Graph Processing
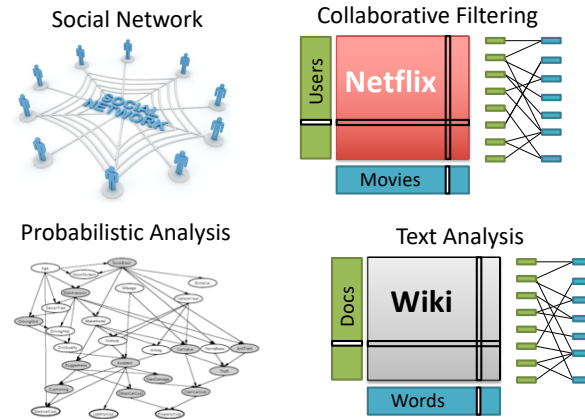
COS 518: Advanced Computer Systems

Lecture 12

Mike Freedman

[Content adapted from K. Jamieson and J. Gonzalez]

Graphs are Everywhere

Social Network

Collaborative Filtering

Users — Netflix — Movies

Probabilistic Analysis

Text Analysis

Docs — Wiki — Words

Concrete Examples

Label Propagation

Page Rank

Label Propagation Algorithm

- Social Arithmetic:

  Sue Ann — 40% — 80% Cameras / 20% Biking

  50% What I list on my profile
  40% Sue Ann Likes
  + 10% Carlos Like
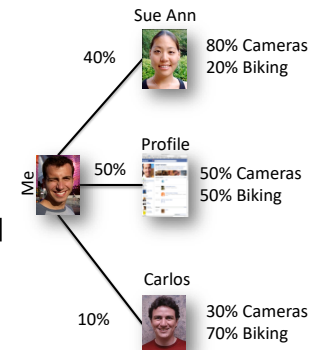
  I Like: 60% Cameras, 40% Biking

  Profile — 50% — 50% Cameras / 50% Biking

- Recurrence Algorithm:

  $$Likes[i] = \sum_{j \in Friends[i]} W_{ij} \times Likes[j]$$
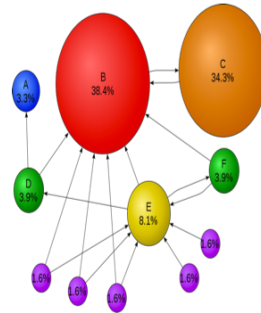
  – iterate until convergence

  Carlos — 10% — 30% Cameras / 70% Biking

- Parallelism:

  – Compute all *Likes[i]* in parallel
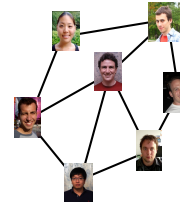
PageRank Algorithm

- PageRank of *u* is dependent on PR of all pages linking to *u*, divided by the number of links from each of these pages

- Recurrence Algorithm:

  $$PR[u] = \Sigma_{v \in Bu} \; PR[v] \; / \; L[v]$$

  – iterate until convergence

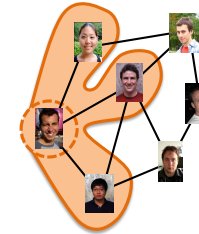- Parallelism:
  – Compute all *PR[u]* in parallel
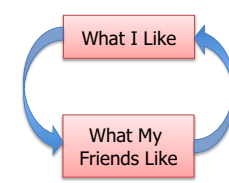

Properties of Graph Parallel Algorithms

Dependency Graph	Factored Computation	Iterative Computation

What I Like

What My Friends Like

Map-Reduce for Data-Parallel ML

- Excellent for **large data-parallel** tasks!

Data-Parallel   Graph-Parallel

MapReduce

Feature Extraction   Algorithm Tuning

Basic Data Processing

MapReduce?

Lasso   Label Propagation

Kernel Methods   Belief Propagation

Tensor Factorization   PageRank

Deep Belief Networks   Neural Networks

7

Problem: Data Dependencies

- MapReduce **doesn't** efficiently <u>express</u> data dependencies
  – User **must code** substantial data transformations
  – Costly **data replication**

Independent Data Rows

Iterative Algorithms

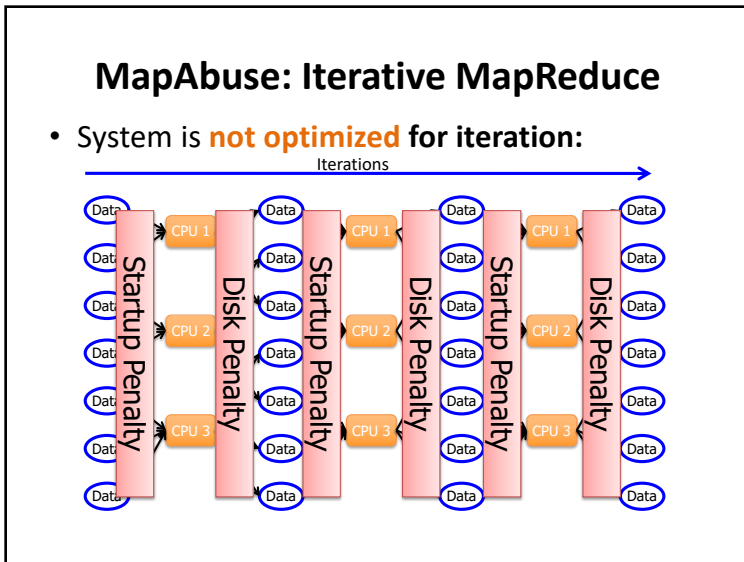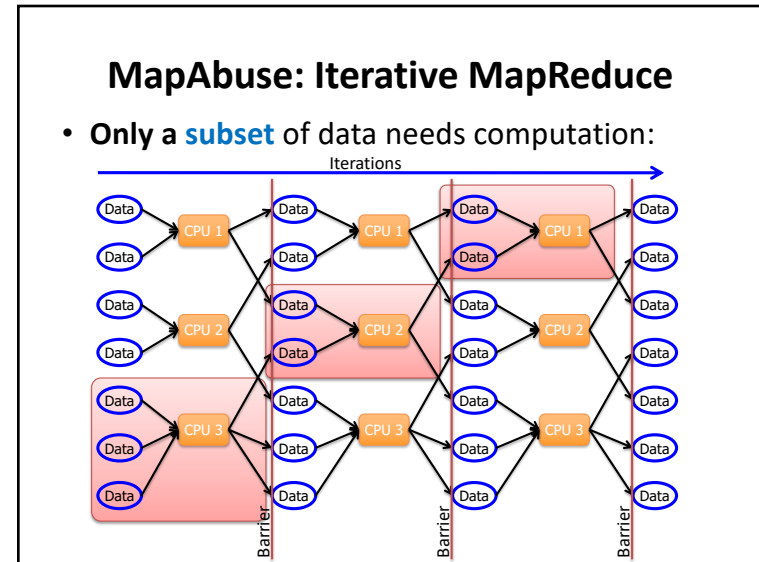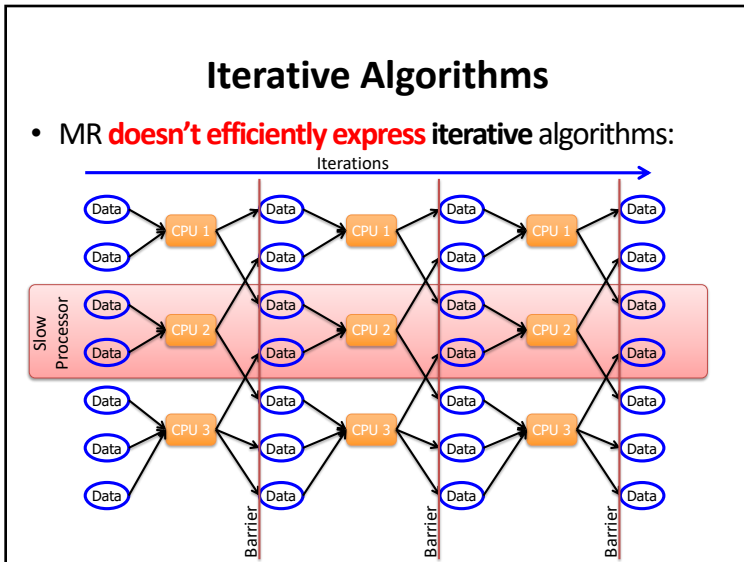- MR **doesn't efficiently express** **iterative** algorithms:

MapAbuse: Iterative MapReduce

- **Only a subset** of data needs computation:

MapAbuse: Iterative MapReduce

- System is **not optimized** for iteration:

ML Tasks Beyond Data-Parallelism

Data-Parallel — Graph-Parallel

Map Reduce

Feature Extraction   Cross Validation

Computing Sufficient Statistics

Graphical Models
Gibbs Sampling
Belief Propagation
Variational Opt.

Collaborative Filtering
Tensor Factorization

Semi-Supervised Learning
Label Propagation
CoEM

Graph Analysis
PageRank
Triangle Counting

12

Next set of readings

- Graph processing
  - Why relationships, sampling, and iterations often use in graph processing not well fit by MapReduce
  - How to take a graph-centric processing perspective

- Machine learning
  - These are solving one type of ML algorithm
  - What other systems are needed, particularly given heavy focus on iterative algorithms

13

Today's readings

- Streaming is about unbounded data sets, not particular execution engines

- Streaming is in fact a strict superset of batch, Lambda Architecture destined for retirement

- Needs of good streaming systems: correctness and tools for reasoning about time.

- Differences between event time and processing time, and the challenges they impose

14