Algorithms for Modern Data Models, Spr 2013-14

Take home exam. Handed 6/4/2014.

The exam is due before midnight, 6/9/2014. No collaboration allowed, but the exam is otherwise open book and open Internet. Please email your answers (scanned docuemnts are ok, but smartphone pictures are not) to the two instructors. Each question is worth 10 points.

1. In class we saw how to compute highly similar pairs of m-dimensional vectors x, y (with entries from $\{0,1\}$) via sampling in the mappers, where the similarity was defined by cosine similarity: $\frac{x^T y}{|x|_2 |y|_2}$. Show how to modify the sampling scheme to work with dice similarity, defined as

$$\text{dice}(x,y) = \frac{2x^T y}{|x|_2^2 + |y|_2^2}$$

   (a) Prove shuffle size is still independent of m, the dimension of x and y.

   (b) Assuming combiners are used with B mapper machines, analyze the shuffle size.

2. The traditional secretary problem assumes that the number of candidates grows to infinity. Under this assumption, the optimal stopping rule to maximize the probability of finding the best candidate is to say 'no' to the first n/e candidates, and then take the next candidate that is better than the first n/e. The probability of finding the best candidate under this rule is $1/e$.

   In the regime where n is small however, the best proportion may not be $1/e$. Under the small n case of $n = 4$, to maximize the probability of finding the best candidate, how many candidates should be interviewed, and what is the probability of finding the best one?

3. Define a simple LSH family for solving the (c, R) near-neighbor problem where each object that you are hashing is a set of elements, and the distance between two sets A and B is given by

$$d(A,B) = \frac{|(A-B) \cup (B-A)|}{|A \cup B|}.$$

   Assume that $cR < 1$. Analyze the performance of the LSH family by giving bounds on the quantity

$$\frac{\log(1/p_1)}{\log(1/p_2)}$$

   where p_1, p_2 are as described in class.

4. Design a sketch s defined over sets that can be incrementally computed, is small, and has the property that given the sketches of two sets, you can compute the size of the symmetric difference of the sets efficiently and with small error. Analyze the performance of your sketch in terms of size, accuracy, the time to compose the sketches of two disjoint sets into a sketch of their union, and the time to estimate the symmetric difference (also, describe the process for estimation). Assume that during your incremental computation, no element is presented twice.