GOOGLE'S PAGERANK

We will study the searching algorithm used by Google called **PageRank** to rank web pages. The name comes from both the term web page and co-founder of Google Larry Page. The goal is to explain why finding the most *important/relevant webpage* is same thing as finding an eigenvector of certain matrix.

1. Markov chain

We give a very brief introduction to a *Markov Chain* to mathematically model web surfing (i.e. navigating through webpages). In a very vague term, a Markov chain is a sequence of vectors that encode the probability to be in a certain state after finitely many steps. Suppose we are in a universe with 5 webpages. When we write the vector

$$x_2 = \begin{bmatrix} 0.1 \\ 0.4 \\ 0.25 \\ 0.25 \\ 0 \end{bmatrix} \tag{1}$$

above encodes the information that there is a 10% chance that a user will be in page 1 after two clicks. Likewise, there is a 25% chance that a user will be in page 4 after two clicks. Since there are only 5 webpages, you end up with 1 when you add all the probabilities to be in webpage k for $k = 1, \dots, 5$, i.e. all the entries. We give those vectors a special name and introduce a cousin matrix.

Definition 1.1. A **probability vector** is a vector with non-negative entries such that the sum of the entries is 1. A matrix whose columns are probability vectors is called a **stochastic matrix**.

Example 1.2. The vector x_2 in (1) above is a probability vector. The matrix

$$P = \begin{bmatrix} 0 & 0.7 & 0.2 \\ 0.2 & 0 & 0.8 \\ 0.8 & 0.3 & 0 \end{bmatrix}$$

is an example of a stochastic matrix. Note that the sum of the entries in a row does not have to be 1.

What is the role of a *stochastic matrix*? The (j, i)-th entry of the stochastic matrix is the probability of state i changing to state j. In our web surfing example, the (j, i)-th entry is the probability that a user currently in webpage i will move to webpage j.

Example 1.3. Let's consider a new universe with only 3 websites. Suppose a user begins web surfing at webpage 1. This initial condition can be expressed as a vector as

$$x_0 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$$

Suppose the probabilities of moving between pages are encoded in the matrix P above in Example 1.2. Then the ith entries of the vectors

$$Px_0 = \begin{bmatrix} 0 \\ 0.2 \\ 0.8 \end{bmatrix}, \quad P^2x_0 = \begin{bmatrix} 0.3 \\ 0.64 \\ 0.06 \end{bmatrix}, \quad P^3x_0 = \begin{bmatrix} 0.46 \\ 0.108 \\ 0.432 \end{bmatrix}, \quad \dots, \quad P^kx_0$$

tell us the probability of the user reaching webpage i after the k-clicks.

We can now mathematically formulate the above phenomena as following.

Definition 1.4. A **Markov chain** is a sequence of probability vectors $x_0, x_1, x_2, \ldots$ with a stochastic matrix P such that

$$x_1 = Px_0, \quad x_2 = Px_1, \quad \ldots, \quad x_{k+1} = Px_k \text{ for all } k \geq 0$$

In fact, the sequence is determined by x_0 and P, as $x_k = P^k x_0$ for all $k \geq 0$.

Recall that we want to determine the most *important webpage*. We try to solve this problem under the assumption that the webpage with the highest probabilty of reaching after *a very large number* of steps is the most important webpage. Borrowing the terminology from Calculus, we want to see whether the sequence $\{x_k\}_{k\geq 0}$ converges, or equivalently, the limit

$$\lim_{k\to\infty} x_k = \lim_{k\to\infty} P^k x_0$$

exists. If such limit converges, say q, then we have

$$Pq = P \lim_{k\to\infty} P^k x_0 = \lim_{k\to\infty} P^{k+1} x_0 = q.$$

Therefore, q is an eigenvector of P for eigenvalue 1. We call such vector a **steady-state vector** as the stochastic matrix P does not change q.

Details 1.5.

(a) We did not explain what it means for a sequence of vectors to converge. Roughly, $\{x_k\}_{k\geq 0}$ converges to a vector q if the the norm $||x_k - q||$ can be small as you want for big enough k. Alternatively, one can say that the limit exists entry-wise. If $x_k = \begin{bmatrix} x_{k1} \\ \vdots \\ x_{kn} \end{bmatrix}$ and $q = \begin{bmatrix} q_1 \\ \vdots \\ q_n \end{bmatrix}$, then

$$\lim_{k\to\infty} x_k = q \text{ if and only if } \lim_{k\to\infty} x_{kj} = q_j \text{ for all } 1 \leq j \leq n$$

(b) For any stochastic matrix P and a probability vector x, one can check that Px is again a probability vector. In particular, if the limit exists, q is a probability vector. (Here we used the fact that the limit of probability vector is again a probability vector.)

(c) In general, the limit q may be different as x_0 varies. What is amazing about the upcoming theorem is that if P is a *positive* matrix, q exists and is unique.

We should establish some basic facts about stochastic matrices.

Theorem 1.6. *Let P be a stochastic matrix. Then 1 is an eigenvalue for P.*

Proof. Let P^T be the transpose of P. Then the sum of the entries in every row is equal to 1. Therefore the vector $\begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}$ with all entries equal to 1 is a eigenvector of P^T with eigenvalue 1. Since P^T and P have the same eigenvalues, P has eigenvalue 1. To see why, observe that

$$\det(P - \lambda I) = \det((P - \lambda I)^T) = \det(P^T - \lambda I)$$

which tells us that the characteristic polynomial of P and P^T are the same. □

Theorem 1.7 (Perron-Frobenius Theorem)**.** *If A is a $n \times n$ positive stochastic matrix (i.e. all entries are positive), then it admits a unique steady state vector q which spans the 1-eigenspace. Furthermore, for any probability vector $x_0 \in \mathbb{R}^n$,*

$$\lim_{k\to\infty} P^k x_0 = q.$$

The proof of this fundamental theorem on Markov chain is too lengthy for this handout, so we will use the theorem without proof. By our assumption above, the *importance* of the webpage is given by the entry of the steady-state vector q of a Markov chain $\{x_k\}_{k\geq 0}$ and P. In fact, by Perron-Frobenius, only P determines q. We now explain how to construct P in the scenario of web surfing.

2. Random Walk on Directed Graphs

A **graph** is collection of points and edges. The chain is equally likely to move from vertex to vertex on the graph.