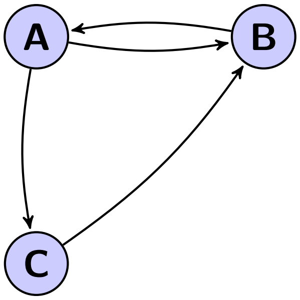
Sketch of Lecture 18

Fri, 2/24/2023

Application: PageRank

Example 99. Suppose the internet consists of only the three webpages A, B, C.



We wish to rank these webpages in order of "importance".

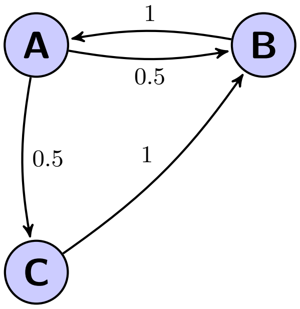
The idea. Instead of analyzing each webpage (which would be a lot of work!) we will try to only use the information how the pages are linked to each other. The idea being that an "important" page should be linked to from many other pages.

A and B have a link to each other. Also, A links to C and C links to B. If you keep randomly clicking from one webpage to the next, what proportion of the time will you be at each page?

The idea. We will assign ranking to the pages according to how frequently such a random surfer would visit these pages.

Comment. Before we start computing, stop for a moment, and think about how you would rank the webpages.

Solution. Let a_t be the probability that we will be on page A at time t. Likewise, b_t, c_t are the probabilities that we will be on page B or C.



The transition from one state to the next now works exactly as in the previous example. We get the following transition matrix:

$$\begin{bmatrix} a_{t+1} \\ b_{t+1} \\ c_{t+1} \end{bmatrix} = \begin{bmatrix} 0 \cdot a_t + 1 \cdot b_t + 0 \cdot c_t \\ \frac{1}{2} \cdot a_t + 0 \cdot b_t + 1 \cdot c_t \\ \frac{1}{2} \cdot a_t + 0 \cdot b_t + 0 \cdot c_t \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 \\ \frac{1}{2} & 0 & 1 \\ \frac{1}{2} & 0 & 0 \end{bmatrix} \begin{bmatrix} a_t \\ b_t \\ c_t \end{bmatrix}$$

To find the equilibrium state, we again determine an appropriate 1-eigenvector.

The 1-eigenspace is $\operatorname{null}\left(\begin{bmatrix} -1 & 1 & 0 \\ \frac{1}{2} & -1 & 1 \\ \frac{1}{2} & 0 & -1 \end{bmatrix}\right)$ which has basis $\begin{bmatrix} 2 \\ 2 \\ 1 \end{bmatrix}$.

The corresponding equilibrium state is $\frac{1}{5}\begin{bmatrix} 2 \\ 2 \\ 1 \end{bmatrix}$. In this context, this is also known as the **PageRank vector**.

In other words, after browsing randomly for a long time, there is (about) a $\frac{2}{5} = 40\%$ chance to be at page A, a $\frac{2}{5} = 40\%$ chance to be at page B, and a $\frac{1}{5} = 20\%$ chance to be at page C.

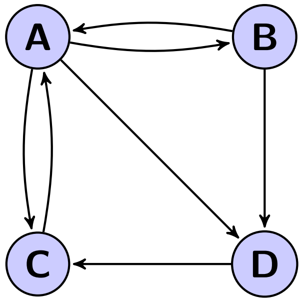
We therefore rank A and B highest (tied), and C lowest.

Just checking. Maybe we were expecting B to be ranked above A, because B is the only page that has two incoming links. However, if we are at page B, then our next click will be to page A, which is why A and B receive equal ranking.

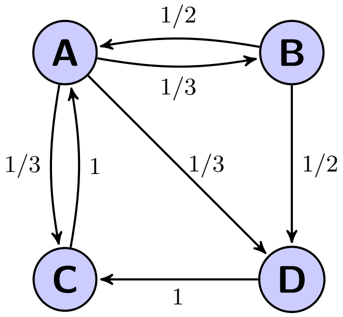
This method of ranking is the famous **PageRank** algorithm (underlying Google's search algorithm).

By the way, the algorithm is named, not after ranking web"pages", but after Larry Page (who founded Google in 1998 together with Sergey Brin).

Example 100. 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 is getting more secretive about this kind of data, so the numbers are estimates from a while ago.]

- 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 websites per month; wikipedia.org is one website, consisting of many webpages (more than 2,000,000).]

Gory details. (extra) 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 linear algebra: we often deal with big but sparse matrices.

Another practical comment. It's not an issue in our simple example, but what if our random surfer gets stuck on a webpage without links? Or, similarly, gets stuck in a loop of links? To deal with these, it is customary to include "teleportation". That is, each time, one of two things happens: with probability p (typically, something like $p = 0.85$) our surfer clicks a link as before; otherwise, with probability $1 - p$, he is teleported to some unrelated other page. Further, if the surfer comes to a page without links, he would teleport away.

A final 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 Example 94).