THE $25,000,000,000* EIGENVECTOR
THE LINEAR ALGEBRA BEHIND GOOGLE

KURT BRYAN† AND TANYA LEISE‡

Abstract. Google's success derives in large part from its PageRank algorithm, which ranks the importance of webpages according to an eigenvector of a weighted link matrix. Analysis of the PageRank formula provides a wonderful applied topic for a linear algebra course. Instructors may assign this article as a project to more advanced students, or spend one or two lectures presenting the material with assigned homework from the exercises. This material also complements the discussion of Markov chains in matrix algebra. Maple and Mathematica files supporting this material can be found at www.rose-hulman.edu/∼bryan.

Key words. linear algebra, PageRank, eigenvector, stochastic matrix

AMS subject classifications. 15-01, 15A18, 15A51

1. Introduction. When Google went online in the late 1990's, one thing that set it apart from other search engines was that its search result listings always seemed deliver the "good stuff" up front. With other search engines you often had to wade through screen after screen of links to irrelevant web pages that just happened to match the search text. Part of the magic behind Google is its PageRank algorithm, which quantitatively rates the importance of each page on the web, allowing Google to rank the pages and thereby present to the user the more important (and typically most relevant and helpful) pages first.

Understanding how to calculate PageRank is essential for anyone designing a web page that they want people to access frequently, since getting listed first in a Google search leads to many people looking at your page. Indeed, due to Google's prominence as a search engine, its ranking system has had a deep influence on the development and structure of the internet, and on what kinds of information and services get accessed most frequently. Our goal in this paper is to explain one of the core ideas behind how Google calculates web page rankings. This turns out to be a delightful application of standard linear algebra.

Search engines such as Google have to do three basic things:

1. Crawl the web and locate all web pages with public access.
2. Index the data from step 1, so that it can be searched efficiently for relevant keywords or phrases.
3. Rate the importance of each page in the database, so that when a user does a search and the subset of pages in the database with the desired information has been found, the more important pages can be presented first.

This paper will focus on step 3. In an interconnected web of pages, how can one meaningfully define and quantify the "importance" of any given page?

The rated importance of web pages is not the only factor in how links are presented, but it is a significant one. There are also successful ranking algorithms other than PageRank. The interested reader will find a wealth of information about ranking algorithms and search engines, and we list just a few references for getting started (see the extensive bibliography in [9], for example, for a more complete list). For a brief overview of how Google handles the entire process see [6], and for an in-depth treatment of PageRank see [3] and a companion article [9]. Another article with good concrete examples is [5]. For more background on PageRank and explanations of essential principles of web design to maximize a website's PageRank, go to the websites [4, 11, 14]. To find out more about search engine principles in general and other ranking algorithms, see [2] and [8]. Finally, for an account of some newer approaches to searching the web, see [12] and [13].

2. Developing a formula to rank pages.

*THE APPROXIMATE MARKET VALUE OF GOOGLE WHEN THE COMPANY WENT PUBLIC IN 2004.

†Department of Mathematics, Rose-Hulman Institute of Technology, Terre Haute, IN 47803; email: kurt.bryan@rose-hulman.edu; phone: 812) 877-8485; fax: (812)877-8883.

‡Mathematics and Computer Science Department, Amherst College, Amherst, MA 01002; email: tleise@amherst.edu; phone: (413)542-5411; fax: (413)542-2550.

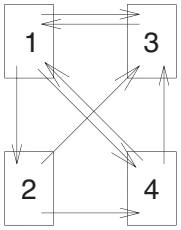

FIG. 2.1. *An example of a web with only four pages. An arrow from page A to page B indicates a link from page A to page B.*

2.1. The basic idea. In what follows we will use the phrase "importance score" or just "score" for any quantitative rating of a web page's importance. The importance score for any web page will always be a non-negative real number. A core idea in assigning a score to any given web page is that the page's score is derived from the links made to that page from other web pages. The links *to* a given page are called the *backlinks* for that page. The web thus becomes a democracy where pages vote for the importance of other pages by linking to them.

Suppose the web of interest contains n pages, each page indexed by an integer k, $1 \leq k \leq n$. A typical example is illustrated in Figure 2.1, in which an arrow from page A to page B indicates a link from page A to page B. Such a web is an example of a *directed graph.*[1] We'll use x_k to denote the importance score of page k in the web. The x_k is non-negative and $x_j > x_k$ indicates that page j is more important than page k (so $x_j = 0$ indicates page j has the least possible importance score).

A very simple approach is to take x_k as the number of backlinks for page k. In the example in Figure 2.1, we have $x_1 = 2$, $x_2 = 1$, $x_3 = 3$, and $x_4 = 2$, so that page 3 would be the most important, pages 1 and 4 tie for second, and page 2 is least important. A link to page k becomes a vote for page k's importance.

This approach ignores an important feature one would expect a ranking algorithm to have, namely, that a link to page k from an important page should boost page k's importance score more than a link from an unimportant page. For example, a link to your homepage directly from Yahoo ought to boost your page's score much more than a link from, say, `www.kurtbryan.com` (no relation to the author). In the web of Figure 2.1, pages 1 and 4 both have two backlinks: each links to the other, but page 1's second backlink is from the seemingly important page 3, while page 4's second backlink is from the relatively unimportant page 1. As such, perhaps we should rate page 1's importance higher than that of page 4.

As a first attempt at incorporating this idea let's compute the score of page j as the sum of the scores of all pages linking to page j. For example, consider the web of Figure 2.1. The score of page 1 would be determined by the relation $x_1 = x_3 + x_4$. Since x_3 and x_4 will depend on x_1 this scheme seems strangely self-referential, but it is the approach we will use, with one more modification. Just as in elections, we don't want a single individual to gain influence merely by casting multiple votes. In the same vein, we seek a scheme in which a web page doesn't gain extra influence simply by linking to lots of other pages. If page j contains n_j links, one of which links to page k, then we will boost page k's score by x_j/n_j, rather than by x_j. In this scheme each web page gets a total of one vote, *weighted by that web page's score*, that is evenly divided up among all of its outgoing links. To quantify this for a web of n pages, let $L_k \subset \{1, 2, \ldots, n\}$ denote the set of pages with a link to page k, that is, L_k is the set of page k's backlinks. For each k we require

$$x_k = \sum_{j \in L_k} \frac{x_j}{n_j}, \tag{2.1}$$

where n_j is the number of outgoing links from page j (which must be positive since if $j \in L_k$ then

[1] A graph consists of a set of *vertices* (in this context, the web pages) and a set of *edges*. Each edge joins a pair of vertices. The graph is *undirected* if the edges have no direction. The graph is *directed* if each edge (in the web context, the links) has a direction, that is, a starting and ending vertex.

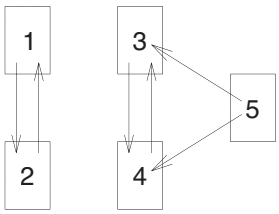

FIG. 2.2. *A web of five pages, consisting of two disconnected "subwebs" W_1 (pages 1 and 2) and W_2 (pages 3, 4, 5).*

page j links to at least page k!). We will assume that a link from a page to itself will not be counted. In this "democracy of the web" you don't get to vote for yourself!

Let's apply this approach to the four-page web of Figure 2.1. For page 1 we have $x_1 = x_3/1 + x_4/2$, since pages 3 and 4 are backlinks for page 1 and page 3 contains only one link, while page 4 contains two links (splitting its vote in half). Similarly, $x_2 = x_1/3$, $x_3 = x_1/3 + x_2/2 + x_4/2$, and $x_4 = x_1/3 + x_2/2$. These linear equations can be written $\mathbf{Ax} = \mathbf{x}$, where $\mathbf{x} = [x_1\ x_2\ x_3\ x_4]^T$ and

$$\mathbf{A} = \begin{bmatrix} 0 & 0 & 1 & \frac{1}{2} \\ \frac{1}{3} & 0 & 0 & 0 \\ \frac{1}{3} & \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{3} & \frac{1}{2} & 0 & 0 \end{bmatrix}. \tag{2.2}$$

This transforms the web ranking problem into the "standard" problem of finding an eigenvector for a square matrix! (Recall that the eigenvalues λ and eigenvectors $\mathbf{x}$ of a matrix $\mathbf{A}$ satisfy the equation $\mathbf{Ax} = \lambda\mathbf{x}$, $\mathbf{x} \neq \mathbf{0}$ by definition.) We thus seek an eigenvector $\mathbf{x}$ with eigenvalue 1 for the matrix $\mathbf{A}$. We will refer to $\mathbf{A}$ as the "link matrix" for the given web.

It turns out that the link matrix $\mathbf{A}$ in equation (2.2) does indeed have eigenvectors with eigenvalue 1, namely, all multiples of the vector $[12\ 4\ 9\ 6]^T$ (recall that any non-zero multiple of an eigenvector is again an eigenvector). Let's agree to scale these "importance score eigenvectors" so that the components sum to 1. In this case we obtain $x_1 = \frac{12}{31} \approx 0.387$, $x_2 = \frac{4}{31} \approx 0.129$, $x_3 = \frac{9}{31} \approx 0.290$, and $x_4 = \frac{6}{31} \approx 0.194$. Note that this ranking differs from that generated by simply counting backlinks. It might seem surprising that page 3, linked to by all other pages, is not the most important. To understand this, note that page 3 links only to page 1 and so casts its entire vote for page 1. This, with the vote of page 2, results in page 1 getting the highest importance score.

More generally, the matrix $\mathbf{A}$ for any web must have 1 as an eigenvalue if the web in question has no *dangling nodes* (pages with no outgoing links). To see this, first note that for a general web of n pages formula (2.1) gives rise to a matrix $\mathbf{A}$ with $A_{ij} = 1/n_j$ if page j links to page i, $A_{ij} = 0$ otherwise. The jth column of $\mathbf{A}$ then contains n_j non-zero entries, each equal to $1/n_j$, and the column thus sums to 1. This motivates the following definition, used in the study of Markov chains:

DEFINITION 2.1. *A square matrix is called a* ***column-stochastic matrix*** *if all of its entries are nonnegative and the entries in each column sum to one.*

The matrix $\mathbf{A}$ for a web with no dangling nodes is column-stochastic. We now prove

PROPOSITION 1. *Every column-stochastic matrix has* 1 *as an eigenvalue. Proof.* Let $\mathbf{A}$ be an $n \times n$ column-stochastic matrix and let $\mathbf{e}$ denote an n dimensional column vector with all entries equal to 1. Recall that $\mathbf{A}$ and its transpose $\mathbf{A}^T$ have the same eigenvalues. Since $\mathbf{A}$ is column-stochastic it is easy to see that $\mathbf{A}^T\mathbf{e} = \mathbf{e}$, so that 1 is an eigenvalue for $\mathbf{A}^T$ and hence for $\mathbf{A}$. □

In what follows we use $V_1(\mathbf{A})$ to denote the eigenspace for eigenvalue 1 of a column-stochastic matrix $\mathbf{A}$.

2.2. Shortcomings. Several difficulties arise with using formula (2.1) to rank websites. In this section we discuss two issues: webs with non-unique rankings and webs with dangling nodes.

2.2.1. Non-Unique Rankings. For our rankings it is desirable that the dimension of $V_1(\mathbf{A})$ equal one, so that there is a unique eigenvector $\mathbf{x}$ with $\sum_i x_i = 1$ that we can use for importance scores. This is true in the web of Figure 2.1 and more generally is always true for the special case of

a strongly connected web (that is, you can get from any page to any other page in a finite number of steps); see Exercise 10 below.

Unfortunately, it is not always true that the link matrix $\mathbf{A}$ will yield a unique ranking for all webs. Consider the web in Figure 2.2, for which the link matrix is

$$\mathbf{A} = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & \frac{1}{2} \\ 0 & 0 & 1 & 0 & \frac{1}{2} \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}.$$

We find here that $V_1(\mathbf{A})$ is two-dimensional; one possible pair of basis vectors is $\mathbf{x} = [1/2, 1/2, 0, 0, 0]^T$ and $\mathbf{y} = [0, 0, 1/2, 1/2, 0]^T$. But note that any linear combination of these two vectors yields another vector in $V_1(\mathbf{A})$, e.g., $\frac{3}{4}\mathbf{x} + \frac{1}{4}\mathbf{y} = [3/8, 3/8, 1/8, 1/8, 0]^T$. It is not clear which, if any, of these eigenvectors we should use for the rankings!

It is no coincidence that for the web of Figure 2.2 we find that $\dim(V_1(\mathbf{A})) > 1$. It is a consequence of the fact that if a web W, considered as an undirected graph (ignoring which direction each arrows points), consists of r disconnected subwebs $W_1, \ldots, W_r$, then $\dim(V_1(\mathbf{A})) \geq r$, and hence there is no unique importance score vector $\mathbf{x} \in V_1(\mathbf{A})$ with $\sum_i x_i = 1$. This makes intuitive sense: if a web W consists of r disconnected subwebs $W_1, \ldots, W_r$ then one would expect difficulty in finding a common reference frame for comparing the scores of pages in one subweb with those in another subweb.

Indeed, it is not hard to see why a web W consisting of r disconnected subwebs forces $\dim(V_1(\mathbf{A})) \geq r$. Suppose a web W has n pages and r component subwebs $W_1, \ldots, W_r$. Let n_i denote the number of pages in W_i. Index the pages in W_1 with indices 1 through n_1, the pages in W_2 with indices $n_1 + 1$ through $n_1 + n_2$, the pages in W_3 with $n_1 + n_2 + 1$ through $n_1 + n_2 + n_3$, etc. In general, let $N_i = \sum_{j=1}^{i} n_j$ for $i \geq 1$, with $N_0 = 0$, so W_i contains pages $N_{i-1} + 1$ through N_i. For example, in the web of Figure 2 we can take $N_1 = 2$ and $N_2 = 5$, so W_1 contains pages 1 and 2, W_2 contains pages $3, 4$, and 5. The web in Figure 2.2 is a particular example of the general case, in which the matrix $\mathbf{A}$ assumes a block diagonal structure

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_1 & \mathbf{0} & \ldots & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_2 & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \vdots & \ddots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{A}_r \end{bmatrix},$$

where $\mathbf{A}_i$ denotes the link matrix for W_i. In fact, W_i can be considered as a web in its own right. Each $n_i \times n_i$ matrix $\mathbf{A}_i$ is column-stochastic, and hence possesses some eigenvector $\mathbf{v}^i \in \mathbb{R}^{n_i}$ with eigenvector 1. For each i between 1 and r construct a vector $\mathbf{w}^i \in \mathbb{R}^n$ which has 0 components for all elements corresponding to blocks other than block i. For example,

$$\mathbf{w}^1 = \begin{pmatrix} \mathbf{v}^1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \quad \mathbf{w}^2 = \begin{pmatrix} 0 \\ \mathbf{v}^2 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \quad \ldots$$

Then it is easy to see that the vectors $\mathbf{w}^i$, $1 \leq i \leq r$, are linearly independent eigenvectors for $\mathbf{A}$

with eigenvalue 1 because

$$\mathbf{A}\mathbf{w}^i = \mathbf{A}\begin{pmatrix} 0 \\ \vdots \\ 0 \\ \mathbf{v}^i \\ 0 \\ \vdots \\ 0 \end{pmatrix} = \mathbf{w}^i.$$

Thus $V_1(\mathbf{A})$ has dimension at least r.

2.2.2. Dangling Nodes. Another difficulty may arise when using the matrix $\mathbf{A}$ to generate rankings. A web with dangling nodes produces a matrix $\mathbf{A}$ which contains one or more columns of all zeros. In this case $\mathbf{A}$ is *column-substochastic*, that is, the column sums of $\mathbf{A}$ are all less than or equal to one. Such a matrix must have all eigenvalues less than or equal to 1 in magnitude, but 1 need not actually be an eigenvalue for $\mathbf{A}$. Nevertheless, the pages in a web with dangling nodes can still be ranked use a similar technique. The corresponding substochastic matrix must have a positive eigenvalue $\lambda \leq 1$ and a corresponding eigenvector $\mathbf{x}$ with non-negative entries (called the *Perron eigenvector*) that can be used to rank the web pages. See Exercise 4 below. We will not further consider the problem of dangling nodes here, however.

EXERCISE 1. *Suppose the people who own page 3 in the web of Figure 1 are infuriated by the fact that its importance score, computed using formula (2.1), is lower than the score of page 1. In an attempt to boost page 3's score, they create a page 5 that links to page 3; page 3 also links to page 5. Does this boost page 3's score above that of page 1?*

EXERCISE 2. *Construct a web consisting of three or more subwebs and verify that* $\dim(V_1(\mathbf{A}))$ *equals (or exceeds) the number of the components in the web.*

EXERCISE 3. *Add a link from page 5 to page 1 in the web of Figure 2. The resulting web, considered as an undirected graph, is connected. What is the dimension of* $V_1(\mathbf{A})$*?*

EXERCISE 4. *In the web of Figure 2.1, remove the link from page 3 to page 1. In the resulting web page 3 is now a dangling node. Set up the corresponding substochastic matrix and find its largest positive (Perron) eigenvalue. Find a non-negative Perron eigenvector for this eigenvalue, and scale the vector so that components sum to one. Does the resulting ranking seem reasonable?*

EXERCISE 5. *Prove that in any web the importance score of a page with no backlinks is zero.*

EXERCISE 6. *Implicit in our analysis up to this point is the assertion that the manner in which the pages of a web W are indexed has no effect on the importance score assigned to any given page. Prove this, as follows: Let W contains n pages, each page assigned an index 1 through n, and let $\mathbf{A}$ be the resulting link matrix. Suppose we then transpose the indices of pages i and j (so page i is now page j and vice-versa). Let $\tilde{\mathbf{A}}$ be the link matrix for the relabelled web.*

- *Argue that $\tilde{\mathbf{A}} = \mathbf{PAP}$, where $\mathbf{P}$ is the elementary matrix obtained by transposing rows i and j of the $n \times n$ identity matrix. Note that the operation $\mathbf{A} \to \mathbf{PA}$ has the effect of swapping rows i and j of $\mathbf{A}$, while $\mathbf{A} \to \mathbf{AP}$ swaps columns i and j. Also, $\mathbf{P}^2 = \mathbf{I}$, the identity matrix.*
- *Suppose that $\mathbf{x}$ is an eigenvector for $\mathbf{A}$, so $\mathbf{Ax} = \lambda\mathbf{x}$ for some λ. Show that $\mathbf{y} = \mathbf{Px}$ is an eigenvector for $\tilde{\mathbf{A}}$ with eigenvalue λ.*
- *Explain why this shows that transposing the indices of any two pages leaves the importance scores unchanged, and use this result to argue that any permutation of the page indices leaves the importance scores unchanged.*

3. A remedy for $\dim(V_1(\mathbf{A})) > 1$**.** An enormous amount of computing resources are needed to determine an eigenvector for the link matrix corresponding to a web containing billions of pages. It is thus important to know that our algorithm will yield a unique set of sensible web rankings. The analysis above shows that our first attempt to rank web pages leads to difficulties if the web isn't connected. And the worldwide web, treated as an undirected graph, contains many disjoint components; see [9] for some interesting statistics concerning the structure of the web.

Below we present and analyze a modification of the above method that is guaranteed to overcome this shortcoming. The analysis that follows is basically a special case of the Perron-Frobenius theorem, and we only prove what we need for this application. For a full statement and proof of the Perron-Frobenius theorem, see chapter 8 in [10].

3.1. A modification to the link matrix A. For an n page web with **no dangling nodes** we can generate unambiguous importance scores as follows, including cases of web with multiple subwebs.

Let $\mathbf{S}$ denote an $n \times n$ matrix with all entries $1/n$. The matrix $\mathbf{S}$ is column-stochastic, and it is easy to check that $V_1(\mathbf{S})$ is one-dimensional. We will replace the matrix $\mathbf{A}$ with the matrix

$$\mathbf{M} = (1-m)\mathbf{A} + m\mathbf{S}, \tag{3.1}$$

where $0 \leq m \leq 1$. M is a weighted average of $\mathbf{A}$ and $\mathbf{S}$. The value of m originally used by Google is reportedly 0.15 [9, 11]. For any $m \in [0,1]$ the matrix $\mathbf{M}$ is column-stochastic and we show below that $V_1(\mathbf{M})$ is always one-dimensional if $m \in (0,1]$. Thus $\mathbf{M}$ can be used to compute unambiguous importance scores. In the case when $m = 0$ we have the original problem, for then $\mathbf{M} = \mathbf{A}$. At the other extreme is $m = 1$, yielding $\mathbf{M} = \mathbf{S}$. This is the ultimately egalitarian case: the only normalized eigenvector $\mathbf{x}$ with eigenvalue 1 has $x_i = 1/n$ for all i and all web pages are rated equally important.

Using $\mathbf{M}$ in place of $\mathbf{A}$ gives a web page with no backlinks (a dangling node) the importance score of m/n (Exercise 9), and the matrix $\mathbf{M}$ is substochastic for any $m < 1$ since the matrix $\mathbf{A}$ is substochastic. Therefore the modified formula yields nonzero importance scores for dangling links (if $m > 0$) but does not resolve the issue of dangling nodes. In the remainder of this article, we only consider webs with no dangling nodes.

The equation $\mathbf{x} = \mathbf{M}\mathbf{x}$ can also be cast as

$$\mathbf{x} = (1-m)\mathbf{A}\mathbf{x} + m\mathbf{s}, \tag{3.2}$$

where $\mathbf{s}$ is a column vector with all entries $1/n$. Note that $\mathbf{S}\mathbf{x} = \mathbf{s}$ if $\sum_i x_i = 1$.

We will prove below that $V_1(\mathbf{M})$ is always one-dimensional, but first let's look at a couple of examples.

Example 1: For the web of four pages in Figure 2.1 with matrix $\mathbf{A}$ given by (2.2), the new formula gives (with $m = 0.15$)

$$\mathbf{M} = \begin{bmatrix} 0.0375 & 0.0375 & 0.8875 & 0.4625 \\ 0.32083\bar{3} & 0.0375 & 0.0375 & 0.0375 \\ 0.32083\bar{3} & 0.4625 & 0.0375 & 0.4625 \\ 0.32083\bar{3} & 0.4625 & 0.0375 & 0.0375 \end{bmatrix},$$

and yields importance scores $x_1 \approx 0.368$, $x_2 \approx 0.142$, $x_3 \approx 0.288$, and $x_4 \approx 0.202$. This yields the same ranking of pages as the earlier computation, but the scores are slightly different.

Example 2 shows more explicitly the advantages of using $\mathbf{M}$ in place of $\mathbf{A}$.

Example 2: As a second example, for the web of Figure 2.2 with $m = 0.15$ we obtain the matrix

$$\mathbf{M} = \begin{bmatrix} 0.03 & 0.88 & 0.03 & 0.03 & 0.03 \\ 0.88 & 0.03 & 0.03 & 0.03 & 0.03 \\ 0.03 & 0.03 & 0.03 & 0.88 & 0.455 \\ 0.03 & 0.03 & 0.88 & 0.03 & 0.455 \\ 0.03 & 0.03 & 0.03 & 0.03 & 0.03 \end{bmatrix}. \tag{3.3}$$

The space $V_1(\mathbf{M})$ is indeed one-dimensional, with normalized eigenvector components of $x_1 = 0.2, x_2 = 0.2, x_3 = 0.285, x_4 = 0.285$, and $x_5 = 0.03$. The modification, using $\mathbf{M}$ instead of $\mathbf{A}$, allows us to compare pages in different subwebs.

Each entry M_{ij} of $\mathbf{M}$ defined by equation (3.1) is strictly positive, which motivates the following definition.

DEFINITION 3.1. *A matrix* $\mathbf{M}$ *is* **positive** *if* $M_{ij} > 0$ *for all* i *and* j. This is the key property that guarantees $\dim(V_1(\mathbf{M})) = 1$, which we prove in the next section.

3.2. Analysis of the matrix M. Note that Proposition 1 shows that $V_1(\mathbf{M})$ is nonempty since $\mathbf{M}$ is stochastic . The goal of this section is to show that $V_1(\mathbf{M})$ is in fact one-dimensional. This is a consequence of the following two propositions.

PROPOSITION 2. *If* $\mathbf{M}$ *is positive and column-stochastic, then any eigenvector in* $V_1(\mathbf{M})$ *has all positive or all negative components. Proof.* We use proof by contradiction. First note that in the standard triangle inequality $|\sum_i y_i| \leq \sum_i |y_i|$ (with all y_i real) the inequality is strict when the y_i are of mixed sign. Suppose $\mathbf{x} \in V_1(\mathbf{M})$ contains elements of mixed sign. From $\mathbf{x} = \mathbf{M}\mathbf{x}$ we have $x_i = \sum_{j=1}^n M_{ij}x_j$ and the summands $M_{ij}x_j$ are of mixed sign (since $M_{ij} > 0$). As a result we have a strict inequality

$$|x_i| = \left| \sum_{j=1}^n M_{ij}x_j \right| < \sum_{j=1}^n M_{ij}|x_j|. \tag{3.4}$$

Sum both sides of inequality (3.4) from $i = 1$ to $i = n$, and swap the i and j summations. Then use the fact that $\mathbf{M}$ is column-stochastic ($\sum_i M_{ij} = 1$ for all j) to find

$$\sum_{i=1}^n |x_i| < \sum_{i=1}^n \sum_{j=1}^n M_{ij}|x_j| = \sum_{j=1}^n \left(\sum_{i=1}^n M_{ij} \right) |x_j| = \sum_{j=1}^n |x_j|,$$

a contradiction. Hence $\mathbf{x}$ cannot contain both positive and negative elements. If $x_i \geq 0$ for all i (and not all x_i are zero) then $x_i > 0$ follows immediately from $x_i = \sum_{j=1}^n M_{ij}x_j$ and $M_{ij} > 0$. Similarly $x_i \leq 0$ for all i implies that each $x_i < 0$. □

The following proposition will also be useful for analyzing $\dim(V_1(\mathbf{M}))$:

PROPOSITION 3. *Let* $\mathbf{v}$ *and* $\mathbf{w}$ *be linearly independent vectors in* $\mathbb{R}^m$, $m \geq 2$. *Then, for some values of* s *and* t *that are not both zero, the vector* $\mathbf{x} = s\mathbf{v} + t\mathbf{w}$ *has both positive and negative components. Proof.* Linear independence implies neither $\mathbf{v}$ nor $\mathbf{w}$ is zero. Let $d = \sum_i v_i$. If $d = 0$ then $\mathbf{v}$ must contain components of mixed sign, and taking $s = 1$ and $t = 0$ yields the conclusion. If $d \neq 0$ set $s = -\frac{\sum_i w_i}{d}$, $t = 1$, and $\mathbf{x} = s\mathbf{v} + t\mathbf{w}$. Since $\mathbf{v}$ and $\mathbf{w}$ are independent $\mathbf{x} \neq \mathbf{0}$. However, $\sum_i x_i = 0$. We conclude that $\mathbf{x}$ has both positive and negative components. □

We can now prove that using $\mathbf{M}$ in place of $\mathbf{A}$ yields an unambiguous ranking for any web with no dangling nodes.

LEMMA 3.2. *If* $\mathbf{M}$ *is positive and column-stochastic then* $V_1(\mathbf{M})$ *has dimension* 1. *Proof.* We again use proof by contradiction. Suppose there are two linearly independent eigenvectors $\mathbf{v}$ and $\mathbf{w}$ in the subspace $V_1(\mathbf{M})$. For any real numbers s and t that are not both zero, the nonzero vector $\mathbf{x} = s\mathbf{v} + t\mathbf{w}$ must be in $V_1(\mathbf{M})$, and so have components that are all negative or all positive. But by Proposition 3, for some choice of s and t the vector $\mathbf{x}$ must contain components of mixed sign, a contradiction. We conclude that $V_1(\mathbf{M})$ cannot contain two linearly independent vectors, and so has dimension one. □

Lemma 3.2 provides the "punchline" for our analysis of the ranking algorithm using the matrix $\mathbf{M}$ (for $0 < m < 1$). The space $V_1(\mathbf{M})$ is one-dimensional, and moreover, the relevant eigenvectors have entirely positive or negative components. We are thus guaranteed the existence of a unique eigenvector $\mathbf{x} \in V_1(\mathbf{M})$ with positive components such that $\sum_i x_i = 1$.

EXERCISE 7. *Prove that if* $\mathbf{A}$ *is an* $n \times n$ *column-stochastic matrix and* $0 \leq m \leq 1$, *then* $\mathbf{M} = (1-m)\mathbf{A} + m\mathbf{S}$ *is also a column-stochastic matrix.*

EXERCISE 8. *Show that the product of two column-stochastic matrices is also column-stochastic.*

EXERCISE 9. *Show that a page with no backlinks is given importance score* $\frac{m}{n}$ *by formula (3.2).*

EXERCISE 10. *Suppose that* $\mathbf{A}$ *is the link matrix for a strongly connected web of* n *pages (any page can be reached from any other page by following a finite number of links). Show that* $\dim(V_1(\mathbf{A})) = 1$ *as follows. Let* $(\mathbf{A}^k)_{ij}$ *denote the* (i,j)*-entry of* $\mathbf{A}^k$.

- *Note that page* i *can be reached from page* j *in one step if and only* $A_{ij} > 0$ *(since* $A_{ij} > 0$ *means there's a link from* j *to* i*!) Show that* $(\mathbf{A}^2)_{ij} > 0$ *if and only if page* i *can be reached from page* j *in exactly two steps. Hint:* $(\mathbf{A}^2)_{ij} = \sum_k A_{ik}A_{kj}$*; all* A_{ij} *are non-negative, so* $(\mathbf{A}^2)_{ij} > 0$ *implies that for some* k *both* A_{ik} *and* A_{kj} *are positive.*

- *Show more generally that* $(\mathbf{A}^p)_{ij} > 0$ *if and only if page* i *can be reached from page* j *in EXACTLY* p *steps.*
- *Argue that* $(\mathbf{I} + \mathbf{A} + \mathbf{A}^2 + \cdots + \mathbf{A}^p)_{ij} > 0$ *if and only if page* i *can be reached from page* j *in* p *or fewer steps (note* $p = 0$ *is a legitimate choice—any page can be reached from itself in zero steps!)*
- *Explain why* $\mathbf{I} + \mathbf{A} + \mathbf{A}^2 + \cdots + \mathbf{A}^{n-1}$ *is a positive matrix if the web is strongly connected.*
- *Use the last part (and Exercise 8) so show that* $\mathbf{B} = \frac{1}{n}(\mathbf{I} + \mathbf{A} + \mathbf{A}^2 + \cdots + \mathbf{A}^{n-1})$ *is positive and column-stochastic (and hence by Lemma 3.2,* $\dim(V_1(\mathbf{B})) = 1$*).*
- *Show that if* $\mathbf{x} \in V_1(\mathbf{A})$ *then* $\mathbf{x} \in V_1(\mathbf{B})$*. Why does this imply that* $\dim(V_1(\mathbf{A})) = 1$*?*

EXERCISE 11. *Consider again the web in Figure 2.1, with the addition of a page 5 that links to page 3, where page 3 also links to page 5. Calculate the new ranking by finding the eigenvector of* $\mathbf{M}$ *(corresponding to* $\lambda = 1$*) that has positive components summing to one. Use* $m = 0.15$*.*

EXERCISE 12. *Add a sixth page that links to every page of the web in the previous exercise, but to which no other page links. Rank the pages using* $\mathbf{A}$*, then using* $\mathbf{M}$ *with* $m = 0.15$*, and compare the results.*

EXERCISE 13. *Construct a web consisting of two or more subwebs and determine the ranking given by formula (3.1).*

At present the web contains at least eight billion pages—how does one compute an eigenvector for an eight billion by eight billion matrix? One reasonable approach is an iterative procedure called the *power method* (along with modifications) that we will now examine for the special case at hand. It is worth noting that there is much additional analysis one can do, and many improved methods for the computation of PageRank. The reference [7] provides a typical example and additional references.

4. Computing the Importance Score Eigenvector. The rough idea behind the power method[2] for computing an eigenvector of a matrix $\mathbf{M}$ is this: One starts with a "typical" vector $\mathbf{x}_0$, then generates the sequence $\mathbf{x}_k = \mathbf{M}\mathbf{x}_{k-1}$ (so $\mathbf{x}_k = \mathbf{M}^k\mathbf{x}_0$) and lets k approaches infinity. The vector $\mathbf{x}_k$ is, to good approximation, an eigenvector for the dominant (largest magnitude) eigenvalue of $\mathbf{M}$. However, depending on the magnitude of this eigenvalue, the vector $\mathbf{x}_k$ may also grow without bound or decay to the zero vector. One thus typically rescales at each iteration, say by computing $\mathbf{x}_k = \frac{\mathbf{M}\mathbf{x}_{k-1}}{\|\mathbf{M}\mathbf{x}_{k-1}\|}$, where $\|\cdot\|$ can be any vector norm. The method generally requires that the corresponding eigenspace be one-dimensional, a condition that is satisfied in the case when $\mathbf{M}$ is defined by equation (3.1).

To use the power method on the matrices $\mathbf{M}$ that arise from the web ranking problem we would generally need to know that any other eigenvalues λ of $\mathbf{M}$ satisfy $|\lambda| < 1$. This assures that the power method will converge to the eigenvector we want. Actually, the following proposition provides what we need, with no reference to any other eigenvalues of $\mathbf{M}$!

DEFINITION 4.1. *The* 1*-norm of a vector* $\mathbf{v}$ *is* $\|\mathbf{v}\|_1 = \sum_i |v_i|$.

PROPOSITION 4. *Let* $\mathbf{M}$ *be a positive column-stochastic* $n \times n$ *matrix and let* V *denote the subspace of* $\mathbb{R}^n$ *consisting of vectors* $\mathbf{v}$ *such that* $\sum_j v_j = 0$*. Then* $\mathbf{M}\mathbf{v} \in V$ *for any* $\mathbf{v} \in V$*, and*

$$\|\mathbf{M}\mathbf{v}\|_1 \leq c\|\mathbf{v}\|_1$$

for any $\mathbf{v} \in V$*, where* $c = \max_{1\leq j\leq n} |1 - 2\min_{1\leq i\leq n} M_{ij}| < 1$. *Proof.* To see that $\mathbf{M}\mathbf{v} \in V$ is straightforward: Let $\mathbf{w} = \mathbf{M}\mathbf{v}$, so that $w_i = \sum_{j=1}^n M_{ij}v_j$ and

$$\sum_{i=1}^n w_i = \sum_{i=1}^n \sum_{j=1}^n M_{ij}v_j = \sum_{j=1}^n v_j \left(\sum_{i=1}^n M_{ij}\right) = \sum_{j=1}^n v_j = 0.$$

Hence $\mathbf{w} = \mathbf{M}\mathbf{v} \in V$. To prove the bound in the proposition note that

$$\|\mathbf{w}\|_1 = \sum_{i=1}^n e_i w_i = \sum_{i=1}^n e_i \left(\sum_{j=1}^n M_{ij}v_j\right),$$

[2]See [15] for a general introduction to the power method and the use of spectral decomposition to find the rate of convergence of the vectors $\mathbf{x}_k = \mathbf{M}^k\mathbf{x}_0$.

where $e_i = \text{sgn}(w_i)$. Note that the e_i are not all of one sign, since $\sum_i w_i = 0$ (unless $\mathbf{w} \equiv \mathbf{0}$ in which case the bound clearly holds). Reverse the double sum to obtain

$$\|\mathbf{w}\|_1 = \sum_{j=1}^{n} v_j \left(\sum_{i=1}^{n} e_i M_{ij} \right) = \sum_{j=1}^{n} a_j v_j, \tag{4.1}$$

where $a_j = \sum_{i=1}^{n} e_i M_{ij}$. Since the e_i are of mixed sign and $\sum_i M_{ij} = 1$ with $0 < M_{ij} < 1$, it is easy to see that

$$-1 < -1 + 2 \min_{1 \le i \le n} M_{ij} \le a_j \le 1 - 2 \min_{1 \le i \le n} M_{ij} < 1.$$

We can thus bound

$$|a_j| \le |1 - 2 \min_{1 \le i \le n} M_{ij}| < 1.$$

Let $c = \max_{1 \le j \le n} |1 - 2 \min_{1 \le i \le n} M_{ij}|$. Observe that $c < 1$ and $|a_j| \le c$ for all j. From equation (4.1) we have

$$\|\mathbf{w}\|_1 = \sum_{j=1}^{n} a_j v_j = \left| \sum_{j=1}^{n} a_j v_j \right| \le \sum_{j=1}^{n} |a_j||v_j| \le c \sum_{j=1}^{n} |v_j| = c\|\mathbf{v}\|_1,$$

which proves the proposition.

Proposition 4 sets the stage for the following proposition.

PROPOSITION 5. *Every positive column-stochastic matrix* $\mathbf{M}$ *has a unique vector* $\mathbf{q}$ *with positive components such that* $\mathbf{Mq} = \mathbf{q}$ *with* $\|\mathbf{q}\|_1 = 1$. *The vector* $\mathbf{q}$ *can be computed as* $\mathbf{q} = \lim_{k \to \infty} \mathbf{M}^k \mathbf{x}_0$ *for any initial guess* $\mathbf{x}_0$ *with positive components such that* $\|\mathbf{x}_0\|_1 = 1$. *Proof.* From Proposition 1 the matrix $\mathbf{M}$ has 1 as an eigenvalue and by Lemma 3.2 the subspace $V_1(\mathbf{M})$ is one-dimensional. Also, all non-zero vectors in $V_1(\mathbf{M})$ have entirely positive or negative components. It is clear that there is a unique vector $\mathbf{q} \in V_1(\mathbf{M})$ with positive components such that $\sum_i q_i = 1$.

Let $\mathbf{x}_0$ be any vector in $\mathbb{R}^n$ with positive components such that $\|\mathbf{x}_0\|_1 = 1$. We can write $\mathbf{x}_0 = \mathbf{q} + \mathbf{v}$ where $\mathbf{v} \in V$ (V as in Proposition 4). We find that $\mathbf{M}^k \mathbf{x}_0 = \mathbf{M}^k \mathbf{q} + \mathbf{M}^k \mathbf{v} = \mathbf{q} + \mathbf{M}^k \mathbf{v}$. As a result

$$\mathbf{M}^k \mathbf{x}_0 - \mathbf{q} = \mathbf{M}^k \mathbf{v}. \tag{4.2}$$

A straightforward induction and Proposition 4 shows that $\|\mathbf{M}^k \mathbf{v}\|_1 \le c^k \|\mathbf{v}\|_1$ for $0 \le c < 1$ (c as in Proposition 4) and so $\lim_{k \to \infty} \|\mathbf{M}^k \mathbf{v}\|_1 = 0$. From equation (4.2) we conclude that $\lim_{k \to \infty} \mathbf{M}^k \mathbf{x}_0 = \mathbf{q}$. □

Example: Let $\mathbf{M}$ be the matrix defined by equation (3.3) for the web of Figure 2.2. We take $\mathbf{x}_0 = [0.24, 0.31, 0.08, 0.18, 0.19]^T$ as an initial guess; recall that we had $\mathbf{q} = [0.2, 0.2, 0.285, 0.285, 0.03]^T$. The table below shows the value of $\|\mathbf{M}^k \mathbf{x}_0 - \mathbf{q}\|_1$ for several values of k, as well as the ratio $\|\mathbf{M}^k \mathbf{x}_0 - \mathbf{q}\|_1 / \|\mathbf{M}^{k-1} \mathbf{x}_0 - \mathbf{q}\|_1$. Compare this ratio to c from Proposition 4, which in this case is 0.94.

k	$\|\mathbf{M}^k \mathbf{x}_0 - \mathbf{q}\|_1$	$\frac{\|\mathbf{M}^k \mathbf{x}_0 - \mathbf{q}\|_1}{\|\mathbf{M}^{k-1} \mathbf{x}_0 - \mathbf{q}\|_1}$
0	0.62	
1	0.255	0.411
5	0.133	0.85
10	0.0591	0.85
50	8.87×10^{-5}	0.85

It is clear that the bound $\|\mathbf{M}^k \mathbf{x}_0 - \mathbf{q}\|_1 \le c^k \|\mathbf{x}_0 - \mathbf{q}\|_1$ is rather pessimistic (note 0.85 is the value $1 - m$, and 0.85 turns out to be the second largest eigenvalue for $\mathbf{M}$). One can show that in general the power method will converge asymptotically according to $\|\mathbf{M}\mathbf{x}_k - \mathbf{q}\|_1 \approx |\lambda_2| \|\mathbf{x} - \mathbf{q}\|_1$, where λ_2

is the second largest eigenvalue of $\mathbf{M}$. Moreover, for $\mathbf{M}$ of the form $\mathbf{M} = (1-m)\mathbf{A} + m\mathbf{S}$ with $\mathbf{A}$ column-stochastic and all $S_{ij} = 1/n$ it can be shown that $|\lambda_2| \leq 1-m$ (see, e.g., [1], Theorem 5.10). As a result, the power method will converge much more rapidly than indicated by $c^k\|\mathbf{x}_0 - \mathbf{q}\|_1$. Nonetheless, the value of c in Proposition 4 provides a very simple bound on the convergence of the power method here. It is easy to see that since all entries of $\mathbf{M}$ are at least m/n, we will always have $c \leq 1 - 2m/n$ in Proposition 4.

As a practical matter, note that the $n \times n$ positive matrix $\mathbf{M}$ has no non-zero elements, so the multiplication $\mathbf{Mv}$ for $\mathbf{v} \in \mathbb{R}^n$ will typically take $O(n^2)$ multiplications and additions, a formidable computation if $n = 8,000,000,000$. But equation (3.2) shows that if $\mathbf{x}$ is positive with $\|\mathbf{x}\|_1 = 1$ then the multiplication $\mathbf{Mx}$ is equivalent to $(1-m)\mathbf{Ax} + m\mathbf{s}$. This is a far more efficient computation, since $\mathbf{A}$ can be expected to contain mostly zeros (most web pages link to only a few other pages). We've now proved our main theorem:

THEOREM 4.2. *The matrix $\mathbf{M}$ defined by (3.1) for a web with no dangling nodes will always be a positive column-stochastic matrix and so have a unique $\mathbf{q}$ with positive components such that $\mathbf{Mq} = \mathbf{q}$ and $\sum_i q_i = 1$. The vector $\mathbf{q}$ may be computed as the limit of iterations $\mathbf{x}_k = (1-m)\mathbf{Ax}_{k-1} + m\mathbf{s}$, where $\mathbf{x}_0$ is any initial vector with positive components and $\|\mathbf{x}_0\|_1 = 1$.*

The eigenvector $\mathbf{x}$ defined by equation (3.2) also has a probabilistic interpretation. Consider a web-surfer on a web of n pages with no dangling nodes. The surfer begins at some web page (it doesn't matter where) and randomly moves from web page to web page according to the following procedure: If the surfer is currently at a page with r outgoing links, he either randomly chooses any one of these links with uniform probability $\frac{1-m}{r}$ OR he jumps to any randomly selected page on the web, each with probability $\frac{m}{n}$ (note that $r\frac{1-m}{r} + n\frac{m}{n} = 1$, so this accounts for everything he can do). The surfer repeats this page-hopping procedure ad infinitum. The component x_j of the normalized vector $\mathbf{x}$ in equation (3.2) is the fraction of time that the surfer spends, in the long run, on page j of the web. More important pages tend to be linked to by many other pages and so the surfer hits those most often.

EXERCISE 14. *For the web in Exercise 11, compute the values of $\|\mathbf{M}^k\mathbf{x}_0 - \mathbf{q}\|_1$ and $\frac{\|\mathbf{M}^k\mathbf{x}_0 - \mathbf{q}\|_1}{\|\mathbf{M}^{k-1}\mathbf{x}_0 - \mathbf{q}\|_1}$ for $k = 1, 5, 10, 50$, using an initial guess $\mathbf{x}_0$ not too close to the actual eigenvector $\mathbf{q}$ (so that you can watch the convergence). Determine $c = \max_{1\leq j\leq n} |1 - 2\min_{1\leq i\leq n} M_{ij}|$ and the absolute value of the second largest eigenvalue of $\mathbf{M}$.*

EXERCISE 15. *To see why the second largest eigenvalue plays a role in bounding $\frac{\|\mathbf{M}^k\mathbf{x}_0 - \mathbf{q}\|_1}{\|\mathbf{M}^{k-1}\mathbf{x}_0 - \mathbf{q}\|_1}$, consider an $n \times n$ positive column-stochastic matrix $\mathbf{M}$ that is diagonalizable. Let $\mathbf{x}_0$ be any vector with non-negative components that sum to one. Since $\mathbf{M}$ is diagonalizable, we can create a basis of eigenvectors $\{\mathbf{q}, \mathbf{v}_1, \ldots, \mathbf{v}_{n-1}\}$, where $\mathbf{q}$ is the steady state vector, and then write $\mathbf{x}_0 = a\mathbf{q} + \sum_{k=1}^{n-1} b_k\mathbf{v}_k$. Determine $\mathbf{M}^k\mathbf{x}_0$, and then show that $a = 1$ and the sum of the components of each $\mathbf{v}_k$ must equal 0. Next apply Proposition 4 to prove that, except for the non-repeated eigenvalue $\lambda = 1$, the other eigenvalues are all strictly less than one in absolute value. Use this to evaluate $\lim_{k\to\infty} \frac{\|\mathbf{M}^k\mathbf{x}_0 - \mathbf{q}\|_1}{\|\mathbf{M}^{k-1}\mathbf{x}_0 - \mathbf{q}\|_1}$.*

EXERCISE 16. *Consider the link matrix*

$$\mathbf{A} = \begin{bmatrix} 0 & \frac{1}{2} & \frac{1}{2} \\ 0 & 0 & \frac{1}{2} \\ 1 & \frac{1}{2} & 0 \end{bmatrix}.$$

Show that $\mathbf{M} = (1-m)\mathbf{A} + m\mathbf{S}$ (all $S_{ij} = 1/3$) is not diagonalizable for $0 \leq m < 1$.

EXERCISE 17. *How should the value of m be chosen? How does this choice affect the rankings and the computation time?*

REFERENCES

[1] A. BERMAN AND R. PLEMMONS, *Nonnegative Matrices in the Mathematical Sciences*, Academic Press, New York, 1979.

[2] M. W. BERRY AND M. BROWNE, *Understanding Search Engines: Mathematical Modeling and Text Retrieval, Second Edition*, SIAM, Philadelphia, 2005.

[3] M. BIANCHINI, M. GORI, AND F. SCARSELLI, *Inside PageRank*, ACM Trans. Internet Tech., 5 (2005), pp. 92–128.

[4] S. Brin and L. Page, *The anatomy of a large-scale hypertextual web search engine*, http://www-db.stanford.edu/~backrub/google.html (accessed August 1, 2005).
[5] A. Farahat, T. Lofaro, J. C. Miller, G. Rae, and L.A. Ward, *Authority Rankings from HITS, PageRank, and SALSA: Existence, Uniqueness, and Effect of Initialization*, SIAM J. Sci. Comput., 27 (2006), pp. 1181-1201.
[6] A. Hill, *Google Inside Out*, Maximum PC, April 2004, pp. 44-48.
[7] S. Kamvar, T. Haveliwala, and G. Golub, *Adaptive methods for the computation of PageRank*, Linear Algebra Appl., 386 (2004), pp. 51–65.
[8] A. N. Langville and C. D. Meyer, *A survey of eigenvector methods of web information retrieval*, SIAM Review, 47 (2005), pp. 135-161.
[9] A. N. Langville and C. D. Meyer, *Deeper inside PageRank*, Internet Math., 1 (2005), pp. 335–380.
[10] C. D. Meyer, *Matrix Analysis and Applied Linear Algebra*, SIAM, Philadelphia, 2000.
[11] Cleve Moler, *The world's largest matrix computation*, http://www.mathworks.com/company/newsletters/news_notes/clevescorner/oct02_cleve.html (accessed August 1, 2005).
[12] Mostafa, J., *Seeking better web searches*, Sci. Amer., 292 (2005), pp. 66-73.
[13] Sara Robinson, *The Ongoing search for efficient web search algorithms*, SIAM News, 37 (Nov 2004).
[14] Ian Rogers, *The Google Pagerank algorithm and how it works*, http://www.iprcom.com/papers/pagerank/ (accessed August 1, 2005).
[15] W. J. Stewart, *An Introduction to the Numerical Solution of Markov Chains*, Princeton University Press, Princeton, 1994.