RESEARCH UPDATE III

10-620 | Independent Study: Research | Andrea Klein

THE DATA

- Recall: I'm working with halos with more than 1000 particles (i.e. the 60,351 halos with masses from 10^12.7 to 10^15.0 Msun).

- Hy has some concerns about the large spread in the velocity distributions. I am concerned because they do not match some of my expectations for a halo like Via Lactea II.

- Some of these effects are likely because the gravitational softening length was too low and the two-body relaxation was too high. Hy is running new simulations that correct these effects, but in the meantime, I am working with the data I have.

- Some of the halos in the catalog have undergone recent mergers. This may blur out their velocity distributions, since you're basically seeing two (or conceivably more) velocity profiles superimposed. I have not tried to correct for this.

- Also, some halos straddle the simulation boundaries. I intend to filter these out (or simply correct the velocities), but meanwhile they are adding a bit of additional noise to the data.

DATA CHUNKING

- Primary issue at the moment (from my perspective): **the particle array is too large to fit in memory,** so it's inefficient to locate the particles that go with a particular halo. This is ok, if suboptimal, for SVR, since I need merely go through once and extract a reasonably-sized set for training. However, it still gets in the way when I want to study individual halos in a particular mass range.

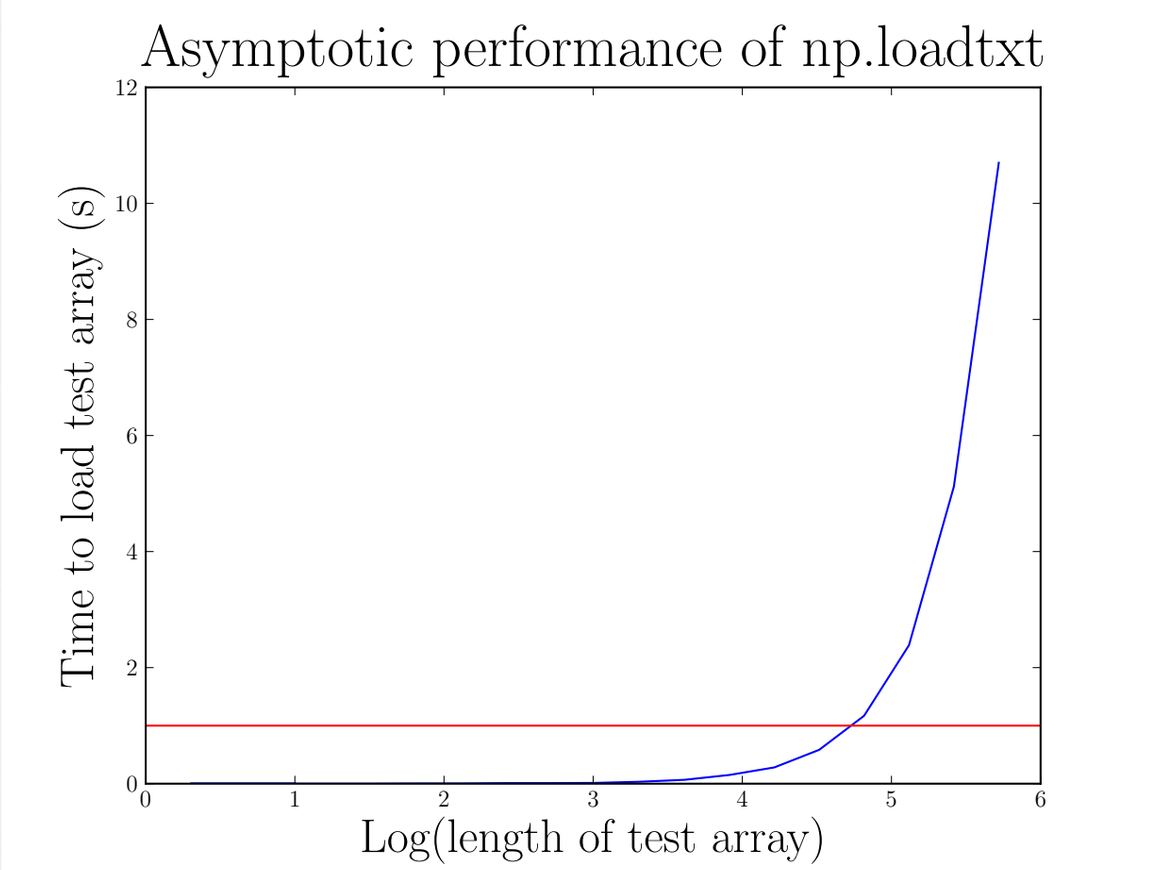
- Dumb solution: divide up the particle data into reasonably-sized "chunks," stored individually in text files that can be loaded into memory one-at-a-time. (No single halo has more particles than can be loaded simultaneously into memory.) When I need particles for a certain halo (or range of halo indices), I can identify the relevant text file(s), temporarily load them, extract the particles, then unload them again.

- I've timed numpy's loadtxt function on test arrays of different lengths in order to determine a good size for the text chunks (see next slide). Note that there's a tradeoff between having more chunks (which can be loaded faster, but the range of interest might straddle several chunks) or fewer. For convenience/speed, I choose the chunk boundaries so that no single halo straddles multiple chunks (so the chunks are not all quite the same size).

DATA CHUNKING

- I tested np.loadtxt() on width-10 arrays of random floats. Each timing is averaged over 3 trials at that array length.

- I don't want loading to take longer than O(1) second, which suggests a max chunk size of about 10^5 (for comparison, the particle data is about 3,000 times that length). In practice I use the number of particles in the largest halo, 10 ^ 5.37, as the upper bound on the chunk size.

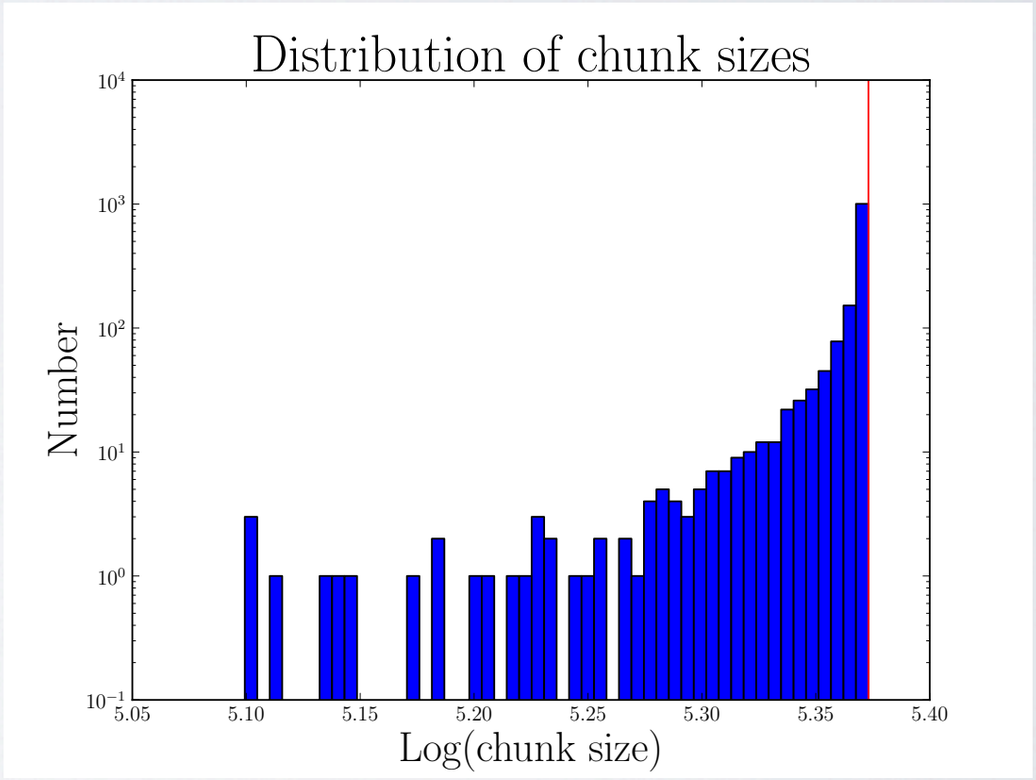


DATA CHUNKING

- Based on these considerations, I divide the halos into 1,463 chunks with sizes from about .5 - 1 times the maximum chunk size (see histogram at right; the red line indicates the maximum chunk size).

- The halos are retrieved by hashing from each halo ID to the name of the file that contains it. The files are also named according to the range of halos they contain, so that if the hashtable were somehow "lost," you could also find the halo relatively quickly by scanning the directory with the chunks.

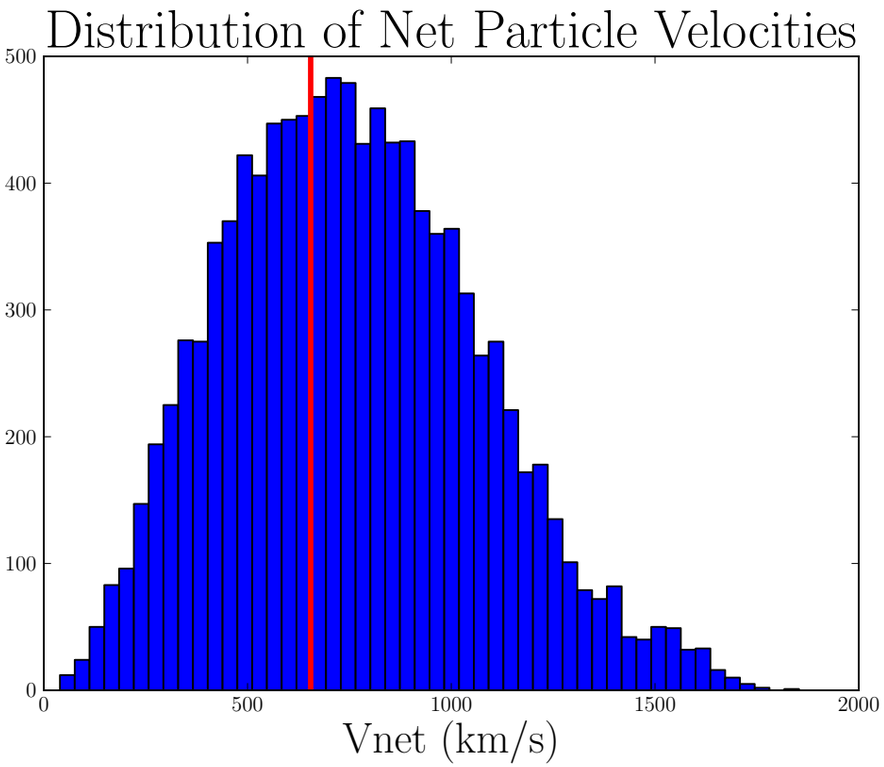
- The net time to load a halo's particles - including finding the right chunk file, selecting by halo ID, and rearranging the columns for indexing purposes - is now about 2-4 seconds regardless of the halo (i.e. no slower for any halo than it is for the first one).

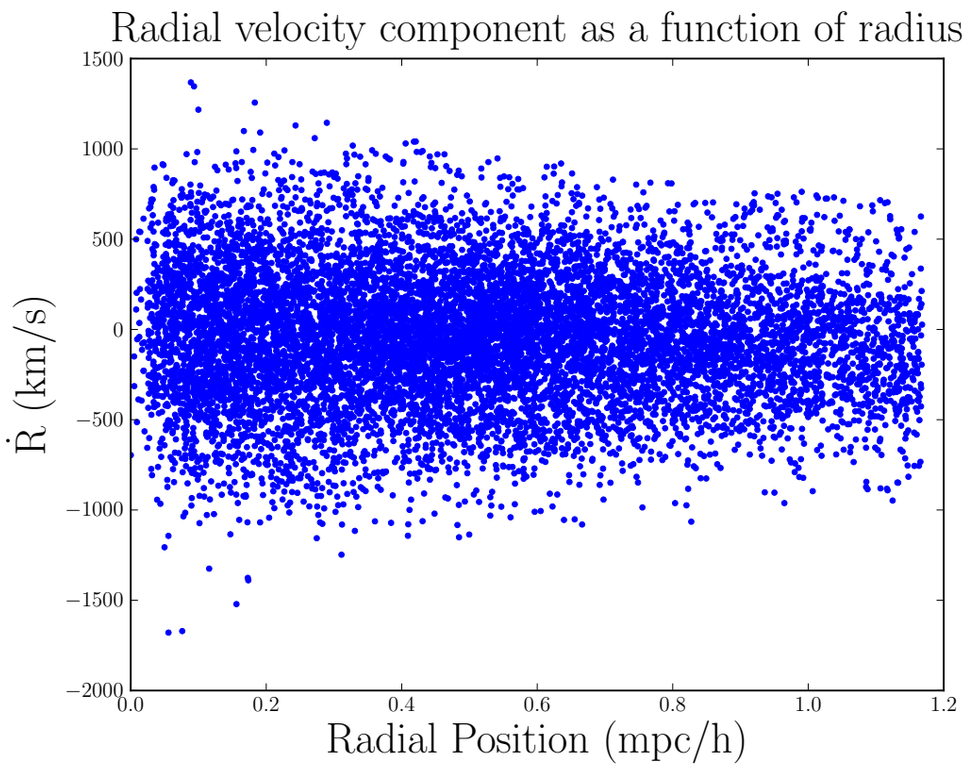


A CLOSER LOOK AT A MID-WEIGHT HALO

Prior to panicking about the unexpectedly large spread in particle velocities (observed in high-mass test halos), I take a closer look at the comparatively medium-sized halo 3145, which has a mass of 10^14 Msun. (Small halos are pending completion of data chunking.)

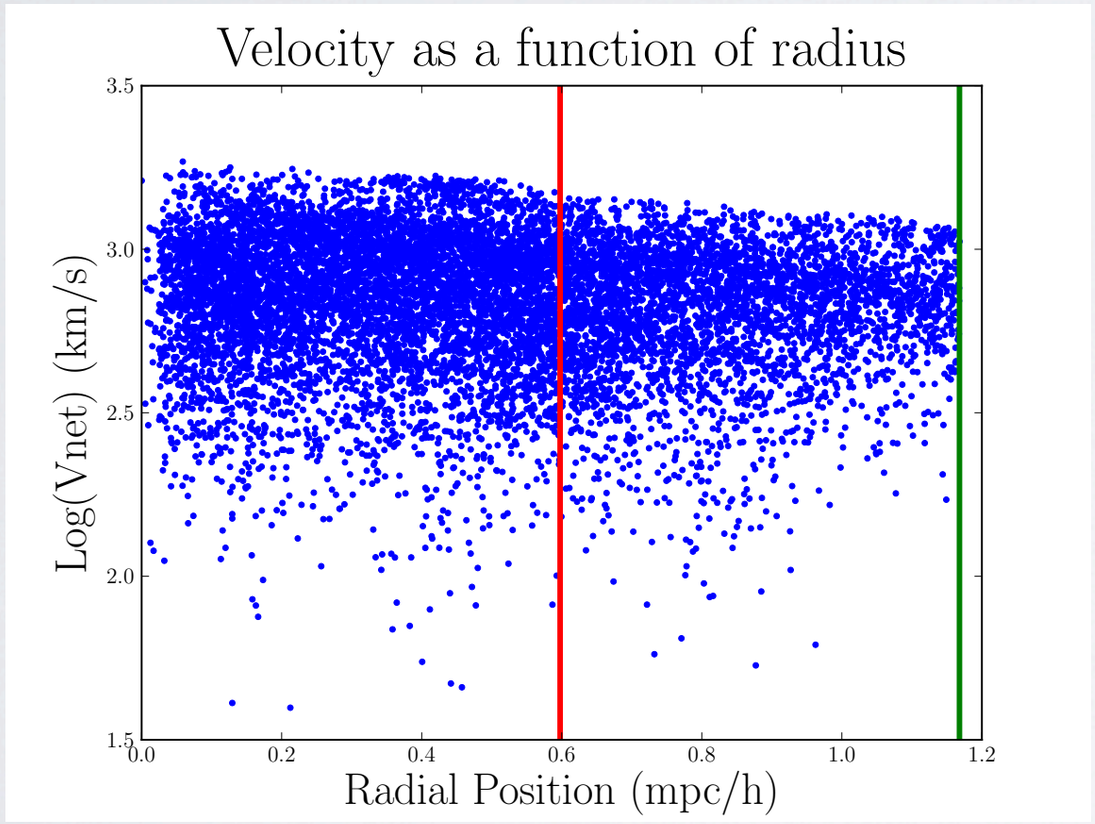
Here are some representative velocity plots (vcmax indicated with red line at left). Overall the scatter still seems fairly high, and there's still not much radial dependence. I intend to re-check these features at 10^13 and 10^12 Msun once chunking is complete. (I currently have about 10,000 halos in chunked form, all from the high-mass end.)

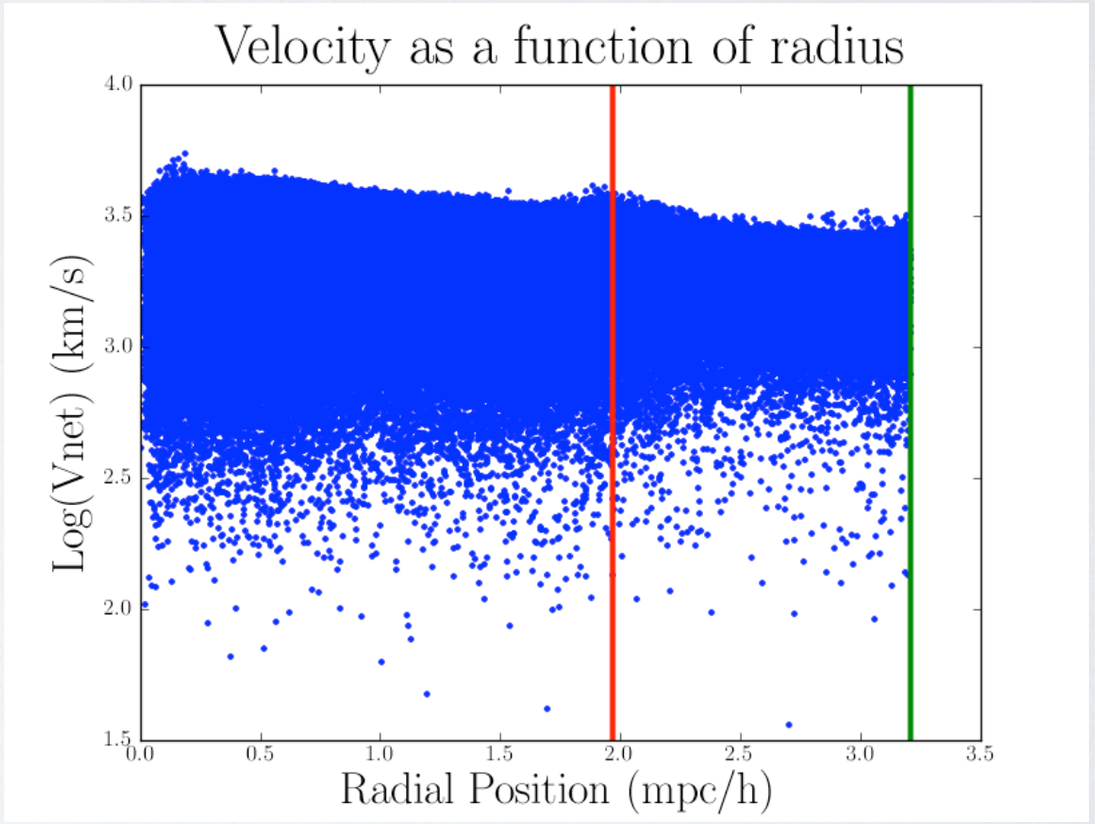




A CLOSER LOOK AT A MID-WEIGHT HALO

Here's a comparison with the most massive halo, which I examined previously (shown at right). Again, similar scatter - just less saturated in the lighter halo. There is, however, a clear offset in the mean value.





Note: rvmax is indicated in red, and R200a is indicated in green.
Axes are different, sorry :(

PRELIMINARY RESULTS

- To start, I am attempting to learn the following function via SVR:
  - [<Mass, vcmax, Mcmax, Rcmax>, (radial position)/Rvir] -> |V_R| (norm of radial velocity)

- My training set consists of a randomly selected set of N particles (I'll be working with powers of ten and making use of the new particle fetching system). All input parameters have been normalized to a scale from 0-1.

- I do not necessarily expect the radial position to be very informative, especially for very massive halos with large velocity scatter; however, I expect to be able to learn something like the mean and scatter in the velocities for those halos, and perhaps a better-defined distribution for lighter (VLII-scale) halos.