E2.1

E2

(2)

Exercise session 17/2 | Markov chains

Exercise 1.

If you think about it, web browsing is basically a Markov chain—the page w_{n+1} you will go to next depends on the page w_n you are currently at. Suppose our web server has three pages, and we have the following transition probabilities:

$$\begin{aligned} p_{1,1} &= 0, & p_{1,2} &= x, & p_{1,3} &= 1-x, \\ p_{2,1} &= y, & p_{2,2} &= 0, & p_{2,3} &= 1-y, \\ p_{3,1} &= 0, & p_{3,2} &= 1, & p_{3,3} &= 0. \end{aligned}$$

The transition probability $p_{i,j}$ represents $\Pr[W_{n+1} = j | W_n = i]$ for all $n \in \mathbb{N}$, which is the probability that I will next ask for page j, given that I am currently at page i. Assume that $0 < x < y < \frac{1}{2}$.

Web browsers cache pages so that they can be quickly retrieved later. We will assume that the cache has enough memory to store two pages. Whenever a request comes in for a page that is not cached, the browser will store that page in the cache, replacing the page least likely to be referenced next based on the current request. For example, if my cache contained pages $\{2, 3\}$ and I requested page 1, the cache would now store $\{1, 3\}$ (because $x < 1 - x$).

(a) Find the proportion of time that the cache contains the pages (i) $\{1, 2\}$, (ii) $\{2, 3\}$, and (iii) $\{1, 3\}$.

(b) Find the proportion of requests that are for cached pages.

Exercise 2.

Data centres alternate between two states: "working" and "down". There are many reasons why data centres can be down, but for the purpose of this problem we mention only two reasons: (i) a backhoe accidentally dug up some cable, or (ii) a software bug crashed the machines. Suppose that a data centre that is working today will be down tomorrow due to backhoe reasons with probability $\frac{1}{6}$, but will be down tomorrow due to a software bug with probability $\frac{1}{4}$. A data centre that is down today due to backhoe reasons will be up tomorrow with probability 1. A data centre that is down today due to a software bug will be up tomorrow with probability $\frac{3}{4}$.

(a) Draw a DTMC for this problem.

(b) Is your DTMC ergodic? Why or why not?

(c) What fraction of time is the data centre working?

(d) What is the expected number of days between backhoe failures?

Exercise 3.

The Google search engine is the most popular search engine on the web. Its popularity is, for a large part, due to the way the search engine orders its results. Different ranking algorithm exist, Google uses its top secret PageRank algorithm. While the (important) details of the PageRank algorithm are Google's most valuable company secret, the working principle of the algorithm is known.

In PageRank, the rank r_w of a web page w is calculated recursively: r_w is high if the sum of the ranks r_i of the web pages linking to w is high. This recursive calculation is written formally as

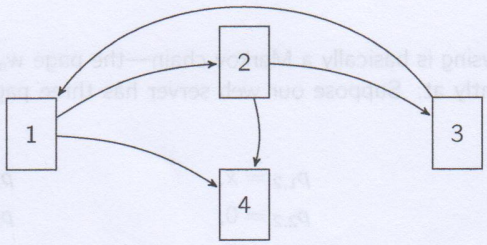
$$r(k+1) = r(k)P,$$

where $r(k)$ is the k-th iterate of the row vector of ranks and P is a matrix that formalises links to web pages. The recursive calculation is just equal to the evolution of a DTMC with probability matrix r.

The question remains of how to construct the initial probability vector $r(0)$ and the transition matrix P. Let N_i be the total number of outgoing links on page i. The (i, j)-th element of P, denoted by P_{ij} is set to be n_i/N_i if there are n_i outgoing links from page i to j, and zero if there is no link from i to j. The transition matrix P constructed this way corresponds to a random walk over the world wide web. Note that if the transition matrix P is ergodic, then we know that r will converge to the stationary distribution π, regardless of the initial distribution $r(0)$.

→ limiting distr. will exist and equals stat. distr.

Consider the following web, which consists of 4 web pages:



(a) Construct the transition matrix P for this web.

(b) Classify the states of the corresponding Markov chain. Is the chain irreducible?

(c) What is the stationary distribution? What is the limiting distribution?

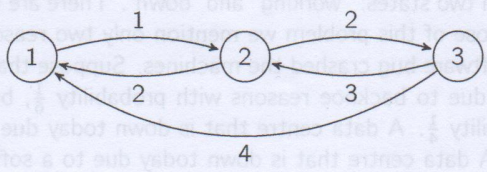
(d) Add teleportation: if a page has no outbound links, then we 'add' links such that someone who clicks links at random ends up at a different page.

(e) Same as in (b), but then with teleportation.

(f) Same as in (c), but then with teleportation.

Exercise 4.

Consider the following continuous-time Markov chain (CTMC):



Convert this chain to a discrete-time Markov chain (DTMC).

Exercise 5.

Secure Storage, a low-budget data storage provider, uses *Crappydisk* hard drives. The time until failure of a *Crappydisk* hard drive can be as an exponentially distributed random variable with rate $\lambda \in \mathbb{R}_{>0}$. In order to cope with this hard drive failure, *Secure Storage* uses stacks of $N \in \mathbb{N}$ *Crappydisk* hard drives, retaining multiple copies of the data in a RAID fashion. When a hard drive fails, it is replaced and the data is copied to this new drive from a different, still functioning hard drive. The time it takes to copy the data to a new disk is exponentially distributed with rate $\mu \in \mathbb{R}_{>0}$.

(a) Create a continuous-time Markov chain that models *Secure Storage*'s storage solution.

(b) What is the limiting probability mass function of the continuous-time Markov chain you constructed in (a).

(c) Assume that the *Secure Storage* system has been running for a very long time, but that the system could not fail yet. Calculate the failure rate of the storage solution in terms of the number of disks N.