3. GRAPH ANALYSIS

Sangmi Lee Pallickara
Computer Science, Colorado State University
http://www.cs.colostate.edu/~cs535

This material is built based on,

- Joseph E. Gonzalez, Reynold S. Xin, Ankur Dave, Daniel Crankshaw, Michael J. Franklin, Ion Stoica, "GraphX: Graph Processing in a Distributed Dataflow Framework" 11th USENIX, OSDI 2014
- GraphX: Unifying Data-Parallel and Graph-Parallel Analytics, presented by Joseph Gonzalez, at Strata 2014
- GraphX: Graph Processing in a Distributed Dataflow Framework, presented by Joseph Gonzalez, at OSDI 2014

Many Graph-Parallel Algorithms

- Collaborative Filtering
  - Alternating Least Squares
  - Stochastic Gradient Descent
  - Tensor Factorization
- Structured Prediction
  - Loopy Belief Propagation
  - Max-Product Linear Programs
  - Gibbs Sampling
- Semi-supervised ML
  - Graph SSL
  - CoEM
  - Community Detection
  - Triangle-Counting
- K-core Decomposition
- K-Truss
- Graph Analytics
- PageRank
- Personalized PageRank
- Shortest Path
- Graph Coloring
- Classification
- Neural Networks

FAQs

- November 30: Exam review
  - Presentation Schedule is ready
- [12/5, Tuesday 8:45AM ~ 10:45AM]
  - Team Rocky Mountain Juniper
  - Team Peachleaf Willow
  - Team Narrowleaf Cottongrass
- [12/7, Thursday 8:45AM ~ 10:45AM]
  - Team Limber Pine
  - Team Engelmann Spruce
  - Team Chokecherry
  - Team Boxelder
- Final report/software submission deadline: 5PM December 4
  - Software demo: December 7, 8
PageRank on the Live-Journal Graph

- GraphLab is 60x faster than Hadoop
- GraphLab is 10x faster than Spark

View a Graph as a Table

Property Graph Data Model

Separate Properties and Structure

Triplets Join Vertices and Edges

- The `triplets` operator joins vertices and edges:

```
SELECT t1.id, reduce(DF(merge(t1))) AS sum
FROM triplets AS t1
GROUP BY t1.id
```
Distributed Graphs as Tables (RDDs)

Where,

- \( p \) is the number of partitions
- \( n \) is the number of vertices
- \( (u, v) \) is an edge in the graph
- \( col = (u \times \text{mixingPrime}) \mod n \)
- \( row = (v \times \text{mixingPrime}) \mod n \)

- 2D hash algorithm

- For the cluster with \( p \) nodes (\( p = \frac{n}{p} \)), the edge \( (u, v) \) is assigned to

\( \text{row} \times \text{ceilSqrtNumParts} + \text{col} \mod \text{numParts} \)

Where,

- \( \text{row} = (u \times \text{mixingPrime}) \mod \text{numParts} \)
- \( \text{col} = (v \times \text{mixingPrime}) \mod \text{numParts} \)

Caching for Iterative mrTriplets

- Identifying and bypassing joins for unused triplets fields

- Example: PageRank only accesses source attribute

Reduction in Communication Due to Cached Updates

- Example: PageRank on Twitter

Join Elimination

- Identify and bypass joins for unused triplets fields
Performance Comparisons

GraphLab is roughly 3x slower than GraphX.

GraphX scales to larger graphs

GraphX is roughly 2x slower than GraphLab.

GraphX performs comparably to state-of-the-art graph processing systems.