ACHIEVING PROBABILISTIC CLASSIFICATION OF COSMIC WEB PARTICLES USING RAPIDLY GENERATED TRAINING DATA: A METHOD FOR CLASSIFYING GALAXIES INTO THEIR COSMIC WEB STRUCTURAL GROUPS USING SUPERVISED MACHINE LEARNING

Allocation: Exploratory/35 Knh
PI: Matias Carrasco Kind[1]
Co-PI: Brandon Buncher[1]

[1]University of Illinois at Urbana–Champaign

EXECUTIVE SUMMARY

The cosmic web consists of a network of galaxies and dark matter. Long, strandlike filaments connect between spheroidal galaxy clusters, leaving underdense void regions in between. Knowing whether a galaxy is a member of a halo (the dark matter clump around which clusters form), filament, or void provides substantial information about its surrounding environment, which helps in understanding how it formed and will evolve.

However, current methods are limited: direct classification algorithms are generally inefficient, while deep learning-based methods are inconsistent with one another. Therefore, the research group created a novel classification method using supervised machine learning. They train the algorithm using quickly generated data that visually approximates the cosmic web using predetermined generation algorithms. Then, with the help of the high memory capacity of Blue Waters, the researchers use the trained algorithm to classify galaxies in simulated data. While the training data lack much of the detail seen in observed/simulated data, it takes substantially less computational power to create. A simulation of cosmic web formation with 16 million particles requires tens of thousands of node-hours on a multinode cluster, whereas the team's method can generate training data with the same number of particles in less than an hour on a laptop computer. The researchers have demonstrated that this method provides enough information for the machine learning algorithm to "learn" to correctly classify particles in more realistic data sets. Although it is trained using simpler data, the robustness of machine learning helps the algorithm bridge the information gap, providing classification at a substantially cheaper cost.

RESEARCH CHALLENGE

Current methods used to classify galaxies into cosmic web structural groups typically utilize neural networks, as direct algorithms are too computationally intense. However, arbitrary hyperparameters and inconsistent cosmic web class definitions lead to substantial disagreement among methods. In addition, owing to these hyperparameters, it is difficult to establish self-consistency for a given method.

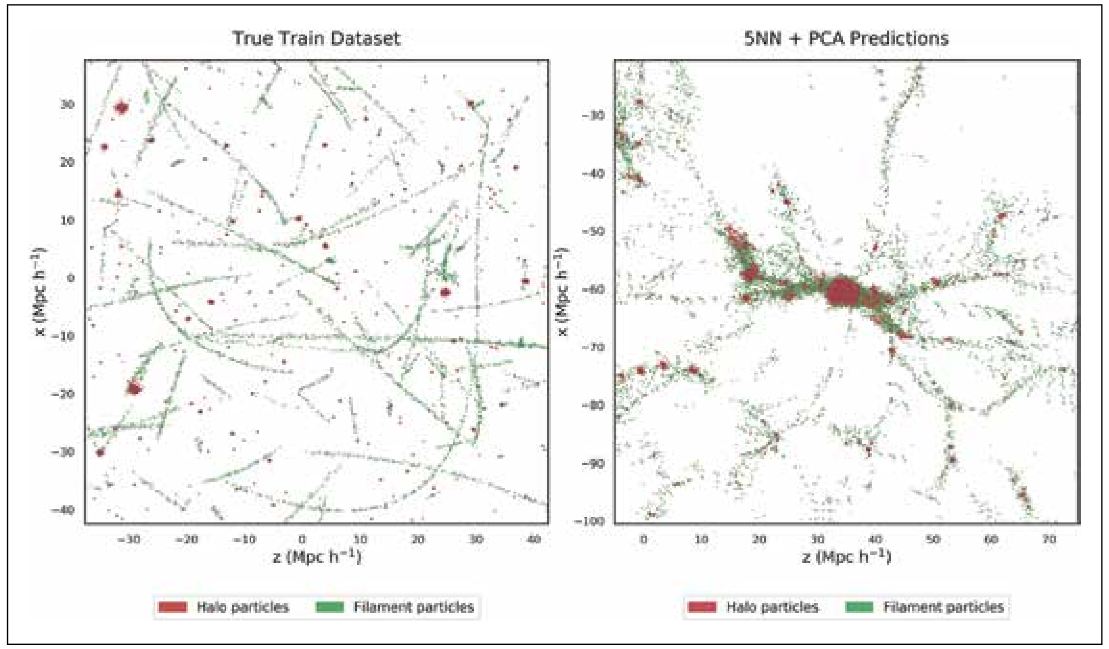

Figure 1: (left) A two Mpc-thick slice of the 3D training data set with 660,000 particles: red particles are members of halos, green of filaments, and blue of voids. (right) A two Mpc-thick slice of the 3D N-body simulation data set with 16.7 million particles; training was performed exclusively using the training data set on the left. A megaparsec (Mpc) is a unit of length used in astronomy.

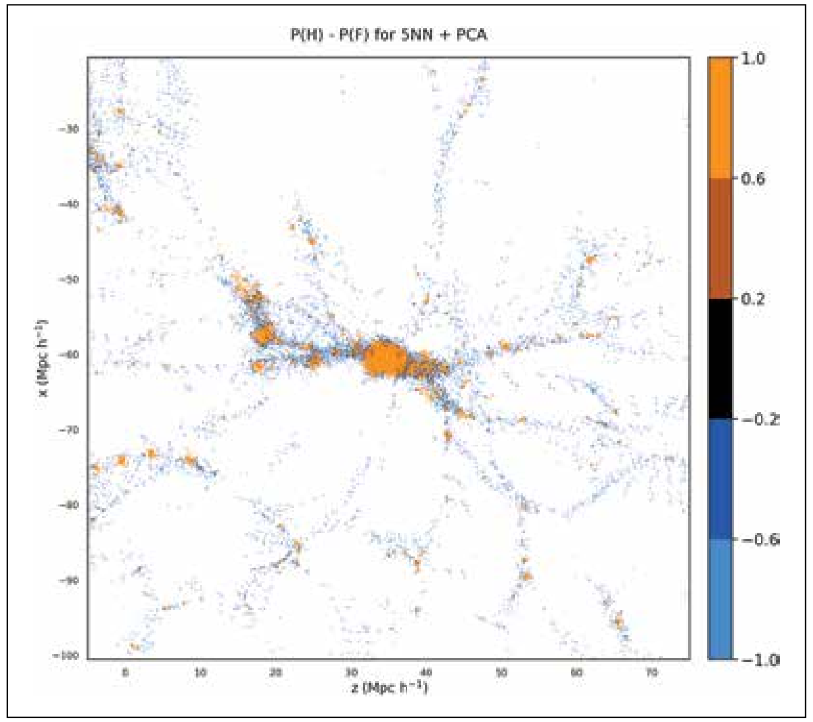

Figure 2: The particle probabilities for N-body simulation halos/filaments. Light-colored particles had a high probability for halo (orange) or filament (blue) classification, while black particles were ambiguous. Note that the majority of ambiguous particles appeared on the border between halos and filaments, where ambiguity naturally exists owing to the low-density contrast.

A particle's cosmic web class membership provides substantial information about its local environment, which plays a great role in its history and future evolution. The research team aims to improve cosmic web classification through implementing a supervised algorithm, which is easier to verify, lacks arbitrary hyperparameters, and, by simplifying training data generation, is substantially less expensive and time-consuming than direct methods.

METHODS & CODES

Training data is generated by sampling particles from predetermined toy model structures. Halos are produced by sampling particles from a spherical Gaussian distribution and then distributed randomly throughout the region. Filaments consist of particles populated in a uniform cylinder around randomly generated Bezier curves (https://github.com/dhermes/bezier), while voids are a uniform background distribution. Measurements of the k-nearest neighbors provide information about the local density magnitude, while the explained variance ratio calculated from a principal component analysis decomposition measures the density field directionality. A random forest algorithm was trained using the results of these measurements on each particle, and the trained algorithm was used to classify particles in an N-body simulation. Though these classifications cannot be directly verified owing to a lack of true class values, the researchers demonstrated the robustness of the method by comparing these predictions to those made on another toy model data set.

RESULTS & IMPACT

Currently, the team is working on a final write-up of the results. Work to date has demonstrated that the methodology does provide a robust method for classifying particles at a much lower computational cost. In addition, the team's verification methods have shown that the method achieves probabilistic classification, providing additional information to use when studying the formation and evolution of galaxies. Although the optimal configuration of measurements and toy model parameters remains to be found, the research group has demonstrated the effectiveness of using a supervised method trained with fast generated data, and they expect that future optimization will lead to improvements in particle classification that is naturally accompanied by information on the classification confidence. This will change the field of cosmology by simplifying and improving the efficiency of particle classification, enabling greater understanding of the cosmic web and its effects on other structures.

WHY BLUE WATERS

Blue Waters provided substantial benefit through its high-memory nodes. Although toy model generation is extremely fast and efficient, performing measurements on the N-body simulation requires considerably more memory than is available on a laptop computer or a single Blue Waters node. Segmenting the particle field does not avoid this issue because of the size of the required files, and analysis would take an unreasonable amount of time. Blue Waters' nodes had enough memory to load the segmented files and analyze them, and by parallelizing the analysis across several nodes, it allowed computations that would normally require several weeks on a high-capacity remote machine to be completed in less than a day.