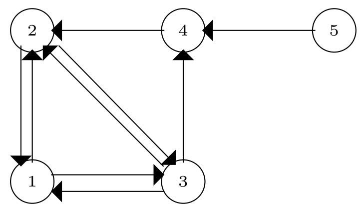
Math 212/266 Groupwork

How google took over the world!

One of the original ideas behind Google was a way of ranking the importance of every website. The basic premise was that the importance should be proportional to how often someone randomly surfing the internet would hit that site. So it just depended on the links on the internet, no human judgement was needed!

Let's see how it worked for a very small internet, one with 5 pages. The arrows show links, so, for instance 4 links to 2, but 2 does not link to 4.



Assume someone starts at 4 and follows links randomly. So first they go to 2, then they could go to 1 or 3, each with probability 0.5. Record the probabilities of where they could be after n steps as a vector

$$\begin{pmatrix} p_n^1 \\ p_n^2 \\ p_n^3 \\ p_n^4 \\ p_n^5 \end{pmatrix}.$$

So, the paragraph above means that

$$\begin{pmatrix} p_0^1 \\ p_0^2 \\ p_0^3 \\ p_0^4 \\ p_0^5 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 0 \end{pmatrix}, \quad \begin{pmatrix} p_1^1 \\ p_1^2 \\ p_1^3 \\ p_1^4 \\ p_1^5 \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \quad \begin{pmatrix} p_2^1 \\ p_2^2 \\ p_2^3 \\ p_2^4 \\ p_2^5 \end{pmatrix} = \begin{pmatrix} 0.5 \\ 0 \\ 0.5 \\ 0 \\ 0 \end{pmatrix}.$$

1. Assume the surfer starts at 1.
   (a) Where could they be after 1 step? How likely are they to be at each website? Express this as above.
   (b) Where could they be after 1 step? How likely are they to be at each website? Express this as above.

2. Write down a matrix M such that

$$M \begin{pmatrix} p_n^1 \\ p_n^2 \\ p_n^3 \\ p_n^4 \\ p_n^5 \end{pmatrix} = \begin{pmatrix} p_{n+1}^1 \\ p_{n+1}^2 \\ p_{n+1}^3 \\ p_{n+1}^4 \\ p_{n+1}^5 \end{pmatrix}$$

   Hint: question 1 (a) should give you the first column, and we found the fourth column together. **Question 1b is not relevant!**

3. A computer calculates

$$M^{10} \simeq \begin{pmatrix} 0.267 & 0.266 & 0.266 & 0.267 & 0.266 \\ 0.332 & 0.333 & 0.334 & 0.333 & 0.334 \\ 0.300 & 0.300 & 0.300 & 0.300 & 0.300 \\ 0.100 & 0.100 & 0.100 & 0.100 & 0.100 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

Using this, answer:

(a) How likely is it to be at site 3 in 10 steps if you are now at site 1?

(b) How likely is it to be at site 4 in 10 steps if you are now at site 5?

(c) How likely is it to be at site 4 in 10 steps if you are now at site 1? 2? 3? 4?

(d) Why are the columns all about the same?

(e) What does it mean that the last row is all 0s?

4. If you ask wolframalpha to diagonalize M, it tells you:

$$P = \begin{pmatrix} -1 & -2 & 2.66 & -0.887-0.48i & -0.87+0.48i \\ 1 & 2 & 3.33 & 0.63+1.45i & 0.63-1.45i \\ 0 & -3 & 3 & -0.75-0.97i & -0.75+0.97i \\ 0 & 2 & 1 & 1 & 1 \\ 0 & 1 & 0 & 0 & 0 \end{pmatrix}$$

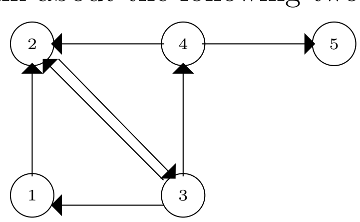
$$D = \begin{pmatrix} -0.5 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & -0.25-0.32i & 0 \\ 0 & 0 & 0 & 0 & -0.25+0.32i \end{pmatrix}$$

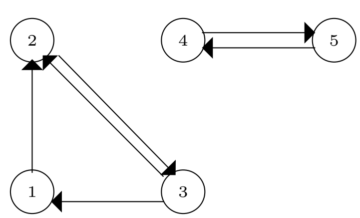
Using this, answer:

(a) How likely are you to be at site 1 after a long time?

(b) What is the most popular site?

5. Think about the following two modified graphs:

(a) 

(b) 

Where would the random surfer be after a long time in these cases? Google was unhappy with these examples. Why? **Don't calculate the matrix, just think about moving around these internets.**

6. To fix the issues like those in (5), google introduced a small chance, which they set at 0.15, that, at any step, the surfer does not follow a link, but instead just goes to a new website totally at random, with all websites on the internet being equally likely. Explain why, with this setup,

$$(0.85M + 0.03R)\begin{pmatrix} p_n^1 \\ p_n^2 \\ p_n^3 \\ p_n^4 \\ p_n^5 \end{pmatrix} = \begin{pmatrix} p_{n+1}^1 \\ p_{n+1}^2 \\ p_{n+1}^3 \\ p_{n+1}^4 \\ p_{n+1}^5 \end{pmatrix},$$

   where R is the matrix with all 1s.

7. Using WA, for the original problem, $(0.85M + 0.03R)$ has largest eigenvalue 1, and the corresponding eigenvector is $(8.13, 10.76, 9.03, 4.4, 1)^T$. Using this, what is the probability that the random surfer is at each site after say 1000 time steps? Does it matter where they started?

8. Look at the probability of being at site 5 in your answer to the previous part. Does it make sense?

9. Google always used the eigenvector of eigenvalue 1. Why will there always be an eigenvalue of 1 for matrices you get this way? Why it is always the most important?