Advanced Algorithms COMP4121 2021
Practice Problems

Here are a few problems of the kind you will be asked to solve on the final exam.

1. You are watching traffic on a busy road and you notice that on average three out of every four trucks on the road are followed by a car, while only one out of every five cars is followed by a truck. What fraction of vehicles on the road are trucks?

   Solution: Let the fraction of cars be x and fraction of trucks y; then the probability to see a car is x and the probability to see a truck is y. These probabilities must satisfy

   $$(x, y)\begin{pmatrix} 4/5 & 1/5 \\ 3/4 & 1/4 \end{pmatrix} = (x, y)$$
   $$x + y = 1$$

   One of the two homogeneous equations is redundant, so we get a solution by solving

   $$4/5x + 3/4y = x$$
   $$x + y = 1$$

   Solution: $x = 15/19$ and $y = 4/19$.

2. How many balls in $\mathbb{R}^{1000}$ of radius $1/2$ does it take so that these balls together have the same volume as a unit ball in this space?

Answer: 107150860718626732094842504906 000181056140481170553360744375038837035105112493612249319837881569585812759467291755314682518714528569231404359845775746985748039345677748242309854210746050623711418779541821530464749835819412673987675591655439460770629145711964776865421676 604298316526243868372056680693 76

We use the fact that $V_{1/2}(1000) = (1/2)^{1000}V(1000)$;

Thus, $V(1000) = 2^{1000}V_{1/2}(1000)$.

3. (A tiny bit harder) Let G be a strongly connected directed graph, i.e., a graph such that for any two vertices u, v there exists a path from u to v. A vertex w in G is periodic if there exists an integer k such that every loop containing w has length divisible by k. Show that if one vertex of a strongly connected graph is periodic, then all vertices must also be periodic.

   Solution: Assume v is periodic and let w be an arbitrary vertex in G. Then there is a path from v to w and a path from w to v. These two paths together form a loop containing v. Thus, the length of this loop must be divisible by k. Consider now an arbitrary loop l containing w. Then the path consisting of the path from v to w, concatenated with the loop l, concatenated with the path from w to v also forms a loop containing v and its length is thus divisible by k. However, since the sum of the lengths of the paths from v to w and from w to v is divisible by k also the length of the loop l must itself be divisible by k.

4. (Basic probability and the Markov inequality) You are presented with n fair coins, $n \geq 1$; you flip each and if it comes tail you get the coin, if it comes head you do not get it.

   (a) Use the Markov inequality to show that the probability to win at least $3/4$ of all coins is not more than $2/3$.

   (b) Show that the probability to win at most $1/4$ of all coins is also not more than $2/3$.

   (c) Find a formula for finding the exact probability to win at least k, ($0 \leq k \leq n$) out of n available coins and plot its values for all $0 \leq k \leq n$ and $n = 100$. (Your formula can contain summations without a closed form; you can use any software you like for computing and plotting.)

   (d) Was Markov inequality useful in this problem?

5. Consider the deterministic Linear Time Algorithm for Order Statistic.

   (a) Assume we split the numbers in groups of seven elements. Derive the asymptotic run time of the algorithm, following closely what we did in case when we split the input numbers into groups of five.

       Hint: You have to figure out what the recursion should look like and for what K you can derive $T(n) < K \cdot C \cdot n$. In case of splitting into groups of five $K = 11$ worked. If you split into groups of seven you will need a different K. Try setting K a variable and see what you get if you substitute this into the right side of the recurrence; it should be easy to figure out what K works.

   (b) Assume we split the numbers in groups of three elements. Explain why the proof breaks down.

       Hint: Show that no K works.

6. Modify the SkipList data structure so that finding the k^{th} smallest element can also be done in expected time $O(\log n)$.

7. Consider Karger's MinCut algorithm. How many repetitions of the 4-Contract(G) would you need to make to guarantee that MinCut is found with a probability of $1 - \frac{1}{n^2}$ where n is the number of vertices of G. How does the asymptotic run time of the whole algorithm change?

8. The present day "publish or perish" madness in academia involves counting number of papers researchers have published, as well as the number of citations their papers got.

   (a) One might argue that not all citations are equally valuable: a citation in a paper that is itself often cited is more valuable than a citation in a paper that no one cites. Design a PageRank style algorithm which would rank papers according to their "importance", and then use such an algorithm to rank researchers by their "importance".

   (b) Assume now that you do not have information on the citations in each published paper, but instead you have for every researcher a list of other researchers who have cited him and how many times they cited him. Design again a PageRank style algorithm which would rank researchers by their importance.

9. You are watching traffic on a busy road and you notice that on average three out of every four trucks on the road are followed by a car, while only one out of every five cars is followed by a truck. What fraction of vehicles on the road are trucks?

10. You are monitoring Sydney summer weather and you noticed that on average seven out of every 10 pleasant weather days are followed by another pleasant wether day and only one out of every 10 pleasant weather days is followed by a day with a summer shower. You also noticed that on average 2 out of every 15 days with a summer shower are followed by another day with a summer shower and that 4 out of every 9 very hot days are followed by a day with a pleasant wether. You also know that on average out of 90 summer days 50 have pleasant wether, 10 have summer showers and 30 are very hot. Set up the equations which can be used to determine what fraction of days with a shower are followed by a very hot day. Note that you are not asked to solve such a system of equations, just to set up all the equations needed to solve the problem.

11. (a) Recall that a matrix consisting of non-negative reals is *row-stochastic* if in each row all the entries in that row sum up to 1; thus if

$$\mathbf{M} = \begin{pmatrix} s_{11} & s_{12} & s_{13} & \dots & s_{1n} \\ s_{21} & s_{22} & s_{23} & \dots & s_{2n} \\ s_{31} & s_{32} & s_{33} & \dots & s_{3n} \\ \dots & \dots & \dots & & \\ s_{n1} & s_{n2} & s_{n3} & \dots & s_{nn} \end{pmatrix}$$

then for all $1 \leq i \leq n$ we have $\sum_{j=1}^{n} s_{ij} = 1$. Assume that $x^\top = (x_1, x_2, \dots, x_n)$ is a vector such that $x_i \geq 0$ and $\sum_{i=1}^{n} x_i = 1$. Show that then vector $y = x^\top M$ also has the property that $y_i \geq 0$ and that $\sum_{i=1}^{n} y_i = 1$. Thus, if x is a vector of probabilities of states of a Markov chain, then so is vector y.

(b) In computing the Google PageRank iteratively we start with vector $(1/N, 1/N, \dots, 1/N)^\top$ giving each page the same initial rank of $1/N$. Prove that after the iteration has stopped, the sum of the PageRanks of all web pages will be equal to 1.

12. Below is given the graph of the Internet one millisecond after the Big Bang.

    (a) Construct the corresponding Google matrix with $\alpha = 7/8$.

    (b) Explain what property the PageRank satisfies (i.e., how it is related to the corresponding Google matrix).

    (c) Find the PageRank of all nodes (you do not need an iterative algorithm to find the PageRank for such a small matrix; you can solve the corresponding system of linear equations directly, keeping in mind that the PageRanks of all pages can be interpreted as probabilities and thus must sum up to 1)

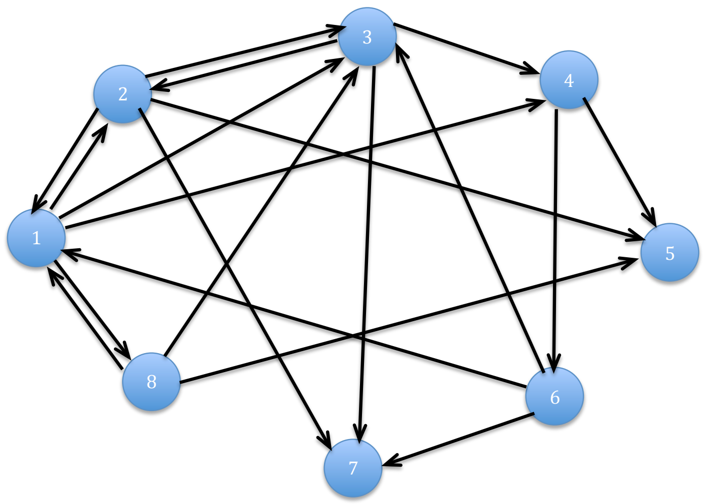

Figure 1: The Internet 1ms after the Big Bang

13. (This is a somewhat tricky one!) Prove that if in an irreducible Markov Chain one state is aperiodic, then all states must be aperiodic. Thus, equivalently, assume you have a directed strongly connected graph G which has one vertex v which satisfies that there is NO number K such that the length of every loop containing v is divisible by K. Show that in this case all vertices have the same property.

14. One of the most common hash functions used is defined as $h(x) = x \mod p$ where p is a prime. This function has a drawback that it is not local, in the

sense that two numbers can be close to each other but their hash values can be far away, for example if $p = 31$ and $x_1 = 30$ and $x_2 = 31$ then $h(x_1) = 30 \mod 31 = 30$ but $h(31) = 31 \mod 31 = 0$. Can you slightly alter this hash function so that numbers which are close to each other have hash values that are also close to each other?

15. Assume that you are using a spherical Gaussian of dimension 10,000 with a unit variance in all directions to produce 3 random vectors, $x, y, z \in \mathbb{R}^{10,000}$. Compute the expected circumference of the triangle whose vertices are given by these 3 vectors. Explain your answer.

16. Modify the deterministic algorithm for order statistic by splitting the set of input numbers into groups of 9 elements instead of 5 and derive its run time estimate.

17. Extend the (doubly linked) SkipList data structure so that finding a median of all elements present in the SkipList runs in constant time.

18. Modify the Karger MinCut algorithm so that it produces the correct value of the minimal cut with probability of at least $1 - \frac{1}{n^{\log n}}$ and which runs in time $O(n^2(\log n)^4)$.

19. Consider the following hash function: Let m be a prime; choose r such that the size of the universe V of all possible keys satisfies $|V| < m^{2(r+1)}$. Randomly and independently choose TWO vectors from $\{0, 1, 2, \ldots, m-1\}^{r+1}$, say they are $\vec{a} = (a_0, a_1, a_2, \ldots, a_r)$ and $\vec{b} = (b_0, b_1, b_2, \ldots, b_n)$. As before, let $\vec{x} = (x_0, x_1, \ldots, x_r)$ be the digits of the representation of x in basis m, i.e., such that $x = x_0 + x_1 m + \ldots + x_r^r$. Define a hash function $h_{a,b}(x)$ mapping V into $\{1, 2, \ldots, m-1\} \times \{1, 2, \ldots, m-1\}$ by

$$\begin{aligned} h_{a,b}(x) &= \left(\sum_{0 \le i \le r} x_i a_i \mod m, \quad \sum_{0 \le i \le r} x_i b_i \mod m \right) \\ &= (\langle \vec{x}, \vec{a} \rangle \mod m, \ \langle \vec{x}, \vec{b} \rangle \mod m) \end{aligned}$$

Show that such a family of functions is universal, i.e., that any two keys collide with probability of $1/m^2$. Note that in this case the size of the hash table is $m \times m$ i.e., with m^2 many slots.

20. Assume you generate two random vectors $\vec{a}, \vec{b}$ in a d-dimensional vector space, where d is large, by choosing their coordinates independently using a Gaussian

of zero mean and a unit variance. Consider vectors $\vec{d} = \vec{a} - \vec{b}$ and $\vec{s} = \vec{a} + \vec{b}$. Find the expected values of the lengths $|\vec{d}|$ and $|\vec{s}|$ of these two vectors as well as the expected value of the angle between $\vec{d}$ and $\vec{s}$.

21. Describe the Google matrix produced by taking into account for each webpage and each outgoing link on that webpage the number of times such a link has been clicked on during a (long) period of time. (provided to Google by user's Chrome browsers). So we no longer assume that every link on a webpage is equally likely to be followed as every other link on the same webpage.

22. You are sitting in a veterinary clinic and want to figure out the ratio of dog owners versus cat owners in your neighbourhood, but you do not want to count the large number of dog and cat users visiting the clinic with their pets on that day. Instead you notice that on average every 3 out of 4 dog owners are followed by another dog owner, and only one out of every 3 cat owners are followed by another cat owner. Use such data to estimate the ratio of dog owners to cat owners.