Computational Mathematics/Information Technology

Dr Oliver Kerr

2009–10

Recall

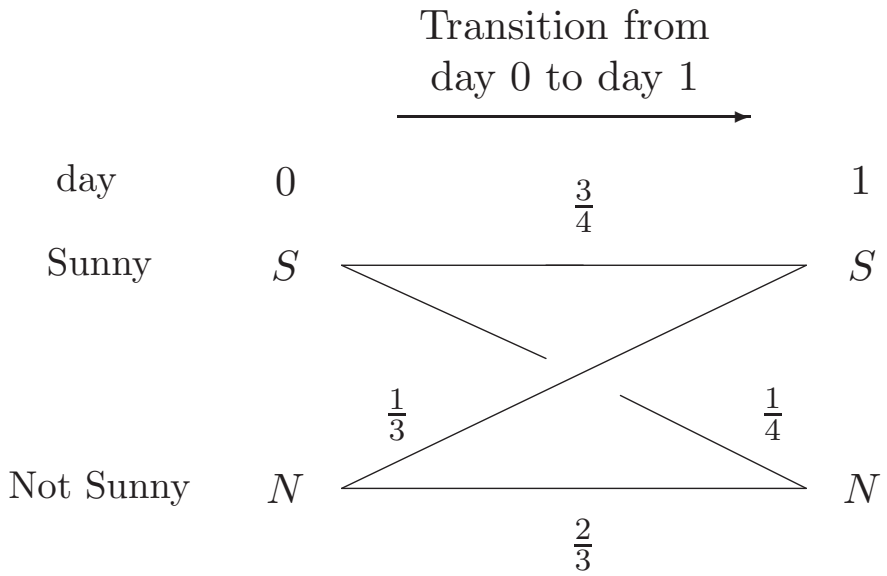
We looked at problems of the sort

Days are either "sunny" or "not sunny". One of the better ways of predicting the weather for tomorrow is to say the weather will be just the same as today's.

Question: The probability of a correct forecast if today is sunny is 3/4, and if today is not sunny is 2/3. What is the probability of the weather being sunny in 4 days time if today is sunny? What is the probability in 100 days time?

This sort of problem has the important property — the **Markov property** — that tomorrow's weather only depends on today's weather. What happened yesterday is not important.

We can display the information for the weather problem in the following transition diagram:



We can also use express the problem in a matrix formulation.

Denoting the probability of being sunny on day k by $P_k(S)$ and not sunny by $P_k(N)$ we know that:

$$P_1(S) = \frac{3}{4}P_0(S) + \frac{1}{3}P_0(N)$$

and

$$P_1(N) = \frac{1}{4}P_0(S) + \frac{2}{3}P_0(N)$$

which can be written in matrix form as

$$\begin{pmatrix} P_1(S) \\ \\ P_1(N) \end{pmatrix} = \begin{pmatrix} \frac{3}{4} & \frac{1}{3} \\ \\ \frac{1}{4} & \frac{2}{3} \end{pmatrix} \begin{pmatrix} P_0(S) \\ \\ P_0(N) \end{pmatrix} = M\mathbf{P}_0$$

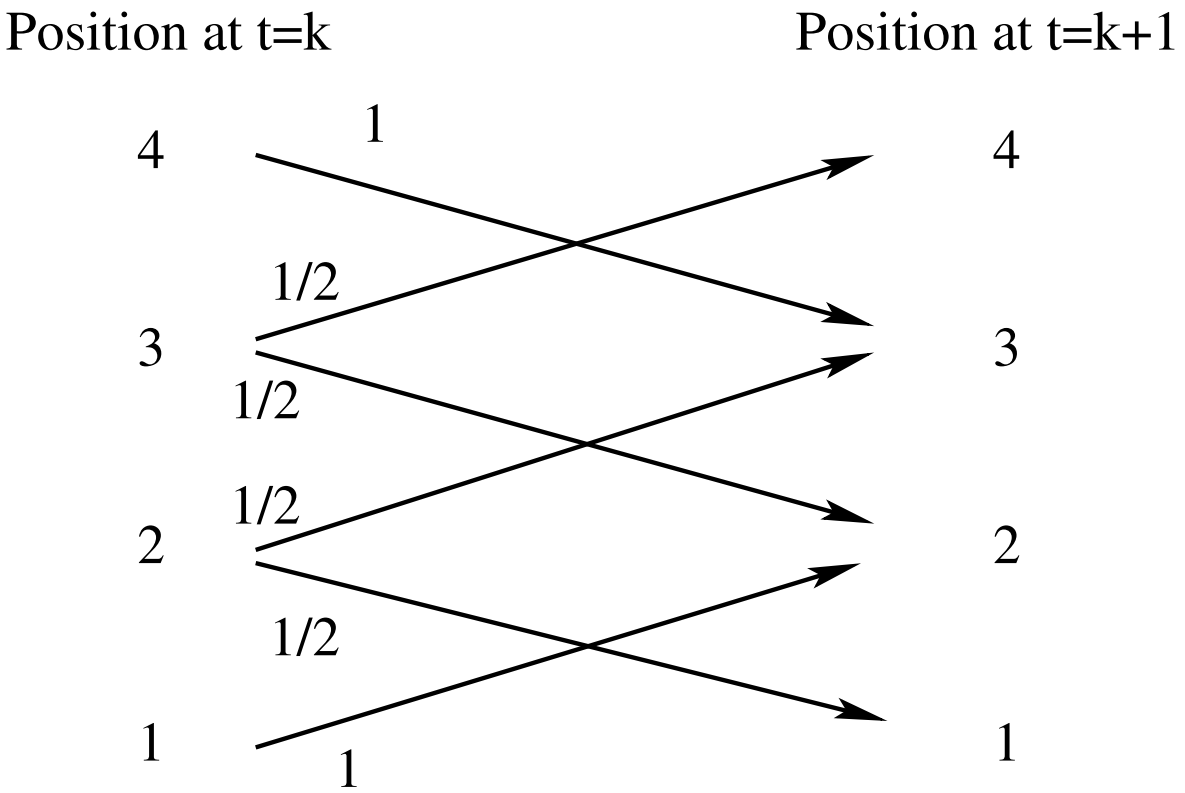
We saw last week that whatever the weather today the probability of it being sunny in 100 days is $4/7$ and not sunny is $3/7$ (Give or take an error of around 10^{-38}).

If a problem does settle down to a steady state then we will have

$$\mathbf{P} = M\mathbf{P}$$

In Linear Algebra terminology: M has an eigenvector $\mathbf{P}$ with eigenvalue 1.

However this doesn't always work, for example a finite "Drunkard's Walk"



Question: If I start in position 2, what is the probability I am in position 2 after 100 steps? What is the probability I will be in position 2 after 101 steps?

Question: If I start in position 2, what is the probability I am in position 2 after 100 steps? What is the probability I will be in position 2 after 101 steps?

The second of these is easier to answer: The probability is 0.

After one step you are on 1 or 3, after 2 steps you are on 2 or 4, and after 3 steps you are on 1 or 3 again.
After an odd number of steps you are on an odd number, and after an even number of steps an even number.

For this problem

$$M = \begin{pmatrix} 0 & 1/2 & 0 & 0 \\ 1 & 0 & 1/2 & 0 \\ 0 & 1/2 & 0 & 1 \\ 0 & 0 & 1/2 & 0 \end{pmatrix}$$

For this problem

$$M = \begin{pmatrix} 0 & 1/2 & 0 & 0 \\ 1 & 0 & 1/2 & 0 \\ 0 & 1/2 & 0 & 1 \\ 0 & 0 & 1/2 & 0 \end{pmatrix}$$

To find the solution for 100 steps consider the problem

$$\mathbf{P} = M^2\mathbf{P}$$

$$M^2 = \begin{pmatrix} 1/2 & 0 & 1/4 & 0 \\ 0 & 3/4 & 0 & 1/2 \\ 1/2 & 0 & 3/4 & 0 \\ 0 & 1/4 & 0 & 1/2 \end{pmatrix}$$

For this problem

$$M = \begin{pmatrix} 0 & 1/2 & 0 & 0 \\ 1 & 0 & 1/2 & 0 \\ 0 & 1/2 & 0 & 1 \\ 0 & 0 & 1/2 & 0 \end{pmatrix}$$

To find the solution for 100 steps consider the problem

$$\mathbf{P} = M^2\mathbf{P}$$

$$M^2 = \begin{pmatrix} 1/2 & 0 & 1/4 & 0 \\ 0 & 3/4 & 0 & 1/2 \\ 1/2 & 0 & 3/4 & 0 \\ 0 & 1/4 & 0 & 1/2 \end{pmatrix}$$

But even this has problems.

$$\mathbf{P} = M^2\mathbf{P} = \begin{pmatrix} 1/2 & 0 & 1/4 & 0 \\ 0 & 3/4 & 0 & 1/2 \\ 1/2 & 0 & 3/4 & 0 \\ 0 & 1/4 & 0 & 1/2 \end{pmatrix} \mathbf{P}$$

has solutions

$$\mathbf{P}_E = \begin{pmatrix} 1/3 \\ 0 \\ 2/3 \\ 0 \end{pmatrix} \qquad \mathbf{P}_O = \begin{pmatrix} 0 \\ 2/3 \\ 0 \\ 1/3 \end{pmatrix}$$

or any vector of the form

$$\mathbf{P} = \alpha\mathbf{P}_E + \beta\mathbf{P}_O$$

Excel cannot solve problems like this as it stands. You will have to write difficult programs to solve this, or you can use an iteration approach:
Start of with

$$\mathbf{P} = \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \end{pmatrix}$$

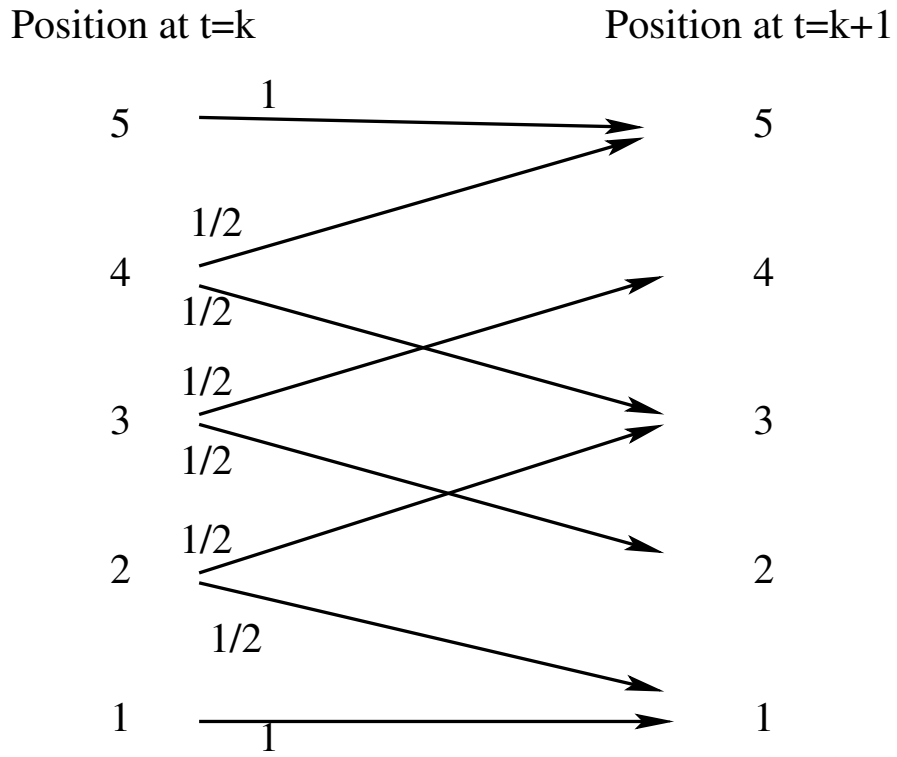
and calculate

$$\mathbf{P}_{2n} = M^{2n}\mathbf{P}_0$$

until $\mathbf{P}_{2n}$ converges.

Other possibilities exist, for example the probabilities may settle down to a steady state, but depend on the initial position:

If you have a series of positions numbered 1 to 5 and you place a counter in one of the positions 2 to 4. Toss a coin: if it is heads move up, and tails move down. If you get to 5 you win, and if you get to 1 you lose.



$$M = \begin{pmatrix} 1 & 1/2 & 0 & 0 & 0 \\ 0 & 0 & 1/2 & 0 & 0 \\ 0 & 1/2 & 0 & 1/2 & 0 \\ 0 & 0 & 1/2 & 0 & 0 \\ 0 & 0 & 0 & 1/2 & 1 \end{pmatrix}$$

$$M = \begin{pmatrix} 1 & 1/2 & 0 & 0 & 0 \\ 0 & 0 & 1/2 & 0 & 0 \\ 0 & 1/2 & 0 & 1/2 & 0 \\ 0 & 0 & 1/2 & 0 & 0 \\ 0 & 0 & 0 & 1/2 & 1 \end{pmatrix}$$

The steady states are easy to spot:

$$\mathbf{P}_W = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} \qquad \mathbf{P}_L = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 1 \end{pmatrix}$$

but working out the probability of winning if you start out in position 2 is not quite so straight forward. Again you can check for convergence of $M^n\mathbf{P}_0$ in Excel. Or do a little thinking ...

Google, PageRank and Markov Chains

Google is the most popular search engine on the web. It delivers search results quickly and, perhaps more importantly, of high quality. It delivers many possible web sites or pages that might be of interest. The addresses of the pages plus other information are ordered according to how *good* they are.

How does it manage to do this? How do you measure goodness?

Speed: It used vast numbers of computers.

Ranking of pages: It used the PageRank algorithm.

Speed: It used vast numbers of computers.

Ranking of pages: It used the PageRank algorithm.

Larry Page — co-founder of Google

The basic idea:

A page's rank is based on the number of pages that link to it

and

the PageRank of these pages.

The basic idea:

A page's rank is based on the number of pages that link to it

and

the PageRank of these pages.

This may seem like a circular argument, but mathematically it is like being asked to solve

$$\mathbf{x} = \mathbf{F}(\mathbf{x})$$

and we have seen problems like this before.

How do we set up an appropriate function?

How important is a page?

How do we set up an appropriate function?

How important is a page?

- If lots of pages link to your web page then it is likely to be important.

How do we set up an appropriate function?

How important is a page?

- If lots of pages link to your web page then it is likely to be important.
- If important web pages link to yours the it is likely to be more important.

  If the BBC has a prominent link to your web page then this will have more impact than a link from your mum's personal web page.

This still isn't very mathematical!

Random Surfer

Consider a "random surfer". The random surfer is condemned for ever to move about the internet by:

(a) Clicking arbitrarily from one page to another using existing page links.

(b) If no link exists from a page he just arbitrarily moves to another page.

(c) If at any time he gets bored using page links he again just arbitrarily moves to another page.

In (a) it is assumed that no preference is given to any one particular link on the current page, the choice of link is completely arbitrary.

(b) is important as the surfer would get stuck on a page with no links!

Similarly (c) is important otherwise you may get stuck in a group of pages that only connect to each other.

In (b) and (c) the initial model assumes that he can arbitrarily select a web site from a complete list of sites. This simplifies the model though of course it is not practically possible for a real surfer to do this.

As the random surfer clicks on into infinity each site develops a probability of being visited. (This will depend on the probability of the surfer getting bored.)

We define the PageRank of a site by:

The PageRank of a site is the probability that it will be visited by the random surfer.

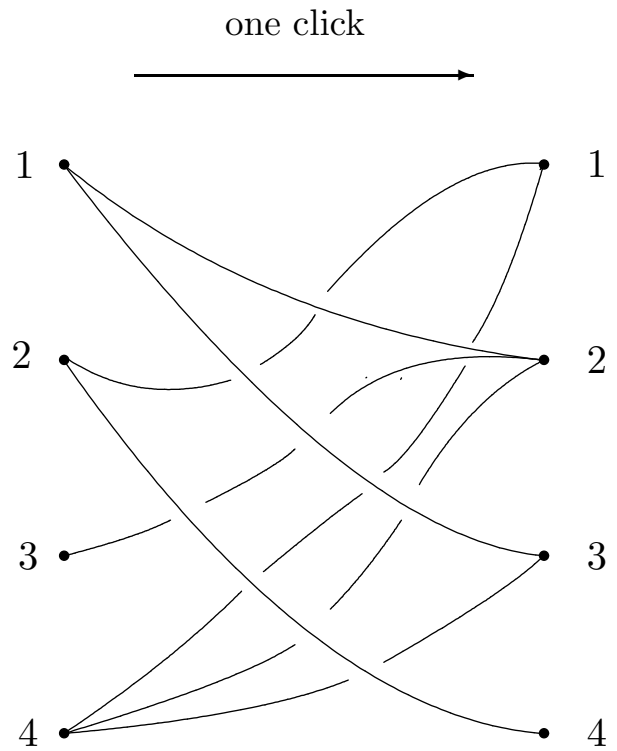
A simple World Wide web

To illustrate the ranking we will look at a small version of the World Wide Web. It only consists of four sites, each of which links to at least one other site, so we can neglect getting stuck for now.

Let the four sites be linked as follows:

- site 1 is linked to sites 2 and 3
- site 2 is linked to sites 1 and 4
- site 3 is linked to site 2 only
- site 4 is linked to all the other sites

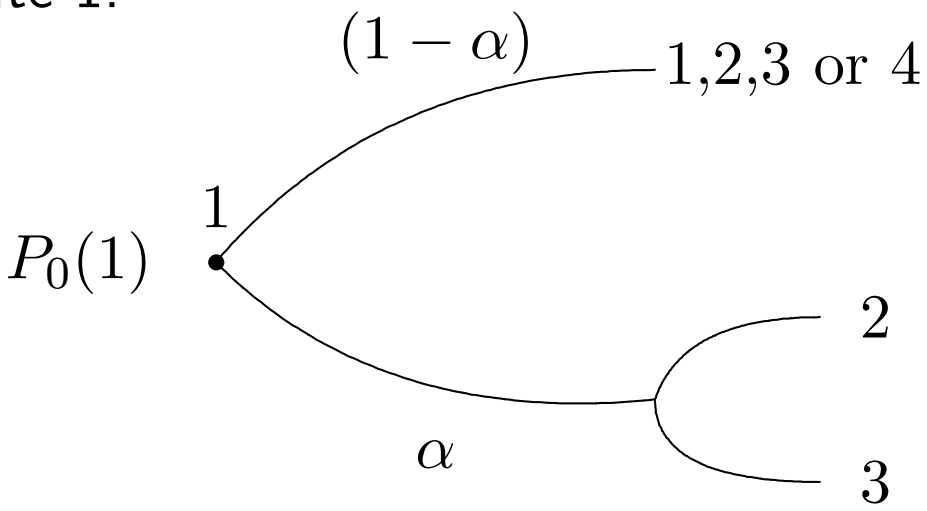
The following diagram illustrates the above linked structure:



Sites on the left are linked, via the arrows, to sites on the right.

Let us assume that the surfer decides to click to another site with a probability of α then there is a probability of $(1-\alpha)$ that he selects a site because of boredom.

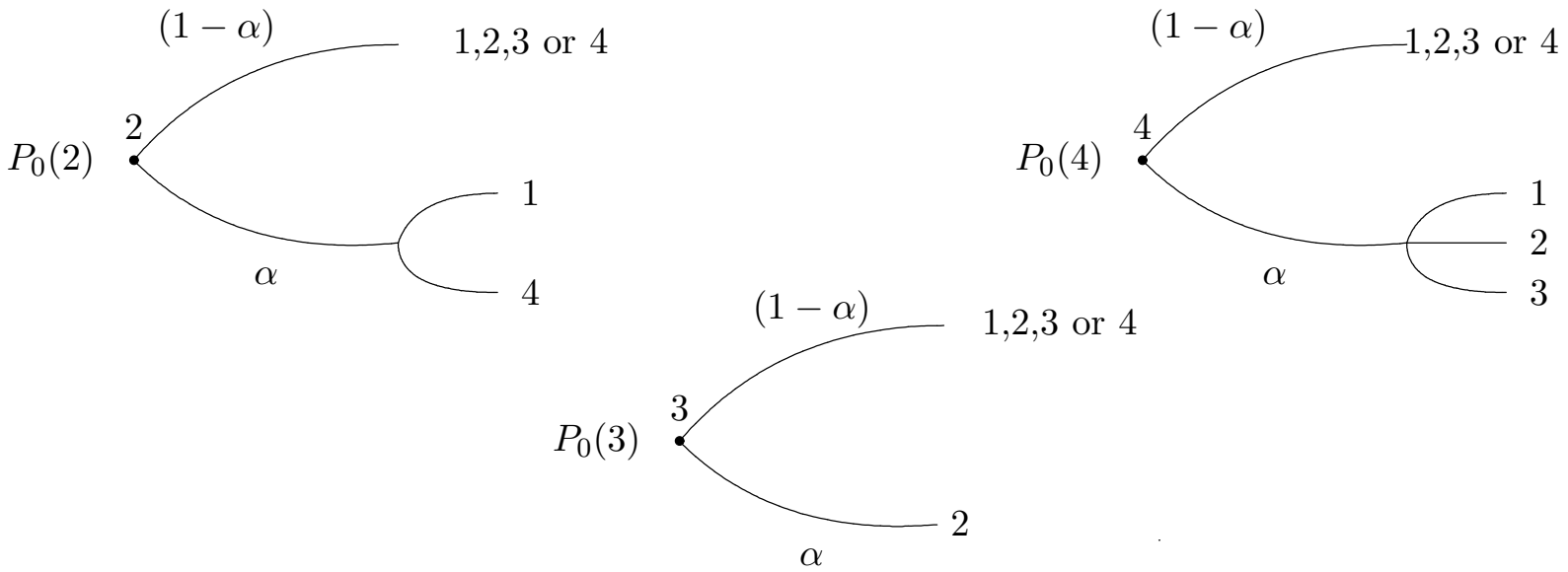
For a surfer at site 1:



$$P(1|1) = (1-\alpha)/4, \qquad P(2|1) = (1-\alpha)/4 + \alpha/2$$

$$P(3|1) = (1-\alpha)/4 + \alpha/2, \qquad P(2|1) = (1-\alpha)/4$$

For surfers at 2, 3 and 4:



The probabilities for the other transitions are

$$P(1|2) = (1-\alpha)/4 + \alpha/2, \qquad P(2|2) = (1-\alpha)/4$$
$$P(3|2) = (1-\alpha)/4, \qquad P(4|2) = (1-\alpha)/4 + \alpha/2$$

$$P(1|3) = (1-\alpha)/4, \qquad P(2|3) = (1-\alpha)/4 + \alpha$$
$$P(3|3) = (1-\alpha)/4, \qquad P(4|3) = (1-\alpha)/4$$

$$P(1|4) = (1-\alpha)/4 + \alpha/3, \qquad P(2|4) = (1-\alpha)/4 + \alpha/3$$
$$P(3|4) = (1-\alpha)/4 + \alpha/3, \qquad P(4|4) = (1-\alpha)/4$$

Denoting the probability that the surfer is currently at site n by $P_0(n)$ the probability of selecting site 1 is given by:

$$P_1(1) = P(1|1)P_0(1) + P(1|2)P_0(2) + P(1|3)P_0(3) + P(1|4)P_0(1)$$

$$= P_0(1)\frac{(1-\alpha)}{4} + P_0(2)\frac{(1-\alpha)}{4} + P_0(3)\frac{(1-\alpha)}{4} + P_0(4)\frac{(1-\alpha)}{4}$$

$$+ P_0(2)\frac{\alpha}{2} + P_0(4)\frac{\alpha}{3}$$

Using the fact that the sum of the probabilities at a given stage equals 1, i.e.

$$P_0(1) + P_0(2) + P_0(3) + P_0(4) = 1$$

this gives

$$P_1(1) = \frac{(1-\alpha)}{4} + \alpha\left\{\frac{1}{2}P_0(2) + \frac{1}{3}P_0(4)\right\}$$

Similarly

$$P_1(2) = \frac{(1-\alpha)}{4} + \alpha\left\{\frac{1}{2}P_0(1) + P_0(3) + \frac{1}{3}P_0(4)\right\}$$

$$P_1(3) = \frac{(1-\alpha)}{4} + \alpha\left\{\frac{1}{2}P_0(1) + \frac{1}{3}P_0(4)\right\}$$

$$P_1(4) = \frac{(1-\alpha)}{4} + \alpha\left\{\frac{1}{2}P_0(2)\right\}$$

In matrix form

$$\begin{pmatrix} P_1(1) \\ P_1(2) \\ P_1(3) \\ P_1(4) \end{pmatrix} = \frac{1-\alpha}{4} \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} + \alpha \begin{pmatrix} 0 & \frac{1}{2} & 0 & \frac{1}{3} \\ \frac{1}{2} & 0 & 1 & \frac{1}{3} \\ \frac{1}{2} & 0 & 0 & \frac{1}{3} \\ 0 & \frac{1}{2} & 0 & 0 \end{pmatrix} \begin{pmatrix} P_0(1) \\ P_0(2) \\ P_0(3) \\ P_0(4) \end{pmatrix}$$

Which gives:

$$\begin{pmatrix} P_1(1) \\ P_1(2) \\ P_1(3) \\ P_1(4) \end{pmatrix} = \left\{ \frac{1-\alpha}{4} \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{pmatrix} + \alpha \begin{pmatrix} 0 & \frac{1}{2} & 0 & \frac{1}{3} \\ \frac{1}{2} & 0 & 1 & \frac{1}{3} \\ \frac{1}{2} & 0 & 0 & \frac{1}{3} \\ 0 & \frac{1}{2} & 0 & 0 \end{pmatrix} \right\} \begin{pmatrix} P_0(1) \\ P_0(2) \\ P_0(3) \\ P_0(4) \end{pmatrix}$$

It may be easier to see how to get this from the expression 3 pages back!

This is of the form $\mathbf{P}_1 = M\mathbf{P}_0$.

The matrix M has each of its columns summing to 1. Hence M is a Markov matrix and the random surfing defines a Markov chain.

For any initial site, $\mathbf{P}_n$ will tend to some steady-state vector, $\mathbf{P}_\infty$. This vector represents the probabilities of a site being visited by the surfer.

The components of $\mathbf{P}_\infty$ are the PageRanks of each site.

Using Excel with $\alpha = 0.85$ the steady-state vector to which the process converges is given by

$$\mathbf{P}_\infty = \begin{pmatrix} 0.247 \\ 0.364 \\ 0.197 \\ 0.192 \end{pmatrix}$$

Thus the PageRanks are: site 1 $= 0.247$, site 2 $= 0.364$, site 3 $= 0.197$ and site 4 $= 0.192$