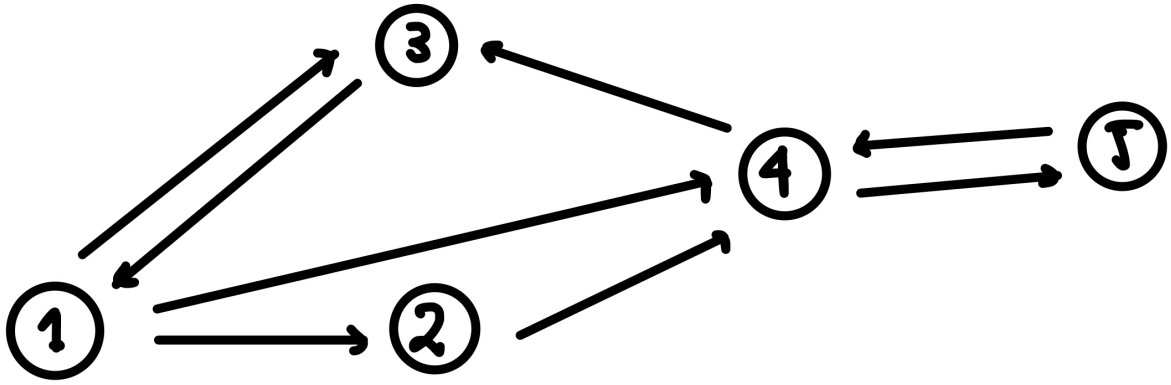
Applied Stochastic Processes
Assignment II

INTRODUCTION

PageRank is a ranking algorithm for graph databases. It is named after the application that sparked its creation: to rank web pages presented by the Google search engine.

A **directed graph** is defined by a set of **nodes** and **arrows**. In the original PageRank application, websites represent nodes. An arrow exists from a node (website) pointing to another node, if there is a link from the first website to the latter. Consider the following directed graph.



A directed graph can be viewed as a **state space diagram** of a **Markov chain.** Each **node** represents a **state** of the chain. Each **arrow** represents a **non zero one step transition probability**. For each state, **the arrows exiting the state are assumed to have equal weight/ probability**. For example, in the one step transition matrix P of the Markov chain associated to the diagram above we have $P_{12} = P_{13} = P_{14} = 1/3$ and $P_{11} = P_{15} = 0$.

The **PageRank algorithm** consists of **computing the limiting distribution of the Markov chain associated to a directed graph**. **The limiting probability of a state represents its ranking**.

The ranking of a page is hence boosted by the number and importance of incoming links. In practice, PageRank outputs a probability distribution that represents the likelihood of visiting any particular node by walking randomly around the graph.

More cutting edge applications of the PageRank algorithm include network optimisation (e.g. high ranking nodes in an energy distribution network can be critical in cascade failures), data lineage, cyber security,…

PART I - Theory Refresh

Consider a Markov chain (X_t) with state space $\chi = \{1,2,\ldots,n\}$ and transition matrix P. Define the following object in full generality.

1. The **invariant distribution** of (X_t)
2. The **limiting distribution** of (X_t)
3. The **asymptotic frequencies/distribution** of (X_t)

Finally, state the **Convergence Theorem** and explain how it links the distributions above.

PART II - Computation

Consider the Markov chain (X_t) associated with the **directed graph reported in the introduction**.

1. Describe (X_t) by providing its state space χ and transition matrix P. Does (X_t) meet the assumption of the Convergence Theorem?
2. Using R or your favourite programming language, compute the invariant/limiting/ asymptotic distribution of (X_t) according to the following three methods:
   i) By finding the suitable normalised left eigenvector of P (**invariant distribution method**)
   ii) By computing the distribution of X_t for t large enough and the initial condition $X_0 = 1$. Does the result change if $X_0 = 5$? (**limiting distribution method**)
   (*hint: to compute the distribution of X_t just apply P repeatedly to the initial distribution vector*)
   iii) By simulating the chain for a long enough time (e.g. $N = 1000$ or longer) and compute its asymptotic frequencies. For each asymptotic frequency also propose a suitable plot (cumulative average) to inspect empirically whether the estimated asymptotic frequency has converged. (**asymptotic distribution method**)
   (*hint: you can apply the command table() on your simulated chain to inspect its absolute frequencies, then divide by the chain length. Use sort() function for ranking*)
3. Report the final ranking obtained with the three methods. <u>It's important you report the code use for each of the method, too</u>.

(*hint: we did a similar exercise in the Limiting Behaviour Showcase script*)

PART III - More Computation

Load in your R Studio environment the matrix `Net` which is contained in the file **Net.Rdata**, in the resources folder for this assignment. The file is also available in csv format under the name **Net.csv**. The matrix `Net` represents a directed network made of 50 nodes numbered from 1 to 50. If $\text{Net}_{ij} = 1$ then there is an arrow pointing from i to j.

In the following, assume that the network is <u>too large for the invariant distribution method and the limiting distribution method to work in a convenient time.</u> This only happens with much larger matrices ($N = 10000$ might be too much!), but I did not want to be too annoying :). We focus therefore on the **asymptotic distribution method**. (*hint: Feel free to check your results with the other methods though!*)

1. Obtain the PageRank transition matrix P associated to the directed network described by `Net`. Make sure that the sum of the entries of each row of P sum up to 1 in order to describe a well defined Markov chain (X_t)
   (*Hint: a possible approach is to divide the matrix* `Net` *by its* `rowSums(Net)`)

2. Calculate the asymptotic frequencies of (X_t) first by setting $X_0 = 1$ and then by setting $X_0 = 50$. Compare and explain thoroughly the result. Does the Convergence Theorem apply to this (X_t)? (*Hint: to visualise a sparse matrix try the R command* `image(P)`).

For a reducible (X_t) with transition matrix P and state space $\{1,2,\ldots,n\}$ a **modified PageRank algorithm** can be proposed. For the same state space the **modified PageRank transition matrix** $P^\star$ is defined:

$$P^\star_{ij} = \alpha P_{ij} + (1-\alpha)\cdot\frac{1}{n},$$

for every $i, j \in \chi$ and for $\alpha \in (0,1)$, typically $\alpha = 0.95$.

Intuition: with probability α the web server moves according to P, with probability $1-\alpha$ they pick a page uniformly at random.

3. Implement the PageRank algorithm for a suitable choice of α. Find and report the ranking of the 50 nodes/pages. Is the the problem highlighted in previous point solved?
4. At least for a few states inspect empirically whether the estimated asymptotic probabilities have converged.