Distributed Computing, Spring 2017

Assignment 1

March 1, 2017

Problem Description

For a pair of (multi-)sets A and B, we say that A supports B if $B \subseteq A$. Given two collections of sets $\mathcal{T} = \{T_1, T_2, \cdots, T_n\}$, and $\mathcal{I} = \{I_1, I_2, \cdots, I_m\}$, we want to find out for each I_i, $1 \leq i \leq m$, the number of sets in $\mathcal{T}$ that support it, namely, the *total support* from $\mathcal{T}$. In real applications, $\mathcal{T}$ may stand for a collection of transactions and $\mathcal{I}$ for a set of interested patterns which may occur in the transactions. Note that a set in $\mathcal{T}$ and $\mathcal{I}$ is a multiset as they may contain the same element multiple times.

You are required to write an efficient parallel algorithm to solve the above problem and then implement your algorithm using C on your personal computer to simulate MPI on a virtual parallel system.

In your experimentation you must validate your algorithm by testing it on the input data sets provided and also measure the running time of your algorithm for the input datasets when the algorithm runs on a single node, 2 nodes, 4 nodes and 8 nodes respectively (4 runs in total). The output of your algorithm should contain for each set I_i in $\mathcal{I}$ 1) the indices of all the supersets of $\mathcal{T}$ containing I_i (e.g. SupSet(2)=\{24,31,48\} indicating I_2 is contained in T_{24}, T_{31} and T_{48}), and 2) the number of sets in $\mathcal{T}$ which contain I_i (e.g. |SupSet(2)|=3).

Submission of Assignment

You must submit both the *source code* of your program, the output and timing results of the test runs. Submissions of the assignment should be made online to the TA before 10pm on Tuesday 28 March 2017.