Google Page Rank

Theorem

If P is a regular $m \times m$ stochastic matrix with $m \geq 2$, then:

- for any initial probability vector $\vec{x}_0$, $\lim_{n\to\infty} P^n \vec{x}_0 = \vec{q}$
- P has a unique eigenvector, $\vec{q}$, which has eigenvalue $\lambda = 1$
- there is a stochastic matrix Π such that $\lim_{n\to\infty} P^n = \Pi$
- each column of Π is the same probability vector $\vec{q}$
- the eigenvalues of P satisfy $|\lambda| \leq 1$

The PageRank algorithm is based on a mathematical model that assumes that we have:

- a collection of web pages that have links to each other
- users who are navigating the web
- a set of rules that govern how the users navigate the web

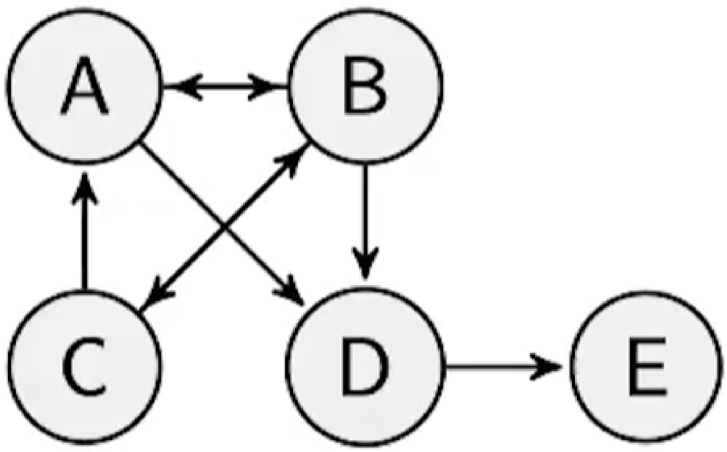
We impose assumptions about how the users navigate the web:

a) A user on a web page is equally likely to go to any page that their page links to.

b) If a user is on a page that does not link to other pages, the user stays at their page.

c) The distribution of users can be modeled using a Markov process, $\vec{x}_{k+1} = P\vec{x}_k$, where
   - $\vec{x}_k \in \mathbb{R}^n$ is a probability vector, gives the proportion of users on each page at iteration k
   - P is an $n \times n$ stochastic matrix
   - n is the number of pages in the web

If we have:



P will be:

$$P = \begin{array}{c} \begin{array}{ccccc} A & B & C & D & E \end{array} \\ \begin{pmatrix} 0 & 1/3 & 1/2 & 0 & 0 \\ 1/2 & 0 & 1/2 & 0 & 0 \\ 0 & 1/3 & 0 & 0 & 0 \\ 1/2 & 1/3 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 \end{pmatrix} \end{array}$$

Problem

- We do not have a unique unique steady-state for all pages
- pages that do not link to other pages can have the largest importance, or highest PageRank

Fix

- If a user reaches a page that does not link to other pages, the user will choose any page in the web, with equal probability, and move to that page.

- Corresponding fixed matrix:

$$P_* = \begin{pmatrix} 0 & 0 & 1 & 0 & .2 \\ .5 & 0 & 0 & 0 & .2 \\ 0 & 1 & 0 & 0 & .2 \\ .5 & 0 & 0 & 0 & .2 \\ 0 & 0 & 0 & 1 & .2 \end{pmatrix}$$

Adjustment 2

A user at any page will navigate to any page among those that their page links to with equal probability p, and to any page in the web with equal probability $1-p$. The transition matrix becomes

$$G = pP_* + (1-p)K.$$

All the elements of the $n \times n$ matrix K are equal to $1/n$.

Why this works

This works because we are taking each column and reducing its sum by $(1-p)$. Then we are adding $(1-p)$ back to the column, but evenly distributed over the whole column. This insures that the sum of the column of the final matrix stays $= 1$.