Math 2B, Wtr 2013

What is Google Up To?

The Google search engine, begun in 1998 by two Stanford grad students (Brin and Page), uses the mathematics of eigenvalues and eigenvectors as the basis for its search engine. Imagine all the sites that are indexed (about 40 billion as of this month) arranged in a giant 40 billion by 40 billion square matrix A. The *i,j* the entry of A is 1 if site j contains a link to site i, and it is 0 otherwise. For example, the following matrix shows the connections between web sites for a company with seven employees, each of whom maintains a web site with links to coworkers:

$$A = \begin{array}{c} \\ E_1 \\ E_2 \\ E_3 \\ E_4 \\ E_5 \\ E_6 \\ E_7 \end{array} \begin{array}{c} \begin{array}{ccccccc} E_1 & E_2 & E_3 & E_4 & E_5 & E_6 & E_7 \end{array} \\ \begin{bmatrix} 0 & 1 & 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 1 & 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \end{array}$$

Thus employee 4 has links at her site to employees 1, 2, and 6. Search engines need to rank sites in order of importance, and list the most important sites first. They might simply count the number of sites that link to a given site, and use that count to rank the sites, but that does not take into account how high or low are the ranks of the sites that link to a given site. The Google method assigns an "importance" number x_i to the *i*th employee's site; this number is proportional to the sum of the "importances" of all the sites that link to that employee's site. (This is based on a mathematical technique developed by Kendall and Wei in the 1950's). Thus, if K is that constant of proportionality, then for the matrix above,

$$x_1 = K(x_2 + x_4 + x_5)$$
$$x_2 = K(x_1 + x_3 + x_4 + x_6)$$
$$etc.$$

If we let $\vec{x} = [x_1, x_2, ..., x_7]^*$ be the "importance vector" for all the employees, then we get the matrix equation $\vec{x} = KA\vec{x},\ \ or\ \ \frac{1}{K}\vec{x} = A\vec{x}$, which is clearly an eigenvector calculation! The search software looks for an eigenvector with all positive entries, and assigns rankings to the employees' sites according to the entries in this eigenvector (why positive?). For the above matrix, this eigenvector is (approximately)

X^*=[.4261, .4746, .2137, .3596, .4416, .4214, .2137].

(1) Find the value of the associated eigenvalue 1/K = ___________

Then the entry in that eigenvector with the greatest value is the most important web site! Why might this kind of calculation be useful? Multiplying a matrix by itself magnifies the importance of the direction of the eigenvector, and the coordinate which is largest has the greatest effect on this process. Notice that multiplying A by itself also leads to a new matrix in which the *i,j* entry represents the total number of links of length two from employee *j*'s site to employee *i*'s site (we can say "from site *j* to site *i*"):

$$[A^2]_{i,j} = a_{i,1}a_{1,j} + a_{i,2}a_{2,j} + ... + a_{i,k}a_{k,j} + ... + a_{i,n}a_{n,j}$$

For example, $a_{i,k}a_{k,j}$ =1 if and only if $a_{i,k}$ and $a_{k.j}$ both equal 1, indicating that there is a link from site *j* to site *k* and then a link from site *k* to site *i*.. In the same way, the *i,j* entry of A^n represents the total number of links of length n from site j to site i. Use this information to find:

(2) The total number of links of length two beginning or ending at employee 5's site: ___________

Sources: Elem. Linear Algebra, by Kolman and Hill (2004), and "Searching the Web with eigenvectors," by Herbert S. Wilf (2001, at http://www.math.upenn.edu/~wilf/website/KendallWei.pdf). Brin and Page's paper is also available – look for it with Google! Size of the web at www.worldwidewebsize.com. See also *Google's PageRank and Beyond* (2006), and *Who's #1?* (2012), both by Amy Langville and Carl Meyer.