S_i | initial state = S_i)
Since $(M^n)_{ij}$ represents the probability of reaching state j starting from i in n steps, then
$\mathbb{P}$(final state = S_i | initial state = S_i)= $(M^n)_{ii}$
$\Rightarrow \mathbb{P}$(returning to the same state)=$\frac{1}{|S|}\sum_{i=1}^{|S|}(M^n)_{ii}$
$\Rightarrow \mathbb{P}$(returning to the same state)=$\frac{1}{|S|}tr(M^n)$
also we proved in above part that

$$\Rightarrow tr(M^n) = \sum_{i=1}^{|S|} {\lambda_i}^n$$

therefore
$\Rightarrow \mathbb{P}$(returning to the same state)=$\frac{1}{|S|}\sum_{i=1}^{|S|}\lambda_i{}^n$, where λ_i's are eigen values of M