Assignment 3 : Map Reduce

Post Date: Tuesday, October 13, 2015
Due Date: Wednesday, October 28, 2015 11:59 PM
TA Office Hours: Monday, October 26, 2015 4:30PM - 6:00PM (LSRC D309)

In this assignment, you will practice writing Map Reduce programs on Amazon Web Services (AWS).The following tutorial will help you (a) get started with AWS, (b) provide locations of the input datasets on the cloud, and (c) point to tutorials on how to write Map Reduce code on AWS using the `mrjob` python package. Please debug your programs in local mode first before running on AWS. Please include the scripts used in the report that you submit.
There are two inputs datasets to this assignment:

- G_T: Twitter follows graph (*id1, id2*) (24.3GB)
- C: List of ids corresponding to celebrities (2.9MB)

Problems

1. Compute the following on G_T
   - Number of nodes in the graph
   - Average (and median) indegree and outdegree
   - # nodes with indegree $> 10,000$
2. Compute the number of connected components in G_T and their sizes.
3. Construct the subgraph induced on C. Repeat problems 1 and 2 on this subgraph.
4. For every node $x \in G_T$, identify nodes $y \in G_T$ such that $sim(x, y) > 0.7$, where

$$sim(x, y) = \frac{|N(x) \cap N(y)|}{|N(x) \cup N(y)|}$$

   and $N(x)$ is the set of in and out neighbors of x.