

## 1 Introduction to Google's "page rank"

The aspect of the Google search engine that we will discuss is how it ranks web pages. Suppose that you search with Google for the term "eigenvector". On Tuesday, Dec. 1, 2009 at 7:29 pm the number of hits listed was "about 812,000". Number 1 and 2 were Wikipedia articles. Number 3 was an article about eigenvectors on [www.mathworld.com](http://www.mathworld.com), a math website run by the people who sell the program Mathematica.

Numbers 4 and 5 were to a company called "Eigenvalue research", which appears to be something to do with chemistry. Number 6 is a link to the paper "The \$25,000,000,000 Eigenvector: the Linear Algebra behind Google." The reference is to the net worth of the Google company when the paper was written, in 2006. It is way out of date now.

So why was Wikipedia first and second, mathworld third, and so forth? This is the question we will explore.

Imagine a world wide web with four websites. They link to each other as shown below:

$$\begin{aligned}1 &\rightarrow 2, 1 \rightarrow 3, 1 \rightarrow 4 \\2 &\rightarrow 3, 2 \rightarrow 4 \\3 &\rightarrow 1 \\4 &\rightarrow 3, 4 \rightarrow 1\end{aligned}$$

We can draw what is called a "graph" to depict this. (This meaning of graph has nothing to do with the "graph of a function" that you are used to.) I will do this in class. You can do it by writing 1,2,3,4 at the 4 corners of a square, and drawing lines connecting these numbers, with arrows showing the direction of the link, depicting the connections shown above. If a link goes both ways, two lines are needed. Some lines will cut across the middle of the square. The study of such diagrams is called "graph theory", and linear algebra is an important tool. However, we do not have time to explore this important topic. (This kind of graph is called a "directed graph", since each connecting line has an arrow indicating a direction.)

In our first attempt at a ranking, we simply count the number of incoming links for each site:

site number	number of incoming links	
1	2	
2	1	(1)
3	3	
4	2	

So #3 gets the highest “importance score” and should be first on the list.

But one of #3’s links is from the apparently unimportant #2, while #1 has links from the more important #3 and #4. While some sort of circular reasoning seems to be involved, we may be tempted to boost #1’s score, and indeed, Google does.

So as a second attempt, we assign a score equal to the sum of the scores of all the inward linking sites. In other words, we get the following equations:

$$\begin{aligned}
 x_1 &= x_3 + x_4 \\
 x_2 &= x_1 \\
 x_3 &= x_1 + x_2 + x_4 \\
 x_4 &= x_1 + x_2.
 \end{aligned}$$

We can see linear algebra sneaking in. This system is of the form  $A\mathbf{x} = \mathbf{x}$ . Here,

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix}, \quad A = \begin{pmatrix} 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & 0 & 0 \end{pmatrix}.$$

Notice that we are therefore looking for an eigenvector of  $A$  corresponding to an eigenvalue  $\lambda = 1$ . Hence we consider

$$\begin{aligned}
 A - I &= \begin{pmatrix} -1 & 0 & 1 & 1 \\ 1 & -1 & 0 & 0 \\ 1 & 1 & -1 & 1 \\ 1 & 1 & 0 & -1 \end{pmatrix} \rightarrow \begin{pmatrix} -1 & 0 & 1 & 1 \\ 0 & -1 & 1 & 1 \\ 0 & 1 & 0 & 2 \\ 0 & 1 & 1 & 0 \end{pmatrix} \\
 &\rightarrow \begin{pmatrix} -1 & 0 & 1 & 1 \\ 0 & -1 & -1 & -1 \\ 0 & 0 & 1 & 3 \\ 0 & 0 & 2 & 1 \end{pmatrix}
 \end{aligned}$$

and we see that the rows are linearly independent, so the rank is 4 and therefore  $\lambda = 1$  is not an eigenvalue. This does not result in a ranking of the pages. There is no solution to the given equations except  $\mathbf{x} = \mathbf{0}$ .

There is another problem with the first ranking we gave, shown in table 1. Suppose that the owners of page #2 look at the system above and see that it is last in the ranking. So it creates some new web pages, say numbers 5,6,7 each of which links to #2. This boosts the score of #2 above that of #1. But more than one can play at that game, and the web would spin out of control. This problem of trying to trick the ranking system persists in every known scheme.

Let's return to our second attempt, where we tried to set up a system of equations. It had the defect that there was no non-zero solution, and the zero solution is of no help in ranking. In a modification of this method, which also tries to fix the problem of people trying to trick the system, Google puts a cap on the total influence of any page on the whole web. Instead of simply adding the scores of the inlinking pages, we will divide each of those scores by the total number of outlinks from that page. For example, since #4 outlinks to two pages, each  $x_4$  on the right side of the equations above will be replaced by  $\frac{x_4}{2}$ . Page#1 links to three other pages, so we replace  $x_1$  with  $\frac{x_1}{3}$ , etc. We get

$$\begin{aligned}x_1 &= x_3 + \frac{x_4}{2} \\x_2 &= \frac{x_1}{3} \\x_3 &= \frac{x_1}{3} + \frac{x_2}{2} + \frac{x_4}{2} \\x_4 &= \frac{x_1}{3} + \frac{x_2}{2}.\end{aligned}$$

Another interpretation of this approach is that a web surfer, currently in a certain page, is equally likely to choose any of the available links from that page.

Now we get

$$A = \begin{pmatrix} 0 & 0 & 1 & \frac{1}{2} \\ \frac{1}{3} & 0 & 0 & 0 \\ \frac{1}{3} & \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{3} & \frac{1}{2} & 0 & 0 \end{pmatrix}.$$

Note that each column adds to one. We find that

$$A - I = \begin{pmatrix} -1 & 0 & 1 & \frac{1}{2} \\ \frac{1}{3} & -1 & 0 & 0 \\ \frac{1}{3} & \frac{1}{2} & -1 & \frac{1}{2} \\ \frac{1}{3} & \frac{1}{2} & 0 & -1 \end{pmatrix}$$

Observe that each column of  $A - I$  adds to zero. This means that the sum of the rows is zero, and hence that the rank is less than 4. So the matrix is singular and there is nonzero solution  $\mathbf{x}$  to the system

$$(A - I) \mathbf{x} = \mathbf{0}.$$

So  $\lambda = 1$  is an eigenvalue. If  $\mathbf{x}$  is an eigenvector, then we just find the largest component of  $\mathbf{x}$  and rank that first, and so forth.

It turns out in the example above that there is only one linearly independent eigenvector corresponding to  $\lambda = 1$ . It is convenient to multiply this by a constant so that each element is an integer. Then we get

$$\mathbf{v} = \begin{Bmatrix} 12 \\ 4 \\ 9 \\ 6 \end{Bmatrix}$$

Now page #1 is ranked 1, page #3 is ranked 2, and so forth.

We can ask if this is liable to the same sort of gamesmanship we described earlier. Suppose page #3 adds page #5, which links to it, and it links to #5. The new matrix is

$$\begin{pmatrix} 0 & 0 & \frac{1}{2} & \frac{1}{2} & 0 \\ \frac{1}{3} & 0 & 0 & 0 & 0 \\ \frac{1}{3} & \frac{1}{2} & 0 & \frac{1}{2} & 1 \\ \frac{1}{3} & \frac{1}{2} & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{2} & 0 & 0 \end{pmatrix}$$

The columns still add up to one. We know that  $\lambda = 1$  is an eigenvalue. The quickest way to find the eigenvector is to find the echelon form of  $A - I$ . I'll use a

computer for this. Actually, the computer finds the reduced row echelon form, which is unique.

$$A - I = \begin{pmatrix} -1 & 0 & \frac{1}{2} & \frac{1}{2} & 0 \\ \frac{1}{3} & -1 & 0 & 0 & 0 \\ \frac{1}{3} & \frac{1}{2} & -1 & \frac{1}{2} & 1 \\ \frac{1}{3} & \frac{1}{2} & 0 & -1 & 0 \\ 0 & 0 & \frac{1}{2} & 0 & -1 \end{pmatrix}$$

reduced row echelon form:  $\begin{pmatrix} 1 & 0 & 0 & 0 & -\frac{4}{3} \\ 0 & 1 & 0 & 0 & -\frac{4}{9} \\ 0 & 0 & 1 & 0 & -2 \\ 0 & 0 & 0 & 1 & -\frac{2}{3} \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$  .. We can therefore read off the eigenvector (I'll multiply by 9 to get integers):

$$\begin{pmatrix} 12 \\ 4 \\ 18 \\ 6 \\ 9 \end{pmatrix}$$

Sure enough, #3 can boost its rank by this method.

The excellent article reference 9 in the “\$25 billion paper” contains a discussion of so-called “link farms” which spammers use to artificially increase the ranking of client pages, and the counter-measures taken by Google. (With work you could all read this reference, I believe. But it would be a major project, as it is 55 pages long and pretty dense. However I think this article is better than its predecessor, reference #3. It’s interesting that it is now ranked #6 among all of the 800,000 pages found when searching for eigenvector. I first used this paper in class in 2006, before it was so well known! Indeed, it was my inspiration for including this topic in a linear algebra course. I knew nothing about Google’s ranking method before seeing it. )

There are some other problems with this method. One is the possibility of “hanging nodes”. This is a website that doesn’t link to anything, but has sites linking to it. There are many such pages. The matrix  $A$  in that case has an all zero column. This difficulty is not dealt with in this paper, except for a few remarks and some references. Reference 9 contains a discussion.

Another difficulty is that there could more than one linearly independent eigenvector. Which do we choose to get our ranking? The rank of  $A - I$  might be less than  $n - 1$ . (The paper denotes the null space of  $A - \lambda I$  by  $V_\lambda(A)$ .)

Is this possible? Consider the following matrix:

$$A = \begin{pmatrix} 0 & \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} \\ 0 & 0 & 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} & 0 \end{pmatrix}$$

You will easily see that every column sums to 1. So  $\lambda = 1$  is an eigenvalue and this matrix could arise for some web. It is not hard to come up with an equivalent directed graph. Perhaps you can see that this graph would have two separate pieces. It would be “disconnected”. It turns out that the world wide web has a great many separate components! (I have been unable to find an estimate of just how many, though I’m sure such estimates exist.)

What are the eigenvectors associated with  $\lambda = 1$ ? We can easily see that

$$\begin{pmatrix} 1 \\ 1 \\ 1 \\ 0 \\ 0 \\ 0 \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 1 \\ 1 \end{pmatrix}$$

are linearly independent eigenvectors. Hence we get no unique eigenvector. This problem must be dealt with. But first we give a general discussion of the kind of matrix we encountered above.

## 2 Markov processes

**Definition:** A probability vector is a vector  $\begin{pmatrix} u_1 \\ u_2 \\ \dots \\ u_n \end{pmatrix}$  such that

$$(i) \quad u_i \geq 0 \text{ for } i = 1, \dots, n$$

$$(ii) \quad \sum_{i=1}^n u_i = 1$$

The vector  $\mathbf{u}$  gives the probabilities that each of a set of  $n$  outcomes will occur.

**Definition:** A probability matrix is a matrix

$$A = (a_{ij})$$

such that

$$(i) \quad a_{ij} \geq 0 \text{ for } i, j = 1, \dots, n$$

$$(ii) \quad \sum_{i=1}^n a_{ij} = 1 \text{ for } j = 1, \dots, n.$$

The second condition says that the sum of each column is one. If  $\mathbf{u}$  is a probability vector and  $A$  is a probability matrix, then  $A\mathbf{u}$  is also a probability vector.

Example: Suppose you know the following two facts:

1. It is sunny today
2. If it is sunny one day, the chances are 70% it will be sunny the next day,  
but if it is cloudy one day, the chances are 80% it will be cloudy the next day.

Question: What are the odds that it will be sunny 10 days from now.

We can start out on a solution:

---

It is sunny today so the chances are 70 % it will be sunny tomorrow, and 30% it will be cloudy. If it is sunny tomorrow the chances are 70% it will be sunny

the next day. The odds of :sunny today, sunny tomorrow, sunny the next day are:  $\left(\frac{7}{10}\right)\left(\frac{7}{10}\right) = \frac{49}{100}$ . (We already know it is sunny today.) Also, the odds of sunny, cloudy, sunny are  $\left(\frac{3}{10}\right)\left(\frac{2}{10}\right) = \frac{3}{50}$ , etc. It will take a long while to get to ten days for now. We will use linear algebra to simplify this problem.

We set the problem up using the probability vector

$$\mathbf{u}_k = \begin{pmatrix} s_k \\ c_k \end{pmatrix}$$

where  $s_k$  is the probability that it will be sunny on the  $k^{\text{th}}$  day and  $c_k$  is the probability that it will be cloudy on the  $k^{\text{th}}$  day. Since sunny and cloudy are the only possibilities, we must have  $s_k + c_k = 1$ .

We also have a “transition matrix ”

$$T = \begin{pmatrix} t_{11} & t_{12} \\ t_{21} & t_{22} \end{pmatrix}$$

where  $t_{11}$  is the probability that it will be sunny tomorrow if it is cloudy today,  $t_{21}$  is the probability that it will be cloudy tomorrow if it is sunny today, and  $t_{12}, t_{22}$  are the probabilities that it will be sunny or cloudy tomorrow if it is cloudy today. Our description above gives that

$$T = \begin{pmatrix} .7 & .2 \\ .3 & .8 \end{pmatrix}.$$

Note that each column has sum 1.

Suppose that we have a given  $\mathbf{u} = \begin{pmatrix} s \\ c \end{pmatrix}$  with  $0 \leq s \leq 1, 0 \leq c \leq 1$ , and  $s + c = 1$ . Then

$$T\mathbf{u} = \begin{pmatrix} .7 & .2 \\ .3 & .8 \end{pmatrix} \begin{pmatrix} s \\ c \end{pmatrix} = \begin{pmatrix} .7s + .2c \\ .3s + .8c \end{pmatrix}.$$

We easily see that the sum of the components of  $T\mathbf{u}$  equals 1:  $(.7s + .2c) + (.3s + .8c) = s + c = 1$ .

The essence of a Markov process is that the probability of an outcome at step  $n + 1$  depends only on the state at step  $n$ , and not on any of the previous states. Tossing of a “fair” coin is a classic example.

To solve our problem we have to apply, or iterate, the linear transformation defined by the matrix  $T$  ten times. Thus we need to compute

$$T^{10}\mathbf{u}_0.$$

Here, from the statement of the problem,  $\mathbf{u}_0 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$ , since we know that it is sunny when we start the process. What is needed is a quick way to calculate  $T^{10}$ . We saw in some homework that this can be done by diagonalizing the matrix. I will not take the time to do this here. The result is that

$$M^{-1}TM = D$$

where

$$M = \begin{pmatrix} 2 & 1 \\ 3 & -1 \end{pmatrix}, \quad D = \begin{pmatrix} 1 & 0 \\ 0 & .5 \end{pmatrix}.$$

Recall that the columns of  $M$  are the eigenvectors of  $T$ , and the elements on the diagonal of  $D$  are the eigenvalues of  $T$ . We saw above in the Google part that 1 is always an eigenvalue of a probability matrix.

Now we can calculate  $T^{10}$  :

$$T^{10} = MD^{10}M^{-1} = \begin{pmatrix} \frac{2051}{5120} & \frac{1023}{2560} \\ \frac{3069}{5120} & \frac{1537}{2560} \end{pmatrix}$$

Look carefully at those fractions, and you see that this is not too far from something simpler:

$$\begin{pmatrix} \frac{2}{5} & \frac{2}{5} \\ \frac{3}{5} & \frac{1}{5} \end{pmatrix}.$$

Using this simpler matrix to find our final answer, we can say that the probability vector  $\mathbf{u}^{10}$  is approximately

$$\begin{pmatrix} \frac{2}{5} & \frac{2}{5} \\ \frac{3}{5} & \frac{1}{5} \end{pmatrix} \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} \frac{2}{5} \\ \frac{3}{5} \end{pmatrix}.$$

So there is about a 40% chance that the weather will be cloudy on day 10.

### 3 Application to Google

We now give the most important theorem about probability matrices. But first we need a definition.

**Definition:** A probability matrix  $A$  is called “regular” if there is a  $k$  such that  $A^k$  has all entries positive.

**Perron Frobenius Theorem:** If  $A$  is a regular probability matrix (or ‘transition matrix’), then  $\lambda = 1$  is an eigenvalue of geometric multiplicity one. Further any other eigenvalue satisfies  $|\lambda| < 1$ . Finally,

$$\lim_{k \rightarrow \infty} A^k$$

is the matrix whose columns are all the eigenvector  $\mathbf{v}$  corresponding to the eigenvalue 1.

**Definition:** The principle eigenvalue of a probability matrix is the eigenvalue of largest absolute value. (When the matrix is regular, this is 1.) The corresponding eigenvector is called the principle eigenvector.

---

We have seen that the existence of a unique eigenvector (except for multiplication by a constant) is important if we are to be able to rank web pages. Unfortunately, we cannot expect to have a unique eigenvector for the Google matrix. To see why, consider the following matrix which we discussed a few pages back.

$$A = \begin{pmatrix} 0 & \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} \\ 0 & 0 & 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} & 0 \end{pmatrix}$$

In class, we will draw the graph showing the web connections that would yield this

matrix. Here I will depict it as we did our first “web” above:

$$\begin{aligned}1 &\rightarrow 2, 1 \rightarrow 3 \\2 &\rightarrow 1, 2 \rightarrow 3 \\3 &\rightarrow 1, 3 \rightarrow 2 \\4 &\rightarrow 5, 4 \rightarrow 6 \\5 &\rightarrow 4, 5 \rightarrow 6 \\6 &\rightarrow 4, 6 \rightarrow 5\end{aligned}$$

From this, or more easily from the graph, we see that there are two separate pieces to this web, and no links between these two pieces. This is reflected in the fact that  $A$  is a block diagonal matrix. And we saw earlier that there are two eigenvectors associated with  $\lambda = 1$ , and therefore no unique page ranking.

The Google method for coping with this is to cheat, with a “fudge factor”. Let  $S$  denote the matrix with every entry equal to  $\frac{1}{n}$ , where  $A$  is  $n \times n$ . (Remember,  $n$  is a number in the billions!) Instead of  $A$  we consider the matrix

$$M = (1 - m)A + mS,$$

where  $m$  is some number between 0 and 1.

Obviously information is being lost here. The matrix  $M$  does not reflect exactly the link structure of the web. But if  $m$  is small, then  $M$  is close to  $A$ . The choice of  $m$  is important. For reasons described in the reference 9 of the \$25 billion paper, Google sets  $m = .15$ .

$M$  is obviously regular, since every entry is now positive. So, there is a unique principle eigenvector, and we can find it by computing

$$\lim_{k \rightarrow \infty} M^k.$$

We hope that this converges quickly, so a small value of  $k$ , say three or four, will give a good approximation. In fact this does occur. Nevertheless, even finding  $M^2$  seems formidable, recalling the size of  $M$ . This is especially so because  $M$  has no zero elements. This means that compute  $M^2$  requires  $n^3$  multiplications. For example,

$$\begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}$$

requires 8 multiplications.

It becomes feasible when we realize, first, that we do not have to compute  $\lim_{k \rightarrow \infty} M^k$ . If  $\mathbf{x}$  is any probability vector, and  $M_\infty$  is a matrix with every column equal to the principle eigenvector  $\mathbf{v}$ , then  $M_\infty \mathbf{x} = \mathbf{v}$ . Therefore, we choose a probability vector  $\mathbf{x}$ , and compute  $M\mathbf{x}, M(M\mathbf{x}),$  etc. Now each stage involves  $n^2$  multiplications.

But when  $n$  is in the billions, this is still a lot of multiplications. But we realize that

$$M\mathbf{x} = (1 - m) A\mathbf{x} + mS\mathbf{x} = (1 - m) A\mathbf{x} + m\mathbf{s},$$

where  $\mathbf{s}$  is the vector with all entries equal to  $\frac{1}{n}$ . This now becomes feasible, because  $A$  will have only a relatively small number of nonzero entries (those which have direct links to the pages containing the search phrase.) For example, there are around 800,000 pages with the word eigenvector. Many will have no other page linking to them. A few, like the Wikipedia article, may have more, but not a huge number. (We are talking about the links to a site from other sites, **not** the number of visitors to a site.) So  $A$  will have a very large percentage of zeros. This is what makes calculation of the page rank feasible.

---

Homework, due next Friday, Dec. 9.

Problems from Google paper: (These start in section 2.2.2 .)

Exercise 4, Exercise 11 (pg. 578), Your calculator can probably find the eigenvectors for you.