Velocity Based Feature Extraction of Multistreaming Events in Cosmological Simulations

Uliana Popov, Eddy Chandra, Katrin Heitmann, Salman Habib,
James Ahrens, and Alex Pang

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 events were studied using scalar density field only. In this paper, we present a new approach that takes into account the velocity field information in finding these multistreaming events. Six different velocity based features are defined and studied. We find that these velocity based feature extractors show different aspects of multistreaming and provide us with a richer knowledge about the formation of LSS.

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 dark matter component (CDM), and only 0.4% of the remaining 5% of ordinary (atomic) matter is visible. 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 such as galaxy clusters (0D), filaments (1D), and surface-like pancakes (2D) can be considered to correspond to nodes, edges, and faces respectively, in a tessellation of the topology of the universe. 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 [4],

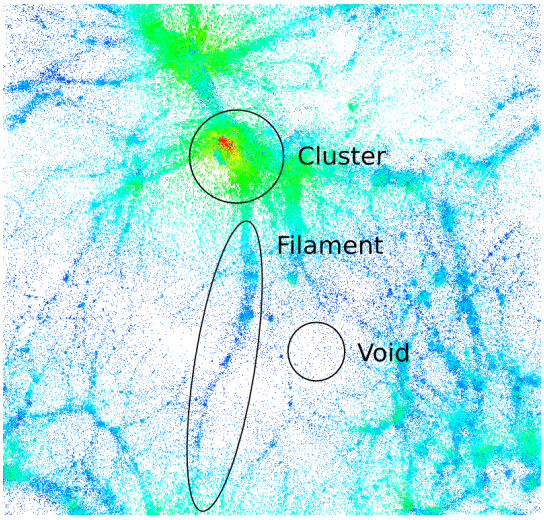

Figure 1: Large scale cosmological structures of the universe.

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.

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 submanifold 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, consider a key cosmological statistic measured from simulations: the density fluctuation power spectrum. The 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., [7]). 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. In particular, we are interested to know what additional information may be gleamed from the velocity information.

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 [9]. 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 because the mapping from phase space to physical space becomes multivalued. 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" [8]. 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 con-

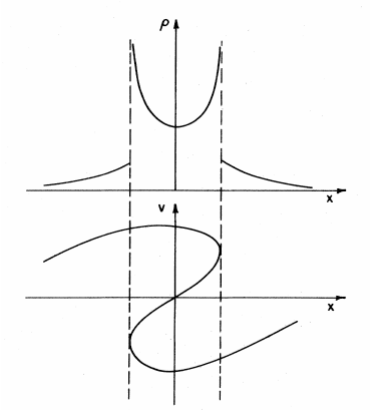

Figure 2: 1-D illustration of multi-stream flow. Top panel: Over-dense region with three-stream flow confined between the dashed lines. Bottom panel: The corresponding phase space plot showing the different stream regions [9].

gregrate, 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 subset of these include works by Li et al. [6] which explored how to display positional and trajectory uncertainties in astrophysical data sets; and Fraedrich et al. [3] 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 identifying multistreaming events in such data sets.

Multistreaming events have been explored in the past years. For example, Yano et al. [12] 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 [2]. 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 [11].

More recently, Shandarin [10] 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 location 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. A related work that focuses on tracking the evolution of multistreaming regions is also under review.

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 [5] 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. In addition, the grid size has to be small enough to resolve the features of interest at certain length scales. 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. For our investigation, we choose a grid resolution such that on average there are 64 particles contributing to each grid point. For the 512^3 particle data set, this goal is achieved by a regular grid with 256^3 cells. As we explain later, this grid size also allows us to find multistreaming regions early on in

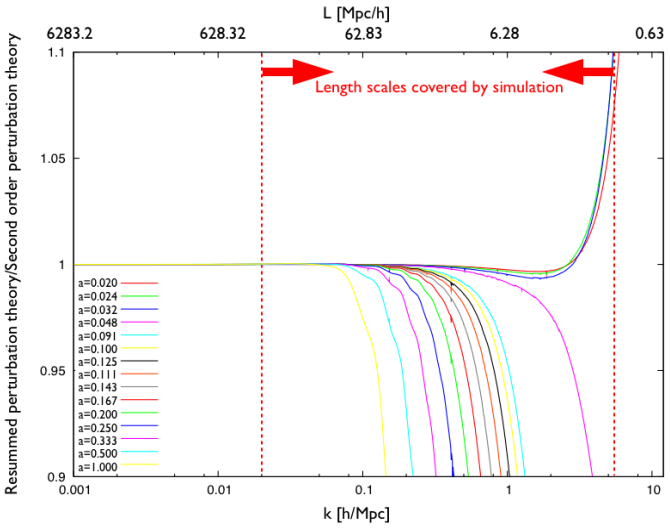

Figure 3: Breakdown of perturbation theory at different scale factors and different length scale. The curves are the ratio of two different perturbation theories. As the ratio deviate from one, perturbation theory is not valid anymore. Each curve shows the result for one time snapshot. At the top of 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.

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.

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

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

scale for the onset of multistreaming.

Following Carlson et al. [1], 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 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 ≈ 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 examine several methods that look for multistreaming regions using the particle velocity information. We use ParaView to visualize our results, and compare them to the output of ParaView's HaloFinder filter. Figure 4 shows the halo finder results. Frame 70 is when the first halo is found. Based on our analysis, we actually expect the multistreaming to occur earlier, at around frame 21. For each of our methods, we show the zoomed in views of two frames from a partial volume of the simulation: (a) frame 70, for comparison with the halo finder results, and (b) frame 499, the last frame of the simulation. In addition, we include plots that show the relative values of the time-varying threshold and the volume average of the different metrics over time.

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. To find the maximum shear stress, we first calculate the velocity gradient tensor of the velocity field, then find its symmetric tensor component, and then the associated eigenvalues λ_1, λ_2, and

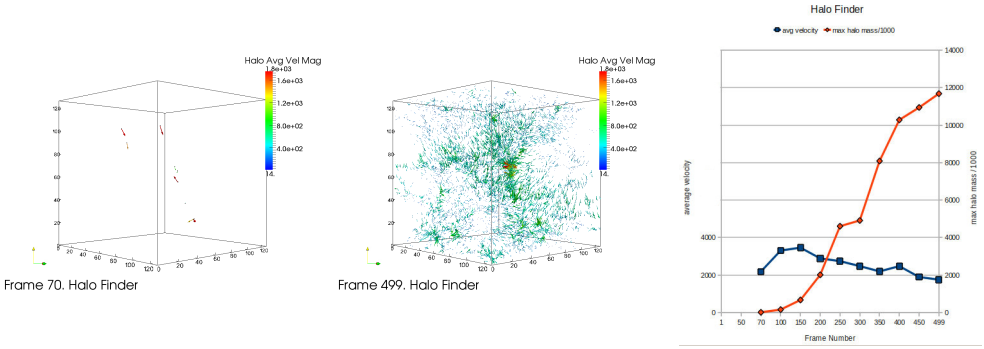

Figure 4: Halo Finder output from ParaView 3.10 using the friends of friends (FOF) algorithm. Arrows represent average velocity of halos. The parameters use a link length of 0.2 and a minimum of 100 particles. This is our reference for comparing velocity based feature extractors. We can observe the rapid increase in average velocities around frame 70 when the first halos are found. Velocities then gradually tapers off, while the concentration of particles in the halos increase over time. Several discernible structures can also be observed by frame 499.

λ_3. We use the von Mises criterion 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. Figure 5 shows the results of this metric. We can see that the general structure in frame 499 has some qualitative similarities with those from the halo finder. Frame 70 seems to capture larger high shear regions than the corresponding halos, while frame 499 captures a subset of the corresponding halos. Looking at shear stress alone, we can observe that the average and median shear stress converge rapidly and approach a value of around 20 with very small standard deviation. The threshold needed to obtain multistreaming regions of the appropriate length scale is decreasing over time, and approaches the average shear stress value. Within the high shear stress regions, we can readily observe that stress is decreasing over time suggesting that differences in velocity magnitudes may be decreasing. This information is not available using the halo finder alone. Later on, we investigate differences due to velocity directions using the normalized dot product metric. In summary, the maximum shear stress criterion is able to qualitatively identify the general structure of multistreaming regions and also provide information about the behavior of maximum shear within such regions.

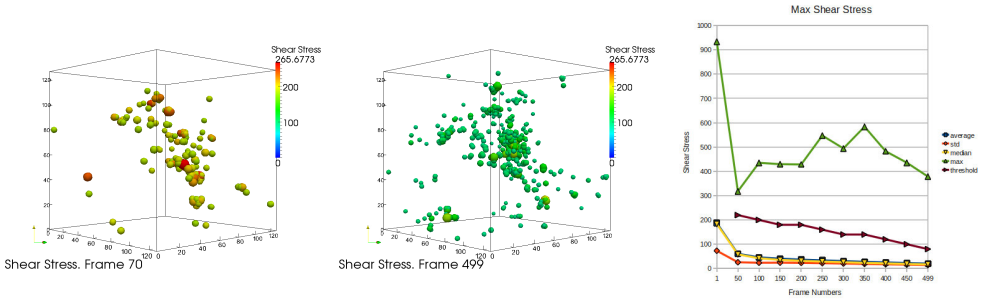

Figure 5: Maximum shear stress. Glyphs are colored and sized by maximum shear stress. The time-varying thresholds are 180 and 80 for frames 70 and 499 respectively.

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

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

The results of the metric are shown in Figure 6. Comparing these images to the halo finder results, we can see that locations of large negative divergence correspond to large clusters, particularly for frame 499. This indicates that those regions are still drawing in particles and smaller regions from its nearby surroundings, though at a diminished rate relative to frame 70. One can observe that earlier in the simulation, the negative divergence is stronger, but the regions are not as dense. Later in the simulation, the multistreaming regions have weaker negative divergence indicating that they are not drawing in nearby particles as aggressively, perhaps leading to a more stability in the LSS. The plot on the right shows that the threshold for regions with negative divergence is becoming less negative over time, confirming that indeed the multistreaming regions is becoming weaker at attracting new particles. This metric shows additional information not available from the halo finder – the dynamical behavior of halo clusters.

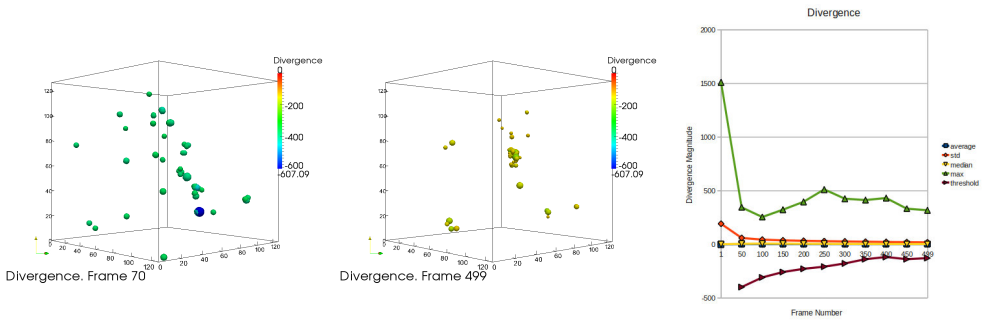

Figure 6: Divergence. The time-varying thresholds are -360 and -130 for frames 70 and 499 respectively. The glyphs are colored and sized according to divergence. Since we are interested in regions with negative divergence only, the colormap of the 2 figures are bounded by 0 and the largest negative value -607. More negative divergence are mapped to larger balls.

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

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

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. The results are illustrated in Figure 7. We see the regions found in frames 70 and 499 are also similar to those of the halo finder. More importantly, we can see that over time, indeed the vorticity magnitudes in the multistreaming regions are getting lower and have less vorticity. This is confirmed by the plot on the right as well. At the same time, the threshold is still above the average vorticity due to regions with very high vorticity. If one extrapolates the curves on the plot, a possible conclusion is that the system appears to be reaching some form of dynamic equilibrium where few large multistreaming regions with low vorticity are balanced by other areas with high vorticity. Such behavior is impossible to observe from the halo finders alone.

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

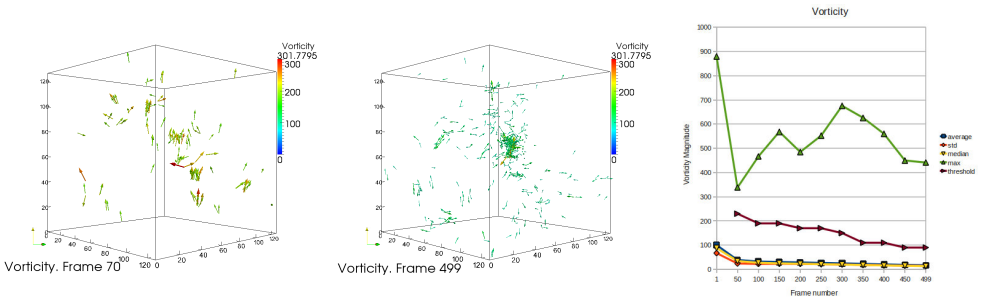

Figure 7: Vorticity vectors colored by their magnitude. The time-varying thresholds are 180 and 90 for frames 70 and 499 respectively. We hypothesize that multistreaming regions will have low vorticity magnitudes. The scale on the 2 images are the vorticity magnitudes found in multistreaming regions over the course of the simulation, while the plot on the right includes all regions including those in non-multistreaming regions.

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

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 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 represents the directional similarity among the particles in the vicinity of the grid vertex.

Multistreaming regions composed of particles going in directions that tend to cross each other can be characterized by low values of the absolute value of the normalized dot product. We can observe from the 2 images that this quantity starts out as fairly high in frame 70 and become smaller in frame 499 indicating that the velocities within the multistreaming regions are becoming more dissimilar. The overall structure is still recognizable when compared to the results of the halo finder, although the overlapped regions are different. There are more regions that are flagged as multistreaming early on that were not found by the halo finder. On the other hand, in later parts of the simulation, the reverse is true. By itself, this metric can find overlaps with the halo finder, but its discriminatory power seem to vary over time. Thus, while it shows information about the degree of dissimilarity

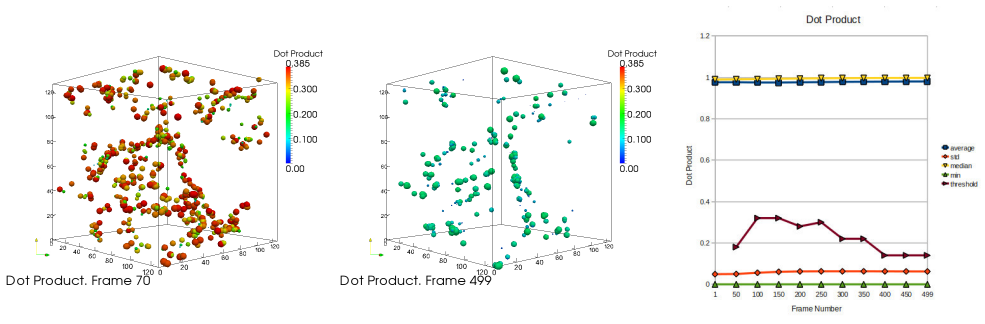

Figure 8: Normalized dot product. The time-varying thresholds are 0.385 and 0.15 for frames 70 and 499 respectively. Spherical glyphs are colored and sized by the absolute value of the normalized dot products. The plots are also based on the absolute value of the normalized dot products.

within regions, it may be best to use other metrics to obtain a more definitive locations of the regions first.

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 multi-streaming 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 [10] also used velocity variance in his analysis, but our formulation differs slightly. Given n numbers $x_1, x_2, ..x_n$ and a mean μ, the variance σ^2 is defined as

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

Since we are interested in velocity variance in 3D, 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.

Comparing the images to the halo finder results, we can observe similar structures in frame 70, but just the core of the largest halo region is extracted in frame 499. Compared to the normalized dot product results, we see a similar pattern of decrease in the size of the multistreaming regions. One difference is that with normalized dot product, the dissimilarity of crossing vectors increased over time, while the velocity variance decreased over time. This difference may be attributable to the fact that the normalized dot product focuses on crossing vectors while the variance includes all types of vectors. This suggests that most of variance in later

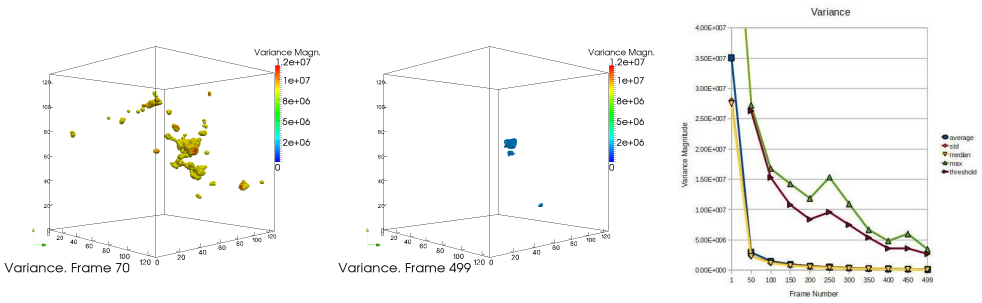

Figure 9: Variance. The time-varying thresholds are 8,500,000 and 1,500,000 for frames 70 and 499 respectively. Glyphs are colored and sized by velocity variance. The maximum variance in the interval from frame 1 to 50 is an order of magnitude higher and truncated so that we can see more detail in the other curves.

frames is dominated by crossing vectors and less by shearing vectors. The results using the shear stress metric supports this observation.

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 get. 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. Figure 10 shows results of running the linearity test on the our data set. The overall structure of the nonlinear regions are consistent with those found with other methods described earlier. Here, we can see that in the earlier frame, the size of the nonlinear region

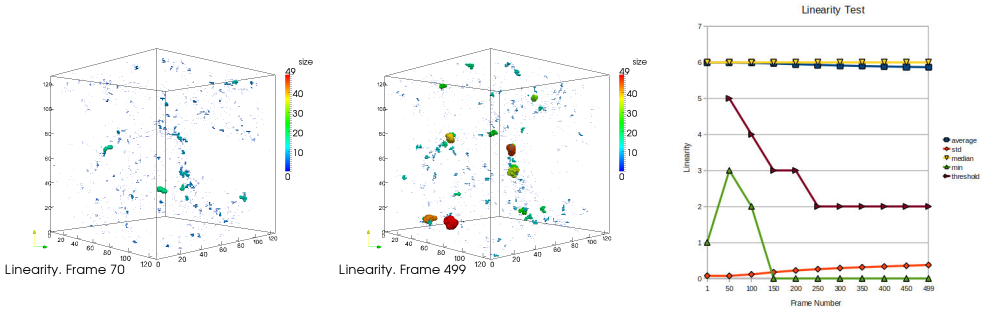

Figure 10: Nonlinear regions in frames 70 and 499. First, we apply the time-varying thresholds of 5 and 2 respectively. These are the number of neighbors voting that the vertex velocity is linear. Applying this threshold identifies vertex velocities that are nonlinear. These are then grouped together into larger connected regions. The glyph size and color are mapped to the size of nonlinear regions. Note that while the images show size of connected nonlinear regions, the plots show the linearity measures. We can see that majority of the volume is linear (high average linearity with small standard deviation) except for the identified nonlinear regions in the images.

are smaller. This corresponds to the expected length scale of the multistreaming regions. At the last frame, we can see much larger regions, together with smaller nonlinear regions. This also corresponds to the multi-scale nature of multistreaming regions. We also note that while the prominent halo found by the halo finder and other methods described earlier is the one near the center of the volume, the region closer to the bottom left also figures prominently in terms of size based on the nonlinearity criterion.

6 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 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, and also against the work by Shandarin [10] which is based on velocity variance. Our findings indicate that: (i) the different velocity based methods not only find the multistreaming regions, but they also provide additional information about the dynamic behavior within and in the vicinity of the regions. These behaviors include how shear, vorticity, divergence, vector similarity in direction, velocity variance, and linearity, change over time; (ii) there is good qualitative correspondence between the regions found using our velocity

based methods and those found by ParaView's halo finder; (iii) the relationship between the regions found using velocity variance (and other velocity based methods) and density 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.

While one may wonder which is the best method, we are not ready to provide an answer yet as the methods have different strengths and weaknesses with respect to the different properties of multistreaming. We plan to investigate if a machine learning approach may yield a superior feature extractor using some combination of density and velocity based methods. Another area of future research is to compare the percolation statistics of multistreaming regions obtained by density and velocity based feature extractors.

References

[1] Jordan Carlson, Martin White, and Nikhil Padmanabhan. Critical look at cosmological perturbation theory techniques. *Physics Review D*, 80(4):043531, Aug 2009.

[2] 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.

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

[4] 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.

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

[6] Hongwei Li, Chi-Wing Fu, Yinggang Li, and Andrew Hanson. Visualizing large-scale uncertainty in astrophysical data. *IEEE Transactions on Visualization and Computer Graphics*, 13(6):1640–1647, 2007.

[7] Zarija Lukic, Darren Reed, Salman Habib, , and Katrin Heitmann. The structure of halos: Implications for group and cluster cosmology. *The Astrophysical Journal*, 692(1):217–228, 2009.

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

[9] 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.

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

[11] Jonathan Woodring, Katrin Heitmann, James Ahrens, Patricia Fasel, Chung-Hsing Hsu Salman Habib, and Adrian Pope. Analyzing and visualizing cosmological simulations with ParaView. http://arxiv.org/abs/1010.6128v1.

[12] Taihei Yano, Hiroko Koyama, Thomas Buchert, and Naoteru Gouda. Universality in the distribution of caustics in the expanding universe. *The Astrophysical Journal Supplement Series*, 151:185–192, 2004.