Data Mining and Analysis

MapReduce + Spark: 3

CSCE 676 :: Fall 2019
Texas A&M University
Department of Computer Science & Engineering
Prof. James Caverlee

Pregel: Large-Scale Graph Processing

- Single machine tools can’t scale
- Parallel tools didn’t address fault tolerance
- MapReduce is not directly suitable
- Need a framework that can enable writing many graph algorithms while taking care of infrastructure -> Pregel

Think like a vertex!

- Message passing with "supersteps"
- Vertex program:
  - processes messages
  - Updates vertex state
  - Send messages

Think like a vertex

Aggregators

- User specific function
- Each vertex sends it a value
- Each vertex receives aggregate (vals)
- Pregel keeps track of superstep **S** and **S+1**

Today's Plan

- PageRank in RDD!
- PageRank in Pregel
- PageRank in graphframes
- Connected Components in Pregel

See spark-graphs.ipynb