Algorithms over Spark and Pegasus

Max Nicosia

Exploratory Approach

- Hadoop YARN
- Pegasus
- Spark

Hadoop YARN

- Yet-Another-Resource-Negotiator
- No JobTracker and TaskTracker
- Run directly on "own" ApplicationManger
  - Handles Scheduling, Failures, etc
- More scalable as no single scheduler
- YARN allows any distributed application to run on it

Pegasus

- May run on top of YARN - Need to investigate!
- Several already graph analysis algorithm already built
- Uses HDFS file system

Spark

- Uses Hadoop libraries
  - HDFS

- Can be set up in "stand alone" mode or YARN
  - Investigate!

- Has several algorithms already implemented

Aim

- Investigate the two systems
- Test PageRank on both and compare execution and performance
- Try other algorithms such as community detection / betweeness centrality to evaluate
- See any possibilities for changes or modifications