# 9 Operational Research

# 9.5 The Google PageRank Algorithm (5 units)

*This project requires an understanding of the Part IB Markov Chains course. Familiarity with Linear Algebra is desirable.*

## 1 Introduction

*PageRank* is a link analysis algorithm, operating on a database of documents connected to each other via directional *hyperlinks*. It was developed to measure the relative importance of a webpage in the World Wide Web, and with minor variations has also been employed in the context of assigning importance to academic journal publications.

**Graph-theoretic terminology** We will represent a collection of hyperlinked documents (webpages, academic journals, etc.) as a *directed graph* $G = (V, E)$, where $V$ is the set of documents $\{d_1, \ldots, d_N\}$ and the edge set $E \subseteq V \times V$ can be represented by an $N \times N$ *adjacency matrix* $A$, where $A_{ij} = 1$ iff $d_j \to d_i$ (i.e., iff $(d_j, d_i) \in E$). The *out-degree* of a node $i$ is the number of outgoing edges $d_i \to d_j$. A node with out-degree 0 is called a *dangling* node. Multiple edges can be incorporated by letting $A_{ij} = d$ when there are $d$ edges $j \to i$.

## 2 The PageRank algorithm

PageRank may be motivated as a *voting system*. Each webpage can distribute a total vote of 1 to other webpages, and votes themselves are weighted according to the importance of the respective voter, giving rise to the following recursion for the score $w_i$ of the $i$th webpage:

$$w_i = \sum_{j=1:N} S_{ij} w_j, \quad \text{where } S_{ij} = \frac{A_{ij}}{\sum_{q=1:N} A_{qj}} \quad \text{and} \quad w_i > 0. \tag{1}$$

A normalisation constraint $\sum_i w_i = N$ is also employed, to ensure an average score of 1. Moreover, we assume that "everyone votes," i.e., that there are no dangling nodes. We may interpret $S$ as the transition matrix of a Markov chain that describes the behaviour of a surfer who chooses where to go next by picking one of the available outgoing links at random. Recursion (1) then characterises the score vector $w$ as an invariant measure for this Markov chain.

> **Question 1** Produce an adjacency matrix for which recursion (1) fails to converge when initialised at $w = (1, 1, \ldots, 1)$ and iterated. Assume that everyone votes.

To avoid having to enforce assumptions on the edge structure of the document collection, we may assume that the surfer occasionally gets bored following links and starts anew, selecting a random webpage from $V$ to visit next according to some "default" distribution $\pi$ on $V$. This is often referred to as *damping*. If we handle dangling nodes in a similar way, we obtain the *random surfer* model of Figure 1.

 

**Random Surfer** $[(V, A), \pi, d]$

**At** $t = 0$**,** choose a random webpage from $V$ according to $\pi$.
**At** $t > 0$**, if there are no outgoing links,**
choose a random webpage from $V$ according to $\pi$;
**else**
with probability $d$
choose an outgoing link uniformly at random among available links,
with probability $(1 - d)$
choose a random webpage from $V$ according to $\pi$.

Figure 1: Description of the random surfer model for user behaviour.

**Question 2** Simulate 100 sample paths of the Markov chain of Figure 1 on the following example graph, with $\pi$ uniform and $d = 0.85$:

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

For the $j$th sample path, denote the average time spent on the $k$th node from the beginning of that sample path until time t by $\mu_{jt}^{(k)}$. For a fixed sample path of your choice, and for each node, plot $\mu_{jt}^{(k)}$ against $t$. For each node, and for each value of $t$, compute the variance of $\mu_{jt}^{(k)}$ over different sample paths and plot it against $t$.

**Question 3** Modify (1) to incorporate damping and handle dangling nodes as described above. Assuming that $1 > d > 0$, and $\pi_i > 0$ for all $i$, use standard Markov chain results to establish that the recursion you have obtained

- has a unique solution $p$, such that $p$ is a distribution (i.e., $\sum_i p_i = 1, p_i \geqslant 0$) and $p_i$ represents the average time the surfer spends visiting webpage $i$; and
- converges to $p$.

The *PageRank scores* are then given by $w = Np$.

**Question 4** Write a procedure that implements PageRank with $d = 0.85$ and $\pi$ uniform. Your procedure should take as input an adjacency matrix and a maximum number of iterations, and output a column vector of PageRank scores.

- Test your procedure on $A$ given in question 2, and compare with your results there.
- Construct an example for which node 1 has a larger number of both incoming and outgoing links than node 2, but a smaller PageRank score.

**Question 5** Write a procedure that generates a random adjacency matrix of size $N$, such that the out-degree of each node is an independent Poisson random variable with mean $k$, and conditional on the sequence of out-degrees all graphs are equiprobable.

- Generate an example with $N = 1000$ and $k = 100$, and convince yourself that your implementation of PageRank is correct by inspecting the eigenvectors of the modified transition matrix. Carefully explain your reasoning. You may use the MATLAB function *eig*.

 

File *citations.dat*:

| | |
|---|---|
| ... | ... |
| 9408099 | 9204102 |
| 9408099 | 9211097 |
| 9408099 | 9402002 |
| 9408099 | 9402005 |
| ... | ... |

File *articlejids.dat*:

| | |
|---|---|
| ... | ... |
| 9204102 | 82 |
| 9204103 | 62 |
| 9205001 | 65 |
| 9205002 | 65 |
| ... | ... |

Figure 2: A few entries from the two files forming the citation dataset.

- What happens to the empirical distribution of scores as $k$ decreases?
- Describe one aspect of this model that provides an unrealistic description of real-life web networks.

# 3 Ranking academic journals.

Let us represent a collection of academic journals as a directed graph, letting $A_{ij}$ be equal to the number of times an article published in journal $j$ cited an article from journal $i$. We force $A_{ii} = 0$, disregarding citations within the same journal. The *Eigenfactor* (EF) score of each journal is then computed by applying PageRank to the graph described, with $d = 0.85$ and the following choice of default distribution $\pi$ intended to represent journal size or *popularity*:

$$\pi_i = \frac{z_i}{\sum_i z_i}, \text{ where } z_i \text{ is the number of articles in journal } i \text{ in the given time period.} \quad (3)$$

Before the introduction of the EF score, the industry standard for ranking academic journals was the *Total Citations* (TC) score, which in this representation is the in-degree of a node. To separate journal prestige from journal size or popularity, the TC score is commonly reported as an Impact Factor (IF), obtained by dividing the in-degree of node $i$ by $z_i$. By analogy, the Article Influence (AI) score is obtained by dividing the EF score of a journal by $z_i$.

## 3.1 Real data

The files *citations.dat* and *articlejids.dat* on the CATAM website contain citation data from the Arxiv high energy physics theory section (also see Figure 2). Each article is represented by a 7 digit identifier, leading zeros being omitted without risk of confusion. Each line of the file *citations.dat* is of the form '[article $i$] [article $j$]', and represents a citation from article $i$ to article $j$. In *articlejids.dat*, each article is assigned a *journal identifier* ranging from 1 to 272.

> **Question 6** Verify that the same set of articles appears in both files, and that each article is assigned a *unique* journal identifier. Then retrieve the (multiple-edge) journal adjacency matrix $A$, and the vector $z$ of articles per journal. In MATLAB, files can be loaded using the function *load*. The following functions might also be useful: *isequal*, *ismember*, *find*, *unique*.

Following the introduction of EF alongside TC scores (and of AI alongside IF scores), a debate ensued as to the relative merits of the two approaches. Central to this debate is the statistical question of whether the two offer similar information.

 

**Question 7** On the basis of the citations dataset, discuss the question whether EF and TC scores are practically indistinguishable. Your answer should consider

- the correlation $\rho_{\mathrm{EF,TC}}$ between EF and TC scores and
- the differences in journal ranking for each of the two scores.

Is the correlation $\rho_{\mathrm{AI,IF}}$ between AI and IF scores relevant to this question? If so, how?

 