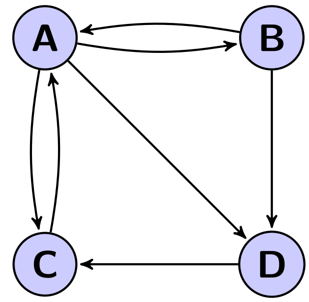
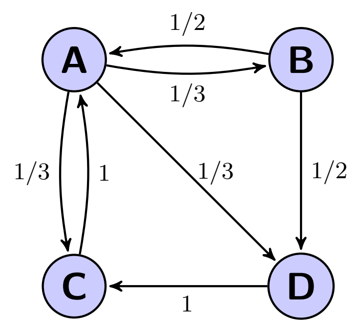
Example 132. Suppose the internet consists of only the four webpages A, B, C, D which link to each other as indicated in the diagram. Rank these webpages by computing their PageRank vector.



Solution. Recall that we model a random surfer, who randomly clicks on links. Let a_t be the probability that such a surfer will be on page A at time t. Likewise, b_t, c_t, d_t are the probabilities that the surfer will be on page B, C or D.

The transition probabilities are indicated in the diagram to the right. As in the previous example, we obtain the following transition behaviour:



$$\begin{bmatrix} a_{t+1} \\ b_{t+1} \\ c_{t+1} \\ d_{t+1} \end{bmatrix} = \begin{bmatrix} 0 \cdot a_t + \frac{1}{2} \cdot b_t + 1 \cdot c_t + 0 \cdot d_t \\ \frac{1}{3} \cdot a_t + 0 \cdot b_t + 0 \cdot c_t + 0 \cdot d_t \\ \frac{1}{3} \cdot a_t + 0 \cdot b_t + 0 \cdot c_t + 1 \cdot d_t \\ \frac{1}{3} \cdot a_t + \frac{1}{2} \cdot b_t + 0 \cdot c_t + 0 \cdot d_t \end{bmatrix} = \underbrace{\begin{bmatrix} 0 & \frac{1}{2} & 1 & 0 \\ \frac{1}{3} & 0 & 0 & 0 \\ \frac{1}{3} & 0 & 0 & 1 \\ \frac{1}{3} & \frac{1}{2} & 0 & 0 \end{bmatrix}}_{=T} \begin{bmatrix} a_t \\ b_t \\ c_t \\ d_t \end{bmatrix}$$

To find the equilibrium state, we determine an appropriate 1-eigenvector of the transition matrix T.

The 1-eigenspace is $\operatorname{null}(T - 1 \cdot I) = \operatorname{null}\left(\begin{bmatrix} -1 & \frac{1}{2} & 1 & 0 \\ \frac{1}{3} & -1 & 0 & 0 \\ \frac{1}{3} & 0 & -1 & 1 \\ \frac{1}{3} & \frac{1}{2} & 0 & -1 \end{bmatrix} \right)$

To compute a basis, we perform Gaussian elimination: $\begin{bmatrix} -1 & \frac{1}{2} & 1 & 0 \\ \frac{1}{3} & -1 & 0 & 0 \\ \frac{1}{3} & 0 & -1 & 1 \\ \frac{1}{3} & \frac{1}{2} & 0 & -1 \end{bmatrix} \overset{\text{RREF}}{\rightsquigarrow} \begin{bmatrix} 1 & 0 & 0 & -2 \\ 0 & 1 & 0 & -\frac{2}{3} \\ 0 & 0 & 1 & -\frac{5}{3} \\ 0 & 0 & 0 & 0 \end{bmatrix}$

We conclude that the 1-eigenspace has basis $\begin{bmatrix} 2 \\ \frac{2}{3} \\ \frac{5}{3} \\ 1 \end{bmatrix}$. (Note that its entries add up to $2 + \frac{2}{3} + \frac{5}{3} + 1 = \frac{16}{3}$.)

The corresponding equilibrium state is $\frac{3}{16} \begin{bmatrix} 2 \\ \frac{2}{3} \\ \frac{5}{3} \\ 1 \end{bmatrix} \approx \begin{bmatrix} 0.375 \\ 0.125 \\ 0.313 \\ 0.188 \end{bmatrix}$. This is the **PageRank vector**.

[For instance, after browsing randomly for a long time, there is (about) a 12.5% chance to be at page B.]
Correspondingly, we rank the pages as $A > C > D > B$.

The real internet.

- Google reports (2016) doing "trillions" of searches per year. [2 trillion means 63,000 searches per second.]
- Google's search index contains almost 50 billion pages (2016). [Estimated to exceed 100,000,000 gigabytes.]
- More than 1,000,000,000 websites (i.e. hostnames; about 75% not active)

[The "average" user apparently only visits about 100 per month; wikipedia.org is one website, consisting of many webpages (more than 2,000,000).]

Armin Straub
straub@southalabama.edu

Gory details. There's nothing interesting about the Gaussian elimination above. Here are the full details:

$$\begin{bmatrix} -1 & \frac{1}{2} & 1 & 0 \\ \frac{1}{3} & -1 & 0 & 0 \\ \frac{1}{3} & 0 & -1 & 1 \\ \frac{1}{3} & \frac{1}{2} & 0 & -1 \end{bmatrix} \underset{\rightsquigarrow}{\substack{R_2+\frac{1}{3}R_1 \Rightarrow R_2 \\ R_3+\frac{1}{3}R_1 \Rightarrow R_3 \\ R_4+\frac{1}{3}R_1 \Rightarrow R_4}} \begin{bmatrix} -1 & \frac{1}{2} & 1 & 0 \\ 0 & -\frac{5}{6} & \frac{1}{3} & 0 \\ 0 & \frac{1}{6} & -\frac{2}{3} & 1 \\ 0 & \frac{2}{3} & \frac{1}{3} & -1 \end{bmatrix} \underset{\rightsquigarrow}{\substack{R_3+\frac{1}{5}R_2 \Rightarrow R_3 \\ R_4+\frac{4}{5}R_2 \Rightarrow R_4}} \begin{bmatrix} -1 & \frac{1}{2} & 1 & 0 \\ 0 & -\frac{5}{6} & \frac{1}{3} & 0 \\ 0 & 0 & -\frac{3}{5} & 1 \\ 0 & 0 & \frac{3}{5} & -1 \end{bmatrix}$$

$$\underset{\rightsquigarrow}{R_4+R_3 \Rightarrow R_4} \begin{bmatrix} -1 & \frac{1}{2} & 1 & 0 \\ 0 & -\frac{5}{6} & \frac{1}{3} & 0 \\ 0 & 0 & -\frac{3}{5} & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix} \underset{\rightsquigarrow}{\substack{-1R_1 \Rightarrow R_1 \\ -\frac{6}{5}R_2 \Rightarrow R_2 \\ -\frac{5}{3}R_3 \Rightarrow R_3}} \begin{bmatrix} 1 & -\frac{1}{2} & -1 & 0 \\ 0 & 1 & -\frac{2}{5} & 0 \\ 0 & 0 & 1 & -\frac{5}{3} \\ 0 & 0 & 0 & 0 \end{bmatrix} \underset{\rightsquigarrow}{\substack{R_1+R_3 \Rightarrow R_1 \\ R_2+\frac{2}{5}R_3 \Rightarrow R_2}} \begin{bmatrix} 1 & -\frac{1}{2} & 0 & -\frac{5}{3} \\ 0 & 1 & 0 & -\frac{2}{3} \\ 0 & 0 & 1 & -\frac{5}{3} \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

$$\underset{\rightsquigarrow}{R_1+\frac{1}{2}R_2 \Rightarrow R_1} \begin{bmatrix} 1 & 0 & 0 & -2 \\ 0 & 1 & 0 & -\frac{2}{3} \\ 0 & 0 & 1 & -\frac{5}{3} \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

Practical comment. The transition matrix we would get for the entire internet indexed by Google is prohibitingly large (a 50 billion by 50 billion matrix). While gigantic in size, it is a very **sparse matrix**, meaning that almost all of its entries are zero (each column has 50 billion entries but only a handful are nonzero, namely those corresponding to a link to another webpage). This is typical for many applications in matrix: we often deal with big but sparse matrices.

Another practical comment. In practical situations, the system might be too large for finding the equilibrium vector by elimination, as we did above. An alternative to elimination is the power method: it is based on the idea that the equilibrium vector is what we expect in the long-term. We can approximate this "long-term" behaviour by simulating a few transitions. For instance, in our example, if we start with the state $[\,1/4 \;\; 1/4 \;\; 1/4 \;\; 1/4\,]^T$, which corresponds to equal chances of being on each webpage, then the next state (that is, after one random click) is

$$T\begin{bmatrix} 1/4 \\ 1/4 \\ 1/4 \\ 1/4 \end{bmatrix} = \begin{bmatrix} 0 & \frac{1}{2} & 1 & 0 \\ \frac{1}{3} & 0 & 0 & 0 \\ \frac{1}{3} & 0 & 0 & 1 \\ \frac{1}{3} & \frac{1}{2} & 0 & 0 \end{bmatrix}\begin{bmatrix} 1/4 \\ 1/4 \\ 1/4 \\ 1/4 \end{bmatrix} = \begin{bmatrix} 3/8 \\ 1/12 \\ 1/3 \\ 5/24 \end{bmatrix} = \begin{bmatrix} 0.375 \\ 0.083 \\ 0.333 \\ 0.208 \end{bmatrix}.$$

Note that the ranking of the webpages is already A, C, D, B if we stop right here.

The state after that (that is, after two random clicks) is $T^2\begin{bmatrix} 1/4 \\ 1/4 \\ 1/4 \\ 1/4 \end{bmatrix} = \begin{bmatrix} 0.375 \\ 0.125 \\ 0.333 \\ 0.167 \end{bmatrix}$, and $T^3\begin{bmatrix} 1/4 \\ 1/4 \\ 1/4 \\ 1/4 \end{bmatrix} = \begin{bmatrix} 0.396 \\ 0.125 \\ 0.292 \\ 0.188 \end{bmatrix}$.

Observe how we are (overall) approaching the equilibrium vector $\begin{bmatrix} 0.375 \\ 0.125 \\ 0.313 \\ 0.188 \end{bmatrix}$.

Iterating like this is guaranteed to converge to a 1-eigenvector under mild technical assumptions on the transition matrix (for instance, that all its entries be positive; in that case, the other eigenvalues λ satisfy $|\lambda| < 1$ so that their contributions go to zero exponentially, as in our example from Lecture 21).