

CS 378 – Big Data Programming

Lecture 28

Page Rank

An Iterative Algorithm in Spark

Review

- Assignment 12
 - Read data into a dataset
 - Queries
- Questions?

Example - Page Rank

- Walk through page rank algorithm for Spark
- See a more complex algorithm using Spark
 - Iterative

Basic Page Rank Algorithm

From Learning Spark, pp. 66-67

- Give each page an initial rank of 1
- On each iteration, have page p send a contribution of $\text{rank}(p) / \text{numNeighbors}(p)$ to its neighbors
- Set each page's rank to $0.15 + 0.85 * \text{contributionsReceived}$

Page Rank - Example

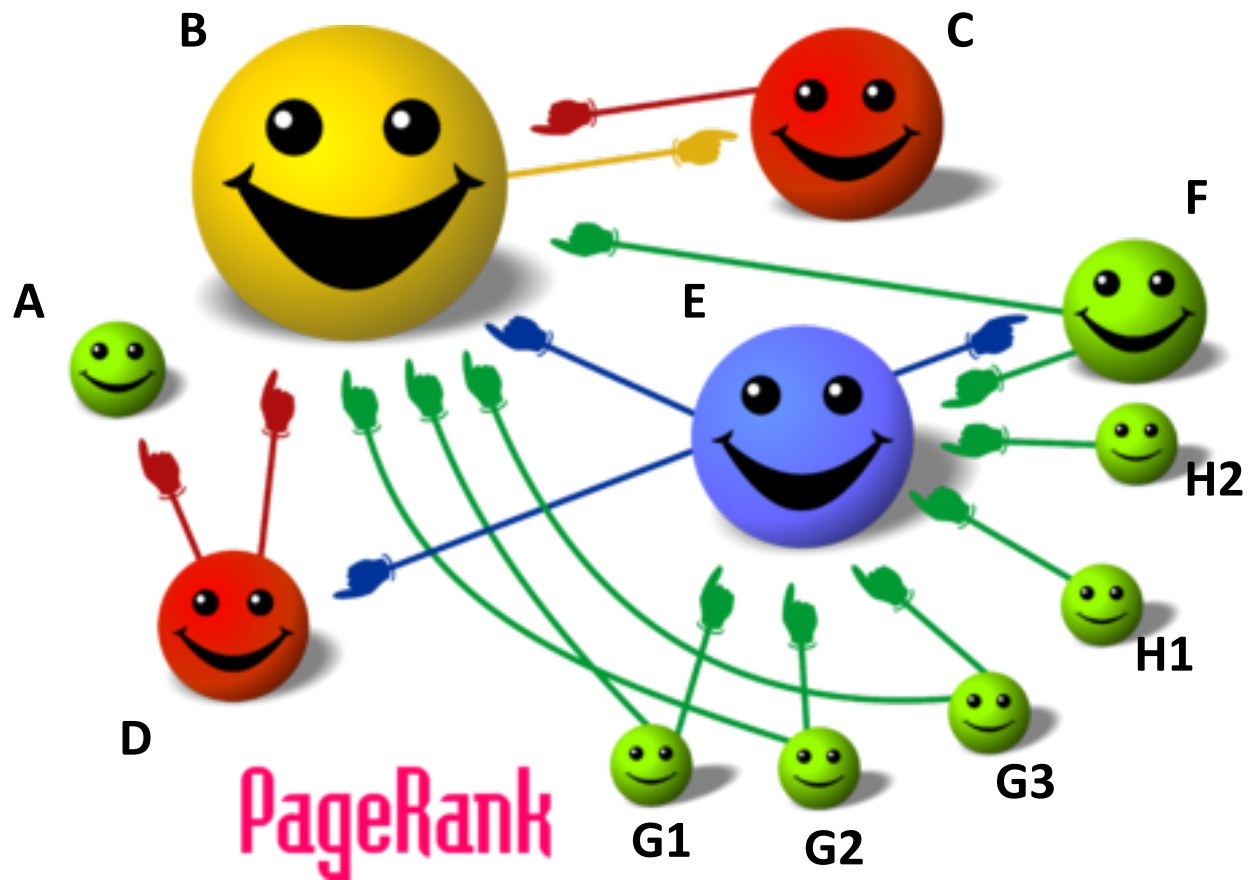


Image from: en.wikipedia.org/wiki/File:PageRank-hi-res.png

Page Rank

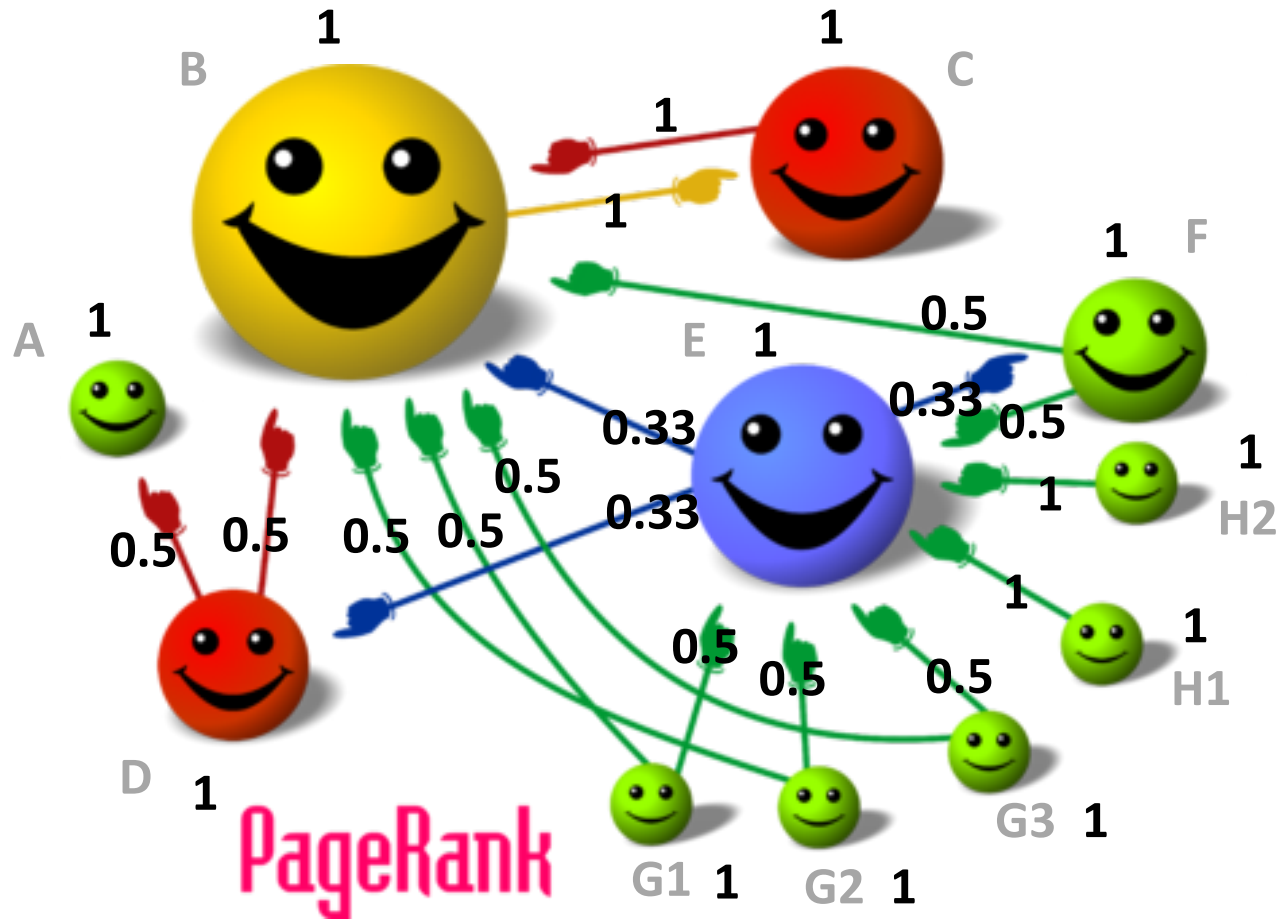


Image from: en.wikipedia.org/wiki/File:PageRank-hi-res.png

Page Rank

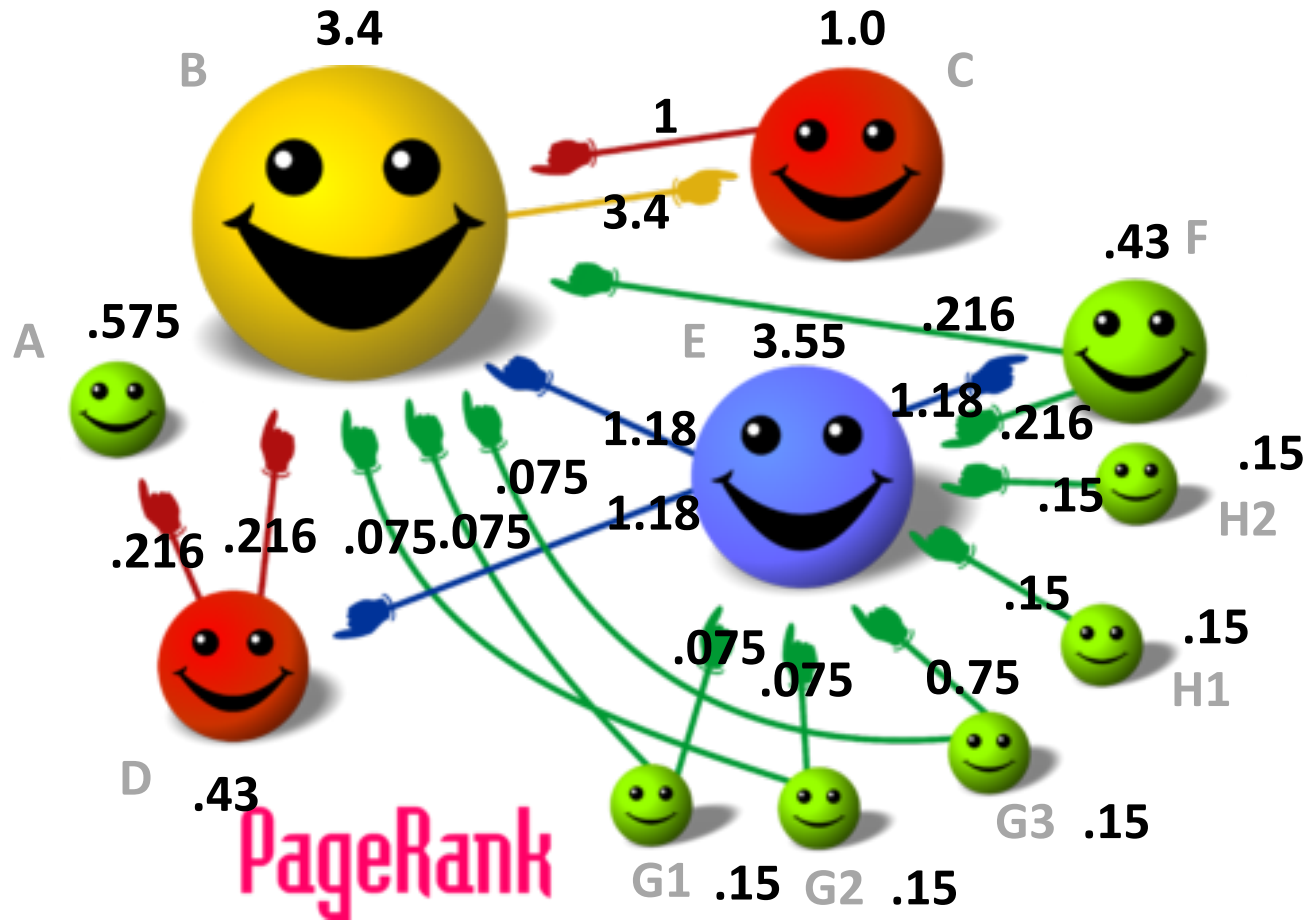


Image from: en.wikipedia.org/wiki/File:PageRank-hi-res.png

Page Rank

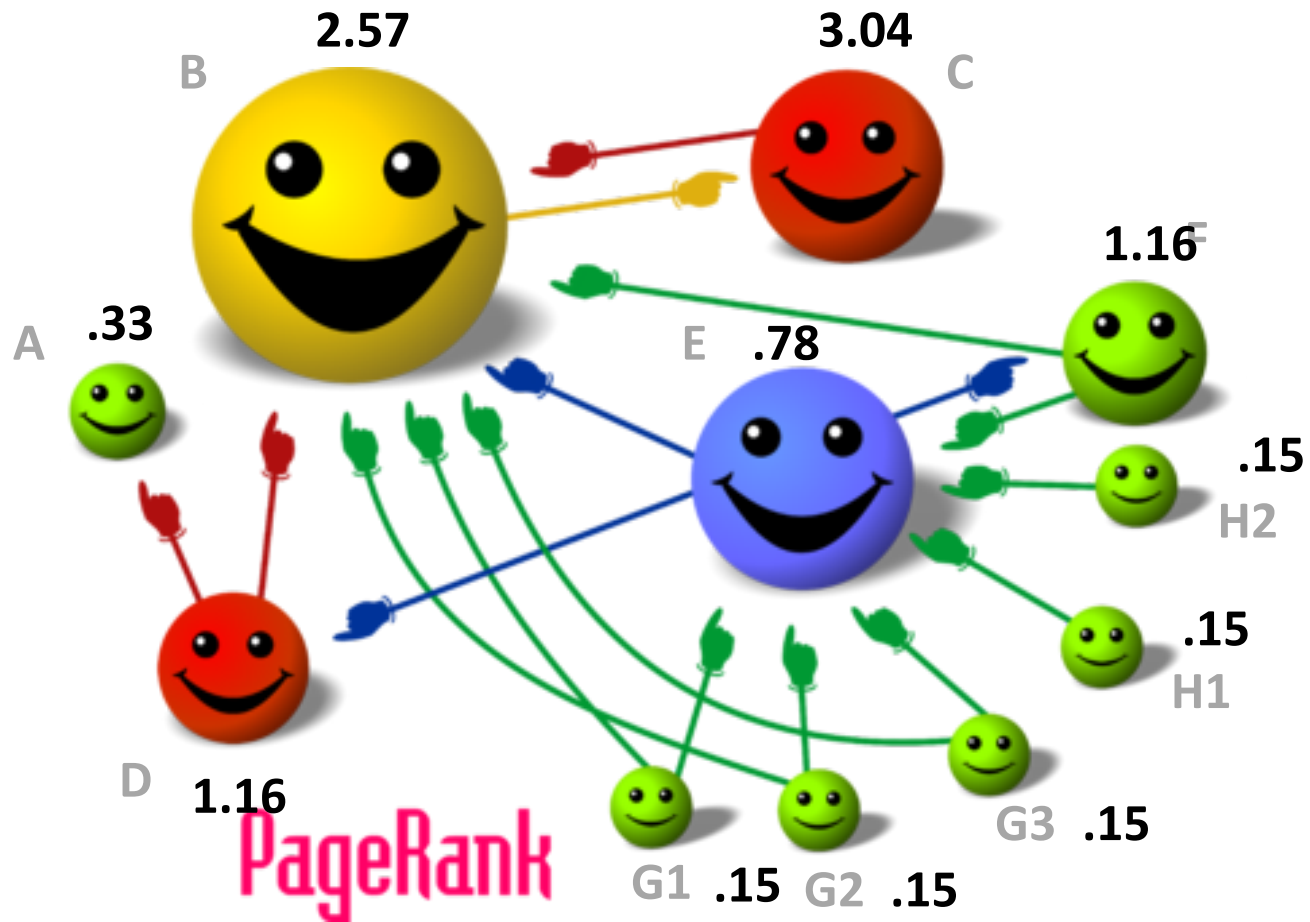
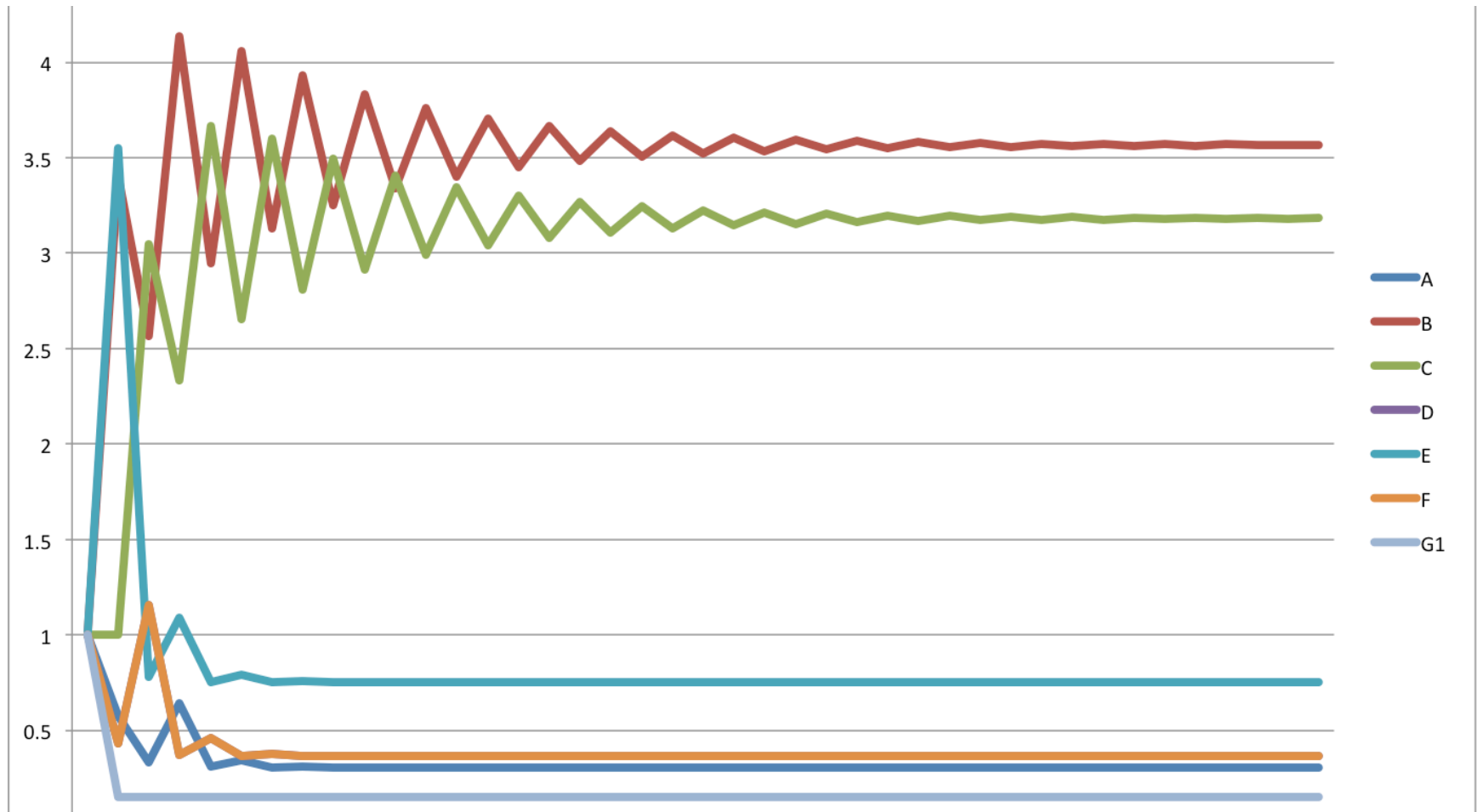


Image from: en.wikipedia.org/wiki/File:PageRank-hi-res.png

Page Rank - Results



Other Topics

for Further Reading

- Discussed in the textbook
- Other file systems
 - HDFS, S3, ...
- Machine learning – MLLib
 - Many algorithms implemented
 - See: spark.apache.org/mllib

Other Topics

for Further Reading

- GraphX – Graph processing
 - Algorithms:
 - PageRank
 - Connected components
 - Label propagation
 - SVD++
 - Strongly connected components
 - Triangle count