15-319 / 15-619 Cloud Computing

Recitation 15

April 28st & 30th 2015

Overview

- **Last week’s reflection**
  - Spark Program
- **This week’s schedule**
  - Project 4.3
- **Demo**

Reflection on P4.2

- Implement a search engine in Spark
  - Wikipedia Page Dataset
  - TF-IDF
  - PageRank
- Issues
  - Scala as a new Language
  - Spark cluster management
  - Jobs taking too long to run

Survey!

- Time for you to reflect on the course
- Anonymous survey will be mailed to you
- System keeps track of survey responders
- 2% bonus to sweeten the deal
- We want to know:
  - Course content, quality, improvements
  - Projects, quality, experience, fun factor, time investment
  - Logistics, course support, improvements
  - How would you improve the course?
- The course relies on feedback for improvement semester to semester!

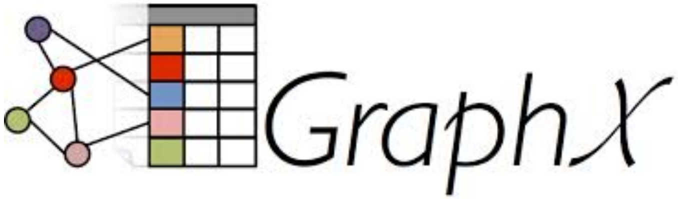
Project 4

- Project 4.1
  - MapReduce Programming Using YARN
- Project 4.2
  - Iterative Programming Using Apache Spark
- Project 4.3
  - Graph Programming Using GraphLab

Graph Computation

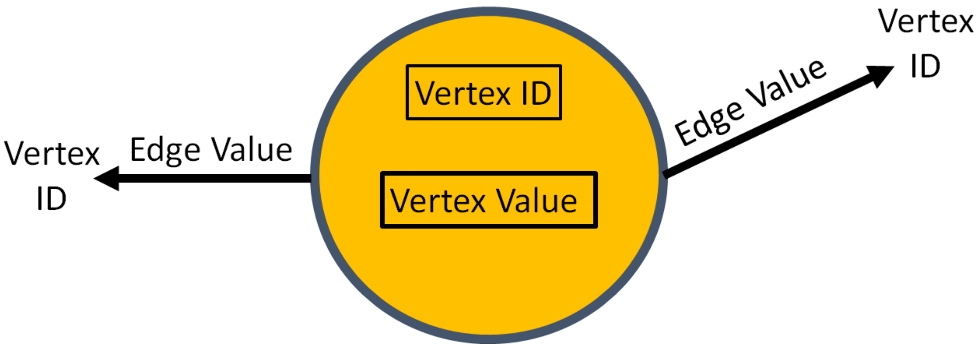
- Some types of data are best expressed using graphs
    - Eg: Social Networks, Transportation Grids…
- There are many computations that can be expressed as graph computations:
    - Eg: PageRank, Traversal, Min Cut/Max Flow etc..
- How about an efficient framework to execute graph-based computation?

Take your pick...



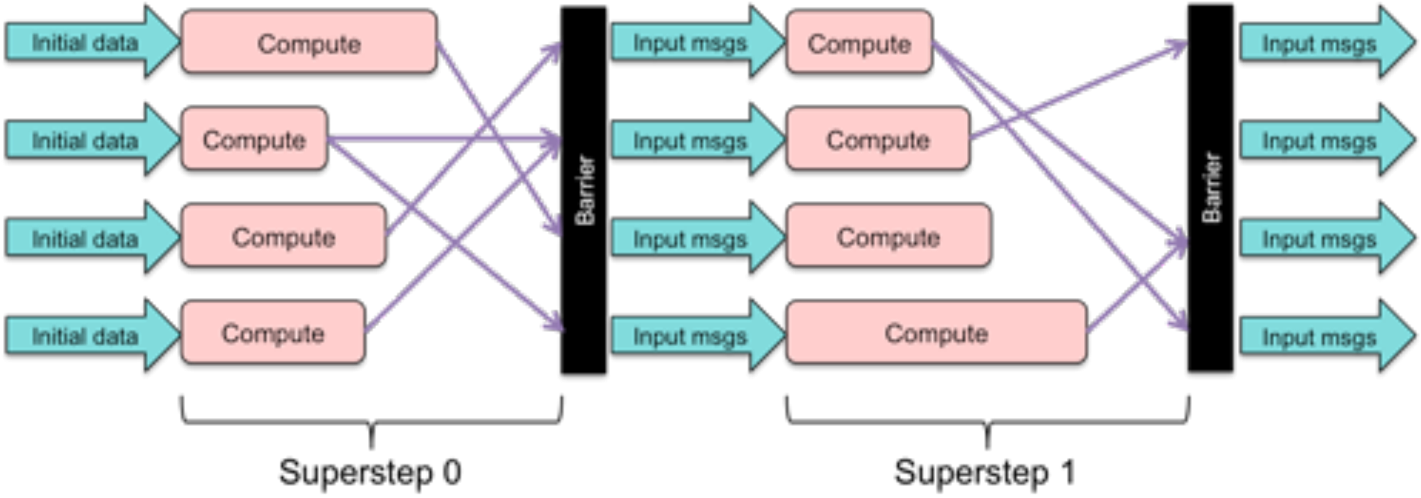
Pregel and Company

- Graph processing framework introduced by Google
- Programs are expressed as operations to be performed on a vertex
- Programs are executed in iterative, bulk-synchronous (lock-step) fashion



Pregel's Performance

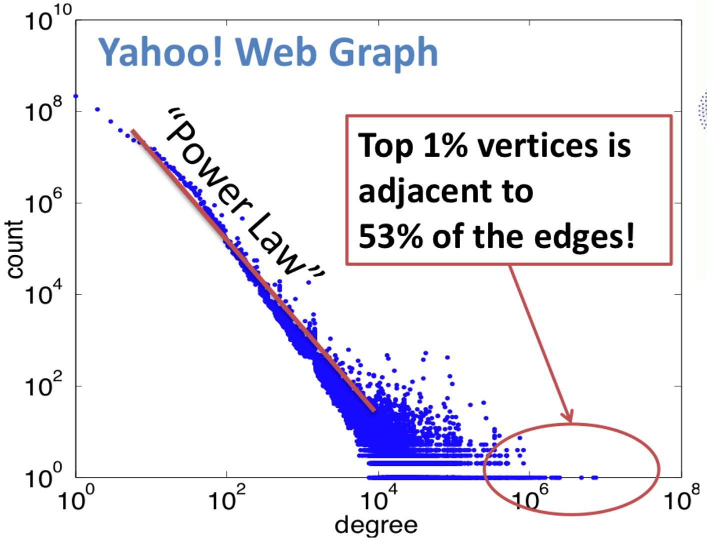
- Synchronous execution can be a performance bottleneck



GraphLab

- Graph processing framework
- Supports both synchronous and asynchronous execution
- Optimized for power-law graphs

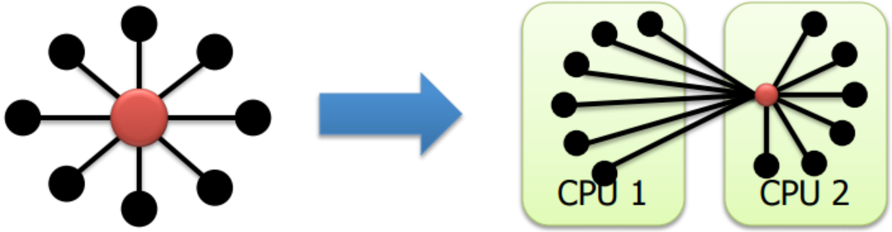
- Natural Graphs:



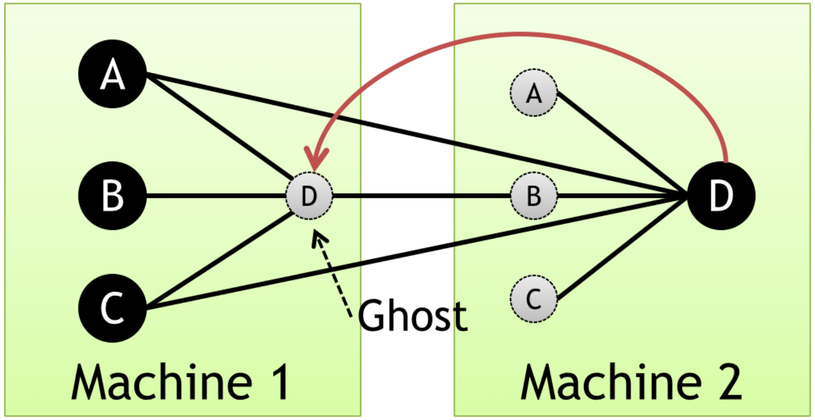
Yahoo! Web Graph

“Power Law”

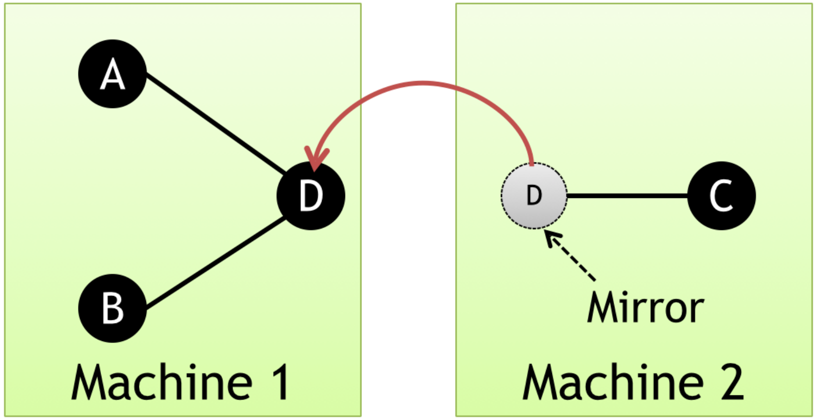
Top 1% vertices is adjacent to 53% of the edges!



Graph Partitioning in GraphLab



Edge Cut

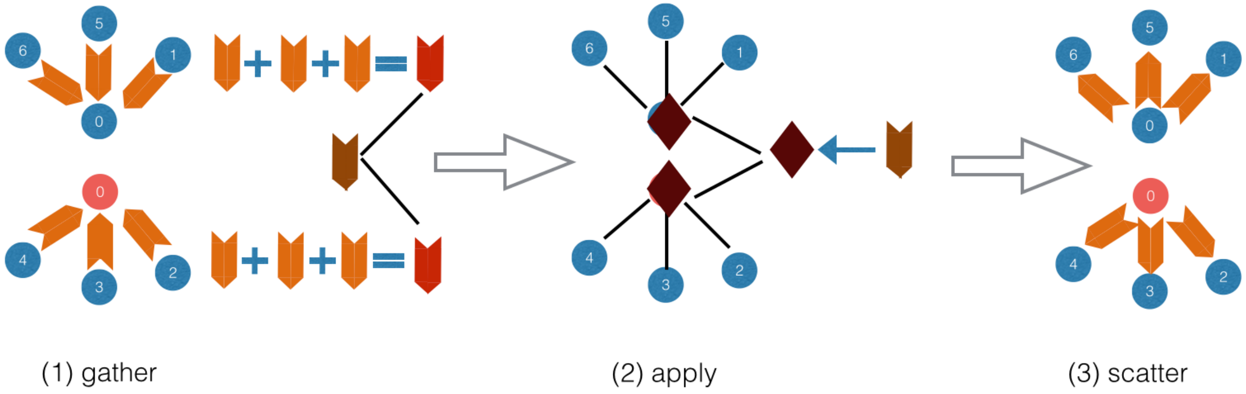


Vertex Cut

- Vertex-cut approach added in GraphLab to handle power-law graphs

How to write a GraphLab program?

- Write three functions that execute on every vertex of a graph
  - Gather
  - Apply
  - Scatter



Example Program - PageRank

$$R[i] = 0.15 + \sum_{j \in \mathrm{Nbr}(i)} w_{\mathrm{ji}} \times R[j]$$

```
GraphLab_PageRank(i)
  // Compute sum over neighbors
  total = 0
  foreach( j in in_neighbors(i)):
    total = total + R[j] * w_ji
```

Gather Information About Neighborhood

```
  // Update the PageRank
  R[i] = 0.1 + total
```

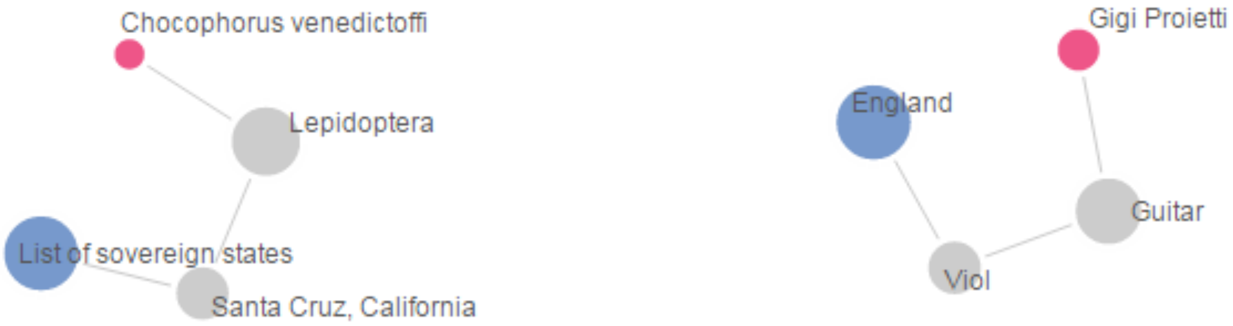
Update Vertex

```
  // Trigger neighbors to run again
  if R[i] not converged then
    foreach( j in out_neighbors(i))
      signal vertex-program on j
```

Signal Neighbors & Modify Edge Data

Project 4.3

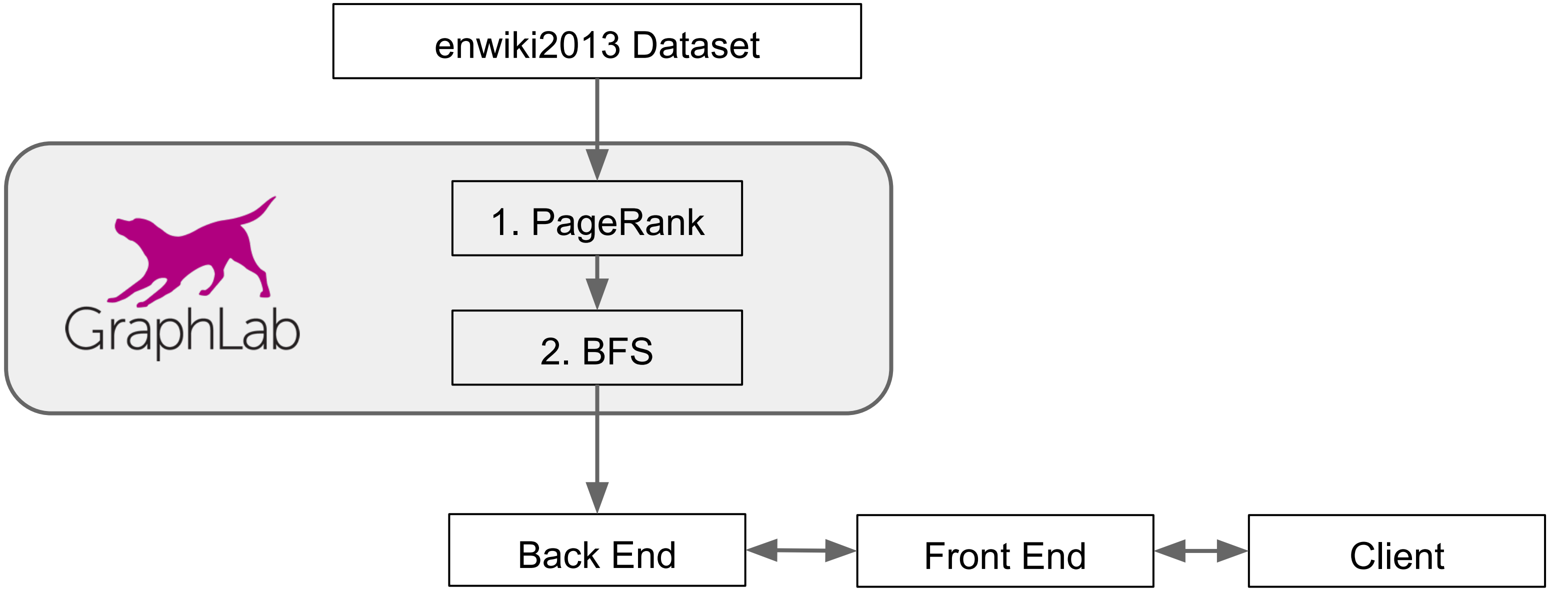
- Find relevant topics that connect two terms
  1. Process the wikipedia graph dataset: PageRank
  2. Breadth First Search (BFS) using GraphLab
  3. Visualize the results



(a) (b)

Project 4.3 - Overview

- Use the enwiki2013 graph dataset
- Run pagerank to find the popular pages
- Find connections between the pages using BFS



Demo

To launch a GraphLab cluster, do the following steps:

- Launch three m3.large instances with **`ami-e697958e`** which has GraphLab installed.
- Upload your key pair file (.pem file) into each instance and create a config file in **`/home/ubuntu/.ssh/`**. Add the following lines into this config file:

```
Host *.compute-1.amazonaws.com
  IdentityFile <path of your .pem file such as~/mykey.pem>
```

Demo cont.

- Log into one instance and create a file called machines in the home directory (**`/home/ubuntu/`**)
- Put the DNS of the current instance in the first line of this file, and the DNSs of another two instances in the subsequent lines.
- Now you can launch your GraphLab applications from this instance!

Upcoming Deadlines

- The end is near...
- Course Survey
  - Due: 11:59PM ET May 01st (Friday)
- Project 4.3
  - Due: 11:59PM ET May 03rd (Sunday)
    - 10% bonus if submitted by Friday

NOT SURE IF LIGHT AT THE END OF TUNNEL
OR A FREIGHT TRAIN COMING MY WAY
quickmeme.com