CS 493: Algorithms for Massive Data Sets
Homework 4
Due: Tue, May 14

1. *(20 points)*
   Consider the algorithm for finding approximate nearest neighbors for binary vectors in $\{0,1\}^d$ using multiple sorted lists; each list is obtained by randomly permuting the d coordinates and sorting lexicographically. Recall that the algorithm discussed in class located the position of the query vector in each of the sorted lists and searched the vectors close to the query vector, in decreasing order of the length of the common prefix. Suppose instead, we simply examine the vectors immediately before and after the query vector in each of the sorted lists.

   (a) Explain why the analysis of the previous algorithm does not apply to this simpler variant.

   (b) Construct an example to demonstrate that the number of sorted lists you need in order to guarantee that you find an approximate nearest neighbor is high compared to the number of lists required for the algorithm in class.

2. *(20 points)*
   Give a randomized algorithm that maps R^n to R^d (d should be chosen appropriately), such that for any two points $x, y \in R^n$, the Euclidean distance between them is distorted by a factor of at most $1+\epsilon$, with probability $1-\delta$. Given an upper bound on d in terms of n, ϵ and δ.

3. *(20 points)*
   This question requires you to analyze the stationary distribution for different random walks on the web graph. Recall that the stationary distribution is a distribution on pages, such that if we draw a page from this distribution and perform one step of the random walk, the resulting distribution on pages is identical to the stationary distribution. Under certain conditions, when you perform a random walk for sufficiently many steps, the resulting distribution on pages will converge to the stationary distribution.

   (a) Suppose we perform a random walk by starting from a fixed page and repeatedly following a random outlink from the current page. Construct a graph (i.e.a set of n pages and links between them) such that the the probability of being at one of the pages in the stationary distribution is exponentially small in n.

   (b) Consider the random walk in part 3a. Construct a graph where the distribution on pages after performing t steps of the random walk does not converge to a limit.

   (c) Now consider the PageRank random walk (i.e. we jump to a random page with probability α and follow a random outlink with probability $1-\alpha$. Is it possible to have a graph where the situation discussed in parts 3a or 3b occur ? Construct such examples or explain (without proof) why this is not possible. You can assume that α is a constant.

4. *(20 points)*
Consider the SALSA random walk. For hub scores, a single step of the random walk consists of following a random outlink and then a random inlink from the resulting page. For authority scores, a single step of the random walk consists of following a random inlink and then a random outlink from the resulting page. The hub scores and authority scores are the stationary distribution for the corresponding random walk. Prove that the hub score for a page is proportional to the number of inlinks to the page and the authority score is proprtional to the number of outlinks. Do not look at the SALSA paper for the proof of this.