CS203A: Assignment 2 (Spring 2022)

Submission Deadline: 22 April 2022 23:59 IST. Total Marks:100

---

1. (15+15 marks) **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.

   (a) 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 $\frac{m}{2}$.

   (b) 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.

   Solution:

   (a) Total number of connections $= m$.
   Let the two sets of paper be S_1 and S_2.

   Let say that a connection i is between p_i and p_j.
   Since the papers are distributed randomly, the probability that student i gets S_1 or S_2 is equal(i.e $\frac{1}{2}$).

   If we assume that the Paper that student i gets is independent of the paper that student j gets, means if we assume the paper distribution to be pairwise independent, then

   $$P(Paper_{pi} = S_k \bigcap Paper_{pj} = S_m) = P(Paper_{pi} = S_k) * P(Paper_{pj} = S_m)$$

   where $P(Paper_{pi} = S_k)$ represent the probability that student p_i received S_k paper(which is $\frac{1}{2}$ here).

   Probability that their connection gets disrupted is

Possible Cases		
Paper for p_i	Paper for p_j	Probability
S_1	S_1	$\frac{1}{4}$
S_1	S_2	$\frac{1}{4}$
S_2	S_1	$\frac{1}{4}$
S_2	S_2	$\frac{1}{4}$

   Favourable cases are (S_1, S_2) and (S_2, S_1).

   $$\text{P}(i^{th} \text{ connection breaks}) = 2 * \tfrac{1}{4} = \tfrac{1}{2}$$

   Let us define a **Variable** X_i such that

   $$X_i = \begin{cases} 1 & i^{th}\_connection\_breaks(p = \frac{1}{2}) \\ 0 & otherwise(p = \frac{1}{2}) \end{cases}$$

$$E(X_i) = 1 * \frac{1}{2} = \frac{1}{2}$$

Let N be the number of friendship connections broken.

$$N = \sum_{i=1}^{m} X_i$$

By **Linearity of Expectation**, we know

$$E[N] = \sum_{i=1}^{m} E[X_i]$$

Substituting the value of $E[X_i]$, we get

$$\boxed{E[N] = \frac{m}{2}}$$

(b) Let us try to generate a corollary between the above case of paper distribution and the construction of pairwise independence.

Construction of Pairwise Independence:
Let $X = \{x_1, x_2, ..., x_k\}$ are k mutually independent bits.
Total subsets of X excluding the empty subset $= 2^k - 1$.Let them be $\{S_1, S_2, ..., S_{2^k-1}\}$
Define $Z_i = (\sum_{x_m \in S_i} x_m)\%2$ where $i \in [1, 2^k - 1]$.

We have proved in class that

$$P(Z_i = 1) = P(Z_i = 0) = \frac{1}{2}$$

$$P(Z_i = k/Z_j = m) = P(Z_i = k) = \frac{1}{2}$$

This is similar to the above condition where we say

$$P(Student\_i = Paper\_1) = P(Student\_i = Paper\_2) = \frac{1}{2}$$

$$P(Student\_i = Paper\_k)/P(Student\_j = Paper\_m) = P(Student\_i = Paper\_k) = \frac{1}{2}$$

So let us convert our above problem into the pairwise independent sets we have generated.
$Z_i = 0$ means that student i gets Paper1.
$Z_i = 1$ means that student i gets Paper2.

We have n students, so we want $\boxed{k = log_2(n+1)}$ independent bits to generate n pairwise independent Z.

In the above case when we distributed papers randomly to students, the expected number of broken bonds $= \frac{m}{2}$. In that case the allocation of papers was pairwise independent. So is the case here.

$$E(broken\_friendship\_bonds) = \frac{m}{2}$$

Now in all possible set of $k(log_2(n+1) \approx log_2(n))$ mutually independent bits, $\exists$ at least one set A , such that the number of broken bonds(N(Broken Bonds) $\geq \frac{m}{2}$).

Algorithm for Allocation of Papers:

i. Let M contain all the possible sets of k mutually independent bits.

$$|M| = 2^k \approx 2^{log_2 n} \approx n$$

$\exists A \in M$, **such that N(Broken Bonds)** $\geq \frac{m}{2}$

ii. We will iterate over all entries of M(**Time taken :** $\mathcal{O}(n)$) (which are basically values assigned to k mutually independent bits).
Calculate Z_i for that distribution. Z_i represent set of Paper with Student i.(**Time Taken :** $\mathcal{O}(nlog_2 n)$) as total number of Z_i is n and maximum time to calculate will be $log_2 n$, the size of largest subset.

iii. For each allocation of Z_i, we will then count the number of bonds broken. This will require us to traverse over all the edges(can see students as vertices and bond between them as edges). (**Time Taken : O(m)**) as there are m bonds.

iv. If we find N(Broken Bonds) $\geq \frac{m}{2}$, we stop , otherwise we continue in the same way for other set in M.
Since there exists at least one set A, we can gurantee that the algorithm **terminates**.

Time Complexity Analysis:

Time taken = $\mathcal{O}(n(nlog_2 n + m))$
To calculate worst time complexity, m can take maximum value $\binom{n}{2}$ $(m = \mathcal{O}(n^2))$

$$\boxed{\text{Time taken} = \mathcal{O}(n^3)}$$

Hence the algorithm takes poly-N time

□

2. (5+10+10+15 marks) **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:

(a) 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 $\tilde{N}$.

(b) 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)

(c) 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)

(d) 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.

Solution:

(a) Let X be a Random Variable, such that

$$1 \leq X \leq N$$

Since we are drawing tickets from the bag with replacement, each label ticket occurs with equal probability. This means

$$P(X = i) = \frac{1}{N}$$

Let us calculate, Expected value of X

$$E[X] = \sum_{i=1}^{N} iP(X = i)$$

$$= \sum_{i=1}^{N} i * \frac{1}{N}$$

$$= \frac{1}{N} \sum_{i=1}^{N} i$$

$$\boxed{E[X] = \frac{N+1}{2}}$$

Now, Andrew drew out k tickets. Let the label on the tickets be,

$$X_1, X_2, X_3, ..., X_k$$

These are k independent copies of X.

$$Y = \sum_{i=1}^{k} X_i$$

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

$$E[\bar{X}] = \frac{N+1}{2}$$

Let us say that expected value of mean is M.

$$E[\bar{X}] = M$$

$$M = \frac{\tilde{N}+1}{2}$$

$$\boxed{\tilde{N} = 2M - 1}$$

(b) **Hoeffding's inequality:**

$$P(|S_n - E[S_n]| > t) < 2\exp(-\frac{2t^2}{\sum_{i=1}^{n}(b_i - a_i)^2})$$

where $a_i \leq X_i \leq b_i$.

From the above notations,$\forall i$,

$$1 \leq X_i \leq N$$

$$P(|Y - E[Y]| > t) < 2\exp(-\frac{2t^2}{\sum_{i=1}^{k}(N-1)^2})$$

$$P(|Y - E[Y]| > t) < 2\exp(-\frac{2t^2}{k*(N-1)^2})$$

$$P(|\frac{Y}{k} - \frac{E[Y]}{k}| > \frac{t}{k}) < 2\exp(-\frac{2t^2}{k*(N-1)^2})$$

$$P(|\bar{X} - E[\bar{X}]| > \frac{t}{k}) < 2\exp(-\frac{2t^2}{k*(N-1)^2})$$

Let error be N be dN. It is given that

$$dN \leq \delta N$$

From the (a) part,

$$N = \tilde{N}$$

$$N = 2M - 1$$

$$dN = 2 * dM$$

$$\boxed{dM = d\bar{X} \leq \frac{\delta N}{2}}$$

We will find probability that error in calculating is greater than $\frac{\delta N}{2}$. Let it be p. Then the final answer will be $(1-p)$.

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

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

$$P(|\bar{X} - E[\bar{X}]|\frac{\delta N}{2} >) < 2\exp(-\frac{2\frac{(k\delta N)^2}{4}}{k*(N-1)^2})$$

$$P(|\bar{X} - E[\bar{X}]|\frac{\delta N}{2} >) < 2\exp(-\frac{k\delta^2 N^2}{2(N-1)^2})$$

So the lower bound the probability is

$$\boxed{P(Error(N) < \delta N) \geq 1 - 2\exp(-\frac{k\delta^2 N^2}{2(N-1)^2})}$$

(c) Let us calculate the expected value of median.
Properties of Distribution:

i. Distribution is uniform with each value from $\{1, 2, ..., N\}$ occuring with probability $\frac{1}{N}$.

This means that the expected value of median should be the center value(i.e $\frac{N+1}{2}$).

$$E[median] = K = \frac{N+1}{2}$$

$$\tilde{N} = 2K - 1$$

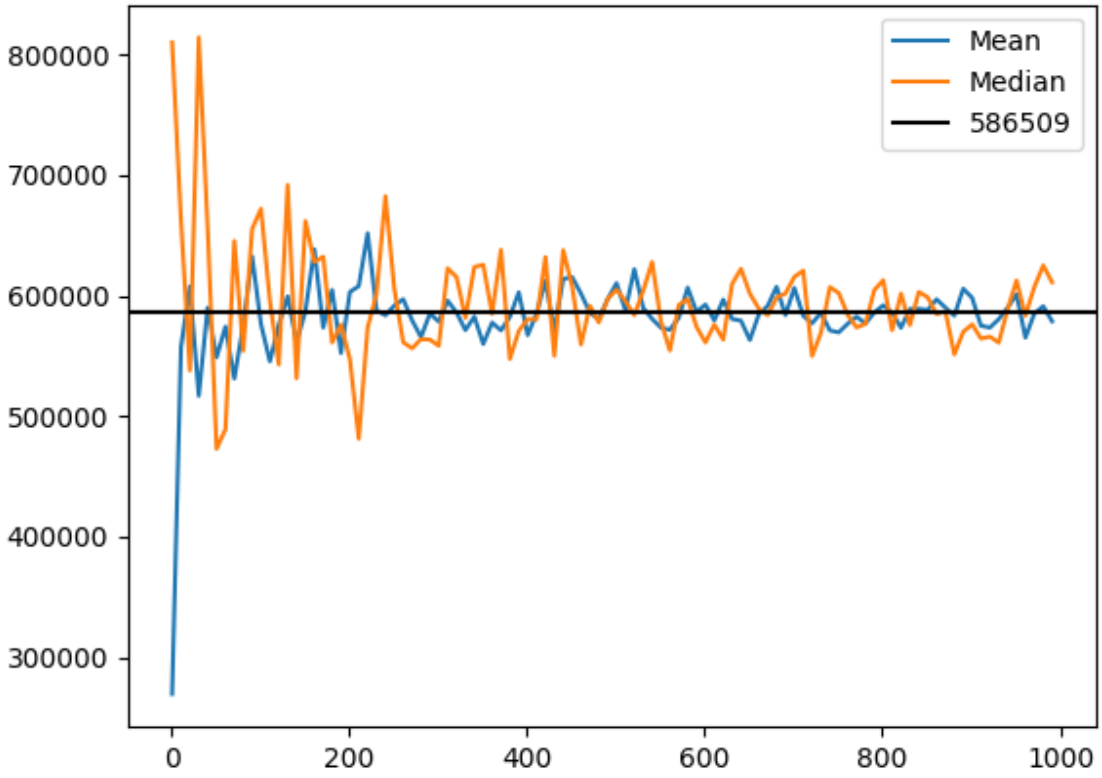

Figure 1: Figure1

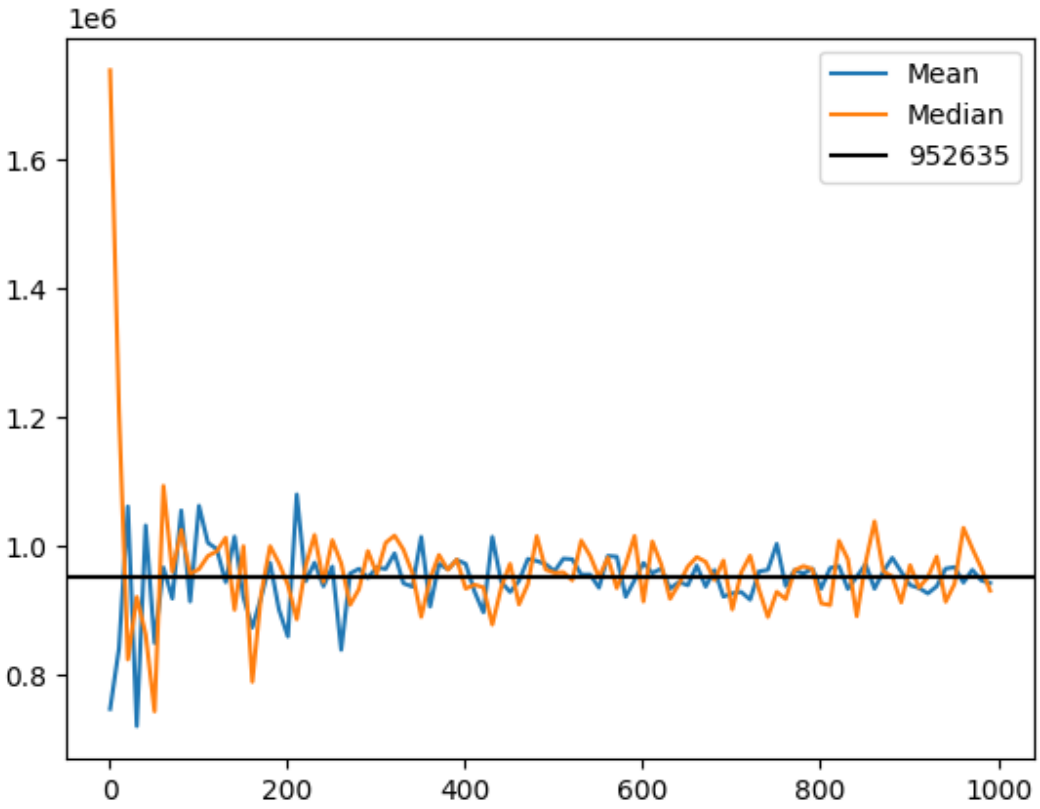

Figure 2: Figure2

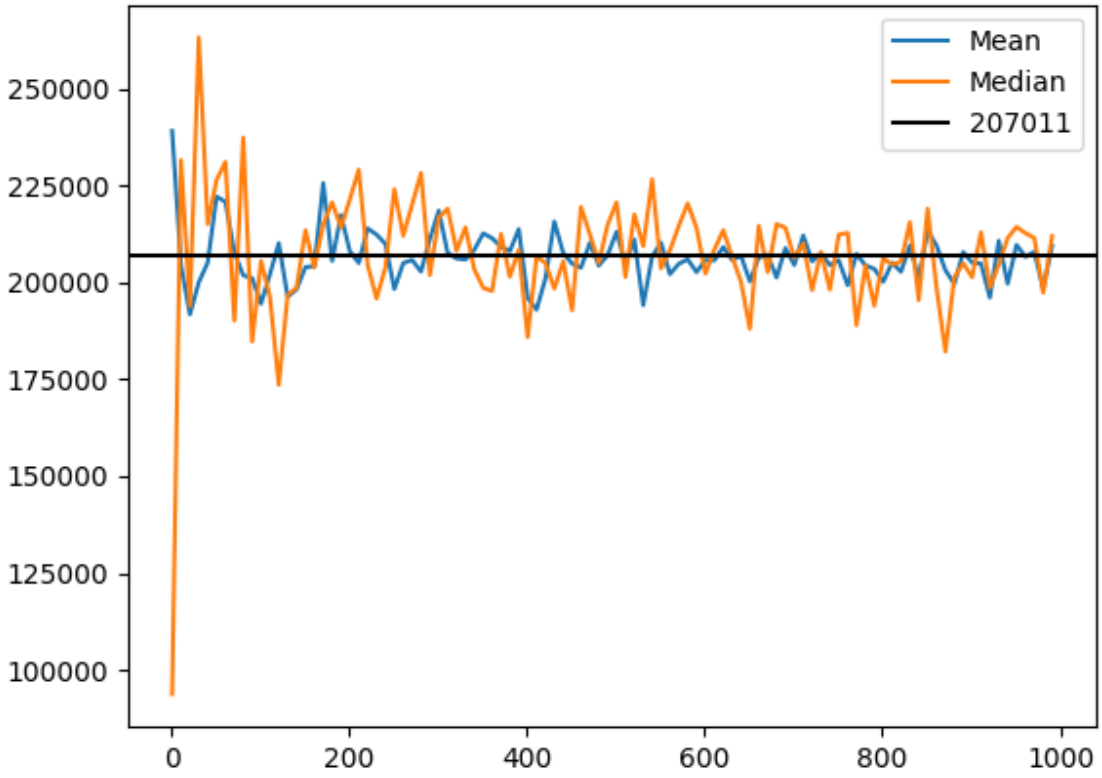

Figure 3: Figure3

□

3. (15+15 marks) **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.

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

(b) 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 $n \to \infty$.

You might find the second part to be easier than the first. Feel free to assume the first part and finish the second part (even when you can't prove the first part).

Solution:

(a) Given transition Matrix M, such that M is symmetric and entry-wise positive.

M is of dimension $|S| * |S|$.

Let λ be eigen value of M and X be the corresponding eigen vector.

$$MX = \lambda X$$

$$X = \begin{bmatrix} \mathrm{x}_1 \\ \mathrm{x}_2 \\ \vdots \\ \mathrm{x}_{|S|} \end{bmatrix} \tag{1}$$

Let k be such that $x_j \leq x_k$,$\forall$ j ,$1 \leq j \leq |S|$

Then equating the k^{th} component on each side, we get

$$\sum_{j=1}^{|S|} M_{kj} x_j = \lambda x_k$$

Since all entries of M are positive, lets increase the value of LHS by keeping all $x_j = x_k$,$\forall j$ since X_k is the largest entry of X.

We also know that sum of rows of M is 1.(M is symmetric)

$$x_k \sum_{j=1}^{|S|} M_{kj} \geq \lambda x_k$$

$$x_k \geq \lambda x_k$$

Take mod on both side

$$|x_k| \geq |\lambda x_k|$$

$$\boxed{|\lambda| \leq 1}$$

This shows that eah eigen value of matrix M is bounded by 1.

Stationary Distribution of M:

M is a regular Markov Chain because:

i. $M_{ij} > 0, \forall i, j$
ii. Sum of rows and columns of M is equal to 1.(Since M is symmetric).
iii. $\exists n$, such that $M_{ij}^n > 0, \forall i, j$. Here n = 1

We know that , if there exists a u such that u satisfies the below property, then u is stationary probability distribution of M.

$$u^T = u^T M$$

In other words, the stationary distribution of M is the eigen vector u of M^T, with eigen value 1.

Let us consider a vector V, where

$$V = \begin{bmatrix} \frac{1}{|S|} \\ \frac{1}{|S|} \\ \vdots \\ \frac{1}{|S|} \end{bmatrix}_{|S| \times 1} \tag{2}$$

Let us check if this vector is eigen vector of matrix M with eigen value 1. This means that V should satisfy

$$MV = V$$

Compare the i^{th} entry on both sides.

$$\sum_{j=1}^{|S|} M_{ij} v_j = v_j$$

Here $v_j = \frac{1}{|S|}, \forall j$

$$\frac{1}{|S|} \sum_{j=1}^{|S|} M_{ij} = \frac{1}{|S|}$$

We know that the sum of rows of M is equal to 1. Hence LHS=RHS.

This proves that V is eigen vector of M with eigen value 1. Hence the stationary probability distribution of M is a uniform distibution V.

Proof that V is unique(there is only one eigen vector with eigen value 1)

Assume that there exists another eigen vector P of M such that eigen value of P is 1.

$$MP = P$$

We know that eigen vectors are linearly independent which means that P will be non-uniform eigen vector.

$$P = \begin{bmatrix} p_1 \\ p_2 \\ \vdots \\ p_{|S|} \end{bmatrix}_{|S| \times 1} \tag{3}$$

We know, $\exists k$ such that $p_j \leq p_k, \forall j$.This is because P is non-uniform.

Equating the k^{th} entry on both sides of $MP = P$, we get

$$\sum_{j=1}^{|S|} M_{kj} p_j = p_k$$

We know that all entries of M are positive, so the largest value LHS can take is when we make all $p_j = p_k$, otherwise the value of LHS will be smaller than RHS .

This means that LHS=RHS, iff $p_j = p_k, \forall j$.This suggests that P is uniform distribution which is a **contradiction**.

Hence, there exists **only one** Eigen vector with eigen value 1.

(b) Starting from a stationary distribution means, the probability of starting from any state is same.

$$P(start\_state = i) = \frac{1}{|S|}$$

Stationary probability distribution vector μ, where

$$\mu = \begin{bmatrix} \frac{1}{|S|} \\ \frac{1}{|S|} \\ \vdots \\ \frac{1}{|S|} \end{bmatrix}_{|S| \times 1} \tag{4}$$

M_{ij}^t gives the probability of going from state i to state j in t steps, given you start from i^{th} state.

Proof: Let us assume that M_{ij}^k represents the probability of going from **state i** to **state j** in k steps, given that the starting state is i.

$$M_{ij}^k = P(S_k = j / S_0 = i)$$

So for time step $k+1$,

$$M_{ij}^{k+1} = \sum_{l=1}^{|S|} M_{il}^k M_{lj}$$

Probability of reaching State j from State i in k+1 steps is given by

$$P(S_{k+1} = j / S_0 = i) = \sum_{l=1}^{|S|} P(S_k = l / S_0 = i) * P(S_{k+1} = j / S_k = l)$$

$$P(S_k = l / S_0 = i) = M_{il}^k$$

We know that M is a homogenous Markov Distribution Matrix, which means

$$P(S_{k+1} = j / S_k = l) = P(S_1 = j / S_0 = l) = M_{lj}$$

$$P(S_{k+1} = j / S_0 = i) = \sum_{l=1}^{|S|} M_{il}^k M_{lj} = M_{ij}^{k+1}$$

Hence M_{ij}^{k+1}, gives the probability of going from state i to state j in $k+1$ steps, given you start from i^{th} state.

We want to calculate the probability(P) to return to the same state as the state at $t = 0$ after $n \in N$ steps.

$$P = \sum_{i=1}^{|S|} P(start\_state = i) M_{ii}^n$$

$$P = \frac{1}{|S|} \sum_{i=1}^{|S|} M_{ii}^n$$

$\sum_{i=1}^{|S|} M_{ii}^n$ is the sum of diagonal elements of M^n and by properties of eigen vectors and eigen values we know that, **The sum of Eigen Values of matrix A is equal to the trace of matrix A.**So,

$$\sum_{i=1}^{|S|} M_{ii}^n = Sum\_of\_eigen\_values\_of\_M^n$$

If a matrix M has eigen values $\{\lambda_1, \lambda_2, ..., \lambda_{|S|}\}$, then M^n has eigen values $\{\lambda_1^n, \lambda_2^n, ..., \lambda_{|S|}^n\}$.

$$\sum_{i=1}^{|S|} M_{ii}^n = \sum_{i=1}^{|S|} \lambda_i^n$$

$$\boxed{P = \frac{1}{|S|} \sum_{i=1}^{|S|} \lambda_i^n}$$

As $n \to \infty$, as all $|\lambda_i| < 1$, except one (let it be $\lambda_1 = 1$)

$$\lim_{n \to \infty} \lambda_i^n \to 0, \forall i, i \neq 1$$

As $n \to \infty$,

$$P = \lim_{n \to \infty} (\frac{1}{|S|})(1 + \sum_{i=2}^{|S|} \lambda_i^n)$$

$$P = \frac{1}{|S|} + \lim_{n \to \infty} (\sum_{i=2}^{|S|} \lambda_i^n)$$

$$\boxed{P = \frac{1}{|S|}}$$

□