Exercise Sheet 5

Due date: Monday, June 29 until 15:00

- Please upload your solutions to Moodle.
- Hand in your solutions in groups of **two to three students**.
- Please hand in the solutions of your group as a single PDF file.
- Mark your answers clearly.
- The solutions of this exercise will be discussed live via Zoom on **Friday, July 3 at 12:30**.

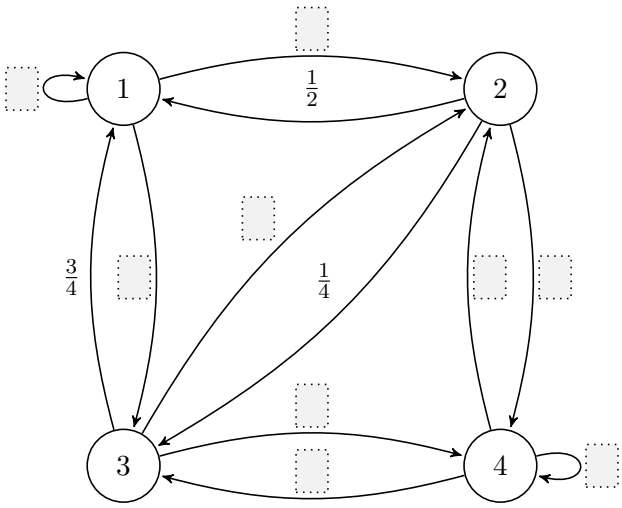
Exercise 1 (Symmetric Markov Chains) — 4 points

A Markov chain $\mathcal{Q}$ is called *symmetric* if its transition matrix Q is symmetric.

Show that there exists a unique probability vector $\boldsymbol{\pi} \in \mathbb{R}^{1\times n}$ such that $\boldsymbol{\pi}$ is the stationary distribution of $\mathcal{Q}$ for all connected, symmetric Markov chains $\mathcal{Q}$ with state space $[n]$.

Exercise 2 (Completing Markov Chains) — 5 points

Consider the following incomplete graphical representation of a Markov chain with 4 states. Therein, all missing edges have probability 0.



Fill in the gaps in the graphical representation such that the stationary distribution of the Markov chain becomes $\boldsymbol{\pi} = (\pi_1, \pi_2, \pi_3, \pi_4) = \left(\frac{1}{2}, \frac{1}{12}, \frac{1}{4}, \frac{1}{6}\right)$ where π_i denotes the probability of being in state i. Justify your solution. Discuss whether the solution is unique.

Exercise 3 (Metropolis-Hastings Sampling of Independent Sets) 1+2+3=6 points

A set $I \subseteq V$ of nodes in a graph $G = (V, E)$ is called *independent*, if $(v, w) \notin E$ for all $v, w \in I$. Denote the independent sets of G by $\mathcal{I}_G \subseteq 2^{V(G)}$.

Fix some graph G. In this exercise, we want to uniformly sample from $\mathbb{U} \coloneqq \mathcal{I}_G$ using Metropolis-Hastings sampling. Recall that the Metropolis-Hastings algorithm performs a random walk on an undirected, connected graph. To avoid confusion with G, we call this graph $\mathcal{H} = \big(V(\mathcal{H}), E(\mathcal{H})\big)$. (It is called $\mathcal{G}$ in the lecture.)

a) Define $V(\mathcal{H})$ and $E(\mathcal{H})$ for sampling from $\mathcal{I}_G$.

Hint: The edge relation $E(\mathcal{H})$ describes which elements of $\mathcal{I}_G$ are "neighbours". Think of a useful notion of "being a neighbour" for sets of nodes of G.

b) Find the maximum degree Δ of $\mathcal{H}$ and show that $\mathcal{H}$ is connected.

c) Define transition probabilities $q_{I,J}$ for $I, J \in V(\mathcal{H})$. Show that the stationary distribution of the resulting Markov chain is the uniform distribution over $\mathcal{I}_G$.

Hint: Use slide 6.35 of the lecture and recall that $\mathcal{D} = Z\mathcal{P}$ where $\mathcal{P}$ is the target distribution and Z is some constant. For our exercise, we know that $\mathcal{P}$ is the uniform distribution on $\mathcal{I}_G$.

Exercise 4 (Gibbs Sampling) 4 points

Prove Theorem 6.13 from the lecture (slide 6.38). That is, show that the intended probability distribution $\mathcal{P}$ is the unique stationary distribution of the Markov chain defined on slide 6.38.

Hint: Use Lemma 6.4 (slide 6.11).

Exercise 5 (Node Failures in Computing Clusters) **5 points**

We want to process a job consisting of n tasks on a computing cluster. The individual tasks can be run in parallel and each of the tasks takes time t to complete. We make the following assumptions regarding cluster node failures:

- Any node failure initiates a recovery procedure on the node taking $10t$ to complete.
- Node failures may also happen during the recovery itself.
- Once a recovery is completed, the node continues exactly where it was interrupted.
- The probability of node failure during the execution of a task is $p_f \in [0, 1)$. The probability of node failure during recovery is also p_f.

Let R denote the *total accumulated execution time* R over all cluster nodes. For example, if no node failures occur, then $R = nt$. If exactly two nodes failures occur, $R = nt+10t+10t$. This is independent of the nodes that failed and when they failed.

Calculate the expected total accumulated runtime $\mathrm{E}(R)$ of the job.

Hint: If needed, you may use that

$$\sum_{i=1}^{\infty} i \cdot x^{i-1} = \tfrac{1}{(1-x)^2} \qquad \text{for all } x \in \mathbb{R} \text{ with } |x| < 1. \qquad (*)$$

Exercise 6 (Facebook Friends) **3+3=6 points**

This exercise is about the friendship relation in social networks such as Facebook. We assume friendship is a symmetric relation: if A is a friend of B, then B is a friend of A. We want to implement some of the related features using Map-Reduce.
The tuples which are the input to the first MAP function are of the form

$$(\texttt{PersonA}, \{\texttt{PersonB}, \texttt{PersonC}, \texttt{PersonD}, \ldots\})$$

where the value part consists of the friends of the person stored in the key.

Implement the following features using Map-Reduce by specifing the MAP and REDUCE functions. You may do so in the style of (for example) slide 7.34.

a) *Listing mutual friends.* When you visit a profile page on Facebook, you are shown a list of persons that are both friends of you, and the person whose profile you are looking at.

Specify a *single round* Map-Reduce algorithm that computes the mutual friends for all pairs of users. The key of the output should be a pair of persons, and the values part should contain the mutual friends of this pair of persons.

b) *Suggesting new friends.* For this exercise, we assume that the suggestions are taken randomly from the friends of your friends. To add some variance in the suggestions, only one person per friend is suggested and there are only 10 suggestions per run of the algorithm.

Specify a *single round* Map-Reduce algorithm that computes 10 suggested friends for every user. The key of the output should be a person, and the values part should contain 10 suggested friends.

(The algorithm is allowed to output duplicates when a suggested friend shares more than one common friend. You may assume that everybody has at least 10 friends).