Page Rank

(Section 3.3 - kind of)

Google's Original Search Algorithm

In a paper by Brin & Page (the founders of Google) from 1998. Since been improved from its original framework.

Step 1: Make an $n \times n$ matrix A where $n =$ the number of webpages (url's) in existence (~ 1.5 billion),

$$A = (A_{i,k})_{i,k=1}^{n} \text{ where}$$

$$\boxed{A_{i,k} = \begin{cases} 1, & \text{if page } k \text{ links to page } i \\ 0, & \text{otherwise} \end{cases}}$$

By convention, a page doesn't link to itself.

Example 1: (3 pages) Given pages

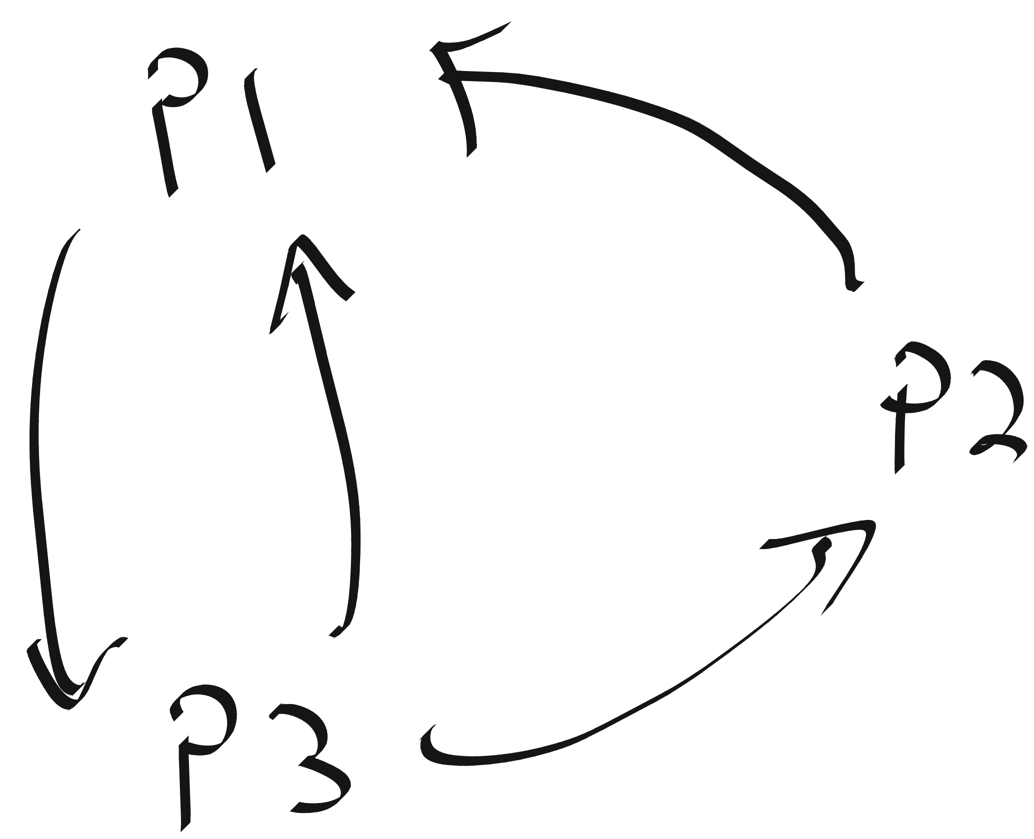
P1, P2, and P3, suppose

P1 links to P3

P2 links to P1

P3 links to P1 and P2

Picture:



Then

$$A = \begin{array}{c} \begin{array}{ccc} P1 & P2 & P3 \end{array} \\ \begin{bmatrix} 0 & 1 & 1 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix} \end{array}$$

Step 2: Add the entries of the k^{th} column of A for all $1 \le k \le n$ and divide the $A_{i,k}$ entries of A by this number. Call this new matrix B.

Back to Example 1:

$$A = \begin{bmatrix} 0 & 1 & 1 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix}$$

Sum of entries in column $1 = 1$

Sum of entries in column $2 = 1$

Sum of entries in column $3 = 2$

Can leave entries in columns 1 and 2 alone, divide the entries in column 3 by 2 to get B:

$$\boxed{B = \begin{bmatrix} 0 & 1 & 1/2 \\ 0 & 0 & 1/2 \\ 1 & 0 & 0 \end{bmatrix}}$$

Note: the entries in every column of B add up to one.

Problem: What if all the entries in one column are zero? Then we're dividing by zero...

Solution: Change all the entries in that column to ones. When constructing B, the entries will now be $1/n$ where $n =$ number of webpages.

As $n \to \infty$, $1/n \to 0$, so this is (somewhat) representative of the original matrix if n is large.

Step 3: Build in the possibility that people will randomly start their search over with probability $1-d$ (usually $d = .85$) and let C be the $n \times n$ matrix with entries

$$C_{i,k} = d\,B_{i,k} + \frac{(1-d)}{n}$$

for all $1 \le i,k \le n$.

Note: every column of C still sums to one.

Back to Example 1

Use fractions: $\alpha = .85 = \frac{85}{100} = \frac{17}{20}$

$1-\alpha = .15 = 3/20$

Why? Sometimes Wolfram Alpha will make mistakes with decimals.

$$C_{i,k} = \frac{17}{20} B_{i,k} + \frac{\frac{3}{20}}{3}$$

$$C_{i,k} = \frac{17}{20} B_{i,k} + \frac{1}{20}$$

$1 \leq i,k \leq 3$.

Can write this using matrices:

$$C = \frac{17}{20} B + \frac{1}{20}\begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}$$

Recall $B = \begin{bmatrix} 0 & 1 & 1/2 \\ 0 & 0 & 1/2 \\ 1 & 0 & 0 \end{bmatrix}$, so

$$C = \frac{1}{20}\left(17B + \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}\right)$$

$$= \frac{1}{20}\left(\begin{bmatrix} 0 & 17 & 17/2 \\ 0 & 0 & 17/2 \\ 17 & 0 & 0 \end{bmatrix} + \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}\right)$$

$$= \frac{1}{20}\begin{bmatrix} 1 & 18 & 19/2 \\ 1 & 1 & 19/2 \\ 18 & 1 & 1 \end{bmatrix}$$ (can stop here)

$$= \begin{bmatrix} 1/20 & 18/20 & 19/40 \\ 1/20 & 1/20 & 19/40 \\ 18/20 & 1/20 & 1/20 \end{bmatrix}$$

Observe that every column still adds up to one!

Step 4: C will always have $\lambda = 1$ as its largest eigenvalue.

Take the corresponding eigenvector $\vec{v}$ and divide every entry of $\vec{v}$ by

$$\|\vec{v}\|_1 = \sum_{i=1}^{n} |v_i| .$$

The i^{th} entry of this vector is the PageRank of page P_i for all $1 \le i \le n$.

Back to Example 1:

$$C = \frac{1}{20}\begin{bmatrix} 1 & 18 & 19/2 \\ 1 & 1 & 19/2 \\ 18 & 1 & 1 \end{bmatrix}$$

Eigenvalue $\lambda = 1$ has

associated eigenvector

$$\vec{v} = \begin{bmatrix} 703/686 \\ 190/343 \\ 1 \end{bmatrix}$$ (entries are all positive)

$$\|\vec{v}\|_1 = \frac{703}{686} + \frac{190}{343} + 1$$

$$\|\vec{v}\|_1 = \frac{1769}{686}$$

Divide $\vec{v}$ by this number.

$$\frac{\vec{v}}{\|\vec{v}\|_1} = \frac{686}{1769} \begin{bmatrix} 703/686 \\ 190/343 \\ 1 \end{bmatrix}$$

$$= \begin{bmatrix} 703/1769 \\ 380/1769 \\ 686/1769 \end{bmatrix}$$

Observe that the entries of $\frac{\vec{v}}{\|\vec{v}\|_1}$

sum to one:

$$\frac{703}{1769} + \frac{380}{1769} + \frac{686}{1769} = \frac{1769}{1769} = 1$$

Each entry represents the probability of landing at the associated page, called the PageRank.

$$\frac{\vec{v}}{\|\vec{v}\|_1} = \begin{bmatrix} 703/1769 \\ 380/1769 \\ 686/1769 \end{bmatrix} \begin{array}{l} \leftarrow \text{PageRank of P1} \\ \leftarrow \text{PageRank of P2} \\ \leftarrow \text{PageRank of P3} \end{array}$$

Example 2: (four pages) Find the PageRank of P_2 if

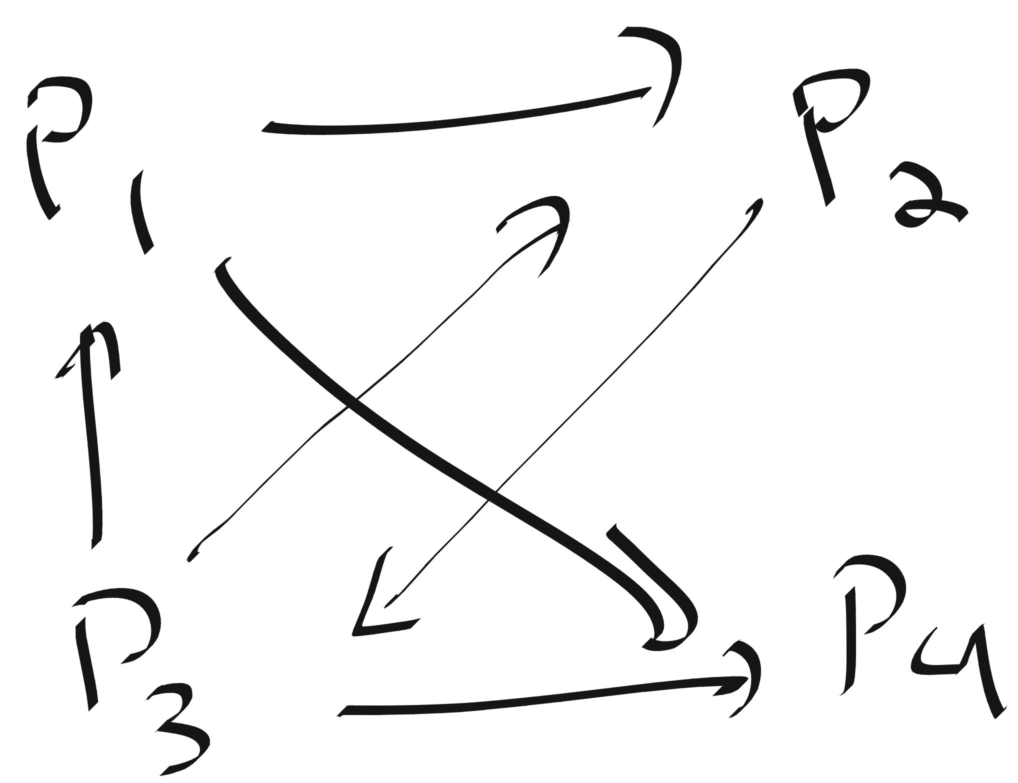
P_1 links to P_2 and P_4

P_2 links to P_3

P_3 links to $P_1, P_2,$ and P_4

P_4 links to nothing.

Solution: Picture



Step 1: Make A.

$$A = \begin{array}{c} \begin{matrix} P_1 & P_2 & P_3 & P_4 \end{matrix} \\ \begin{bmatrix} 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \end{bmatrix} \end{array}$$

Step 2: Make B. Note column 4 is all zeros, so change the zeros to ones in this column:

$$A \rightarrow \begin{bmatrix} 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 \end{bmatrix}$$

$$B = \begin{bmatrix} 0 & 0 & 1/3 & 1/4 \\ 1/2 & 0 & 1/3 & 1/4 \\ 0 & 1 & 0 & 1/4 \\ 1/2 & 0 & 1/3 & 1/4 \end{bmatrix}$$

Note every column of B adds up to one.

Step 3: Make C.

$$d = .85 = \frac{17}{20}$$

$$C_{i,k} = \frac{17}{20} B_{i,k} + \frac{\frac{3}{20}}{4}$$

$$C_{i,k} = \frac{17}{20} B_{i,k} + \frac{3}{80}$$

$$1 \le i,k \le 4.$$

$$C = \frac{17}{20} B + \frac{3}{80} \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix}$$

$$= \frac{1}{20} \left(17B + 3/4 \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix} \right)$$

$$= \frac{1}{20} \left(\begin{bmatrix} 0 & 0 & 17/3 & 17/4 \\ 17/2 & 0 & 17/3 & 17/4 \\ 0 & 17 & 0 & 17/4 \\ 17/2 & 0 & 17/3 & 17/4 \end{bmatrix} + \right.$$

$$\left. 3/4 \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix} \right)$$

$$= \frac{1}{20} \begin{bmatrix} 3/4 & 3/4 & 77/12 & 5 \\ 37/4 & 3/4 & 77/12 & 5 \\ 3/4 & 71/4 & 3/4 & 5 \\ 37/4 & 3/4 & 77/12 & 5 \end{bmatrix}$$

Observe each column of C adds up to one.

Step 4: Compute PageRanks by finding the eigenvector associated to $\lambda = 1$ for C.

We get the eigenvector

$$\vec{v} = \begin{bmatrix} 40/57 \\ 1 \\ 1769/1463 \\ 1 \end{bmatrix}$$

$$\|\vec{v}\|_1 = 40/57 + 1 + \frac{1769}{1463} + 1$$

$$\|\vec{v}\|_1 = \frac{17{,}165}{4389}$$

$$\frac{\vec{v}}{\|\vec{v}\|_1} = \frac{4389}{17{,}165}\begin{bmatrix} 40/57 \\ 1 \\ 1769/1463 \\ 1 \end{bmatrix}$$

Only asked for the PageRank of P2, which is $\boxed{\dfrac{4389}{17{,}165}}$

For large values of n, how Page and Brin find this eigenvector is Google's secret!

Some questions:

1) How do we know $\lambda = 1$ is always an eigenvalue for C?

2) How do we know $\lambda = 1$ is the largest eigenvalue?

3) The associated eigenvector $\vec{v}$ had better have positive entries. How do we know this is the case?

Answer for 1): eigenvalues of C = eigenvalues of C^t, and

$$\vec{w} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}$$ is

an associated eigenvector for $\lambda = 1$ and C^t.

If we look at (for any $n \times n$ matrix A)

$$\det(A^t - \lambda I_n)$$

$$= \det(A^t - \lambda I_n^t)$$

$$= \det((A - \lambda I_n)^t)$$

$$= \det(A - \lambda I_n) \quad \text{(determinant property 4)}$$

So $0 = \det(A^t - \lambda I_n) = \det(A - \lambda I_n)$,

which tells us that the eigenvalues of A and A^t are the same.

Now consider C^t. Remember that

$$(C^t)_{i,k} = C_{k,i} \text{ for all } 1 \le i,k \le n$$

and that $\displaystyle\sum_{k=1}^{n} c_{k,i} = 1$ for all fixed i, $1 \le i \le n$.

Then

$$C^t \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}$$

$$= \begin{bmatrix} \sum_{k=1}^{n} (C^t)_{1,k} \\ \sum_{k=1}^{n} (C^t)_{2,k} \\ \vdots \\ \sum_{k=1}^{n} (C^t)_{n,k} \end{bmatrix}$$

$$= \begin{bmatrix} \sum_{k=1}^{n} C_{k,1} \\ \sum_{k=1}^{n} C_{k,2} \\ \vdots \\ \sum_{k=1}^{n} C_{k,n} \end{bmatrix} \begin{matrix} = 1 \\ = 1 \\ \vdots \\ = 1 \end{matrix}$$

(all columns of C sum to 1)

$$= \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}$$

$$= 1 \cdot \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}$$

This shows $\lambda = 1$ is an eigenvalue for C^t, so $\lambda = 1$ is an eigenvalue for C but not with the vector of all ones!

Answer for 2): norm inequality, takes a little bit of work

Answer for 3): Perron-Frobenius Theorem!