Analyzing the Evolution of Large Scale Structures in the Universe with Velocity Based Methods.

Category: Application

ABSTRACT

Multistreaming events are of great interest to astrophysics because they are associated with the formation of large scale structures (LSS) such as halos, filaments and sheets. Until recently, these structures were studied using scalar density field only. This is an application paper where we applied appropriate data analysis and visualization techniques to cosmological simulations. The original problem that was presented to us by the cosmologist we worked with was whether it is possible to find multistreaming regions based on velocity information only. Compared to the current practice of using density information to find these regions (e.g. with halo finders), we show that the velocity based methods that we propose produce good agreement in general. More interestingly, there are regions that are found by halo finders but not by velocity based methods, as well as regions that are found by velocity based methods but not by halo finders. Further analysis of our results indicate that not all halos are interesting, while there are some interesting regions that are missed by halo finders. In addition, the velocity based feature extractors show dynamical behavior not possible with halo finders. Thus, while we started with the simple question of finding multistreaming regions using velocity information, we are re-examining the relationship between halos and multistreaming regions, and discovering new properties of these regions that can be used for classification purposes.

1 INTRODUCTION

Over the last two decades cosmology has made extremely rapid progress. There now exists a cosmological "Standard Model" that is in very good agreement with a large number of observational datasets at better than the $5-10\%$ level of accuracy. A key feature of the model is the existence of a "dark" sector that is not directly observable by emission or absorption of light but may be inferred via effects such as gravitational lensing and by its dynamical effects, especially in the formation of cosmic structure. Observations indicate that 70% of the Universe consists of a mysterious dark energy, 25% of a yet unidentified cold dark matter (CDM), and only 0.4% of the remaining 5% of ordinary (atomic) matter is visible [12]. Understanding the physics of the dark sector is the foremost challenge in cosmology today.

The evolution and dynamics of the dark matter distribution can be investigated by following the formation of LSS as observed in the distribution of galaxies today, and in the past. LSS are often characterized by the dimensionality of their spatial distribution e.g. galaxy clusters (0D), filaments (1D), and surface-like pancakes (2D). A hint of the complex geometry and topology of cosmic structure is illustrated in Figure 1.

Precision dark matter simulations are a key foundation of cosmological studies. These simulations track the evolution of the dark matter with very high resolution in time, force, and mass. At the scales of interest to structure formation, a Newtonian approximation in an expanding universe is sufficient to describe gravitational dynamics. The evolution is given by a collisionless Vlasov-Poisson equation [6], a six-dimensional partial differential equation. This is solved using an N-body approach. The six-dimensional phase space distribution is sampled by "tracer" particles and these particles are evolved by computing the inter-particle gravitational forces.

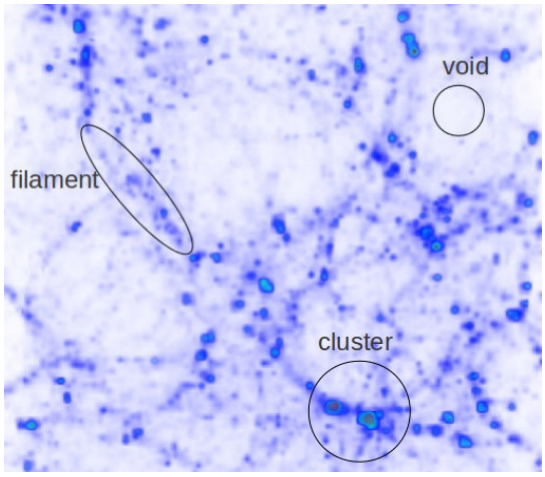

Figure 1: Large scale cosmological structures of the universe.

The starting point of the simulations is a Gaussian random density field which imprints small perturbations on a uniform density, isotropic universe. The simulations start in the linear regime of the density fluctuations which then evolve under the influence of gravity. At any given length scale, during the early stages, the evolution remains linear but as time progresses, evolution first enters the quasi-linear regime (where perturbation theory can be applied) before finally reaching the fully nonlinear regime at which point all analytic descriptions break down. There is substantial interest in determining and characterizing the transitions between linear, quasi-linear, and nonlinear dynamics in the simulations by tracking the dynamics of dark matter tracer particles. At the start of the simulation, the velocity dispersion is initially zero, and the phase-space distribution is a three-dimensional sub-manifold of the phase space (only one velocity direction at a given spatial point). As the 3-hypersurface evolves, it folds, leading to the occurrence of singularities in the density field corresponding to the appearance of regions with multistream flow.

Finding multistreaming regions in cosmological simulations is an important endeavor for several reasons. The onset of multistreaming and the evolution of multistreaming regions as part of the theory of nonlinear structure formation is certainly interesting in of itself. Additionally, it is becoming an increasingly important aspect in understanding the formation of galaxy clusters where several "cold flows" combine. Different cosmological models and theories of structure formation will make different predictions for multistreaming.

The determination of the onset of multistreaming with respect to time and length scale is important in predicting the validity of approximate methods such as perturbation theory. Since running large cosmological simulations is very costly, cosmologists are always searching for methods that provide accurate answers at certain scales that do not require expensive simulations. For example, con-

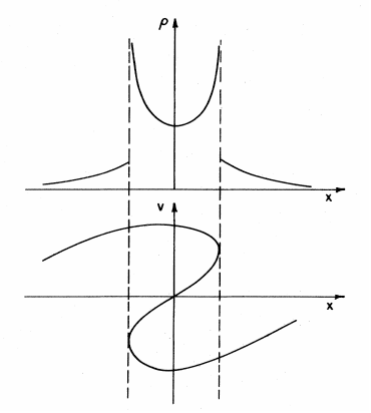

Figure 2: Illustration of 1-D multi-stream flow. Top panel: High density region with three-stream flow confined between the dashed lines. Bottom panel: The corresponding phase space plot showing the different stream regions [15].

sider a key cosmological statistic measured from simulations: the density fluctuation power spectrum. This *power spectrum* can be predicted over a range of (large) length scales by perturbation theory. Multistreaming, however, cannot be described within perturbation theory. Thus, it is important to study the relationship of the breakdown of perturbation theory and the onset of multistreaming.

Finally, a robust method to capture the onset of multistreaming across multiple scales will help to set the initial cosmological time for starting cosmological simulations. The initial conditions for cosmological simulations are based on the Zel'dovich approximation, which is only valid if the paths of tracer particles do not cross (i.e., before multistreaming). Therefore, the simulations have to be started sufficiently before the occurrence of multistreaming events in order to guarantee accurate results.

Traditionally, LSS is investigated primarily by considering the distribution of dark halos. Although there are differences between methods, halos are typically identified by thresholding on the density of tracer particles (for a description of halo finders, see, e.g., [10]). In this paper, we are concerned not so much with density, but how the velocity information of tracer particles can find and characterize multistreaming regions.

2 Multistreaming

What is multistreaming? Unfortunately, there is no single precise mathematical description of multistreaming. As such, the problem is reminiscent of finding vortex core lines in vector field analysis. What do exist in literature are phenomenological descriptions of multistreaming events. In this paper, we derive several velocity based multistreaming extractors based on such descriptions.

Multistreaming is said to occur when there are multiple velocities at a given spatial point. A simple example is illustrated in Figure 2 for a one-dimensional cold and collisionless medium [15]. In the phase space plot (bottom panel), the boundary between a three-stream flow and a single stream is denoted by the dashed lines. At the boundaries, there is a shell-crossing singularity (caustic) in the density field (ρ) because the mapping from phase space to physical space becomes multi-valued. This picture generalizes to higher dimensions.

One can also find additional clues for finding multistreaming from the following description — "If the dark matter is a cold, collisionless fluid, then at any given spatial point, at early times, there is a unique fluid velocity. However, as evolution proceeds, the map connecting initial to final positions develops singularities (caustics) corresponding to multiple flow directions at a given spatial point. Regions of multistream flow form, and even though each stream is irrotational (curl-free), the velocity field is no longer a potential flow. Because of the large density of particles near caustics and other dynamical complexities associated with multistreaming, it is expected that perturbative methods will tend to break down in these regions" [13]. This description suggests additional avenues for finding multistreaming events via velocity based analyses. For example, we can look for regions where the flow is irrotational, examine the divergence field to see where particles may possibly congregrate, examine the linearity of the flow field, check similarity of velocities as well as velocity dispersion. The following flow behaviors may also account for multistreaming: (i) particle flows have different speed and direction, or (ii) particles flows have the same speed but different direction, or (iii) particles flows have different speeds but the same direction. So, checking the shear in the flow may provide some information as well. We explore these in Section 5.

3 Previous Work

The visualization of cosmological data sets has received significant attention. Most cosmological simulations are particle-based. The size of these simulations, measured by the number of particles, have increased with better computing resources, allowing us to capture physical phenomenon at a much wider range of length scales.

Within the visualization community, there have been several works focusing on astrophysics data sets. A small subset of these include works by Li et al. [9] which explored how to display positional and trajectory uncertainties in astrophysical data sets; and Fraedrich et al. [5] focused on scalable rendering of large cosmological simulations using a combination of hierarchical level-of-detail approach and GPU accelerations. While these works have studied the issues related to visualizing cosmological simulation data, they are different from the work in this paper in that we are primarily interested in feature extraction of multistreaming events in such data sets.

Multistreaming events have been explored in recent years. For example, Yano et al. [18] investigated the distribution of caustics in the expanding Universe. In this work, the model describe continuous matter density fields, such as singularities of density field or density perturbations. Regions demarcated by high density contrast are associated with multistreaming and results in structures such as halos. The density contrast is defined as, for a given time and region, how much does the density changes with respect to mean density. Depending upon the types of the Universe simulated or the halo structure of interests, such as inner halo parts or halo boundary, the minimum value of density contrast varies. One of the common approaches for finding halos uses the Friends-Of-Friends (FOF) group finder [4]. The basic idea is that given a simulation with N particles with a fixed volume, the average inter-particle spacing is first calculated. Then, pairs of particles that are closer than some fraction of the average inter-particle spacing are linked together, resulting in a network of linked particles. We compare our results to the FOF halo finder implemented in ParaView 3.10 [17]. An extensive survey of other halo finders can be found in [1].

While there are a number of density based analysis similar to [18], there are much fewer works based on the analysis of the velocity field. The concept of phase space plots shown in Figure 2 easily extends from 1-D to 3-D, but using this approach to detect multistreaming in 3-D has not been fruitful so far.

In flows with high shear, one can observe that locally, particles will move at drastically different rates and direction. One way to detect such regions in 3-D is to form tetrahedral volume elements from 4 nearby particles. By tracking these tetrahedral volume elements over time, multistreaming regions can be detected when there is a change in the sign of the tetrahedral volume. This happens be-

cause some members of the tetrahedral volume moved in such a manner to penetrate through the tetrahedral shape causing its volume to flip sign [3]. However, this approach tends to be more expensive since it requires multiple time frames of large data sets.

More recently, Shandarin [14] proposed a new approach to identify the cosmic web based on finding multistreaming flows. Instead of relying solely on the density of particles, Shandarin's technique incorporates velocity information of particles along with their positional information. He used the local velocity variance to identify multistreaming events. Prior to his work, we have also worked with particle velocity information to identify multistreaming events. In that work, our analysis was based on simulations consisting of 64^3 and 256^3 particles. The analysis and results reported in this paper are based on simulations consisting of 512^3 particles within a box that is 256 h^{-1}Mpc along each side. Higher resolution data sets allow us to resolve multistreaming regions at a wider range of length scales.

4 Time and Scale Dependent Thresholds

Our approach to finding velocity based multistreaming regions assumes a continuous velocity field is available. There are a number of options available for converting the discrete particle velocity information into a gridded velocity format where we can assume some form of continuity. These options range from the simple nearest-grid-point (NGP) method that assigns particle velocities to the nearest grid point, to more sophisticated methods such as those that use radial basis functions to provide a smooth velocity field. NGP has some drawbacks such as abrupt changes between nearby grid points, while more sophisticated methods are also more expensive as the number of particles and spatial resolution increase. In this paper, we use the cloud-in-cell (CIC) method to generate a velocity field from the particle velocity. CIC [7] uses a weight factor to account for the distance of the particle to its closest grid points. That is, the velocity of each particle is distributed, using a distance based weight factor, amongst the grid points of the cell containing the particle. This method is a good compromise in terms of speed and smoothness of the resulting field. It is also the same method used in the simulation code to resolve the influence of the gravitational field on the particles.

The choice of grid resolution is quite important. If the grid is too coarse, the resampling process will smooth out the data too much and we may miss the multistreaming event. On the other hand, if the grid is too fine, it would result in a low particle count and confidence, not to mention the extra computational expense. The main requirement is that the grid size has to be small enough to resolve the features of interest at certain length scales. For our investigation, we choose a grid resolution such that on average there are 64 particles contributing to each grid point. The demonstration data set provided to us for this investigation contained 512^3 particles. The simulation consisted of 499 time frames from redshift Z=50 to Z=0, covering a box domain measuring 256 h^{-1}Mpc per side, and with the entire data set stored in 2 terabytes. To achieve the desired average particle density per cell, we used a regular grid with 256^3 cells for converting the particle velocities into a continuous velocity field. Note that there are much larger data sets such as the Bolshoi simulation [8] where our proposed techniques can be applied later on. As we explain later, this grid size also allows us to find multistreaming regions early on in the evolution. At the start of the simulation, each grid cell contains 8 particles on average. Therefore, the 8 cells sharing a grid point contain 64 particles on average. The simulation uses periodic boundaries. Note that as time progresses, some regions become more dense while others become more sparse or even empty. Empty cells as well as those in their immediate vicinity must be treated with care and are specially marked so that they do not produce erroneous results in the analysis.

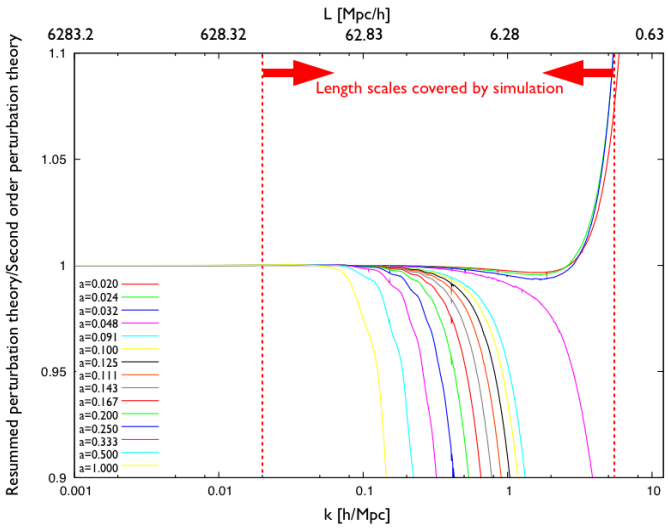

Figure 3: Breakdown of perturbation theory at different scale factors (a) and different length scales (L). The curves are the ratio of two different perturbation theories. As the ratio deviates from one, perturbation theory is not valid anymore. Each curve shows the result for one time snapshot. At the top of the plot we indicate length scales, at the bottom we indicate wave numbers. The dashed lines in red show the scales that can be resolved by the simulation data.

One may also wonder whether particles actually collide in such simulations. To determine this, we examined the most dense cell at a later time in the simulation, and found that the average distance between any two particles is 0.09 h^{-1}Mpc. We also searched for the two closest particles and the distance between them was 0.0018 h^{-1}Mpc, about 50 times smaller than the average distance. Since these tracer particles do not actually collide, although the "streams" around them conceptually do, we perform our analysis using a single velocity field.

Based on previous studies, we know that multistreaming happens at different scales and increases over time. Initially, small multistreaming regions form, which later coalesce in a complex manner into larger multistreaming regions. Hence, an important parameter in searching for multistreaming regions is estimating the length scales for different times. In order to do this, we examine when perturbation theory fails. The perturbative treatment of gravitational clustering should break down in regions where multistreaming events occur. To predict these events, we make the following simple argument based on an internal check within the perturbative analysis. To do this, we note that perturbation theory can be carried out at different orders in the density perturbation. In the regimes where perturbation theory works, higher-order corrections serve to improve the lower-order results. However, once the fluctuations are too large, consistency between orders no longer exists, and different order results can disagree strongly. By investigating at what scales two different approaches at different orders diverge from each other, we can estimate the scale where perturbation theory fails, and hence produce a candidate scale for the onset of multistreaming.

Following Carlson et al. [2], we calculate the matter power spectrum for second order perturbation theory and a re-summed scheme with a code provided by the authors. We then take the ratio of these power spectra at different epochs. The results are shown in Figure 3 for scale factors between $a = 0.02$ and $a = 1.0$. Note that redshift is related to the time dependent scale factor a. An estimate of when and at what length scales multistreaming will occur can be

a	Frame #	Redshift Z	L scale 10%	L scale 5%
0.500	248	1	37 h^{-1}Mpc	43 h^{-1}Mpc
0.333	165	2	30 h^{-1}Mpc	34 h^{-1}Mpc
0.250	123	3	24 h^{-1}Mpc	27 h^{-1}Mpc
0.200	95	4	18 h^{-1}Mpc	24 h^{-1}Mpc
0.167	80	5	14 h^{-1}Mpc	19 h^{-1}Mpc
0.111	70	6	10 h^{-1}Mpc	12 h^{-1}Mpc
0.125	60	7	8 h^{-1}Mpc	10 h^{-1}Mpc
0.111	52	8	6 h^{-1}Mpc	8 h^{-1}Mpc
0.100	47	9	5.8 h^{-1}Mpc	7 h^{-1}Mpc
0.091	43	10	5.6 h^{-1}Mpc	6 h^{-1}Mpc
0.045	21	20	2 h^{-1}Mpc	3 h^{-1}Mpc
0.033	13	30	1 h^{-1}Mpc	1.05 h^{-1}Mpc
0.020	7	50	0.9 h^{-1}Mpc	0.98 h^{-1}Mpc

Table 1: This table shows the relationship between the scale factor a and the frame number of the simulation. It also shows length scale for two different tolerances at which perturbation theory breaks down. The tolerances are at 10 and 5 percent from the ratio of one between the two perturbation calculations seen in Figure 3. When choosing the grid size for calculating the continuous fields it is important that the smallest length scale of interest is resolved. For example, with redshift 30 at a tolerance of 10 percent, the scales of interest are at 1 h^{-1}Mpc. With a box size of 256 h^{-1}Mpc the grid size has to be at least 256^3 to resolve these scales. If the grid is coarser, the length scale that can be resolved increases and therefore multistreaming events could only be resolved at a later time step. To determine the thresholds for the different methods described in the next section, we use a tolerance of 10 percent, which has a finer grid requirement than using a tolerance of 5 percent.

obtained by measuring the scales at which the curves deviate from unity in Figure 3. The figure indicates that these scales vary with time. Multistreaming regions that are relevant to the breakdown of perturbation theory start out as small structures which grow bigger over time. The dashed line on the right indicates the resolution limits due to smoothing from the density calculation. It can be easily varied by reducing or increasing the grid size for the CIC (an increase moves the cutoff lower and a reduction moves it higher), although one cannot increase it beyond a certain point set by particle spacing limits in the simulation. For the data set being presented in this paper, the smallest wave-number ($k = 2\pi/L$) is k $\approx$ 0.02 h^{-1}Mpc, and the corresponding smallest length scales we can resolve is 0.256 h^{-1} Mpc. Using a grid of 256^3 cells, we can resolve length scales of 1 h^{-1} Mpc. However, when coupled with CIC with window size equal to one cell, our resolution drops to length scales of 2 h^{-1}Mpc.

Table 1 is created based on the predictions from Figure 3. It lists the expected size of the multistreaming scales for different snapshots in the simulation data. The time stepping unit is measured with respect to the scale factor a. Given that there are 500 time steps in the simulation, $\Delta a = 0.002$ from one frame to the next. In short, this table provides us with information as to the timing and length scale of multistreaming. The next section focuses on finding their location.

This table is instrumental in determining the threshold values used by the different feature extractors. As can be seen in this table, multistreaming regions grow over time. We therefore use the information from Table 1 to guide us in finding a time-varying threshold appropriate for the epoch in the simulation. For example, if we are searching for regions of interest at frame 250, we expect these regions to have length scale of about 37 h^{-1}Mpc. Therefore, we want to find a threshold value that will produce regions of this expected size. Since the regions may come in a variety of shapes, and because the length scale itself does not fully capture shape information, we use it as an indicator of a region size rather than a strict length scale. In this regard, region size is taken to mean the number of connected grid points that are above the current threshold. To determine the appropriate threshold for a given frame, the initial threshold $threshold_0$, is set to a value that will result in all points being classified as multistreaming according to the feature extractor. We then adjust the current threshold by a small amount, which is some fraction of the range of values for the particular feature extractor, and restart the scanning process. Once we find at least one region with the expected feature size, we finalize the threshold value for that frame. Because the growth of region size is fairly well behaved, we can use the final threshold value of the current frame as the initial guess for the next frame.

5 Velocity Based Extractors

In this section, we present six velocity based extractors to detect multistreaming. They are based on the phenomenological descriptions of multistreaming found in astrophysics literature. Since our science question is whether one can find such events using velocity information instead of density information, we used fairly straightforward formulations using just local information.

5.1 Maximum Shear Stress

Particles going in opposite directions or even in the same direction but at different speeds lead to shear in the velocity field. We hypothesize that shear in the velocity field can be one of the mechanisms for multistreaming. This method should be able to catch two sheets moving in opposite directions. To find the maximum shear stress, we first calculate the velocity gradient tensor of the velocity field, then find its symmetric tensor component, and the associated eigenvalues λ_1, λ_2, and λ_3. We use the von Mises criterion [16] for maximum shear stress which is defined as:

$$MS = \sqrt{\frac{(\lambda_1-\lambda_2)^2+(\lambda_1-\lambda_3)^2+(\lambda_2-\lambda_3)^2}{2}} \quad (1)$$

Note that as flows become isotropic i.e. $\lambda_1 = \lambda_2 = \lambda_3$, the maximum shear stress goes to zero. So, this particular feature detector looks for regions that exhibit high shear as indicated by highly anisotropic regions. Note that other types of anisotropic measures could possibly be used in place of the von Mises criterion.

5.2 Divergence

Divergence is a scalar quantity that measures the degree to which a vector field is a source or a sink at a given location. Positive values indicate a source-like behavior, while negative values indicate a sink-like behavior. The motivation for using divergence for finding multistreaming is that it finds regions where particles congregate, as in caustics. The more negative the divergence value, the stronger that region attracts nearby particles. Given a vector field $\vec{V} = (V_x, V_y, V_z)$ and operator $\nabla = (\frac{\partial}{\partial x}, \frac{\partial}{\partial y}, \frac{\partial}{\partial z})$, divergence is defined as

$$\nabla \cdot \vec{V} = \frac{\partial V_x}{\partial x} + \frac{\partial V_y}{\partial y} + \frac{\partial V_z}{\partial z} \quad (2)$$

5.3 Vorticity

In fluid dynamics, the rotation of vector field is well studied, and is called vorticity. It determines the tendency of an object to rotate at a given location (x,y,z). The vorticity at a point is a vector and is defined as the curl of the velocity field. Given a vector field $\vec{V} = (V_x, V_y, V_z)$ and the operator $\nabla = (\frac{\partial}{\partial x}, \frac{\partial}{\partial y}, \frac{\partial}{\partial z})$, the curl is

$$\nabla \times \vec{V} = \left(\frac{\partial V_z}{\partial y} - \frac{\partial V_y}{\partial z}\right) i + \left(\frac{\partial V_x}{\partial z} - \frac{\partial V_z}{\partial x}\right) j + \left(\frac{\partial V_y}{\partial x} - \frac{\partial V_x}{\partial y}\right) k \quad (3)$$

Since multistreaming regions are suppose to remain curl-free (irrotational), this metric provides an indication of how well this condition is satisfied. Regions of interest are those with very small rotational motions and not their particular orientations. Therefore, the key variable is the vorticity magnitude.

5.4 Dot Product

Particles inside multistreaming regions have different velocities. We can measure the degree to which a set of vectors are similar or different using dot products. Given two vectors: $\vec{V} = (v_1, v_2, ..., v_n)$ and $\vec{U} = (u_1, u_2, ..., u_n)$, the dot product is defined as

$$\vec{V} \cdot \vec{U} = \sum_{i=1}^{n} v_i \cdot u_i \tag{4}$$

Note that this term measures similarity of both vector directions and magnitudes. For this metric, we are primarily interested in similarity of directions. Hence, the vectors should be normalized first. This normalized dot product measures the angular difference between pairs of vectors. The range of this metric is [-1..1] corresponding to 180 degrees (opposite direction) to 0 degrees (same direction). Since the maximum shear metric already accounts for situations where vectors are going in the same direction with different speeds or opposite directions with the same speeds, we tailor the normalized dot product metric to find only those regions where vectors are crossing each other at large angles. Specifically, we use the absolute value of the normalized dot product. Values closer to zero therefore indicate regions with crossing vectors.

To calculate this metric, we first calculate the average normalized velocity of all the particles in neighboring cells that share a grid vertex. We then calculate the dot product of each of the normalized particle velocity against the vertex velocity. These dot products are finally averaged together and represent the directional similarity among the particles in the vicinity of the grid vertex.

5.5 Variance

Variance is a measure of how different and spread out a set of numbers are from each other. Velocity variance then measures the spread of velocities. Since multistreaming regions are characterized as having different velocities (also referred to as the velocity dispersion property in literature), velocity variance is intuitively a good measure for finding these regions. Shandarin [14] also used velocity variance in his analysis, but our formulation differs slightly. We iterate over particles, while Shandarin iterates over flows.

Given n numbers $x_1, x_2, ..x_n$ and a mean μ, the variance σ^2 is defined as

$$\sigma^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu)^2 \tag{5}$$

The CIC weighted velocity at each grid point is μ. While the velocities of the particles inside the 8 cells containing the grid point are represented by x_i. Since we are interested in velocity variance in 3D, the calculation is performed for each velocity component. The variance extends to a symmetric covariance matrix where the diagonals are the variance of each velocity component. Treating each component as an independent random variable, the net velocity variance is simply the sum of the diagonals. If this sum is high it indicates high velocity variance. Unlike the normalized dot product measure described earlier, this measure captures the variance of both the direction and magnitude of the velocity field.

5.6 Linearity Test

Another test for multistreaming is to check if the velocity field is still linear. This is motivated by the description that the simulations start out being linear, then transition through a quasi-linear, and finally to a nonlinear behavior. Detecting changes in the linearity of the velocity field may be an indicator of multistreaming.

Given a velocity field $\vec{V}$, position p, and velocity gradient J, we can obtain the velocity of a nearby point that is δp away using first order approximations, if the field is linear. From the velocity at p, we can obtain the velocities around it through

$$\vec{V}(p + \delta p) = \vec{V}(p) + J(p) \cdot \delta p \tag{6}$$

δp is set to one of $[\pm 1, 0, 0]$, $[0, \pm 1, 0]$, or $[0, 0, \pm 1]$ depending on which neighboring velocity we want. To check whether the velocities around p are linear or nonlinear, we compare the first order approximation of $\vec{V}(p + \delta p)$ against the original velocity $V_o(p + \delta p)$ at each of the 6 orthogonal neighbors. We use the normalized dot product to see if the directions of two vectors are similar, and use the absolute value of the difference of their velocity magnitudes to see if their magnitudes are similar. Note that the simple unnormalized dot product will consider velocities that agree in direction but not in magnitude as being similar, which is not what we want in this case. The vector pair is considered similar if their normalized dot product is at least 0.90 (i.e. less than 25.8 degrees). If a vector pair is similar according to this criterion, we assign it a value of 1, else a 0. A grid vertex is determined to be nonlinear based on the number of neighboring cells that are dissimilar. An aggregate value of 0 means the cell is highly nonlinear, while a value of 6 means the cell is linear. Note that if any neighbor of p is empty, we skip the calculation of $J(p)$ and do not apply the linearity test at p.

6 Comparison with Halos and Discussion

In this section, we examine if the feature extractors were able to find multistreaming regions. This question can be answered to a large extent by comparing their output against those of a density based halo finder. Due to space limitations, we only show images based on the last frame of the simulation.

Figure 5 is the output from the Halo Finder filter of ParaView applied on the last frame of the simulation. We use simple colored ellipsoidal glyphs to identify and distinguish the different regions found by the different extractors. In Figures 4 to 10, the glyphs are scaled and colored by the size of the regions that are detected by their respective extractors. In Figure 11, the color of the glyphs indicate the type of extractor. In all cases, we vary the orientation of the ellipsoids simply to distinguish among overlapping ellipsoids. The orientations do not have any physical meaning.

The size of the extracted region is closely tied to the power spectrum prediction shown in Figure 3, which was used to determine the time-dependent threshold of a particular method. There are many definitions of a region size. It can be the smallest bounding box around a region, or a diagonal of a box that bounds the region, or number of cells that are contained inside the region. We chose the last one, as it is the most accurate measure and easy to estimate, since all our calculations are performed on a grid.

Maximum Shear Stress. The regions with high shear stress are depicted in Figure 4. We can see that there is fairly good correspondence between regions with high shear stress and high density regions. But there are quite a few smaller regions with high shear stress but insufficiently high density to be detected by the halo finder. Likewise, we can also observe some high density region where the shear stress is not as pronounced (either the ellipsoidal glyph is absent or very small – the spheres are rendered transparently). Using the maximum shear stress criterion alone. we can say that not all halos have high shear stress, some potentially interesting regions are missed by halo finders, and at a qualitative level, high density regions have good correspondence with high shear regions.

Divergence. The results of the divergence extractor are shown in Figure 6. Comparing this figure with Figure 5, we see that the

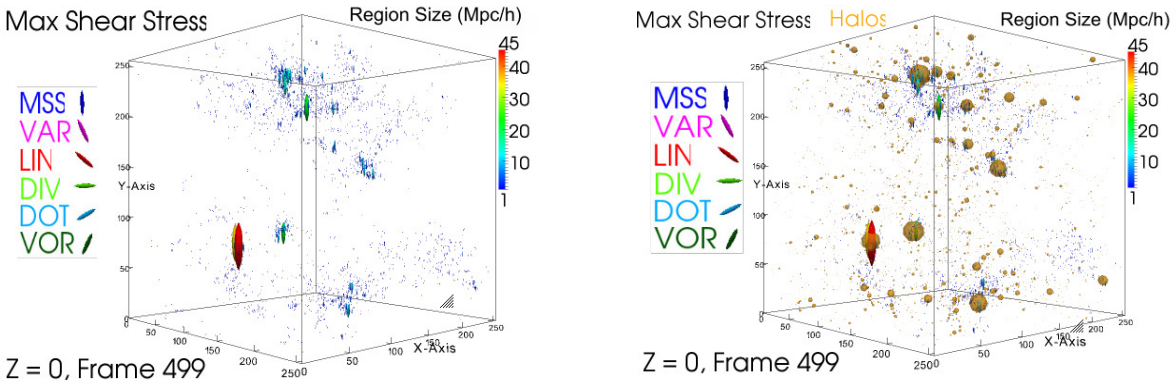

Figure 4: Maximum Shear Stress. Region size is number of 1 h^{-1}Mpc cells. Left: Ellipsoids scaled and colored by size of region with high shear stress. Right: Overlaid with transparent brown spherical halos.

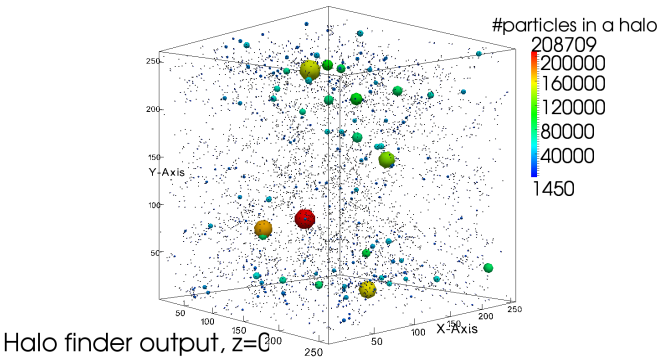

Figure 5: Halo Finder output from ParaView. Spheres are placed at the center of halo mass and their radius and color are mapped to the number of particles in a halo. The parameters use a link length (*bb*) of 0.2 and a minimum (*pmin*) of 1450 particles. This is our reference for comparing velocity based feature extractors. The image shows the last frame of the simulation, frame 499 with redshift Z = 0.

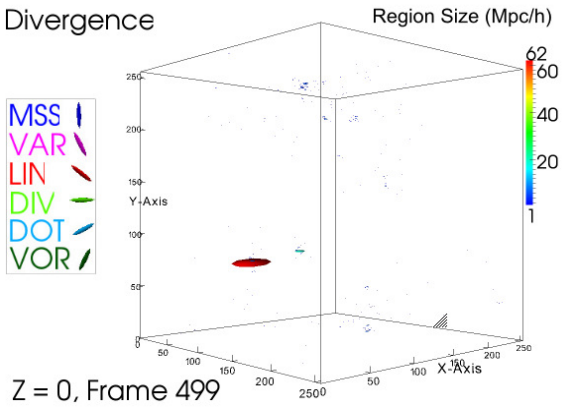

Figure 6: Ellipsoids scaled and colored by size of region with negative divergence. Region size is number of 1 h^{-1}Mpc cells.

2nd largest halo corresponds to the largest region with highly negative divergence (sink-like behavior). This indicates that the halo is still very active in recruiting new particles from its surrounding and increasing its density. The locations of other regions with negative divergence roughly correspond to where the other more prominent halos are located. However, the obvious lack of other significant regions with negative divergence seem to indicate that at this stage of the simulation accretion has subsided for most halos.

Vorticity. Figure 7 shows regions with low vorticity, and again we can see the correspondence of such regions with the halos in Figure 5. This time however, the largest region in 7 corresponds to the 3rd largest halo in Figure 5. What one can surmise from this particular frame is that the regions identified by this extractor are curl-free. But it is harder to see if there are halos that are also curl-free or if there are curl-free regions that are not halos. For a more quantitative assessment, we defer that to Table 2. Due to space limitations, we cannot show earlier frames of the simulation. But we do notice that as time progresses, the vorticity decreases confirming the general observation associated with multistreaming events.

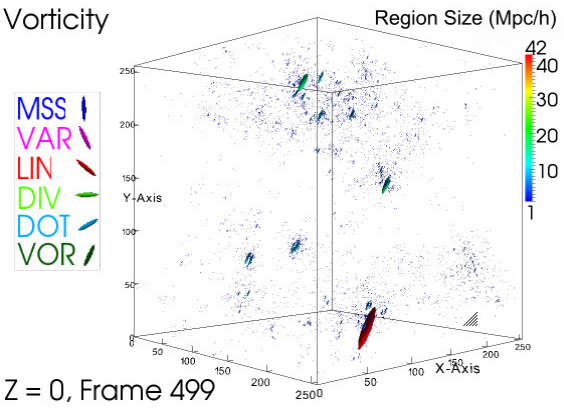

Figure 7: Ellipsoids scaled and colored by size of region with low vorticity. Region size is number of 1 h^{-1}Mpc cells.

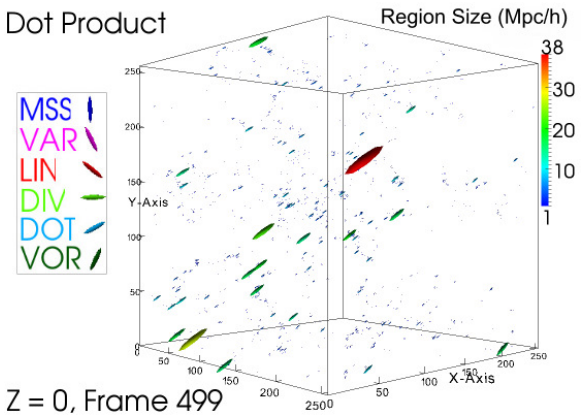

Figure 8: Ellipsoids scaled and colored by size of region with low absolute values of dot product. Region size is number of 1 h^{-1}Mpc cells.

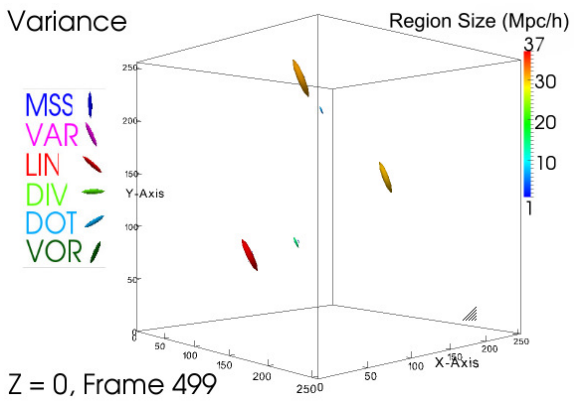

Figure 9: Ellipsoids scaled and colored by size of region with high variance. Region size is number of 1 h^{-1}Mpc cells.

Dot Product. Since we are using normalized quantities for this extractor, we are primarily looking for regions where particles tend to cross each other's paths – characterized by low values of the absolute value of this measure. In terms of correspondence between the regions with low dot products in Figure 8 and large halos in Figure 5, this measure seem to offer the least agreement. The only exception is the region that corresponds to the high variance in Figure 9, which also has a high divergence in Figure 6. A possible explanation for the apparent lack of crossing paths in the other halos is that indeed their particle velocities are fairly homogeneous in direction. That is, majority of the halos are fairly "boring" in terms of activity and range of particle directions.

Variance. Comparing Figure 9 to the halo finder result, we can observe that the main structures are similar. The four biggest halos agree with the four biggest regions with high variance. The number of regions high variance regions is very low, indicating that the rest of the space is fairly uniform within its local region. This includes those particles contained within the found halos.

Linearity Test. Figure 10 shows results of running the linearity test. We can observe that nonlinear regions, shown by ellipsoids, corresponds to the halos in Figure 5, particularly those in the upper

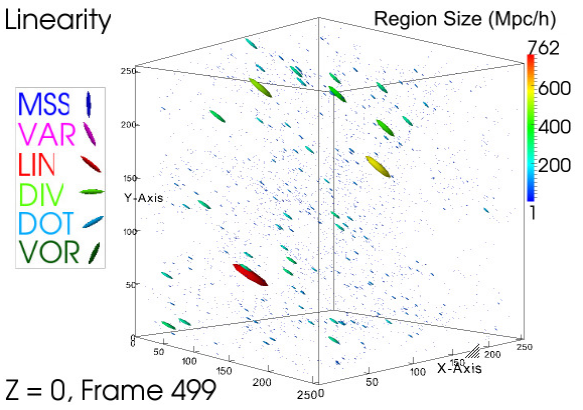

Figure 10: Ellipsoids scaled and colored by size of region with high nonlinearity. Region size is number of 1 h^{-1}Mpc cells.

half of the box. The overall structure of the nonlinear regions are consistent with those found with other methods described earlier.

Particles, Halos, and Regions. Beyond the qualitative visual agreement, we also look at some quantitative information with the following caveat: halos represent over-dense regions, while the regions found by the velocity based methods represent certain dynamic activity. The full relationship between halos and such regions has not been established, although it is believed that multistreaming is directly related to the formation of LSS, such as halos.

The halo finder found is 5036 halos. The number of particles inside these 5036 halos is 25,314,910. To compare these against the multistreaming regions, we define and measure two quantities: accuracy and coverage. By accuracy we mean how many regions out of those found by a method where also identified by halo finder. For example, if method A found ten regions and nine of them intersected (had common particles) with fifteen halos, then accuracy is 90%. While method A is pretty accurate its coverage is very low, only 15/5036, which is less than 0.3%. Our definition of coverage is the number of halos that were identified as interesting by both halo finder and our method. Results are summarized in Table 2. The two extreme cases are the variance method which found very few regions but they are all halos, and the linearity method which found a lot of nonlinear regions but only a few of them were found by halos. These results also hint at the different dynamic behavior of different halos and suggest a possible classification of halos based on their behavior.

The methods are faster than the halo finder. On a desktop with a 6 core AMD phenom II X6 processor and 16gb ram, it takes about 4 minutes to read a frame and another 17 minutes to find halos using ParaView. On the other hand, the pre-processing time to convert discrete particle velocities to a continuous field using CIC took 12.5 minutes. The different methods took varying times: maximum shear stress: 60s; divergence field: 30s; vorticity: 65s; dot product: 7 minutes; variance: 6.5 minutes; linearity test: 25s.

7 Conclusions and Future Work

We started this investigation with a general question of whether we can use the particle velocity information to detect and characterize multistreaming events. We hypothesized how the flow field should behave given the various descriptions of multistreaming in the cosmology literature and formulated ways to extract regions with those behaviors. These methods require a threshold value to determine if a region is multistreaming or not. For that, we used a physics based

Method Name	# Regions Found	# Intersections with halos	Accuracy (%)	Coverage (%)
Max Shear Stress	6576	1838	28.0	36.5
Divergence	616	529	86.0	10.5
Vorticity	6770	2082	41.9	41.3
Dot Product	1368	1349	98.6	26.8
Variance	14	14	100.0	0.3
Linearity	122189	15240	12.5	100.0

Table 2: Overalps between multistreaming regions and halos using different velocity based methods.

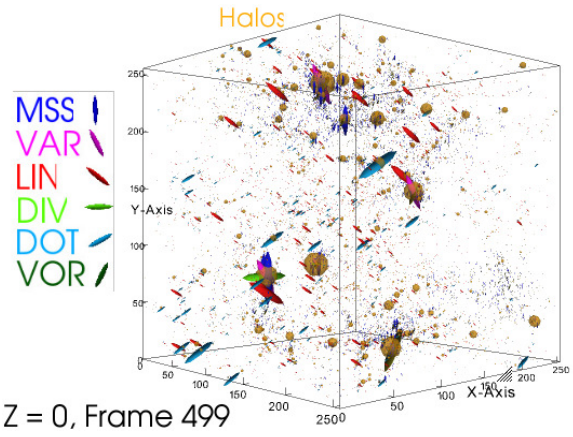

Figure 11: Overlay of halos as spheres and the different regions found by velocity based regions as oriented ellipsoids. Different colors indicate different methods.

approach of determining a time-varying threshold for the different methods that would capture the multi-scale multistreaming events.

We have compared our results against the popular density based halo finder as implemented in ParaView. Our findings indicate that: (i) the different velocity based methods found multistreaming regions and also provide additional information about their dynamic behavior; (ii) there is good qualitative correspondence between the regions found using our velocity based methods and those found by ParaView's halo finder; (iii) not all halos are the same in terms of their dynamic qualities, and that the velocity based methods could be used to classify halo types; (iv) the relationship between multistreaming regions and halos reflect those observed by Shandarin: "While there are differences in the locations and peaks of over density and high velocity variance regions (or those by other velocity based methods), we posit that these can be resolved by analyzing the evolution of these regions as opposed to studying individual frames of the simulation. In particular, we hypothesize that high velocity variance regions (or those by other velocity based methods) may at a later time lead to high over density regions, and vice versa." Further investigations along this line will require feature tracking tools which we are currently pursuing. No fancy visualization was necessary to answer our science question in this investigation. We will endeavor to keep things simple as well in the future.

There are several avenues of future research: a pushback to the scientist is the question on how multistreaming relates to halos, and what is the significance of halos with different behaviors? In terms of the technical side, it would be interesting to see if a machine learning approach may yield a superior feature extractor using some combination of density and velocity based methods. Furthermore, the methods we have explored so far are based on local information. Some halo finders now use the 6-D phase space (3 position and 3 direction) information to detect subhalos as well phase-space structures such as pure streams and candidate caustics [11]. How can one utilize phase-space analysis, beyond clustering, to detect multistreaming regions in a global manner?

References

[1] A. Knebe at el. Haloes gone mad: The halo-finder comparison project. *arXiv:1104.0949v1*, 4 2011.

[2] J. Carlson, M. White, and N. Padmanabhan. Critical look at cosmological perturbation theory techniques. *Physics Review D*, 80(4):043531, Aug 2009.

[3] E. Chandra. Exploring multistreaming in the universe. Master's thesis, University of California Santa Cruz, 2009.

[4] M. Davis, G. Efstathiou, C. S. Frenk, and S. D. M. White. The evolution of large scale structure in a universe dominated by cold dark matter. *Astrophysical Journal*, 292:371–394, May 1985.

[5] R. Fraedrich, J. Schneider, and R. Westermann. Exploring the millennium run - scalable rendering of large-scale cosmological datasets. *IEEE TVCG*, pages 1251–1258, 2009.

[6] K. Heitmann, P.M. Ricker, M.S. Warren, and S. Habib. Robustness of cosmological simulations I: Large scale structure. *Astrophysics Journal Supplement*, 160(28), 2005.

[7] R. W. Hockney and J. W. Eastwood. *Computer simulation using particles*. Taylor & Francis, Inc., Bristol, PA, USA, 1988.

[8] A. Klypin, S. Trujillo-Gomez, and J. Primack. Halos and galaxies in the standard cosmological model: results from the Bolshoi simulation. http://arxiv.org/abs/1002.3660v4.

[9] H. Li, C.-W. Fu, Y. Li, and A. Hanson. Visualizing large-scale uncertainty in astrophysical data. *IEEE TVCG*, 13(6):1640–1647, 2007.

[10] Z. Lukic, D. Reed, S. Habib, and K. Heitmann. The structure of halos: Implications for group and cluster cosmology. *The Astrophysical Journal*, 692(1):217–228, 2009.

[11] M. Maciejewski, S. Colombi, V. Springel, C. Alard, and F. R. Bouchet. Phase-space structures II. hierarchical structure finder. *Monthly Notices of the Royal Astronomical Society*, 196:1329–1348, 2009.

[12] P.J.E. Peebles. The standard cosmological model, 1998. http://nedwww.ipac.caltech.edu/level5/Peebles1/paper.pdf.

[13] V. Sahni and P. Coles. Approximation methods for non-linear gravitational clustering. *Physics Reports*, 262:1–135, November 1995.

[14] S. Shandarin. The multi-stream flows and the dynamics of the cosmic web, 2010. http://arxiv.org/abs/1011.1924.

[15] S. F. Shandarin and Ya. B. Zeldovich. The large-scale structure of the universe: Turbulence, intermittency, structures in a self-gravitating medium. *Rev. Modern Physics*, 61(2):185–220, 1989.

[16] R. von Mises. Mechanik der festen körper im plastisch deformablen zustand. *Göttin. Nachr. Math. Phys.*, 1:582–592, 1913.

[17] J. Woodring, K. Heitmann, J. Ahrens, P. Fasel, C.-H. Hsu, S. Habib, and A. Pope. Analyzing and visualizing cosmological simulations with ParaView. http://arxiv.org/abs/1010.6128v1.

[18] T. Yano, H. Koyama, T. Buchert, and N. Gouda. Universality in the distribution of caustics in the expanding universe. *The Astrophysical Journal Supplement Series*, 151:185–192, 2004.