Assignment #2

1. **[XML and RDF] (40)**
   (a) (10) Consider the database instance you gave in Assignment 1 Question 2 (a). Assume now that you don't have any schema. Give an XML document to represent the tuples as the fact about the airports.
   (b) (10) Consider the relational schemas you gave in Assignment 1 Question 2 (b). Give an XML schema representation of each relational schema. How do you encode keys? Foreign keys?
   (c) (20) Consider a set of natural language sentences collected from Web pages.
      i. A human can like another human.
      ii. A human can have a sex property of a man or a woman.
      iii. A man can be the father of another human.
      iv. A woman can be the mother of another human.
      v. A human can be married to another human.
      vi. A human can have a BirthYear property of type "xs:Year".
      vii. If a human is married to another, then they like each other.
      viii. If a human is a mother or father, the human is a parent.

   Write a RDF schema and give a graphical presentation to describe these relationships.

2. **[Graph Algorithm]** (30) The following questions test your understanding on basic graph algorithms
   a. (15) 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 node s to a target node t, which consists of edges having a label from a label set M. Give an algorithm (pseudo-code) to answer query Q.
   (hint: A straightforward way is to revise BFS or DFS traversal)

   b. (15) Consider a network $G(V, E)$ of servers, where each edge $e = (u, v)$ 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 server u to server v will not fail. Assume that these probabilities are independent. Give an algorithm (pseudo-code) to find the most reliable path between two given servers. Give the complexity (in Big O notation) of your algorithm.
   (hint: transform the weight to non-negative numbers and the problem may become very familiar to you).

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-minute interval. For example, in the first 5 minutes, 127 contacts are observed. In the next 5 minutes which is time interval [5, 10], 71 contacts, …
   a. Give the Haar decomposition and draw a corresponding error tree for the contacts data vector.
   b. Give the process and result for reconstructing the frequency during time interval [15, 20] using Haar decomposition (explain the path in a top-down fashion).
   c. Use Haar decomposition and error tree to compute the total number of communications between time interval [15, 30] (explain the path in a top-down fashion).