Using Networking Algorithms to Assess the Environment of Galaxy Groups

Ali M. Bramson, Eric M. Wilcots

Department of Astronomy, University of Wisconsin-Madison, Madison, WI

Abstract
Understanding the environment in which a galaxy resides is crucial to our understanding of galaxy evolution. Most galaxies (~70%) live in groups and it is important to develop a quantitative understanding of how galaxies are distributed within groups and how groups are distributed in the larger scale structure. In addition to the traditional friends-of-friends algorithm, I have applied various concepts from networking algorithms to understand the network of galaxy groups in clusters, and to understand the substructure within groups. Doing so provides a new method of gauging the large scale environments of galaxies and groups, and will result in a more concrete way to define a group and to quantify the strength of the community structure that exists.

Background

Galaxy structures Galaxies are the fundamental unit of the universe, and most reside in groups, seen in Figure 1, or the even bigger clusters. Our own Milky Way Galaxy resides in a small group of galaxies creatively called the "Local Group". Approximately 70% of galaxies reside in groups, qualitatively defined as a gravitationally bound system with a mass of $10^{13} – 10^{14}$ solar masses, in between that of an individual galaxy and a cluster.

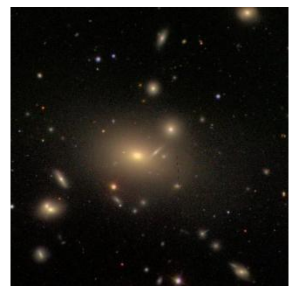

Figure 1: Optical image from Sloan Digital Sky Survey (SDSS) of a galaxy group.

The evolution of galaxies occurs in groups. The first step for learning about galaxy evolution is to establish what these groups are. The environment of a galaxy greatly affects its development, and the properties of galaxies, including luminosity, color/morphology and star formation rate, depend on their surroundings.

Current group finding techniques Currently there are several ways of quantitatively identifying groups, but these methods do not yield consistent results. In addition, the distribution of galaxies within a group is not homogenous; groups themselves can be clumpy. Different techniques might identify the clumps as separate groups while another technique would simply find the larger group (Fig. 2). Other techniques depend on just looking at the number of bright galaxies within a specific radius [10].

A common technique uses the friends-of-friends (FoF) algorithm [4]. FoF starts with one galaxy and then looks for nearby galaxies that satisfy the linking criterion. It then looks at these galaxies, iterating until the nearest galaxy exceeds some specified threshold of distance or velocity [4]. The galaxy group is then classified as all of the linked galaxies. FoF succeeds at finding concentrations of galaxies but falls short in determining where one group of “friends” ends and another begins.

This research was supported by the Wisconsin Space Grant Consortium Undergraduate Research Scholarship and the UW-Madison Undergraduate Research and Mentoring Program (NSF grant DBI-0731564).

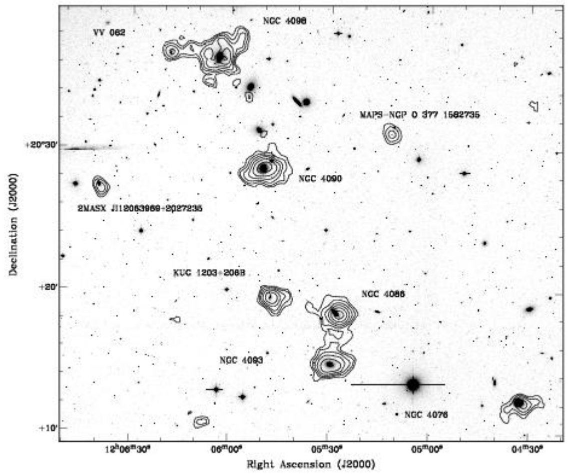

Figure 2: Image showing contours of neutral hydrogen gas with grayscale being optical observations from the Digitized Sky Survey. This is one group as determined by the FoF method but could be broken into two groups [3].

Problems with the current methods While useful, this FoF algorithm cannot adequately distinguish between two close groups. Another downfall to the current methods of identifying galaxy groups is that the various different techniques yield different sets of groups, and many automated group finders identify far too few galaxies as members. For example, most of the groups in the recent catalogs consist of only three or four galaxies [1] while nearby groups are known to contain tens of galaxies.

Objectives for the current work Determining which galaxies belong in which group is not the only important thing to learn. Knowing how tightly linked galaxies are within and between groups will be important for figuring out the scales on which galaxy interactions occur. Finding the most important spatial scale on which the environment influences galaxy evolution is crucial to understanding how galaxies evolve. Subgroups indicate that a group is dynamically young and, possibly, more susceptible to galaxy-galaxy mergers. Determining substructure is important for learning if groups are built up of subgroups like clusters are built up by groups.

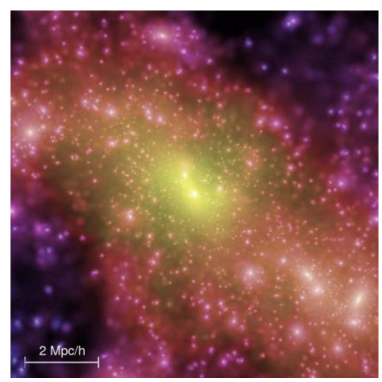

Figure 3: This simulation from the Millenium Simulation Project [8] shows the complicated network of galaxies that exists. Bright spots indicate areas of high mass, with the brightest spots representing galaxy groups. The boundaries between the galaxies are not well defined and there are many connections between.

Science of networks as a solution Simulations of the formation of structure in the universe, as shown in Figure 3, show that links between galaxies are numerous and complicated. Determining the strength of these links and the breakage points in which we can definitively say where galaxy groups separate is important in learning about the environment galaxies live in. What is really needed are observational methods to understand how linked galaxies and groups are, and that is where this project comes in. The science of networks has been developing algorithmic detection methods for determining community structure in various settings for over a decade. Seeing what networking algorithms can tell us about the physical world is a very new idea with many potential applications. I have applied these techniques to the distribution of galaxies, but the applicability could be much broader.

Networking algorithms Community structure in networks can be seen across all disciplines: natural, social and information sciences [5]. An example is in the case of the popular Zachary's Karate Club network, a benchmark network [5]. In this real-world example, social interactions were observed between an American university's club members, including when the club split into two during a dispute between two of the club's leaders [6]. Algorithms developed were able to correctly model these interactions and properly showed a strong natural division when the club split. This has been applied to many other data [2], including communities on social networking websites such as Facebook (www.facebook.com) as well as analyzing individual and group voting dynamics in Congress [5]. Community structure and networks have been worked on for a long time in these other disciplines; now we have applied them to astronomy!

Methods

Modifying friends-of-friends to identify centers of groups I wrote an algorithm that works like the traditional friends-of-friends algorithm [4], searching out galaxies that are connected to a given "seed" galaxy through some path of other galaxies that are all linked by a given distance. This distance criterion is chosen based on the average separation between the galaxies occupying a given area in the data set. My algorithm keeps track of the galaxies found at each step away from the seed galaxy. It sums the total number of friends every galaxy has at each step, as a measure of how closely connected the group is to the seed galaxy. Outputting the results in this way (Figure 4) has provided us with a new way to look at FoF, and be able to use it to look for the best central galaxy as a start to finding structure.

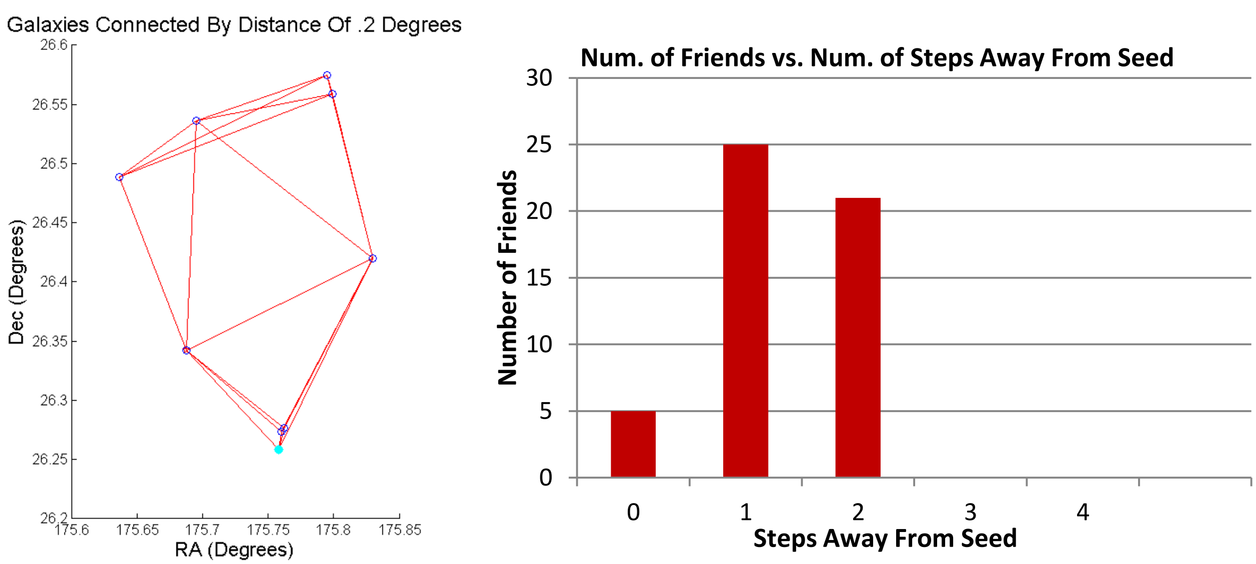

Figure 4: Diagrams showing connections (left) starting with the seed galaxy (bottommost galaxy) and the corresponding histogram (right) showing the numbers of "friends" at each step away from the seed galaxy.

Interpretation One example of a galaxy group structure that we have looked at with the current algorithm is NGC 2563. Zabludoff & Mulchaey [9] identified well over 20 individual galaxies that belong to the group. The results of our algorithm suggests that NGC 2563 is likely two separate groups, with a few galaxies connecting the two, and a few others not being included at all. The shape of histograms showing the number of connections at steps away from the "seed" galaxy can help identify the best central galaxy. Histograms that peak at one step away and fall rapidly indicate good "seeds", while distributions that are shifted indicate that the galaxy is on the outskirts. These histograms for NGC 2563 are shown in Figure 5.

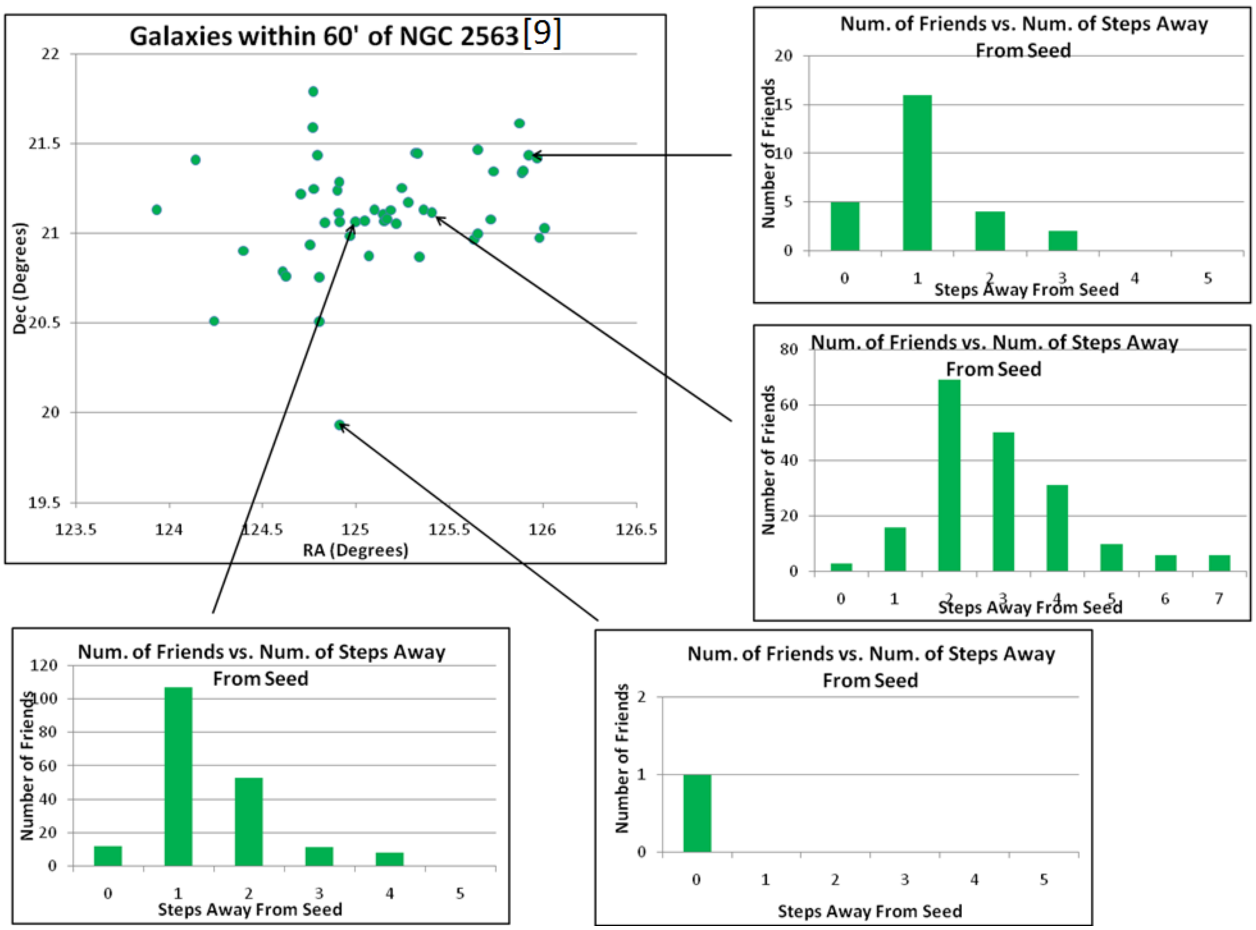

Figure 5: This shows the outputted histograms from our friends-of-friends algorithm for galaxies located in different parts of the group. These histograms show the number of friends that the galaxies at a given step number away from the "seed" have, in sum. Good "seed" galaxies at the center of a group peak at one step away and fall off rapidly, while galaxies on the edge of the group peak further away. The histogram on the top right shows that a good seed galaxy may exist far away from what appears to be the center of the group. This could indicate a separate community structure from the main group. We are also able to find outliers, galaxies that do not belong in any group.

Distance matrices Many social networking methods use the concept of distance matrices to assess community structure. These are matrices of "similarity", with each element being the distance between two sets of nodes (galaxies). Calculating the similarity between galaxies is simply how far away they are from each other. The separation in the projection of the sky is easy to calculate, but the third dimension is trickier. Velocities and redshifts are used as a measure of how far away galaxies are from us. To combine velocity with separation in right ascension (RA) and declination (Dec) requires a special calculation due to the difference in units. The 3D measure of distance that is used is zeta [7], defined in equation 1.

$$\xi = \sqrt{(\frac{\Delta R}{\Delta Rmax})^2 + (\frac{\Delta vr}{\Delta vmax})^2}$$

Equation 1: The calculation of zeta, a 3D measure of distance between two galaxies.

Each element of the distance matrix (Figure 6) is the value of zeta calculated between the galaxies associated with that element's row and column. These values are then quantized on a scale related to the average size of a group to correspond to numbers between 0 and 1, with 1 meaning the two galaxies are extremely close and likely interacting and 0 meaning the galaxies are far apart and not interacting. The element corresponding to a galaxies distance from itself is assigned to be 0 so as to have a galaxy not appear to be interacting with itself.

	Gal 1	Gal 2	Gal 3	Gal 4	Gal 5	Gal 6	Gal 7	Gal 8
Gal 1	0	0.5	0.5	0.8	0.5	0.5	0.5	0.8
Gal 2	0.5	0	1	0.8	0.8	0.5	0.8	0.5
Gal 3	0.5	1	0	0.8	0.5	0.5	0.8	0.5
Gal 4	0.8	0.8	0.8	0	0.5	0.5	0.5	0.8
Gal 5	0.5	0.8	0.5	0.5	0	0.5	0.5	0.5
Gal 6	0.5	0.5	0.5	0.5	0.5	0	0.5	1
Gal 7	0.5	0.8	0.8	0.5	0.5	0.5	0	0.5
Gal 8	0.8	0.5	0.5	0.8	0.5	1	0.5	0

Figure 6: Example of a distance matrix, a matrix of values corresponding to how close or far away two galaxies are from each other. A 1 implies that galaxies are very close to each other and likely interacting while a 0 implies the galaxies are not interacting, with varying values between according to the separation.

Viewing distance matrices This matrix can be converted into a grayscale image so that it can be easily viewed. The columns are organized by increasing RA and the rows are organized by increasing Dec so that the matrix is roughly related to the positions on the sky. Using these distance matrix images, we can look for bright areas which indicate groups and look for connections between groups. In Figure 7, I made a fake data set with two groups and a "bridge" of galaxies connecting them. These two groups show up very clearly in the distance matrix image, with elements between galaxies in the same group being bright, and elements between galaxies of different groups being dark. The groups are easy to pick out.

Distance matrices for large sections of the sky Real data sets, such as for a subsection of the Abell 1367 Coma Supercluster (Figure 8), are much more complicated. However, it is still possible to see structure within the distance matrix. Bright areas indicate groups and gray areas suggest structures exist linking the groups.

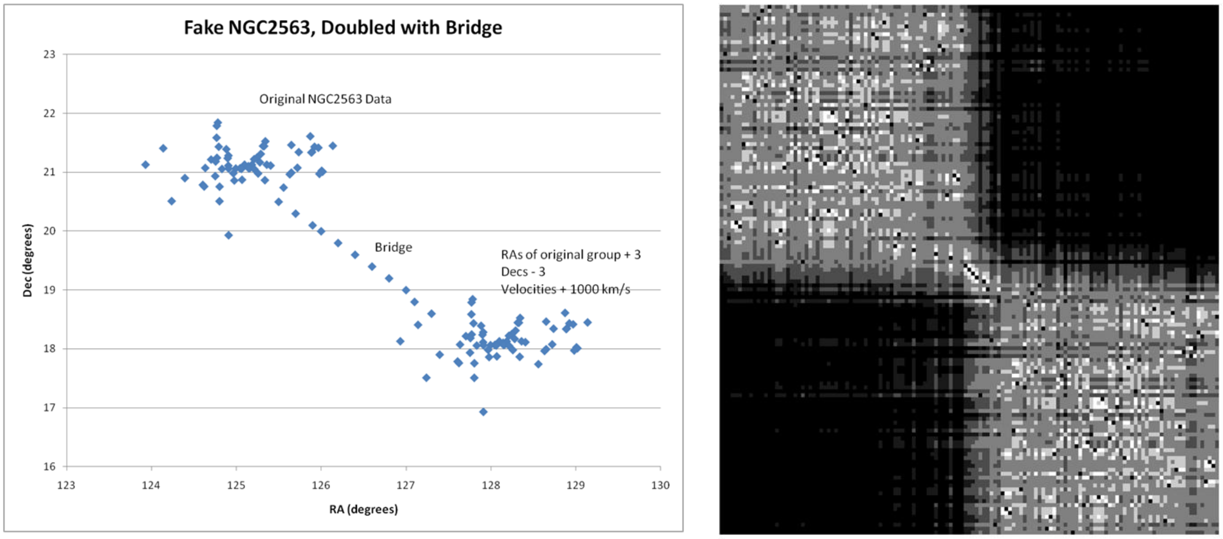

Figure 7: The distance matrix converted into a grayscale image (right) that corresponds to a fake data set (left). Here, the group NGC2563 has been duplicated and shifted, with a fake "bridge" of galaxies connecting the two. The distance matrix is sorted by RA and Dec and the two separate groups are very apparent as bright regions, with a small bridge of points between them. The dark areas consist of elements between two galaxies from different groups, which are not interacting.

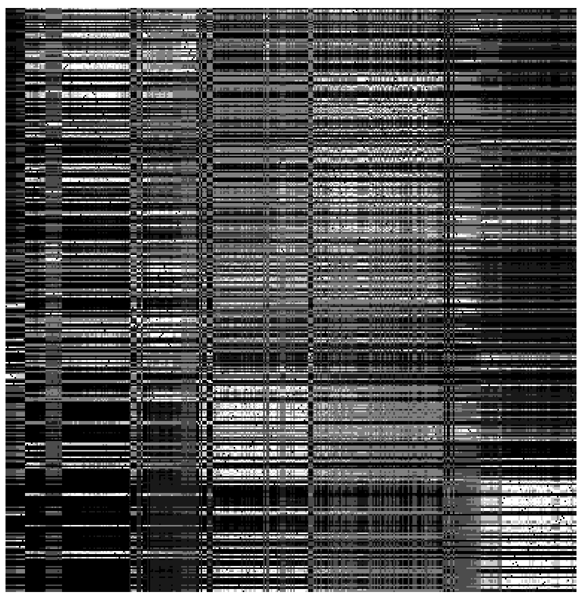

Figure 8: Distance matrix image for a subsection of the Abell 1367 Coma Supercluster. This large area of the sky contains 383 galaxies and connections are much more complicated than for the fake data set in Figure 7. However, bright areas containing groups can still be seen and grey areas suggest linkages between groups. Dark areas indicate no galaxy interactions there.

Agglomerative Hierarchical Clustering One networking technique that uses distance matrices is hierarchical clustering. This technique iteratively builds a hierarchy of clusters by either starting with all nodes (galaxies, in our case) connected in one big group and splitting them into separate groups (divisive clustering) or starting with all nodes as separate groups and connecting them until they are all in the same group (agglomerative). The agglomerative method runs as follows:

1. Assign each galaxy to a separate cluster
2. Create a distance matrix of the distances between all pairs of galaxies
3. Merge the two galaxies that are the closest into one cluster
4. Remove the galaxies from the distance matrix and add in our new cluster
5. Iterate until all have been merged into one cluster

Dendrograms Using the agglomerative hierarchical clustering above, I then create a dendrogram. Dendrograms show the order in which galaxies have been merged together in this technique (Figure 9).

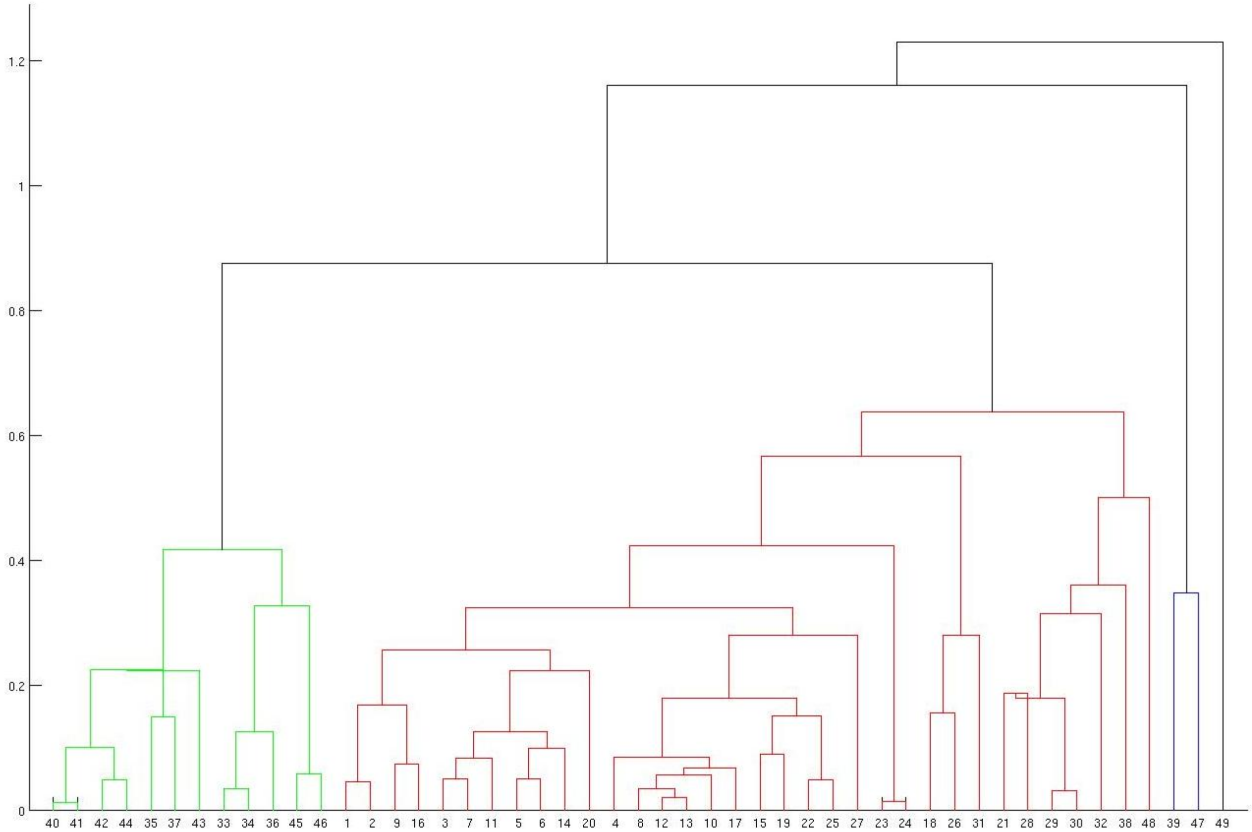

Figure 9: Dendrogram showing connections between galaxies in NGC2563. This shows that NGC2563 should perhaps be split into two groups, with three outliers not in either group. Outliers are identified by galaxies that are connected much higher than the rest. The dendrogram also shows substructure within identified groups.

The value on the left side of the dendrogram, referred to as the modularity, is a way to quantify the value at which structures are merged. This value is simply the value in the distance matrix between the two groupings or galaxies.

Conclusions

Many techniques exist for determining galaxy groups. However, the techniques are not consistent and yield different results. These techniques are poor at addressing the following questions:

1. Which galaxies belong in which group?
2. How tightly linked are galaxies within and between groups?

3. What is the most important spatial scale on which the environment influences galaxy evolution?
I have combined various concepts from the science of networks to address these questions.

Using the traditional group-finding friends-of-friends technique, I can identify good “seed” galaxies, galaxies that are central to the group. I can also identify galaxies that are on the outskirts. Identifying these is important because they will have different properties due to their differences in environment.

I can also identify which galaxies belong in which group by looking at the dendrograms in the agglomerative hierarchical clustering technique. Using the modularity that the technique uses to create the dendrograms, we can quantify levels at which galaxies become connected with each other, as well as when groups become connected to each other. We can further identify outliers that should not be included in any group. These isolated galaxies will have completely different properties than the galaxies that exist in groups.

Next Step The next step is to combine all of the techniques used to address the third question above. By analyzing the results of the various techniques as applied to several different groups and clusters, I will look at the spatial scales that affect the properties of galaxies. This will lead me to identify the scales on which galaxy evolution occurs.

References

1. Berlind, A.A., et al. 2006 ApJ, **167**, 1.
2. Fortunato, S. 2009, arXiv:0906.0612v1.
3. Freeland, E.E., Stilp A., & Wilcots, E.M. 2009 AJ, **138,** 295.
4. Li, I. H., Yee H. K. C. 2008 AJ, **135,** 809.
5. Mucha, P.J., Richardson, T., Macon, K., Porter, M..A., & Onnela, J-P. 2009, arXiv:0911.1824.
6. Newman, M.E.J. & Girvan, M. 2004 Phys. Rev. E, **69,** 026113.
7. Rasmussen, J., et al. Astrophysical Journal submitted 2011
8. Springel, V., et al. 2005 Nature, **435,** 7042, 629.
9. Zabludoff, A. & Mulchaey, J. S.1998 ApJ, **496**, 39.
10. Zirbel, E.L. 1997 ApJ, **476,** 489.