

Abstract

Working on big data is challenging and interesting. Size of the Social Networking data is growing bigger everyday and analysis on such data is getting more complicated. Twitter which has 500 million users has an impressive growth rate. Social network can be viewed as a map of the individuals, and the ways how they are related to each other. We exploit this mapping paradigm to use it in our analysis. Map Reduce is a buzzword in Big Data industry and a perfect framework to perform analysis on such data. We used Hadoop and Amazon Elastic Map Reduce Cluster for our analysis on Community Detection using Label Propagation Algorithm. Different notion of analysis can be derived from this work.

Objective

To detect communities using the edge data of user connections in social networks using Map Reduce like Framework.

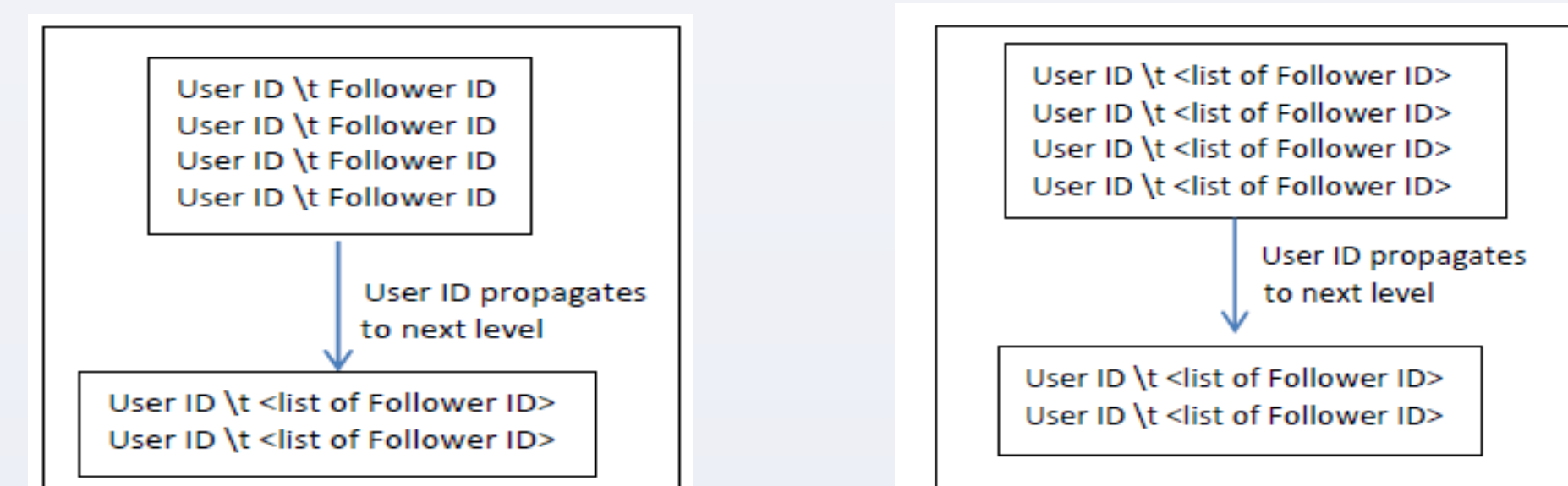
DATA SET

- Social Graphs format
 - USER \t FOLLOWER \n
- Directional or Un-directional
- Kaist Dataset for twitter data
 - <http://an.kaist.ac.kr/traces/WWW2010.html>
 - Huge data 26.4GB
- Graphs generated by LFR benchmark
 - Open source code for generation of Graph DB
 - Testing accuracy using artificially generated communities

Label Propagation Algorithm

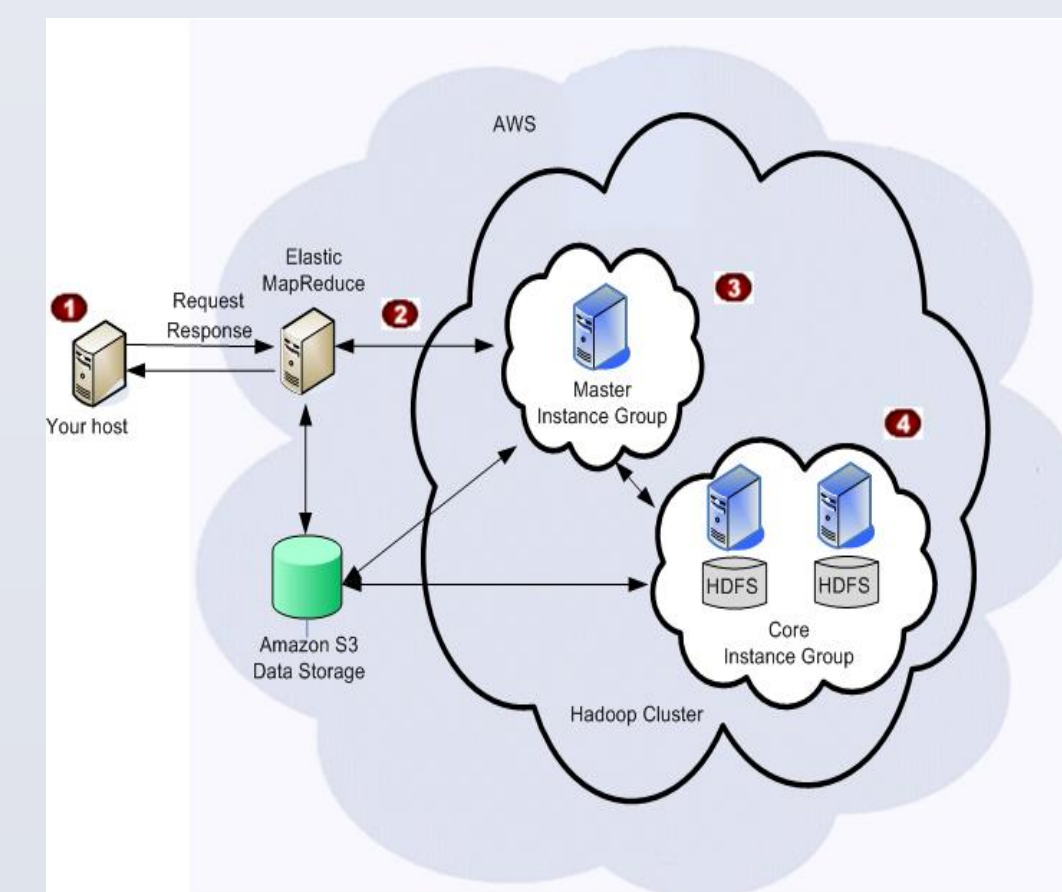
- Initialize labels at all nodes in the network. For a given node x , $C_x(0) = x$.
- Set $t = 1$.
- For each $x \in X$ chosen in that specific order, let $C_x(t) = f(C_x(t-1), \dots, C_{xk}(t-1))$ f returns the label occurring with the highest frequency .
- If every node has a label that the maximum number of their neighbors have, then stop the algorithm(no further change in labels). Else, set $t = t + 1$ and go to (3).

Community Detection with Map Reduce



- Collect all ids followed by a user
user id <list of following all users(tab separated)>
- Initially run mapper with all nodes with labels same as node id
- Apply Label Propagation Algorithm
- Connected nodes with the same label form a community at end of iterations

Infrastructure



- EMR – elastic map reduce**
 - Open source code for generation of
 - One master with multiple number of slaves
 - Number of mappers and reducers can be controlled
 - CLI and Web Interface Available
- BOTO**
 - Python integrated interface to Amazon Web Services
 - Command line or python script to launch map reduced jobs on EMR

Correctness on LFR benchmark

- LFR benchmark (A. Lancichinetti, S. Fortunato, and F. Radicchi) used to create artificial graphs
- Generated graph data using graph bench <https://github.com/tinkerpop/tinkubator/tree/master/graphdb-bench>
- Community detection algorithms: a comparative analysis (Andrea Lancichinetti and Santo Fortunato)
- Generated directional graph using benchmark tool having 20 groups, 1000 nodes and around 2000 edges
- Results 22 groups after 11 iterations

Results

- LFR Benchmark

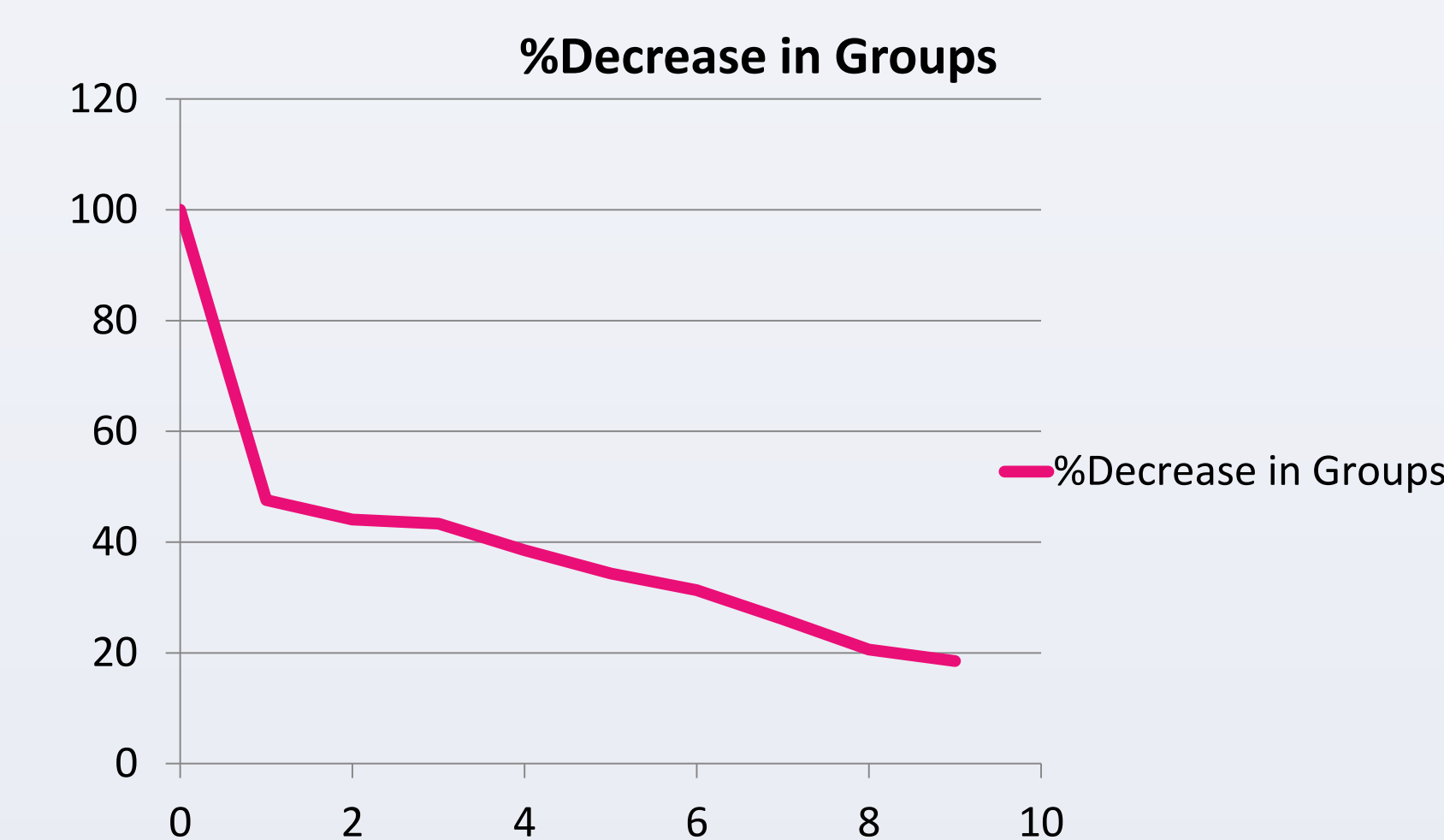


Fig: % Reduction in Communities with Label Propagation in each iteration

Iteration	Groups Detected
0	1000
1	524
2	293
3	166
4	102
5	67
6	46
7	34
8	27
9	22

- Twitter Data

% Reduction in Comm. Detection for every iteration

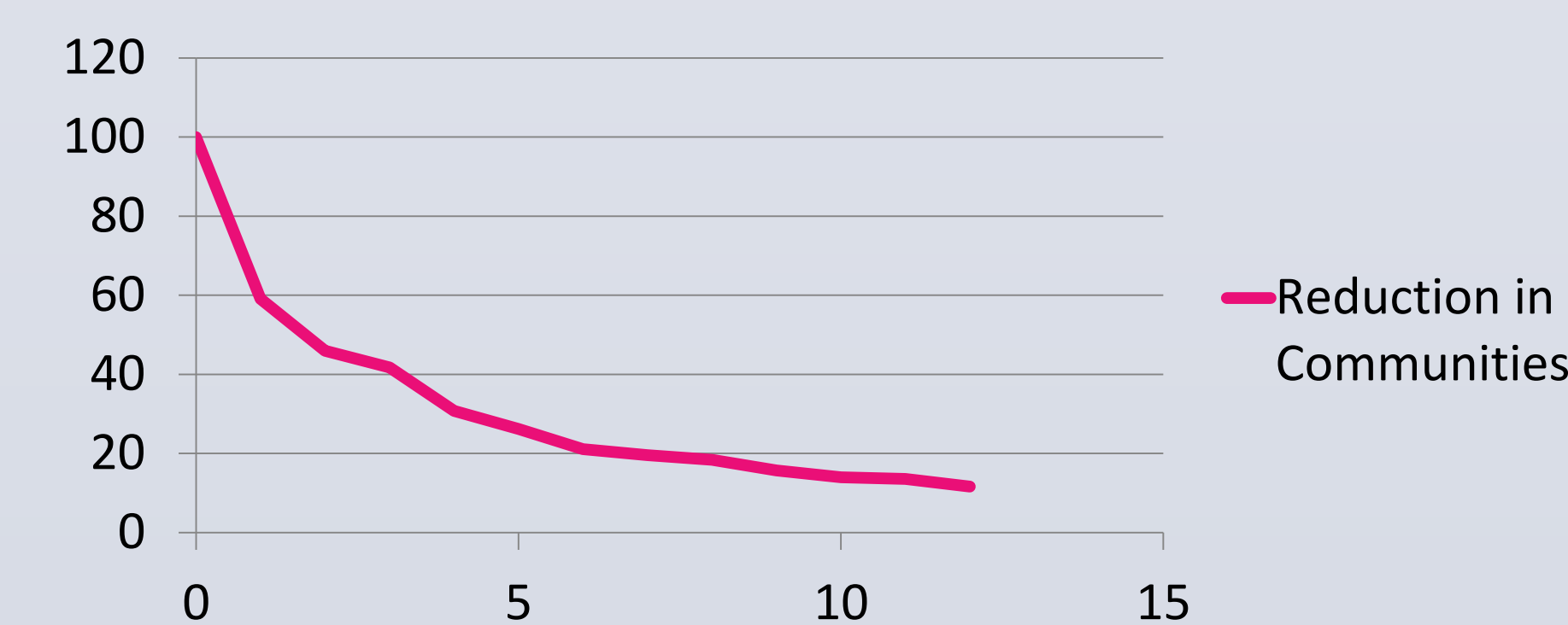


Fig: % Reduction in Communities with Label Propagation in each iteration

Iteration	Communities
0	100000
1	40943
2	22921
3	12884
4	8918
5	6585
6	5197
7	4180
8	3412
9	2876
10	2475
11	2139
12	1891

Working on AWS

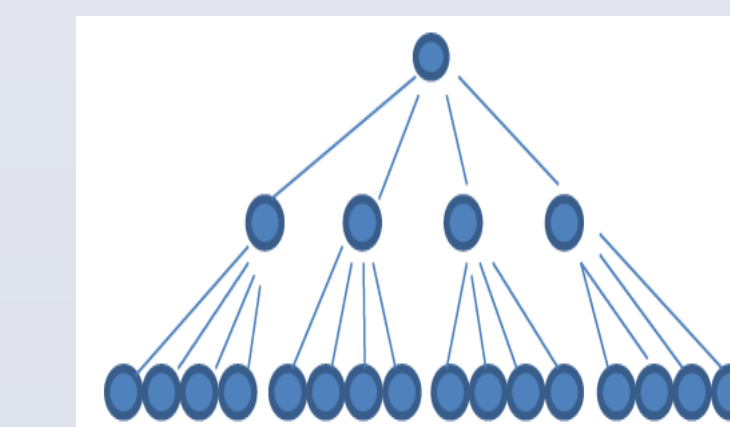
- Web Interface can only run Single step MR job
- Command line can run Multistep but still multiple issues
- No tight coupling between services like EC2 ,S3 ,EMR
- 5gb upload limit on files
- Unzipping files on S3 is also a challenge
- Cannot use normal commands that can be used with normal Hadoop clusters as all options are not supported by EMR
- Charged for no result or failures
- Always good to try with Hadoop and then shift to AWS

Conclusion

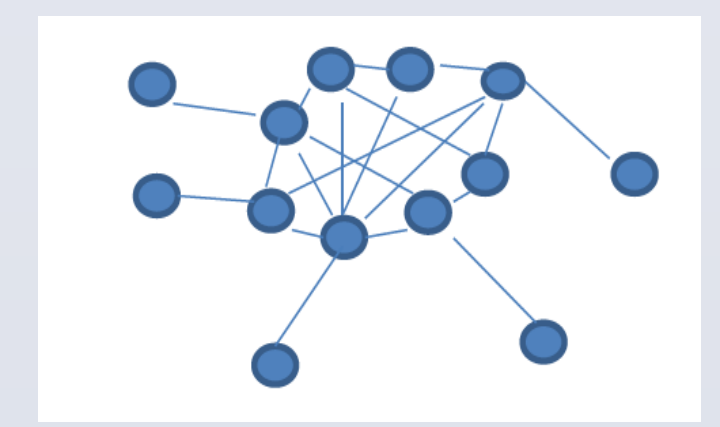
- The AWS EMR infrastructure can be used for detecting communities in huge datasets using Label propagation method
- Reduction in number of communities is faster at initial iterations as compared to later iterations
- The same method can also be used to detect influential users in communities

Future Work

- Link Prediction Mechanism
- Evaluation by comparing result from Data Set snapshots taken at different times



- Link Prediction with 2-level BFS
- Exhaustive
- Not possible to run on he dataset



- Link Prediction using community detection
- Fast and Simple

References

- U.N. Raghavan, R. Albert and S. Kumara. Near linear time algorithm to detect community structures in large-scale networks. Physical Review E, 76:036106, 2007
- I.X.Y. Leung, P. Hui, P. Liò and J. Crowcroft. Towards real-time community detection in large networks. Physical Review E, 79:066107, 2009
- H. Kwak, C. Lee, H. Park, and S. Moon. What is Twitter, a social network or a news media?. In WWW '10: Proceedings of the 19th international conference on World Wide Web, pages, 591-600
- A Lancichinetti, S Fortunato, Community detection algorithms: a comparative analysis, Phys. Rev. E, 2009

Acknowledgements

- We thank Dr. Daisy Zhe Wang (daisyw@cise.ufl.edu), for her guidance and support during the entire course of the project.