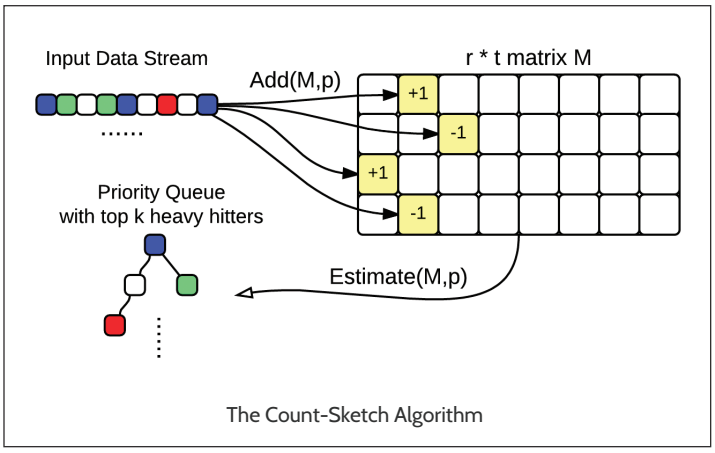
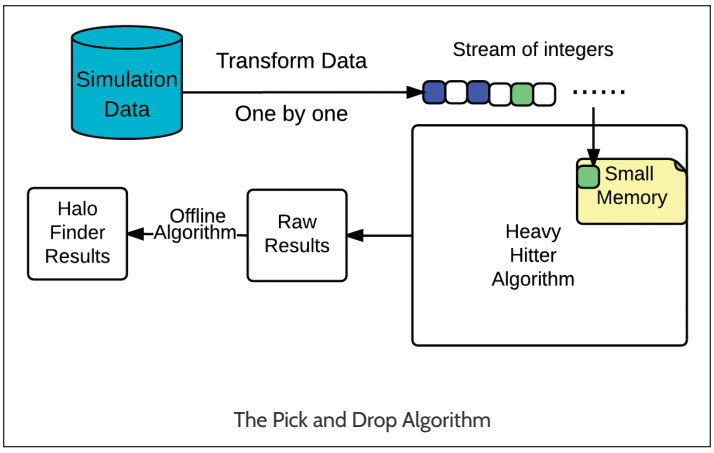
Clustering Algorithms for Data Streams and Related Applications

Cosmologists and Computer Scientists Collaborate on Scalable Algorithms for Big Data

Cosmological simulations are among the largest computer experiments currently run in science, the largest of them currently using over a trillion particles to represent the matter content in synthetic universes that model significant subsets of the observable universe. The traditional way of analyzing their results generally requires machines with memory sufficient to hold complete snapshots of the data. At 10s of Terabytes this can only be achieved on machines similar to the parallel super-computers capable of producing the simulations in the first place and is open to only the handful of specialist programmers of the simulation codes. This, together with the fact that I/O is poses a severe performance bottle neck at these data volumes has meant that the analysis of recent simulations has been lagging behind their production more and more.

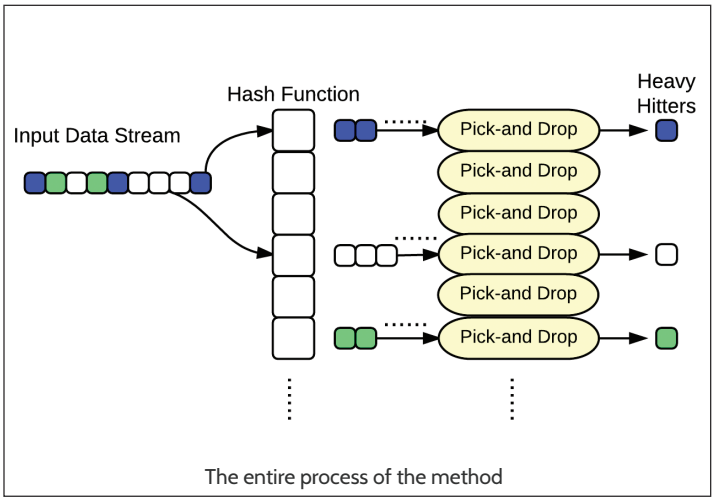


The Count-Sketch Algorithm

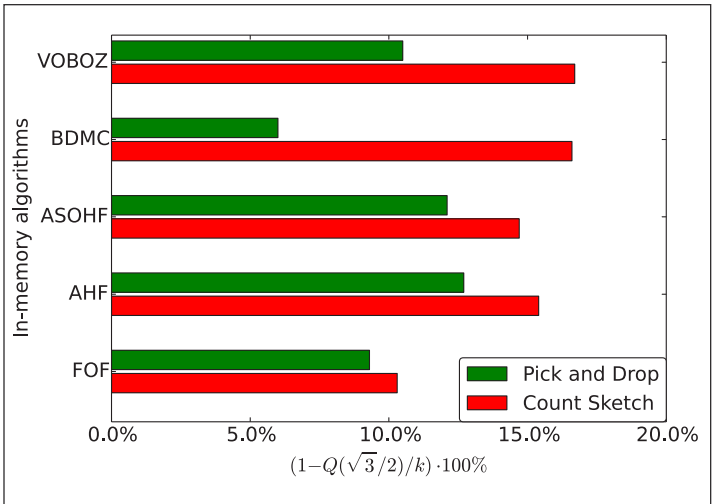


The Pick and Drop Algorithm

One way to address this problem is to try to find algorithms that put lower demands on resources whilst still producing scientifically meaningful results. We have started a collaboration between the CS and Cosmology groups at JHU with the specific aim of analyzing streaming and randomized algorithms. In particular we want to apply these to clustering problems. An important first step in much of the analysis of cosmological simulations is to identify clusters of particles and derive properties of these. Analyzing such cluster catalogues, which generally are 2-3 orders of magnitudes smaller in size than the raw data, is more feasible and on themselves they provide valuable starting points for further analysis and model comparison.



The entire process of the method



Error rate comparison between different algorithms

In a first attempt we have used two algorithms, known as Count-Sketch (Figure 1) and Pick-and-Drop (Figure 2) to find the 1000 most massive clusters. Comparisons to traditional "offline" implementations shows that our randomized streaming approach finds more than 90 percent of all these 1000 most massive clusters, using a small fraction of the memory. This is significant step forward bringing scalable methods of computations to cosmological simulations.

This project is funded by NSF Award # 1447639 - Principal Investigator - Vladimir Braverman; Co-Principal Investigators - Alexander Szalay, Randal Burns, Tamás Budavári, Benjamin Van Durme; Sr. Personnel - Gerard Lemson, Mark Neyrinck.