HW3: PageRank & MapReduce

De Huang UID: 2018510
email: dhuang@caltech.edu

1 Warmup with PageRank and stationary distributions

Let $P \in \mathbb{R}_+^{n\times n}$ be a nonnegative stochastic matrix such that

$$\sum_{j=1}^{n} P_{ij} = 1, \quad i = 1, 2, \ldots, n. \tag{1}$$

We have

1. $\lambda = 1$ is an eigenvalue of P.
   Let $\mathbf{1} = (1, 1, \cdots, 1)^T \in \mathbb{R}^{n\times 1}$, then due to property (1) we have

   $$P\mathbf{1} = \mathbf{1}.$$

   Therefore $\lambda = 1$ is an eigenvalue of P.

2. If μ is an eigenvalue of P, then $|\mu| \le 1$.
   Let μ be an eigenvalue of P, and $x = (x_1, x_2, \cdots, x_n) \in \mathbb{R}^{1\times n}\backslash\{\mathbf{0}\}$ be a left eigenvector associated with μ. Then

   $$xP = \mu x.$$

   Define $|x| = (|x_1|, |x_2|, \cdots, |x_n|)$, then

   $$|\mu||x|\mathbf{1} = |\mu x\mathbf{1}| = |xP\mathbf{1}| = |x\mathbf{1}| \le |x|\mathbf{1}.$$

   Since $x \ne 0$, $|x|\mathbf{1} \ge 0$, we have $\mu \le 1$.

3. There exists some $\pi \in \mathbb{R}^{1\times n}\backslash\{\mathbf{0}\}$ such that $\pi P = \pi$, $\pi\mathbf{1} = 1$, $\pi \ge 0$(entrywise).
   Let $x = (x_1, x_2, \cdots, x_n) \in \mathbb{R}^{1\times n}\backslash\{\mathbf{0}\}$ be a left eigenvector of P associated with $\lambda = 1$. If $x \ge 0$(or $x \le 0$), we can simply take

   $$\pi = \frac{x}{x\mathbf{1}}.$$

   Other wise, without loss of generality, we may assume that $x_i > 0, i = 1, 2, \cdots, k$, $x_j \le 0, j = k+1, k+2, \cdots, n$ for some $1 \le k \le n$. Let $y = (x_1, x_2, \cdots, x_k) \in \mathbb{R}^{1\times k}$, $z = (x_{k+1}, x_{k+2}, \cdots, x_n) \in \mathbb{R}^{1\times n-k}$. Consider the division of P

   $$P = \left[\begin{array}{cc} A_{k\times k} & B_{k\times n-k} \\ C_{n-k\times k} & D_{n-k\times n-k} \end{array}\right].$$

Since $xP = x$, we have

$$yA + zC = y, \quad yB + zD = z.$$

Since $y > 0$, $z \leq 0$, $B, C \geq 0$, we have

$$yB\mathbf{1}_k \geq 0, \quad zC\mathbf{1}_{n-k} \leq 0.$$

Since

$$\sum_{j=1}^{k} A_{ij} \leq \sum_{j=1}^{n} P_{ij} = 1, \quad i = 1, 2, \cdots, k,$$

$$\sum_{j=1}^{n-k} D_{ij} \leq \sum_{j=1}^{n} P_{i+k,j} = 1, \quad i = 1, 2, \cdots, n-k,$$

we have

$$yA\mathbf{1}_k \leq y\mathbf{1}_k, \quad zD\mathbf{1}_{n-k} \geq z\mathbf{1}_{n-k}.$$

Then

$$y\mathbf{1}_k = yA\mathbf{1}_k + zC\mathbf{1}_k \leq yA\mathbf{1}_k \leq y\mathbf{1}_k,$$

$$z\mathbf{1}_{n-k} = yB\mathbf{1}_{n-k} + zD\mathbf{1}_{n-k} \geq zD\mathbf{1}_{n-k} \geq z\mathbf{1}_{n-k}.$$

That is, we can only have

$$zC = 0, \quad yB = 0, \quad yA = y, \quad zD = z.$$

Now let

$$\pi = \frac{(y, \mathbf{0}_{n-k})}{y\mathbf{1}_k} \in \mathbb{R}^{1\times k},$$

then $\pi \geq 0$, $\pi\mathbf{1} = 1$, and

$$\pi P = \frac{(y, \mathbf{0}_{n-k})}{y\mathbf{1}_k} \left[\begin{array}{cc} A & B \\ C & D \end{array} \right] = \frac{(yA, yB)}{y\mathbf{1}_k} = \frac{(y, \mathbf{0}_{n-k})}{y\mathbf{1}_k} = \pi.$$

4. Let $\lambda_1, \lambda_2, \cdots, \lambda_n$ be all eigenvalues of P. Assume that $|\lambda_1| = |\lambda_2| = \cdots = |\lambda_k| = 1$, $|\lambda_{k+1}|, |\lambda_{k+2}|, \cdots, |\lambda_n| < 1$, then P has a Jordan normal formal

$$P = X \left[\begin{array}{cc} J_{k\times k} & \mathbf{0}_{k\times n-k} \\ \mathbf{0}_{n-k\times k} & \tilde{J}_{n-k\times n-k} \end{array} \right] X^{-1},$$

where $J, \tilde{J}$ are Jordan matrices, the diagonal entries of J are $\lambda_1, \lambda_2, \cdots, \lambda_k$, the diagonal entries of $\tilde{J}$ are $\lambda_{k+1}, \lambda_{k+2}, \cdots, \lambda_n$; the j_{th} column of X is a right eigenvector of P associated with λ_j, $j = 1, 2, \cdots, n$, the i_{th} row of X^{-1} is a left eigenvector of P associated with λ_i, $i = 1, 2, \cdots, n$. Then

$$\begin{aligned} \lim_{N\to\infty} P^N &= \Big(\lim_{N\to\infty} X \left[\begin{array}{cc} J & \mathbf{0} \\ \mathbf{0} & \tilde{J} \end{array} \right] X^{-1} \Big)^N \\ &= X \left[\begin{array}{cc} \lim_{N\to\infty} J^N & \mathbf{0} \\ \mathbf{0} & \lim_{N\to\infty} \tilde{J}^N \end{array} \right] X^{-1} \\ &= X \left[\begin{array}{cc} \lim_{N\to\infty} J^N & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{array} \right] X^{-1}. \end{aligned}$$

If $\lambda_1 = \lambda_2 = \cdots = \lambda_k = 1$, then $\lim_{N\to\infty} J^N$ exists and is an upper triangular matrix with diagonal entries equal to 1. If at least one of $\{\lambda_1 = \lambda_2 = \cdots = \lambda_k\}$ is -1, then $\lim_{N\to\infty} J^N$ doesn't exist.

In particular, if $\lambda_1 = 1$, $|\lambda_2|, |\lambda_3|, \cdots, |\lambda_n| < 1$, we may assume that the first column of X is $\mathbf{1}$, and the first row of X^{-1} is π such that $\pi \geq 0$, $\pi\mathbf{1} = 1$. Then

$$\lim_{N\to\infty} P^N = X \begin{bmatrix} 1 & \mathbf{0} \\ \mathbf{0} & \lim_{N\to\infty} \tilde{J}^N \end{bmatrix} X^{-1} = X \begin{bmatrix} 1 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} X^{-1} = \mathbf{1}\pi.$$

Therefore, for any $\pi_0 \in \mathbb{R}^{1\times n}$ such that $\pi_0 \geq 0$, $\pi_0\mathbf{1} = 1$, we have

$$\lim_{N\to\infty} \pi_0 P^N = \pi_0 \mathbf{1}\pi = \pi.$$

(a) $P = \begin{pmatrix} 0 & 3/8 & 5/8 \\ 2/3 & 1/4 & 1/12 \\ 4/9 & 0 & 5/9 \end{pmatrix}$. By solving $\pi P = \pi$, $\pi\mathbf{1} = 1$, we obtain

$$\pi = (\frac{1}{3}, \frac{1}{6}, \frac{1}{2}).$$

The eigenvalues of P are $\lambda_1 = 1, \lambda_2 = 0.3543, \lambda_3 = -0.5488$. Using the previous results, we have

$$\lim_{N\to\infty} \pi_0 P^N = \pi, \quad \text{for } \forall \pi_0 \text{ s.t. } \pi_0 \mathbf{1} = 1.$$

That is to say, the convergence does not depend on π_0. No matter what distribution π_0 we start from, we finally end up at the stationary distribution π, which explains the word "stationary".

(b) $P = \begin{pmatrix} 0 & 1/4 & 0 & 3/4 \\ 1/2 & 0 & 1/2 & 0 \\ 0 & 3/4 & 0 & 1/4 \\ 1 & 0 & 0 & 0 \end{pmatrix}$. By solving $\pi P = \pi$, $\pi\mathbf{1} = 1$, we obtain

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

The eigenvalues of P are $\lambda_1 = 1, \lambda_2 = 0.5, \lambda_3 = -0.5, \lambda_4 = -1$. Since we have $\lambda_4 = -1$, by the previous results, the limit $\lim_{N\to\infty} P^N$ does't exist, thus we can not expect $\lim_{N\to\infty} \pi_0 P^N = \pi$ for $\forall \pi_0$. Indeed we have

$$\lim_{N\to\infty} P^{2N} = \begin{pmatrix} 5/6 & 0 & 1/6 & 0 \\ 0 & 1/3 & 0 & 1/6 \\ 5/6 & 0 & 1/6 & 0 \\ 0 & 1/3 & 0 & 1/6 \end{pmatrix},$$

$$\lim_{N\to\infty} P^{2N+1} = \begin{pmatrix} 0 & 1/3 & 0 & 1/6 \\ 5/6 & 0 & 1/6 & 0 \\ 0 & 1/3 & 0 & 1/6 \\ 5/6 & 0 & 1/6 & 0 \end{pmatrix},$$

therefore,

$$\lim_{N\to\infty}(\frac{1}{4}\ ,\ \frac{1}{4}\ ,\ \frac{1}{4}\ ,\ \frac{1}{4})P^N = (\ \frac{5}{12}\ ,\ \frac{1}{6}\ ,\ \frac{1}{12}\ ,\ \frac{1}{3}),$$
$$\lim_{N\to\infty}(\frac{1}{2}\ ,\ 0\ ,\ \frac{1}{2}\ ,\ 0)P^{2N} = (\frac{5}{6}\ ,\ 0\ ,\ \frac{1}{6}\ ,\ 0),$$
$$\lim_{N\to\infty}(\frac{1}{2}\ ,\ 0\ ,\ \frac{1}{2}\ ,\ 0)P^{2N+1} = (0\ ,\ \frac{1}{3}\ ,\ 0\ ,\ \frac{2}{3}).$$

Indeed, let $y \in \mathbb{R}^{1\times 4}$ and $x \in \mathbb{R}^{4\times 1}$ be the left eigenvector and right eigenvector, respectively, of P associated with $\lambda_4 = -1$ such that $yx = 1$, then we have

$$\lim_{N\to\infty} P^N = \mathbf{1}\pi + (-1)^N xy.$$

If the initial distribution π_0 satisfies $\pi_0 x = 0$, then

$$\lim_{N\to\infty} \pi_0 P^N = \pi_0\mathbf{1}\pi + (-1)^N \pi_0 xy = \pi.$$

Otherwise

$$\lim_{N\to\infty} \pi_0 P^{2N} = \pi_0\mathbf{1}\pi + \pi_0 xy = \pi + \pi_0 xy,$$
$$\lim_{N\to\infty} \pi_0 P^{2N+1} = \pi_0\mathbf{1}\pi - \pi_0 xy = \pi - \pi_0 xy.$$

But we can still get π by doing

$$\lim_{N\to\infty} \frac{1}{2}\pi_0(P^{2N} + P^{2N+1}) = \pi.$$

2 Training to be a farmer

(a) After adding the new page X, the new transition matrix for the $n+1$ nodes is

$$\widetilde{G} = \alpha\widetilde{P} + \frac{1-\alpha}{n+1}(\mathbf{1}_{(n+1)\times(n+1)}),$$

where

$$\widetilde{P} = \begin{pmatrix} P & \mathbf{0} \\ \mathbf{0} & 1 \end{pmatrix}.$$

Since $(\tilde{r}, x) = (\tilde{r}_1, \tilde{r}_2, \cdots, \tilde{r}_n, x)$ is the stationary distribution of $\widetilde{G}$, we have

$$(\tilde{r}, x)\mathbf{1}_{n+1} = \tilde{r}\mathbf{1}_n + x = 1,$$

where $\mathbf{1}_n = (1, 1, \cdots, 1)^T \in \mathbb{R}^{n\times 1}$, and

$$\begin{aligned}(\tilde{r}, x) &= (\tilde{r}, x)\widetilde{G} \\ &= \alpha(\tilde{r}, x)\widetilde{P} + \frac{1-\alpha}{n+1}(\tilde{r}, x)(\mathbf{1}_{(n+1)\times(n+1)}) \\ &= \alpha(\tilde{r}P, x) + \frac{1-\alpha}{n+1}\mathbf{1}_{n+1}^T.\end{aligned}$$

That is

$$x = \alpha x + \frac{1-\alpha}{n+1} \quad \Rightarrow \quad x = \frac{1}{n+1},$$

$$\tilde{r} = \alpha \tilde{r} P + \frac{1-\alpha}{n+1}\mathbf{1}_n^T.$$

Then we have

$$\tilde{r}\mathbf{1}_n = 1 - x = \frac{n}{n+1} \quad \Rightarrow \quad \tilde{r}\mathbf{1}_{n\times n} = \frac{n}{n+1}\mathbf{1}_n^T.$$

Therefore

$$\tilde{r} = \alpha \tilde{r} P + \frac{1-\alpha}{n}\frac{n}{n+1}\mathbf{1}_n^T = \alpha \tilde{r} P + \frac{1-\alpha}{n}\tilde{r}\mathbf{1}_{n\times n} = \tilde{r}\big(\alpha P + \frac{1-\alpha}{n}\mathbf{1}_{n\times n}\big).$$

That is

$$\frac{n+1}{n}\tilde{r}\mathbf{1}_n = 1, \quad \frac{n+1}{n}\tilde{r} = \frac{n+1}{n}\tilde{r}G,$$

thus $(n+1)\tilde{r}/n$ is a stationary distribution of G. But since G is strict positive, the stationary distribution of G is unique, hence we have

$$\frac{n+1}{n}\tilde{r} = r \quad \Rightarrow \quad \tilde{r} = \frac{n}{n+1}r.$$

(b) After adding the page Y, we still use $\widetilde{G}$ to denote the transition matrix of the $n+2$ nodes, use $\tilde{r}$ to denote the new PageRanks of the n node old pages, and use x, y to denote the PageRanks of X, Y. Then we have

$$\widetilde{G} = \alpha \begin{pmatrix} P & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & 1 & 0 \\ \mathbf{0} & 1 & 0 \end{pmatrix} + \frac{1-\alpha}{n+2}\mathbf{1}_{(n+2)\times(n+2)},$$

and

$$\begin{aligned}(\tilde{r}, x, y) &= (\tilde{r}, x, y)\widetilde{G} \\ &= \alpha(\tilde{r}P, x+y, 0) + \frac{1-\alpha}{n+2}(\tilde{r}, x, y)\mathbf{1}_{(n+2)\times(n+2)} \\ &= \alpha(\tilde{r}P, x+y, 0) + \frac{1-\alpha}{n+2}\mathbf{1}_{n+2}^T.\end{aligned}$$

Therefore

$$y = \frac{1-\alpha}{n+2},$$

$$x = \alpha x + \alpha y + \frac{1-\alpha}{n+2} \quad \Rightarrow \quad x = \frac{1+\alpha}{n+2}$$

$$\tilde{r}\mathbf{1}_n = 1 - x - y = \frac{n}{n+2}$$

$$\tilde{r} = \alpha \tilde{r} P + \frac{1-\alpha}{n+2}\mathbf{1}_n^T = \alpha \tilde{r} P + \frac{1-\alpha}{n}\tilde{r}\mathbf{1}_{n\times n} = \tilde{r}G.$$

Similarly we have

$$\tilde{r} = \frac{n}{n+2}r.$$

If $\frac{1+\alpha}{n+2} > \frac{1}{n+1}$, then the PageRank of X in (b) is better than in (a). Indeed, for large n, we have

$$\frac{x_{(b)}}{x_{(a)}} = \frac{n+1}{n+2}(1+\alpha) \approx 1+\alpha.$$

(c) After adding the page Z, we still use $\widetilde{G}$ to denote the transition matrix of the $n+3$ nodes, use $\tilde{r}$ to denote the new PageRanks of the n node old pages, and use x, y, z to denote the PageRanks of X, Y, Z. Assume that

$$\widetilde{G} = \alpha \begin{pmatrix} P & \mathbf{0} \\ \mathbf{0} & Q \end{pmatrix} + \frac{1-\alpha}{n+3}\mathbf{1}_{(n+3)\times(n+3)},$$

where $Q \in \mathbb{R}^{3\times 3}$ is the transition matrix for (x, y, z). Then we have

$$\begin{aligned} (\tilde{r}, x, y, z) &= (\tilde{r}, x, y, z)\widetilde{G} \\ &= \alpha(\tilde{r}P, (x, y, z)Q) + \frac{1-\alpha}{n+3}\mathbf{1}_{n+3}^T, \end{aligned}$$

$$\begin{aligned} \Rightarrow \qquad \tilde{r} &= \alpha\tilde{r}P + \frac{1-\alpha}{n+3}\mathbf{1}_n^T, \\ (x, y, z) &= \alpha(x, y, z)Q + \frac{1-\alpha}{n+3}(1, 1, 1). \end{aligned}$$

We need to choose a proper transition matrix Q so as to maximize x. Notice that

$$\begin{aligned} x + y + z &= (x, y, z)\mathbf{1}_3 \\ &= \alpha(x, y, z)Q\mathbf{1}_3 + \frac{1-\alpha}{n+3}\mathbf{1}_3^T\mathbf{1}_3 \\ &= \alpha(x, y, z)\mathbf{1}_3 + \frac{3-3\alpha}{n+3}, \end{aligned}$$

$$\Rightarrow \qquad x + y + z = \frac{3}{n+3},$$

and that

$$\begin{aligned} y &= \alpha(Q_{12}x + Q_{22}y + Q_{32}z) + \frac{1-\alpha}{n+3}, \\ z &= \alpha(Q_{13}x + Q_{23}y + Q_{33}z) + \frac{1-\alpha}{n+3}, \end{aligned}$$

$$\begin{aligned} \Rightarrow \qquad y &= \frac{1}{1-\alpha Q_{22}}\left[\alpha(Q_{12}x + Q_{32}z) + \frac{1-\alpha}{n+3}\right] \geq \frac{1-\alpha}{n+3}, \\ z &= \frac{1}{1-\alpha Q_{33}}\left[\alpha(Q_{13}x + Q_{23}y) + \frac{1-\alpha}{n+3}\right] \geq \frac{1-\alpha}{n+3}. \end{aligned}$$

Therefore

$$x = \frac{3}{n+3} - y - z \leq \frac{3}{n+3} - 2\frac{1-\alpha}{n+3} = \frac{1+2\alpha}{n+3}.$$

Especially, if we choose

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

we have

$$x = \frac{1+2\alpha}{n+3}, \quad y = z = \frac{1-\alpha}{n+3}.$$

Thus the best case is that X links to no page, both Y and Z only link to X.

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

$$\widetilde{G} = \alpha \begin{pmatrix} P & \mathbf{0} \\ W & \widetilde{Q} \end{pmatrix} + \frac{1-\alpha}{n+3}\mathbf{1}_{(n+3)\times(n+3)}.$$

Since the matrix W takes some row-sum from $\widetilde{Q}$, we have

$$\widetilde{Q}\mathbf{1}_3 < \mathbf{1}_3.$$

Still we have

$$\begin{aligned} (\tilde{r}, x, y, z) &= (\tilde{r}, x, y, z)\widetilde{G} \\ &= \alpha(\tilde{r}P + (x,y,z)W, (x,y,z)\widetilde{Q}) + \frac{1-\alpha}{n+3}\mathbf{1}_{n+3}^T, \end{aligned}$$

$$\Rightarrow \quad (x,y,z) = \alpha(x,y,z)\widetilde{Q} + \frac{1-\alpha}{n+3}(1,1,1).$$

$$\begin{aligned} \Rightarrow \quad x+y+z &= (x,y,z)\mathbf{1}_3 \\ &= \alpha(x,y,z)\widetilde{Q}\mathbf{1}_3 + \frac{1-\alpha}{n+3}\mathbf{1}_3^T\mathbf{1}_3 \\ &< \alpha(x,y,z)\mathbf{1}_3 + \frac{3-3\alpha}{n+3}, \end{aligned}$$

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

Also we have

$$\begin{aligned} y &= \alpha(Q_{12}x + Q_{22}y + Q_{32}z) + \frac{1-\alpha}{n+3}, \\ z &= \alpha(Q_{13}x + Q_{23}y + Q_{33}z) + \frac{1-\alpha}{n+3}, \\ \Rightarrow \qquad y &= \frac{1}{1-\alpha Q_{22}}\left[\alpha(Q_{12}x + Q_{32}z) + \frac{1-\alpha}{n+3}\right] \geq \frac{1-\alpha}{n+3}, \\ z &= \frac{1}{1-\alpha Q_{33}}\left[\alpha(Q_{13}x + Q_{23}y) + \frac{1-\alpha}{n+3}\right] \geq \frac{1-\alpha}{n+3}. \end{aligned}$$

Therefore

$$x < \frac{3}{n+3} - y - z \leq \frac{3}{n+3} - 2\frac{1-\alpha}{n+3} = \frac{1+2\alpha}{n+3},$$

which means adding links from X, Y, Z to the n older pages can not improve the PageRank of X but do the inverse thing.

(e) Assume that we add m pages, $X_j, j = 1, 2, \cdots, m$, to the web graph. Let $x = (x_1, x_2, \cdots, x_m) \in \mathbb{R}^{1\times m}$ denote the PageRanks of our new m Pages. Our purpose is to maximize x_1, the PageRank of X_1. According to the result of (d), we should not add links from the our new pages to the old pages. Thus the transition matrix of the $n+3$ nodes should be

$$\widetilde{G} = \alpha \begin{pmatrix} P & \mathbf{0} \\ \mathbf{0} & Q \end{pmatrix} + \frac{1-\alpha}{n+m}\mathbf{1}_{(n+m)\times(n+m)},$$

where $Q \in \mathbb{R}^{m\times m}$ is the transition matrix for our new m pages. Again we have

$$\begin{aligned}
(\tilde{r}, x) &= (\tilde{r}, x)\widetilde{G} \\
&= \alpha(\tilde{r}P, xQ) + \frac{1-\alpha}{n+m}\mathbf{1}^T_{n+m}, \\
\Rightarrow \quad x &= \alpha x Q + \frac{1-\alpha}{n+m}\mathbf{1}^T_m,
\end{aligned}$$

$$\Rightarrow \quad \sum_{j=1}^m x_j = x\mathbf{1}_m = \alpha x Q \mathbf{1}_m + \frac{1-\alpha}{n+m}\mathbf{1}^T_m\mathbf{1}_m = \alpha x\mathbf{1}_m + \frac{(1-\alpha)m}{n+m},$$

$$\Rightarrow \quad \sum_{j=1}^m x_j = \frac{m}{n+m}.$$

Using the same argument in (c) we have

$$x_j = \alpha\sum_{i=1}^m Q_{ij}x_i + \frac{1-\alpha}{n+m},$$

$$\Rightarrow \quad x_j = \frac{1}{1-\alpha Q_{jj}}\Big[\alpha\sum_{i\neq j} Q_{ij}x_i + \frac{1-\alpha}{n+m}\Big] \geq \frac{1-\alpha}{n+m}.$$

Therefore,

$$x_1 = \frac{m}{n+m} - \sum_{j=2}^m x_j \leq \frac{m}{n+m} - \frac{(1-\alpha)(m-1)}{n+m} = \frac{1+(m-1)\alpha}{n+m}.$$

Especially, if we choose

$$Q = (\mathbf{1}_{m\times 1}, \mathbf{0}_{m\times(m-1)}),$$

we have

$$x_1 = \frac{1+(m-1)\alpha}{n+m}, \quad x_j = \frac{1-\alpha}{n+m}, \; j = 2, 3, \cdots, m.$$

Therefore, to maximize x_1, the new m pages should have the structure that X_1 links to no page, and all the other $m-1$ new pages only link to X_1.

3 Beyong PageRank

(a) First we will show that for a connected, undirected graph G(V,E), the **Degree centrality** and the **PageRank**(obtained with $\alpha = 1$) rank the nodes in the same order of importance. That is, we will show that

$$\frac{r_i}{d_i} = \frac{1}{2m}, \quad i = 1, 2, \cdots, n,$$

where m denotes the number of edges in the graph. Now define

$$A \in \mathbb{R}^{n\times n} : A_{i,j} = \begin{cases} 1, & (i,j) \in E; \\ 0, & (i,j) \notin E, \end{cases}$$

$$P \in \mathbb{R}^{n\times n} : P_{i,j} = \begin{cases} 1/d_i, & (i,j) \in E; \\ 0, & (i,j) \notin E, \end{cases}$$

$$D \in \mathbb{R}^{n\times n} : D_{i,j} = \begin{cases} 1/d_i, & i = j; \\ 0, & i \neq j, \end{cases}$$

Here, A is the adjacency matrix for G, P is the transition matrix for PageRank with $\alpha = 1$, D is a diagonal matrix. Since the graph G is connected and undirected, it's easy to prove that P and A have the following properties:

1. $A = A^T$.

2. $P = DA$, $P^T = AD$.

3. P is irreducible, 1 is a simple eigenvalue of P associated with a unique left eigenvector(up to a constant) and a unique right eigenvector(up to a constant).(Perron-Frobenius Theorem)

4. Any right eigenvector of P associated with eigenvalue 1 is a multiple of $\mathbf{1}_n = (1, 1, \cdots, 1)^T \in \mathbb{R}^{n\times 1}$.

Let $r = (r_1, r_2, \cdots, r_n) \in \mathbb{R}^{1\times n}$ be the PageRanks obtained with P, that is

$$r = rP.$$

Define $x = (r_1/d_1, r_2/d_2, \cdots, r_n/d_n)^T \in \mathbb{R}^{n\times 1}$, then

$$x^T = rD = rPD = rDAD = rDP^T = x^T P^T,$$

$$\Rightarrow \quad x = Px.$$

That is to say, x is a right eigenvector of P associated with eigenvalue 1, which means x is a multiple of $\mathbf{1}_n = (1, 1, \cdots, 1)^T$. Therefore

$$\frac{r_1}{d_1} = \frac{r_2}{d_2} = \cdots = \frac{r_n}{d_n} = C,$$

where C is a constant to be determined. Further, we know that

$$1 = \sum_{i=1}^{n} r_i = C \sum_{i=1}^{n} d_i = 2mC,$$

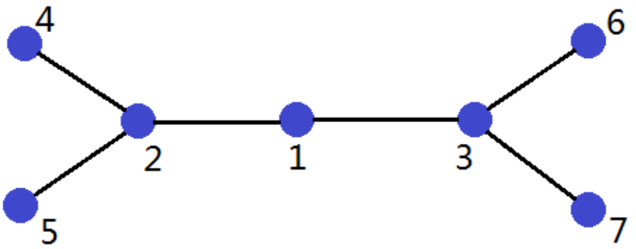
where m denotes the number of edges in G. Therefore

$$\frac{r_i}{d_i} = C = \frac{1}{2m}, \quad i = 1, 2, \cdots, n,$$

the PageRank r_i is proportional to the degree centrality $C_D(i)$ in a connected, undirected graph.

Next, we only need to contrast **Degree centrality**, **Closeness centrality** and **Betweenness centrality**.

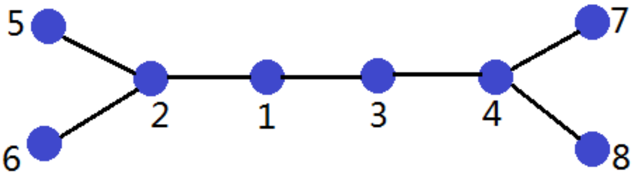
- **Degree centrality v.s. Closeness centrality**
  Consider a graph with 7 nodes below



$$C_D(1) = \frac{2}{6} = \frac{1}{3}, \qquad C_C(1) = \frac{6}{1+2+2+1+2+2} = \frac{3}{5},$$
$$C_D(2) = \frac{3}{6} = \frac{1}{2}, \qquad C_C(2) = \frac{6}{1+1+1+2+3+3} = \frac{6}{11},$$

we have $C_D(1) < C_D(2)$ but $C_C(1) > C_C(2)$.

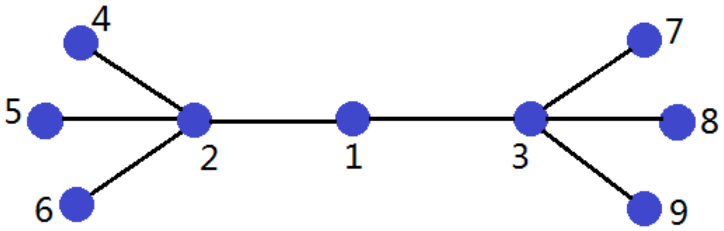
- **Degree centrality v.s. Betweenness centrality**
  Consider a graph with 8 nodes below



$$C_D(1) = \frac{2}{7}, \qquad C_B(1) = \frac{3\times 4}{21} = \frac{12}{21},$$
$$C_D(2) = \frac{3}{7}, \qquad C_B(2) = \frac{1+5+5}{21} = \frac{1}{21},$$

we have $C_D(1) < C_D(2)$ but $C_B(1) > C_B(2)$.

- **Closeness centrality v.s. Betweenness centrality**
  Consider a graph with 9 nodes below



$$C_C(1) = \frac{8}{2+6\times 2} = \frac{4}{7}, \qquad C_B(1) = \frac{4\times 4}{21} = \frac{16}{21},$$
$$C_C(2) = \frac{8}{1+2+3\times 1+3\times 3} = \frac{8}{15}, \qquad C_B(2) = \frac{3\times 5+3}{21} = \frac{6}{7},$$

we have $C_C(1) > C_C(2)$ but $C_B(1) < C_B(2)$.

(d)

1. **Degree centrality: social network**
   Let G be a social network. Each node denotes a person, and each edge connecting two nodes means these two people are friends.
   A person with a greater degree has more friends, and generally is more popular and has more influence on the whole group. Thus the degree centrality represents the importance of a person in a social network.

2. **Closeness centrality: bus station network**
   Let G be a bus station network of a city. Each node denotes a bus station, and each edge denotes a bus route between two stations.
   If a bus station has, on average, shorter paths to the other stations, it would be more convenient to live near this station, the area round this station would also be more prosperous. Therefore, the closeness centrality represents the importance of a bus station in the transportation network.

3. **Betweenness centrality: subway network**
   Let G be a subway network of a city. Each node denotes a subway station, and each edge denotes a direct subway line between two stations.
   Assume that between station A and station B there are several shortest paths, some of these paths pass through a station C. If C is out of service, then the passengers traveling from A to B will be forced to take the other alternative shortest paths or even longer paths, which will increase the transportation pressure of the other paths. If there are a lot of shortest paths passing through station C, its role of sharing transportation pressure will be seriously important. Therefore, the betweenness centrality describes the ability of a station in increasing the transport capacity of the subway network.

4. **PageRank: web pages network** Let G ba a network of web pages. Each node denotes a web page, and each directed edge denotes a link from one page to another. By definition, a web page with high PageRank is likely to have more in-links or be linked to by some other important pages, and as a result it may be visited more frequently. Therefore, the PageRank naturally represents the importance of a web page.

4 MapReduce: Warming up for rankmaniac

(a) **Common friends:**

```
function main() {
    listOfFriendship = list of (Person, [list of Friends])

    Create a list called listOfFriends
    for each (keyA, valueA) pair in listOfFriendship {
        for each keyB in valueA {
            if keyB is lexicographically bigger than keyA {
                Find the pair (keyB, valueB) in listOfFriendship
                key = (keyA, keyB)
                value = (valueA, valueB)
                Append (key, value) to listOfFriends
            }
        }
    }
   # Each element of listOfFriends is a pair like ((personA, personB), ([list of
   # friends of A], [list of friends of B])).

    listOfOutputs = MapReduce(listOfFriends)

    return listOfOutputs
}

function Map(key, value) {
    # Here the input is the pair ((keyA, keyB), (valueA, valueB)).
    keyA = the first item of key
    keyB = the second item of key
    valueA = the first item of value
    valueB = the second item of value
    for each keyC in valueA {
        if keyC is in valueB {
            EmitIntermediate((keyA,keyB), keyC)
        }
    }
}

function Reduce(key, listOfValues) {
    # It seems that we have nothing more to do in this problem. Maybe we can sort
    # the common friends in some order of importance.
    return (key, sort(listOfValues))
}
```

(b) High school days:

```
function main() {
    listOfFiles = list of pairs (Filename, [list of Scores])

    listOfOutputs = MapReduce(listOfFiles)

    for score = 100 to 1 {
        Find (score, value) in listOfOutputs
        print (score, value)
    }
}

function Map(key, value) {
    # The input is a pair (Filename, [list of Scores]).
    for each score in value {
        EmitIntermediate(score, 1)
    }
}

function Reduce(score, listOfValues) {
    return (score, length(listOfValues))
}
```

(c) Good old pie

We can calculate the proportion p of random points that fall inside the unit circle. Then π = 4p/n, where n is the total number of random points that we create.

```
function main() {
    n = a given large number
    Create a list called listOfRands
    for i = 1 to n {
        x = rand()
        y = rand()
        key = i
        value = [x,y]
        append a pair (key, value) to listOfRands
    }

    (key, value) = MapReduce(listOfRands)

    return pi=4*value/n
}
```

```
function Map(key, value) {
    x = value[1]
    y = value[2]
    if xˆ2+yˆ2 <= 1 {
        EmitIntermediate( ‘hit’ , 1)
        # Count the number of random points inside the unit circle.
    }
}

function Reduce(key, listOfValues) {
    return (key, length(listOfValues))
}
```

(d) Gauge the distance

```
function main() {
    G[] = Adjacency list of the graph
    n = length(G)
    dist[] = (0, 0, 0, ..., 0)
    distUpdated[] = (0, ∞, ∞, ..., ∞)

    while( NOT stoppingCriterion(dist, distUpdated)) {
        dist = distUpdated
        Create a list listTempG
        for i = 1 to n {
            key = i
            value = (dist[i], G[i])
            append a pair (key, value) to listTempG
        }
        distUpdated = MapReduce(listTempG).
    }
    print the list dist
}

function stoppingCriterion(dist1, dist2){
    n = length(dist1)
    for i = 1 to n {
        if dist1[i] != dist2[i] {
            return 0
        }
    }
    return 1
}
```

```
function Map(key, value){
    # Here the input key is an index, and the value is a tuple (dist[index], G[index])
    distSoFar = the first item of value # which is a distance
    listNeighbor = the second item of value # which is a list
    for each tuple (neighbor, distance) in listNeighbor {
        EmitIntermediate(neighbor, distSoFar+distance)
        # Create a possible shorter distance of this neighbor.
        # Here the neighbor is also a node index.
    }
}

function Reduce(key, listOfValues){
    return (key, min(listOfValues))
    # Get the shortest distance between key and node 1 found so far.
}
```