CS 224W Final Project: Predicting Galactic Properties from Network Structure of Cosmological Simulations

Ryan Gao (rgao)

December 8, 2015

1 Introduction

This project focuses on an application of network science to the field of cosmology, one of the less well-explored cross-domain collaborations. Cosmological simulations aim to reproduce the evolution of the universe, from its early homogeneous state 13.8 billion years ago, to the current highly inhomogeneous arrangement of galaxies separated by large voids of intergalactic space. In recent years, as computational power has increased exponentially quickly, the granularity and detail of cosmological simulations has likewise improved. Since the earliest large-scale gravitational N-body simulations in 1985 [1], cosmologists have been tweaking their models to better simulate the observed universe. The influential Millennium simulations [7] used N-body simulations coupled with smoothed particle hydrodynamics, which models the fluid behavior of gases, to better resolve small-scale structure.

The prevailing standard model of cosmology, the Lambda Cold Dark Matter (ΛCDM) model, describes a universe dominated by the unexplained mechanisms of dark matter and dark energy, with a small contribution from the familiar baryonic matter made of protons and neutrons. There are 6 unknown parameters of the ΛCDM model, and by tuning simulations to observations, we can find the cosmological parameters that best fit our universe. However, not only do empirical observations inform our understanding of the simulation results, so too do the simulations inform our understanding of what we see in the night sky. Simulations have explained how galaxies come to be arranged in a "Cosmic Web," a series of filaments and sheets that intersect at dense clusters and are separated by sparse voids. This was first noted in observational data by de Lapparent et al. in 1985 [3].

This project aims to predict properties of galaxies seen in cosmological simulations, based on the galaxy's network properties within the Cosmic Web. High-quality observations of galaxies are time-consuming and expensive, especially ones at high redshift, i.e., distant in both time and space. Thus, we can apply the patterns learned from simulation data to observational data, in order to fill in missing fields or to improve on noisy observations.

We will review the literature for cosmological applications of network analysis in Section 2, describe the Illustris simulation dataset in Section 3, and explain the methods and results in Section 4.

2 Literature Review

Here, we review two of the few instances in the literature of network theory applied to cosmology.

Ueda et al. [8] compare some network characteristics derived from the eigenvalues of a simulated adjacency matrix to the values seen in real sky data. This adjacency matrix represents the undirected constellation graph (k-nearest neighbor network for $k = 1$) of a 2-D projection of galaxies. This network is an example of a proximity graph, or a set of points with edges based solely on the geometric layout of the nodes. The network edges have weights w_{st} proportional to r_{st}^j, where r_{st} is the Euclidean distance between nodes s and t.

As described in Section 4, we will use a connected proximity graph, rather than a constellation graph. Proximity graphs like Delaunay triangulation are always connected, whereas $k = 1$ nearest neighbor

graphs tend to be highly disconnected, even when there is large-scale spatial structure. This choice will also keep the network resilient against missing or additional nodes, which happens often in observational data when dim or obscured galaxies go undetected. We employ a similar system for weighting edges, with $j \leq 0$, to indicate that proximate nodes should have a stronger connection than distant nodes.

In a more recent and thorough study, Hong and Dey [2] measure degree centrality, closeness centrality, and betweenness centrality on a network derived from observational Cosmological Evolution Survey (COSMOS) data. Nodes are galaxies, and two nodes are connected by an edge if they are closer than a linking length parameter ℓ. This proximity graph requires parameter-tuning and can produce unconnected networks, so we prefer connected proximity graphs, like the Delaunay triangulation.

Using the three measures of centrality, Hong and Dey classify nodes into 8 categories, which correspond to different topological features of the Cosmic Web. Degree centrality is used to assign the categories 'cluster,' 'wall,' and 'void,' as it's a proxy for local node density. Betweenness centrality assigns the labels 'main branch' and 'dangling leaf,' which distinguishes between galaxies forming the main filaments and galaxies at the periphery. Closeness centrality is used to assign labels 'fracture' (non-giant component nodes), 'backbone' (majority of giant component nodes), and 'kernel' (highest closeness giant component nodes).

Hong and Dey then look for differences in the distributions of galaxy properties like color, star formation rate, and stellar mass between the 8 galaxy types. They compare this network-based predictor against a baseline predictor that uses only Voronoi tessellation density, a measure of local node density. Their result was that network centrality works better for older, quiescent, "red" galaxies, while the baseline Voronoi cell density is a better predictor for more active, star-forming, "blue" galaxies. We will take a similar approach for this project, however, on a simulated dataset with other kinds of proximity graphs. We borrow their evaluation methodology of comparing new predictors to the baseline Voronoi predictor, using the Kolmogorov-Smirnov statistic (described in section 4.3).

3 Data

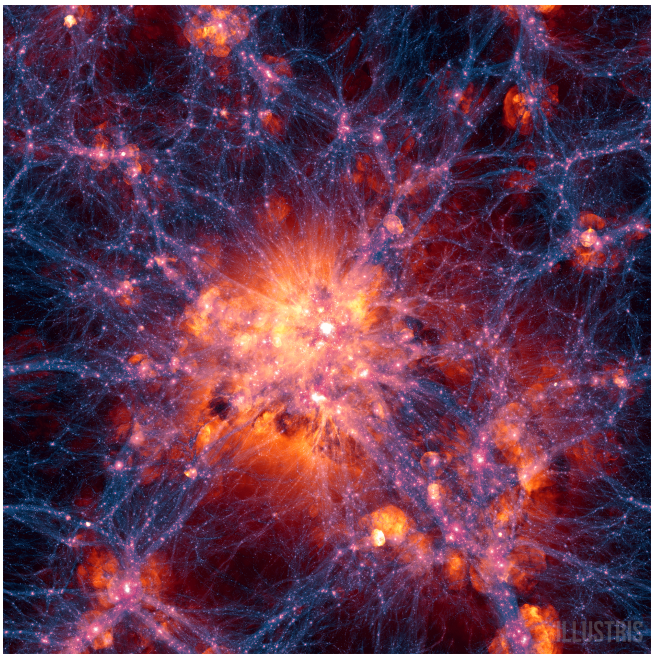

Figure 1: Illustris simulation at current epoch $z = 0$, showing dark matter density (in blue) overlaid with the gas velocity field (in orange).

The Illustris cosmological simulations [9] model a full suite of physical phenomena that can affect the evolution of galaxies, including gravity, hydrodynamics, gaseous chemistry, magnetic fields, stellar formation, galactic winds, and black hole-driven heating. There were six runs of Illustris, three with only dark matter under the force of gravity, and three with full hydrodynamical modeling, at varying levels of granularity. Due to computational complexity, we will use data from the coarsest granularity simulations, Illustris-3-Dark (only dark matter) and Illustris-3 (full hydrodynamic). For a sense of the model complexity, Illustris-3 tracked more than 282 million particles in a volume of $(106.5 \text{ Mpc})^3$, equivalent to a cube with side length 350 million light-years.

In early 2015, the Illustris team publicly released all 265 TB of Illustris simulation data [5]. An Illustris-3 snapshot, or all data at a given time, occupies over 20 GB, which is prohibitively large. Instead,

we use group catalog data, which has aggregated the roughly 282 million particles into a more manageable 112,000 clusters. These clusters, called subhaloes, correspond roughly to galaxies. Clustering is done via the subfind algorithm [6], which identifies locally overdense, gravitationally self-bound groups of particles. The group catalog has computed properties for each subhalo, such as mass, velocity (mass-weighted average of all constituent particle velocities), luminosity (magnitude of stellar light, measured at eight wavelengths), black hole mass, star formation rate (rate at which stars are newly formed), and metallicity (the proportion of matter consisting of elements heavier than helium).

4 Methods and Results

4.1 Data Cleaning

The process by which these subhaloes form, gravitational agglomeration, is a rich-get-richer process. The more massive a subhalo, the higher the rate of accretion, i.e., the more smaller subhaloes it subsumes. Thus, it's unsurprising that subhalo masses follow a power law distribution, as shown in green in figure 2. However, the power law fails at the low end, as there are too few subhaloes with mass below about $10^{10} M_\odot$ (solar masses). This is likely due to peculiarities in the subhalo clustering algorithm, so we clean the data by removing all subhaloes with mass less than $2 \cdot 10^{10} M_\odot$, shown in blue in figure 2.

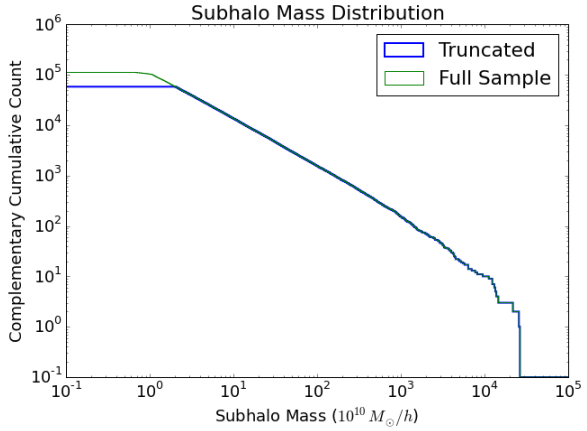

Figure 2: Distribution of subhalo masses. For comparison, the mass of the Milky Way is estimated to be $8.5 \cdot 10^{11} M_\odot$.

In order to better visualize the data, we work with $1/10^{\text{th}}$ slices of Illustris-3-Dark data. Suppose the simulation coordinates are $[0, L] \times [0, L] \times [0, L]$; then all points with $iL/10 \leq z < (i+1)L/10$ would comprise the i-th slice. To simplify the analysis, we project points into 2D by ignoring the z-coordinate.

4.2 Predicting Subhalo Mass

One of the most fundamental characteristics of a subhalo is its mass. Traditionally, galaxy mass measurements are computed from the orbital speed of stars around the galaxy center, which is inferred from Doppler shifts in the galaxy's spectrum. Our goal is to provide a way to differentiate galaxy masses based on only the most fundamental characteristic, the galaxy position. We encode the galaxy positions in proximity graphs; some examples were discussed in section 2. For this project, we consider the Delaunay triangulation in section 4.4 and the Gabriel graph in section 4.7.

4.3 Voronoi Tessellation Density Baseline

Given n nodes $\{x_i\}$, the Voronoi tessellation is a unique partitioning of the plane into n regions $\{R_i\}$, such that all points in R_i are closer to node x_i than to any other node. The volume of the Voronoi cell associated with a node provides an estimate for the local node density. Larger Voronoi cells mean a more sparse, void-like region. Just as Hong and Dey did, we will take Voronoi cell density to be the baseline, purely geometric, non-network predictor of subhalo properties.

Applied to the Illustris data, Voronoi cell area seems to do a reasonable job at differentiating the clusters (high density regions, in blue), the filaments (intermediate density regions, in red), and the void (low density regions, in green), as shown in Figure 3.

We use the 2-sample Kolmogorov-Smirnov (K-S) statistical test to evaluate the degree to which a variable x is good for distinguishing subhalo masses. The two samples are masses of subhaloes in the lowest quartile of x and subhalo masses in the highest quartile of x. The K-S test computes the maximum difference between the cumulative distribution functions (CDFs) of the two samples.

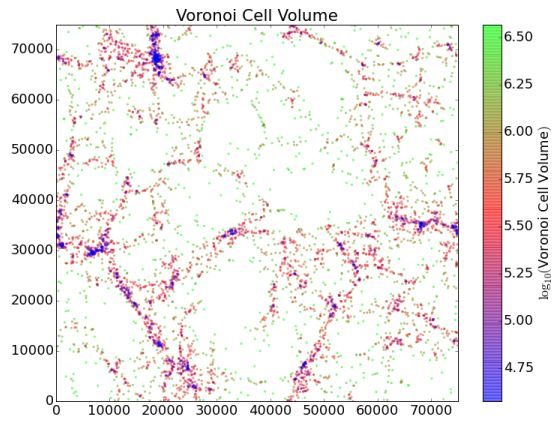

Figure 3: Nodes colored by Voronoi cell volume. Green represents large Voronoi cells and blue represents small Voronoi cells.

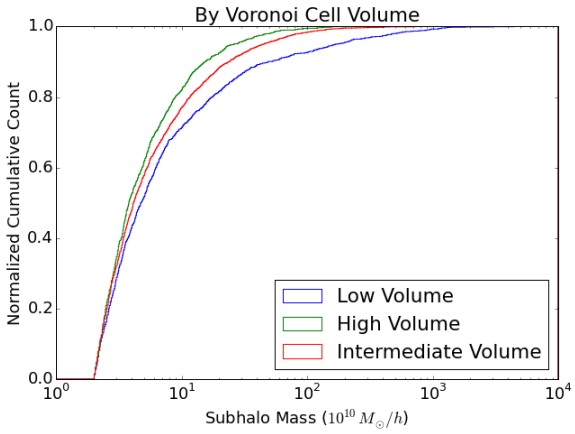

Figure 4: Distribution of subhalo mass, by Voronoi cell density

Higher K-S statistic means that the distributions are more different, or specifically, x is a better distinguishing variable for subhalo mass.

As shown in Figure 4, the mass distribution of lowest quartile Voronoi cell volume subhaloes (in blue) - those in locally dense regions - is centered higher than that of high Voronoi cell volume subhaloes (in red) - those in locally sparse regions. This is consistent with the rich-get-richer feedback of gravitational agglomeration. High-mass subhaloes tend to gravitationally bind more smaller subhaloes, and eventually, merge with them to grow even larger. Thus, we should expect to see high-mass subhaloes in denser regions, like clusters and filaments. The test statistic of the 2-sample K-S test is 0.1249, which translates to a p-value of $5.2 \cdot 10^{-10}$. This means that the difference between mass distributions in high and low density regions is statistically significant.

4.4 Delaunay Triangulation

The Delaunay triangulation is a unique partitioning of the entire space into triangles with the nodes as vertices, such that no node lies within the circumcircle of any triangle. We use existing Python libraries, but implement a modification that allows space to have toroidal topology, i.e., the bottom and top edges are associated and the left and right edges are associated. This matches the topology of space in the Illustris simulation. If a galaxy cluster contains subhaloes near $x = 0$, it will also likely contain subhaloes near $x = 75000$ (side length of the Illustris simulation cube). Our modified Delaunay triangulation allows edges between those groups of subhaloes, while the standard implementation does not, as they're considered very distant.

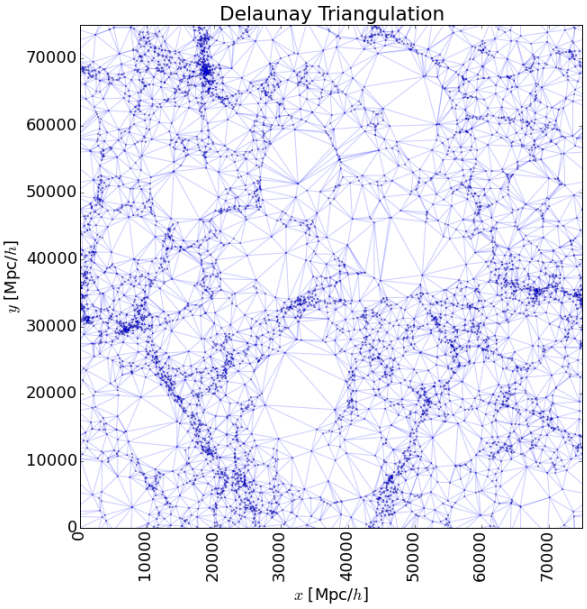

Figure 5: Delaunay triangulation of 2D slice of Illustris-3-Dark run.

4.5 Degree Centrality

We take the edges of the Delaunay triangulation to be edges in an undirected network. We also assign

edge weights w_{st} proportional to r_{st}^j, where r_{st} is the Euclidean distance between nodes s and t.

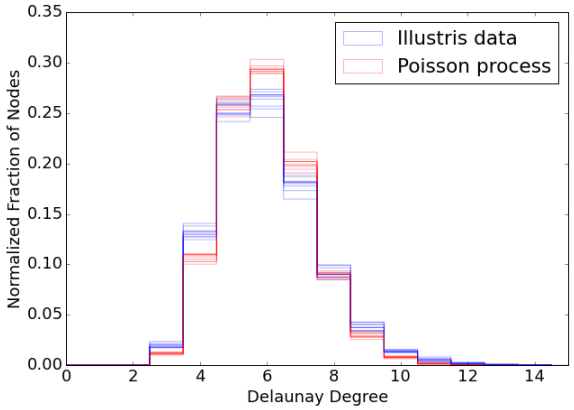

Figure 6: Unweighted degree distribution ($j = 0$) of Delaunay triangulation, for each of 10 slices.

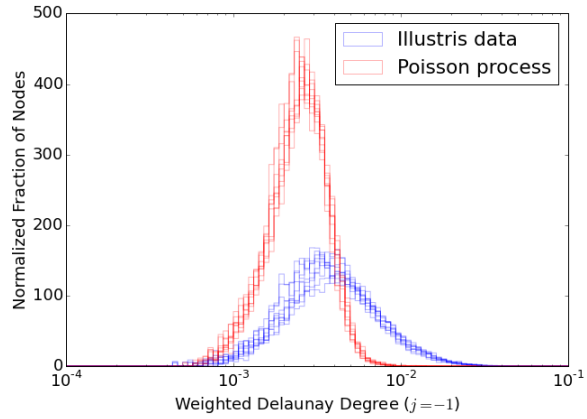

Figure 7: Weighted degree distribution ($j = -1$) of Delaunay triangulation, for each of 10 slices.

In figures 6 and 7, we compare weighted ($j = -1$) and unweighted ($j = 0$) degree distributions of the Delaunay triangulation of Illustris subhaloes with that of an homogenous Poisson process of the same size, i.e. equal number of uniformly randomly distributed subhaloes. It's clear that for Illustris data, the degree distributions (both weighted and unweighted) are wider and heavier-tailed than for the Poisson process. This is exactly the highly skewed degree distribution that arises from a rich-get-richer, preferential attachment model. There are large "hub" subhaloes, which tend to be in dense regions and close to other nodes, which

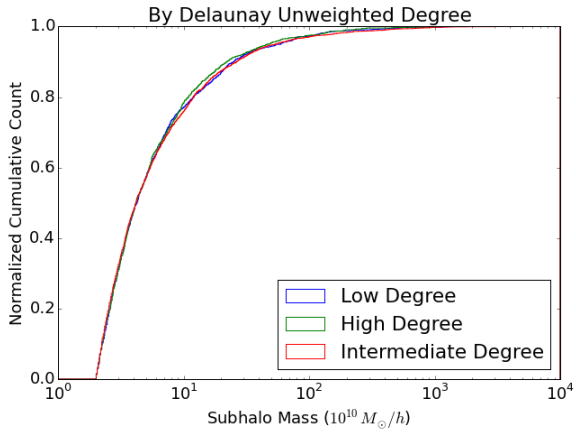

Figure 8: Distribution of subhalo mass, by unweighted degree.

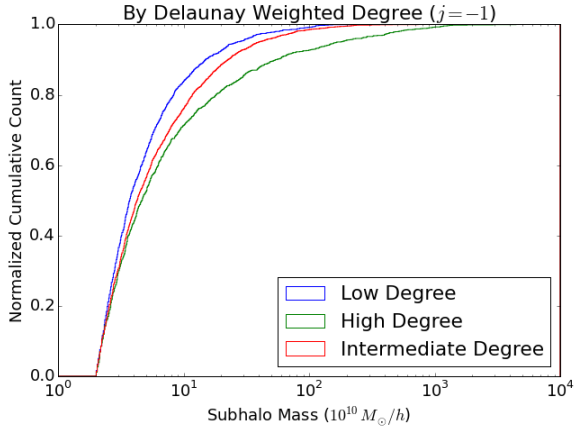

Figure 9: Distribution of subhalo mass, by weighted degree.

leads to higher weighted degree.

From figure 8, we get no statistically significant differentiation of subhalo mass when categorizing by unweighted degree: K-S statistic 0.0243, p-value 0.80. However, as shown in figure 9, we get great differentiation by $j = -1$ weighted degree: K-S statistic 0.131, p-value $5.1 \cdot 10^{-11}$. In fact, this is better differentiation than the baseline of categorizing by Voronoi cell volume.

We perform the same analysis from $j = -9$ (heavily discount long edges) to $j = 0$ (all edges weighted equally), and across all 10 slices of the simulation cube. We compare the K-S statistic for subhalo mass distribution using weighted degree centrality, versus using the baseline Voronoi cell volume. The

difference in K-S statistics is plotted in figure 10, with higher values indicating that degree centrality is better at distinguishing subhalo mass than Voronoi volume. The advantage peaks near $j = -2$, but is fairly stable for $j < -1$.

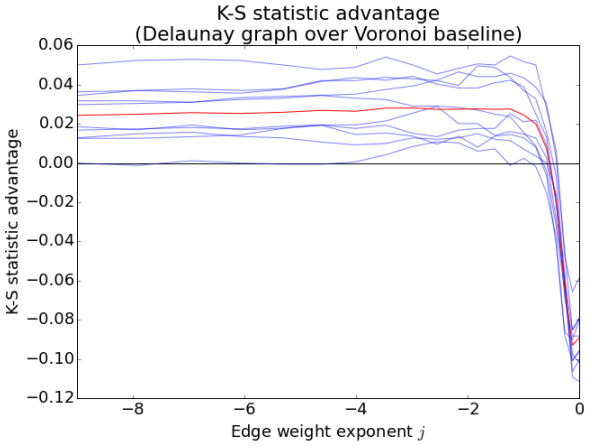

Figure 10: K-S test statistic advantage for distinguishing subhalo mass; weighted degree centrality compared to baseline Voronoi cell density. The average advantage over all 10 slices is shown in red.

4.6 Predicting Subhalo Luminosity: a second-order effect

Many of the other subhalo properties one might want to predict, such as luminosity, black hole mass, star formation rate, metallicity, and maximum radial velocity (speed of stars orbiting galactic center), are strongly correlated with mass. We've shown that degree centrality is a good predictor of subhalo mass, so it will also likely be a good predictor of these other properties. However, we now wish to investigate a second-order effect: how well does degree centrality distinguish residual luminosity (residual compared to other galaxies of the same mass)?

First, we filter out all subhaloes without luminosity estimates, and all subhaloes with mass less than $10^{11} M_{\odot}$ (luminosity estimates are often unreliable at low masses). In figure 11, we see the expected strong correlation between luminosity and mass - more massive galaxies are brighter. The red curve is computed using locally weighted scatterplot smoothing (LOWESS), and we take the residual luminosity with respect to that LOWESS estimate.

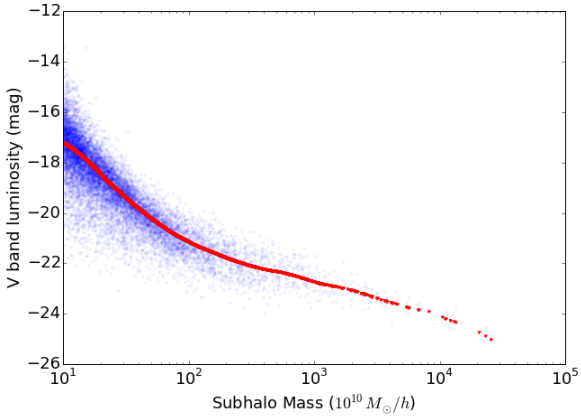

Figure 11: Luminosity, as a function of subhalo mass. Note that luminosity units (mag) are on a logarithmic scale, and lower values indicate higher luminosity.

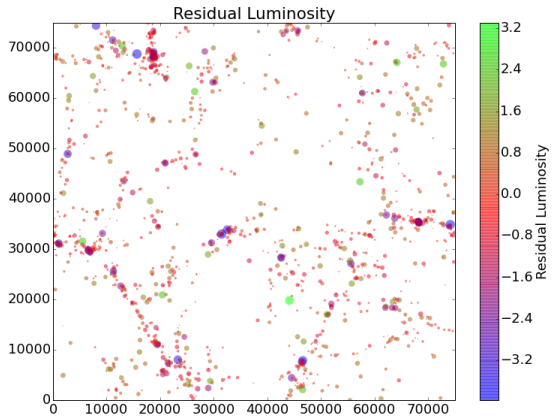

Figure 12: Residual luminosity in a slice of Illustris-3 data. Nodes are larger if the residual is further from zero.

In figure 12, we see that subhaloes with negative residual luminosity (brighter than expected for its mass) tend to lie in higher density clusters, while dimmer-than-expected subhaloes tend to lie in lower density filaments and voids. This can be explained physically, because high density clusters tend to have higher densities of coalescing gas, which drives star formation. Star-forming regions contain many hot, bright stars, which burn out quickly, while quiescent regions contain mostly old, cool, dim stars. Figure 13 shows that even for this kind of second-order property, degree centrality outperforms the baseline Voronoi predictor.

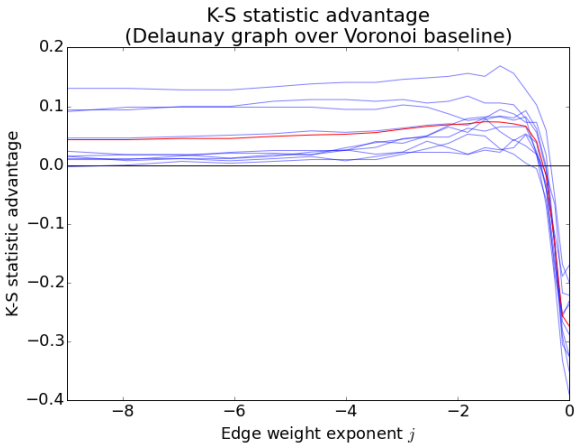

Figure 13: K-S test statistic advantage for distinguishing residual luminosity; weighted degree centrality compared to baseline Voronoi cell density.

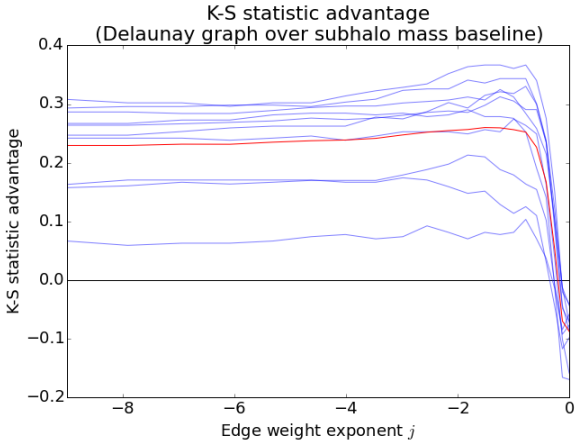

Figure 14: K-S test statistic advantage for distinguishing residual luminosity; weighted degree centrality compared to subhalo mass.

One might posit that the heteroskedasticity evident in figure 11 might result in high K-S statistic. This might be due to differences in distribution variance, rather than differences in distribution mean, which are much more useful when trying to distinguish two populations. However, figure 14 refutes this claim, by comparing the distinguishing power of degree centrality compared to subhalo mass. Categorizing by degree centrality results in a much higher K-S statistic than categorizing by mass, so the heteroskedasticity contributes negligibly to the K-S statistic.

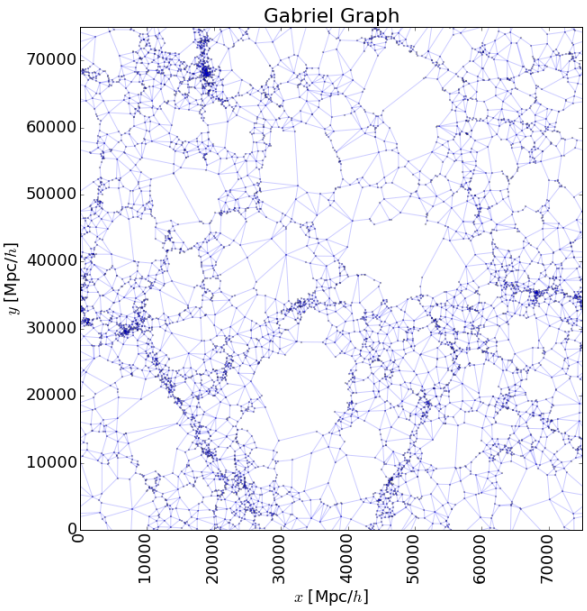

Figure 15: Gabriel graph of 2D slice of Illustris-3-Dark run.

4.7 Gabriel Graph

The Gabriel graph is a subgraph of the Delaunay triangulation. It is also a proximity graph, but with the rule that an edge between node S and node T exists iff the closed disk with diameter ST intersects no other nodes. We implemented in Python a linear time construction, proposed by Matula & Sokul [4], that takes the Delaunay triangulation as input.

Comparing figures 5 and 15, we see that the Gabriel graph is considerably more sparse. For planar uniformly distributed points, the average degree of the Delaunay triangulation is 6, while the average Gabriel graph degree is 4. However, it's clear visually that mostly long edges have been removed, so the important filamentary structure is still expressed.

Shown in figure 16, weighted degree centrality on the Gabriel graph shows slightly better K-S advantage than on the Delaunay triangulation. However, this difference is not statistically significant. This is expected, as the removal of long edges has little effect on weighted degree centrality for small $j << 0$. Thus, there is little difference between using the Delaunay triangulation and the Gabriel graph.

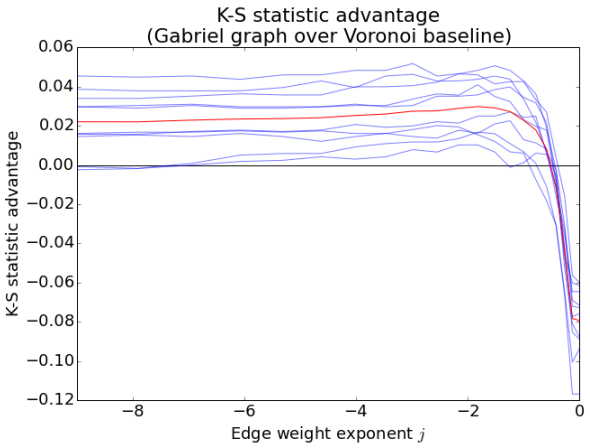

Figure 16: K-S test statistic advantage, weighted degree centrality on Gabriel graph compared to baseline Voronoi cell density. The average advantage over all 10 slices is shown in red.

5 Conclusion

We have shown that, for the task of distinguishing subhaloes with different masses, a spatial network-based subhalo property, like weighted degree centrality on the Delaunay triangulation, will outperform the baseline non-network-based Voronoi cell volume property. We have shown that degree centrality distinguishes both first-order properties like mass, as well as second-order properties like residual luminosity. This work hopefully demonstrates that there is room to explore the sparse juncture between network analysis and cosmology.

References

[1] Marc Davis et al., "The Evolution of Large-Scale Structure in a Universe Dominated by Cold Dark Matter". The Astrophysical Journal 292 (May 1985), 371-394.

[2] Sungryong Hong and Arjun Dey, "Network analysis of cosmic structures: network centrality and topological environment". Monthly Notices of the Royal Astronomical Society 450 (June 2015), 1999-2015.

[3] Valérie de Lapparent, Margaret J. Geller, John P. Huchra, "A Slice of the Universe". The Astrophysical Journal 302 (March 1986), L1-L5.

[4] David W. Matula and Robert R. Sokal, "Properties of Gabriel Graphs Relevant to Geographic Variation Research and the Clustering of Points in the Plane". Geographical Analysis 12 (July 1980), 205-222.

[5] Dylan Nelson et al., "The Illustris Simulation: Public Data Release". Astronomy and Computing 13 (November 2015), 12-37.

[6] Volker Springel et al., "Populating a cluster of galaxies - I. Results at z = 0". Monthly Notices of the Royal Astronomical Society 328 (2001), 726-750.

[7] Volker Springel et al., "Simulations of the formation, evolution and clustering of galaxies and quasars". Nature 435 (June 2005), 629-636.

[8] H. Ueda, T. T. Takeuchi, and M. Itoh, "A graph-theoretical approach for comparison of observational galaxy distributions with cosmological N-body simulations". Astronomy & Astrophysics 399 (2003), 1-7.

[9] Mark Vogelsberger et al., "Properties of galaxies reproduced by a hydrodynamic simulation". Nature 509 (May 2014), 177-182.