CPTS 415 Big Data

Assignment 3

Instructor: Yinghui Wu

1. [Join operators] (60) This sets of questions test the understanding of basic database search operators. Consider a join $R.A=S.B$. 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. How many pages


d. (15) If $R.A$ is a foreign key that refers to $S.B$, what’s the cost of the block nested join algorithm in (1)?
[Hint: The foreign key constraint tells us that for every $R$ tuple there is exactly one matching $S$ tuple (because $S.B$ is a key).]

2. [Graph algorithms] (40) 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 to answer query $Q$. [*A straightforward way is to revise BFS or DFS traversal*].

b. (10) Denote the diameter of a tree $T=(V, E)$ as the largest of all shortest-path distances in the tree $T$. Use BFS only to give an algorithm to compute the diameter of a tree. Give a correctness and complexity analysis of your algorithm (in Big O notation)

c. (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 efficient 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., $-\log r(u,v)$; you want to find path $p$ from $s$ to $t$ with maximized $\Pi r(u,v)$ for each edge $(u,v)$ on $p$].

3. [Approximate query processing] * (bonus)

We introduced the definition of graph simulation as a relaxed subgraph matching semantics. Assume all pattern nodes in a pattern $P$ and the nodes in data graph $G$ have a same label. Prove the following two results: (1) if there is a cycle in a pattern $P$, then there must exist a cycle in the data graph $G$, if $G$ simulates $P$. (2) For any subgraph $Gs$ that is isomorphic to the pattern $P$, every node and edge of $Gs$ must be contained in a simulation relation from $P$ to $G$. 