Networks: Structure & Economics — Caltech CMS/CS/EE 144
Miniproject 1 — February 8rd, 2016

Churn

John Li, Albert Ge, Timothy Chou, Kevin Chang

Rankmaniac Report

1 Overview

During this RankManiac project, our group "churn" implemented PageRank on the given network via Map Reduce through Hadoop on Amazon AWS Clusters. Two sample datasets, "GNPn100p05" and "EmailEnron", were distributed to us for the purpose of local testing. Our Rankmaniac score was determined by our PageRank's elapsed time and the correctness on a larger, hidden dataset on these Amazon AWS clusters. This report's purpose is to illustrate the process for our implementation of PageRank.

2 Initial Framework

Development

The data that was given to us was in the form:

NodeId:[NODEID]   [CURRENT_RANK],[PREVIOUS_RANK],[ADJACENCY_LIST]

After a basic understanding of the structure of the project, our initial framework of MapReduce was as follows:

- PAGERANK_MAP:
  The input of this mapping will either be in the above format, or the output of PROCESS_REDUCE, which prepends a [ITERATION] value in front of the current values. This mapping function will take in the key/value pair, increment [ITERATION] by 1, and
- PAGERANK_REDUCE:
- PROCESS_MAP:
- PROCESS_REDUCE:

Problems

3 Refactored Framework

Refactoring

Optimizations

4 Testing

5 Contributions

6 References