CS 378 – Big Data Programming

Lecture 28

Page Rank

An Iterative Algorithm in Spark

Review

- Assignment 12
  - Create user sessions
  - Order events by timestamp
  - Order sessions by user ID, then referring domain
  - Partition sessions by referring domain
  - Sample OTHER sessions (1 in 1,000)

- Questions?

Example - Page Rank

- Walk through page rank algorithm for Spark

- See a more complex algorithm using Spark
  - Iterative

- Show benefits of partitioning, persistence

What is Page Rank?

Algorithm for weighting linked documents

Part of Google's ranking algorithm – lots of other stuff included

Basic idea

Rank++ for inbound links

Rank++ for high rank links

In this image:

Size proportional to # inbound links

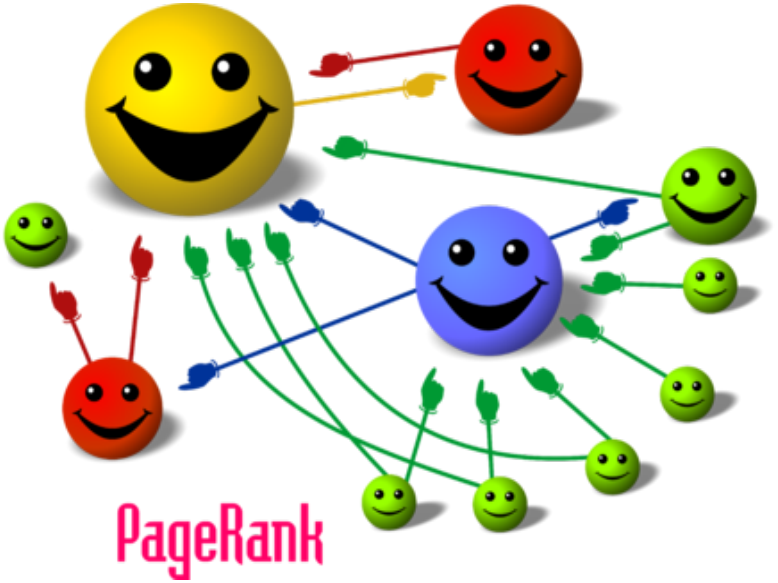


Image: en.wikipedia.org/wiki/File:PageRank-hi-res.png

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 `rank(p)/numNeighbors(p)` to its neighbors
- Set each page's rank to

  `0.15 + 0.85 * contributionsReceived`

Page Rank - Example

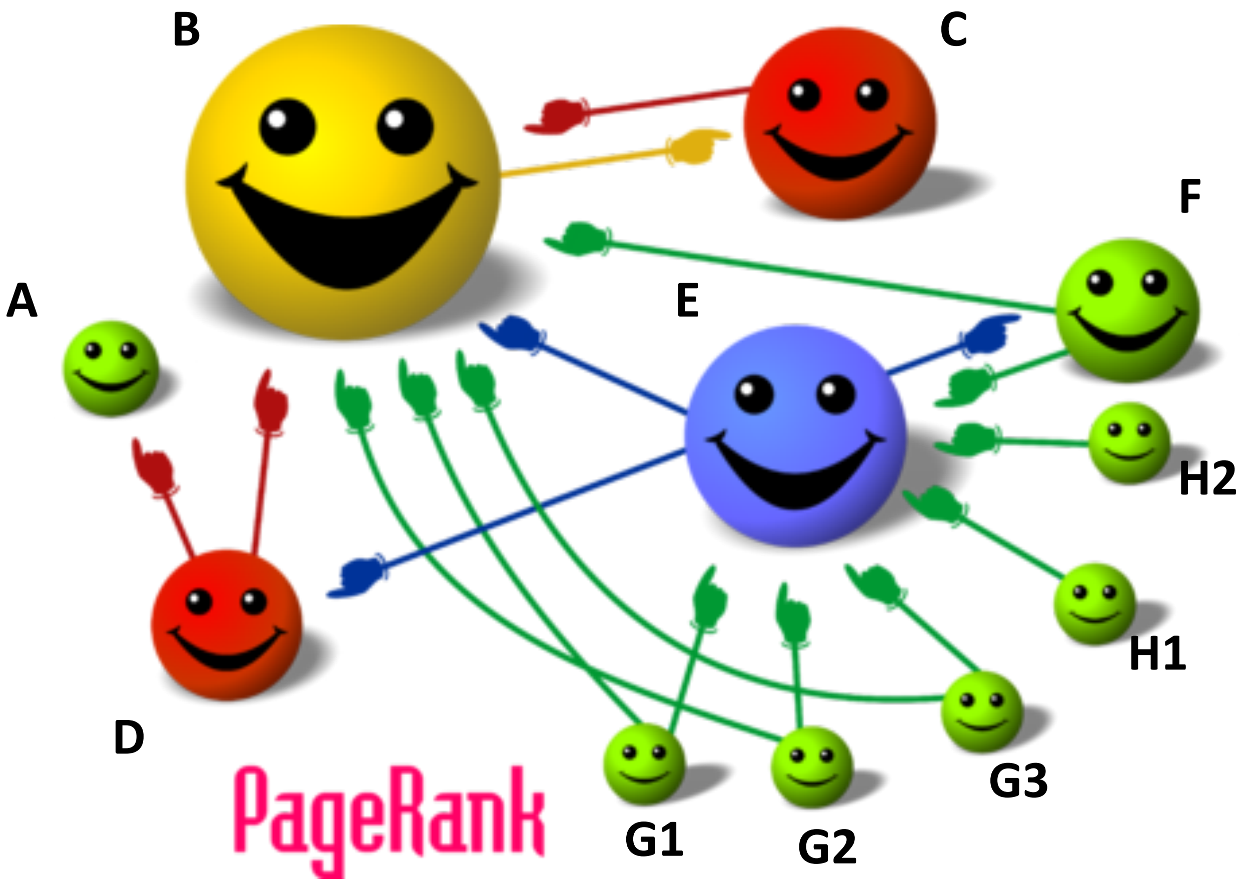
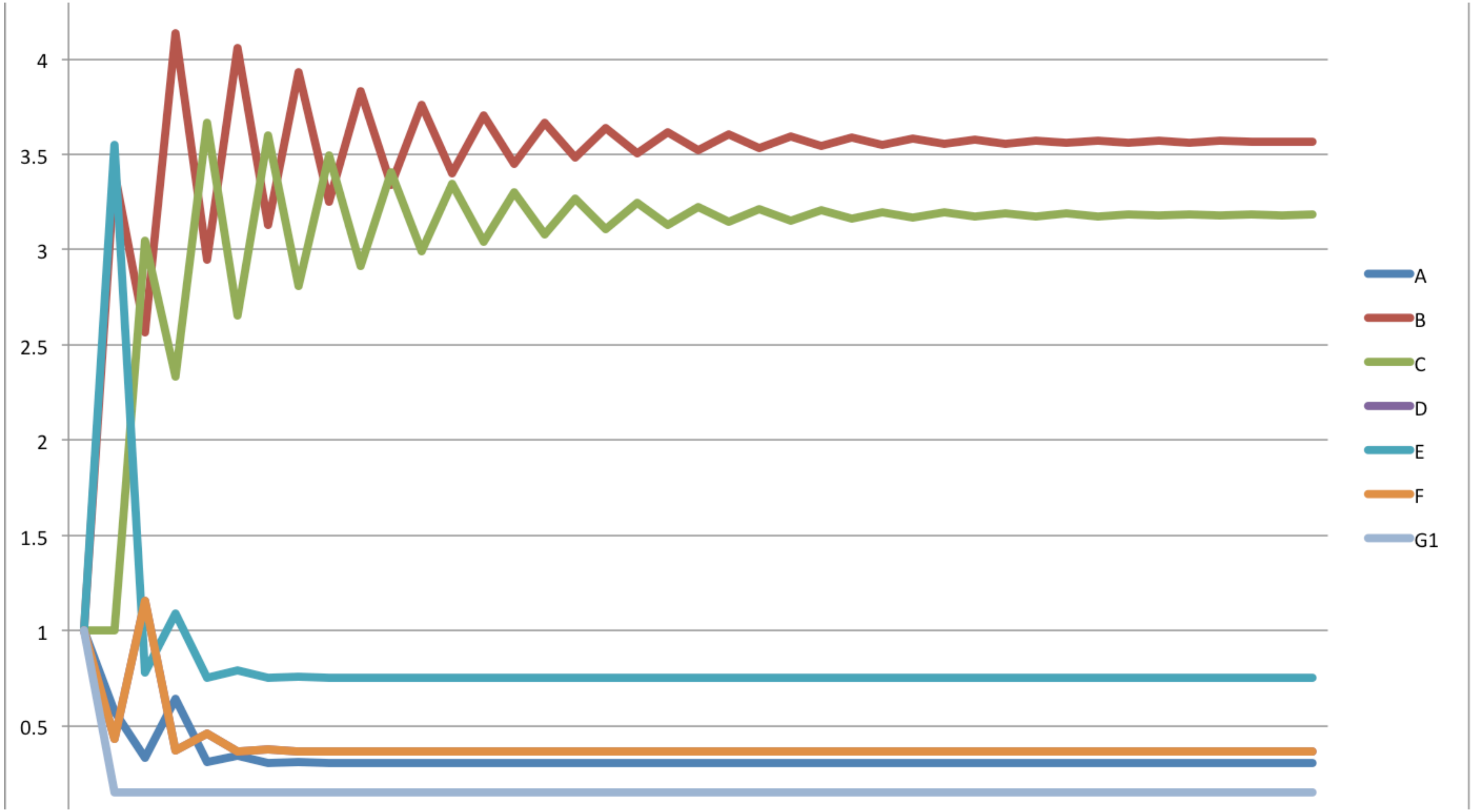


Image from: en.wikipedia.org/wiki/File:PageRank-hi-res.png

Page Rank - Results



Other Topics

for Further Reading

- Discussed in the textbook
- Other file systems
  - HDFS, S3, …
- Database – Spark SQL
- Streams – Spark Streaming

Other Topics

for Further Reading

- 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