Label Propgation algorithm implementation using GraphLab

Label Propagation by Raghvan et. al. [1] is a near linear time algorithm for detecting communities in Mega scale network. Having tested a simple map reduce based implementation of the algorithm, I was interested in testing parallel implementation of the algorithm. By keeping the entire network in memory and performing all computations on a single machine, significant performance improvements can be achived.

This page describes preliminary results using the GraphLab API.

Code

Since the code requires 32 Giga bytes of RAM for performing communitiy detection on Twitter's social network, I have created an AWS (Amazon Web Services) snapshot, which contains both code as well as data for testing.

- For using the Snapshot, create a High-Memory Double Extra Large Amazon EC2 instance with AMI 08f40561 (Official 64 bit Ubuntu Image).
- Create an EBS instance using snapshot **snap-5615833a** and mount it as an ext3 filesystem.
- execute run.sh, it will download the new version of the code and will perform community detection on Twitter social network containing users with less than 900 followers.
- The results are stored in res.txt, you can also access previously calculated results from seventh and fourteenth iteration from oldresults directory.
- Each line in the result file has following format:
  User \t Community \n
  In case there is no screen_name available for a given user, then the userid is present.

Additionally you can just download the latest version of the code from http://dl.dropbox.com/u/11064717/graphlabapi.tar.gz

Dataset

This Network dataset collected by researchers from KAIST is used for testing the code, Currently it has follower information for 40 Million users collected in June 2009. Also for purpose of converting userids to screen names we use dataset provided by Prof. Jure Leskovec and J. Yang.
Both datasets can be obtained from following link http://an.kaist.ac.kr/traces/WWW2010.html

Please have a look at the terms under which the data is being distributed by the researchers, It might NOT be suitable for commercial or non-research applications.

References:

1. Usha Nandini Raghavan, Réka Albert, and Soundar Kumara, Near linear time algorithm to detect community structures in large-scale networks. Phys. Rev. E 76, 036106 (2007)
2. Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon, What is Twitter, a Social Network or a News Media? WWW 2010, April 26–30, 2010

Template modified version of a design by David Kohout