Representing galaxies and their surroundings with graph neural networks

PI: John F. Wu, Assistant Astronomer at STScI (email, website)
Group Website: Chesapeake ML-Astro, ISM*@ST
Project Duration: ~~Three~~ Two separate 6 month rotation projects; can lead to thesis project.

Background and Context

Galaxies grow and evolve in dark matter halos amidst a complex, large-scale environment. These galaxies (or dark matter subhalos) can be represented as nodes on a graph, and their relationships or effective interactions can be represented as edges on a graph. By train graph neural networks (GNNs), we can learn the physical relationships between galaxies, subhalos, and their surroundings directly from large data sets, such as hydrodynamic simulations (see Figure 1 below; Wu & Jespersen 2023). This has implications for modeling the galaxy-halo connection (e.g., Wechsler & Tinker 2018) and large-scale correlations of galaxy properties (e.g., Hearin et al. 2016, Wu et al. 2024).

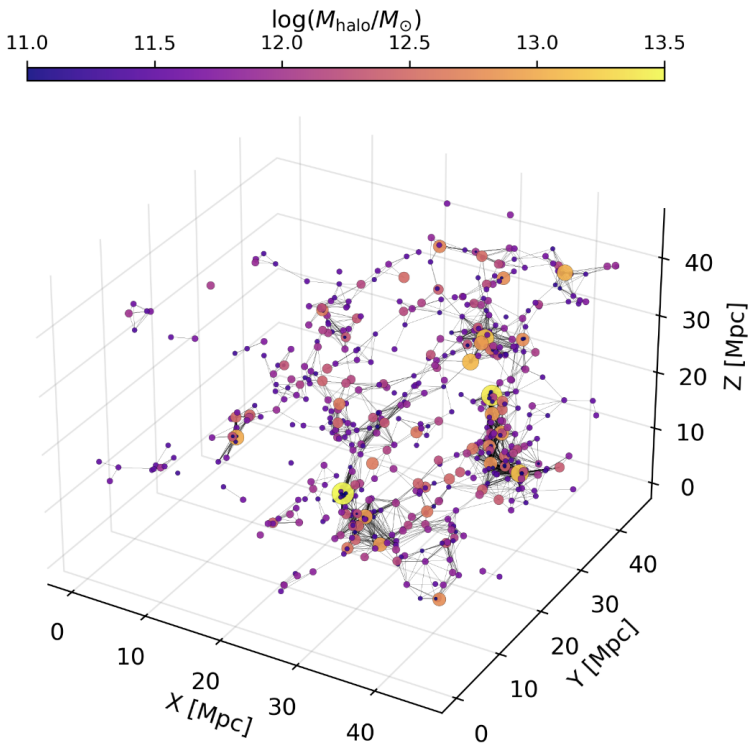

Figure 1. A cosmic graph showing galaxies/subhalos from the Illustris TNG300 simulation. From Wu & Jespersen (2023).

Available Projects

There are still many open questions about how to improve the use of GNNs to model cosmic structures. A prospective study can select *one of ~~three~~ two projects* listed below:

1. **~~Augmenting GNNs with morphological information.~~** ~~One promising avenue is to include the morphological information of galaxies while trying to learn the galaxy-halo environment connection, since galaxy mergers and evolutionary processes leave imprints on galaxy appearances. There are several ways to quantify galaxy morphology, and in this project, we will compare a baseline GNN against GNN models augmented with morphological information. All data here will come from publicly accessible cosmological simulations.~~
2. **Applying GNNs to observations of galaxy clusters.** We can also apply trained GNN models to observations of galaxy clusters in order to estimate the dark matter mass profiles. Many lensing clusters have been imaged by HST, JWST, and ground-based telescopes, providing an orthogonal constraints on the cluster masses. In this case, GNNs trained on simulation data would be adapted to galaxy catalogs taken from real observations.
3. **Using symbolic regression to interpret scaling laws**. By using symbolic regression (e.g., PySR) to derive analytic formulas in place of neural networks, we can learn an interpretable form of GNN outputs, e.g., the dark matter halo mass as a function of galaxy properties.

Student Work

Planned work:

- Download and preprocess large hydrodynamic simulation data (i.e. TNG300).
- Implement and train a GNN to estimate dark matter halo masses from baryonic properties (starter code is available).
- ~~**Project 1**~~
  - ~~Augment GNN using added information from galaxy morphology classifications.~~
  - ~~Augment GNN using convolutional neural network (CNN) extracted morphological features from synthetic galaxy images.~~
- **Project 2**
  - Obtain galaxy catalog data for clusters (e.g. JWST galaxy catalogs for SMACS 0723 or Abell 2744).
  - Compare GNN predictions against lensing measurements.
- **Project 3**
  - Extract analytical equation for M_halo as a function of M_star and other environmental parameters by using symbolic regression.
  - Compare against other standard tools such as halo occupation distribution modeling.
- Write and publish a short paper detailing findings.

If a student is interested in doing multiple projects, then there is an option to turn these projects into part of a thesis.