1. [Approximate query processing]. (20) 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].

2. [Parallel Data models] [20]

   a. What’s linear speedup and scaleup? Give three reasons why we cannot do better than linear speedup.

   b. Describe and compare the pros and cons of the three architectures for parallel systems.

3. [MapReduce] [30] This sets of questions test the understanding and application of MapReduce framework.

   a. Facebook updates the “common friends” of you and response to hundreds of millions of requests every day. The friendship information is stored as a pair (Person, [List of friends]) for every user in the social network. Write a MapReduce program to return a dictionary of common friends of the form ((User i, User j), [List of Common friends of i and j]) for all pairs of i and j who are friends. The order of i and j you returned should be the same as the lexicographical order of their names.

   You need to give the pseudo-code of a main function, and both Map() and Reduce() function. Specify the key/value pair and their semantics (what are they referring to?)

   b. Top-10 keywords. Search engine companies like Google maintains hot webpages in a set R for keyword search. Each record r from set R is an article, stored as a sequence of keywords. Write a MapReduce program to report the top 10 most frequent keywords appeared in the webpages in R. (hint: you need two rounds of MR process). Give the pseudo-code of your MR program.

   c. MapReduce for graph algorithms may not be efficient. Give an example to explain why. For your course project, discuss if MapReduce will be a good fit or not – this will be stated in your project milestone.

4. [Graph parallel models] [30] This sets of questions relates to three graph parallel models.

   a. What are the two commonly used parallel models designed for graph processing, beyond MapReduce? Give a comparison of the four parallel models: MapReduce, BSP, vertex centric, and GRAPE.

   b. Consider the common friends problem in problem 2.a. We study a “2-hop common contactor problem”, where a list should be returned for any pair of friends i and j, such that the list contains all the users that can reach both i and j within 2 hops. Write a pregel program to solve the problem and give the pseudo code.
c. We described how to compute reachability queries using GRAPE framework. Consider a class of $d$-bounded reachability as follows. Given a graph $G$, two nodes $u$ and $v$ and an integer $d$, it returns a Boolean answer YES, if the two nodes in $G$ can be connected by a path of length no greater than $d$. Otherwise, it returns NO.

Write a GRAPE-style algorithm to compute the query $G(u, v, d)$. You may use the template provided in the slide. Provide necessary correctness and complexity analysis.

* [course project] read the tutorial [http://web.stanford.edu/class/cs246/homeworks/tutorial.pdf](http://web.stanford.edu/class/cs246/homeworks/tutorial.pdf) and test MapReduce program in pseudo-distributed environment. If you use MapReduce for your course project, you should get to know how to execute your MR code now.

* Please send your homework to yinghui@eecs.wsu.edu. Do not send it to yinghui.wu@wsu.edu.