Journal of Student Research

Volume 10 Issue 3 (2021)

Detection and Analysis of Galaxy Clusters Via a Hierarchical Algorithmic Approach

Samantha Liu[1], Pranav Eswaran[1] and Shyamal Mitra[1#]

[1]University of Texas, Austin, TX, US
[#]Advisor

ABSTRACT

This paper is a discussion of our analysis of galaxy clustering in space using an algorithmic approach. Our algorithmic galaxy clustering analysis and galaxy morphology analysis produced promising results in identifying galaxy clusters at different scales, and we used these clusters to draw correlations between cluster membership and galaxy properties such as size and color. We also compare our work in algorithmic galaxy clustering to existing work using machine learning, showing where our results are consistent with previous work, and where they differ from previous work. Overall, we found our research to be insightful into how algorithms perform when finding clusters of galaxies, and we find many possible follow up questions to explore in the future.

Introduction

Beginning in the 20th century, there have been great efforts to catalog galaxies and clusters of galaxies (Abell, 1958). The categorization of clusters of galaxies in the Universe has greatly aided astronomers in understanding the formation of stars, galaxies, and the Universe itself, informing much of the current knowledge on dark matter and dark energy, among other topics (Allen et al., 2011). As a part of this ongoing effort to characterize the clustering behavior of galaxies, we aim to evaluate the effectiveness of algorithmic techniques to organize galaxies into clusters; what algorithms do best at finding clusters, in the quality of the clusters found, the time taken to find said clusters, and how these algorithms scale to accommodate the vast amount of astronomical data there is to be analyzed. We analyze some correlations between galaxy clustering and galaxy properties/morphology, such as size and color. We also endeavor to compare our algorithmic technique's results to existing catalogs of galaxies, and to see how our algorithmic techniques could be improved in the future.

Methods

The Sloan Digital Sky Survey Dataset

For our analysis, we used the Sloan Digital Sky Survey (SDSS) dataset. According to their website, the Sloan Digital Sky Survey has created "the most detailed three-dimensional maps of the Universe ever made." With the publishing of Data Release 16 in December 2019, the SDSS galaxy database is one of the most detailed data sets of surveyed galaxies in the Universe, making it very suitable for our research.

We used this dataset in previous work in our preliminary analysis of galaxy clustering phenomena in the area of the sky between a right ascension 100 and 250 degrees, but found that we restricted the size and brightness of our data subset too greatly to really extract quality algorithmically identified galaxy clusters that are on par with existing

human-created galaxy cluster catalogs; to remedy this issue of inadequate density of galaxies in our dataset, we greatly loosened the criteria for a galaxy in our dataset in terms of its brightness and size, expanding our dataset from 46,171 galaxies to over 652,701 galaxies.

We chose the particular region of the sky between a right ascension of 100 and 250 degrees because this is the region of the sky that has previously been cataloged by astronomers like Abell in the past. This provides us a very useful comparison and validation tool which we will discuss later in this paper.

We use galaxy redshift as a proxy for a galaxy's distance from the Earth. This follows from Hubble's law, which states that an object's redshift as viewed from Earth is proportional to its distance from Earth. Due to the Malmquist Bias, the phenomenon where perceived galaxy density decreases as redshift increases, galaxies with large differences in redshifts cannot easily be directly compared, so we restricted our dataset to galaxies with redshifts between 0.02 and 0.04. This takes the number of galaxies our algorithms are working on from 652,701 down to 49,987. This slice still has an adequate density of galaxies for our algorithmic approach because we increased the number of galaxies as mentioned earlier. This also greatly helps the time it takes to run our analysis with our algorithm. We don't have statistics to back up this time-save claim because our algorithm never was able to fully run on the largest dataset, but seeing as how our algorithm ran on this smaller slice, we believe we made the right choice in using this slice.

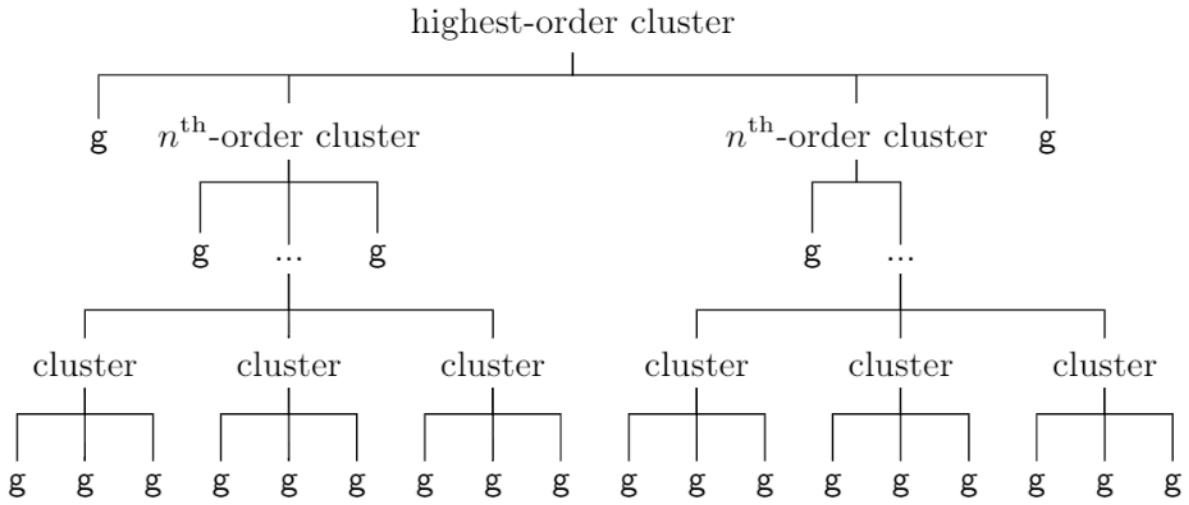

Figure 1. Toy Hierarchy Tree Diagram. This is a simplified example of what our hierarchy tree looks like at a high level. The root node of this tree is the highest-order cluster which contains every galaxy in the dataset we are analyzing. Each leaf node g represents an individual galaxy. An n-th order cluster is a cluster that contains within it clusters of order n-1, and a cluster of galaxies is considered a first order cluster. Subtrees of this hierarchy tree can be collapsed to their respective root nodes when graphing and analyzing our algorithm's cluster outputs; this can help comparisons between the algorithmically identified clusters and existing cluster catalogs like that of Abell. Note that this is an n-ary tree, so not every cluster has the same number of subclusters, and first order clusters have varying numbers of member galaxies; subtrees can also be of varying depth (since some regions may have more higher-order clustering phenomena than others

The OPTICS Algorithm

The OPTICS (Ordering Points To Identify the Clustering Structure) algorithm works very similarly to the better-known density-based algorithm known as DBSCAN, the basic idea of which is to mark points as a cluster if there are more than a certain number of points within a set distance epsilon of each other. We use the OPTICS algorithm to cluster based on the physical location of galaxies (as given by declination, right ascension, and redshift). In our usage of OPTICS, the epsilon parameter corresponds to the maximum physical distance between two galaxies to be considered in the same cluster. Unlike DBSCAN however, OPTICS also maintains a hierarchy of clusters, i.e. it stores a list of clusters and then a cluster of those clusters and so on; every datapoint is part of the most expansive supercluster. This avoids the need to provide an epsilon hyperparameter for the algorithm, as it figures out the optimal epsilon through this hierarchical classification instead. The OPTICS algorithm is also preferable to other algorithms such as k-means and mean-shift because it elegantly handles cases with clusters of greatly varying densities and shapes, which naturally occur in our dataset. Furthermore, OPTICS is less memory-intensive than DBSCAN, making it suitable for larger datasets such as ours.

Methodology

We preprocessed our data by converting the polar celestial coordinates (right ascension, declination, and z) of the galaxies to cartesian coordinates, which we then provided to the OPTICS algorithm as the metric to use for distance computation. After comparing different values for the hyperparameters of the OPTICS algorithm, we settled on using min_samples=10 and xi=0.03. In other words, our algorithm detects clusters with a minimum size of 10 galaxies, and the xi parameter determines the minimum change in reachability between two galaxies that constitutes a cluster boundary to be 1 minus xi, or 0.97 in our case. We found this configuration to have the best level of sensitivity for our purposes.

We first ran the OPTICS clustering algorithm on our SDSS dataset to get clusters and a cluster hierarchy of galaxies. We converted this cluster hierarchy into a tree with the root node containing every galaxy in the dataset and each leaf node containing only galaxies classified together as a low-level cluster. Internal nodes represent mid-level clusters in the cluster hierarchy. Then, using this tree, we categorized each galaxy's "depth" based on how deeply into the tree it was placed into a cluster. For example, if a galaxy was recognized as part of a cluster at depth 1 but not at depth 2, that galaxy had a clustering depth of 1. If the galaxy was recognized as part of a cluster at depth 2 but not at depth 3, then the galaxy had a depth of 2. Figure 1 illustrates this cluster hierarchy.

Results

Overall, our algorithm classified 29,443 of 49,987 galaxies as part of 765 highest-order clusters; in other words, we found that 59% of galaxies are in clusters, and that the average cluster size was about 40 galaxies. We detected 6 levels of hierarchy in clustering.

We looked at the size and color of galaxy clusters to see how the order of a cluster affected the average size and color of its member galaxies. We categorized each galaxy's "depth" based on how deeply into the tree it was placed into a cluster. Clustering depth essentially separated the galaxies into categories which could be used as proxies for how central they were in their clusters and how densely clustered they were. We used these categories to analyze correlations between clustering centrality and galaxy properties like size and color. Since the last depth, depth 7, only had 11 galaxies in it, we deemed that the sample size was too small to draw meaningful results from, so we ignored it for the purposes of our analysis. Note that the deeper into the hierarchical tree a cluster is, the smaller and more dense that cluster is; high order clusters are found at lower depths in the hierarchical tree.

Table 1. Mean cluster size at each depth of the tree, root excluded. The average top-level galaxy cluster had a size of about 40 galaxies.

Clustering Depth	Mean Size of Cluster (# Galaxies)
1	38.49
2	31.45
3	29.17
4	26.98
5	28.27
6	20.73

Size of Member Galaxies

The size of galaxies seems to trend upwards as depth increases, but then flattens out and dips back down, as seen in Figure 2. We expected galaxy size to positively correlate with being more centrally located within clusters, because bigger galaxies will have greater forces of gravitational attraction to themselves and to other galaxies, potentially making them more prone to clustering closer to other galaxies. More central galaxies in more dense areas of space are also more likely to collide to form larger galaxies. However, our analysis does not indicate a positive correlation between cluster depth and galaxy size. Since our analysis is inconsistent with previous findings, we speculate that our sample size may have been too small to detect a meaningful pattern in galaxy size, especially at higher depths in the clustering tree.

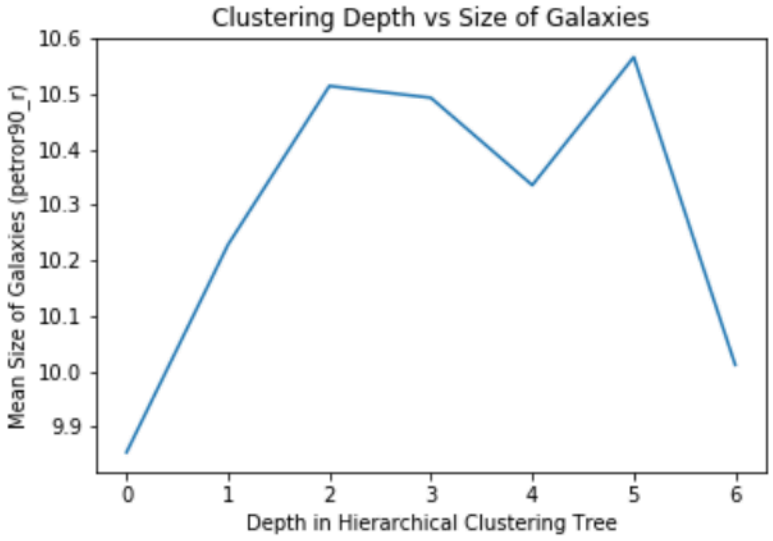

Figure 2. Graph of Clustering Depth versus the Size of Galaxies. Contrary to expectation, the mean size of galaxies seems to be positively correlated with the depth of clusters in the hierarchical clustering tree only up until depth 2, then drops off.

Color of Member Galaxies

We found a correlation between the depth of clusters in our hierarchical cluster tree and the average color of member galaxies in these clusters, as seen in Figure 3. We see that the g-r and r-i values of a galaxy cluster are positively correlated with the depth in the hierarchical clustering tree of that galaxy cluster. The u-g colors are positively correlated with the depth of the cluster as well, with the exception of an extreme jump in mean u-g color from depth 0 to 1. A higher u-g and g-r color value indicates that a galaxy is redder; thus, our analysis suggests that galaxies that are more central in their clusters and more densely clustered tend to be redder. This suggests that galaxies in dense clusters are more likely to be older, elliptical galaxies, because elliptical galaxies tend to be redder than spiral galaxies. This is consistent with our knowledge about how galaxy clusters form; the gravitational pull of the galaxies causes the galaxies to collide together into large, red elliptical galaxies.

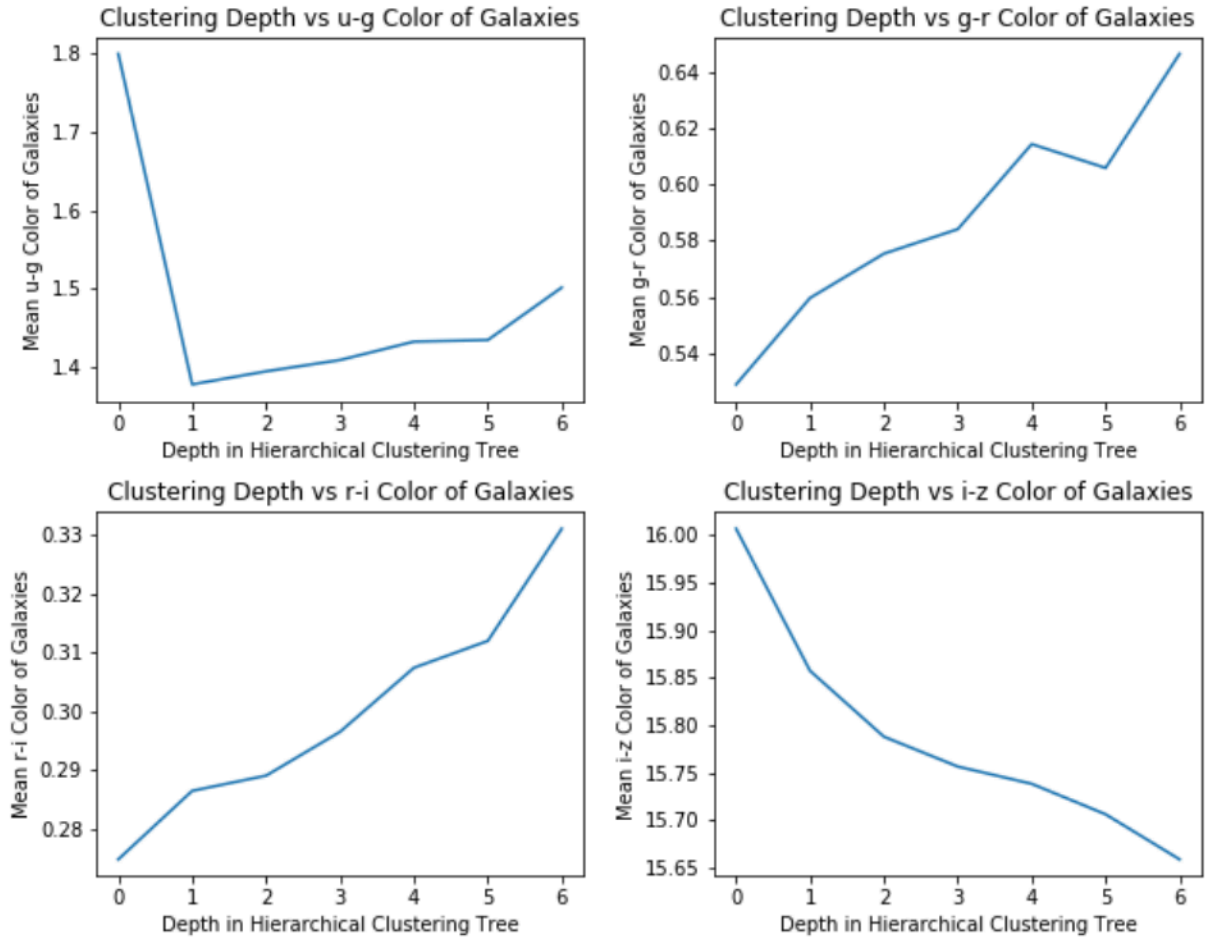

Figure 3. Graph of Clustering Depth versus the u-g, g-r, r-i, and i-z Colors of Galaxies. This shows the correlation between the clustering depths of galaxies and the four galaxy colors. We see that in general, more central and densely clustered galaxies tend to be redder.

Discussion

To assess the efficacy of our algorithmic approach to identifying galaxy clustering phenomena, we compare the clusters output by our algorithmic techniques to existing work on galaxy clusters. The Abell catalog is a highly regarded catalog of 4,073 rich galaxy clusters (Abell 1958). Compared to Abell's catalog, our method tended to break up large Abell clusters into smaller subclusters, especially along the redshift axis. This is partly because our method used redshift to calculate distance, while Abell did not. This result is shown in Figure 4. We also detected many more clusters than Abell overall, as shown in Figure 5, mostly owing to a more modern dataset. For the highest order clusters from the OPTICS hierarchical clustering output for our redshift slice, 29,443 galaxies or 59.05% of galaxies were part of a cluster, which is consistent with the current understanding that the majority of galaxies are members of a cluster. Our results are also consistent with previous work done by Santiago-Bautista et al. (2020) in finding that central galaxies in clusters and filaments tend to be redder and thus more elliptical than outer galaxies and field galaxies. However, we failed to replicate their finding that central galaxies are more massive than outer galaxies.

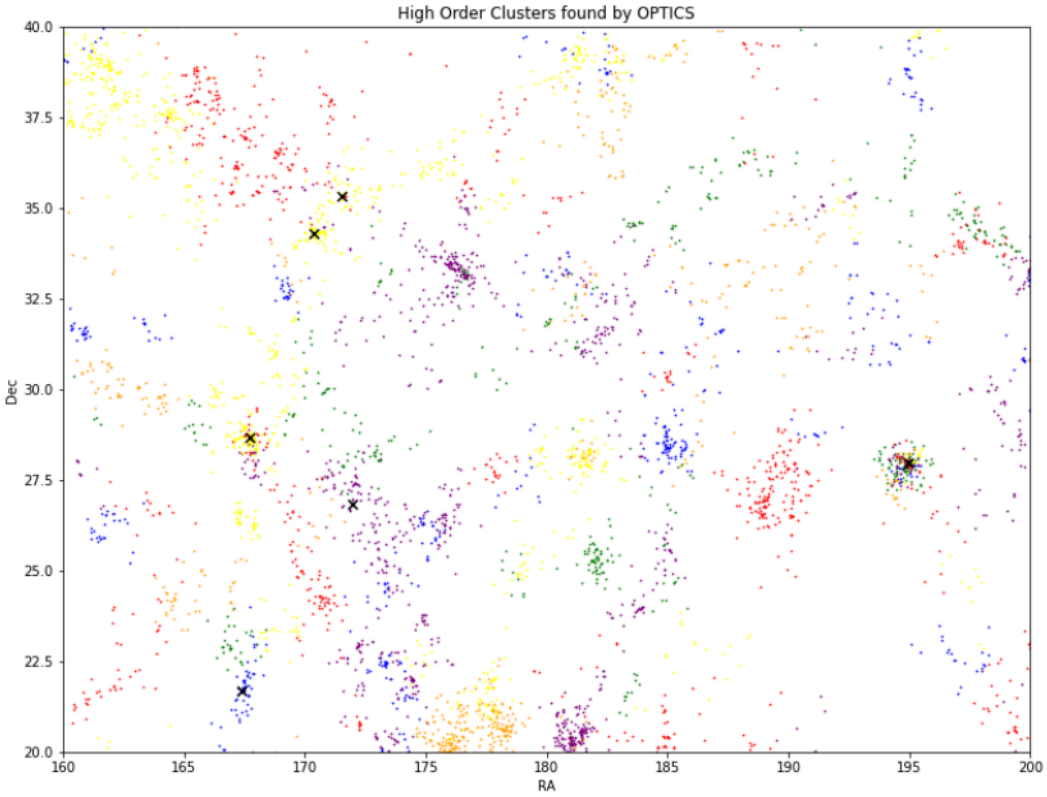

Figure 4. Zoomed-in plot of Abell clusters (black X's) and galaxies (dots) in the region of the sky between 160 and 200 degrees right ascension and between 20 and 40 degrees declination. Galaxies near other galaxies of the same color marker indicate that these galaxies are part of the same higher-order cluster. Clusters shown are the highest order clusters from the OPTICS hierarchy that aren't the root node of the hierarchy tree.

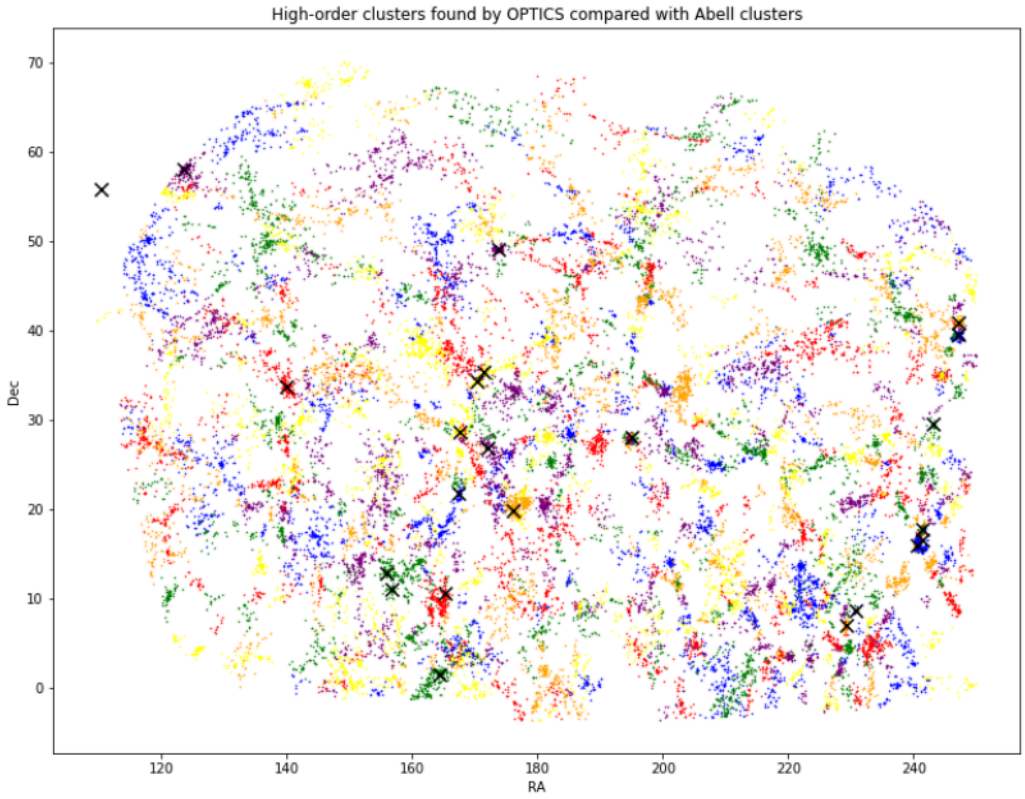

Figure 5. The region of the sky between 100 and 260 degrees right ascension and between -10 and 80 degrees declination. Galaxies are represented by dots and Abell clusters are represented by black X's. Galaxies near other galaxies of the same color marker indicate that these galaxies are part of the same higher-order cluster. Clusters shown are the highest order clusters from the OPTICS hierarchy that aren't the root node of the hierarchy tree.

Conclusion

By our analysis, the average galaxy cluster contains around 40 galaxies, and shows up to 6 hierarchical levels of clustering. Our research has contributed to the body of evidence for the claim that central galaxies in galaxy clusters tend to be redder and more elliptical and that the majority of galaxies are parts of clusters. However, we failed to replicate the finding of others that central galaxies are more massive than outer galaxies and field galaxies.

We found that many typical machine learning algorithms such as k-means, mean-shift, and DBSCAN were unsuitable for analyzing astronomical datasets, and even with OPTICS we had to greatly reduce the size of our dataset due to scalability issues. In future work, we hope to expand our analysis to a larger dataset and to test out other clustering algorithms and combinations of algorithms such as agglomerative hierarchical clustering and minimum spanning trees. More work is also needed to determine a good way to remove the Fingers of God effect and other redshift distortions in preprocessing. We would also like to work more on overcoming the limitations posed by

Malmquist bias, which poses a selection effect on distant galaxies for intrinsically brighter galaxies. Finally, we would like to explore more correlations between cluster membership, different cluster types, and galaxy morphology.

Acknowledgements

We would like to thank the Geometry of Space Freshman Research Stream at the University of Texas at Austin for giving us the opportunity and support to pursue this research. It has truly been an amazing experience for us!

References

Abell, G. O. (1958). The Distribution of Rich Clusters of Galaxies. The Astrophysical Journal Supplement Series, 3, 211. https://doi.org/10.1086/190036

Allen, S. W., Evrard, A. E., & Mantz, A. B. (2011). Cosmological Parameters from Observations of Galaxy Clusters. Annual Review of Astronomy and Astrophysics, 49(1), 409–470. https://doi.org/10.1146/annurev-astro-081710-102514

Santiago-Bautista, I., Caretta, C. A., Bravo-Alfaro, H., Pointecouteau, E., & Andernach, H. (2020). Identification of filamentary structures in the environment of superclusters of galaxies in the Local Universe. Astronomy & Astrophysics, 637, A31. https://doi.org/10.1051/0004-6361/201936397

SDSS. (n.d.). https://www.sdss.org/