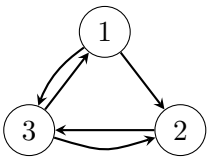
Notes for Recitation 9

When you search for a broad term (e.g., "mathematics') on Google, there are typically millions of matches; many webpages contain the word mathematics! In order to give useful results, Google needs to find a good way of ranking these results.

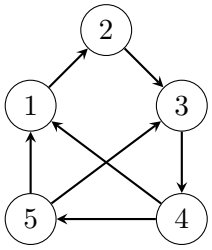
The success of Google is founded on a good way to do this, using the digraph structure of the web. The "PageRank" algorithm was invented by Larry Page and Sergey Brin. There are a lot more complications than we'll discuss here, but we will be able to understand the core of the idea.

The web graph

We can construct a digraph from the web very easily: the vertices of the graph are webpages, and there is a directed edge from webpage i to webpage j if i has a link to j. Here are two (very small) web graphs:



(a)



(b)

Just for today's recitation, we will assume that (i) the web graph has no self loops (we could just remove them, anyway), and (ii) it is strongly connected. This is very far from being true. The real PageRank algorithm does not need this, but it will simplify things a bit for us.

Problem 1: A first try

How can we use the web graph to determine the importance of a page? A natural idea is to look at the number of links to that page; the more incoming links, the higher the rank!

(a) What ranking does this give us in the two graphs above?

Solution. The ranking is just given by the indegree of the webpage. So it works fairly well in these examples, though there are some ties. ■

(b) Give some reasons why this ranking will not work well, in general. (How could you game the system?)

Solution. If you wanted to improve the ranking of your website, you could create a whole bunch of other dummy websites which just link to your main website. This will improve its indegree, and hence its ranking. This approach would never survive the advertisers... ■

Problem 2: PageRank

Let's try something more sophisticated. We would like links from "important" webpages to count more than links from unimportant ones. But this sounds self-referential, since "importance" is what we're trying to determine!

Let's try an iterative process. Let's give each webpage a million dollars (!). On the hour, each webpage does the following: it takes all its money, divides it equally amongst the webpages it links to, and sends it along to those pages. The process continues, and we hope that things settle down eventually, so that after a while the amount of money a given webpage has stays essentially constant. We then say that the importance (or PageRank) of the webpage is how much money it ends up with.

(a) Consider graph (a) shown earlier. Suppose we have x_i millions of dollars at vertex i. Find a formula for the amount of money (in millions) x_i' at each node i after 1 hour.

Solution.

$$\begin{aligned} x_1' &= x_3/2 \\ x_2' &= x_1/2 + x_3/2 \\ x_3' &= x_1/2 + x_2. \end{aligned}$$

■

(b) Figure out a formula for the amount of money $x_i^{(n)}$ (in millions) that node i has after n hours. Prove that your formula is correct using induction. *Hint: the formula for* $x_3^{(n)}$ *is* $\frac{1}{3}\left(4 - (-\frac{1}{2})^n\right)$*; this should help you get started.*

Solution. The formula is:

$$\begin{aligned} x_1^{(n)} &= \tfrac{1}{3}\left(2 + (-\tfrac{1}{2})^n\right) \\ x_2^{(n)} &= 1 \\ x_3^{(n)} &= \tfrac{1}{3}\left(4 - (-\tfrac{1}{2})^n\right). \end{aligned}$$

Use induction on n.

Base case: it is true for $n = 0$, since the formulae yield $x_i^{(0)} = 1$ for $i = 1, 2, 3$.

Suppose the formula holds for n. Applying the previous part, we see that

$$\begin{aligned} x_1^{(n+1)} &= x_3^{(n)}/2 = \tfrac{1}{3}\left(2 + (-\tfrac{1}{2})^{n+1}\right) \\ x_2^{(n+1} &= x_1^{(n)}/2 + x_3^{(n)}/2 = 1 \\ x_3^{(n+1)} &= x_1^{(n)}/2 + x_2^{(n)} = \tfrac{1}{2}\tfrac{1}{3}\left(2 + (-\tfrac{1}{2})^{n}\right) + 1 \\ &= \tfrac{1}{3}\left(4 - (-\tfrac{1}{2})^{n+1}\right). \end{aligned}$$

Thus the formula holds for all $n \geq 0$ by induction. ■

(c) What happens as n goes to infinity? Hence determine the PageRank of the webpages.

Solution. Since $(-1/2)^n \to 0$ as $n \to \infty$, we obtain the limiting amounts $p_1 = 2/3$, $p_2 = 1$ and $p_3 = 4/3$. ■

(d) The formula for $\vec{x}'$ in terms of $\vec{x}$ can be written as a matrix product: $\vec{x}' = W\vec{x}$, for some matrix W (we'll call this the *update matrix*). Determine W.

Solution.

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

■

(e) Check that it satisfies the equation $\vec{p} = W\vec{p}$, where $\vec{p}$ is the vector of PageRanks[1]. (In other words, p_i is the PageRank of page i.)

This last fact is true in general: for any strongly connected web graph G, with W being its update matrix, the equation

$$\vec{p} = W\vec{p} \tag{1}$$

is satisfied by the vector $\vec{p}$ of PageRanks. We won't prove this, but assume this for the next question.

(f) Determine the update matrix W for the web graph (b) shown earlier. Hence determine the PageRank vector $\vec{p}$ by finding a non-zero solution to (1). *(The solution is not unique; but if you add the requirement that, e.g., $p_1 = 1$, then it will be unique).*

Solution.

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

[1] We snuck in some linear algebra; some of you may recognize that p is an eigenvector of W, with associated eigenvalue 1.

The equation $\vec{p} = W\vec{p}$ yields the following system of equations:

$$\begin{aligned} p_1 &= p_4/2 + p_5/2 \\ p_2 &= p_1 \\ p_3 &= p_2 + p_5/2 \\ p_4 &= p_3 \\ p_5 &= p_3/2. \end{aligned}$$

Putting in the normalizing constraint that $p_1 = 1$, we can solve this system to obtain

$$p_1 = p_2 = 1 \qquad p_3 = p_4 = 4/3 \qquad p_4 = 2/3.$$

■