1 Warmup with PageRank and stationary distributions:

(a)

(i) calculate stationary distribution:

using $\pi P = \pi$, we have,

$$(\pi_1, \pi_2, \pi_3)\begin{pmatrix} 0 & 3/8 & 5/8 \\ 2/3 & 1/4 & 1/12 \\ 4/9 & 0 & 5/9 \end{pmatrix} = (\pi_1, \pi_2, \pi_3)$$

then:
$$\begin{cases} \frac{2}{3}\pi_2 + \frac{4}{9}\pi_3 = \pi_1 \\ \frac{3}{8}\pi_1 + \frac{1}{4}\pi_2 = \pi_2 \\ \frac{5}{8}\pi_1 + \frac{1}{12}\pi_2 + \frac{5}{9}\pi_3 = \pi_3 \end{cases} \Rightarrow \begin{cases} \pi_1 = \frac{1}{3} \\ \pi_2 = \frac{1}{6} \\ \pi_3 = \frac{1}{2} \end{cases}$$

and $\pi_1 + \pi_2 + \pi_3 = 1$

therefore, the stationary distribution is:

$$\pi = (1/3, 1/6, 1/2)$$

(ii) Show $\pi_0 P^n$ converges to π (stationary distribution) as $n \to \infty$

We first calculate that the eigen values of P, only one equals 1 and others less than 1 (>0)

Therefore, we can represent P as follows.

$$P = X \Lambda X^{-1}$$

here, $X = (x_1, x_2, x_3)$ when x_i is right eigenvector of P.

$$\Lambda = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ 0 & 0 & \lambda_3 \end{pmatrix}$$ where λ_2, λ_3 are eigen values < 1

therefore,

$$P^n = \underbrace{(X\Lambda X^{-1})(X\Lambda X^{-1}) \cdots (X\Lambda X^{-1})}_{n} = X\Lambda^n X^{-1} = X\begin{pmatrix} 1 & & \\ & \lambda_2^n & \\ & & \lambda_3^n \end{pmatrix} X^{-1}$$

therefore, the stationary distribution is:

$$\pi = \left(\frac{5}{12}, \frac{1}{6}, \frac{1}{12}, \frac{1}{3} \right)$$

(ii) Now we want to show that $\lim_{n\to\infty} \pi_0 P^n$ not always converges.
Similar as (a), we calculate the eigen values of P as construct our Λ

$$\Lambda = \begin{pmatrix} 1 & & & \\ & -1 & & \\ & & 0.5 & \\ & & & -0.5 \end{pmatrix}$$

and $\lim_{n\to\infty} P^n = \lim_{n\to\infty} X \Lambda^n X^{-1} = X \begin{pmatrix} 1 & & & \\ & \lim_{n\to\infty}(-1)^n & & \\ & & 0 & \\ & & & 0 \end{pmatrix} X^{-1}$

where it depends on whether n is oval or even, $\lambda_2 = (-1)^n$ shift from 1 to -1
therefore,

$$\lim_{n\to\infty} \pi_0 P^n = \pi_0 x_1 y_1 + (-1)^n \pi_0 x_2 y_2$$

where x_1, x_2 denotes the 1st, 2nd column of X,
y_1, y_2 denotes the 1st, 2nd row of X^{-1}
we observe that if we choose π_0 wisely that makes $\pi_0 x_2 = 0$.

$x_2 = \begin{pmatrix} 1 \\ -1 \\ 1 \\ -1 \end{pmatrix}$ and $x_1 = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}$, $y_1 = \pi$ (stationary distribution)

★ therefore, if we choose initial $\pi_0 = \left(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4} \right)$
then $(-1)^n \pi_0 x_2 y_2 = 0$, then $\lim_{n\to\infty} \pi_0 P^n = \pi_0 x_1 y_1 = \pi$ converges.

★ if we choose initial $\pi_0 = \left(\frac{1}{2}, 0, \frac{1}{2}, 0 \right)$
then $(-1)^n \pi_0 x_2 y_2 \neq 0$, then $\lim_{n\to\infty} \pi_0 P^n$ does not converge to a single value.

Conclution: $\lim_{n\to\infty} \pi_0 P^n$ does not always converge,
it depends on π_0, when $\pi_0 = \left(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4} \right)$
it converges to π

$$\lim_{n\to\infty} \frac{n}{n+1} = 1, \quad \lim_{n\to\infty} \frac{1}{n+1} = 0$$

which means new added page has very little impact on original web graph.

(b) when adding another page Y that links to X.

$$\tilde{P} = \begin{pmatrix} P & 0 \\ 0 & \begin{matrix} 1 & 0 \\ 1 & 0 \end{matrix} \end{pmatrix}$$

where:

$$(\hat{r}, x, y) = (\hat{r}, x, y)\tilde{G}$$
$$= (\hat{r}, x, y)\left(d\tilde{P} + \frac{1-d}{n+2} \mathbb{1}_{(n+2)\times(n+2)}\right)$$

here we have:

$$\begin{cases} d\hat{r}P + \frac{1-d}{n+2}\mathbb{1}_n^T = \hat{r} \\ d(x+y) + \frac{1-d}{n+2} = x \\ 0 + \frac{1-d}{n+2} = y \end{cases} \Rightarrow \begin{cases} x = \frac{1+d}{n+2} \\ y = \frac{1-d}{n+2} \end{cases}$$

Using the same argument in (a), we then have $\hat{r} = \frac{n}{n+2} r$

Then, the page rank in $n+2$ pages now are:

$$(\hat{r}, x, y) = \left(\frac{n}{n+2} r, \frac{1+d}{n+2}, \frac{1-d}{n+2}\right)$$

Compared with (a),

there is improvement of X pagerank, when n is large enough,

$$\lim_{n\to\infty} \frac{\frac{1+d}{n+2}}{\frac{1}{n+1}} = \lim_{n\to\infty} \frac{1+d}{1} \cdot \frac{n+1}{n+2} = 1+d$$

if $d = \frac{1}{2}$, the pagerank improvement is $\frac{3}{2}$ its original value.

(c) Here we first denote the transition matrix between X, Y, Z is Q, therefore, we have:

$$\tilde{G} = d\begin{pmatrix} P & 0 \\ 0 & Q \end{pmatrix} + \frac{1-d}{n+3}\left(\mathbb{1}_{(n+3)\times(n+3)}\right)$$

where,

$$\lim_{n\to\infty} P^n = \lim_{n\to\infty} X \begin{pmatrix} 1 & & \\ & \lambda_2^n & \\ & & \lambda_3^n \end{pmatrix} X^{-1} = X \begin{pmatrix} 1 & & \\ & \lim_{n\to\infty}\lambda_2^n & \\ & & \lim_{n\to\infty}\lambda_3^n \end{pmatrix} X^{-1}$$

$$= X \begin{pmatrix} 1 & & \\ & 0 & \\ & & 0 \end{pmatrix} X^{-1} = x_1 \cdot (\text{the first row in } X^{-1})$$

The first row in X^{-1} can be π, where $\pi \mathbb{1} = \pi$

As the sum of each row in P equals 1, we can define $x_1 = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$

and

$$P x_1 = \lambda_1 x_1 = x_1 \quad \text{stands.}$$

therefore,

$$\lim_{n\to\infty} \pi_0 P^n = \lim_{n\to\infty} \pi_0 x_1 \pi = \lim_{n\to\infty} \pi_0 \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \pi = \pi$$

here π_0 is initial distribution $\pi_0 = (\pi_{01}, \pi_{02}, \pi_{03})$

we denote $sum(\pi_0) = \pi_{01} + \pi_{02} + \pi_{03}$, therefore $sum(\pi_0) = 1$

and

$$\pi_0 x_1 = (\pi_{01}, \pi_{02}, \pi_{03}) \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} = sum(\pi_0) = 1$$

conclusion: we have $\lim_{n\to\infty} \pi_0 P^n = \pi$, and this is not depend on π_0

(b)

(i) calculate stationary distribution:

Using $\pi P = \pi$, we have,

$$(\pi_1, \pi_2, \pi_3, \pi_4) \begin{pmatrix} 0 & 1/4 & 0 & 3/4 \\ 1/2 & 0 & 1/2 & 0 \\ 0 & 3/4 & 0 & 1/4 \\ 1 & 0 & 0 & 0 \end{pmatrix} = (\pi_1, \pi_2, \pi_3, \pi_4)$$

we have,

$$\begin{cases} 1/2\,\pi_2 + \pi_4 = \pi_1 \\ 1/4\,\pi_1 + 3/4\,\pi_3 = \pi_2 \\ 1/2\,\pi_2 = \pi_3 \\ 3/4\,\pi_1 + 1/4\,\pi_3 = \pi_4 \end{cases} \Rightarrow \begin{cases} \pi_1 = 5/12 \\ \pi_2 = 1/6 \\ \pi_3 = 1/12 \\ \pi_4 = 1/3 \end{cases}$$

and $\pi_1 + \pi_2 + \pi_3 + \pi_4 = 1$

2. Training to be a farmer:

(a) create a web page X which has neither in-links and out-links.

for the new graph, we have:

$$(\tilde{r}, x) = (\tilde{r}, x)\tilde{G} \qquad (1)$$

when $\tilde{G}$ is the new transition matrix.

$$\tilde{G} = \partial \begin{pmatrix} P & O \\ O & 1 \end{pmatrix} + \frac{1-\partial}{n+1}(\mathbb{1}_{(n+1)(n+1)})$$

where O represent all zero matrix.

then we depart (1):

$$\begin{cases} \tilde{r} = \partial \tilde{r} P + \frac{1-\partial}{n+1}(\mathbb{1}_n^T) & (1) \\ x = \partial x + \frac{1-\partial}{n+1} & (2) \end{cases}$$

we have:

$x = \frac{1}{n+1}$ which means in new graph, the page rank of X is $1/_{n+1}$

moreover, as for $\tilde{r}$,

we have,

$$\text{sum}(\tilde{r}) = \tilde{r}\mathbb{1}_n = 1 - x = \frac{n}{1+n}$$

therefore,

$$\tilde{r}\mathbb{1}_{n\times n} = \frac{n}{1+n}\mathbb{1}_n^T$$

substitude $\mathbb{1}_n^T$ to $\frac{n+1}{n}\tilde{r}\mathbb{1}_{n\times n}$ in (1)

we have: $\tilde{r} = \partial \tilde{r} P + \frac{1-\partial}{n+1}\cdot\frac{n+1}{n}\tilde{r}\mathbb{1}_{n\times n}$

$$= \tilde{r}\left(\partial P + \frac{1-\partial}{n}\mathbb{1}_{n\times n}\right) = \tilde{r}G$$

Therefore, we know:

$$r = Gr$$
$$\tilde{r} = G\tilde{r}$$

and $\text{sum}(\tilde{r}) = 1 - x = \frac{n}{n+1}$ $\quad \text{sum}(r) = 1$

Therefore, we have $\tilde{r} = \frac{n}{n+1}r$, $\tilde{r}$ is just the scaling of r, which means when adding a node with no-links and out-links, the pagerank just shrink to its previous $\frac{n}{n+1}$ (for old nodes), when n is very large.

Using $(\tilde{r}, x, y, z) = (\tilde{r}, x, y, z)\,\tilde{G}$

we have:

for example:

for z:

$$d(Q_{13}x + Q_{23}y + Q_{33}z) + \frac{1-d}{n+3} = z$$

$$\Rightarrow \quad z = \frac{dQ_{13}x + dQ_{23}y + \frac{1-d}{n+3}}{1 - Q_{33}}$$

because $Q_{33} < 1$ and also $Q_{13}, Q_{23}, x, y, d > 0$,

we have:

$$z \geq \frac{1-d}{n+3}$$

similar to y we have: $\quad y \geq \frac{1-d}{n+3}$

Back to equation which has all x, y, z:

$$(x, y, z) = (x, y, z)\,dQ + \frac{1-d}{n+3}\mathbb{1}_3^T$$

$$(x, y, z)\begin{pmatrix}1\\1\\1\end{pmatrix} = (x, y, z)\,dQ\begin{pmatrix}1\\1\\1\end{pmatrix} + \frac{1-d}{n+3}(1,1,1)\begin{pmatrix}1\\1\\1\end{pmatrix}$$

$$\Downarrow$$

$$\text{sum}(x, y, z) = d\,\text{sum}(x, y, z) + \frac{1-d}{n+3}\cdot 3$$

$$\cancel{(1-d)}\,\text{sum}(x, y, z) = \frac{3}{n+3}\cancel{(1-d)}$$

Therefore, $x = \text{sum}(x, y, z) - y - z$

$$= \frac{3}{n+3} - y - z$$

$$\leq \frac{3}{n+3} - \frac{2(1-d)}{n+3} = \frac{2d+1}{n+3}$$

x reaches its maxima when Q:

$$Q = \begin{pmatrix}1 & 0 & 0\\1 & 0 & 0\\1 & 0 & 0\end{pmatrix}$$

Which means that Y, Z all connected to X,

then the max pagerank of X

$$x_{max} = \frac{2d+1}{n+3}$$

3. Beyond PageRank:

(a)

We first prove that, for any connected, undirected graph G, The definition 1 (Degree Centrality) and definition 4 (PageRank) are in the same order of importance. That is, the PageRank r_i is proportional to the degree centrality $C_D(i)$, to be simplify, we prove:

$$\frac{r_i}{d_i} = c \text{ (constant)}$$

first, we create transition matrix P, where $P_{ij} = \begin{cases} 1/d_i & \text{if } i \to j \\ 0 & \text{if } i \not\to j \end{cases}$

we also build matrix A,

$$A_{ij} = \begin{cases} 1 & \text{if } i \to j \\ 0 & \text{if } i \not\to j \end{cases}$$

matrix D, where $D_{ii} = \frac{1}{d_i}$

$$D = \begin{pmatrix} \frac{1}{d_1}, & & & \\ & \frac{1}{d_2}, & & \\ & & \cdots & \\ & & & \frac{1}{d_N} \end{pmatrix}$$

Then we have:

$$P = DA$$

$$P^T = A^T D^T$$

because A and D are all symmetrix, then,

$$P^T = AD$$

Therefore: $rD = (r_1, r_2, \cdots r_n) \begin{pmatrix} 1/d_1, & & & \\ & 1/d_2, & & \\ & & \cdots & \\ & & & 1/d_n \end{pmatrix}$

$$= \left(\frac{r_1}{d_1}, \frac{r_2}{d_2}, \cdots \frac{r_n}{d_n}\right)$$

we want to prove $rD = (\underbrace{c, c, \cdots c}_{n})$

Here, using stationary distribution:

$$r = rP$$

$$rD = rPD = rD\underline{AD} = rDP^T \quad \text{we denote } rD = x^T$$

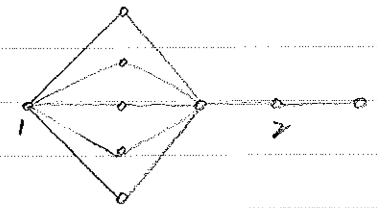
then

$$x^T = x^T P^T$$

$$x = Px$$

Because sum$(P_i) = 1$ (each row), we have $x = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}$ is actually eigenvector of P, then as $rD = x^T = (1, 1, \cdots, 1)$, we prove that $\frac{r_i}{d_i} = c$ (constant)

(3) Difference Between 2 & 3



for node 1:

$$C_C(1) = \frac{n-1}{\sum_{j\neq 1} l(1,j)} = \frac{9-1}{5\times1+2+3+4} = \frac{8}{14}$$

$$C_B(1) = \frac{\sum_{j,k:\, j\neq k,\, j,k\neq 1} \frac{P_1(j,k)}{P(j,k)}}{\binom{n-1}{2}} = \frac{\binom{5}{2}\times\frac{1}{2}}{\binom{8}{2}} = \frac{5}{28}$$

for node 2:

$$C_C(2) = \frac{9-1}{2\times1+5\times2+1\times3} = \frac{8}{15}$$

$$C_B(2) = \frac{7}{28}$$

Therefore: $C_C(1) > C_C(2)$

$C_B(1) < C_B(2)$

(b) For each centrality meatures,

1. degree centrality: Social Network

Let G denotes a social network where each node represents a person, the edge between two nodes indicates that these two are friends. If a person has more degree which means he/she has more friends, that means he/she has more influence to the whole social network and thus is more important.

2. closeness centrality: Information Network

degree centrality has limitations: the measure does not take into consideration the global structure of the network. For example, although are node has many adjacencies, it might not be in the position to reach other quickly to access recourses.

let G denotes a Information Network, where each path between nodes have different weights (costs). Therefore, the node with

(d) Now if we add links from my page X (or Y and Z) to older pages, then the transition matrix of the $n+3$ nodes becomes

$$\tilde{G} = d\begin{pmatrix} P & 0 \\ U & \tilde{Q} \end{pmatrix} + \frac{1-d}{n+3} \mathbb{1}_{(n+3)\times(n+3)}$$

since in this situation, $\text{sum}(U+\tilde{Q}) = 1$,

therefore, we have:

$$\tilde{Q}\mathbf{1}_3 < \mathbf{1}_3$$

similar to calculation in (c)

$$(\tilde{r}, x, y, z) = (\tilde{r}, x, y, z)\,\tilde{G}$$

$$= d(\tilde{r}P + (x,y,z)U, (x,y,z)\tilde{Q}) + \frac{1-d}{n+3}\mathbf{1}^T_{n+3}$$

where,

$$(x,y,z) = d(x,y,z)\tilde{Q} + \frac{1-d}{n+3}(1,1,1)$$

for z:

$$d(Q_{13}x + Q_{23}y + Q_{33}z) + \frac{1-d}{n+3} = z$$

$$\Rightarrow z = \frac{dQ_{13}x + dQ_{23}y + \frac{1-d}{n+3}}{1-Q_{33}} \geq \frac{1-d}{n+3}$$

same as $y \geq \frac{1-d}{n+3}$

$$\text{sum}(x,y,z) = x+y+z = (x,y,z)\mathbf{1}_3$$

$$= d(x,y,z)\tilde{Q}\mathbf{1}_3 + \frac{1-d}{n+3}(1,1,1)\mathbf{1}_3$$

$$< d(x,y,z)\mathbf{1}_3 + 3\frac{1-d}{n+3}$$

$$x+y+z < \frac{3}{n+3}$$

therefore $x < \frac{3}{n+3} - \frac{2(1-d)}{n+3} = \frac{2d+1}{n+3}$, which means adding links from X to older pages will not improve the pagerank of X, it will reduce it conversely. Also, situation will not change if Y or Z is linked to older pages.

Then we give counterexample to show difference between 1(4) and 2 and 3.

(1). Difference Between 1 & 2:

for node 1: $C_D(1) = 3/(7-1) = 1/2$

$$C_C(1) = \frac{7-1}{\sum_{j\neq 1} l(1,j)} = \frac{6}{1\times 3+2+2\times 3} = \frac{6}{11}$$

for node 2: $C_D(2) = 2/(7-1) = 1/3$

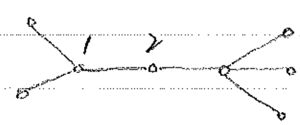
$$C_C(2) = \frac{7-1}{2\times 1+2\times 4} = \frac{6}{10}$$

therefore:

$C_D(1) > C_D(2)$

$C_C(1) < C_C(2)$

(2) Difference Between 1 & 3:



for node 1: $C_D(1) = 3/(7-1) = 1/2$

$$C_B(1) = \frac{\sum_{j,k: j\neq k, k\neq i} \frac{P_i(j,k)}{P(j,k)}}{\binom{n-1}{2}} = \frac{11}{28}$$

for node 2: $C_D(2) = 2/(7-1) = 1/3$

$$C_B(2) = \frac{12}{28}$$

therefore: $C_D(1) > C_D(2)$

$C_B(1) < C_B(2)$

shortest average distance between all other nodes is the most important and has more centrality since it is easier to access any other nodes (information) in the network.

3. betweenness centrality, train network.

let G be a train network of a country, each node denotes a city train station, edges between two nodes are train way between 2 stations. therefore, if a node$^{(A)}$ (city) is important, it must be in the shortest paths of other stations, and meanwhile, the path excluded node$^{(A)}$ is very few, then, if station node$^{(A)}$ is broken and need repairment, it will affect the transportation of many travellers.

4. pagerank centrality: web page network

let G be a web page network and each node is a single web page, Then, by defintion, a page with high PageRank may have more in-links and connected to other important pages and thus has a higher visited frequency. Therefore, pageRank represents the importance of a webpage.

(e) To improve the pageRank of web page X, according to (a) & (b), I will generate many new web pages $(A_1, A_2, \cdots A_{m-1})$ links to X
we have the pageRank of X:

$$x = \frac{1+(m-1)d}{m+n}, \quad a_j = \frac{1-d}{m+n} \quad (j=1,2,\cdots,m-1)$$

(page rank of A_j)

and also X is linked to no page (eg. older page) as stated in (d), that links to older page will not increase but will decrease the pagerank of X.

Prove.

$$\tilde{G} = d\begin{pmatrix} P & 0 \\ 0 & Q \end{pmatrix} + \frac{1-d}{n+m} \mathbb{1}_{(n+m)\times(n+m)}$$

where Q is the transition matrix of the new pages (m)

$$(\tilde{r}, v) = (\tilde{r}, v)\tilde{G}$$
$$= d(\tilde{r}P + vQ) + \frac{1-d}{n+m} 1^T_{n+m}$$

$$\Rightarrow v = dvQ + \frac{1-d}{n+m} 1^T_m \qquad (v = (x, a_1, a_2, a_3, \cdots a_{m-1}))$$

$$sum(v) = v1_m = dvQ1_m + \frac{1-d}{n+m} 1^T_m 1_m = dv1_m + \frac{(1-d)m}{m+n}$$

$$\Rightarrow sum(v) = \frac{m}{m+n}$$

Using argument in (c)

$$x = \frac{m}{n+m} - \sum_{j=1}^{m-1} a_j \leq \frac{m}{n+m} - \frac{(1-d)(m-1)}{n+m} = \frac{1+(m-1)d}{m+n}$$

we have:
when choosing $Q = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 1 & & & \\ \vdots & & & \\ 1 & 0 & \cdots & 0 \end{pmatrix}$

$$x_{max} = \frac{1+(m-1)d}{m+n}$$