Astronomy Application of Map-Reduce: A Distributed Friends-of-Friends Algorithm

Bin Fu, Eugene Fink, Julio López, Garth Gibson

**Long-term Goal**

We are developing algorithms and software tools for massive astronomical computations on large computer clusters.

**Initial Results**

We have developed tools for identifying gravitationally bound clusters of galaxies based on the Friends-of-Friends technique:

- Two galaxies are “friends” if they are close to each other; that is, the distance between them is within a specific global threshold
- We analyze an undirected graph, where galaxies are vertices and their “friendships” are edges. We identify the graph’s connected components, which serve as an approximation of gravitationally bound clusters

**Astronomical Datasets**

- Pan-STARRS (began in 2009): Half-order of magnitude larger than Sloan
- Large Synoptic Survey Telescope (to begin in 2016): Order of magnitude larger than Sloan

**Previous Results**

- Researchers have designed fast sequential Friends-of-Friends algorithms:
  - Exact: $O((n \cdot \log n)^{1.5})$
  - Approximate: $O(n)$
- These algorithms however do not scale to massive astronomical surveys

**Distributed Procedure**

We have developed a Map-Reduce “wrapper” that distributes the Friends-of-Friends computation among multiple cores:

- Divide the space into cubes, where each cube includes about the same number of galaxies, by applying the kd-tree construction to a randomly selected subset of galaxies
- Apply a sequential Friends-of-Friends procedure to find the clusters within each cube
- Identify cross-cube “friendships” and merge the respective clusters, using the union-find algorithm

**Performance**

**Strong scalability**

Dependency of the running time on the number of available cores for 500 million galaxies (dashed line) and 1000 million galaxies (dotted line).

**Weak scalability**

Dependency of the running time on the number of available cores, where the input size is proportional to the number of cores. We show results for 4 million galaxies per core (dashed line) and 8 million galaxies per core (solid line).