Fast Approximate Spectral Clustering Versus Distributed Spectral Clustering

Christopher H. Lin (1121854) **Jijiang Yan (0837209)**

1 Proposal

On very large data sets, spectral clustering can be very slow. We can deal with this "big data" problem in two ways. One way is to parallelize the spectral clustering algorithm to run on many machines. Another way is to use an approximate spectral clustering algorithms that is run on one machine and trades accuracy for speed. We propose to compare these two methods in order to better understand their advantages and disadvantages.

In particular, we propose to compare K-means-based approximate spectral clustering and RP-tree-based approximate spectral clustering [1] against a naive parallelization of spectral clustering on GraphLab. We will test these algorithms on the CIFAR-10 dataset, CIFAR-100 dataset, and the Mouse Visual Cortex Dataset as contributed on the Graphlab datasets page.

By the project milestone, we propose to have completed learning about how to run spectral clustering on Graphlab, and be able to run it on the three proposed datasets.

References

[1] Yan, D. & Huang, L. & Jordan, M.I. (2009) Fast Approximate Spectral Clustering. In KDD.