Final exam for T.61-5060

December 16, 2014

This is an **open book** exam. It is allowed to use any textbook, printed material, or personal notes brought in the room. Using any electronic device is **not** allowed.

There are **five** problems. A yes answer to the first problem receives 20 pts. Each of the other problems receives 25 points. The last problem is open-ended. There is not only one correct answer. Any logically sound answer will be accepted.

The course instructor will visit the exam room to answer clarifying questions at around 10am.

Problem 1

Did you fill the course evaluation form?

Problem 2

Consider two graphs $G_1 = (V, E_1)$ and $G_2 = (V, E_2)$ over the same set of nodes V but different set of edges E_1 and E_2. We define the *symmetric-difference distance* $\Delta(G_1, G_2)$ between G_1 and G_2 to be the number of edges that are different in the two graphs, that is,

$$\Delta(G_1, G_2) = |E_1 \setminus E_2| + |E_2 \setminus E_1|.$$

Question 2.1. Is the distance Δ a metric?

Question 2.2. Given a set of graphs $\mathcal{G} = \{G_1, \ldots, G_k\}$ all of which are over the same set of nodes, design an algorithm for finding the *median graph* of the set $\mathcal{G}$ according to the distance function Δ. In other words, we want to find a *new* graph G (over the same set of nodes as the input graphs $G_1, \ldots, G_k$) that minimizes the sum of distances

$$\sum_{i=1}^{k} \Delta(G, G_i).$$

Problem 3

Consider a set of n objects $X = \{x_1, \ldots, x_n\}$, and a distance function $d : X \times X \to \mathbb{R}$ defined over the objects of X. We are interested in defining a distance function between *subsets* of objects of X. Given two sets $A, B \subseteq X$, the *Hausdorff distance* $d_H(A,B)$ between A and B is defined as a function $d_H : 2^X \times 2^X \to \mathbb{R}$ with

$$d_H(A,B) = \max\left\{\max_{x \in A} D(x,B), \max_{y \in B} D(y,A)\right\},$$

where the distance function $D : X \times 2^X \to \mathbb{R}$ is defined as

$$D(u,C) = \min_{v \in C} d(u,v).$$

Question 3.1. Discuss the intuition behind the definition of the Hausdorff distance. What is the role of the outer max?

Question 3.2. We now want to embed the Hausdorff distance d_H into L_∞^n, the L_∞ norm in the $\mathbb{R}^n$. For each set $A \subseteq X$ we define the mapping

$$f(A) = (D(x_1,A), \ldots, D(x_n,A)).$$

(This is a mapping of the set of objects A to an n-dimensional vector.)

Prove that the mapping f is a *distance-preserving embedding* of d_H to L_∞^n.

Recall that an distance-preserving embedding of a metric space (U,d) to a metric space (U',d') is a mapping $f : U \to U'$ that satisfies $d(z,w) = d'(f(z), f(w))$, for all $z, w \in Z$.

Problem 4

In class we saw how to define a *locality-sensitive hashing* scheme for the Jaccard coefficient.

As a reminder, recall that given a ground set of objects X, the *Jaccard coefficient* $J(A,B)$ for any sets $A, B \subseteq X$ is defined as

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}.$$

Then, given a random permutation $\pi : X \to X$ we defined the *min-wise hash* of a set $A \subseteq X$ as

$$h(A) = \min_{x \in A} \pi(x).$$

To show that this hashing scheme is locality sensitive, we need to show that for any sets $A, B \subseteq X$ it holds

$$\Pr[h(A) = h(B)] = J(A,B), \tag{1}$$

where the probability is taken uniformly over the space of all possible permutations.

Prove Equation (1).

Problem 5

Consider the PageRank algorithm that we studied in class. Given the web graph $G = (W, E)$, where W is the set of all web-pages and E is the set of hyper-links between web-pages, the PageRank algorithm defines a score $PR(w)$, for each $w \in W$, which captures the *importance* of page w.

One nice extension of PageRank is the *personalized PageRank*. This works as follows. Assume that a user u favors a set of pages $U \subseteq W$. Then, the personalized PageRank $PR(w \mid U)$ of a page w for the user u is defined using the same random-surfer model as PageRank, but the random surfer now makes her random restarts only to pages in U, instead of all pages in W. The set U can be defined in many ways, for instance, the set of pages that the user visits more often.

Question 5.1. Discuss the intuition of the definition of personalized PageRank.

Question 5.2. Google would like to compute the personalized PageRank of all users, in order to provide a personalized service to each one of them. However, this plan not feasible: computing 7 billion PageRank vectors is too costly and it would take too much space to store.

Propose some alternative to Google so that they can provide personalized service to their users (based on the concept of personalized PageRank) while the computation/storage would be feasible.