CSC346/646 Parallel Computation Spring 2015

K-means Clustering – Data

For our first project, we will develop a parallel program using OpenMP to do K-means clustering. Your main program should accept the name of a data file and a positive integer k. E.g. `% cluster datafile 12`

The data file is a text file. On the first line, there are two integers m and n, indicating the number of rows and columns respectively. The following m lines each contain n numbers (either 0 or 1).

Test data files are available on our web server; see:

`http://menehune.opt.wfu.edu/csc346`

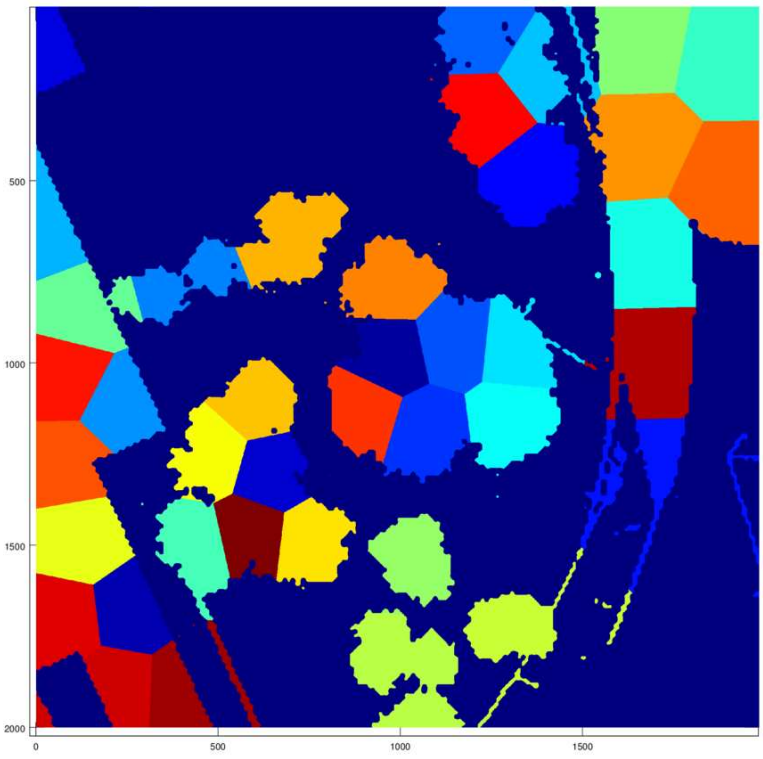
There are two data files you can download: a small one for development and testing, and a large one to measure speedup. A simple sequential implementation of K-means clustering runs the large file with 40 clusters in 84.7 seconds on gottlieb.

kdata.idat Data set size 201×201.

med_kdata.idat Data set size 2000×2000.

Run your parallel program to create 40 clusters. Measure the run time (excluding input and output time) of your program with N threads for $N = 1, 2, 3, ..., 50$. Use the time with $N = 1$ for the sequential time. Compute and plot a speedup curve[1].

A sample result with 40 clusters is shown below:



[1]Suggestion: Use octave for plotting the speedup curve.