DISC-Distances: Analyzing the Distribution of Distances Between Galaxies

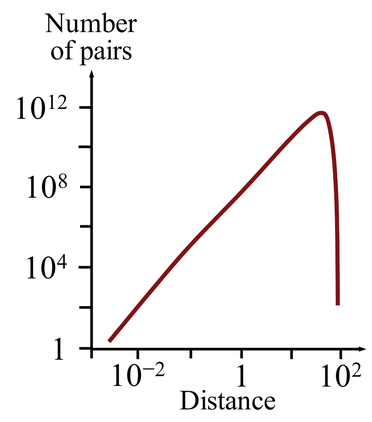
Bin Fu, Eugene Fink, Julio López, Christos Faloutsos, and Garth Gibson

The computation of *correlation functions* is a standard cosmological application for analyzing the distribution of matter in the universe. We have studied several approaches to this problem and developed a distributed approximation procedure based on a combination of these approaches, which scales to massive datasets with tens of billions of galaxies.

Problem

The problem is to build a histogram of all pairwise distances between galaxies. Each bar of the histogram shows the number of galaxy pairs whose distances are within the respective interval. Astrophysicists use such histograms to analyze sky surveys and cosmological simulations. The construction of such histograms is computationally expensive; the complexity of the best exact algorithm is $O(n^{5/3})$.

EXAMPLE: A 577-bar histogram for a cosmological simulation with 4.5 million galaxies in a finite cubic "universe."



Basic Algorithms

NAIVE COUNTING:

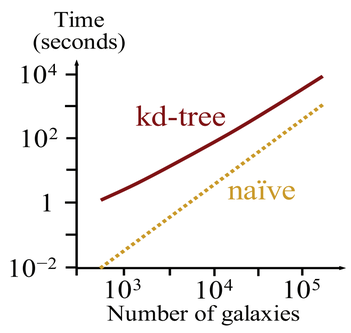
- **Brute-force looping through all galaxy pairs**
- **$O(n^2)$ time complexity with a small constant factor**
- **Time does not depend on number of bars in a histogram**

USING A KD-TREE:

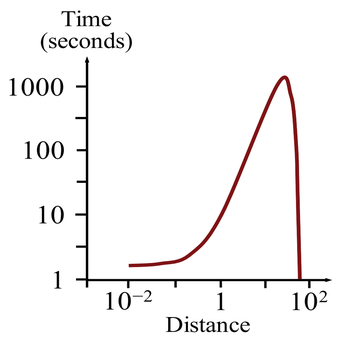
- **Application of the kd-tree indexing to identify galaxy pairs for each given interval of distances**
- **$O(n^{5/3})$ time complexity with a large constant factor**
- **Separate counting for each bar of a histogram; time grows with the number of bars**

The kd-tree algorithm outperforms the naive counting for small and very large distances, but runs slower for the middle of the histogram because of the large constant factor.

EXAMPLE: The time of building a 577-bar histogram.



The naive counting (dotted line) vs. the kd-tree (solid line).

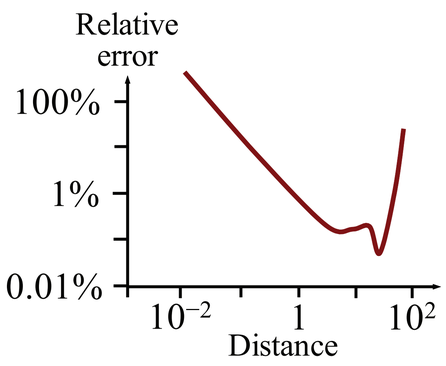


The kd-tree time for each bar of the histogram, on a set of 4.5 million galaxies.

Sampling

- **Select a random sample of galaxies, construct the histogram for the sample, and scale it proportionately to approximate the histogram for the whole set**
- **This approach allows fast approximate computation for arbitrarily large sets of galaxies**
- **The approximation error is inversely proportional to the sample size**
- **The error is higher for small and very large distances, and lower in the middle of the histogram**

EXAMPLE: The dependency of the relative error on the distance for a 577-bar histogram based on a sample of 20 thousand galaxies.

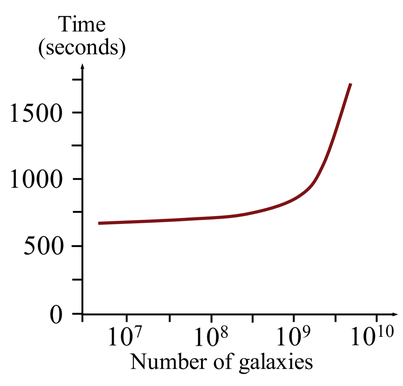


Hybrid Procedure

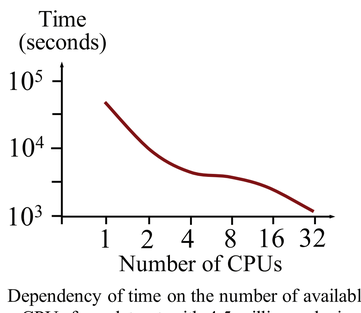
- **Select a random sample of galaxies and apply the naive counting to build a histogram**
- **Select a larger sample and apply the kd-tree to get a more accurate count for small and very large distances. Since the kd-tree is fast at the low and high ends of the histogram, it improves the accuracy without a significant time increase**

Distributed Procedure

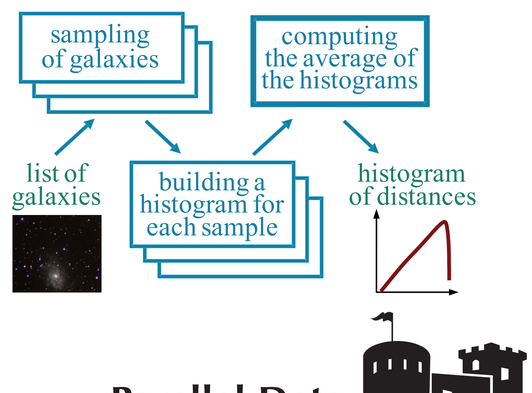
- **Process multiple random samples in parallel and then average the results**
- **The approximation error is inversely proportional to the square root of the number of samples**



Dependency of time on dataset size for the distributed version using 30 CPUs.



Dependency of time on the number of available CPUs for a dataset with 4.5 million galaxies.



Building a histogram with 1% relative error for all bars.

Carnegie Mellon

Parallel Data Laboratory

SVD10