Final exam for CS-E4600

Wed, Feb 14, 2018
09.00 – 12.00

This is a **closed book** exam.

There are **4** problems. Each problem receives an equal number of points.

To get full points you should answer **all 4** problems.

Problem 1

Consider the *Jaccard coefficient* $J(x,y)$ between sets x and y,

$$J(x,y) = \frac{|x \cap y|}{|x \cup y|}.$$

Define the *Jaccard distance* as

$$d_J(x,y) = 1 - J(x,y).$$

Is d_J a metric function? Prove or disprove.

Problem 2

Question 2.1 Consider the nearest-neighbor problem: We are given a set of objects X and a distance function d. At query time we are given an object q and the goal is to find a point $x^* \in X$ that such that

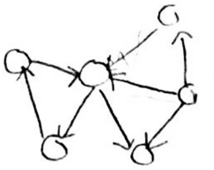
$$d(q,x^*) \leq d(q,x), \text{ for all } x \in X.$$

The *linear-scan* algorithm has complexity $\mathcal{O}(nD)$, where n is the number of objects in X and D is the time required for one distance computation.

In certain cases computing the distance function d is an expensive operation. In those cases it is desirable to have a *lower bound* on the distance d. A lower bound is a function d_L with the property $d(x,y) \geq d_L(x,y)$ for all $x,y \in X$. A lower bound d_L is useful when it is much faster to compute than the function d.

Explain how a lower bound distance function can be used to speed up the linear-scan algorithm.

Question 2.2 Consider the *edit distance* for string comparison: Given two strings x and y the edit distance $d(x,y)$ between x and y is the minimum number of *character operations* (character additions or deletions) needed to transform x to y.



(a) Is edit distance a metric? Prove or disprove your claim.

(b) Devise an algorithm to compute exactly the edit distance between two strings x and y. What is the running time of your algorithm?

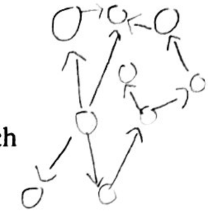
(c) Consider a dataset of strings $\mathcal{D}$, and a query string s. Consider the linear scan algorithm (as in Question 2.1) for finding the nearest neighbor of s in $\mathcal{D}$ with respect to the string edit distance.

Provide run-time analysis for this problem.

(*Hint*: you should introduce variables expressing the size of the input of the problem. The running time of the algorithm should then be expressed using those variables.)

(d) Provide a lower bound for the string edit distance. What is the running time for computing this lower bound?

aaa bbb



Problem 3

We mentioned in class that citation networks (networks representing citations between research articles) are, in theory, *directed and acyclic graphs* (dags).

Explain why we expect citation networks to be directed and acyclic.

However, in practice, we expect that citation networks have a small number of cycles. Why?

Assume that we want to test whether a given citation network has cycles. Propose an algorithm to detect if a directed graph has a cycle. What is the running time of your algorithm?

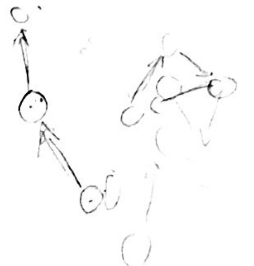
Problem 4

Consider a data stream $X = (x_1, x_2, ..., x_m)$, where each element x_j is a number between 1 and n. The number of occurrences of number i in the stream is $m_i = |\{j : x_j = i\}|$. We are interested in estimating m_i for each number i between 1 and n.

(a) Describe how to compute m_i, for $i = 1, \ldots, n$, with one pass over the stream and available memory $\mathcal{O}(n)$ integers.

(b) Consider the following sketching algorithm, where s is a hash function that maps each number of $[1..n]$ to $\{+1, -1\}$, uniformly at random:

Algorithm FREQSKETCH
$c \leftarrow 0$
for $j \leftarrow 1, \ldots, m$
    $c \leftarrow c + s[x_j]$
return c



DFS (NODE)
MARK NODE AS VISITED
FOR EDGE IN NODE
IF EDGE

Show that $E[c \cdot s[i]] = m_i$, for all $i = 1, \ldots, n$.

(c) Explain how to use the idea of (b) to design a data-stream algorithm for computing m_i, for all $i = 1, \ldots, n$, with one pass over the stream and with memory smaller than $\mathcal{O}(n)$.