12 Google PageRank, Communication Classes

The model that we created for Google PageRank in the last lecture does not exactly work. Some websites may not have any links, other websites may have only one link to itself, "trapping" us in that website.

This means that the transition matrix would not be regular, and so the stable vector might not be unique.

We adjust the Markov chain as follows (with n states):

1. If entry (i, i) is 1, we edit the column with every entry as $\frac{1}{n}$. Call this new matrix $\overline{P}$.
2. For all other entries, we first scale the entry by ρ, a **damping factor** ("normalizing occurrences of other websites") and add an additional $\frac{1-\rho}{n}$ to each value.

   Note: for problems in class, we use $\rho = 0.85$.

Definition 12.1
The **Google Matrix** is

$$G = \rho\overline{P} + (1-\rho)K$$

where $\overline{P}$ is from (1), and K is the $n \times n$ matrix

$$K = \begin{bmatrix} 1/n & 1/n & \cdots & 1/n \\ \vdots & \vdots & \ddots & \\ 1/n & 1/n & \cdots & 1/n \end{bmatrix}$$

What is the behavior of G?

1. States with 100% probability of returning to itself:

$$P = \begin{bmatrix} 0 & 0 & \cdots \\ 0 & 1 & \cdots \\ 0 & 0 & \cdots \end{bmatrix} \implies \overline{P} = \begin{bmatrix} 0 & 1/n & \cdots \\ 0 & 1/n & \\ 0 & \vdots & \end{bmatrix}$$

$$\implies \rho\overline{P} + (1-\rho)K = \begin{bmatrix} \rho/n \\ \rho/n \\ \vdots \end{bmatrix} + \begin{bmatrix} (1-\rho)/n \\ (1-\rho)/n \\ \vdots \end{bmatrix} = \begin{bmatrix} 1/n \\ 1/n \\ \vdots \end{bmatrix}$$

2. Other states:

$$P = \begin{bmatrix} & 1/3 & \\ & 0 & \\ \cdots & 1/2 & \cdots \\ & 1/6 & \\ & 0 & \end{bmatrix}$$

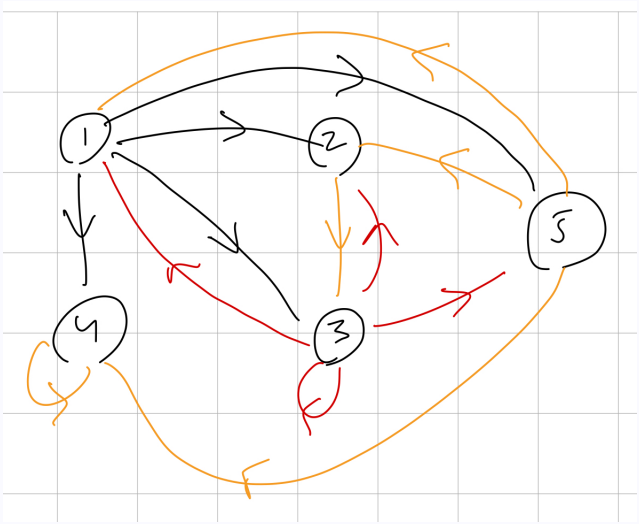
$$\implies G = \rho\overline{P} + (1-\rho)K = \begin{bmatrix} & \rho/3 & \\ & 0 & \\ \cdots & \rho/2 & \cdots \\ & \rho/6 & \\ & 0 & \end{bmatrix} + \begin{bmatrix} & (1-\rho)/n & \\ & (1-\rho)/n & \\ \cdots & (1-\rho)/n & \cdots \\ & (1-\rho)/n & \\ & (1-\rho)/n & \end{bmatrix} = \begin{bmatrix} & \rho/3 + (1-\rho)/n & \\ & (1-\rho)/n & \\ \cdots & \rho/2 + (1-\rho)/n & \cdots \\ & \vdots & \end{bmatrix}$$

Note that now we have made all entries of the transition matrix stricly positive, which ensures that this matrix is regular, meaning that we will have a unique stable vector.

We can go between any two states in any number of transitions, which also means this Markov chain is regular.

Thus we can use the unique stable vector to rank the pages by popularity.

Example 12.2
Suppose we have the following graph (note that if the probabilities are not stated, we assume equal chance):



$$P = \begin{bmatrix} 0 & 0 & 1/4 & 0 & 1/3 \\ 1/4 & 0 & 1/4 & 0 & 1/3 \\ 1/4 & 1 & 1/4 & 0 & 0 \\ 1/4 & 0 & 0 & 1 & 1/3 \\ 1/4 & 0 & 1/4 & 0 & 0 \end{bmatrix} \implies \overline{P} = \begin{bmatrix} 0 & 0 & 1/4 & 1/5 & 1/3 \\ 1/4 & 0 & 1/4 & 1/5 & 1/3 \\ 1/4 & 1 & 1/4 & 1/5 & 0 \\ 1/4 & 0 & 0 & 1/5 & 1/3 \\ 1/4 & 0 & 1/4 & 1/5 & 0 \end{bmatrix}$$

Note that here we do not modify the second column of P because it does not represent a self-loop.

$$\implies G = 0.85\overline{P} + 0.15K \qquad K = \begin{bmatrix} 1/5 & 1/5 & \cdots \\ \vdots & \ddots & \end{bmatrix}$$

Then, solving

$$(G - I)\vec{q} = \vec{0} \qquad q_1 + q_2 + q_3 + q_4 + q_5 = 1$$

We find

$$\vec{q} = \begin{bmatrix} 0.1688 \\ 0.2046 \\ 0.3333 \\ 0.1338 \\ 0.1595 \end{bmatrix}$$

And from the stable vector, we get that the web pages are ranked 3, 2, 1, 5, 4.

We know that with regards to the long term behavior, the start state does not matter. What about problems where the start state matters?

12.1 Communication Classes (Section 3.3)

Definition 12.3
State i and state j are said to **communicate** if there exists non-negative integers m and n such that entry (i, j) of P^n and entry (j, i) of P^m is non-zero.

In other words, two states i and j communicate if we are able to go from i to j and j to i in a certain number of steps.