CPTS 415 Big Data

Assignment 2

Instructor: Yinghui Wu

1. [**Join operators**] (45) This sets of questions test the understanding of basic database search operators. Consider a join $\bowtie_{R.A=S.B}$ S. We ignore the cost of output the result, and measure the cost with the number of I/O. Given the information about relations to be joined below.

   Relation S contains 20K tuples and has 10 tuples per page. Relation R contains
   100K tuples and has 10 tuples per page.
   Attribute B of S is the primary key of S. In total 52 Buffer pages are available in memory.
   Assume neither relation has any index.

   a. (15) Describe a block nested join algorithm. Give the cost of joining R and S with a block nested loops join.

   b. (15) Describe a sort-merge join algorithm. Give the cost of joining R and S with a sort-merge join.

   c. (15) Describe a hash-join algorithm. Give the cost of joining R and S with a hash join.

2. [**Graph algorithms**] (30) The following questions test your understanding on basic graph algorithms.

   a. (10) Given a directed graph G (V, E, L) with V the node set, E the edge set and L a function that assigns to each edge e in E a label L(e). A label constrained reachability query Q(s,t,M) tests if there exists a path from a source s to a target t with a path, which consists of edges having a label from a label set M. Give an algorithm (pseudo-code) to answer query Q. [*_A straightforward way is to revise BFS or DFS traversal_*].

   b. (20) Consider a network G (V, E) of servers, where each edge (u,v) represents a communication channel from a server u to another server v. Each edge has an associated value r(u,v), which is a constant in [0,1]. The value represents the reliability of the channel, i.e., the probability that the channel from u to v will not fail. Assume these probabilities are independent. Give an algorithm (pseudo-code) to find the most reliable path between two given servers. Give a correctness proof and complexity (in Big O notation) of your algorithm. [hint: transform the weight to non-negative numbers, e.g., -log r(u,v) and transform it to a familiar graph problem].

3. **[Approximate query processing]**. (25) This question continues our discussion on using data synopsis for query processing based on data-driven approximation. You are given a vector of numbers: [127, 71, 87, 31, 59,3,43, 99, 100, 42,0, 58, 30, 88, 72,130], each data point records the frequency of communication of a server in a 5 minutes interval. For example, in the first 5 minutes ([0,5]), 127 contacts; in the second 5 minutes ([5, 10]), 71 contacts…

   (1) Give the Haar decomposition and draw a corresponding error tree for the contacts data vector.

   (2) Give the process and result for reconstructing the frequency during time interval [15, 20] using Haar decomposition.

   (3) Use Haar decomposition and error tree to compute the total number of communication between time interval [15, 30].