Bobby’s Big Box of Galaxies: The Categorization and Visualization of the Sloan Digital Sky Survey

by

Robert Baldocchi
Class of 2019

A thesis submitted to the faculty of Wesleyan University in partial fulfillment of the requirements for the Degree of Bachelor of Arts with Departmental Honors in Astronomy

Middletown, Connecticut April, 2019
“Did you ever hear the Tragedy of Darth Plagueis the wise? I thought not. It’s not a story the Jedi would tell you. It’s a Sith legend. Darth Plagueis was a Dark Lord of the Sith, so powerful and so wise he could use the Force to influence the midichlorians to create life... He had such a knowledge of the dark side that he could even keep the ones he cared about from dying. The dark side of the Force is a pathway to many abilities some consider to be unnatural. He became so powerful... the only thing he was afraid of was losing his power, which eventually, of course, he did. Unfortunately, he taught his apprentice everything he knew, then his apprentice killed him in his sleep. It’s ironic he could save others from death, but not himself.” - Sheev Palpatine
Acknowledgements

To Roy, thank you for your patience, encouragement, and wisdom. I could not have written this thesis without you, and I am forever grateful for your mentorship.

To Ed, thank you for introducing me to astronomy. My current passions, goals, and life journey are all a result of your Dark Side of the Universe class four years ago.

To Seth, thank you for the constant encouragement, enthusiasm, and compassion. I am consistently impressed and inspired by your genuine excitement and passion towards teaching others.

To Jack, my thesis brother in arms, I cannot accurately express how grateful I am for the four years of love and support you have unfailingly provided. Though our four year journey is coming to a close, our story is far from over.

To Ricky, thank you for being such a positive influence in my life. You are the one of the most driven, smart, and loving people I know, and I aspire to be more like you.

To Caroline, thank you for being my best friend for as long as I can remember. You have provided over a decade of love and support, and I am forever grateful for your presence in my life.

To Jeff Goldblum, thank you for appearing shirtless in Jurassic Park and showing me what the ideal human form looks like.

To Alexis, my astrobuddy. Thank you for always being at my side and laughing at things that only the two of us find funny. I would not trade your friendship for
the world.

To whomever is in charge of thesis due dates, the fact that you did not check with HBO to see if your due date is right after the premiere date of Game of Thrones Season 8 is very disappointing.

To my Parents, I would not be here if not for your unwavering support and encouragement. From driving me to long beach for practice at 5 AM every Saturday morning to supporting me in my decision to study astronomy, I have always known you both have my back. I love you both, and I could not have asked for better parents.

To my siblings, Brian, Michael, and Catherine. Thank you for always showing interest in my work, even if you don't understand a thing I am saying. You have all been my best friends first and foremost, and I cannot wait to join you all in California. Love you all, and thank you for not being too salty about the fact that I am the best Smash Player in the family.

To Blair and Sammy, thank you for being you. I cannot even list the amount of times I wished you both were here. Blair, you will always be my handsome boy, and Sammy you have only gotten wiser and more majestic with age.

To Nana, I miss you each and every day. Thank you for being the most supportive person in my life, and thank you for being my biggest fan from day one.

And finally, thank you to all those who have supported me. I am so lucky to have such a long list of people to thank, and I wish I had enough space to thank each and every one of you.
# Contents

1 Introduction 1

2 Classification 7
   2.1 Introduction ................................................. 7
      2.1.1 Red, Blue, and Green Galaxies ....................... 9
   2.2 Classification by Activity ............................... 10

3 The Data 12
   3.1 The Sloan Digital Sky Survey ............................. 12
   3.2 Classification Parameters ................................. 12
   3.3 Defining the Sample ....................................... 13
   3.4 Separation by Color ....................................... 14
   3.5 Separation by Activity ................................... 15

4 Visualization of the SDSS in 3D 25
   4.1 Introduction ................................................. 25
   4.2 Scope .......................................................... 25
   4.3 Visualization Process ..................................... 26
   4.4 Why Blender? ............................................... 28

5 The Visualization 30
   5.1 Intro .......................................................... 30
   5.2 The Cosmic Web ............................................ 30
5.3 Distribution of Galaxy type ........................................ 32
5.4 Selection Bias ......................................................... 32
5.5 The Incomplete Nature of the SDSS ................................. 33

6 Conclusion and Future Work .......................... 35
   6.1 Conclusions .......................................................... 35
   6.2 Future Work .......................................................... 36
Chapter 1

Introduction

The Sloan Digital Sky Survey (SDSS) is the largest sample of spectroscopic galaxy redshifts. Since its inception in 2000, the SDSS has imaged a total of 31,637 square degrees. This area covers more than one third of the entire celestial sphere. With the SDSS nearing one million individual imaging fields, a staggering 1,231,051,050 individual objects have been catalogued in the most recent data release, data release 15 (DR15). Included in DR15, the SDSS has useful optical spectra for 4,151,126 objects. This translates to approximately two million galaxies, six hundred eighty thousand quasars, and nearly one million stars. (Aguado et al. 2019; see Figure 1.1).

The main goal and top priority of the SDSS is to conduct wide angle surveys of the sky with an emphasis on assisting science designed to address critical issues in extragalactic astronomy. In order to achieve this, they conducted a multiband photometric survey followed by a multifiber spectroscopic survey of galaxies and quasars at high Northern Galactic latitudes. The photometric surveys not only select galaxies and quasars for the spectroscopic survey, but they also take significant data on these objects. These data are supplemented by a deeper survey in the Southern Galactic hemisphere. (Sloan Digital Sky Survey n.d.)
The SDSS has a core mission to provide the highest possible level of scientific integrity. The intention of the survey is to create a database that will have a long and productive lifetime. To ensure this, the SDSS uses highly efficient instruments and software. The data included in DR15 were taken by three separate telescopes; The Sloan Foundation 2.5m Telescope at Apache Point Observatory, The Ireneu du Pont Telescope at Las Campanas Observatory, and The NMSU 1-Meter Telescope at Apache Point Observatory. The software used for the SDSS is open source and transparent, meaning all researchers using the SDSS data have access to the software available. This is done in order to confirm that the software is working properly as well as encourage researchers to advance the software. (Sloan Digital Sky Survey n.d.)

Sorting these data has been the topic of numerous studies, as accessing and organizing this immense database has countless applications. Many studies have been dedicated to categorizing the galaxies in the SDSS, yet very few have been done to accurately visualize the objects in a three dimensional setting. This is a consequence of a few major challenges: though expansive, the SDSS is still an incomplete map of the sky; redshift distortions are common in the SDSS and must be corrected for; finally, there are inherent difficulties in creating a visual representation of a dataset as large as the SDSS in three dimensions.

The application of visualizing the data from the SDSS in three dimensions was explored in a 2008 study by SubbaRao et al. (2008). The group used the open source three dimensional modeling program Blender to model the large scale structure of the universe by using the available extragalactic data from the SDSS. Blender proved incredibly useful for modeling and visualizing the large scale of the data available. However the incomplete nature of the available data release, DR6,
1. Introduction

Figure 1.1: This image shows both the full sky coverage of SDSS I/II and III imaging, as well as an example of the capabilities of the SDSS, over both large and small scales. The top images show the three different views of M33 as seen by the SDSS, showing both the large and small scale imaging capabilities of the Survey. The bottom two images show an up-to-date sky coverage map of the SDSS. This image has been nicknamed “The Orange Spider” since the observation strips for the scope of DR15 resemble a spider. (Aguado et al. 2019)
limited the application of the visualization (SubbaRao et al. 2008; See Figures 1.2, 1.3). They hypothesized that as large scale structure catalogues increase in size and completeness, scientific visualization will play a larger role in data analysis.
1. Introduction

Figure 1.3: These images consists of views rendered in Blender of the DR6 galaxy dataset from a variety of viewing angles. (SubbaRao et al. 2008)

Since SubbaRao et al. (2008) modeled DR6, more than one million galaxies have been added to the SDSS database, as well as an additional five hundred thousand quasars. In total, the current SDSS database increases the coverage over that of DR6 by 23,237 square degrees. Also, the technology for three dimensional astronomical visualization has been greatly improved. The improved scope and technology accessible create an opportunity to again model the SDSS in a three dimensional environment.

The visualizations presented in this thesis show a culmination of my efforts to categorize the available data in a useful way as well as present these data in a visual medium. I include all of the galaxy and quasar data available in DR15. Galaxies have been divided into eighteen different categories in order to provide a detailed visualization of the SDSS data. This visualization was developed using a modified version of Blender, Astroblend. Astroblend, developed by Naiman
(2016), improves upon the cosmological visualization capabilities of Blender. In the following chapters, I will discuss the methods used to categorize galaxies, the methods through which DR15 was searched, and how the data were processed. I will also describe the process of using Astroblend to visualize cosmological data.
Chapter 2
Classification

2.1 Introduction

As the advance of large scale spectroscopic surveys continues, work concentrating on classifying and analyzing classes of celestial objects has become crucial to understanding the massive amounts of data accumulated. On an extragalactic scale, many studies have focused on dividing and classing objects. One of the main challenges innate to observational cosmology is deciphering the evolutionary connection between various classes of galaxies. The most traditional classification of galaxies, the Hubble Sequence, was created based on galaxies without nuclear activity. The Hubble Sequence includes elliptical galaxies, lenticular galaxies, spiral galaxies, barred spiral galaxies, and irregular galaxies. (Hubble 1936; Sandage 1961; De Vaucouleurs 1974).

The main standard used to classify galaxies via The Hubble Sequence is morphology. The important characteristics include the existence, size ratio, and appearance of spiral arms, disc, bulge, and bar. The objects that fit the Hubble Sequence were referred to as “normal” galaxies. These classifications suffice when studying the morphological differences of galaxies; however, for the diversity of objects within DR15, it is important to classify galaxies in as specific a nature as possible. In DR15, a large number of galaxies with nuclear activity are found,
and these are still referred to as abnormal galaxies. These abnormal galaxies are divided into classes according to emission line features; for example, Seyfert 1 galaxies, Seyfert 2 galaxies, broad-line radio galaxies, narrow-line radio galaxies, and low ionization nuclear emission regions (LINERs). Star forming galaxies, LINERs and Seyferts can be categorized through the Baldwin Phillips Terlevich (BPT) diagnostic diagram (Baldwin, Phillips Terlevich 1981) and the segregation curves determined by Kewley et al. (2001), Kauffmann et al. (2003a) and Kewley et al. (2006).

A BPT diagram uses nebular emission lines to distinguish the ionization mechanisms of nebular gas. A majority of AGNs have nuclear optical spectra that are dominated by emission lines of low ionization species such as \([\text{O} I]\), \([\text{O} II]\), and \([\text{S} II]\) (Kewley et al. 2006). LINERs are often classified by their \([\text{O} I]\) and \([\text{O} II]\) emission line ratios. They are separated from Seyferts by comparing the line ratios of low-ionization species (\([\text{O} I]\) and \([\text{O} II]\)) to the line ratios of high ionization species (\([\text{O} III]\)). BPT diagrams are based on four optical line ratios: 
\[
\frac{[\text{O} III]}{[\text{H} \beta]}, \quad \frac{[\text{N} II]}{[\text{H} \alpha]}, \quad \frac{[\text{S} II]}{[\text{H} \alpha]}, \quad \text{and} \quad \frac{[\text{O} I]}{[\text{H} \alpha]}. 
\]
These diagrams showed that Seyferts and LINERs form two separate branches on the diagram. Also, AGN, AGN + HII, and Star forming galaxies can be easily segregated based on their position on a \(\frac{[\text{O} II]}{[\text{H} \beta]}\) versus \(\frac{[\text{N} II]}{[\text{H} \alpha]}\) diagram (See Figures 3.6, 3.7) (Kewley et al. 2006).

Due to the large volume of data and spectra available in DR15, the classifications of galaxies can be far more specific than the traditional Hubble Sequence. In order to understand the nature of galaxies and the composition of the observable universe in the most comprehensive way possible, it is crucial to examine galaxies with finely divided and specific classes. Attempts to provide a more accurate galaxy classification have included luminosity, abundance of spectral features
from either very old or very young stellar populations, peak wavelengths, and
many more classes. However, in order to deal with the massive amount of data
currently available, many astronomers have simply divided galaxies into the blue
and red sequence, based on their color magnitude relation (Lee et al. 2008).

Rather than classifying galaxies based on the Hubble Sequence, I will be using
a slightly modified version of the color magnitude relation of galaxies. I will first
divide galaxies into three main categorizes: red, blue, and green by their \((u - r)\)
color. Within these three main categories, six subcategories are added. These are
Passive, H\(\alpha\), Star forming, AGN + HII, LINER, and Seyfert. This is to provide the
best accuracy possible. The following sections will provide detailed explanations
of each of the classes.

\section*{2.1.1 Red, Blue, and Green Galaxies}

The first major divide and classification used is to categorize galaxies by their
color, specifically red, blue, or green. Red galaxies are those that occupy the red
sequence. The red sequence is populated by red galaxies, which are galaxies with
primarily old stellar populations or galaxies with significant intrinsic extinction.
These are usually elliptical galaxies. The galaxies in the red sequence usually
have little to no star formation (Dobos et al. 2012). Blue galaxies are galaxies
that occupy the blue cloud. The blue cloud includes mostly actively star forming
galaxies. These are usually spirals (Dobos et al. 2012). Green galaxies occupy
the green valley. The green valley is defined as the transitional state in which blue
galaxies evolve into red galaxies. During this transitional period, star formation
is slowly being quenched (Strateva et al. 2001). The galaxies have been organized
into these three color classes based on their redshift and \((u - r)\) color.
2. Classification

The \((u - r)\) color has proven to be a very useful diagnostic for categorizing galaxies. This was noticed by Strateva et al. (2001). The correlations they presented showed that the \((u - r)\) color distribution is bimodal for galaxies. The two peaks in this galaxy distribution show an optimal distinction between blue and red galaxies. This is confirmed qualitatively. It is expected that elliptical galaxies will get redder with increasing redshift. It is also expected that blue galaxies will get initially redder in \((u - r)\) color by a few tenths of a magnitude but will then stay fairly constant with increasing redshift. (Strateva et al. 2001)

2.2 Classification by Activity

The three classes of galaxies, red, blue, and green, are split into further subclasses based on nuclear activity. This is done through nuclear spectroscopy. Nuclear spectroscopy techniques involve measuring the emission created by excited atomic nuclei (Nuclear Spectroscopy - an overview — ScienceDirect Topics n.d.). These emissions can be used to create multiple subclasses for galaxies. The six subclasses are based on a categorization by star formation and nuclear activity done by (Kewley et al. 2001) and (Kauffmann et al. 2003a). In these studies, star formation rate is measured through \(H\alpha\) emission, since it is the strongest emission line and is the most reliable tracer of the star formation. These subclasses are Passive Galaxies, \(H\alpha\), Star forming galaxies, AGN + HII Galaxies, Liner Galaxies, and Seyfert Galaxies. Passive galaxies have no measurable \(H\alpha\) emission. \(H\alpha\) is a class of galaxies in which hydrogen emission is detected, but the other lines are not strong enough for classification. Star forming galaxies have measured star formation, but no active nucleus. AGN + HII galaxies have both an AGN and starburst signatures. LINER galaxies have emission from weakly ionized atoms,
while the emission from strongly ionized atoms is relatively weak. Seyfert galaxies are galaxies with bright, energetic, and compact cores with strong infrared emission (Dobos et al. 2012). The application of classifications to the SDSS will be discussed in the next chapter.
Chapter 3

The Data

3.1 The Sloan Digital Sky Survey

I use the SDSS DR15 for this study. Like all the previous data releases, DR15 is cumulative, containing spectra taken over the 20 years the SDSS has now been operational. The DR15 imaging data cover 31,637 square degrees in the $ugriz$ bands, covering over a third of the celestial sphere. All photometric and spectroscopic observations were taken using the 2.5-m SDSS telescope at the Apache Point Observatory in New Mexico, USA. The full scope of DR15 is shown in figure 1.1 (Aguado et al. 2019).

3.2 Classification Parameters

As described earlier, galaxies are grouped into three main classes with six subclasses each. The main classes are Red, Blue, and Green, while the subclasses are Passive Galaxies, Hα galaxies, Star forming galaxies, AGN + HII Galaxies, Liner Galaxies, and Seyfert Galaxies.
3.3 Defining the Sample

Before classification, I used a number of parameters in order to only include objects with confident measurements. In order to do that, I made the following cuts. The first cut I made was in regards to the Petrosian Radius of the galaxies. The Petrosian Radius relates to the Petrosian ratio \( R_p \), which is defined by Blanton et al. (2003) as:

\[
R_p(\theta) = \frac{\int_{\alpha_{\text{lo}}}^{\alpha_{\text{hi}}} \frac{d\theta'2\pi\theta' I(\theta')}{(\pi(\alpha_{\text{hi}}^2 - \alpha_{\text{lo}}^2))^{3/2}}}{\int_{0}^{\theta_0} \frac{d\theta'2\pi\theta' I(\theta')}{\pi\theta'^2}}
\]

Where \( I(\theta) \) is the azimuthally averaged surface brightness profile, while \( \alpha_{\text{lo}} \) and \( \alpha_{\text{hi}} \) define the annulus. The SDSS has defined \( \alpha_{\text{lo}} = 0.8 \) and \( \alpha_{\text{hi}} = 1.25 \). The Petrosian radius \( \theta_p \) is the radius in which \( R_p \) falls below a certain value, which is 0.2 for the SDSS. If the Petrosian Ratio is above 0.2, then measurements have a far higher chance of error. In order to keep the Petrosian Ratio below 0.2, I limited the Petrosian Radius to \( r < 18 \). Though this improves the reliability of measurements, it does create a selection bias towards brighter objects.

The next step I took was to split the data into ten separate redshift bins. This was done in order to properly visualize and categorize the data without the issue of oversaturation in plots. The ten redshift bins are \( 0 < z \leq 0.1 \), \( 0.1 < z \leq 0.2 \), \( 0.2 < z \leq 0.3 \), \( 0.3 < z \leq 0.4 \), \( 0.4 < z \leq 0.5 \), \( 0.5 < z \leq 0.6 \), \( 0.6 < z \leq 0.7 \), \( 0.7 < z \leq 0.8 \), \( 0.8 < z \leq 0.9 \), and \( z > 0.9 \). These redshift bins are not science driven. This split was done in order to provide a cursory cut in redshift as well as demonstrating the evolution of galactic activity as a function of redshift. It also makes the navigation and plotting of data far more accessible.
The final constraint I used was to not include any objects that were saturated, too bright, or not fully resolved. In the final filtered set there are 2,283,977 galaxies. This is 89.9 percent of the full set of 2,541,424 galaxies available in DR15.

3.4 Separation by Color

The first step needed to classify the galaxies in my selected volume was to segregate them by color. I slightly modified a technique developed by Strateva et al. (2001). In their study, they determined that the cutoff between blue and red galaxies was located at \((u - r) = 2.22\). Using this constraint, I first segregated galaxies into two categories, Blue and Red, based on their \((u - r)\) color. I used the \((u - r)\) since the r band is the standard band for SDSS photometry. These galaxies were then plotted on a Galaxy Color Magnitude diagram. I defined the Green Valley as all galaxies between the peaks of the blue and red galaxies. Strateva et al. (2001) calculated the lines fit to these peaks as \((u - r)_{\text{blue}} = 2.72 - 0.062g\) for the blue peak and \((u - r)_{\text{red}} = 2.17 + 0.035g\) for the red peak. After confirming that these lined correctly fit the peaks of the DR15 data the galaxies between these two peaks were classified as green valley galaxies. This confirmation was done by plotting the \((u - r)\) vs. \(g\) Color Magnitude Plot and testing whether this segregation is still a good fit through the peaks. Table 3.1 shows the distribution of each type of galaxy by redshift. Figures 3.3, 3.4, and 3.5 show the spectral features of each type of galaxy. In these spectra, one of the clearest indicators of star formation is the \(H\alpha\) line, located at 6500 Å. The blue galaxy has a very strong \(H\alpha\) emission line, while the example red galaxy has a weak \(H\alpha\) line. The strong \(H\alpha\) line in the blue galaxy shows that the galaxy still shows evidence of
3. The Data

<table>
<thead>
<tr>
<th>Redshift</th>
<th>Blue</th>
<th>Green</th>
<th>Red</th>
</tr>
</thead>
<tbody>
<tr>
<td>$0 &lt; z \leq 0.1$</td>
<td>62796</td>
<td>216845</td>
<td>78280</td>
</tr>
<tr>
<td>$0.1 &lt; z \leq 0.2$</td>
<td>16358</td>
<td>155199</td>
<td>200337</td>
</tr>
<tr>
<td>$0.2 &lt; z \leq 0.3$</td>
<td>4235</td>
<td>24074</td>
<td>144262</td>
</tr>
<tr>
<td>$0.3 &lt; z \leq 0.4$</td>
<td>1886</td>
<td>22770</td>
<td>206746</td>
</tr>
<tr>
<td>$0.4 &lt; z \leq 0.5$</td>
<td>2465</td>
<td>84164</td>
<td>262100</td>
</tr>
<tr>
<td>$0.5 &lt; z \leq 0.6$</td>
<td>9457</td>
<td>183401</td>
<td>283027</td>
</tr>
<tr>
<td>$0.6 &lt; z \leq 0.7$</td>
<td>16567</td>
<td>106400</td>
<td>107300</td>
</tr>
<tr>
<td>$0.7 &lt; z \leq 0.8$</td>
<td>13508</td>
<td>34048</td>
<td>26947</td>
</tr>
<tr>
<td>$0.8 &lt; z \leq 0.9$</td>
<td>8219</td>
<td>10364</td>
<td>7655</td>
</tr>
<tr>
<td>$z &gt; 0.9$</td>
<td>6760</td>
<td>5061</td>
<td>3246</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>142251</td>
<td>821826</td>
<td>1319900</td>
</tr>
</tbody>
</table>

Table 3.1: This table contains a count of each galaxy type by color as a function of redshift as discussed in Section 3.4.

star formation, while the weak $H\alpha$ line in the red galaxy shows that there is little to no star formation. The green galaxy shows an $H\alpha$ line that is not as strong as the blue galaxy’s, but stronger than the red. This shows that it is transitioning from a blue to a red galaxy.

3.5 Separation by Activity

Galaxies are often classified using their emission line features in spectroscopic studies. By using spectral line features, galaxies are classified into Passive Galaxies, $H\alpha$ galaxies, star forming galaxies, AGN + HII Galaxies, Liner Galaxies, and Seyfert Galaxies. However, the spectroscopic data are not complete for every object in DR15, so only galaxies between $0 < z \leq 0.50$ were classified based on line features. In Figure 3.6, the BPT diagrams of the galaxies in the sample are
Figure 3.1: This the Color Magnitude diagram of all the galaxies in my sample divided by redshift as discussed in Section 3.4.
3. The Data

Figure 3.2: This the Color-Color diagram of all the galaxies in my sample divided by redshift as discussed in Section 3.4.
Figure 3.3: An example of a Blue Galaxy and its spectrum as discussed in section 3.4.
3. The Data

Figure 3.4: An example of a Red Galaxy and its spectrum as discussed in section 3.4.
Figure 3.5: An example of a Green Galaxy and its spectrum as discussed in section 3.4.
plotted. The blue curve, which divides star forming galaxies (below the curve) and AGN + HII (between the curves), is defined by:

\[ \log_{10}\left[ \frac{[OIII]}{H\beta} \right] = 0.61\left[ \log_{10}\left[ \frac{[NII]}{H\alpha} \right] - 0.05 \right]^{-1} + 1.3 \]

(Dobos et al. 2012)

The Red curve shows the distinction between pure AGN (above the curve) and AGN + HII (between the curves) is defined by:

\[ \log_{10}\left[ \frac{[OIII]}{H\beta} \right] < 0.61\left[ \log_{10}\left[ \frac{[NII]}{H\alpha} \right] - 0.47 \right]^{-1} + 1.19 \]

(Dobos et al. 2012)

Figure 3.7 shows the cut between LINER galaxies (below the line) and Seyfert galaxies (above the line). The cut used to segregate these categories is defined by:

\[ \log_{10}\left[ \frac{[OIII]}{H\beta} \right] = 1.18\log_{10}\left[ \frac{[OI]}{H\alpha} \right] + 1.4 \]

(Dobos et al. 2012)

Galaxies without strong enough emission lines to make it onto the BPT diagram are classified as passive galaxies (no detectable emission lines) or as H\alpha galaxies (measurable H\alpha).
### Table 3.2

This table contains a count of each galaxy type classified by color and nuclear activity as a function of redshift as discussed in Section 3.5. V1, V2, V3, V4, and V5 refer to the redshift cuts of $0 < z \leq 0.1$, $0.1 < z \leq 0.2$, $0.2 < z \leq 0.3$, $0.3 < z \leq 0.4$, and $z > 0.4$ respectively.

<table>
<thead>
<tr>
<th>Galaxy Type</th>
<th>Blue V1</th>
<th>Red V1</th>
<th>Green V1</th>
<th>Blue V2</th>
<th>Red V2</th>
<th>Green V2</th>
<th>Blue V3</th>
<th>Red V3</th>
<th>Green V3</th>
</tr>
</thead>
<tbody>
<tr>
<td>AGN</td>
<td>673</td>
<td>29985</td>
<td>61792</td>
<td>941</td>
<td>57355</td>
<td>51139</td>
<td>608</td>
<td>10015</td>
<td>4504</td>
</tr>
<tr>
<td>AGN + HII</td>
<td>1192</td>
<td>13474</td>
<td>9979</td>
<td>5</td>
<td>30417</td>
<td>5540</td>
<td>14</td>
<td>6909</td>
<td>419</td>
</tr>
<tr>
<td>Star Forming</td>
<td>13689</td>
<td>25570</td>
<td>160399</td>
<td>14719</td>
<td>55820</td>
<td>121031</td>
<td>2849</td>
<td>14961</td>
<td>11994</td>
</tr>
<tr>
<td>Seyfert</td>
<td>10553</td>
<td>29293</td>
<td>102603</td>
<td>5047</td>
<td>60380</td>
<td>111809</td>
<td>843</td>
<td>15736</td>
<td>9922</td>
</tr>
<tr>
<td>Liner</td>
<td>48919</td>
<td>14135</td>
<td>75383</td>
<td>9802</td>
<td>28047</td>
<td>27049</td>
<td>2108</td>
<td>4786</td>
<td>3480</td>
</tr>
<tr>
<td>Passive</td>
<td>301</td>
<td>1499</td>
<td>3034</td>
<td>107</td>
<td>10981</td>
<td>2204</td>
<td>29</td>
<td>9019</td>
<td>259</td>
</tr>
<tr>
<td>H(\alpha)</td>
<td>136</td>
<td>18738</td>
<td>28028</td>
<td>17</td>
<td>75413</td>
<td>9934</td>
<td>15</td>
<td>26569</td>
<td>715</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Galaxy Type</th>
<th>Blue V4</th>
<th>Red V4</th>
<th>Green V4</th>
<th>Blue V5</th>
<th>Red V5</th>
<th>Green V5</th>
<th>all</th>
</tr>
</thead>
<tbody>
<tr>
<td>AGN</td>
<td>237</td>
<td>6260</td>
<td>1002</td>
<td>1</td>
<td>130</td>
<td>15</td>
<td>224657</td>
</tr>
<tr>
<td>AGN + HII</td>
<td>11</td>
<td>4496</td>
<td>178</td>
<td>0</td>
<td>71</td>
<td>6</td>
<td>72711</td>
</tr>
<tr>
<td>Star Forming</td>
<td>680</td>
<td>13270</td>
<td>3156</td>
<td>23</td>
<td>922</td>
<td>135</td>
<td>439218</td>
</tr>
<tr>
<td>Seyfert</td>
<td>281</td>
<td>14683</td>
<td>2274</td>
<td>7</td>
<td>5710</td>
<td>101</td>
<td>369242</td>
</tr>
<tr>
<td>Liner</td>
<td>445</td>
<td>3748</td>
<td>1292</td>
<td>16</td>
<td>246</td>
<td>46</td>
<td>219502</td>
</tr>
<tr>
<td>Passive</td>
<td>9</td>
<td>17467</td>
<td>479</td>
<td>421</td>
<td>27835</td>
<td>4384</td>
<td>78028</td>
</tr>
<tr>
<td>H(\alpha)</td>
<td>18</td>
<td>26661</td>
<td>791</td>
<td>2</td>
<td>461</td>
<td>33</td>
<td>187531</td>
</tr>
</tbody>
</table>
Figure 3.6: This is a BPT diagram of all the galaxies in my sample, as discussed in Section 3.5. The red curve shows the distinction between pure AGN and AGN + HII, with AGN being above the curve. The blue curve shows the segregation line between star-forming galaxies and AGN + HII galaxies, with star forming galaxies below the curve. AGN + HII galaxies lie between the two curves.
Figure 3.7: This is a Liner/Seyfert diagram of all the galaxies in my sample, as discussed in Section 3.5. Galaxies above the blue line are Seyfert Galaxies, while galaxies below are LINERs.
Chapter 4

Visualization of the SDSS in 3D

4.1 Introduction

In this section I will discuss the process and challenges of visualizing the large scale cosmological dataset that is the SDSS. Some of these challenges include the incomplete nature of the survey, the difficult nature of visualizing datasets that span massive cosmological areas, and the inherent challenges in presenting dense datasets in 3D.

4.2 Scope

The visualizations I am presenting were inspired by the work done by SubbaRao et al. (2008). Their study mapped the extragalactic data available in DR6. DR6 only contained 790,000 galaxy redshifts. Seven data releases later, the SDSS contains upwards of a million more galaxy redshifts. DR6 mainly surveyed the Northern Galactic Cap, with three small stripes in the south. DR15 expands upon the Northern Galactic Cap, as well as vastly improving the coverage of the Southern Galactic Cap. This visualization covers 31,637 square degrees and contains 2,283,997 galaxies. Galaxies are divided into blue galaxies, red galaxies, and green galaxies, and are color coded as such.
4. Visualization of the SDSS in 3D

4.3 Visualization Process

Since one of the main goals of this visualization is to illustrate the advantages of visualizing large astronomical datasets in three dimensions, I chose to only visualize the galaxies separated by star formation rate, since it is a much larger dataset than the galaxy spectra available in DR15. Including these data will be discussed in section 6.2.

The Visualization Process was done using a modified version of Blender called Astroblend developed by Naiman (2016). Blender is an open-source 3D computer graphics program designed to create animated films, visual effects, art, 3D printed models, interactive 3D applications, and video games. Astroblend uses Blender version 2.7 as a base, and is modified to function specifically for astronomical data visualization. I made the decision to use Astroblend because it is better suited for visualizing large astronomical datasets.

One of the first challenges was converting Right Ascension and Declination into Cartesian coordinates, as Blender is not suited to create a visualization using polar coordinates. When converting to Cartesian coordinates, redshift needed to be converted to a usable unit, such as Mpc. For this conversion, I used a Hubble constant of \(69.32 \text{ km} \text{ Mpc}^{-1}\). This is the value derived from the Planck Results stored in astropy.cosmology. A script was written in Python to do this conversion.

Once the redshift had been converted to Mpc, and the coordinates converted to Cartesian, the coordinates were uploaded to a csv file. The coordinates were then imported into Astroblend. In the study done by SubbaRao et al. (2008), luminous halos could not be assigned to each individual object, due to the constraints of Blender. Many materials are available in Blender, and they decide how the object
is rendered. For example, choosing the wire material will render an empty object with only a thin wire frame around the vertices. Choosing the Halo material for an object renders each of the objects as glowing dots or small clouds of light, however, they do not cast light. An object that casts light will project their light on to other objects. For example, a rendered lamp that casts light will create shadows. A rendered lamp that does not cast light will be luminous and visible, yet its light will not interact with other objects. Halos are visible, yet have no substance. These qualities are ideal for this massive visualization since the dearth of light casting and substance save valuable memory and shorten rendering time. They settled for only 255 “halo objects” containing a position of a set of galaxies sharing common properties. With Astroblend, I was able to assign a luminous halo to each galaxy. Blue galaxies were assigned blue halos, red galaxies were assigned red halos, and green galaxies were assigned green halos. In order to make the halos opaque and more visible in the final render, a texture was applied to the halo corresponding to their color.

The next challenge faced was the immense scale of data available. With a scale of 1 Blender unit = 1 Mpc, the entire dataset was too large and dispersed to fully render. Therefore, I changed the scale to 1 Blender unit = 100 Mpc. This made the volume of the visualization able to be fully rendered.

I then coded the camera to follow a bezier curve around the visualization. A bezier curve is a parametric curve often used in computer graphics. The curve uses the Bernstein polynomials as a basis. A bezier curve is represented by

\[ r(t) = \sum_{i=1}^{n} b_i B_{i,n}(t) \]

The variables \( b_i \) and \( B_{i,n} \) are the control points that determine the shape of the
curve (Kent and Institute of Physics Great Britain, 2015). Bezier curves are useful in animation in rendering since they create an incredibly smooth trajectory, as well as allowing the animator to specify the velocity a camera travels while still allowing it to focus on a chosen object or location. I rendered it at 30 frames per second over a total of 1800 frames. This was done in order to create a smooth and high resolution rendering without having an excessively long rendering time. This sequence took approximately six hours to render.

4.4 Why Blender?

At first glance, this may seem like an unnecessary amount of work to create a three dimensional map that can easily be made in Python. Using Python on its own is far easier and quicker while the Blender interface is intimidating and difficult to use without practice. However, Blender is far more useful for this visualization as well as any future massive astronomical visualizations. Blender gives scientists the tools to visualize data in a way that is unmatched by Python, especially for large datasets.

Blender is a Python based program, which means it includes all the most useful applications of Python and expands upon them. The models available allow you to modify any object. For example it allows you to add UV maps, surface textures, volume textures, and changes in object density. The lighting engine in Blender allows for ray tracing, diffusion, and light falloff. You are also able to create animated versions of visualizations, allowing for an interactive view of the visualizations created.

Analyzing a large dataset requires efficient data organization and visualization, and Blender allows you to do this in a useful and interactive way. The program
is able to show your visualizations in real time as you create them, and allows you to explore them freely in this 3D setting. This was incredibly useful for this project, as there was no need to constantly run and rerun the code in order to see the visualization. Instead, any changes made showed in real time.

This visualization was done in Blender in order to create the best visualization possible. I made extensive use of the lighting options in Blender that are not available in Python. Also, the real time monitoring of my visualization was crucial for the efficiency of this project. Finally, Blender’s animation feature allowed me to easily view the visualization from any view I chose, as well as create an animation that showed the visualization from every angle on a continuous loop.
Chapter 5

The Visualization

5.1 Intro

This three dimensional visualization offers but a glimpse of the massive scale of the universe. However, as is obvious at a first glance, the sample is largely incomplete. The next few sections will discuss both the positive results from the visualization as well as its shortcomings. Figure 5.1 shows the visualization from six different angles.

5.2 The Cosmic Web

The large scale structure of the universe is thought to have evolved through gravitational instabilities caused by small density fluctuations in the early and mostly homogeneous universe. The resulting structure that the universe evolved into consists of galaxy clusters connected in filaments and sheets, with some areas having a low density of galaxies, while others regions are incredible dense. These filaments weave together to form a web like structure. This structure has been named the Cosmic Web (see figure 5.2), as described in detail by Bond, Kofman, and Pogosyan (1996).

This visualization shows clear and further support for the Cosmic Web, as it is visible from nearly every angle of viewing. As is expected, these filamentary
Figure 5.1: The Volume of the SDSS DR15 in three dimension, as viewed from 6 different angles. The full rendering can be found at https://www.youtube.com/watch?v=ZyDHT29tD9Mfeature=youtu.be
structures become less defined as redshift increases. This is due to the decreasing angular size as redshift increases, as well as the incompleteness in the sample.

5.3 Distribution of Galaxy type

For the first time, the distribution of galaxy type by star formation is able to be explored in a three dimensional environment. The sample shows that a majority of the galaxies observed in the SDSS are red. However, this does not confirm that the majority of galaxies in the universe are red, although that is commonly believed. There also seems to be a lower density of blue galaxies between redshifts of 0.3 and 0.6. However, the density of blue galaxies increases at a redshift greater than 0.6, but drops off again after a redshift of 0.9. Green galaxies are very common, suggesting that many of the galaxies in the universe are transitioning from blue galaxies to red galaxies, as star formation turns off.

5.4 Selection Bias

There are two main observational biases that affect this visualization. The first of these biases is apparent in the radial color gradient that is visible in the visualization. This radial color gradient is caused by the fact that the main galaxy sample is a magnitude limited sample. Bluer, late-type galaxies occupy a vast majority of the fainter end of the spectrum. This means that faint red and green galaxies are unrepresented in this sample. As redshift increases, only the largest red galaxies can be detected, as the smaller are far fainter SubbaRao et al. (2008).

The second of these biases is the Fingers of God that are clearly visible. Fingers
of God are elongated distortions along the line of sight. These are caused by
interactions between galaxies and their environment that are nonlinear. These are
often galaxies that orbit each other in massive clusters. As they orbit each other,
their orbital velocities cause errors in their redshift. This causes some galaxies to
appear to have higher redshifts than others, even if they are at a similar distance
away. This makes their location impossible to pinpoint with data from the SDSS
alone.

5.5 The Incomplete Nature of the SDSS

The other weakness of this visualization is a result of the incomplete nature
of the SDSS. The SDSS is a magnitude and volume limited sample. Since it is
a magnitude limited sample, dim galaxies are underrepresented, making it im-
possible to find the exact number of dim galaxies in the scope of the SDSS by
using only SDSS data. The sample is also volume limited. Though a third of the
celestial sphere has been covered by the survey, there is still a large section of the
sky that needs to be surveyed. Also, there are gaps in which no observations were
taken. These gaps are much more apparent in the Southern hemisphere, as the
SDSS is primarily a northern hemisphere based survey.
Figure 5.2: This is a model of the Cosmic Web as developed by Bond, Kofman, and Pogosyan (1996).
Chapter 6

Conclusion and Future Work

This thesis has categorized the available extragalactic data in DR15 based on star formation and activity. This results in eighteen individual galaxy classes that yield valuable insight towards the quantity of each galaxy type. The visualization process displays the distribution of galaxies by star formation rate and provides an example for how three dimensional visualizations are useful for massive astronomical datasets. Chapter 6 will discuss the results and examine the potential for future work.

6.1 Conclusions

The categorization by star formation rate shown in this thesis are based on the work done by Kewley et al. (2001), Strateva et al. (2001), and Kauffmann et al. (2003a). These studies demonstrate that galaxies can be accurately categorized by star formation rate using their $(u-r)$ color. In this thesis, I was able to categorize a much larger sample of galaxies than the previous studies. Galaxies were also classified based on activity. The categorization based on activity is based on the work done by Dobos et al. (2012). I was able to classify many more galaxies than the Dobos study.

I created color magnitude and color-color diagrams with 2,283,977 galaxies that show the distribution and density of galaxy type at different redshifts. Table
3.1 shows the number of each type of galaxy at different redshifts. In a smaller sample of spectra containing 1,403,358 galaxies, I created BPT and Seyfert/Liner diagrams. These diagrams show the distribution and density of galaxy type at different redshifts. Table 3.2 shows the amount of each type of galaxy based on star formation rate and activity at different redshifts.

Using the categorization by star formation rate, I then visualized the distribution of galaxies in Astroblend, an open source 3D computer graphic program. This visualization is an excellent display of the Cosmic Web as well as redshift dependant star formation.

6.2 Future Work

Future work can be done with the SDSS data. Creating visualizations based on the subcategories discussed in section 3.5 will display the distribution and density of galaxies based on their nuclear activity. This will yield valuable insights towards the nature of the universe on a large scale. Multiwavelength data can also be incorporated in the future, as Astroblend has the capability to include light filters. Visualizing galaxies with actual galaxy models rather than point sources will be useful to create a more accurate image of the universe.

The interactivity of the visualizations in this thesis present a unique opportunity to test how virtual reality can be used to analyze massive sets of astronomical data. The use of Blender in visualizations allows for the integration of virtual reality. It will be beneficial to examine the advantages of analyzing data in a visual and interactive medium.

As the scale and availability of large spectroscopic surveys improves, visualization may play a much larger role in astronomical data analysis. Datasets such as
Gaia, Simbad, and Vizer can be similarly visualized using the methods described in this thesis. As Gaia nears 1.5 billion objects, a similar visualization of the data may yield insights towards the structure of our own galaxy.

Revisiting this visualization with a larger set of data will result in interesting and exciting results. By revisiting the three dimensional visualization of the SDSS done by SubbaRao et al. (2008), I showed that the availability of a much larger data release makes a visualization far more useful and complete. I expect the visualizations in astronomy to improve as datasets become larger and the visualization software becomes more advanced.


Naiman, J. P. 2016. AstroBlend: An Astrophysical Visualization Package for


