THE CLUSTERING OF GALAXIES ON THE SMALLEST SCALES ACROSS COSMIC TIME

By

Jennifer Anne Piscionere

Dissertation

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

in

Physics
August 2015
Nashville, Tennessee

Approved:

Dr. Andreas Berlind

Dr. Kiril Bolotin

Dr. Kelly Holley-Bockelmann

Dr. Robert Scherrer

Dr. Keivan Stassun

DEDICATION

For my family;
Italian or Indian,
Two legs or four.

ACKNOWLEDGEMENTS

This thesis would not be possible without the support of my advisor, Prof. Andreas Berlind, as well as from Dr. Cameron McBride. I would also like to thank Dr. Manodeep Sinha for extensive help with software development without which this thesis would have taken even longer than seven years.

I sincerely thank Prof. Kelly Holley-Bockelmann for extensive mentorship and support.

I thank Dan Foreman-Mackey and Ryan Scranton for valuable discussions and software. I thank the anonymous referee for helping us improve the clarity of Chapter 2. I was supported by the National Science Foundation (NSF) through NSF grant AST-1109789. The simulations used in this paper were produced by the LasDamas project; I thank NSF XSEDE for providing the computational resources for LasDamas. Some of the computational facilities used in this project were provided by the Vanderbilt Advanced Computing Center for Research and Education (ACCRE).

Finally, I would like to thank the first response EMT team, doctors, and therapists that helped me recover after the trauma of my car accident.

TABLE OF CONTENTS

CHAPTER

Appendix

LIST OF TABLES

LIST OF FIGURES

Chapter I

INTRODUCTION

1.1 Observing and Modeling the Large Scale Structure of the Universe

We are privileged to live in a time of vast and fascinating progress in the study of the large scale structure of the Universe. A consistent model of the universe is emerging from extensive observational surveys, a model that includes a mysterious "dark" energy that is driving our accelerated expansion, and matter that cannot be directly detected through our telescopes but has profound implications for cosmology. Understanding this dark matter is key to understanding our cosmological history? without it, we cannot explain how the Universe came to look the way it does. Our theories are being fueled and tested by the new generation of galaxy surveys that provide a wide range of sensitivity and observations heretofore thought impossible.

Galaxies are the objects that make up our observed structure, and studying how they cluster can tell us much about the physical processes of the Universe. On the largest scales, the observed galaxy clustering is a product of early universe physics. Measurements at these scales constrain the cosmological parameters that govern the evolution of our Universe as a whole. However, if we want to understand how dark matter influences the evolution of a single galaxy, we must look at much smaller scales than one Megaparsec. By measuring the very small scale clustering of galaxies and how it relates to the distribution of dark matter in halos we can understand how galaxies evolve over cosmic time. Vital to how we measure clustering is the amount of galaxies in our sample- hence the need for such ambitious surveys. Combined with the ever increasing computational demands of determining dynamical forces for this large sample of galaxies, very small scale galaxy clustering is a sandbox to test the limits of galaxy formation models as well as advanced modeling techniques.

In Section 1.2, there will be an overview of the tools needed, both historical and physics background, to understand the clustering of galaxies. Section 1.3 will cover the details of creating and measuring the autocorrelation function $\xi(r)$, the most powerful tool for quantifying the clustering of galaxies (Peebles, 1980; Padmanabhan, 1993). There will also be a discussion about the technical details of large galaxy surveys, including the Sloan Digital Sky Survey (SDSS) (York et al., 2000). In Section 1.4, we will attempt to uncover the underlying physics behind why the structure of the Universe looks the way it does, specifically how galaxies occupy dark matter halos. This will include a discussion about N-body simulations, a key way to model galaxy clustering. Section 1.5 will include a summary of the subject.

1.2 Background

1.2.1 Historical Perspective

The idea of dark matter and its effect on the observed galaxies from these surveys is not a new one. Zwicky (1937) first drew the conclusion that there must be unseen matter in the Universe through detailed observations of the velocity dispersion of galaxies in the Coma Cluster. Using a simple application of the virial theorem, we can measure total

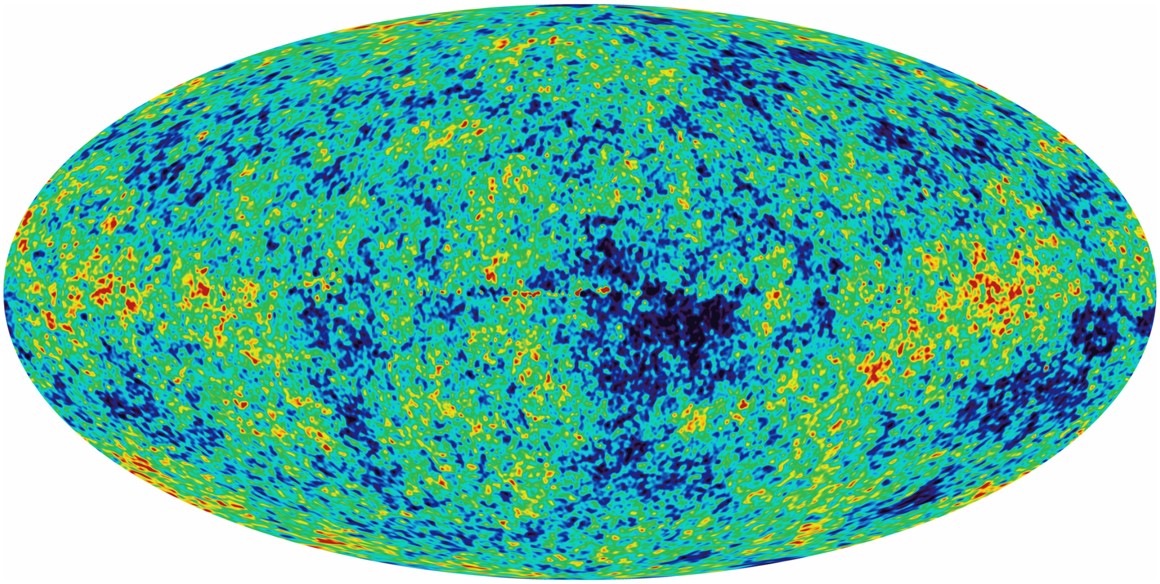

Figure 1.1 An image of the Cosmic Microwave Background (CMB) taken by the Wikilson Microwave Anisotropy Probe (WMAP), scaled to show to the very small temperature anisotropies in the CMB.

mass of a system by the velocity dispersion of its constituent parts (e.g. Ryden, 2003):

$$M = \frac{5\langle v^2\rangle r_h}{3G}, \tag{1.1}$$

where r_h is the half-mass radius of the cluster and $\langle v^2\rangle$ is the mean squared velocity of galaxies. The magnitude of their velocities could not be explained by the amount of observed baryons. Zwicky's detailed observations were groundbreaking and set the stage for future research.

The discovery of the cosmic microwave background (CMB) by Penzias and Wilson in 1960, and the prediction by Sachs & Wolfe (1967) that the anisotropies in the density field of the early universe would be visible as the variations in temperature of the CMB (Figure 1.1) has motivated much of the research attempting to form a cohesive theory of structure formation. This theory must bridge the gap between the tiny density fluctuations imprinted during an event that happened when the Universe was only 0.003% of its current size to the largest and most massive objects known today.

1.2.2 Composition of the Universe

Two of the most significant observations in cosmology are the location of first peak of the Cosmic Microwave Background temperature anisotropy power spectrum, and the distance measurements from far away Type Ia Supernovae (SNeIa). The first tells us that the Universe is flat, which means that the energy density of the universe is at the critical density; see Figure 1.1 for a map of the temperature anisotropies, and §1.4.2 for their origin. This critical density is measured today to be about one hydrogen atom per 200,000 cm^3 (Ryden, 2003). From the SNeIa, we observe that the universe isn't only expanding, something first noted by Hubble in 1929, but the expansion is accelerating. Before these precise measurements were made, we knew the universe was expanding by measuring the $redshifts$ (see equations 1.3) of distant objects. Hubble's Law states

$$cz = H_o d, \tag{1.2}$$

where c is the speed of light and the recession velocity of a galaxy, cz, is equated to its distance multiplied Hubble's constant H_o. H_o denotes the present day value of the constant, which is thought to be 70 km/s/Megaparsec[1]. This expansion is called the *Hubble Flow*.

---

[1] One parsec (pc) = 3x 10^{18}cm. One Megaparsec (Mpc) = $10^6 pc$.

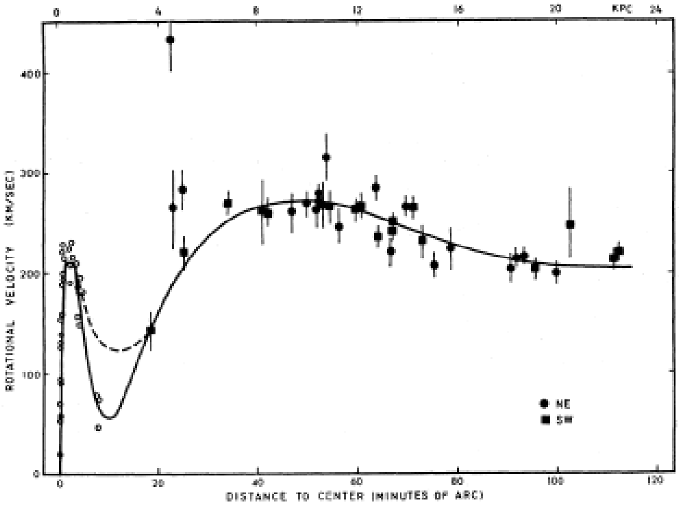

Figure 1.2 The Galaxy Rotation curve from Rubin & Ford (1970) measuring the rotational velocity of stars as a function of their galactocentric distance using the Andromeda Galaxy.

We can tell that there is an energy component which we call the *Cosmological Constant*, or Λ, driving this acceleration. Combined, these two observations point to the rather strange idea that we are 'missing' mass. We know that Λ is the dominant source of energy in the universe, making up $\sim 70\%$ of the energy and that the observed baryons account for only another $\sim\ 4\%$, with radiation only making up another minuscule portion. This means that there is a still more mysterious mass source in the universe. This missing matter is appropriately called *dark matter*, noted earlier by Zwicky, and it plays a critical role in how structure in the universe is formed.

1.2.3 Evidence for Dark Matter

Dark matter does not interact electromagnetically; since our telescopes (and eyes for that matter) can only see photons, we can't directly see dark matter- only its effect on the visible matter of the universe. One of the first, and still most compelling, arguments

Figure 1.3 An image of the cluster Abell 2218 taken with the Hubble Space Telescope. The effect of gravitational lensing is apparent in the light from background galaxies being distorted into arcs surrounding the cluster.

for dark matter came with the measurements of galactic rotation curves of Rubin & Ford (1970), Figure 1.2. If the potential well of our galaxy traced the luminous, baryonic matter the rotational velocity of stars about a disk galaxy should drop off as $v_{rot} \propto 1/r^{1/2}$. Instead, we observe a "flat" rotation curve- indicative of more, unobserved, mass that exists in a halo surrounding the central, luminous galaxy.

More empirical evidence for dark matter can be found by looking at the gravitational lensing of distant galaxies by the potential well of galaxy clusters. As light from these galaxies passes through a cluster, the bending of space time by the gravity of the cluster causes light to bends and distort. This is observed as arcs and in some cases even Einstein rings, Figure 1.3, and it magnifies the brightness of the lensed galaxy as well. The baryonic mass in the cluster cannot account for these observations.

1.2.4 The Physics of Dark Matter

Now that we see the need for dark matter in order to understand the observations, the question arises: what exactly is dark matter? The simple answer is that dark matter is cold, dark and matter. It is cold, meaning it does not move at relativistic speeds. Dark meaning that it does not produce or reflect electromagnetic radiation. We can measure the physical effects that dark matter has on the visible baryons in the universe, but have yet to directly detect dark matter particles though experiments [2]. From these experiments, we know that dark matter must be a very stable, massive particle. The current leading candidate for dark matter is a weakly interacting massive particle (WIMP), but this is still a very active area of inquiry.

We believe a galaxy forms in a diffuse halo of dark matter that is more massive and extends much farther than the luminous matter. For instance, the baryonic mass of the Milky Way is $\sim 6x10^{10} M_{\odot}$ [3] whereas the dark matter has a mass of $\sim 10^{12} M_{\odot}$. Similarly, the stellar disk has a radius of 15 kiloparsecs (kpc), while the dark matter halo extends to about 200 kpc. All these properties make dark matter crucial to structure formation in the universe. Without it, the universe today would be completely unrecognizable.

1.3 Observing the Structure of the Universe

Over the past few decades there have been tremendous leaps in our ability to observe the universe at larger and larger scales. Coming from the first surveys of only a few thousand galaxies, (Figure 1.4, to today's surveys of over one million galaxies, see §1.3.6,

---

[2] See CDMS II Collaboration et al. (2010) for an example of a current experiment.

[3] It is convenient for astronomers to define $M_{\odot}$ as the mass of the sun, $2x10^{33}$g, and use it as a standard unit of mass.

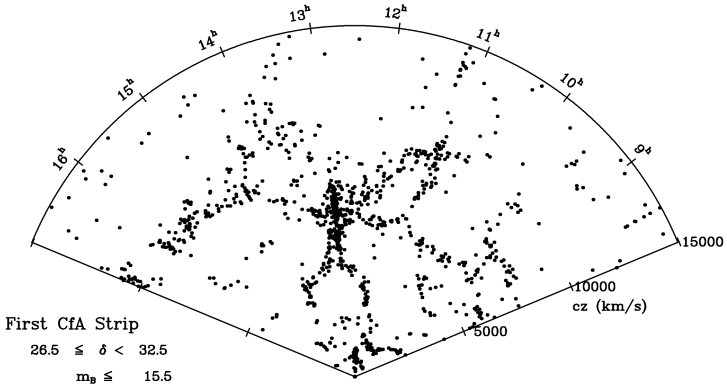

Figure 1.4 The initial galaxy map of the Center for Astrophysics (CfA) Survey (Geller & Huchra, 1989), showing a slice into an area of sky. The earth is at the point of the pie, and distant galaxies are the dots.

we've come along way in understanding our universe. We believe that galaxies trace the underlying dark matter distribution closely, and as such the study of galaxy properties is key to understanding the formation and evolution of the structure of matter as a whole. There are two main aspects to galaxy surveys: (i) imaging, or photometry, of galaxies and (ii) redshift measurements. Redshift surveys measure the distances to galaxies, creating a large sample of galaxies for which clustering statistics can be measured. As with all observations, there are complications to measuring the redshift of a galaxy; for example the proper motion of a galaxy imprints on the redshift.

1.3.1 Galaxy Surveys

Optical surveys of galaxies had been done since before the time of Messier, but the first galaxy survey to attempt a large scale map of the Universe was the Center for Astrophysics

(CfA) Redshift Survey (Geller & Huchra, 1989), started in 1977 (see section 1.3). Galaxy surveys once thought impossible are currently taking data, making unprecedented, precise measurements of our particular cosmology and the structure of the universe. In order to perform these measurements, surveys need to both of in cover large areas of the sky, and observationally deep enough to measur extremely faint and distant objects. The need for such a large sample of galaxies has driven the technology required to complete these surveys, measuring the distance to more than one million galaxies.

1.3.2 Types of Surveys

The first aspect of a galaxy survey, imaging, can give us important information regarding the morphology of galaxies. For instance, the Galaxy Evolution Explorer (GALEX, Martin et al. (2005)) is studying the universe in the ultra violet wavebands to better understand star formation and galaxy merger rates. Angular positions of galaxies can also be used to compute certain clustering statistics, see section 1.3.10.

Redshift surveys can give us three-dimensional positions, as well as detailed information about galaxy composition. Accurate positions are key for measuring the clustering of galaxies. Understanding the spatial distribution of galaxies in three dimensions requires knowledge of the distance to the galaxy. This distance is measured in redshift (z):

$$z = \frac{\lambda_o - \lambda_e}{\lambda_e}, \tag{1.3}$$

where λ_e is the wavelength of the emitted light in the object's rest frame, and λ_o is the wavelength of the observed light. Knowing the redshift of a galaxy requires measuring a spectrum. Details of how spectra are taken can be found in section 1.3.6, but a good estimate is that a spectrum takes about 10 times as long as the photometry for an object (Weinberg et al., 2007). Hubble's Law, equations 1.2, is the key to transforming this measured redshift into a physical distance. For small redshifts, we can take $z \backsimeq v/c = H_o d/c$ (Geller & Huchra, 1989). Conventionally, distances are measured in units of Mpc. More details about the specifics of this calculation can be found sections 1.3.6 & 1.3.7.

1.3.3 Types of Samples for Redshift Surveys

Another decision one must make when generating a galaxy catalogue from available data is whether to use a flux-limited or volume-limited sample. Each has its own benefits depending on the data analysis being done. Flux limited samples use all galaxies observed above the detection limit of the telescope. The advantage to this approach is obvious–there are no brightness cuts on which galaxies can be included in your sample, maximizing the sample size. However, this also introduces a large selection effect, called Malmquist bias, that causes a strong bias against faint objects. The number density of objects observed drops off with distance, until only the brightest objects are observed.

One way of overcoming this bias is to create a volume limited sample. Volume-limited catalogues, Figure 1.5, are complete galaxy catalogues in a defined luminosity bin, containing all galaxies that could be observed in that bin, and are therefore optimal for statistical analysis. They are created so that the faintest galaxy would be above the lower luminosity threshold at the high redshift end and the brightest galaxy is below the high luminosity cutoff at the low redshift end, (e.g. Zehavi et al., 2005b). Volume-limited samples also have, by design, a constant number density of galaxies. Most of the analysis

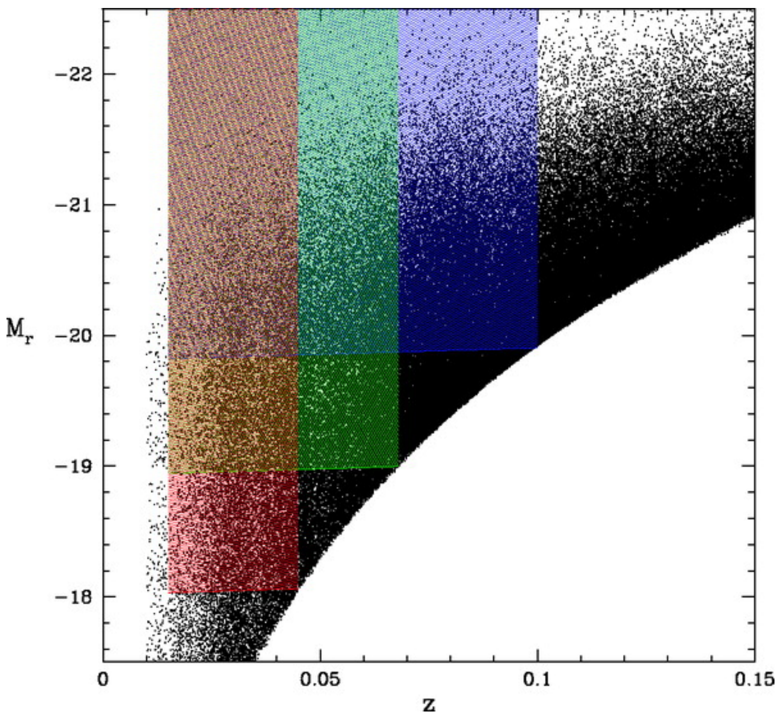

Figure 1.5 Absolute magnitude (M_r) as a function of redshift (z) for galaxies in the SDSS. The three colored boxes represent different volume limited samples, M_r brighter than -18, -19 and -19.9 (Berlind et al., 2006). These lower boundaries evolve with redshift to account for galactic evolution.

discussed herein is done on volume limited samples.

1.3.4 K-corrections

Galaxies emit a continuous spectrum of light with a characteristic shape, but our observations are done in particular photometric filters with a finite waveband. The location of the peak in spectral energy distribution of galaxies is determined by the age of the stellar population, with older stellar populations (and therefore galaxies) appearing more red. When studying galaxies at high redshift, we may be missing much of their light because the peak of their spectrum has been redshifted outside the wavebands of our most red filter; in the Sloan Digital Sky Survey (SDSS; §1.3.6) that filter is the *z*-band and it has an effective wavelength of 8840Å, (York et al., 2000). A K-correction is the difference between the dimming as a function of redshift for an object in a fixed observing waveband, and the bolometric (total) flux expected (Peacock, 1999). It is usually denoted $K(z)$, as it is a function of redshift. For a galaxy whose spectrum peaks in the red, observed at high redshift, this correction can be quite substantial.

1.3.5 Evolution Correction

During the course of their lifetimes, galaxies undergo many traumatic events that cause changes in color and morphology, and possibly even complete destruction. These cataclysmic events can affect the stellar population and luminosity of galaxies, discussed in §1.4.4; but galaxies also undergo what we call 'passive' evolution- over time their luminosity dims, and their light becomes more red solely due to the evolution of the stellar population. After 10 Gyr, all stars more massive than the Sun have left the main sequence and moved onto the giant branch, causing the galaxy to be appear more red. The very massive blue stars that dominate a young galaxy's luminosity are long since gone, leaving the galaxy less luminous. The same galaxy will be intrinsically more luminous at higher redshift, and this must be accounted for when making volume limited galaxy samples, which can be seen in Figure 1.5. The lower boundary of each sample evolves with redshift so that fainter galaxies will be included at lower redshift.

1.3.6 Specifics of the Sloan Digital Sky Survey

The Sloan Digital Sky Survey (SDSS;York et al. (2000)) comprises four separate surveys that use the wide-field 2.5m telescope (Gunn et al., 1998) at Apache Point Observa-

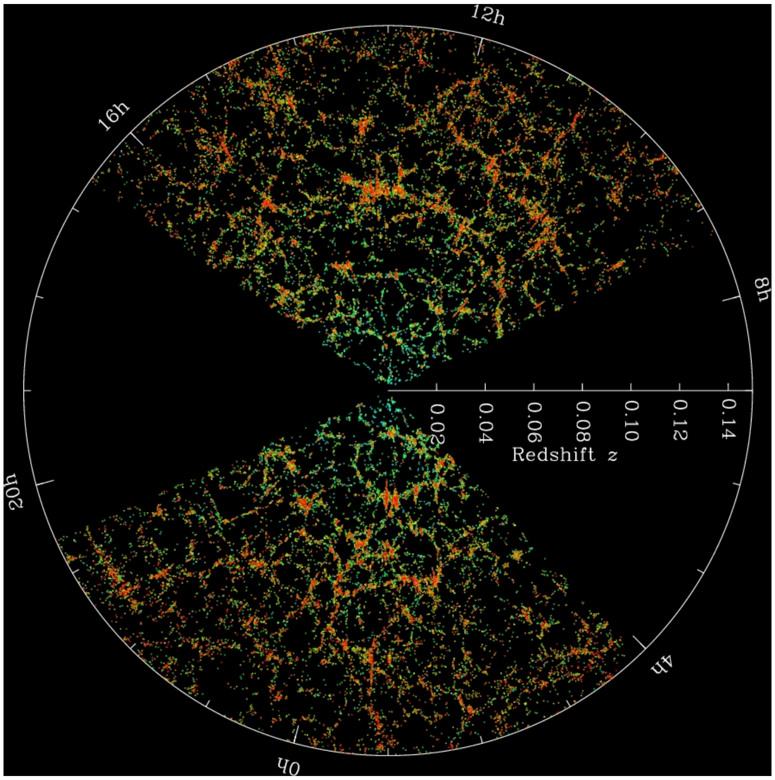

Figure 1.6 A slice of the SDSS galaxy survey, where each point is a galaxy and the earth is located at the center. We can see the filaments and voids that make up the cosmic web of our universe.

tory. One of these surveys, The Baryon Oscillation Spectroscopic Survey or BOSS, focuses on precision measurements of galaxy clustering over a wide range of redshifts, providing us with view of the universe (Weinberg et al., 2007). The original SDSS, as well as its two extensions (SDSS II & III), have made enormous contributions to the understanding of how galaxies cluster. The SDSS telescope has a large field of view and ability to take multiple spectra at one time. This instrumentation is ideal for getting the large sample size and detailed distances needed for precision clustering measurements (see Figure 1.6 for a sample of galaxies observed).

The 2.5m telescope at Apache Point operates in drift mode– the telescope is stationary and takes photometry in great circles as the sky rotates. Each object passes through one column of 5 different CCDs corresponding to the 5 SDSS filters, arranged in 6 columns, for a total of 30 CCDs. The SDSS filters are denoted, from ultraviolet to near-infrared, as u g r i z. The SDSS covered over 8000 deg^2 of sky since.

Object classes are determined by comparing the object's magnitude[4], or flux, within its Point Spread Function (PSF) and the modeled magnitude an extended light profile[5]. To be included in the SDSS main galaxy sample, a galaxy must have an r-band apparent magnitude $r \leq 17.77$. One of the major difficulties for the SDSS is deblending– when the flux of two objects overlap (Strauss et al., 2002). A galaxy's light can be blended with that of a saturated star, causing the galaxy to not have a measured spectrum and introduce incompleteness into the sample. Two galaxies may overlap, and it is the data

---

[4] Astronomers measure the brightness of galaxies on the logarithmic magnitude system. The higher the flux of an object, the more *negative* the apparent magnitude, denoted as m. An intrinsically more luminous object has a more negative absolute magnitude, M.

[5] An exponential or de Vaucouleurs profile.

Figure 1.7 An image of a plug plate and cartridge used used in SDSS. The optical fibers are plugged by hand into the plate (bottom) and lead to the spectrograph. The cartridge is slid behind the focal plane of the telescope to collect light from distant galaxies.

analysis pipeline's job to correctly identify them as such, and distribute flux accordingly to the 'parent' (brighter) and 'child' galaxies. Objects denoted as a child are not included in the main galaxy sample (Strauss et al., 2002).

Once galaxies are detected and classified as extended sources, spectra must be measured to get precise redshifts. For the SDSS, the selected galaxies will be mapped onto the sky and their location drilled onto an aluminum 'plug plate,' Figure 1.7. These plug plates are placed over the focal plane of the telescope; the plates are tiled over the sky and holes are drilled in a way that maximizes the number of galaxies able to have measured spectra (Blanton et al., 2003a). The 640 optical fibers that feed into the spectrographs are plugged into the plates by hand. These fibers cannot be located closer than 55" on the sky from another fiber. This is called a fiber collision, affecting about 6% of galaxies (Strauss et al., 2002) and can influence clustering measurements. It takes approximately 45 minutes to measure a spectrum with an acceptable signal to noise (Strauss et al., 2002).

The measured spectra determine the absolute distance to within a precision of 1.0% at z=0.35 and 1.1% at z=0.6. (Weinberg et al., 2007).

Thus, the main source of error in the galaxy catalogue is incompleteness due either to deblending or to fiber collisions causing a galaxy to be left out of the catalogue- 99.9% of galaxies that are observed spectroscopically yield successful redshifts (Strauss et al., 2002). The SDSS has collected the spectra for almost one million galaxies already, with the new SDSS-III extension aiming to collect 1.5 million spectra, making it the largest galaxy redshift survey.

1.3.7 Redshift Distortions

While redshift space distortions can be seen as a technical aspect of building a galaxy survey, they are the product not of details of the survey but of the physical properties of the cosmological mass distribution. This mass distribution affects a galaxy's peculiar motion along the line of sight, as opposed to its Hubble flow motion, and causes what we call redshift space distortions. What is actually measured when finding redshifts is the combination of the Hubble flow and these peculiar velocities, v_p, so that

$$cz = H_o d + v_p. \tag{1.4}$$

These peculiar motions cause the galaxy's position in redshift space to be displaced.

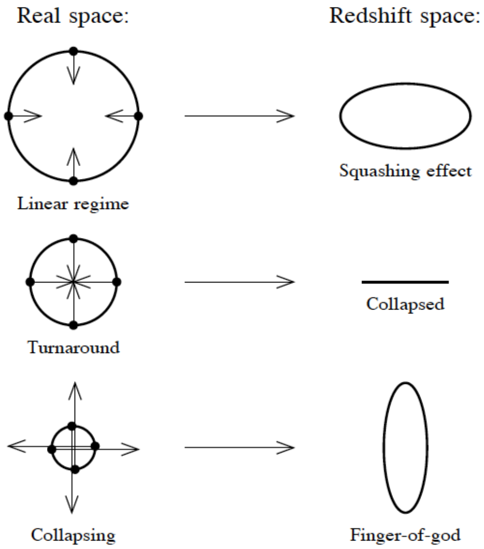

Figure 1.8 A schematic presenting the physical mechanisms and observed effects of redshift space distortions (Hamilton, 1998). The dots are galaxies and the arrows are their peculiar velocity vectors.

1.3.8 Fingers of God

The first of these distortions are the so called “Fingers of God” (Hamilton, 1998) caused by the velocity dispersion of galaxies within their own halo. The proper motions cause a broadening of the galaxy redshift distribution along the line of sight in a given cluster– making some galaxies appear farther away and others closer, depending on if their doppler motion causes their spectrum to be blue or redshifted. They were noticed with the first redshift surveys, and were named as such because this doppler shifting causes all these ‘fingers’ to point directly to us. An illustration of this can be seen in Figure 1.8.

1.3.9 Kaiser infall

The second redshift space distortion is due to large scale coherent galaxy velocities, caused by the bulk motion of groups of galaxies towards a very massive cluster, seen as an attractor (Kaiser, 1987). This is observed as a compression of the observed galaxy distribution. These distortions are much more difficult to detect (Hamilton, 1998), but can manifest themselves in statistical measurements of galaxy clustering (section 1.3.10).

1.3.10 The Galaxy Autocorrelation Function

One of the most useful statistics we have to quantify galaxy clustering is the correlation function. The Galaxy Autocorrelation Function measures the excess of pairs of galaxies as a function of separation when compared to a random distribution (Peebles, 1980; Connolly et al., 2001). The galaxy autocorrelation function is how galaxies are clustered together over a random clustering.

The real-space correlation function can be defined as the Fourier transform of the k-space matter power spectrum $P(\vec{k})$:

$$\xi(r) = \int d^3k P(\vec{k}) exp(i\vec{k}\cdot\vec{r}), \tag{1.5}$$

but it is often more useful to think of $\xi(r)$ in relation to the probability dP of finding two galaxies separated by a distance $\vec{r_2} - \vec{r_1}$ for a given number density $\bar{n}$ of galaxies,

$$dP = \bar{n_1}\bar{n_2}(1 + \xi_{GG}(\vec{r_2} - \vec{r_1}))d^3r_1 d^3r_2. \tag{1.6}$$

Taken in this way, ξ_{GG} is the two-point auto-correlation function of galaxies (Padmanabhan, 1993). A positive value for ξ_{GG}, at a given separation indicates the galaxies are more clustered than random. If ξ_{GG} is zero, the galaxy distribution at that separation is homogenous (Landy & Szalay, 1993). It is also useful to introduce the overdensity parameter

$$\delta(\vec{x}) = \frac{\rho(\vec{x}) - \bar{\rho}}{\bar{\rho}}, \tag{1.7}$$

where $\rho(\vec{x})$ is the density at point $\vec{x}$ and $\bar{\rho}$ is the mean density of the universe. In this formalization, a galaxy is an overdensity at given location, and can be related to the correlation function through

$$\xi(r) \equiv \langle \delta(\vec{r'})\delta(\vec{r'} + \vec{r}) \rangle. \tag{1.8}$$

In this form, the correlation function is the dependence of $\delta(r')$ on $\delta(r'+r)$. The concept of overdensity will be useful in our theoretical understanding of structure formation (§1.4.1). In Figure 1.9, we can see at small separations galaxies are very clustered. That clustering decreases as an approximate power law over many decades of separation, which is true regardless of whether that distance is measured in angular space (Connolly et al., 2001), redshift space or projected space (The SDSS Collaboration et al., 2010).

The most famous (and used) estimator for measuring the correlation function is the Landy-Szalay Estimator (Landy & Szalay, 1993), equation 1.9, where DD is the number of Data-Data pairs at a given separation, RR is the number of Random-Random pairs, and DR is the cross correlation of Data-Random pairs

$$\xi(r) = \frac{DD - 2DR + RR}{RR}. \tag{1.9}$$

This estimator measures the correlation function for any galaxy catalogue as long as we have an appropriate catalogue of random points. When generating these random

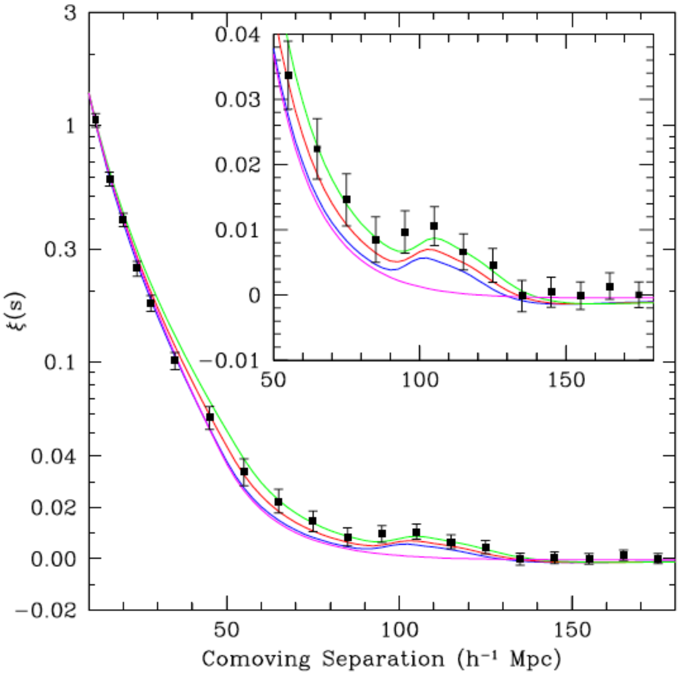

Figure 1.9 The Correlation Function measured in redshift space for a sample of SDSS galaxies (black points). The colored lines correspond to different cosmological models. The Baryon Acoustic Oscillation Peak is highlighted in the inset, see §1.3.11 (Eisenstein et al., 2005).

catalogues one must take into account the geometry of the data sample, as well as its completeness.

The separation can be measured in physical distance $\xi(r)$, comoving distance $\xi(s)$, projected distance $\xi_p(r_p)$ or angular separation on the sky $\omega(\theta)$. The advantage of measuring $\omega(\theta)$ over the other methods is that the calculation can be done without having the redshifts of the observed galaxies- a time consuming process when dealing with the large sample size needed to accurately measure the correlation function.

1.3.11 Recent Observations

As mentioned before, there has been great leaps forward in our understanding of the universe. A success of modern observational cosmology is the confirmation of the Baryon Acoustic Oscillation (BAO) peak in the galaxy correlation function. Sound waves that propagated in the hot plasma of the early universe imprinted a characteristic scale on the clustering of dark matter, galaxies and intergalactic gas (Eisenstein et al., 2005). This measurement can be seen in Figure 1.9. Much of the new extension of the Sloan Digital Sky Survey, SDSS-III, will be dedicated to measuring this ‘bump’ in the correlation function.

1.4 Theoretical Understanding of Observations

In this section, we will be constructing a theory that explains our observations of large scale structure and galaxy clustering. We will go from the initial energy distribution of the universe, to dark matter halos and finally how galaxies inhabit those halos, both in the universe and in our N-body simulations.

Dark matter halos are believed to form in a hierarchical manner, meaning that they start as small over-densities that accrete onto each other, forming larger and larger halos. Our theoretical understanding of galaxy evolution and clustering is based on dark matter halos. Galaxies are thought to form at the center of dark matter halos and, as such, understanding the formation, growth and dynamics of halos is key to understanding the observed properties of galaxies.

To theoretically predict the correlation function $\xi(r)$– our best statistic for measuring the observed clustering of galaxies- we require a spatial and mass distribution of halos (the halo mass function), the density profile of the dark matter halos, and ha description of the number and position of galaxies within a halo of a given mass (the Halo Occupation Distribution). This is called the Halo Model, and encompasses a large range of physics and statistics.

1.4.1 The Growth of Structure

To understand why our universe looks the way it does now, we have to understand where it came from. There must be a theory that connects the dramatic structures we observe today to those small anisotropies observed in the cosmic microwave background. There was an initial energy density perturbation imprinted from the Big Bang that grew differently depending on whether radiation or mass dominated the energy density of universe.

The final halo mass function is a product of both how matter collapses to form halos and how these halos interact and merge. We believe that structure in the universe forms hierarchically- meaning small, overdense objects collapse first and then merge together to form the most massive objects in the universe. The most massive objects in the universe, galaxy clusters, are therefore considered some of the youngest objects in the universe. Press and Schechter (Press & Schechter, 1974) first introduced this theoretical construct of hierarchical structure that has proven to be the foundation of modern structure formation.

1.4.2 Origin of Cosmic Structure

The early universe density perturbations are thought be due to quantum mechanical fluctuations in the early universe frozen in by an extremely accelerated expansion and can be observed in the CMB. These quantum fluctuations are thought to have a scale invariant power spectrum (e.g. Narlikar & Padmanabhan, 1991),

$$P = Ak^n, \tag{1.10}$$

where k is the wavenumber ($k = 2\pi/\lambda$), and the power spectrum index n is thought to be unity (e.g. Binney & Tremaine, 1987), meaning that perturbations on all scales behave the same way. It is important to keep in mind that these anisotropies in the CMB are on order of $\delta T/T \propto 10^{-5}$ (Dodelson, 2003), very small indeed (these anisotropies can be seen in Figure 1.1). So small, in fact, that we need to invoke inflation once again to explain how regions causally unconnected in the universe are so similar. This is called the horizon problem, and is solved by how rapid expansion was during the epoch of inflation, see for example Guth (1981) and Linde (1982).

1.4.3 Evolution of Structure

The largest structures in the Universe form through what is called gravitational instability- the slightly overdense regions (see equations 1.14) become increasingly overdense and eventually break away from the Hubble flow to form virialized dark matter halos. Gravity works like capitalism- the rich get richer and the poorer, poorer. Density perturbations on the scale of dark matter halos can only occur after matter has cooled sufficiently after the Big Bang so that it is no longer supported by radiation pressure. For normal baryonic matter, this occurs at the time of decoupling, z $\sim$ 1100, but for non-baryonic dark matter this collapse can start much earlier.

When these overdensities are small, we can make a linear approximation on how gravitational instability causes an overdense region to grow. In this approach, we see an overdensity as a small perturbation about the mean density field of the universe that can grow over time. Inverting equations 1.14, we see that the density at any location can be approximated as

$$\rho = \bar{\rho}(1 + \delta), \tag{1.11}$$

where $\delta << 1$.

Eventually, these small overdensities become too large ($\delta \approx 1$) to use a linear approximation of gravitational collapse. At this point, dark matter is collapsing into virialized halos- an equilibrium state where the energy of the gravitational collapse is transferred into the kinetic energy of the dark matter, causing pressure support and thus a stable halo of dark matter (Padmanabhan, 1993). See §1.4.9 for a model of this collapse. These halos form the fundamental framework of our current understanding of galaxy formation and clustering. Because dark matter has a 'head start' on collapse, baryonic matter falls into the pre-existing gravitational well of the dark matter halo (e.g. Ryden, 2003; Mo et al., 2010). It can then cool and condense, something dark matter cannot do, and can eventually form galaxies (White & Rees, 1978).

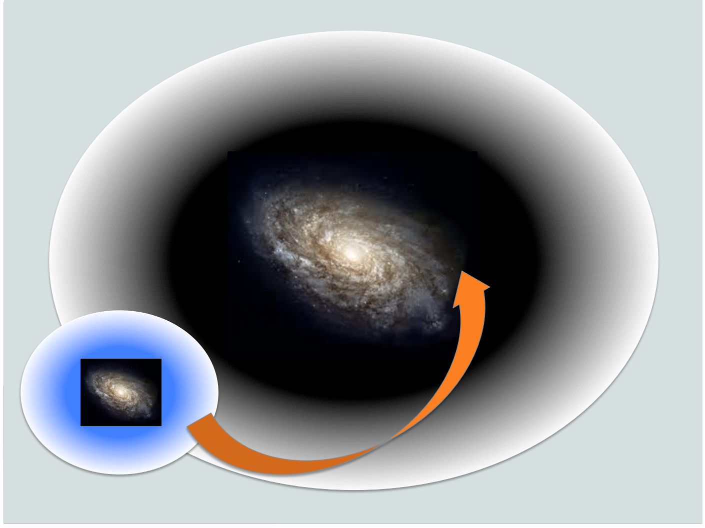

Figure 1.10 An illustration of a subhalo, in blue, merging into a host halo.

1.4.4 Mergers

Up until now, our theory only treats galaxies and dark matter halos as stationary objects, but to fully understand the observed structure of the universe we must understand mergers. Small dark matter halos grow by merging together to form more massive halos. The more massive halo is called the "host" halo and the smaller halo being merged becomes a "subhalo" within the host, see Figure 1.10 for an illustration. Both of these halos can contain baryonic matter, whether it be gas that may condense to form a galaxy or an already formed galaxy (Benson, 2010).

There are two main processes that explain what happens to a subhalo as it falls into and becomes merged into its host halo, dynamical friction and tidal stripping, (e.g. Binney & Tremaine, 1987). As a subhalo, and its galaxy, falls into the potential of the larger host halo, it will experience tidal forces that will strip away its least bound outer regions. The limit at which the self-gravity of an object is equal to the gravitational pull of the host

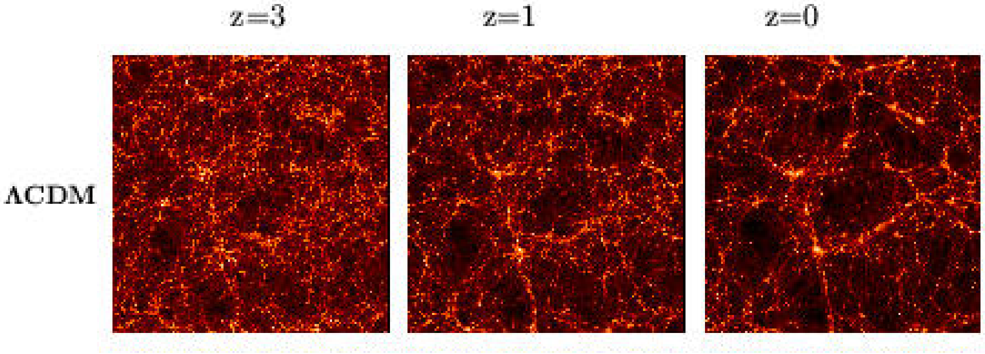

Figure 1.11 A mosaic of the structure formation and evolution of dark matter halos in a Λ cold dark matter cosmological simulation during three different redshifts, from Jenkins et al. (1998) for the VIRGO collaboration.

is called the tidal radius. Any mass outside the tidal radius of the object is able to be stripped away and forms streams of material[6]. As a dark matter halo moves through its host halo's mass, it leaves behind a wake of material. The wake of masses trailing behind the main body exerts a gravitational tug, called dynamical friction, on the main body, slowing it down. Dynamical friction is a complicated physical process that depends on the density profile and velocity dispersion of the host halo, among many other variables, and whose understanding greatly benefits from N-Body simulations.

1.4.5 N-Body Simulations

N-body simulations are a very powerful tool for understanding different physical processes. The equation of motion for gravitational collapse can only be analytically solved in the linear regime, making N-body simulations a necessity for calculating the exact evolution of the density field (e.g. Bertschinger, 1998). See Figure 1.11 for an example of

---

[6] Streams of merged galaxies are observed in the Milky Way- see Belokurov et al. (2006) for an example.

structure evolution in a cosmological N-body simulation. N-body simulations are not just used for cosmological simulations- there are a great variety of physical processes that can be better understood through simulations such as how galactic bars are formed through galaxy flybys (Lang, Holley-Bockelmann & Sinha, 2014). It is ideally suited, though, for studying the evolution of the density field because dark matter particles can be approximated as collisionless, discrete particles. The "N" in N-body refers to the number of particles for which we must compute the equations of motion.

1.4.6 Growth of Structure in a Cosmological N-body Simulation

Cosmological N-body simulations reproduce structure growth in the universe, where the initial conditions are dictated by the calculated power spectrum, the cosmology and the parameters of the simulation. The higher the resolution of the simulation– the larger number and smaller mass of the particles that make up dark matter halos– the more computationally expensive it will be.

Before we can even worry about the time evolution of particle positions and velocities, initial positions and velocities must be assigned. Once the initial mass power spectrum has been calculated, there are a few methods for calculating positions and velocities of the dark matter particles. The first of these approximations is the Zel'dovich approximation (Zel'dovich, 1970) which perturbs an initial Cartesian lattice of dark matter particles in the simulation 'box' (see fig 1.12) based on the growth of linear perturbations; once dark matter halos start forming, this approximation becomes inaccurate. Other methods for determining how structure grows are based on either linear perturbation theory or second order perturbation theory. These methods are also approximations of how structure grows

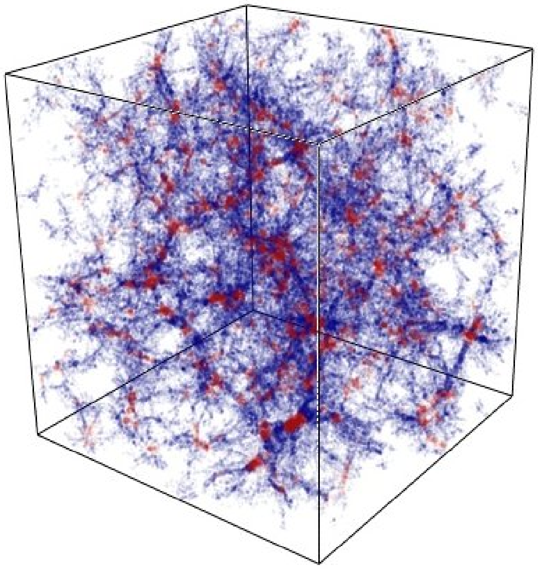

Figure 1.12 A dark matter simulation cube of size 180 Mpc/h containing 512^3 dark matter particles, in blue. The red areas denote clusters (Hahn et al. 2006).

in between the time the initial power spectrum is calculated and when the actual force integrations are done- the higher the redshift (earlier time) the simulation starts, the more accurate it will be.

Unfortunately, force calculations can be very costly. Without any approximations, a cosmological N-body simulation must calculate the force on each particle from every other particle in the simulation in order to determine the equations of motion. This is called an N^2 process, so the amount of time the simulation takes to run increases as roughly the number of particles in the simulation squared. There are thankfully improvements and short-cuts to take, starting with the idea of a mesh or grid. Most of the gravitational acceleration on a particle is due to other particles clustered to it within a distance much smaller than the size of the simulation. Once the force from those particles has been calculated, the rest of the box can be divided up into a grid where the mass of each grid point is the sum total of all the particles within. Another approximation is a tree-algorithm, where particles are recursively placed in cubic boxes based on how far away from the force calculation they are located (Barnes & Hut, 1986). In this way, only particles very close are counted individually while those farther away are grouped together in a box, reducing the time needed for the force calculation.

One of the final caveats to the force calculation is the idea of 'softening.' Dark matter particles in a simulation are very massive discrete points, and the force between particles can increase unphysically at small particle separations $\vec{r}$. To account for this, we 'soften' the gravitational force at a given radial separation so it asymptotes to a finite value as $r \to 0$. The softening radius represents a resolution limit, kiloparsecs for many cosmological simulations, at which we can no longer trust the positions of particles (Bertschinger, 1998).

1.4.7 Defining A Halo

We will see that the mass of the halo is the most important factor in deciding how many galaxies are placed in it, and therefore defining the particles that constitute a single halo is extremely important. A Friends-of-Friends (FoF; Davis et al., 1985) algorithm identifies a halo based on a characteristic 'linking length' between particles, see Figure 1.13 for an illustration of this. This linking length is usually some fraction of the mean inter-particle separation, with 0.2 being a common choice motivated by physical processes, see §1.4.9 for a discussion of these processes. All particles that are within a linking length of another particle are grouped into a halo. When a halo is defined this way, its mass is very dependent on the choice of linking length- a larger linking length will include more particles and will make the halo more massive.

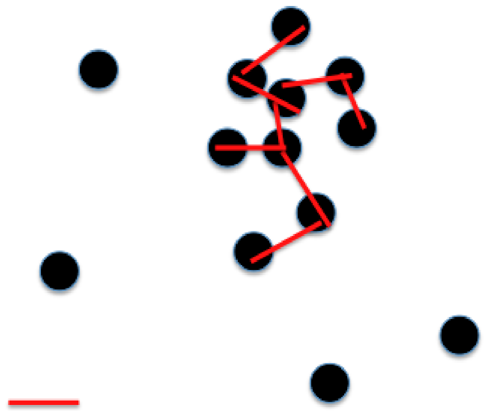

Figure 1.13 An illustration showing how the friends-of-friends algorithm connects dark matter particles in a simulation to form a halo. The black circles are the particles and the red line in the lower left hand corner is the linking length. Particles within a linking length of another particle will be grouped together

1.4.8 Halo Model

1.4.9 Halo and Galaxy Bias

Bias is the idea that an observable is a function of an underlying distribution. The location of cities are biased tracers of the distribution of people- what we call cities are just some location where the density of people is much higher than the average density. Just like cities, dark matter halos are a biased function of the mass distribution (Press & Schechter, 1974; Mo & White, 1996). In the Press-Schechter (PS) formalism, halos start as regions of space that have an overdensity δ equal to some critical overdensity δ_c, as shown in Figure 1.14. The location of these peaks is dependent (or biased) on the background matter distribution, which in turn is a product of the initial density fluctuations. It is assumed that once the density inside an area reaches δ_c, gravitational

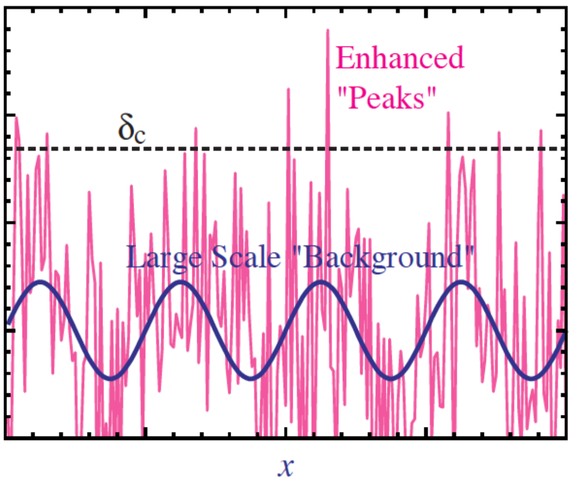

Figure 1.14 Overdensity of dark matter (y-axis) as a function of position. Regions with an overdensity greater than δ_c (dashed line) will collapse into virialized halos. These collapsed regions ("Enhanced Peaks") correspond to peaks in the initial Gaussian density fluctuations ("Large Scale 'Background'"). Image credit: Risa Wechsler

collapse will runaway from the Hubble expansion and the mass inside will eventually form a virialized dark matter halo (Padmanabhan, 1993).

The value of δ_c can be derived through the spherical collapse model. This model assumes that an object whose overdensity is too high for the linear regime (see §1.4.3) can be approximated as a sphere undergoing gravitational collapse, (see Cooray & Sheth, 2002, for a review). One useful result from this formalism is that a region of space which starts with δ_c becomes a virialized halo whose density is Δ_c times the critical density at that redshift[7], where $\Delta_c \simeq 200$ is often used, (Binney & Tremaine, 1987; Bryan & Norman, 1998). This result informs how halos are defined in N-body simulations, §1.4.7; a Friends-of-Friends linking length is calibrated so that each halo has a $\Delta_c \simeq 200$ (Jenkins et al., 2001).

---

[7] Critical density today $\rho_{crit} \simeq 2.7x10^{11} M_{\odot} Mpc^{-3}$. The density of the Milky Way measured at a radius of a few kpc is 10^5 times this (Binney & Tremaine, 1987).

Dark matter halos are not the only biased tracers of the mass distribution. We know that galaxy formation and evolution encompasses much more physics than that of dark matter halos, and that luminous matter may not perfectly follow the distribution of mass, even on the largest scales (Peacock, 1999). It is useful to employ a biasing parameter as a way to scale the underlying mass correlation function so that it corresponds to the correlation function of an object of interest,

$$\xi_{ii}(r) = b_i^2 \xi_{mm}(r), \tag{1.12}$$

where $\xi_{ii}(r)$ can be the halo-halo or galaxy-galaxy correlation function, and $\xi_{mm}(r)$ is the correlation function of the matter distribution, often measured as $\xi(r)$ of particles in an N-body simulation. The environmental bias parameter b_i enables us to encode all the physics that causes δ_i to not be exactly δ_{mass}.

On distance scales larger than that of a dark matter halo, the galaxy correlation function can be seen as a biased tracer of the spatial distribution of dark matter halos, which in turn is a function of the mass clustering. This often referred to as halo bias b_{halo} and can be empirically derived using N-body simulations (§1.4.5). Recent simulations have shown b_{halo} to be an increasing function of the mass of a halo (Seljak & Warren, 2004). Halo bias is a key component to understanding how galaxies cluster.

For scales within the size of a dark matter halo, the difference between the spatial distribution of galaxies and dark matter has its own type of bias that can be parameterized using the Halo Occupation Distribution (§1.4.12). See Figure 1.15 for an illustration of both clustering within one halo and among a distribution of halos.

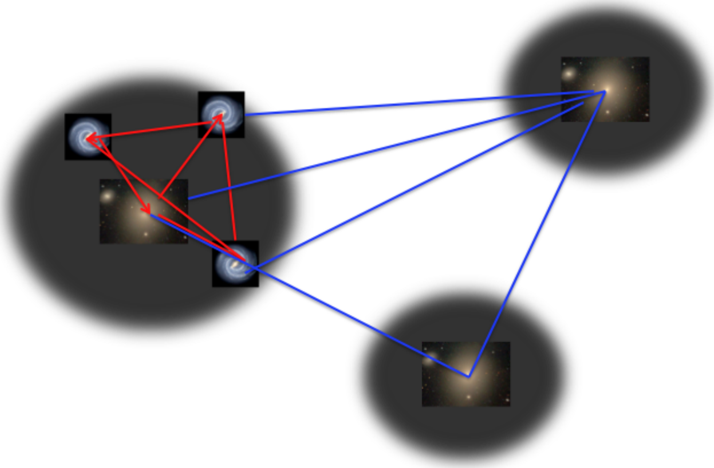

Figure 1.15 An illustration of the parings of galaxies in one halo (red lines) as well as the pairings of galaxies between two halos (blue lines). The pairings of the blue lines depend on the halo distribution in the universe, whereas the pairings of the red lines are determined by the distribution of galaxies within one halo.

1.4.10 Halo Mass Function

We now have the preparation to address our three ingredients to a theoretical galaxy distribution on scales within a single halo. The first of these ingredients is our halo mass function. We assume all mass is associated with virialized dark matter halos that obey a given mass density distribution. This distribution, the halo mass function, measures the number density of halos as a function of mass. A halo mass function will be characterized by a power law slope at low halo masses until a characteristic mass, $\mathrm{M}_{\star}$, at which point there is a exponential cutoff, see Figure 1.16. This mass is roughly $10^{12} M_{\odot}$, about the same mass as the Milky Way's halo.

Analytical calculations of the halo mass function, based on the work of Press and Schechter, have proven fruitful though inexact[8]. In the PS formalism, the spatial bias of

---

[8] These calculations over-predict the number of low mass halos while under-predicting high mass ones

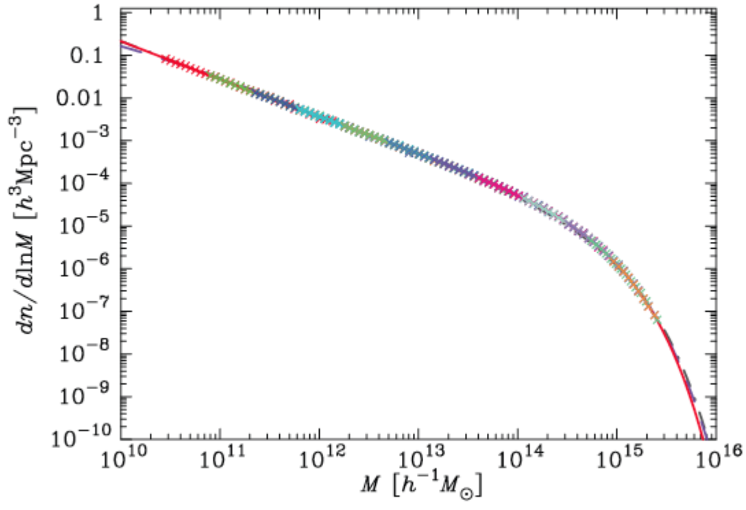

Figure 1.16 The normalized halo mass function from a suite of high resolution cosmological N-body simulations, showing the number density of dark matter halos as a function of their mass (Warren et al., 2006). The different color crosses represent the different simulations, while the the lines are different models and fits to the mass function.

halos (see §1.4.9) is used to estimate the fraction of the volume that has an overdensity $\delta > \delta_c$, and from that the number density of these regions as a function of mass (Cooray & Sheth, 2002; Dodelson, 2003). We now have the spatial distribution (bias) of halos as well as their mass function, but clustering on scales within the size of a dark matter halo must depend how galaxies populate the inner structure of a halo (Cooray & Sheth, 2002). For this, we need the density profile of mass in a halo and the number of galaxies that occupy a halo of a given mass.

1.4.11 Density Profiles

Before the galaxy-mass spatial distribution bias is determined, we first need to derive the density profile of the dark matter. Measuring density profiles of halos comes naturally

---

(Peacock, 1999).

from very high resolution N-body simulations, with one of the most widely accepted density profiles being the Navrro, Frenk and White density profile (NFW; Navarro et al., 1997)

$$\rho(r) = \frac{\rho_s}{(\frac{r}{r_s})(1+\frac{r}{r_s})^2}; \tag{1.13}$$

where ρ_s is the mean overdensity of the halo and r_s is the scale radius at which the slope of the density function changes. This result will prove very important for how the biased galaxy clustering is modeled, §1.4.9. While it may look as if there are two free parameters in an NFW profile, in practice the mean density ρ_s of a halo is determined by the mass and virial radius R_{vir} of a halo in a simulation. The concentration of a halo is defined as $c \equiv R_{vir}/r_s$, and has also been shown using simulations to be a decreasing function of the mass of a halo (Bullock et al., 2001).

Another interesting result from N-body simulations is that the halo density profiles, and specifically their concentration are linked to the assembly history of halos (Wechsler et al., 2002). We can track the assembly history of dark matter halos in a simulation or model through a "Halo Merger Tree", showing the progenitors of a present day halo that have merged together to form the final halo. These trees give us important information about how and when halos grew in mass and the effect this has on their concentration. Halo density profiles most likely have a large effect on galaxy properties, such as galactic rotation curves[9] (Bullock et al., 2001; Wechsler et al., 2002).

---

[9] This has an interesting implication for the Tully-Fisher relation, an empirical relationship between the rotation of a spiral galaxy and its luminosity.

1.4.12 Halo Occupation Distribution

To describe the distribution of galaxies within a dark matter halo, we use the Halo Occupation Distribution (HOD; Berlind & Weinberg, 2002). Within this framework, the halo occupation numbers, the number of galaxies in a halo, can be parameterized, as well as the spatial and velocity bias. Taken together, these three bias parameters represent a complete model of galaxy clustering for a given cosmology, and a testable framework that can provide insight into galaxy formation. When applied to a cosmological N-body simulation, the HOD allows us to populate dark matter halos with galaxies, creating a mock galaxy catalogue that provides the bridge between the theoretical construct of the halo model and observed galaxy clustering.

The first of these biasing parameters, halo occupation number, tells us how many galaxies are in a given halo, solely based on the mass of that halo. This parameter is represented as $P(N_{gal}|M_{halo})$, where the mean number of galaxies per halo of mass M is given as

$$N_{avg} = \sum_{N} NP(N|M). \tag{1.14}$$

The parameterization of N_{avg} has evolved from a simple power law (Berlind & Weinberg, 2002) to complicated functional forms that have been fit to data (Zheng et al., 2007a). To illustrate, a power law form is

$$N_{avg} = \begin{cases} 0 & \text{if} \quad M < M_{min} \\ (M/M_1)^{\alpha} & \text{otherwise.} \end{cases} \tag{1.15}$$

Here M_{min} is the lowest mass halo that can still host a galaxy, M_1 is the halo mass that contains on average one galaxy and α is the rate at which additional galaxies are added. Physically, these parameters will change depending on the particular class of galaxy studied, for instance, the M_1 for halos hosting a very massive, luminous galaxy will be much larger than the M_1 for dwarf galaxies. The other parameters, M_{min} and α, are also product of the physics of galaxy formation and evolution, including to host halo mergers. Additional parameters are introduced to isolate specific physics, for instance a characteristic mass at which a halo has on average one central and one satellite galaxy (Zheng et al., 2005).

Once the halo occupation number has been calculated, we can turn our attention to the other two features of the HOD- the spatial and velocity bias of the galaxies. In a halo with only one galaxy, that galaxy could resides at the halo center- with additional galaxies will act as satellites, that trace the density profile of the halo (Peacock & Smith, 2000). The difference between the density profiles of galaxies and mass can be modeled as

$$\rho_g(r)/\rho_m(r) = r^{\Delta\gamma}, \tag{1.16}$$

where $\rho_m(r)$ can be defined to be an NFW density profile and $\Delta\gamma$ is the biasing parameter that ifs adjusted based on observables. When assigning galaxy positions in an N-body simulation, $\rho_m(r)$ is taken to be that of the particles in the halo. A galaxy is assigned to a particle based on the value of $\Delta\gamma$; if this value is zero, a random dark matter particle is picked to ‘host’ the galaxy.

Once a galaxy has been assigned to a dark matter particle in a simulation, this same approach can be taken for velocity bias, where the velocity of a particular galaxy can be given as

$$\mathbf{v}_{galaxy} = \mathbf{v}_{halo} + \alpha_v(\mathbf{v}_{particle} - \mathbf{v}_{halo}), \tag{1.17}$$

where $\mathbf{v}_{halo}$ is the center-of-mass velocity for the host halo. The velocity and velocity biasing of galaxies are important elements to incorporate in our model because of redshift space distortions. The real-space clustering statistics can be compared to calculations done in redshift space, comparing both to the measured statistics of galaxies. Taken together, the HOD parameters: P(N—M), $\Delta\gamma$, and α_v cover the entirety of galactic physics that can influence the distribution of galaxies within a single halo and provides the necessary link that connects theory to observations. See Figure 1.17 for a comparison of data and theory.

1.5 Summary

The problem with a theory describing the universe as being compared of dark matter halos that host galaxies is that we cannot directly observe these halos, only observe the light- galaxies, stars, quasars. From this light, we can $infer$ the properties of these dark halos but not directly measure its mass, density or radius. Inference is costly; we need a statistically significant population of galaxies to say anything definite about the composition of the universe, driving the need for the next generation of galaxy surveys.

To understand the large scale structure of the universe, we must understand: (i) how dark matter halos populate the universe and (ii) how galaxies populate the dark matter halos. Much theoretical work has been done to analytically determine (i), but the most promising results are coming from cosmological N-body simulations. The second factor

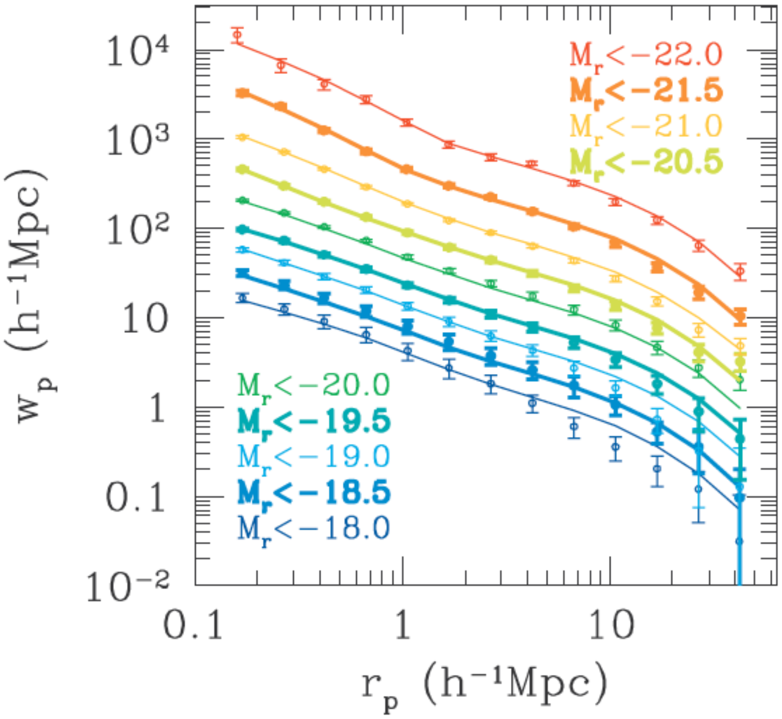

Figure 1.17 The projected correlation $w_p(r_p)$ function for different volume limited samples of galaxies (points) from the SDSS and their corresponding HOD model (lines), (The SDSS Collaboration et al., 2010). The data and corresponding lines are staggered 0.25 dex for clarity. The HOD is able to model the clustering of galaxies across a wide range of luminosities, corresponding to much theoretical physics, including galaxy formation and evolution.

requires more complicated physics that can be parameterized using the Halo Occupation Distribution. The basic equation for recreating observables now is an N-body combined with a HOD gives us a Mock Galaxy Catalogue. By comparing this catalogue with data using the galaxy correlation function, we can constrain our understanding of both halo clustering and as galaxy formation and evolution.

Work is being done on all fronts- new data is being taken by SDSS as we speak, larger N-body simulations are being run, ones that include the gas physics needed for galaxy formation, and the parameters of the HOD are being refined to better represent the underlying physics. Together, the theory and observations are working towards a complete understanding of large scale structure formation and evolution.

Chapter II

THE SPATIAL DISTRIBUTION OF SATELLITE GALAXIES IN THE LOCAL UNIVERSE

2.1 Introduction

One of the best statistical tools we have for an understanding of the galaxy distribution in the universe is the two-point correlation function of galaxies. On large scales (approximately greater than $10h^{-1}$Mpc), galaxies are simple tracers of the underlying matter density field and so the correlation function (or its Fourier equivalent, the power spectrum) can be used to probe the nature of matter fluctuations, and thus yield constraints on cosmological parameters (e.g., Tegmark et al., 2004b,a). At a scale of $\sim 100h^{-1}$Mpc, the Baryon Acoustic Oscillation (BAO) feature in the correlation function also provides strong cosmological constraints (e.g., Eisenstein et al., 2005).

On scales smaller than $\sim 10h^{-1}$Mpc, the galaxy correlation function encodes information about the detailed relationship between the spatial distribution of galaxies and the underlying dark matter, which is substantially more complex than on large scales. Adopting the assumption that all galaxies live within dark matter halos, the halo model provides a useful roadmap for interpreting galaxy clustering on these scales. In the halo model framework, the clustering of galaxies can be calculated from statistical properties of halos, such as their abundance, clustering and internal structure, combined with parameterized relations that describe how galaxies occupy halos. This latter part is referred to as the Halo Occupation Distribution (HOD) and it typically specifies the number of galaxies as a function of halo mass, together with an assumption for the their spatial distribution within halos (e.g., Peacock & Smith, 2000; Scoccimarro et al., 2001; Berlind & Weinberg, 2002; Cooray & Sheth, 2002; Berlind et al., 2003; Kravtsov et al., 2004; Zheng et al., 2005). Several studies have used the measured galaxy correlation function on scales of $\sim 0.1 - 10h^{-1}$Mpc to constrain the HOD and thus illuminate the nature of the connection between galaxies and dark matter (e.g., Zehavi et al., 2005b, 2011; Zheng et al., 2007b; Guo et al., 2014, and references therein).

On very small scales, well within $0.5h^{-1}$Mpc, the typical size of halos that host bright galaxy pairs, the shape of the correlation function is primarily dictated by the spatial distribution of galaxies in each halo (e.g., Berlind & Weinberg, 2002; Zehavi et al., 2005b). Most studies adopt a simple model whereby the first "central" galaxy in each halo lives at the halo center, and subsequent "satellite" galaxies trace the density distribution of the dark matter. Specifically, satellite galaxies are usually assumed to follow a Navarro-Frenk-White (NFW; Navarro et al., 1997) profile, which does a good job of describing the density profiles of halos in pure dark matter N-body simulations. This assumption is theoretically motivated (e.g., Berlind et al., 2003) and it works well in explaining the observed shape of the correlation function on small scales.

The first evidence from galaxy clustering that satellite galaxies might not actually trace mass within halos came from Masjedi et al. (2006) who pushed the measurement of the galaxy correlation function down to scales of $10h^{-1}$kpc. Using a sample of Luminous Red Galaxies (LRGs; Eisenstein et al., 2001) selected from the Sloan Digital Sky Suvey (SDSS; York et al., 2000), Masjedi et al. (2006) found that the correlation function of LRGs at the smallest scales ($\lesssim 30h^{-1}$kpc) was under-predicted by the Zehavi et al. (2005b) HOD model that had successfully fit the clustering of the same galaxies at larger scales. Specifically, the HOD model predicted a r^{-1} slope for the correlation function at the smallest scales (which comes from the inner slope of the NFW profile), whereas Masjedi et al. (2006) found a much steeper r^{-2} slope. Watson et al. (2010) explored this discrepancy in detail by fitting the Masjedi et al. (2006) correlation function measurements with a HOD model that relaxed the assumption that satellite galaxies follow a NFW profile. Instead, they adopted a more flexible profile where the inner slope was allowed to vary. Watson et al. (2010) were able to obtain a good fit to the LRG clustering for a satellite galaxy profile with an r^{-2} inner slope while ruling out the NFW profile at high significance.

Watson et al. (2012) extended this work to a wider range of galaxy luminosities. They fit their flexible HOD model to measurements of the projected correlation function, $w_p(r_p)$, in several SDSS luminosity samples, ranging from absolute r-band magnitude of -18 on the faint end, to LRGs on the bright end. These measurements were made by Jiang et al. (2012) using the same methods as Masjedi et al. (2006) for pushing to very small scales. Watson et al. (2012) found a clear luminosity trend whereby the clustering of galaxy samples with $M_r <$ -20 and brighter demanded steeper density profiles for satellite galaxies than NFW, whereas lower luminosity samples were consistent with NFW satellite profiles. Guo et al. (2014) adopted the same flexible density profile when modeling the clustering of galaxies in the SDSS III (Eisenstein et al., 2011) Baryon Oscillation Spectroscopic Survey (BOSS; Dawson et al., 2013), and also found a significant departure from NFW, albeit only for the reddest galaxies in that survey. Unfortunately, it is difficult to directly compare these results with those of Watson et al. (2012) because of the substantially different sample selections. Using a different technique that does not involve correlation functions, Tal et al. (2012) found that satellite galaxies around LRGs deviate from NFW at very small scales, in agreement with Watson et al. (2010). On the other hand, Guo et al. (2012b) used a similar technique to find that satellite galaxies have density profiles that are consistent with NFW. Deep imaging of satellites around luminous Early-type galaxies at intermediate redshifts have shown an isothermal profile (Nierenberg et al., 2011) with no dependence on host mass (Nierenberg et al., 2012).

Measurements of the galaxy correlation function on such small scales suffer from two potentially severe systematic errors. First, two bright galaxies that are only separated by $\sim 10 - 30h^{-1}$kpc are likely in the process of merging and will have overlapping light profiles. It can be difficult to accurately de-blend the observed light into two separate components and a sufficiently large error in the assigned magnitude of either galaxy can cause the pair to either enter or drop out of a luminosity selected sample. Second, in surveys that use fiber-fed multi-object spectrographs, it is not possible to obtain spectra of both galaxies that are separated by less than the physical diameter of the fibers. In the SDSS, these "fiber collisions" enter at an angular scale of $55''$ (Blanton et al., 2003a). At the typical redshifts of SDSS galaxies, this corresponds to a much larger physical scale than $30h^{-1}$kpc. About a third of these collided galaxy pairs are recovered in the SDSS because part of the survey footprint is observed ("tiled") more than once. However, the spatial distribution of this overlap region is very complex. Incompleteness due to fiber collisions affects the correlation function the most on the smallest scales, but the $55''$ angular scale translates into many different length scales in real and projected space, so even large scales are affected. Various methods have been used to correct for fiber collisions. The simplest method is to assign collided galaxies the redshifts of their nearest neighbors.

This works well on large scales, but not small scales. Masjedi et al. (2006) and Jiang et al. (2012) used an estimator for $w_p(r_p)$ that corrects for fiber collision incompleteness statistically. Guo et al. (2014) used a different method that essentially only considers galaxy pairs in overlap regions (Guo et al., 2012a). It is important to correctly account for these systematic effects before drawing any conclusions about the inner density profile of satellite galaxies.

It should not necessarily come as a surprise that satellite galaxies may not be perfect tracers of dark matter. The spatial distribution of satellite galaxies can be affected both by dynamical mechanisms, such as dynamical friction and tidal stripping of stars due to the host halo potential, and by baryonic processes, such as quenching of satellite star formation. A detection of a departure from the dark matter profile in the satellite density profile can thus serve as a probe of these processes. Theoretical predictions of the satellite galaxy density profile at such small scales are difficult to make because it is challenging to resolve massive distinct satellite halos (i.e., subhalos) so close to the center of a larger host halo. Nevertheless, both pure N-body and hydrodynamic simulations are now achieving the resolutions and volumes necessary to compare with SDSS data (e.g., Pujol et al., 2014; Genel et al., 2014).

In this paper, we test the validity of the Watson et al. (2012) results using the same galaxy selection, but an entirely different methodology. First, we measure the angular correlation function $\omega(\theta)$, instead of the projected function $w_p(r_p)$. In general, $\omega(\theta)$ is a powerful tool for two-dimensional galaxy surveys (see Crocce et al. 2011 and references therein). It has been employed to measure the galaxy clustering in the early data release of the SDSS (Connolly et al., 2002; Scranton et al., 2002; Infante et al., 2002; Budavári et al., 2003), as well many other galaxy surveys (e.g., Groth & Peebles, 1977; McCracken et al., 2001; Maller et al., 2005). The angular function is less sensitive to fiber collisions because the fiber incompleteness enters at a fixed scale and thus does not contaminate larger scales. Moreover, we restrict our samples to survey overlap regions, which reduces the effects of fiber collisions even more. Second, we improve on the HOD modeling by switching to an accurate and fully numerical way of computing clustering predictions, instead of the quick and approximate analytic method that was used in Watson et al. (2012).

The description of our data samples appears in §2.2. The $\omega(\theta)$ measurements, along with power-law fits, are described in §2.3. The description of our modified density profile HOD model is in §2.4, with results of the model fits presented in §3.4. In §2.5 we summarize our results and discuss their implications. Finally, we discuss fiber collision incompleteness in the Appendix.Throughout this paper, we assume a standard ΛCDM cosmology in concordance with the best fit WMAP5 parameters.

2.2 Data Sample

Measuring angular correlations does not usually require galaxies with measured redshifts. However, we wish to constrain the density profile of satellite galaxies within their halos for different luminosity samples so that we can test the Watson et al. (2012) results. We therefore need volume-limited samples built from a spectroscopic sample. We use data from the SDSS Data Release 7 (DR7; Abazajian et al., 2009). Specifically, we use the large-scale structure samples from the NYU Value Added Galaxy Catalog (NYU-VAGC; Blanton et al. 2005), that were built from the SDSS main galaxy sample (Strauss

Table II.1. Volume-limited Samples and Power Law Fits

M_r^{lim}	z_{min}	z_{max}	N_{gal}	Collision Scale (kpc/h)	Median Slope	χ^2dof
-18	0.02	0.042	18690	25.7	-0.70 ± 0.05	1.45
-19	0.02	0.067	41515	39.9	-0.77 ± 0.02	0.70
-20	0.02	0.106	67108	59.9	-0.74 ± 0.02	1.06
-21	0.02	0.165	43528	89.1	-0.92 ± 0.02	0.779

Note. — The table shows the absolute magnitude and redshift limits of each sample, the number of galaxies, the physical scale of fiber collisions at the median redshift of the sample, and the median slope and best-fit χ^2 from fitting a power law to the angular correlation function.

et al., 2002). The main spectroscopic galaxy sample is approximately complete down to an apparent r-band Petrosian magnitude limit of < 17.77. However, we have cut our sample back to $r < 17.6$ so that it is complete down to that magnitude limit across the sky. Galaxy absolute magnitudes are k-corrected (Blanton et al., 2003c) to rest-frame magnitudes at redshift $z = 0.1$.

We construct four volume-limited samples that are complete down to absolute r-band magnitude limits of -18, -19, -20, and -21. When constructing the volume-limited samples, we adopt corrections for passive luminosity evolution (Blanton, 2006), which results in slightly evolving absolute magnitude limits as a function of redshift (the magnitude limits listed above apply at $z = 0.1$). The four volume-limited galaxy samples are shown in Figure 2.1 and their redshift limits and sizes are summarized in Table II.1.

The galaxy redshift sample has an incompleteness due to the mechanical restriction that spectroscopic fibers cannot be placed closer to each other than their own thickness.

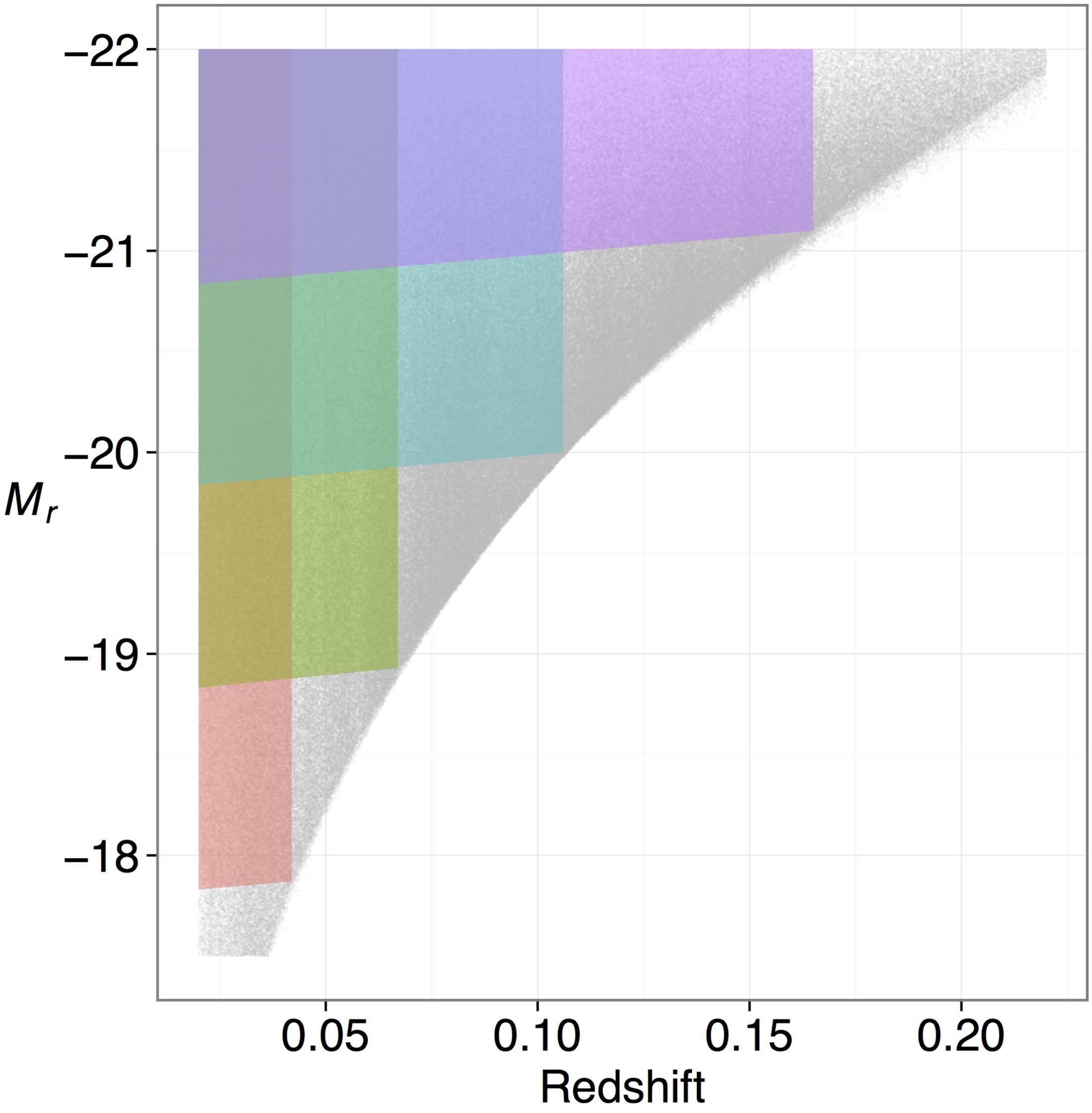

Figure 2.1 The four volume-limited samples that we use in this study, shown in absolute r-band magnitude vs. redshift. Light grey points show the full flux-limited sample from which the volume-limited samples were selected. The absolute magnitude limits of the four samples evolve slightly with redshift to account for passive luminosity evolution in the galaxy population.

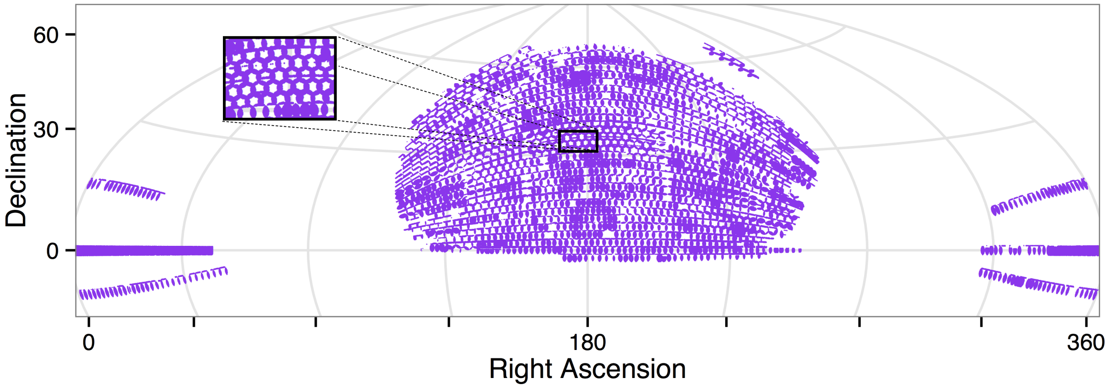

Figure 2.2 The footprint on the sky (Hammer projection) of the SDSS ‘overlap’ sample that we use in this paper. The sample consists only of regions that were spectroscopically observed more than once. The area of this footprint is roughly 40% of the full SDSS DR7 footprint.

This fiber collision constraint makes it impossible to obtain redshifts for both galaxies in pairs that are closer than 55″ on the sky. In the case of a conflict, the target selection algorithm randomly chooses which galaxy gets a fiber (Strauss et al., 2002). Spectroscopic plate overlaps alleviate this problem to some extent, but fiber collisions still account for a $\sim 6\%$ incompleteness in the main galaxy sample. On the very small scales that we are considering in this paper, fiber collision incompleteness is severe. The 55″ angular scale translates to physical scales of $25 - 90h^{-1}$kpc at the median depths of our four samples, which is right in the interesting region we wish to study.

A commonly used correction for fiber collisions in galaxy clustering studies is to assign fiber collided galaxies the redshift of the galaxy they collided with (i.e., the “nearest neighbor correction”; Zehavi et al., 2002). This correction recovers the true correlation function on large scales, but it performs poorly on small scales. Masjedi et al. (2006) and

Jiang et al. (2012) addressed this problem by proposing a new estimator for the projected correlation function. Instead of computing an autocorrelation function of spectroscopic galaxies, they computed a cross-correlation between spectroscopic galaxies and all spectroscopic targets from the imaging survey. For each pair, the imaging galaxy was placed at the same redshift as the spectroscopic galaxy. This procedure recovers all the fiber collided pairs, but it also includes an artificial signal from pairs that are actually uncorrelated. The uncorrelated pairs are then statistically removed from the correlation function using a random catalog. In this paper, we adopt a different approach. We apply the nearest neighbor correction to recover collided galaxies without redshifts, we construct our samples, and then we measure the angular correlation function of galaxies, $\omega(\theta)$. The angular function is significantly less sensitive to errors in the assigned redshifts than the projected function $w_p(r_p)$ because the angular scale θ is not affected by these errors, whereas the physical scale r_p is. Errors in the nearest neighbor correction only affect $\omega(\theta)$ if they cause galaxies to enter or drop out of the volume-limited sample. For most collision pairs, the nearest neighbor correction does not result in the gain or loss of the pair in the sample. This only happens in special cases. For example, when a collision pair straddles the outer redshift limit of a particular volume-limited sample, if the more distant galaxy of the pair did not get a redshift due to the collision, the nearest neighbor correction will bring it into the sample and thus result in a new small-scale pair contributing to $\omega(\theta)$. Alternatively, if the higher redshift galaxy of a collision pair is close to the luminosity limit of the sample and did not get a redshift due to the collision, the nearest neighbor correction could make it exit the sample. This would result in a loss of a small scale pair contributing to $\omega(\theta)$.

The SDSS DR7 sample covers an area on the sky of approximately 8000 square degrees. However, to minimize the errors due to fiber collisions discussed above, we restrict the sample to regions on the sky that have been spectroscopically observed more than once (the so called "plate overlap" regions) as part of the tiling process (Blanton et al., 2003a). In these regions, which cover about 40% of the full SDSS footprint, the vast majority of collided galaxies have been recovered. However, we note that a region that has been tiled twice can only recover close *pairs* of galaxies. In order to measure the redshifts of close *triplets*, a region would have to be tiled thrice. This continues on to higher groups, which represent a small number of the collision groups, but a non-negligible fraction of pairs. The effects of fiber collisions are thus not completely removed from our analysis and we revisit this issue in §2.3 and in the Appendix. The total area of our sample is 3300 square degrees and we refer to it as the 'overlap' sample throughout this paper. We show the sample footprint in Figure 2.2.

2.3 Angular Correlation Function

2.3.1 Measuring $\omega(\theta)$

We measure $\omega(\theta)$ using the Landy & Szalay (1993) estimator

$$w(\theta) = \frac{DD - 2DR + RR}{RR}, \tag{2.1}$$

where DD, DR and RR are the correctly normalized number of data-data, data-random and random-random pairs in each bin of angular separation θ. We construct a random sample that has the same overlap geometry as the data sample and a size such that the amount of shot noise in the inner bins is not dominated by RR or DR.

In order to estimate errors and measure the covariance matrix, we separate the footprint into 100 jackknife samples that represent approximately equal area sections on the sky. For each jackknife sample k, we measure the angular correlation function $\omega^k(\theta)$. The covariance matrix can then be computed as

$$C_{ij} = \frac{N-1}{N} \sum_{k=1}^{N} (\omega_i^k - \bar{\omega_i})(\omega_j^k - \bar{\omega_j}), \tag{2.2}$$

where C_{ij} is the covariance between angular bins i and j, and $\bar{\omega_i}$ is the mean of correlation function measurements in angular bin i computed from the N jackknife samples. We will use the full covariance matrix to fit models to our measurements since neighboring data points in the angular correlation function are highly correlated (Connolly et al., 2002).

The measurement of $\omega(\theta)$ is done using `STOMP`, a C++ library platform for doing fast spatial statistics on arbitrary spherical geometries using 10s of millions of points[1]. The 100 jackknife samples of equal area on the sky are made using the `STOMP` libraries.

2.3.2 Data Results

Figure 2.3 presents our measurements of the angular correlation function $\omega(\theta)$, for the four volume-limited samples described in §2.2 in the range $7'' < \theta < 320''$. We choose this range of scales because on smaller scales photometric deblending effects are expected to be severe, while larger scales no longer probe the clustering of galaxies within a single dark matter halo. Masjedi et al. (2006) quantified the effects of photometric deblending

---

[1] http://code.google.com/p/astro-stomp/

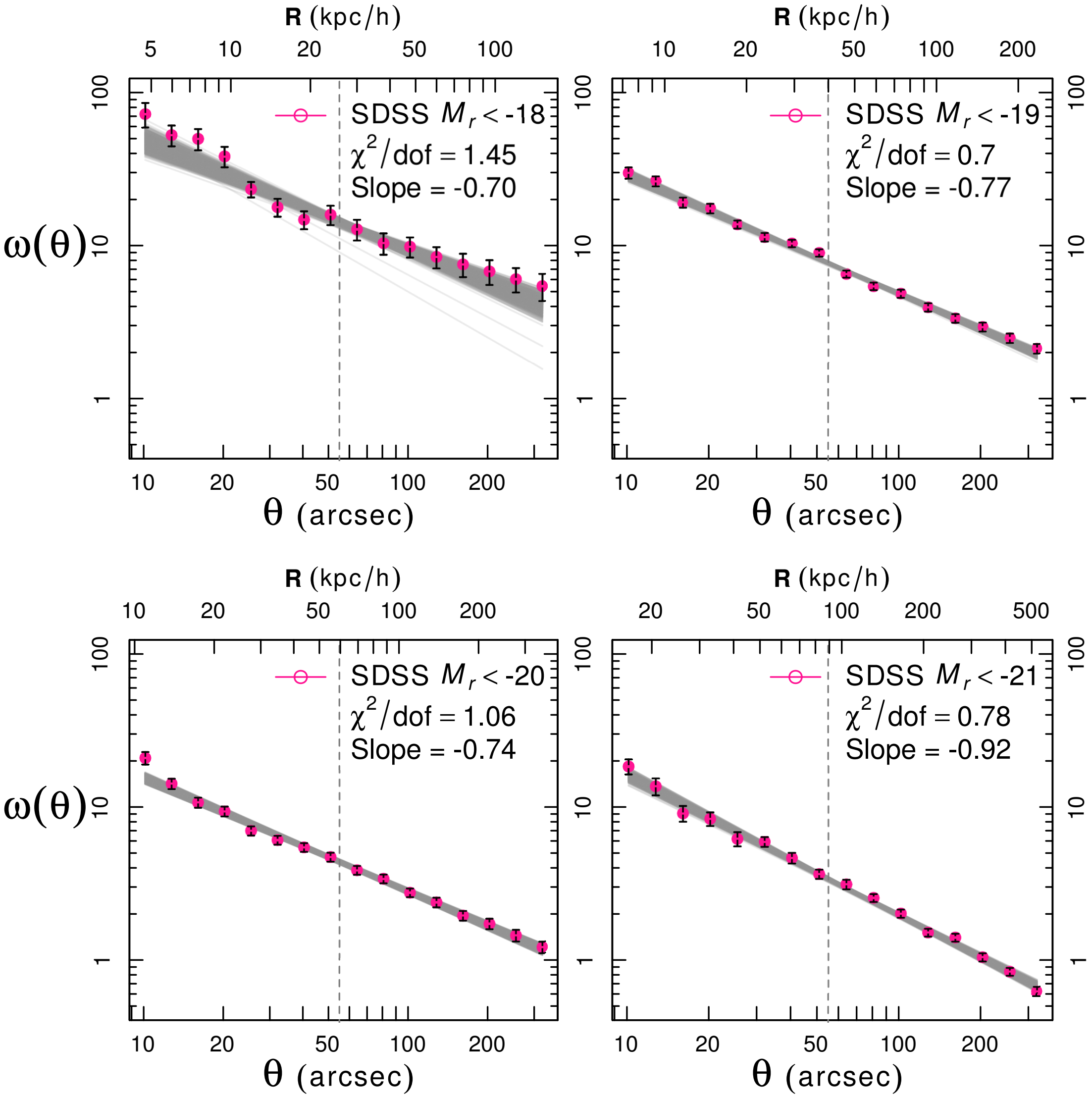

Figure 2.3 The angular correlation function of SDSS galaxies in four volume-limited samples, along with their power-law fits. Each panel shows results for a specific volume-limited sample, as described in §2.2. Points show the measurements, and error bars are estimated from jackknife resampling of the data on the sky. The bottom axis of each panel shows the angular scale in units of arcsec and the top axis shows the corresponding physical scale at the median redshift of each sample. The vertical dashed line in each panel denotes the fiber collision scale of 55″. The gray band shows a selection of power law models that are randomly drawn from the best-fitting 68% of models in the MCMC chain. The median value of the slope and the goodness of fit are listed in each panel.

on the correlation function for LRGs by adding artificial galaxy pairs into the raw SDSS images and studying how well the photometric pipeline recovered the light of each galaxy. They found that the clustering of LRGs is significantly overestimated on scales less than $20h^{-1}$kpc due to deblending errors, while larger scales are mostly unaffected. Since the physical sizes of galaxies decrease rapidly with decreasing luminosity, Jiang et al. (2012) calculated that it is safe to ignore photometric deblending effects for the lower luminosity samples and scales we consider here. The physical scales corresponding to these angular scales for the median redshift of each sample are shown at the top axis of each panel in Figure 2.3. For example, in the case of the $M_r <$ -20 sample, the physical range covered by our measurements is approximately $10h^{-1}\mathrm{kpc} < r < 300h^{-1}\mathrm{kpc}$, which is mainly probing the spatial distribution within halos.

The points in Figure 2.3 show the $\omega(\theta)$ measurements and the error bars are estimated from jackknife resampling, as described in §2.3.1 (they are the diagonal values of the covariance matrix). The M_r <-18 sample is significantly noisier than the other three because it is the smallest of our galaxy samples (see Table II.1). The amplitude of $\omega(\theta)$ is highest for the least luminous sample and drops progressively with luminosity. This is simply due to the fact that more luminous samples extend further in redshift, resulting in more uncorrelated galaxy pairs in each angular bin that dilute the clustering signal.

As we discussed in the previous section, we expect fiber collision errors to be small in these measurements. However, there are still some galaxy pairs lost and gained in special cases where the nearest neighbor correction applied to collision triplets and higher multiplicity collision groups causes galaxies to incorrectly enter or exit the volume-limited sample. One of the advantages of using the angular correlation function is that errors due to fiber collisions should appear as a sharp feature at 55″. An inspection of Figure 2.3 shows no such significant features, except perhaps for a small feature in the case of the M_r <-19 sample. The M_r <-18 sample measurement shows two small discontinuities at small scales, but these occur between bins four and five and again between bins seven and eight, whereas the fiber collision scale occurs between bins eight and nine. We think that it is more likely that these small scale discontinuities are due to noise, given that they are similar in amplitude to the size of the data error bars, and that they occur at the wrong scales to be obviously caused by fiber collisions. We therefore conclude that fiber collision errors are indeed likely small, as expected. However, we emphasize that our analysis method has not eliminated fiber collision incompleteness and that it is definitely present in our measurements, as we discuss in detail in the Appendix.

All four correlation functions look approximately like power laws by eye and the most luminous sample appears to have a somewhat steeper correlation function than that of the lower luminosity samples. Before we can fit any model to our measurements we must first estimate covariance matrices. We do this using jackknife resampling, as described in §2.3.1. Figure 3.18 shows the correlation matrix, which is the covariance matrix normalized by its diagonal elements, in the case of the M_r <-18 sample. The matrix clearly shows that nearby angular bins are highly correlated with each other, especially at large angular scales.

2.3.3 Power Law Fitting

Galaxy correlation functions are approximately shaped like power laws and so the power law model is often used to quantify their shape and amplitude. However, the near power-law shape of the galaxy correlation function is largely a coincidence (Watson et al., 2011), and it has been shown that the correlation function is not well described by a power law in a statistical sense, especially at the scales corresponding to the size of the typical dark matter halos that contain bright galaxies (Zehavi et al., 2004). Power law models are thus not accurate models and they do not directly yield a physical understanding of galaxy clustering. However, they are useful as a descriptive tool for quantifying the overall slope of the correlation function and for comparing the slopes of different galaxy samples. The inner slope of the density profile of satellite galaxies in halos directly affects the slope of the 3D correlation function on small scales, which in turn directly affects the slope of the angular correlation function. A steeper density profile for satellite galaxies should translate into a steeper $\omega(\theta)$ (e.g. Peebles, 1980; Efstathiou et al., 1991; Watson et al., 2010).

We fit a power law model to our $\omega(\theta)$ measurements using the MCMC code `emcee` (Foreman-Mackey et al., 2013), which we describe in more detail in §2.4.3. We calculate the χ^2 value for each model parameter combination using

$$\chi^2 = \sum_{ij} (\omega_i - \omega_{\mathrm{model},i}) C_{ij}^{-1} (\omega_j - \omega_{\mathrm{model},j}), \tag{2.3}$$

where ω_i and $\omega_{\mathrm{model},i}$ are the data and model correlation function in bin i, and C_{ij}^{-1} is the inverse of the jackknife covariance matrix from Equation 2.2.

Figure 2.3 shows a random sampling of power laws drawn from the best-fitting 68% of models in the MCMC chains. We list the median values and 68% confidence intervals of power-law slopes for all four samples in Table II.1, as well as the corresponding values

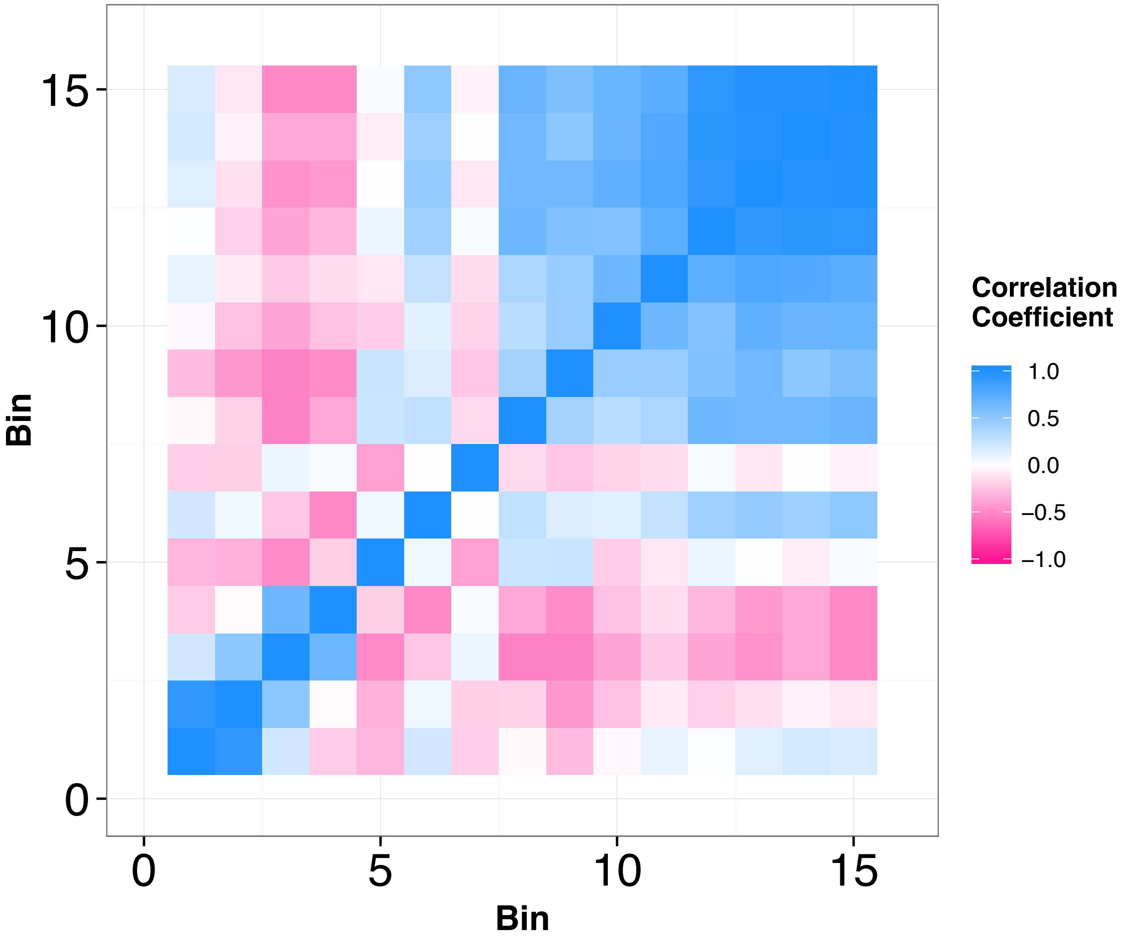

Figure 2.4 Correlation matrix for the angular correlation function of the M_r <-18 sample, derived from 100 jackknife resamplings of the data on the sky. The correlation matrix is simply the covariance matrix normalized by its diagonal elements and we compute it as described in §2.3.1.

of χ^2 per degrees of freedom. The correlation matrix for the M_r <-18 sample is shown in Figure 3.18. Finally, in Figure 2.5, we show the posterior probability density function of slope for each luminosity sample, as given from the MCMC chains. The best-fit χ^2 values indicate that a power law functional form provides a good statistical description of the shape of $\omega(\theta)$ for all four luminosity samples (the $M_r <$ -18 sample has a *p*-value of 0.12). Furthermore, the fit results show that the most luminous galaxy sample, $M_r <$ -21, has a significantly steeper power-law slope than the less luminous samples, while there is no trend in the steepness of the slope for the lower luminosity samples. This result seems to confirm the results of Watson et al. (2012), who found that only luminous galaxies had steep satellite density profiles. In the next section we fit our clustering measurements with a halo model in order to directly probe what constraints we can place on the distribution of satellite galaxies within halos.

2.4 Halo Modeling

Most previous studies fitting halo models to clustering measurements, including Watson et al. (2010, 2012), have used an analytic framework to compute the correlation function. This framework requires analytic approximations for the halo mass function, the large scale bias of halos, and the halo density profile and it combines them together with a parameterized HOD to predict the distribution of galaxy pairs. Analytic halo models are fast and reasonably accurate; however, one should be cautious before trusting them at better than the $\sim 10 - 20\%$ level. In this paper, we adopt a fully numerical procedure that eliminates most of the systematic errors that are present in analytic models. We populate dark matter halos in cosmological N-body simulations with galaxies according

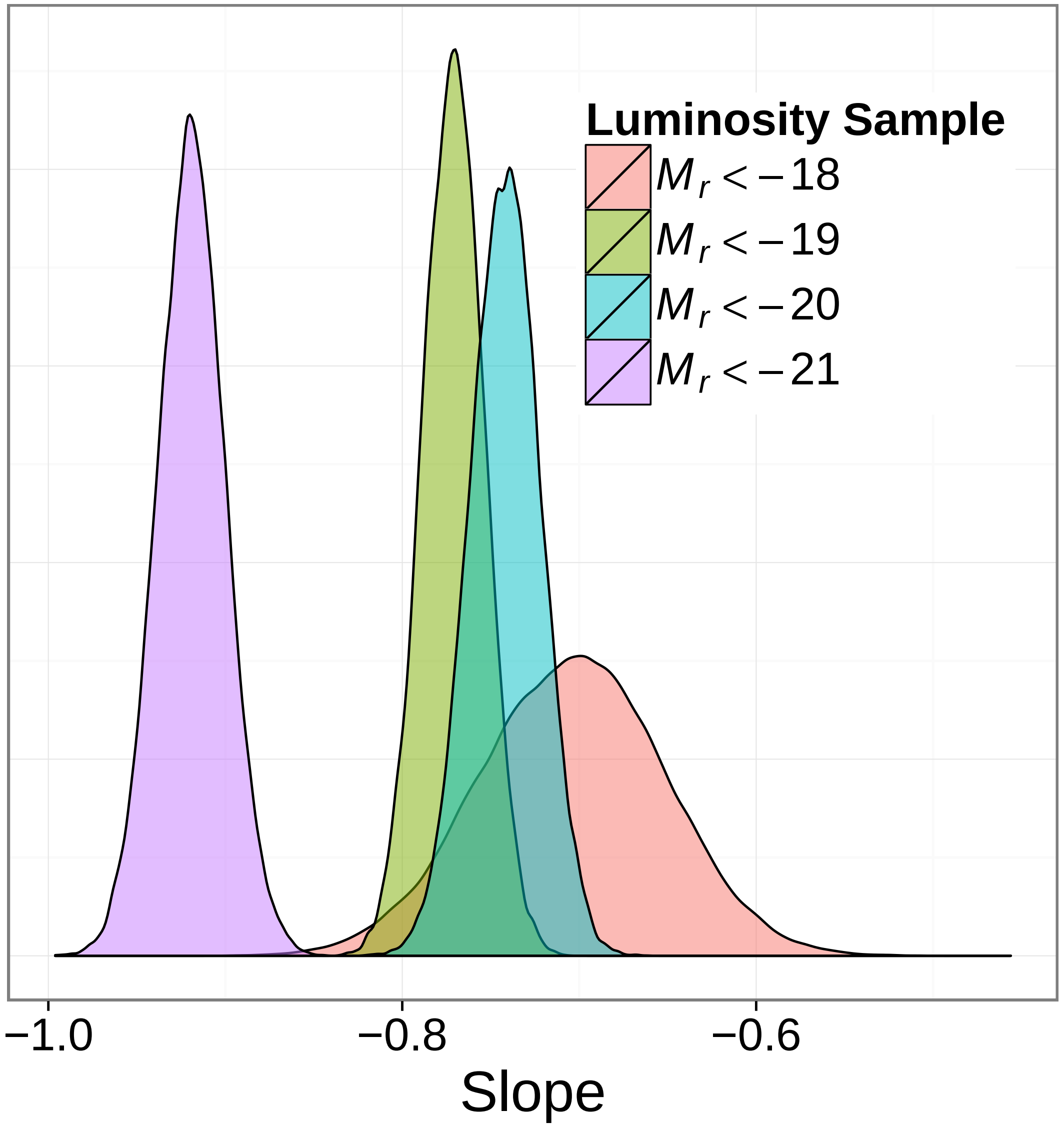

Figure 2.5 Probability density functions of slope for the power-law fits to the angular correlation functions of our four luminosity samples. The clustering of the most luminous galaxies exhibits a significantly steeper slope than that of the less luminous galaxies.

to our adopted HOD, we then construct mock galaxy samples with similar selection as our SDSS samples, and we measure $\omega(\theta)$ from the mocks in the same way as we do from the SDSS data. A few recent studies have used similar numerical modeling to fit galaxy clustering measurements (White et al., 2011; Parejko et al., 2013; Reid et al., 2014).

2.4.1 Simulations and Halo Catalogues

We obtain halo catalogues from the Large Suite of Dark Matter Simulations (LasDamas; McBride et al., 2014) project[2]. LasDamas consists of many realizations of dark matter N-body simulations for a few different box size formats. The goal of the project is to create a large number of realistic mock galaxy catalogs for several luminosity samples in the SDSS in order to assist in the modeling of galaxy clustering measurements. For each luminosity sample that we model, we use a set of LasDamas simulations with appropriate box size and mass resolution, which are listed in Table II.2. All the simulations adopt a ΛCDM cosmological model with parameter values that are consistent with WMAP5 (Dunkley et al., 2009). The particle distributions were evolved using the code `GADGET-2` (Springel, 2005). Halos were identified from the dark matter distributions using a friends-of-friends (FoF) algorithm with a linking length of 0.2 times the mean inter-particle separation. We use halo catalogues from ten independent realizations (seeded with the same primordial power spectrum, but different random phases) when we model our clustering measurements in order to address cosmic variance errors in our analysis. We discuss this further in §2.4.5.

---

[2] http://lss.phy.vanderbilt.edu/lasdamas/overview.html

Table II.2. LasDamas Simulation Properties

Name	Sample	L_{box} (Mpc/h)	N_{part}	m_{part} ($10^{10} M_\odot/h$)	r_{soft} (kpc/h)
Consuelo	-18, -19	420	1400^3	0.187	8
Esmeralda	-20	640	1250^3	0.931	15
Carmen	-21	1000	1120^3	4.938	25

Note. — For each LasDamas simulation, the table lists the absolute magnitude limit for the galaxy sample modeled, the simulation box size, the number of particles, the particle mass, and the force softening scale. Ten realizations of each box were used in the analysis.

2.4.2 HOD Formalism

We use the halo occupation distribution (HOD; e.g., Berlind & Weinberg, 2002) framework to create mock galaxy distributions from the dark matter halo catalogues. The HOD completely describes the bias between galaxies and dark matter by specifying the number and spatial positions of galaxies within halos. We first parameterize the probability distribution $P(N|M)$ that a dark matter halo of mass M contains N galaxies. We adopt the specific formalism introduced by Zheng et al. (2007b), which separates central and satellite galaxies as motivated by theoretical results (Kravtsov et al., 2004; Zheng et al., 2005). The average number of central galaxies as a function of halo mass is essentially a smooth step function that rises from zero to one,

$$\langle N_{\mathrm{cen}}(M) \rangle = \frac{1}{2}\left[1 + \mathrm{erf}\left(\frac{\log M - \log M_{\mathrm{min}}}{\sigma_{\log M}}\right)\right], \tag{2.4}$$

where M_{min} is the mass at which half the halos contain a central galaxy, and $\sigma_{\log M}$ controls the smoothness of the cutoff. The form of this function comes from the assumption that the scatter in the halo mass vs. galaxy luminosity relation has a lognormal form. The average number of satellite galaxies as a function of halo mass is essentially a power law with the same cutoff applied,

$$\langle N_{\rm sat}(M)\rangle = \langle N_{\rm cen}(M)\rangle \left(\frac{M - M_0}{M_1}\right)^{\alpha}, \tag{2.5}$$

where M_0 is the halo mass below which there are no satellite galaxies, M_1 is the halo mass that contains on average one satellite galaxy [3], and α is the slope of the power law relation. Once we have specified the mean number of centrals in a halo using Equation 2.4, we place an actual central galaxy in that halo using a probability equal to $\langle N_{\rm cen}\rangle$ (e.g., if $\langle N_{\rm cen}\rangle = 0.7$, we give the halo a 70% chance of actually containing a central galaxy). Likewise, once we have specified the mean number of satellites in a halo using Equation 2.5, we draw an actual number of satellites for that halo from a Poisson distribution.

2.4.3 Spatial Distribution of Galaxies Within Halos

Once we know how many galaxies a halo receives we have to decide where to put them. We place each central galaxy at the deepest location of its halo's gravitational potential well, which we calculate from the dark matter particles in the halo. For satellite galaxies, we adopt the methodology of Watson et al. (2010) and introduce a Generalized

---

[3] This is not exactly true unless $\langle N_{\rm cen}\rangle = 1$ and $M_1 \gg M_0$. However, for the samples we consider here, this is close to correct.

Navarro-Frenk-White (GNFW) density profile

$$\rho_{gal}(r) = \frac{\rho_s}{(\frac{r}{r_s})^{\gamma}(1+\frac{r}{r_s})^{3-\gamma}}, \tag{2.6}$$

where the slope of the profile transitions from -γ in the inner regions of the halo to -3 in the outer regions. As with a NFW profile, the transition scale depends on the concentration, but we allow the concentration of satellite galaxies to differ from that of dark matter through the parameter $f_{\rm gal}$

$$c_{\rm gal} = f_{\rm gal} \times c_{\rm DM}. \tag{2.7}$$

For the dark matter concentration we adopt the modified Bullock et al. (2001) relation from Zheng et al. (2007b)

$$c_{\rm DM} = 11\left(\frac{M}{M_{\star}}\right)^{-0.13}. \tag{2.8}$$

The GNFW profile thus has two free parameters, γ, and $f_{\rm gal}$, and we draw random radii from this profile to determine the positions of satellite galaxies within each halo. Note that values of $\gamma = f_{\rm gal} = 1$ recover an NFW profile. Drawing satellite positions from an analytic profile instead of using actual particle positions allows us to avoid force resolution errors that occur at the smallest scales we consider. Models of this type have been used to model the inner slope of the dark matter density profile (e.g., Fukushige et al., 2004; Reed et al., 2005).

Before exploring the parameter space of our flexible GNFW model, we briefly investigate whether the NFW model can reproduce our $\omega(\theta)$ measurements. We construct mock catalogs for the $M_r <$ -20 SDSS sample using halo catalogs from the Esmeralda

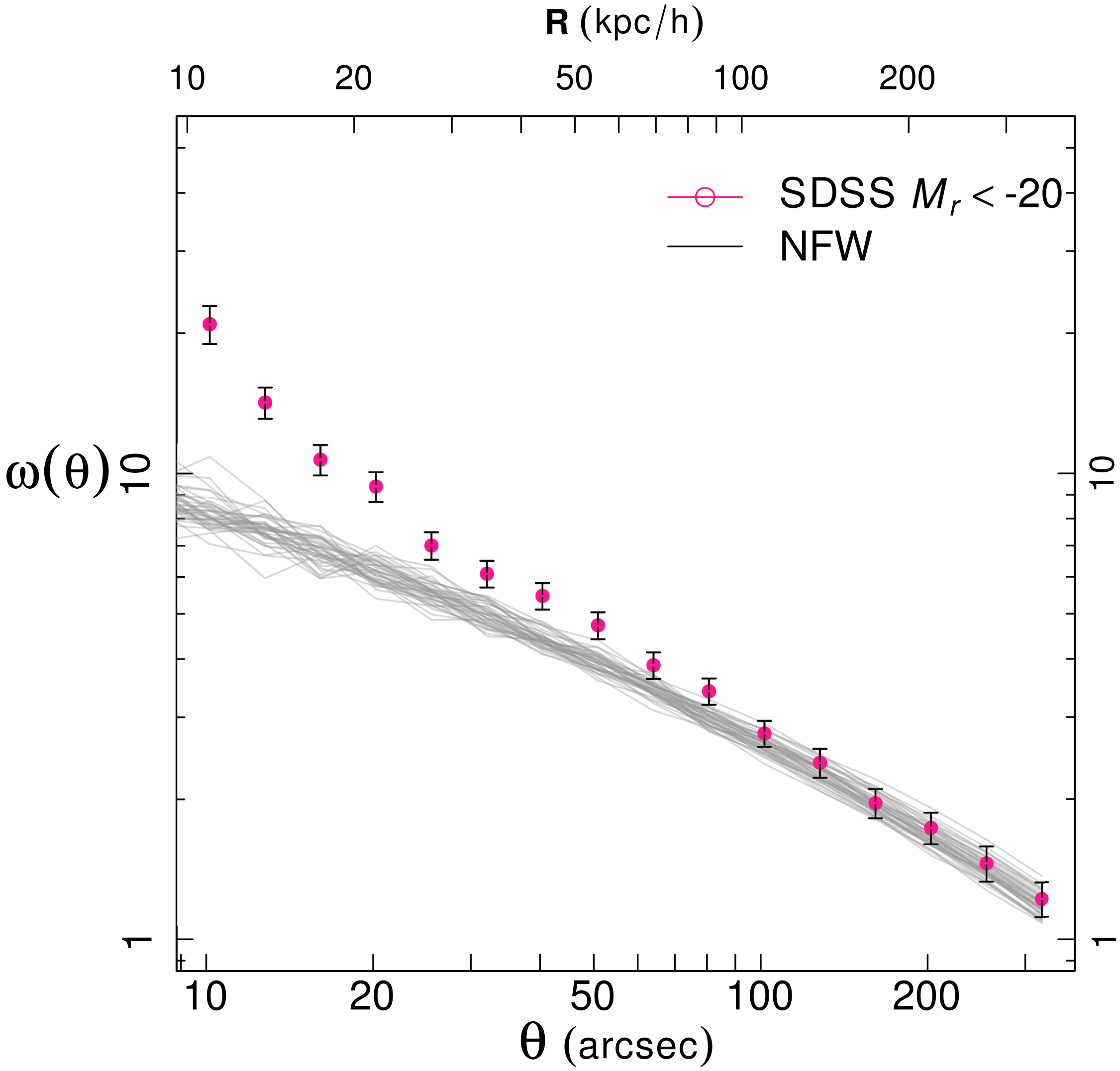

Figure 2.6 Angular correlation function for SDSS galaxies with M_r < -20 compared to a model where satellite galaxies within dark matter halos follow a NFW density profile. Points show $\omega(\theta)$ for SDSS galaxies (also shown in Figure 2.3). Curves show measurement from several independent mock galaxy catalogs that are constructed by populating dark matter halos in N-body simulations with galaxies. Each sufficiently massive halo gets a central galaxy that is placed at the halo center, while any additional satellite galaxies are distributed according to an NFW density profile. The bottom axis shows the angular scale in units of arcsec and the top axis shows the corresponding physical scale at the median redshift of the M_r < -20 sample.

simulations, as detailed above. For this test we adopt values for the HOD parameters outlined in §2.4.2 that have been found to yield a projected correlation function $w_p(r_p)$ that agrees with $M_r <$ -20 SDSS galaxies on scales larger than $100h^{-1}$kpc (McBride, private communication). We then adopt $\gamma = f_{\rm gal} = 1$ for our satellite profile, which corresponds to a NFW profile. Figure 2.6 shows $\omega(\theta)$ for several independent mock catalogs compared to our SDSS measurements. The NFW mock catalogs go from faithfully reproducing the clustering at high angular separations to under-predicting the observed clustering on the very small scales. We therefore see the same tension as Masjedi et al. (2006) and this further motivates us to explore alternative density profiles for satellite galaxies.

2.4.4 Computing $\omega(\theta)$

Once we have populated a N-body simulation with galaxies as outlined above, we place the observer at the center of the box, compute spherical coordinates, and throw out galaxies that lie outside the redshift limits of the sample we are trying to model. We do not include redshift space distortions in our analysis since they do not affect the angular clustering. Each resulting mock catalog covers the full celestial sphere and thus contains about 12 times more volume than the corresponding SDSS sample. This guarantees that the cosmic variance and shot noise in the mock catalog are much lower than in the SDSS and will therefore not significantly degrade the precision of our results.

We compute $\omega(\theta)$ using the natural estimator,

$$w(\theta) = \frac{DD}{RR} - 1. \tag{2.9}$$

Table II.3. Median Values of Halo Model Parameters from MCMC Chains

$M_r^{\rm lim}$	$\sigma_{\log M}$	$\log M_0$	$\log M_1$	α	γ	$f_{\rm gal}$	χ^2dof	$PP_{\rm NFW}$
-18	$0.48\,^{+0.36}_{-0.33}$	$8.14\,^{+2.34}_{-2.15}$	$12.87\,^{+0.16}_{-0.10}$	$1.10\,^{+0.10}_{-0.08}$	$1.80\,^{+0.46}_{-0.56}$	$1.11\,^{+0.62}_{-0.74}$	13.710	3.13
-19	$0.14\,^{+0.19}_{-0.11}$	$9.30\,^{+1.01}_{-0.89}$	$12.90\,^{+0.04}_{-0.03}$	$1.11\,^{+0.03}_{-0.03}$	$0.85\,^{+0.61}_{-0.47}$	$1.33\,^{+0.46}_{-0.64}$	16.210	1.17
-20	$0.49\,^{+0.34}_{-0.35}$	$10.10\,^{+1.59}_{-1.46}$	$13.44\,^{+0.14}_{-0.12}$	$1.33\,^{+0.15}_{-0.12}$	$1.89\,^{+0.27}_{-0.46}$	$0.51\,^{+0.80}_{-0.37}$	9.110	6.33
-21	$0.54\,^{+0.31}_{-0.35}$	$10.72\,^{+1.80}_{-1.90}$	$14.03\,^{+0.10}_{-0.09}$	$1.63\,^{+0.22}_{-0.19}$	$1.80\,^{+0.24}_{-0.40}$	$0.99\,^{+0.65}_{-0.64}$	7.810	9.32

Note. — The median halo model parameter values, along with the middle 68% interval, as measured from the MCMC chains. Also shown are the best-fit value of χ^2, as well as $P/P_{\rm NFW}$, the ratio of probability between the median value of γ and $\gamma = 1$.

We choose this estimator because it does not include a DR term, which is computationally much more expensive than DD since the random catalog is much larger than the data catalog. The RR term only needs to be computed once so when we perform our model parameter search we only have to compute DD for each set of model parameter values. This estimator is different from the one shown in Equation 3.12; however, on small scales and for a full sky geometry, these estimators yield indistinguishable results (Kerscher et al., 2000).

2.4.5 Model Fitting

We are most interested in constraining the inner slope of the satellite galaxy density profile, which is described by the γ parameter. Even though this parameter plays a primary role in setting the shape of the correlation function on the small scales we are examining, it is not easy to disentangle its effect on $\omega(\theta)$ from that of the other HOD parameters. We therefore allow all the following parameters to be free during our parameter

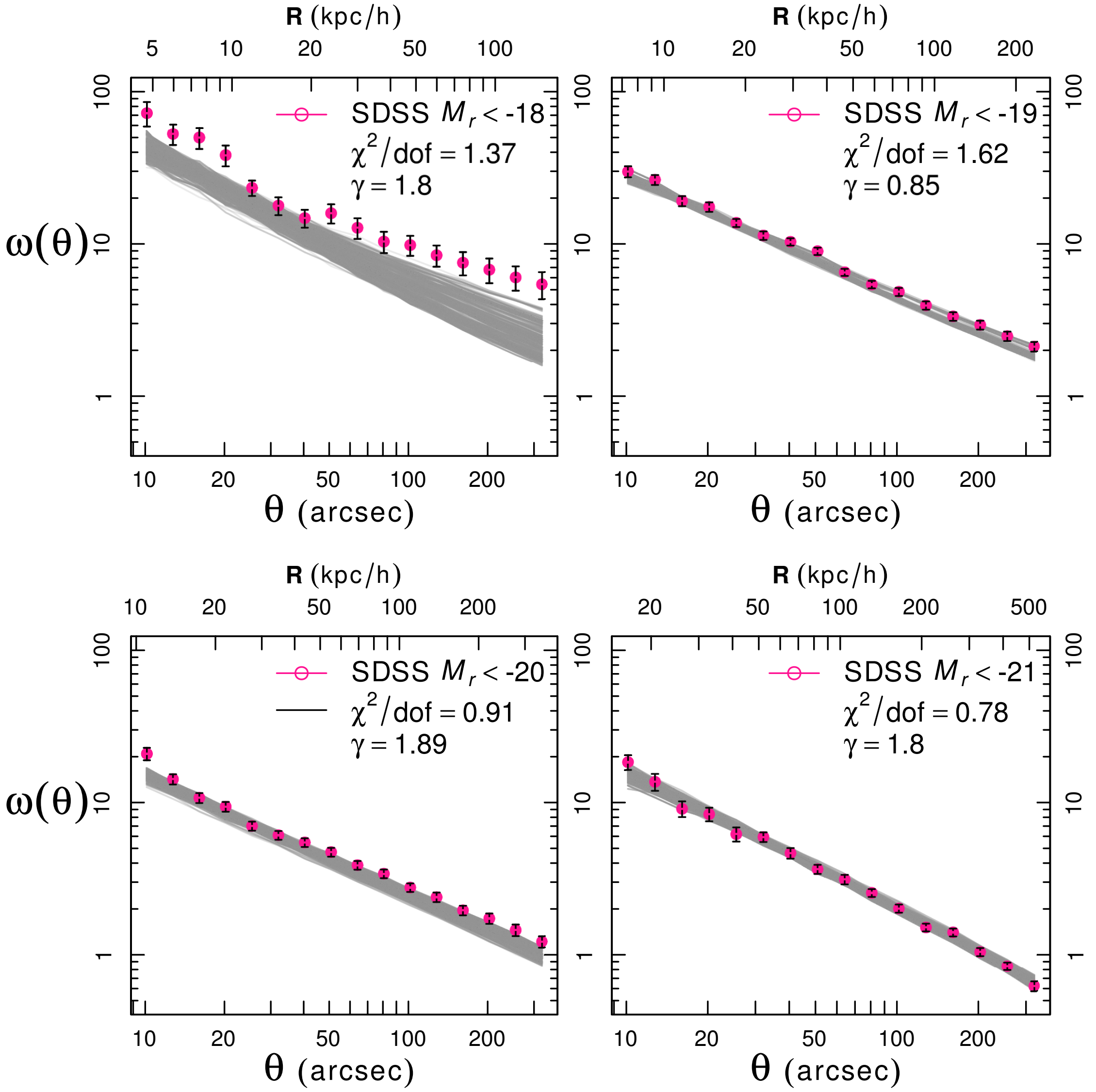

Figure 2.7 The angular correlation function of SDSS galaxies in four volume-limited samples, along with their halo model fits. The data measurements and overall layout are the same as in Fig. 2.3. The gray lines show a selection of model correlation functions that are randomly drawn from the best-fitting 68% of halo models in the MCMC chains. The median value of the satellite galaxy density profile inner slope γ, and the goodness of fit are listed in each panel.

search:

1. $\sigma_{\log M}$: Amount of scatter in the luminosity-mass relation for central galaxies.
2. M_0: Mass below which halos contain no satellite galaxies.
3. M_1: Mass at which halos contain on average one satellite galaxy.
4. α: Slope of the power-law relation between the mean number of satellite galaxies and halo mass.
5. γ: Inner slope of the number density profile for satellite galaxies within their halo.
6. $f_{\rm gal}$: Concentration of satellite galaxies with respect to the dark matter concentration.

For each combination of the above six free parameters, we set $M_{\rm min}$ to the value that recovers a total galaxy number density equal to that observed by the SDSS.

We perform a parameter search using the MCMC `emcee` code and algorithm described by Foreman-Mackey et al. (2013). The algorithm is based on the affine invariant sampling algorithm proposed by Goodman & Weare (2010). It is fast, efficient, and easily parallelized. We use 500 "walkers" to explore the parameter space in parallel. The basic procedure we follow each time we test a new location in our six dimensional parameter space is as follows. We first select a random halo catalog from among ten independent N-body realizations. This builds cosmic variance errors in our theoretical calculations directly into the modeling. We then use the halo catalog to determine the value of $M_{\rm min}$ required to create a galaxy catalog with the observed mean number density. For each halo in the catalog, we use Equation 2.4 to determine whether the halo contains a central galaxy, and Equation 2.5 together with a Poisson distribution to choose the number of satellite galaxies. We then randomly draw satellite positions from the density profile shown in Equations 2.6 and 2.7. We make an all-sky galaxy mock catalog and compute $\omega(\theta)$ as described in §2.4.4. Finally, we estimate χ^2 for the parameter combination using the jackknife covariance matrix described in §2.3.1. All six of our free parameters are given physically motived flat priors. In particular, the satellite profile parameters γ and $f_{\rm gal}$ are allowed to vary over a broad range that includes the NFW profile. For all four luminosity samples, we find that we need approximately one million parameter combinations in order to get MCMC chains that converge.

2.4.6 Halo Modeling Results

Figure 2.7 shows the resulting $\omega(\theta)$ of our halo modeling in each luminosity bin. SDSS measurements and the overall figure layout are the same as in Figure 2.3, while the gray lines show a random sampling of halo model correlation functions drawn from the best-fitting 68% of models in the MCMC chains. The lines thus illustrate the spread in $\omega(\theta)$ for models that are consistent with the SDSS data. Each panel shows the median value of γ, as well as the value of χ^2dof for the best fit model. With 16 angular bins and 6 free parameters, the number of degrees of freedom is equal to 10. Our halo model produces good fits to the angular clustering of all four luminosity samples. There is a slight tension in the case of the $M_r <$ -19 sample, but the difference between the model and the SDSS data is not statistically significant (the p-value for this sample is 0.094). We list the χ^2dof values for all four samples in Table II.3. Though it looks like the model is not a good fit to the data in the case of the M_r <-18 sample, we note that it is very misleading to perform χ by eye because neighboring bins in $\omega(\theta)$ are extremely correlated with each other, as shown in Figure 3.18.

The main focus of this analysis is the inner slope γ of the satellite galaxy density profile within halos. Figure 2.8 shows the marginal distribution of γ values from the MCMC chains as a function of galaxy luminosity. Specifically, the middle line in each box shows the median value of γ, the shaded box shows the 68% confidence interval, and the whiskers show the 95% confidence interval for γ. The two most luminous samples, $M_r <$ -20 and -21, both prefer fairly steep density profiles and are inconsistent with the NFW profile at approximately the 2σ level. Specifically, the fraction of points in the MCMC chain that have $\gamma > 1$ is 97% and 96% for the $M_r <$ -21 and -20 samples, respectively. The less luminous $M_r <$ -19 sample prefers less steep profiles and is perfectly consistent with NFW. The lowest luminosity $M_r <$ -18 sample seems to favor steep profiles, but it has a broad γ distribution and is not significantly inconsistent with NFW. The poor constraints for the least luminous sample are due to the small size of this galaxy sample.

The constraints on γ are consistent with those from Watson et al. (2012), denoted by the asterisks in Figure 2.8, at the 1σ level for the $M_r <$ -19, -20, and -21 samples. The $M_r <$-18 sample is consistent at approximately the 2σ level. The constraints are somewhat weaker in this paper because we reduce our sample size by only considering galaxies in plate overlap regions. Additionally, fiber collisions are a source of systematic error that was not included explicitly in our model. Though we do not expect this error to be large, it might be affecting the $M_r <$-18 measurement in a subtle way that is contributing to the tension with Watson et al. (2012). We list the median and 68% confidence intervals

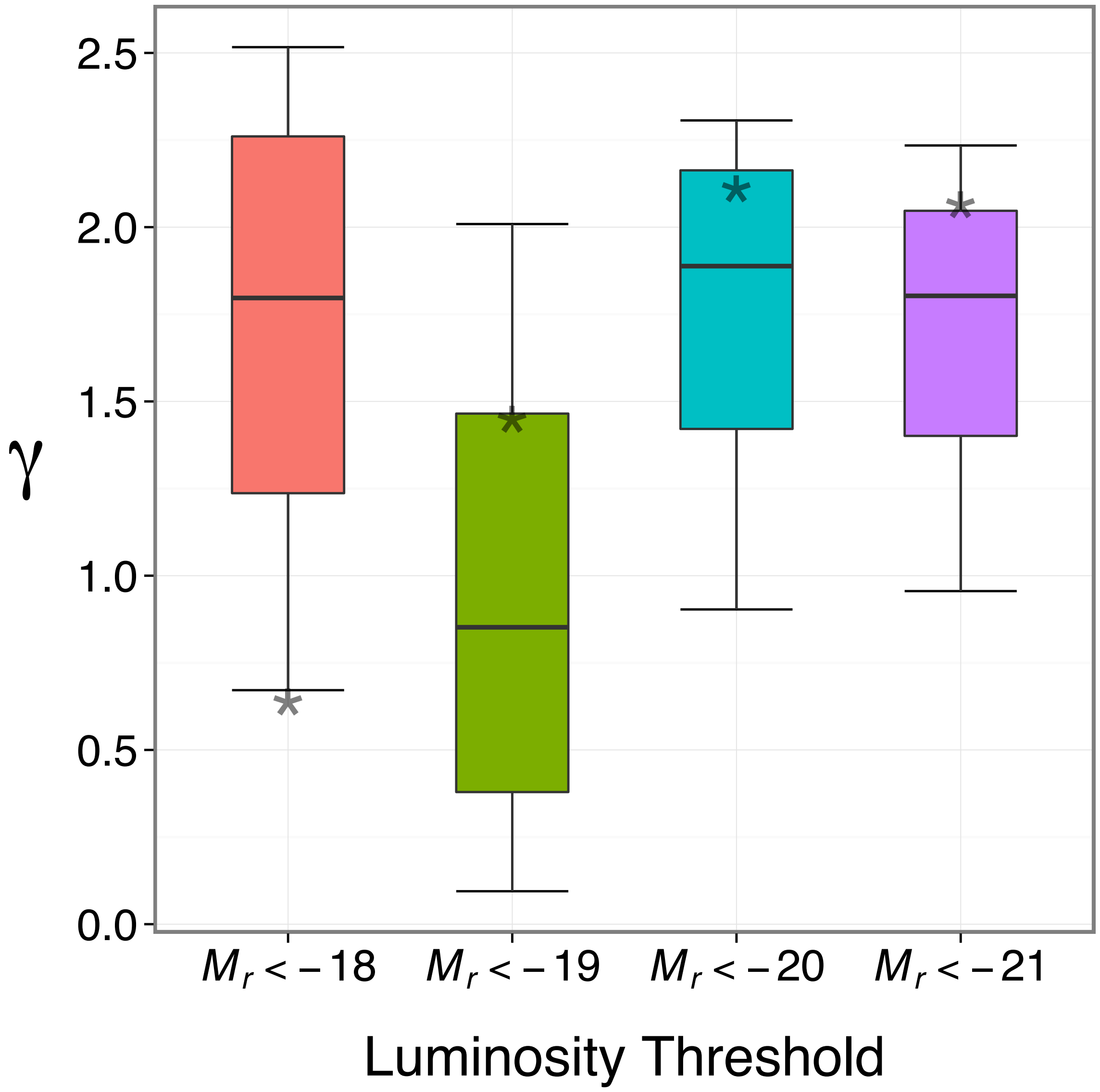

Figure 2.8 Luminosity dependence of the satellite galaxy density profile inner slope. Each box and associated whiskers corresponds to a particular luminosity sample, as shown on the x-axis. The middle line in each box shows the median value of γ from the MCMC chain, the vertical range of the box shows the middle 68% of γ values, and the whiskers extend to the middle 95% of values. For the two high luminosity samples, a value of $\gamma = 1$, corresponding to the NFW density profile, is inconsistent with the SDSS data at approximately the 2σ level. Lower luminosity galaxies do not show this tension. The median values of γ from Watson et al. (2012) are marked as asterisks.

for all four samples in Table II.3. We note that a cursory examination of Figure 2.6 may lead to the impression that the NFW model is ruled out at higher significance than 2σ. However the models shown in Figure 2.6 were fit to $w_p(r_p)$ on larger scales, not $\omega(\theta)$ on very small scales. Moreover, as stated previously, it is misleading to perform χ by eye due to the high degree of correlation between bins.

We now quantify the extent to which the γ values preferred by our clustering measurements are more likely than the $\gamma = 1$ NFW case. We calculate the number of accepted parameter combinations in the MCMC chain that have values of γ in a bin of width ± 0.1 that is centered around the median value of γ. We then do the same for a bin centered around $\gamma = 1$ and take the ratio of these two numbers, which we call $PP_{\rm NFW}$. This yields the relative likelihood of the two γ values given the measured correlation function. We find that the steep slopes measured for the $M_r <$ -20 and -21 samples are 6.3 and 9.3 times more likely than $\gamma = 1$, while the slopes measured for the less luminous samples are only 3.1 and 1.2 times more likely than $\gamma = 1$. We list these values in Table II.3.

Figure 2.9 shows the final probability distributions of the HOD parameters that determine the number of galaxies as a function of halo mass. The parameter controlling the shape of the low mass cutoff for central galaxies $\sigma_{\log M}$ is poorly constrained by the angular correlation function on the small scales we consider in this study. This is due to the fact that on these scales most galaxy pairs come from within a single halo and so the low mass regime where a halo either has zero or one galaxy is not very important. The distributions of $\sigma_{\log M}$ are bound by values of 0 and 1 as this was the prior that we adopted for this parameter. The parameter M_0 is also very poorly constrained because it cuts off the satellite occupation number on a sufficiently small mass scale where the expected

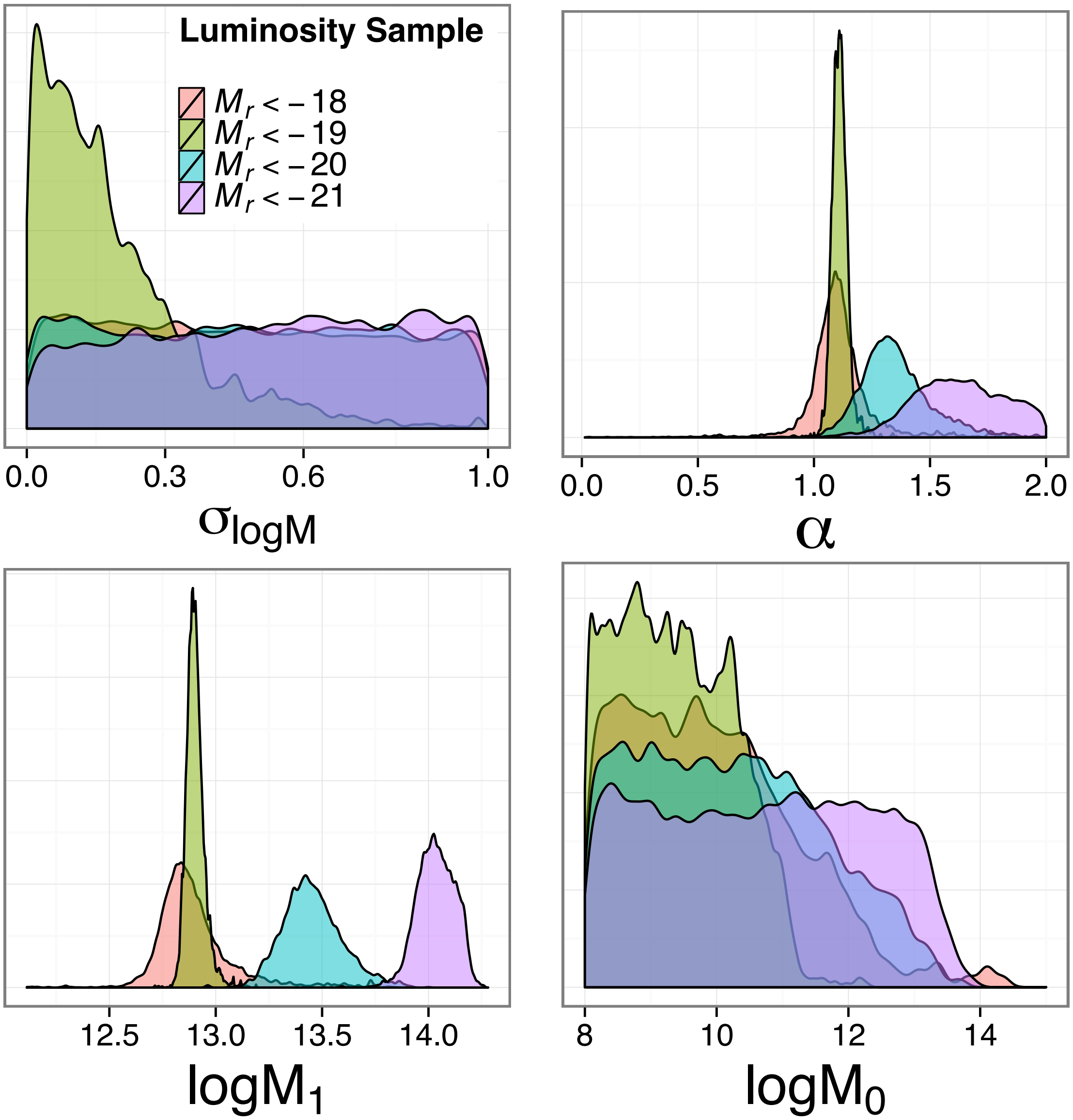

Figure 2.9 The probability distributions for the four HOD parameters that control the number of galaxies a dark matter halo of a given mass will receive. Each panel shows a different parameter, while the four distributions in each panel show results for our four luminosity samples. The parameters $\sigma_{\log M}$ and M_0 are largely unconstrained by our measurements and the bounds of their distributions reflect their prior flat distributions. $\sigma_{\log M}$ was restricted to the range $0 - 1$, while M_0 was restricted to values greater than 10^8 $\mathrm{M}_{\odot}$.

number of satellites is already much less than one. This result is consistent with other studies (e.g., Zehavi et al., 2011). The lower bound for M_0 at 10^8 $\mathrm{M}_\odot$ is once again due to the prior we adopted for this parameter. Though these priors are physically motivated, we have checked that relaxing them does not significantly change our conclusions about the slope γ. The two parameters that control the number of satellites a halo receives, α and M_1, are much better constrained. More luminous samples have a higher mass M_1 at which they typically contain a single satellite and they have a steeper relation between the number of satellites and halo mass. These trends are consistent with Zehavi et al. (2011) though the values we find are somewhat higher for both M_1 and α. This could be due to the different information content of $\omega(\theta)$ compared to $w_p(r_p)$, or it could be due the difference between our numerical modeling compared to the analytic halo model used in Zehavi et al. (2011), or it could be due to the extra freedom we have added to the HOD model by using a GNFW profile. We note that for the purpose of this paper we are mainly interested in the γ parameter and these parameters therefore act as nuisance parameters. We list the median and 68% confidence intervals for all model parameters in Table II.3. We also note that the galaxy concentration parameter f_{gal} is very poorly constrained.

2.4.7 Power Law vs. Halo Model

Both the power-law fits and the halo model have suggested that luminous satellite galaxies have a steeper density profile than lower luminosity galaxies. However, the two analyses also show some differences. In the case of the power law slope, $M_r <$ -21 galaxies have a steeper $\omega(\theta)$ than less luminous samples and this difference is highly significant. On the other hand, when using the halo model, $M_r <$ -20 galaxies have consistent values of γ with $M_r <$ -21 galaxies and discrepancies with less luminous galaxies are less significant. We now take a closer look at these results to determine if they are consistent with each other.

We first make a naive comparison between the power law slope and the value of γ. If we assume that the galaxy correlation function $\xi(r)$ is dominated by central-satellite pairs in halos of mass $\sim M_1$ that only have a single satellite galaxy, then its slope should essentially be the same as the slope of the satellite galaxy density profile within these halos. This is not a bad assumption since more massive halos that contain more than one satellite galaxy are relatively rare. Therefore, the slope of $\xi(r)$ should be equal to -γ. If we further assume that $\xi(r)$ is a perfect power law, then the angular correlation function $\omega(\theta)$ should also be a power law with a slope that is shallower by +1 (Totsuji & Kihara, 1969). Therefore, the slope of $\omega(\theta)$ should be $1 - \gamma$, or conversely, γ should be $1 - slope$. Using this simple transformation, we can check whether the power law slopes that we found in §2.3.3 are consistent with the γ distributions shown in Figure 2.8. The values of γ inferred from the power law slopes of $\omega(\theta)$ are 1.7, 1.77, 1.76, and 1.92 for the $M_r <$ -18, -19, -20, and -21 samples, respectively. These values are perfectly consistent with the broad distributions shown in Figure 2.8.

We next perform a more sophisticated test to assess the relationship between the slope of the power law model and γ. We fit a power law to measurements of $\omega(\theta)$ made from mock galaxy catalogues produced using the best-fit HOD model. We then compare the recovered power law slopes and compare them to the input values of γ. We find that for three out of four samples, $slope \sim 1 - \gamma$, as expected. For the $M_r <$ -19 sample, the power-law slope is somewhat steeper than $1-\gamma$. We conclude from these explorations that our results from fitting power laws and halo models are consistent with each other, and that constraints on γ are weaker than on the power law slope due to marginalization over all the other HOD parameters.

2.5 Summary & Discussion

The goal of this paper is to probe the radial number density profile of satellite galaxies within dark matter halos using SDSS clustering measurements. We wish to test the results of Watson et al. (2012), who found that luminous satellite galaxies (SDSS, $M_r <$ -20) have significantly steeper density profiles than the NFW profile on scales smaller than $\lesssim 40h^{-1}$kpc. Unfortunately, clustering measurements on these scales are strongly affected by fiber collision incompleteness, making it important to verify this result with different measurements and modeling methodology. We used the angular correlation function $\omega(\theta)$ as our clustering statistic of choice because it is fairly insensitive to fiber collision errors. Moreover, we restricted our measurements to plate overlap regions on the sky, where most fiber collided galaxy pairs are recovered because of repeat observations. We measured $\omega(\theta)$ on four volume-limited samples with absolute r-band limits of $M_r <$ -18, -19, -20, and -21, on scales in the range $7-320''$. These angular scales correspond to physical scales that are within the virial radii of dark matter halos expected to host these galaxies, even at the far redshift of each sample. Our measurements thus directly probe the spatial distribution of galaxy pairs within halos.

Motivated by the approximately power-law shape of our correlation function measurements, we first fit a power law function to $\omega(\theta)$ in order to quantify its slope. We found that the most luminous galaxies ($M_r <$ -21) have a significantly steeper correlation function than the lower luminosity samples. We then used the more physically motivated halo model to determine what constraints our $\omega(\theta)$ measurements place on the density distribution of satellite galaxies within halos. We used a fully numerical modeling procedure that populates dark matter halos in N-body simulations with galaxies, creates mock SDSS samples, and computes $\omega(\theta)$ the same way as it is done in the real galaxy data. This method is computationally expensive, but it minimizes systematic errors in the modeling. The key ingredient in our halo model is a generalized density profile for satellite galaxies, whose inner slope is a free parameter. After marginalizing over other parameters in our halo model, we found that the two more luminous galaxy samples ($M_r <$ -20 and -21) prefer a satellite density profile that is substantially steeper than the NFW profile. The NFW profile is discrepant at the 2σ level for these galaxy samples. We found that the lower luminosity samples do not constrain the satellite inner profile slope as well and they are consistent with NFW.

Our results are qualitatively consistent with those of Watson et al. (2012) who also found that satellite galaxies more luminous than $M_r <$ -20 have steeper density profiles than NFW. Our results are also quantitatively consistent, as our marginal distributions of the inner profile slope overlap nicely. The main differences between our two studies are that (1) Watson et al. (2012) found somewhat shallower inner profiles than we did for the least luminous ($M_r <$ -18) galaxies, and (2) their constraints on the inner profile slope of the most luminous galaxies are tighter than ours. These differences allowed them to detect a significant luminosity trend in the spatial distribution of satellite galaxies, while we cannot do the same with confidence. The loss of statistical significance in our study is mainly due to the lower information content of $\omega(\theta)$ compared to $w_p(r_p)$, as well as to the fact that we reduce our sample size by only considering galaxies in plate overlap regions. On the other hand, our constraints are less likely to be affected by errors in the halo modeling. In addition, the two studies are affected by fiber collision incompleteness in different ways. Overall, the agreement between the two studies despite the different measurement and modeling methods lends credibility to the main conclusion that the spatial distribution of luminous satellite galaxies is steeper than that of the underlying dark matter.

Before making claims about how well satellite galaxies trace the dark matter distribution, we need to consider whether the NFW model is itself an accurate representation of the density profile of dark matter. Though the NFW profile has been shown to provide an imperfect description of the structure of dark matter halos in collisionless N-body simulations (e.g., Navarro et al., 2004; Merritt et al., 2005; Gao et al., 2008; Navarro et al., 2010; Ludlow et al., 2013), the departures shown by these studies are not large and the NFW model remains consistent with simulation results at the $\sim 10-20\%$ level (Benson, 2010). However, it is far less safe an assumption that the density profiles of halos in collisionless simulations represent reality given that they completely ignore the effects of baryons. This is especially true for the very small scales we consider in this paper, since baryons dominate the mass budget at the centers of halos. Some theoretical studies argue that the condensation of baryons at the centers of halos should steepen the dark matter density profile (e.g., Gnedin et al., 2011), while others argue the opposite (e.g., Del Popolo, 2012). Observational studies using weak lensing measurements have found that the density profiles at of clusters are either consistent with or shallower than NFW (e.g., Mandelbaum et al., 2008; Newman et al., 2013), though the measurements are noisy on the small scales we care about here.

It is not necessarily surprising that satellite galaxies do not trace the underlying mass distribution. Galaxies are extended massive objects and they should thus experience dynamical effects such as dynamical friction and tidal stripping of mass and stars, which do not affect dark matter particles. These mechanisms could act to steepen the density profile of satellites. Moreover, this could be a luminosity dependent process. To study these effects, it would be illuminating to compare our satellite profile results with the distribution of dark matter subhalos within host halos, since satellite galaxies presumably occupy these systems. However, this comparison will have to wait for simulations of sufficient volume and particle mass resolution to be able to accurately measure the distribution of massive subhalos at scales of only a few tens of kpc from the center of host halos. It would be even better to compare our results with predictions from hydrodynamic simulations that include baryonic processes such as gas cooling and feedback, which can affect the density profiles of halos. Our measurements can help to constrain these processes. Simulations that have both sufficient volume and resolution to make such predictions are now becoming possible. For example, the Illustris simulation has already enabled a prediction of the density profile of luminous satellite galaxies on small scales (Genel et al., 2014). Though this result is a bit too noisy to be tested against our measurements, the next generation of simulations should be more than adequate for making this comparison.

Chapter III

EXTENDING BEYOND THE LOCAL: VERY SMALL SCALE CLUSTERING IN THE DISTANT UNIVERSE

3.1 Introduction

Measuring galaxy clustering is a powerful tool for understanding the fundamental physics of the universe. The current galaxy-halo model, in which galaxies are biased tracers of the underlying dark matter, exists because it has been so successful at predicting galaxy clustering using statistics like the two-point galaxy correlation function (2PCF).

On the largest scales of the universe, the 2PCF has been used to constrain cosmology (Eisenstein et al., 2005). On the smallest, it probes possible deviations of the galaxy distribution from that of dark matter (Masjedi et al., 2006; Watson et al., 2010). Much of our physical interpretation of the 2PCF comes from modeling clustering using cosmological N-body simulations and the Halo Occupation Distribution (HOD; Peacock & Smith, 2000; Seljak, 2000; Berlind & Weinberg, 2002; Zheng et al., 2005). The HOD is a prescription for relating the mass of the host halo to the number of galaxies residing within. There are many assumptions implicit here: from the cosmological parameters of the simulation to the galaxy distribution in one dark matter halo.

Our current theories of galaxy clustering assume that the density distribution of dark matter drives the distribution of galaxies in the universe. However, what if there are scales at which galaxy clustering is determined not by just the dark matter, but by some intrinsic property of the galaxy? The answer to these questions can be found by studying galaxy clustering within an individual dark matter halo. Measuring and modeling galaxy clustering on these scales less than about one Megaparsec requires careful analysis of large data sets, in particular the Sloan Digital Sky Survey (SDSS; York et al., 2000).

The galaxy density profile gal is the radial distribution of galaxies in one halo. In principal, ρ_{gal} should be the product of the dynamical or baryonic processes a galaxy undergoes over its formation and merger history. These histories correlate with the galaxy environment and luminosity of a galaxy (Hogg et al., 2003). Because it is difficult to simulate a cosmological volume that includes detailed enough baryonic physics, ρ_{gal} is a choice. Oftentimes this choice is made unknowingly: galaxy positions are drawn directly from the dark matter distribution. Since the dark matter follows a Navarro-Frenk-White (NFW; Navarro et al., 1997) density profile, this technique generates ρ_{gal} with an NFW profile as well. In the most popular semi-analytic recipes for galaxy clustering, NFW is imposed for ρ_{gal}. This assumption has far reaching consequences, from affecting our understanding of galaxy evolution to how we derive fundamental cosmological parameters from galaxy clustering. Indeed, Watson et al. (2012) found that the deviation of ρ_{gal} from NFW was a function of luminosity. Improving on that work, I have introduced new methods for studying ρ_{gal} as a function of both redshift and luminosity.

In Chapter 2, I have shown that significant deviations from the standard Navarro-Frenk-White density profile are seen in the distribution of satellite galaxies in the local universe. This deviation could even be a function of the luminosity of the galaxy. Instead of looking at how different classes of galaxies occupy their halos at low redshifts, we now investigate how the same mass galaxy changes its radial distribution as a function of time. In principle, this could tell us about the merger history and evolution of a specific type of galaxy. In this case, we focus on using the SDSS-III BOSS CMass sample– specifically designed to be constant mass over redshifts $0.45 < z < 0.6$.

With the new data from the Baryonic Oscillation Spectroscopic Survey (BOSS Dawson et al., 2013), one of the four science projects from the SDSS-III survey (Eisenstein et al., 2011), we can now study galaxies extending back a few billion years at the same statistical level as galaxies in our own backyard. BOSS has measured the spectra of 1.5 million galaxies extending out to a redshift of 0.7- corresponding to when our universe was slightly more than half the size it is today. This chapter is organized as follows. The data sample selection is described in Section 1. In Section 2, we describe the statistical method in which we study the very small scale clustering of BOSS galaxies. In Section 3 we compare our findings to previous results. Section 4 is a discussion of implications and physical interpretations of our findings. The chapter concludes in Section 6. Throughout this paper we assume a spatially flat Λ cold dark matter (ΛCDM) cosmology with $\Omega_m = 0.274$.

3.2 Data Sample

Accurate measurements of galaxy clustering at very small scales requires a large galaxy survey. Data were taken from the five-band *ugriz* imaging from the Sloan Digital Sky Survey (York et al., 2000). The SDSS survey uses a dedicated 2.5m telescope (Gunn et al., 2006) with a wide-field mosaic CCD (Gunn et al., 1998) operating in drift scan mode (Fukugita et al., 1996). Additional photometry for SDSS-III was taken in the southern galactic cap and all imaging data, including the previous SDSS-I,II data, was re-reduced for SDSS-III. The results of this re-reduction are in Aihara et al. (2011).

Data in the sample used in this analysis covers 6500 sq. deg. of the sky. The footprint

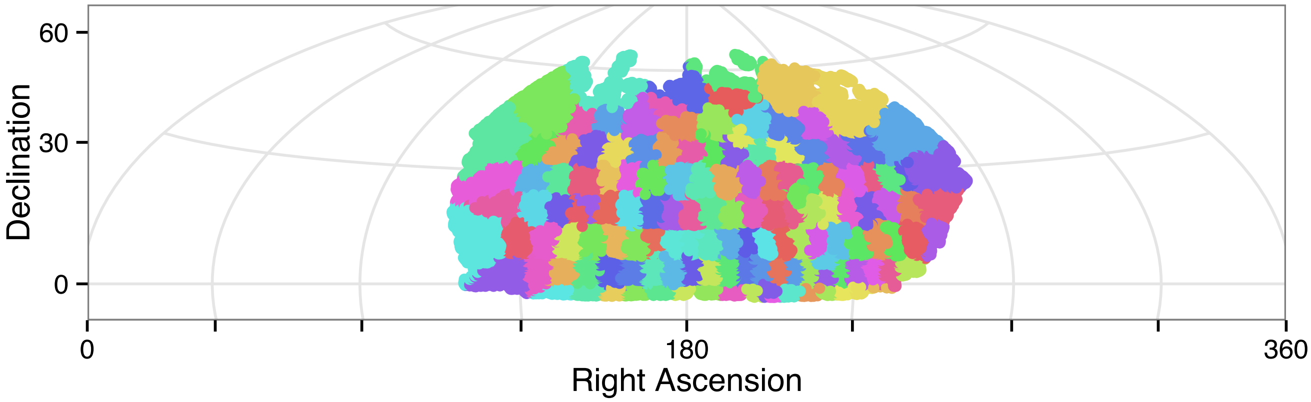

Figure 3.1 Footprint on the sky (Hammer Projection) of the Sloan Digital Sky Survey Data Release 11 BOSS CMASS galaxy sample. The individual jackknife regions are shown in color. Jackknife regions were determined using the underlying sector distribution of targeted galaxies.

is shown in Figure 3.1. Below we will detail the color and magnitude cuts that were used to target the high redshift sample of galaxies used in this analysis.

3.2.1 Target Selection

The SDSS-II targeted Luminous Red Galaxies (LRGs), highly biased traces of the dark matter distribution with very little star formation, in order to better constrain cosmological parameters using the BAO. This initial target selection for (LRGs Eisenstein et al., 2001) was incomplete in mass at the bright end (Tojeiro et al., 2012). SDSS-III sought to construct a mass limited sample with well understood mass completeness limits. In order to create a Constant Mass (CMASS Reid et al., 2015) sample, the color-magnitude cuts were extended to include bluer star-forming galaxies thought to be the progenitors of the lower redshift LRGs.

The selection criterion for the CMASS sample was informed by the galaxy evolution and stellar population synthesis models of Maraston et al. (2009). These models allow us to target a more complete high mass galaxy population over a redshift range of $0.1 < z < 0.7$. The targeting uses four different definitions of magnitude for the central g, r, z magnitude bands in order to target a clean sample, without contamination from stars or low redshift galaxies. All magnitudes are corrected for galactic dust extinction (Schlegel et al., 1998) and are in the observed frame.

The four different definitions of magnitude used to select BOSS CMASS galaxies are as follows:

1. SDSS calibrated model magnitudes, subscript *mod*, from Padmanabhan et al. (2008), denoted with the subscript *mod*. These magnitudes are derived from either an exponential or a DeVaucouleurs light profile fit to the r band.
2. *cmod* magnitudes. The *cmod* are defined first as a flux f:

   $$f_{cmod} = (1 - P)f_{exp} + Pf_{deV} \tag{3.1}$$

   where P is the probability of an exponential or deVaucouleaur profile.
3. Point Spread Function, psf, magnitudes.
4. Fiber Magnitudes in a 2 arcsecond aperture, $fib2$.

For the CMASS sample, magnitudes are all calculated using *cmod* magnitudes while colors are all defined using *mod* magnitudes. The SDSS-II LRG sample included another definition of magnitude, *Petrosian* magnitudes, that are no longer used for galaxy sample selection in SDSS-III.

The following are the selection criteria for BOSS CMASS Galaxies, from Reid et al. (2015) and Tojeiro et al. (2012). First we define an ancillary color, $d_\perp$ as

$$d_\perp = r_{mod} - i_{mod} - (g_{mod} - r_{mod})/8.0. \tag{3.2}$$

This color separates lower redshift galaxies from the CMASS sample by cutting on

$$d_\perp > 0.55. \tag{3.3}$$

An additional sliding color cut

$$i_{cmod} < 19.86 + 1.6(d_\perp - 0.8) \tag{3.4}$$

which selects the brightest and most massive objects at each redshift based on the passively evolving galaxy models of Maraston et al. (2009). Maraston et al. (2013) showed that this cut creates a more complete galaxy sample at $z > 0.6$. Additionally, these color cuts include bluer galaxies making the CMASS sample more complete at the higher mass end(Tojeiro et al., 2012).

The following are all magnitude cuts. This cut,

$$17.5 < i_{cmod} < 19.9, \tag{3.5}$$

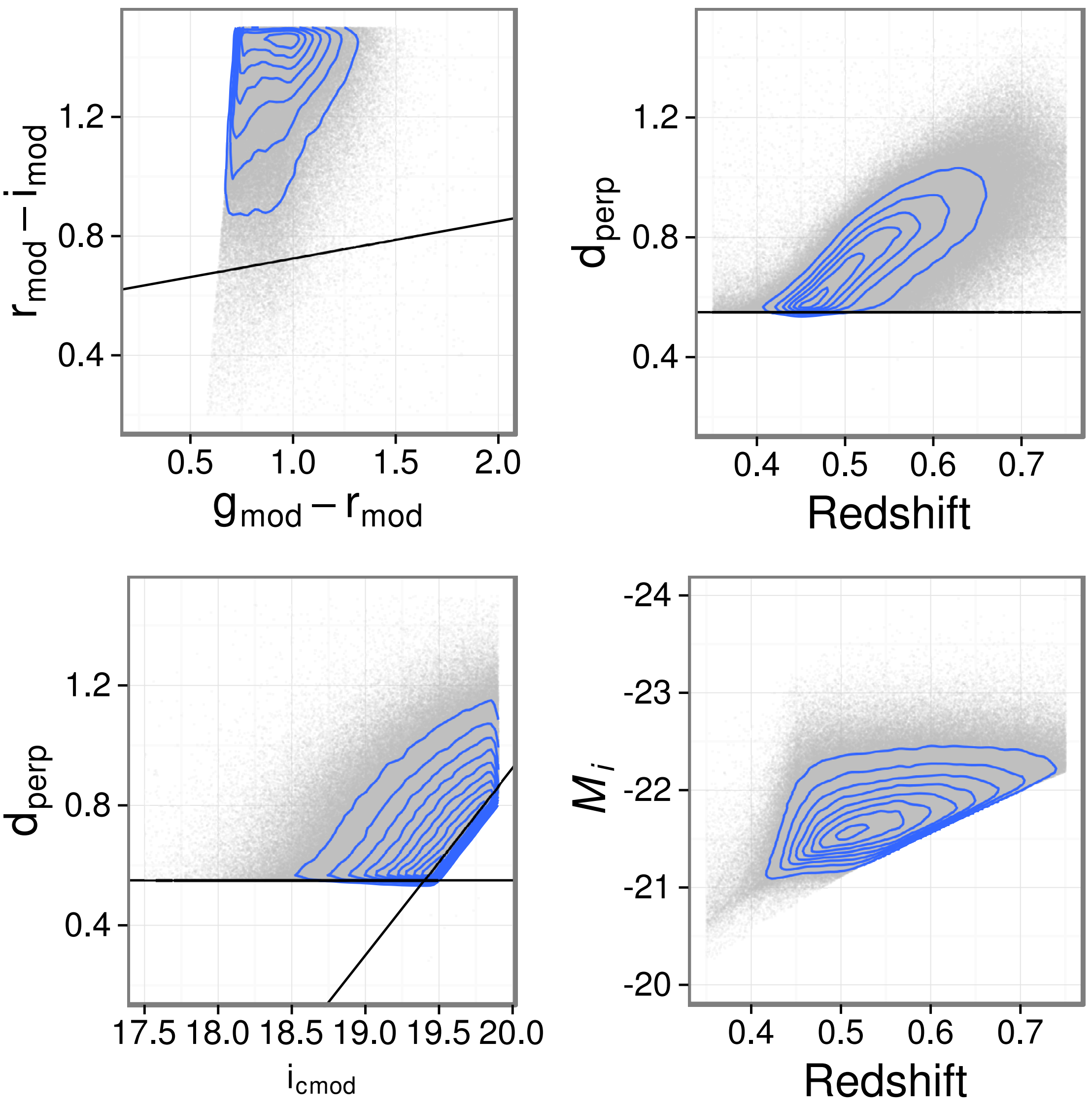

Figure 3.2 The color and magnitude cuts that were used in this paper. Apparent magnitude cuts were done using ‘cmodel’ magnitudes whereas color cuts were done using model magnitudes. The cuts were designed to select the most massive galaxies over a large range of redshifts. There is no absolute magnitude cut, however we show the trend with redshift in the lower right panel.

selects out low redshift interlopers on the bright limit and ensures a high redshift success rate on the faint limits. To increase the redshift success rate, which is based on the observational limits of the telescope, we cut on

$$i_{fib2} < 21.5. \tag{3.6}$$

An additional cut to protect against some outliers is

$$r_{mod} - i_{mod} < 2 \tag{3.7}$$

and

$$i_{rdev} < 20.0 \tag{3.8}$$

where i_{rdev} is the effective radius in the fit to the deVaucouleurs profile for the $i - band$ magnitude. The population of CMASS galaxies is separated from stars using a comparison of model magnitudes and psf magnitudes

$$i_{psf} - i_{mod} > 0.2 + 0.2 * (20.0 - i_{mod}) \tag{3.9}$$

$$z_{psf} - z_{mod} > 9.125 - 0.46 z_{mod}. \tag{3.10}$$

Combined these cuts create a clean and complete sample of galaxies with masses up to 10^{12} $M_{\odot}$ (Maraston et al., 2013). We present these cuts in Figure 3.2. Throughout this chapter, we will refer to all the objects that have passed these color and magnitude cuts as the 'Imaging' Sample.

3.2.2 Spectroscopic Sample

Once the Imaging Sample has been defined, the process of observing galaxy spectra to measure redshifts and absolute magnitudes can begin. The SDSS-III spectrograph (Smee et al., 2013) is an improvement upon on the original SDSS spectrograph (Gunn et al., 1998), allowing the SDSS to observe 1000 objects at a time using a duel channel fiber-fed multi-object spectrograph. The 2” fibers are plugged manual into aluminum plates which subtend 3° on the sky (Smee et al., 2013). In SDSS-III these fiber cannot be placed closer than 62”; targeted objects closer than this cannot be observed on a single plate. This causes a problem called fiber collisions, which in SDSS-II affected up to 10% of targets (Blanton et al., 2003b) and which will be revisited later on. Because of the overlap of the tiles, about one third of fiber collision objects can be recovered. These plates form the bases of how objects are optimally tiled for spectra. Below we summarize the steps taken to create a large scale structure galaxy catalogue.

3.2.2.1 Tiling

The tiling algorithm (Blanton et al., 2003b) is the process of arranging these plates on the sky such that the maximal number of targets are observed with the minimal number of plates. The tiling algorithm adapts to the local angular density of targets to maximize the number of objects receiving spectroscopic fibers. Tiling was done piecemeal over the course of the survey, in ‘chunks’. Each chunk is a spatially contiguous set of overlapping tiles that the targeting algorithm has arranged. These chunks comprise the SDSS footprint, shown in Figure 3.1.

This is going to be a summary of Bolton et al. (2012) and Dawson et al. (2013).

3.2.2.2 Fitting Redshift

The galaxy is observed for four to five 15 minute intervals until a minimum signal-to-noise, $S/N > 15$ (Smee et al., 2013) is reached. A template is then fit to the spectra of the galaxy using a χ^2 minimization. The templates are built up from previous observations and include non-galaxy objects in case of misclassification. The template is spaced out in 45 km/s steps, going from redshift 0.1 to 1, with emphasis placed on fitting the $4000\AA$ break prevalent in CMASS galaxies. The χ^2 minimum is found, and must be much more significant than any other minima that is found – about 7σ (Bolton et al., 2012). CMASS galaxies are dominated by continuum, as opposed to specific spectral features that would cause the redshift fitting to fail. Sky subtraction does not cause redshift failures for the BOSS survey (Bolton et al., 2012).

3.2.2.3 Mask & Weights

The Angular Selection Function, including the observed footprint, of the SDSS survey is described using a mask. These masks mathematically detail the observed regions for the survey. Because the survey is tiled in chunks, the observable region is a complicated footprint that cannot be simply broken up into cut on Right Ascension and Declination. The masks, whose base level pixel structure is the the sector, incorporates this complicated selection function as well as additional information that is necessary to take into account when making a measurement of the galaxy distribution. A sector is a portion of a tile that is unique- two overlapping tiles have three sectors, like a venn diagram. Each galaxy belongs to a single sector which has a weighting associated with it based on completeness. The completeness of a sector is determined by the number of galaxies that were targeted vs. the number of galaxies that actually received redshifts.

In order to account for fiber collisions in the correlation function we assign a fiber collision weight to each galaxy. For each galaxy that did not receive a redshift, its nearest neighbor galaxy's weight increases by one. In a pair of galaxies separated by less than 62", one galaxy has a weight of two. These weights will factor during the pair counting aspect of measuring the correlation function. Additional weights include the sector completeness weights, which are based on the number of targeted galaxies versus the number of galaxies actually observed not due to fiber collisions. The total completeness of SDSS is over 97%, but individual sectors may be lower or higher depending on weather or fibers failings.

3.2.2.4 Calculating Absolute Magnitude

When calculating absolute magnitude, we must take into account how the emitted spectrum of the galaxy has been redshifted. These corrections are called k-corrections, and we use the model of Tojeiro et al. (2012) that models the evolution of the galaxy spectrum. This model corrects all galaxies' spectra to a common redshift of $z = 0.55$ so that the absolute magnitude in the $r - band$ can be compared equally between galaxies, regardless of how the observed flux has been altered by the spectra being redshifted out of the band. Figure 3.3 illustrates the k-correction by showing how the spectra of a galaxy at $z = 0.55$ has been redshifted.

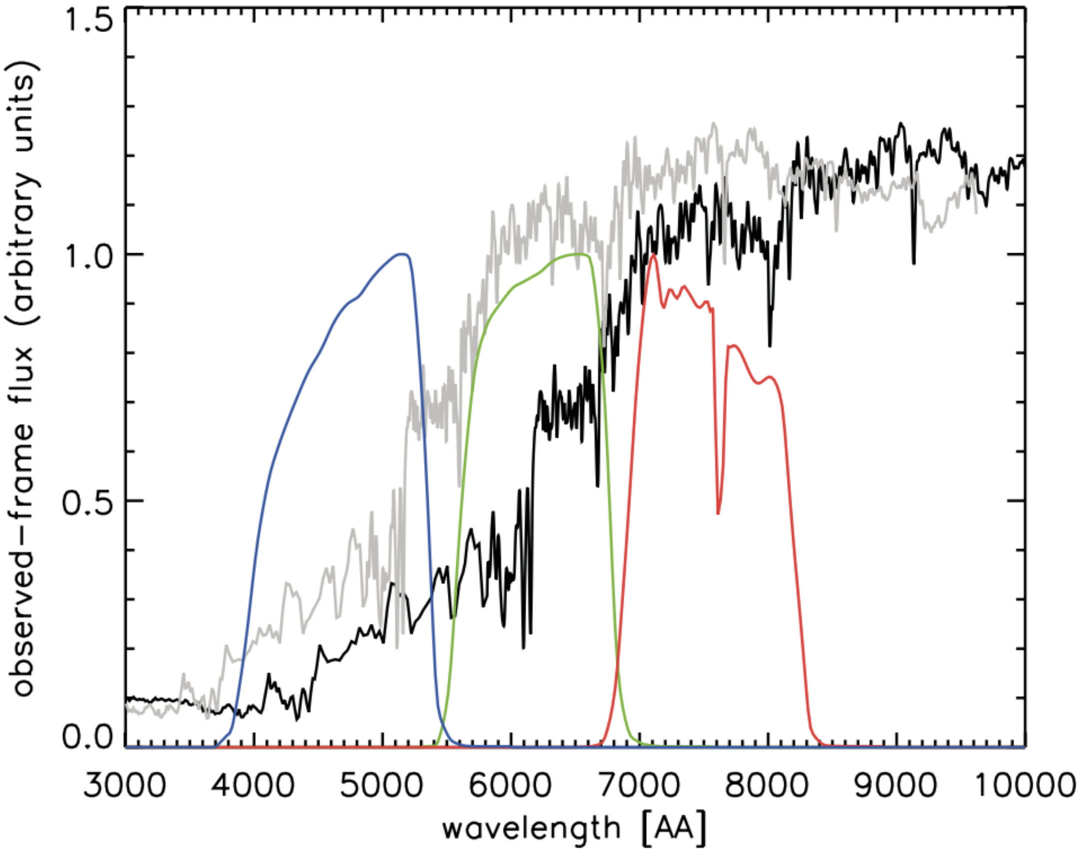

Figure 3.3 Figure 5 from Tojeiro et al. (2012) showing the expected observed spectrum, in black, of a typical BOSS CMASS galaxy at a redshift of $z = 0.55$. This Figure is meant to illustrate how the rest-frame spectrum of a typical galaxy will be redshifted out of the observed $g-band$. We correct for this decrease in flux using a $k-correction$ from Tojeiro et al. (2012).

Table III.1. Galaxy Catalogue Properties

Bin	Redshift Range	N_{gal}	$\bar{n_{gal}}$	$\bar{M_i}$
BinAll	(0.43,0.7]	505261	0.0001847	-21.5
Bin1	(0.43,0.5]	133404	0.0002494	-21.25
Bin2	(0.5,0.55]	143895	0.0003150	-21. 5
Bin3	(0.55,0.6]	116239	0.0002238	-21.6
Bin4	(0.6, 0.7]	111723	0.00009129	-22.0
LRG	(0.16,0.36)	61899	0.000094	-21.65

Note. — Attributes of the CMASS redshift samples that we use. We note that the number density of the total sample is closest to Bin4, the highest redshift sample. For part of our analysis we combine the lower and upper two bins. We have also included values for the LRG sample, from Kazin et al. (2010)

3.2.2.5 Redshift Samples

We choose to make redshift cuts for the spectroscopic sample based on having equal number of galaxies in each bin. A summary of the properties of these samples can be found in Table III.1 and the number distribution is shown in Figure 3.4. The peak of the redshift distribution of galaxies corresponds to our second bin, $0.5 < z < 0.55$, which has the highest number density of all the bins. Figure 3.5 shows the evolution in the color-color targeting cuts as function of redshift. As we move to higher redshift bins the galaxies become redder. Additionally we show the same redshift evolution as a function of $d_\perp$ in Figure 3.6. Higher redshift galaxies have on average higher values of $d_\perp$. Combined, both of these redshift trends are due to needing to get more red galaxies as a function of redshift in order to have a constant number density.

The bins span a large range of cosmic time; long enough to see an evolution in the very

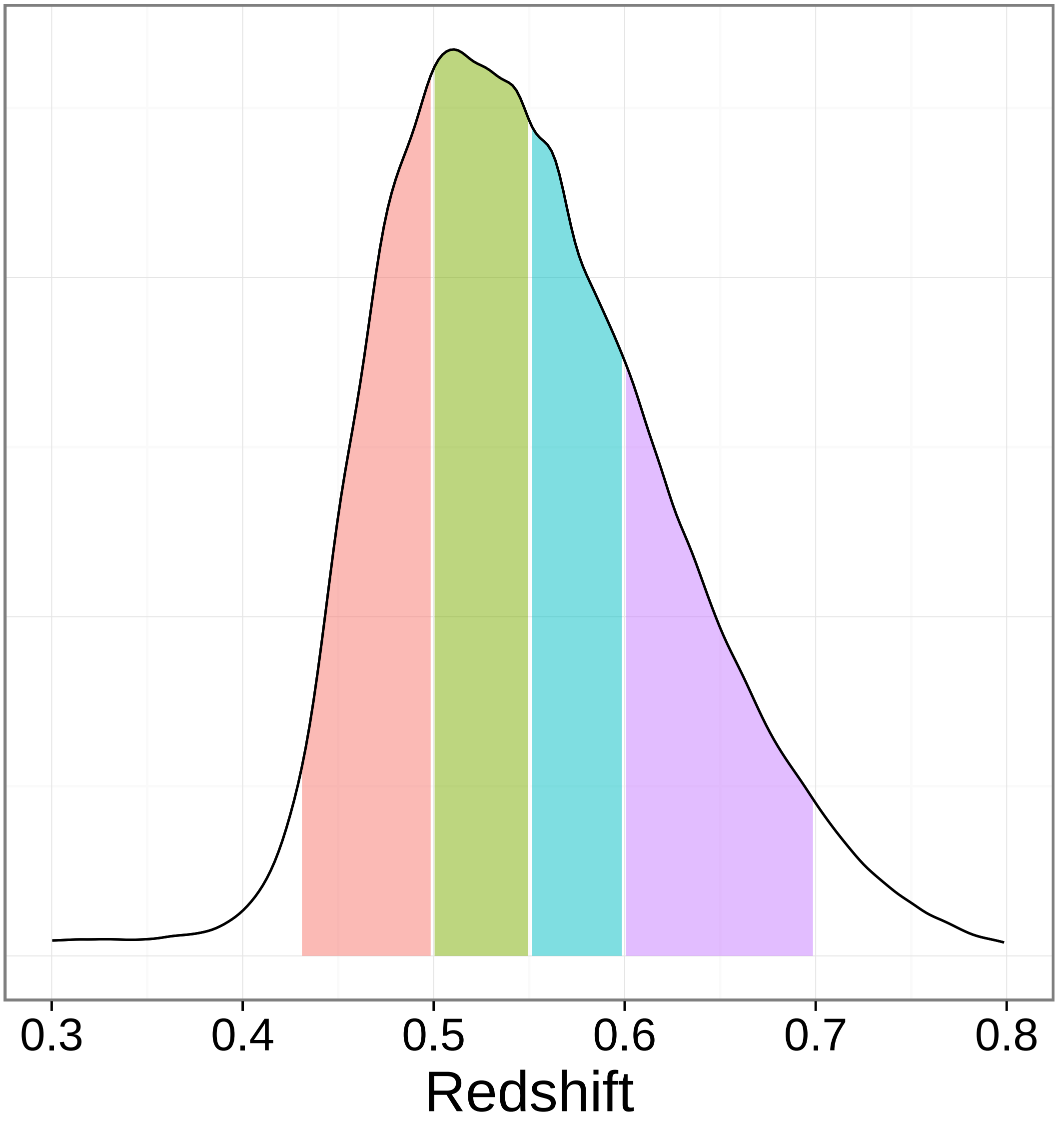

Figure 3.4 The number density of CMASS galaxies as a function of redshift. The redshift range of our four samples are highlighted in different colors. The number of galaxies is roughly equal in each of the redshift bins. Their properties are shown in Table III.1.

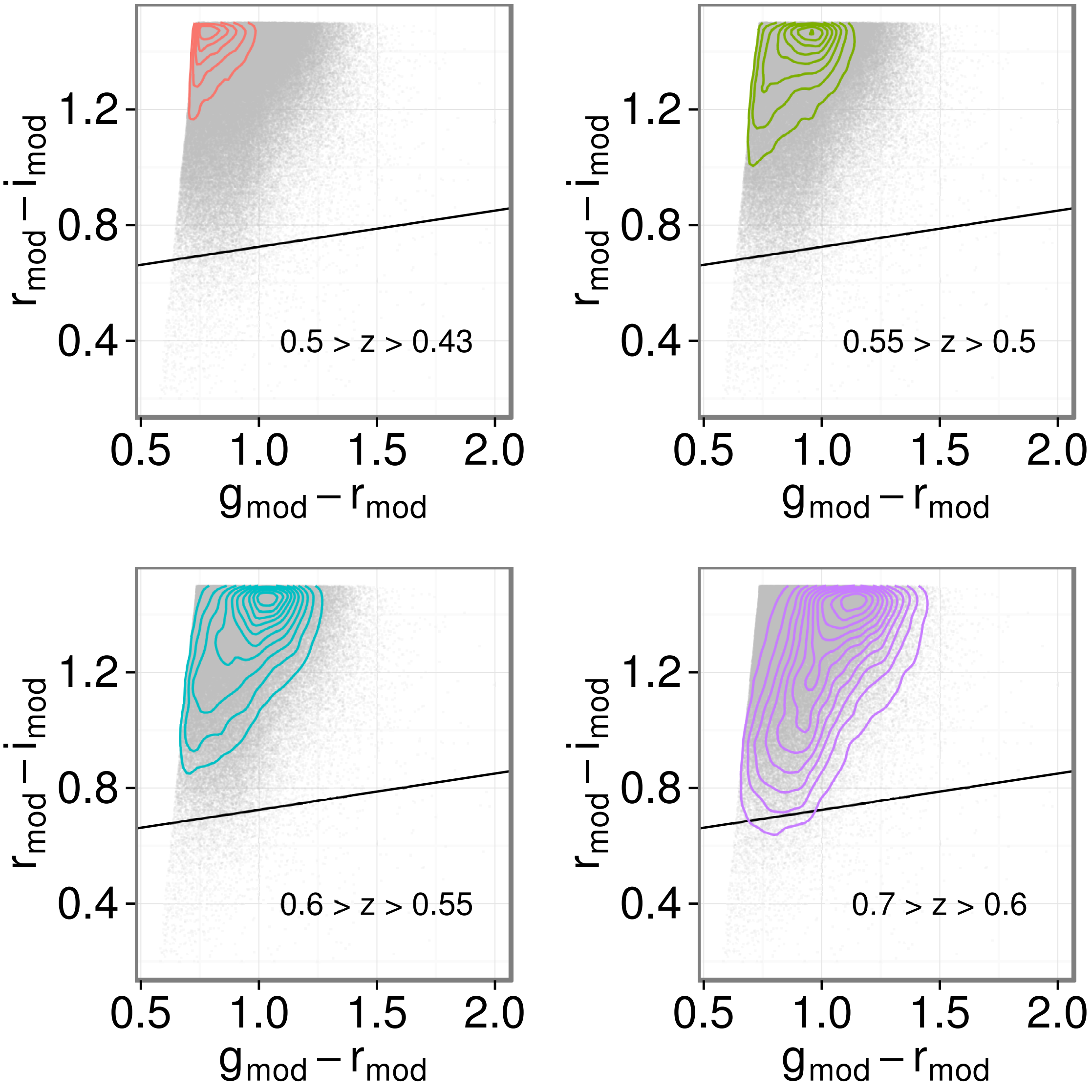

Figure 3.5 The color and magnitude cuts as a function of redshift. The grey points are the same in every panel and show the total underlying color color distribution of CMASS galaxies. The colored lines are the iso-density contours for galaxies in each of the four redshift bins. We can see that the higher redshift galaxies have a larger spread in the color-color space. As the volume increased, redder galaxies had to be included in order to keep the same number density. This is shown in the rotated color $d_{\perp}$ in Figure 3.6.

small scale clustering, if one exists, due to a change in the satellite occupation fraction of halos. The dynamical time for a satellite galaxy falling into a halo at $z = 0.7$ is rough 900Myr. This would put the theoretical satellite's peri-center passage at a redshfit of $z \sim 0.55$, straddling our second and third redshift bin.

3.3 Projected Cross Correlation Function

The project two-point correlation function can be defined as

$$w_{\rm p}(r_{\rm p}) = 2\int_{0}^{\pi_{max}} \pi\, \xi(r_{\rm p}, \pi), \tag{3.11}$$

where $r_{\rm p}$ is the distance perpendicular to the line of sight and π is the line of sight distance. We define ξ using the Landy & Szalay (1993) estimator:

$$\xi = \frac{DD - 2DR + RR}{RR}, \tag{3.12}$$

where DD, DR and RR are the correctly normalized number of data-data, data-random and random-random pairs in each bin of angular separation $r_{\rm p}$. By integrating over a sufficiently large π, all physically correlated pairs are included. One of the benefits of measuring $w_{\rm p}$ instead of the real- or redshift-space correlation function is its insulation from line of sight distance errors due to the peculiar velocity of galaxies. The $r_{\rm p}$ we calculate is the comoving distance between this pair of galaxy at the redshift of the spectroscopic galaxy.

Measuring galaxy clustering on very small scales presents two main challenges– sample

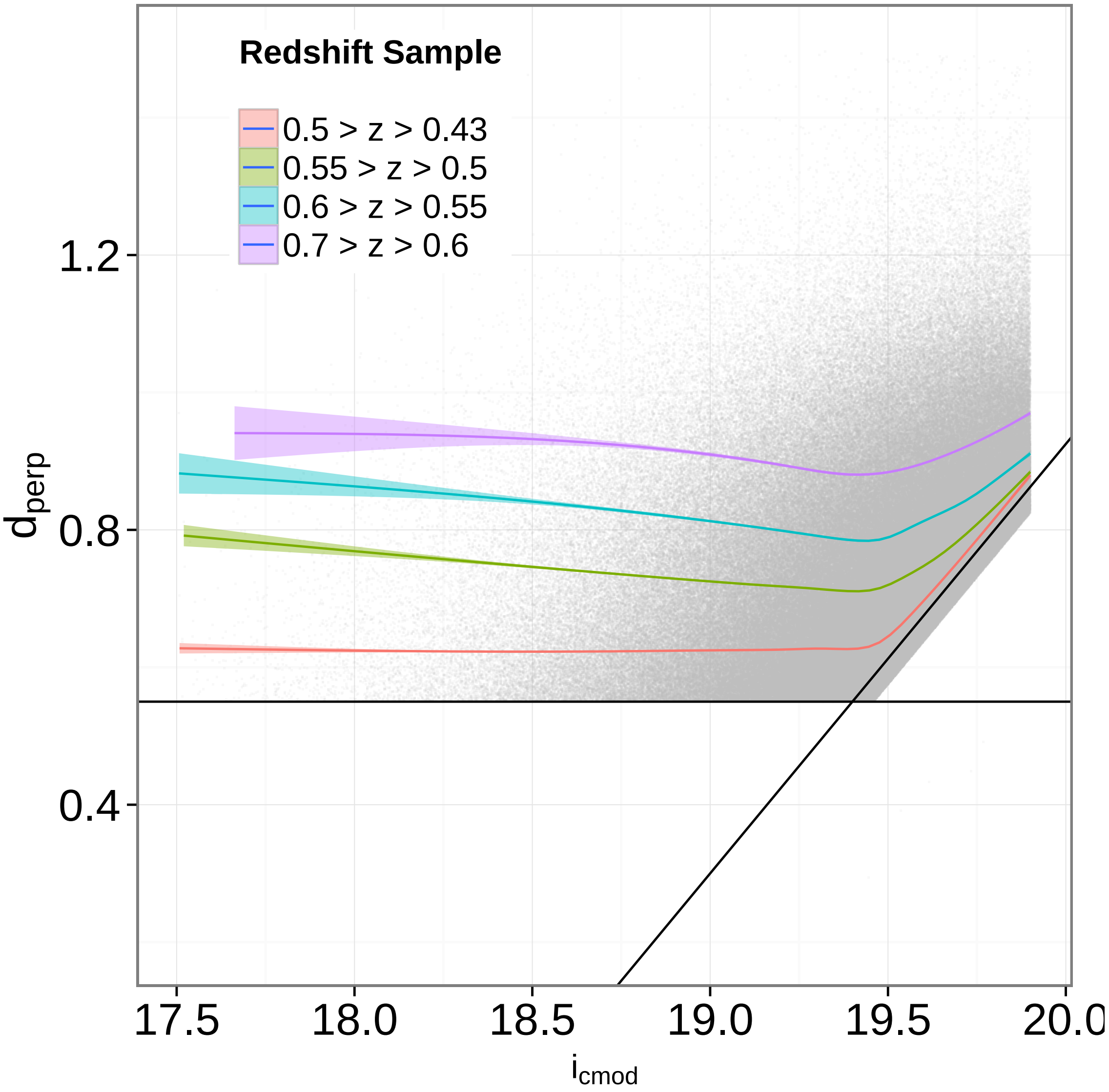

Figure 3.6 The rotated color $d_{\perp}$ vs $i - band$ apparent magnitude of the total CMASS sample, shown with the median values and standard error spread at each redshift. We can see that the higher redshift bins have a higher $d_{\perp}$ on average, which is consistent with being redder, as shown in Figure 3.5. The locus of points that extends outside of the cuts, to the lower right, consists of the commissioning photometry sample which had a slightly different color-magnitude cut. This sample went on to be named the sparse sample.

contamination from survey dependent systematics, such as fiber collisions and deblending, and low number counts causing large noise. Some of these effects can be mitigated by cross correlating a spectroscopic sample, having redshifts, and an imaging sample, which is much larger. Having the redshift of one of the galaxy pairs transforms an angular distance on the sky to a physical separation. Having an imaging sample that includes galaxies missed due to fiber collisions adds reliability to very small scale clustering measurements.

Cross correlation have been shown to measure the small scale clustering of Luminous Red Galaxies (Eisenstein, 2003). Here, we employ the Masjedi et al. (2006) method. This method measured the projected cross correlation function between spectroscopic and imaging samples in order to recover galaxies lost to fiber collisions. Galaxies that do not exist in the spectroscopic, or D_s, catalogue would still be present in the imaging, or D_i, catalogue. The cross correlation allows us to count those pairs by assuming the imaging galaxy is at the same redshift as the spectroscopic galaxy. However, cross correlating D_s and D_i also introduces line of sight uncorrelated pairs. If we assume that those uncorrelated pairs are distributed randomly, we can subtract off their contribution by including cross correlations between random catalogues.

We define the cross correlation function as

$$n\, w_{\mathrm{p}}(r_{\mathrm{p}}) = \frac{\mathrm{D_s D_i}}{\mathrm{D_s R_i}} - \frac{\mathrm{R_s D_i}}{\mathrm{R_s R_i}} \quad , \tag{3.13}$$

where n is the average comoving number density of the spectroscopic catalogue, D_s and D_i are the spectroscopic and imaging data samples, and R_s andR_i are the random spectroscopic and random imaging samples. In follow up work, Masjedi et al. (2008), changed the definition of number density to be that of Imaging data catalogue– a number that cannot be measured as the true number density of the imaging catalogue is not able to be determined without redshifts. While we do note that the theoretically quantity n_{Di} is technically the more accurate term, n_{Ds} should be very similar n_{Di}.

The first half of the estimator cross correlates the spectroscopic data catalogue D_s to determine the signal of clustering. The first numerator is defined as

$$\mathrm{D_sD_i} = \frac{\sum\limits_{j \in D_s D_i \mathrm{pairs}} p_j}{\sum\limits_{j \in D_s} p_j} \quad , \tag{3.14}$$

where p_j is fiber collision weight assigned to each spectroscopic galaxy. This term sums over all the pairs of galaxies, weighting them by their fiber collision weights (explained below), and then normalizes by the sum of fiber collision weight in the sample. Fiber collision weights are very close to unity so this term is effectively counting all the pairs and then dividing by the number of galaxies in the spectroscopic sample.

Fiber collision weights are assigned by running a friends-of-friends group finding algorithm, with the linking length being the fiber collision scale (SDSS-II 55", SDSS-III 62"). Each fiber collision group is then assigned spectroscopic fibers in such as way as to maximize the number of galaxies retained in the sample, echoing the spectroscopic tiling algorithm (Blanton et al., 2003b). The weight p_j is then calculated by counting the number of galaxies lost and assigning those counts to their nearest neighbor with was retained and is very close to unity.

Continuing, the first denominator is defined as,

$$\mathrm{D_sR_i} = \frac{\displaystyle\sum_{j\in D_s R_i \mathrm{pairs}} p_j}{\displaystyle\sum_{j\in D_s} p_j \left(\frac{\Omega}{A}\right)_j \frac{N}{\Omega}} \quad , \tag{3.15}$$

where $\left(\frac{\Omega}{A}\right)_j$ is the inverse square of the comoving distance to spectroscopic galaxy j and $\frac{N}{\Omega}$ is the number density of the random imaging catalog per solid angle. The multiplication of these two terms gives the average number of random imaging objects per unit comoving area around each spectroscopic galaxy. In practice, the numerator is counting all pairs between the D_s and R_i catalogues, weighting by the fiber collision weight of the D_s galaxy pair. The denominator is normalizing the number of D_s-R_i pairs by estimating the number of possible random imaging galaxies around each spectroscopic galaxy.

The second half of the estimator, the cross correlation with the random spectroscopic catalogue R_s, is meant to subtract out physically un-associated pairs that can be seen as noise in the pair counting of the first half. The second numerator is defined as:

$$\mathrm{R_sD_i} = \frac{\displaystyle\sum_{j\in R_s D_i \mathrm{pairs}} f_j}{\displaystyle\sum_{j\in \mathrm{R_s}} f_j} \quad , \tag{3.16}$$

where f_j is the weight given to random spectroscopic galaxy j which accounts for other incompleteness of the spectroscopic survey in that region of the sky not due to fiber collisions. Similarly to Eq. 3.14, this is a sum over all pairs, weighted by the completeness weight associated with the R_s galaxy pair. It is then normalized by the sum of all f_j weights. Completeness in each sector is determined by the number of galaxies that were targeted versus the number of galaxies that actually received good redshift measurements.

Incompleteness that is not fiber collision related could be due to poor signal to noise in a spectra or fibers failing. Completeness is very close to 100% in all sectors, so the f_j completeness weights are often unity.

Finally, the second denominator (similar to the first denominator) is defined as:

$$\mathrm{R_sR_i} = \frac{\displaystyle\sum_{j \in R_s R_i \mathrm{pairs}} f_i}{\displaystyle\sum_{j \in R_s} f_j \left(\frac{\Omega}{A}\right)_j \frac{N}{\Omega}} \quad , \tag{3.17}$$

where the terms are similar to those in Eq 3.15, except f_j is the sector completeness weight.

In order to estimate errors and measure the covariance matrix, we separate the galaxy catalogue footprint into 150 jackknife samples that represent approximately equal area sections on the sky, shown in Figure 3.1. For each jackknife sample k, we measure the projected correlation function w_p^k. The covariance matrix can then be computed as

$$C_{ij} = \frac{N-1}{N} \sum_{k=1}^{N} ({\mathrm{w_p}}_i^k - \bar{\mathrm{w_p}}_i)({\mathrm{w_p}}_j^k - \bar{\mathrm{w_p}}_j), \tag{3.18}$$

where C_{ij} is the covariance between projected distance bins i and j, and $\bar{\mathrm{w_p}}_i$ is the mean of correlation function measurements in projected distance bin i computed from the N jackknife samples. In practice, each jackknife region has the four terms of w_p measured separately and then combined before the variance between jackknife regions can be calculated. Jackknife regions are based on the underlying sector distribution of galaxies.

3.3.1 Tests on Mock Galaxy Catalogues

In order to test the efficacy of this method for estimating the projected correlation function, we perform validation tests on Mock galaxy catalogues. For the first test, we compare the results of the CrossCorrelation estimator to a standard Landy-Szalay estimator. Because of the lack of fiber collisions and deblending errors in mocks the two estimators should be equivalent. The mock galaxy catalogues were creating using dark matter halos from the LasDamas N-body simulation and standard values for the halo occupation distribution (Guo et al., 2014). This allows us to populate dark matter halos with a realistic distribution of galaxies. Because we will be populating halos at very small scales, we assume that the galaxies follow an Navarro-Frenk-White density profile. This allows us to avoid numerical issues, such as softening, that could bias our comparison. See Chapter 2 for a more detailed description of how mock galaxy catalogues are created.

In Figure 3.7, we compare the results of the Landy-Szalay estimator to the Cross Correlation estimator. We performed this test on ten different mock galaxy catalogues for robustness. In Figure 3.7 we see good agreement between the two different methods of estimating clustering on mocks. There is a much larger variance in the cross correlation $w_{\rm p}$ at the largest scales, which we believe is due to treatment of random-random pairs in the estimator. From this test we conclude that the cross correlation estimator is effective at accurately measuring clustering in the absence of systematics.

For the second test, we compare results between mock galaxy catalogues with and without fiber collisions. As stated before, the main motivation for using the cross correlation estimator is that it is able to account for galaxy pairs lost to fiber collisions The mock

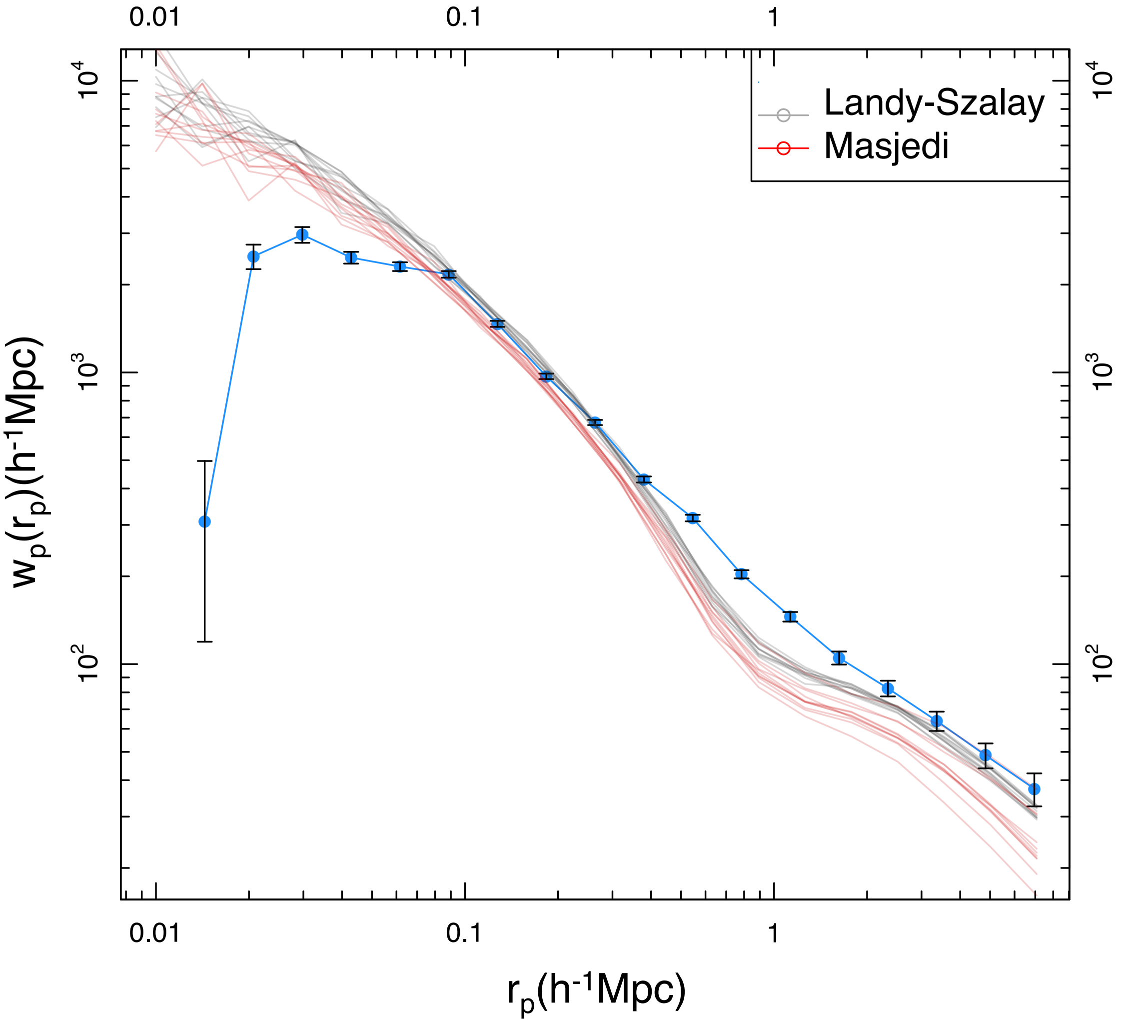

Figure 3.7 Comparison of the cross correlation estimator to the Landy-Szalay estimator. Each line represents a measurement from a single mock galaxy catalogue. The mock galaxy catalogues were created using a standard HOD from the literature and dark matter halos from the LasDamas simulations. We measured the cross correlation and L-S estimator on each catalogue. There is good agreement at all scales, however the large scatter shown in the cross correlation is not present in the L-S estimator. We have added our measurement of BOSS CMASS galaxies to guide the eye. Please see Section 3.4 for a description of the data measurement.

galaxy catalogues are populated the same way as before, but we then run a collision group finder which groups galaxies using a linking length of 62". We assign redshifts optimally so that the most number of galaxies in a collision group receive redshifts. Often, these groups are simply pairs where the pair galaxy that gets to keep its redshift is randomly chosen. Our imaging sample D_i includes all galaxies, while the spectroscopic sample, D_s, only includes galaxies that kept their assigned redshifts.

We note that this is only an approximation of how fiber collisions occur in observations- there are additional constraints such as foreground objects or Quasars, which have target priority over galaxies. As shown in Appendix A, this approximation only recovers only half the number of fiber collisions that occur in data. The correct way to add fiber collisions to mocks is to create a full flux limited mock catalogue that includes both foreground and background uncorrelated objects. This is not feasible at this time because of numerical constraints and is not necessary for this test.

Once fiber collisions are included in our mocks, we can test how well the cross correlation estimator is able to recover the correct clustering where the correct clustering is the clustering measured on the mocks without fiber collisions. This is a fair comparison as the only difference between individual mocks is the effect of fiber collisions. We see good agreement between these two measurements, lending veracity to the claims that the cross correlation estimator is able to take into account data systematics. For comparison, we have also used the L-S estimator on the fiber collided mocks as well as showing the clustering of a mock without any galaxies separated by less than the fiber collision scale to show maximum effect. This shows that the CC estimator is the better choice for measuring very small scale structure.

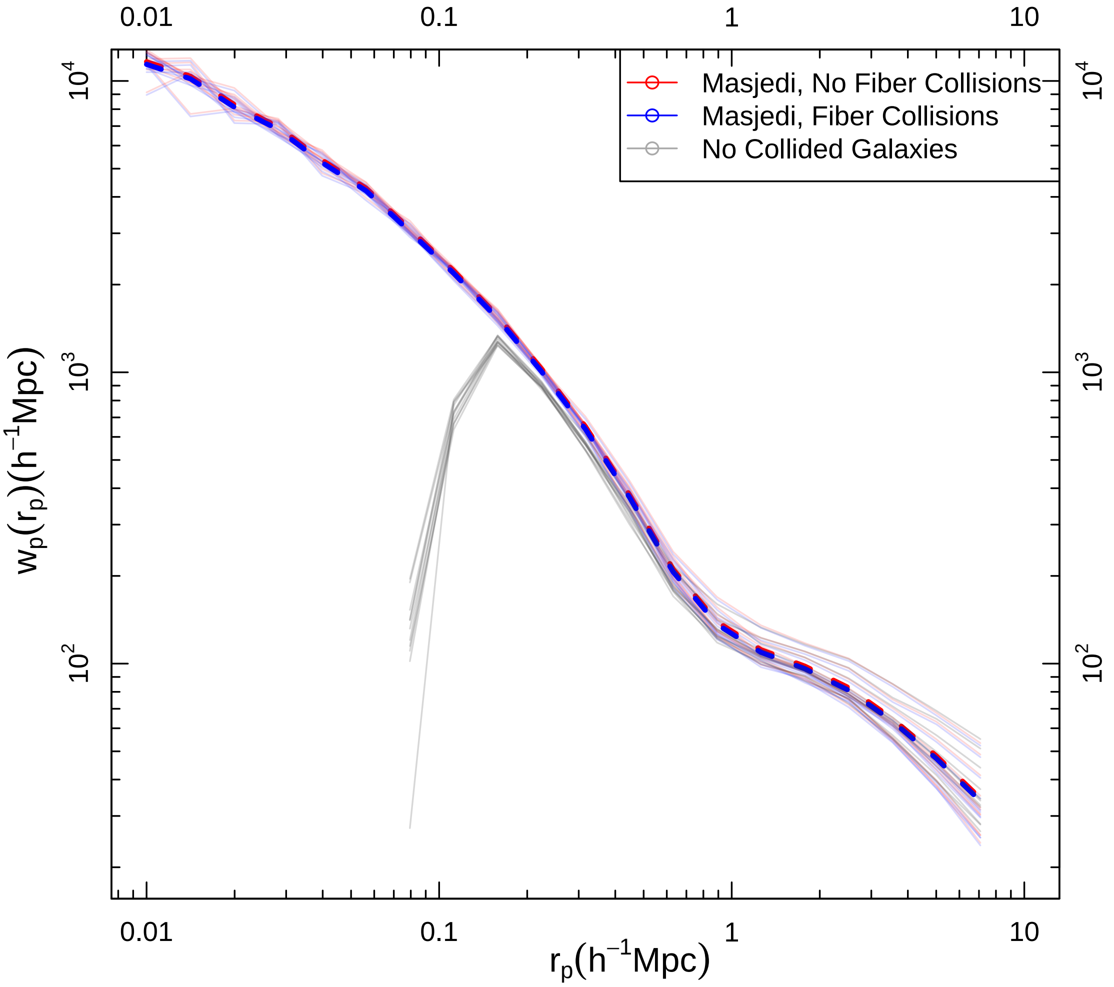

Figure 3.8 Comparison of the cross correlation estimator on mocks with and without fiber collisions. Each line is a different mock galaxy catalogue made from different realizations of the same simulation. The red lines are the CC on mock catalogues without any fiber collisions, the same as in the previous plot. The blue lines are the CC measurement on mock catalogues that have fiber collision incompleteness introduced through a friends-of-friends collision group finding algorithm. The CC estimator effectively recovers the correct clustering at very small scales. We have also included lines showing the maximum affect of fiber collisions, where all galaxies without measured redshifts are removed from the sample. We can see that fiber collision play a very large role starting around $r_{\mathrm{p}} < 0.3h^{-1}\mathrm{Mpc}$

3.3.2 Comparison to Masjedi et al. 2006

In order to verify our implementation of the CC estimator, we reproduced the measurements of both Masjedi et al. (2006), hereafter M06, and Zehavi et al. (2005a), hereafter Z05. M06 found a good agreement in the real space correlation function $\xi(r)$ with the results of Z05. Here, we will compare the measurement, of $w_{\rm p}$. The steps to recreate these measurements are as follows: 1) Locate the correct spectroscopic catalogues, both data and randoms, 2) Create the appropriate imaging catalogues and 3) Calculate $w_{\rm p}$ in the exact same way as M06, including making the same absolute magnitude cuts on the imaging catalogue needed to make this measurement an autocorrelation between samples with the same luminosities. This last point requires validation of our calculation of absolute magnitude. Once we have verified our method, we will apply it to measuring the very small scale clustering of SDSS DR7 LRGs, a much larger sample than those used in either M06 or Z06, and with better calibrated magnitudes.

We use data and random spectroscopic catalogues from Kazin et al. (2010) for both Data Release 3 and Data Release 7, which already have accurately computed fiber collision and completeness weights. DR3 is very similar in footprint and number of galaxies to the Sample 14 that both M06 and M05 used in their analysis, which is no longer available. We used OmegaM=0.3 OmegaL=0.7 for DR3 and for DR7 we used OmegaM=0.25 and OmegaL=0.75 to be consistent with both previous analysis.

We constructed our imaging catalogues using the SDSS Catalogue Archive Server (CAS), selecting objects that passed the correct LRG primtarget flag from the photoObj schema. We return the objects's RA, Dec, r-band magnitude and r-band extinction. The necessity of the last two items is discussed below. This query is equivalent to creating a query using the full LRG sample selection, which is included in the appendix.

```
SELECT
  ra, dec, petroMag_r, extinction_r into
        mydb.lrg_dr3 from photoObj
  WHERE ( PrimTarget=32 )
```

We made the random imaging catalogue using the mangle ransak function and the appropriate polygon files describing the survey footprint.

Unlike CMASS galaxies, the sample of LRGs used in this analysis has an absolute magnitude cut. In order to make the same absolute magnitude and redshift cuts from Masjedi et al. (2006), we calculated the absolute g-band Magnitude of each imaging galaxy at the redshift of its spectroscopic pair. The m_r apparent magnitude is used because the rest frame g-Band absolute Magnitude of the galaxy is redshifted into the r-band in the observed frame (Zehavi et al., 2005a). We follow the methodology of Kazin et al. (2010) outlined below.

$$M_g = r_{petro} - extinction_r - 5log_{10}\left(\frac{D_L(z_i)}{10pc}\right) - K_e(z). \tag{3.19}$$

Where $K_e(z)$ are color, evolution and k-corrections based on the models of Eisenstein et al. (2001)

$$K_e(z) = \delta_g - (g - r) + z_{calibration}. \tag{3.20}$$

In practice, these values are interpolated from the table in Eisenstein et al. (2001)

and are solely a function of redshift. The value of $z_{calibration}$ is a constant 0.2, taking into account evolution. The values of $K_e(z)$ are meant to evolve all galaxies to a rest frame redshift of $z = 0.35$. We show in Figure 3.9 that our calculation of absolute magnitude is consistent with that of Kazin et al. (2010).

We present our measurement of very small scale LRG clustering in Figure 3.10. We see a good agreement with previous results on all scales, thus verifying our algorithm as well as sample selection for the imaging sample. We also present, for the first time, a measurement of the very small scale clustering of SDSS DR7 LRGs. We see that there is little to no difference in the small scale pairwise distribution, excluding a bump at about 100kpc. We do include in our measurements the correction M06 used to compensate for the affect of deblending, which will be discussed in Section 3.4.

3.4 Clustering Results

In Figures 3.11 and 3.12 we present the measurements of the projected correlation function for the samples described above in the range $0.01h^{-1}\text{Mpc} < r_{\text{p}} < 5h^{-1}\text{Mpc}$. We have go much smaller than the standard separation of $0.1h^{-1}$Mpc to emphasize the clustering due to central-satellite pairs, which make up the majority of counts in the innermost bins. The gray dotted line in each panel of Figure 3.11 represents the fiber collision scale in real space separation at the innermost redshift. It is clear that the scale at which fiber collisions become a source of error also corresponds to roughly the beginning of the one-halo term around $r_{\text{p}} = 0.45h^{-1}$Mpc. This further motivates our use of the cross correlation estimator. We also note, with the golden dashed line, the smallest separation that M06 found to be trustworthy based on the seeing of the telescope.

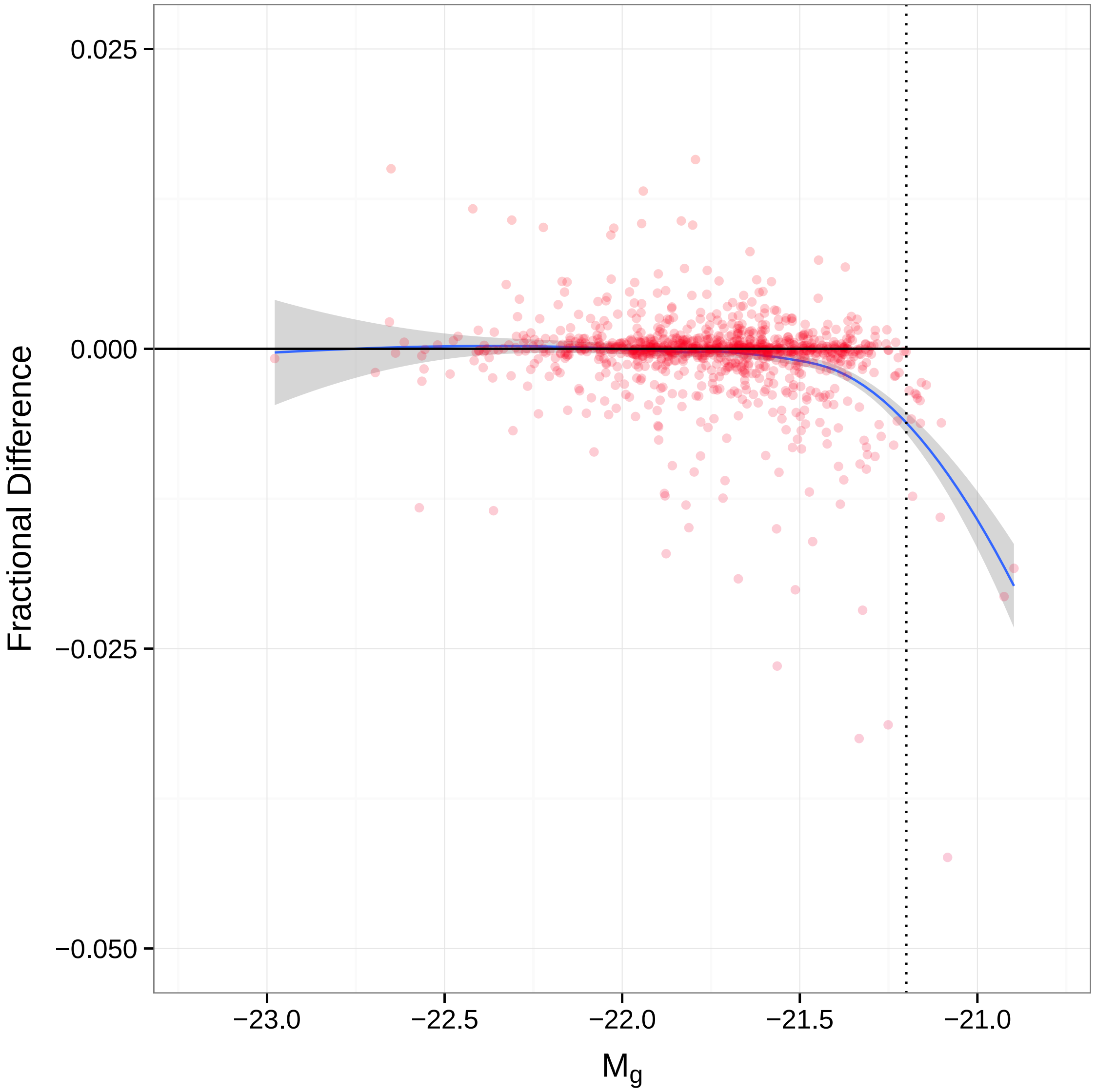

Figure 3.9 This figure shows the comparison between our calculation of M_g to those stated by Kazin et al. (2010). The bottom axis shows our calculated M_g while the vertical axis shows the fractional difference between our calculation and those stated in the literature. The blue line shows the smoothed mean and the grey shaded region shows the standard error. This comparison was necessary in order to be confident we were recreating the measurement of M06, who used absolute Magnitude cuts to create their sample. We note the lower luminosity limit of the sample with the solid black line.

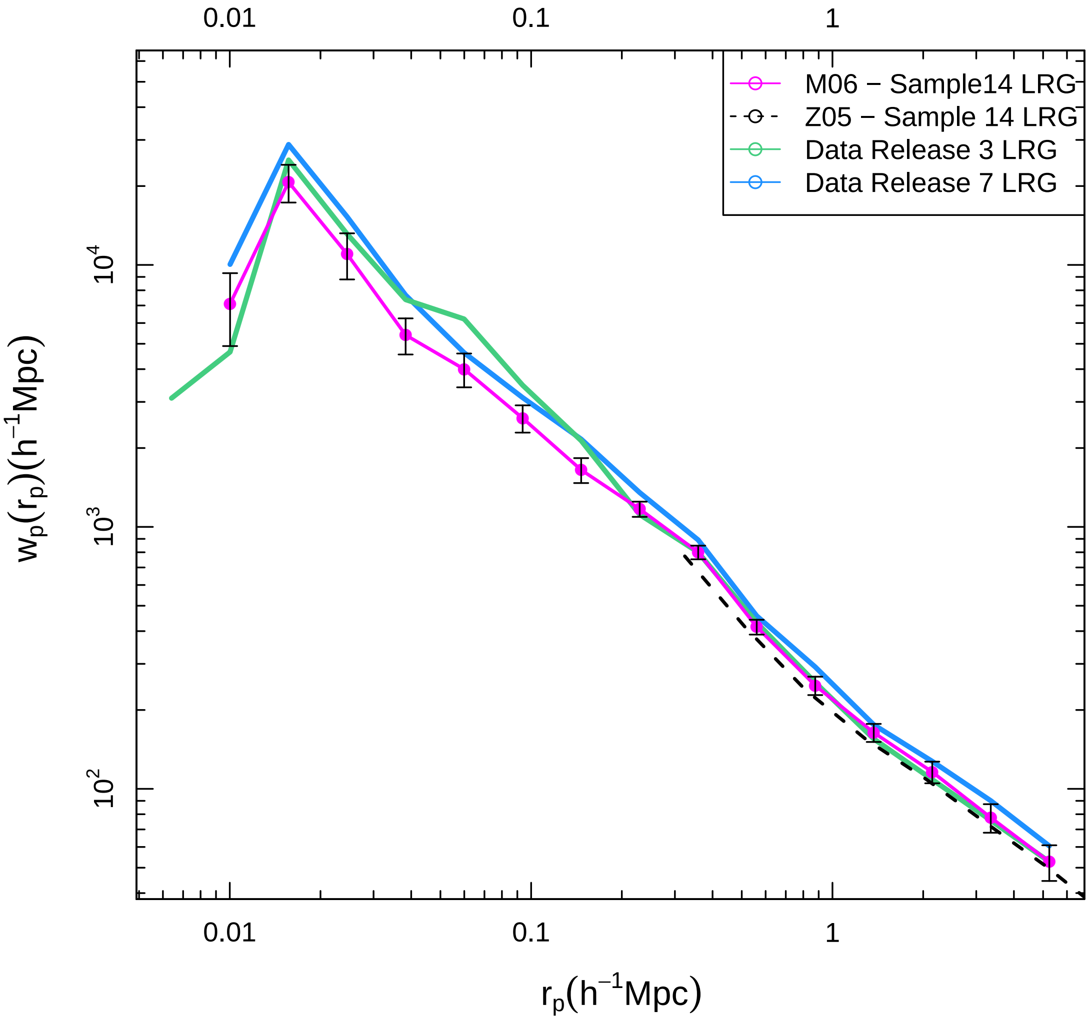

Figure 3.10 The projected cross correlation function w_{p} of the very small scale clustering of SDSS-II Luminous Red Galaxies. We plot the literature values for M06 and Z05, as well as our measurement of LRGs from Data Releases 3 & 7. Our measurements agree very well with those previously published. Our measurements of LRGs are also consistent with each other, except for a bump at $r_{\mathrm{p}} = 0.1h^{-1}$Mpc, and show no discrepancy on the smallest scales. The slight difference at the larger scales could be due to cosmic variance. The errors are taken from M06.

We show the measurement of the very small scale clustering of BOSS CMASS galaxies in four redshift bins in Figure 3.11. The points are the mean value of w_p measured in each jackknife region and the error bars are the diagonal elements of the covariance matrix estimated using jackknife regions and Equation 3.18. We see hints of evolution in the inner bins, with the highest redshift bin having a very flat shape at separations smaller than about $r_\mathrm{p} < 0.1h^{-1}$Mpc. The lowest redshift bin show a smoothly increasing function with a much less pronounced turn over at smallest scales. Regardless of bin, one can see that there is a flattening in the CMASS clustering.

In order to better measure any trend in the shape of the correlation function at very small scales, we combine to four redshift bins into two. This has the benefit of beating down noise at the smallest scales. We present this w_p measurement for two total samples of CMASS galaxies ranging from $0.43 < z < 0.55$ and $0.55 < z < 0.7$ in Figure 3.12. Additionally, we present the measurement for the SDSS DR7 LRGs. Previous clustering measurements of LRGs, in M06 and Z05, have been performed on the smaller Sample14 data release. We have shown in Section 3.3.2 that our measurements are consistent with previous findings. The slight offset at larger scales between our DR3 and DR7 LRG measurement is most likely due to cosmic variance; the DR3 sample is both smaller and covers a smaller footprint on the sky. Again, we note the fiber collision and deblending scales, calculated at $z = 0.43$.

There is a statistically significant different in the very small scale clustering of LRGs and CMASS galaxies, thought to be the progenitors of the LRGs (Maraston et al., 2009; Tojeiro et al., 2012). Additionally, there is a smaller but still significant difference between the clustering of the different CMASS bins, with the largest downturn happening in the

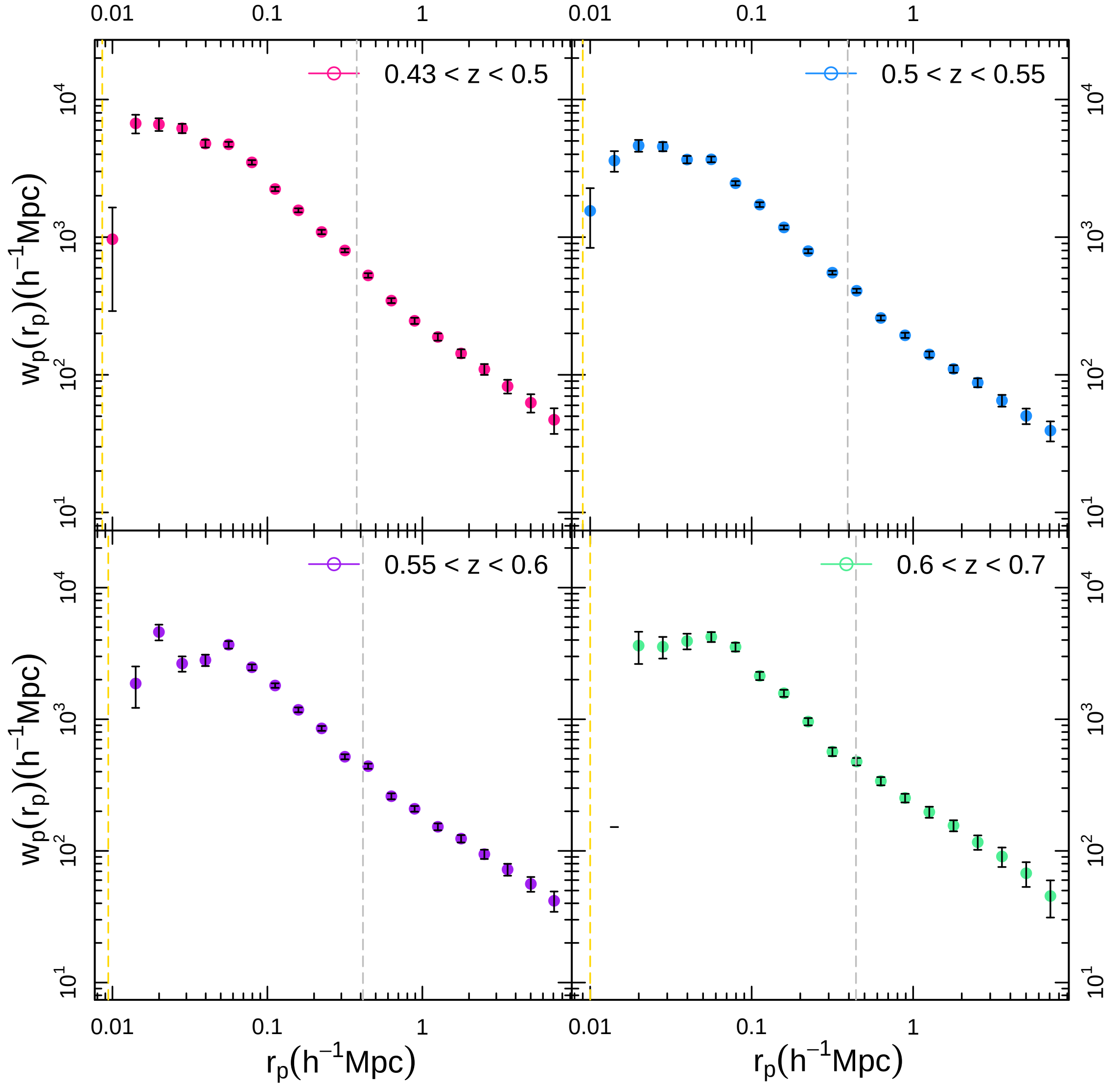

Figure 3.11 We show here the very small scale clustering in four different redshift bins of BOSS CMASS galaxies. The gray dotted lines denote the fiber collision limit at the inner edge of each redshift bin. The golden dotted line is the effective radius of the LRG galaxies from M06 transformed into a physical radius at the outmost edge of the redshift bin. There does appear to be a trend towards flattening of the inner slope of w_p as we go out to higher redshifts.

higher redshift sample. This is consistent to what was hinted at in the previous Figure 3.11 but is now made clearer due a decrease in the errors.

One possible explanation for this turn down is an observational systematic called 'deblending'. When two extended objects are overlapping on the plane of the sky, the PHOTO pipeline has to decide how to allocate flux to each of the galaxies. Naively, one could think that this would cause galaxies to be lost from the sample thus decreasing the pair counts at very small separations. Howeve, M06 showed that the downturn is not due to deblending, as deblending tends to boost clustering at small scales by adding lower luminosity galaxies into the sample. The PHOTO pipeline assigns additional flux to the fainter galaxy, augmenting its brightness enough to pass the lower luminosity threshold. Additionally M06 estimated a deblending correction term for the smallest bins that pushes the correlation downwards. For our measurements of LRGs we do implement this correction term in order to better compare our measurements with M06.

Additionally, M06 also estimated the minimum projected separation that one may trust the distribution of fluxes that occurs due to deblending. We note this separation one our plots with the gold dashed line. This calculation is based on the angular scale of the half light radius of the galaxies transformed into a physical scale at the median redshift of the survey. We do stress that the turnover in the last bin of the LRG measurement is most likely due to deblending, our measurement turns over far before this scale. We do not implement the same correction as M06 for our CMASS galaxies; had we done this, the turnover would be more pronounced. While we cannot run the same estimation of the effect of deblending, Beifiori et al. (2014) has measured half light radii of BOSS CMASS galaxies and have found that these radii are substantially smaller than our smallest bin–

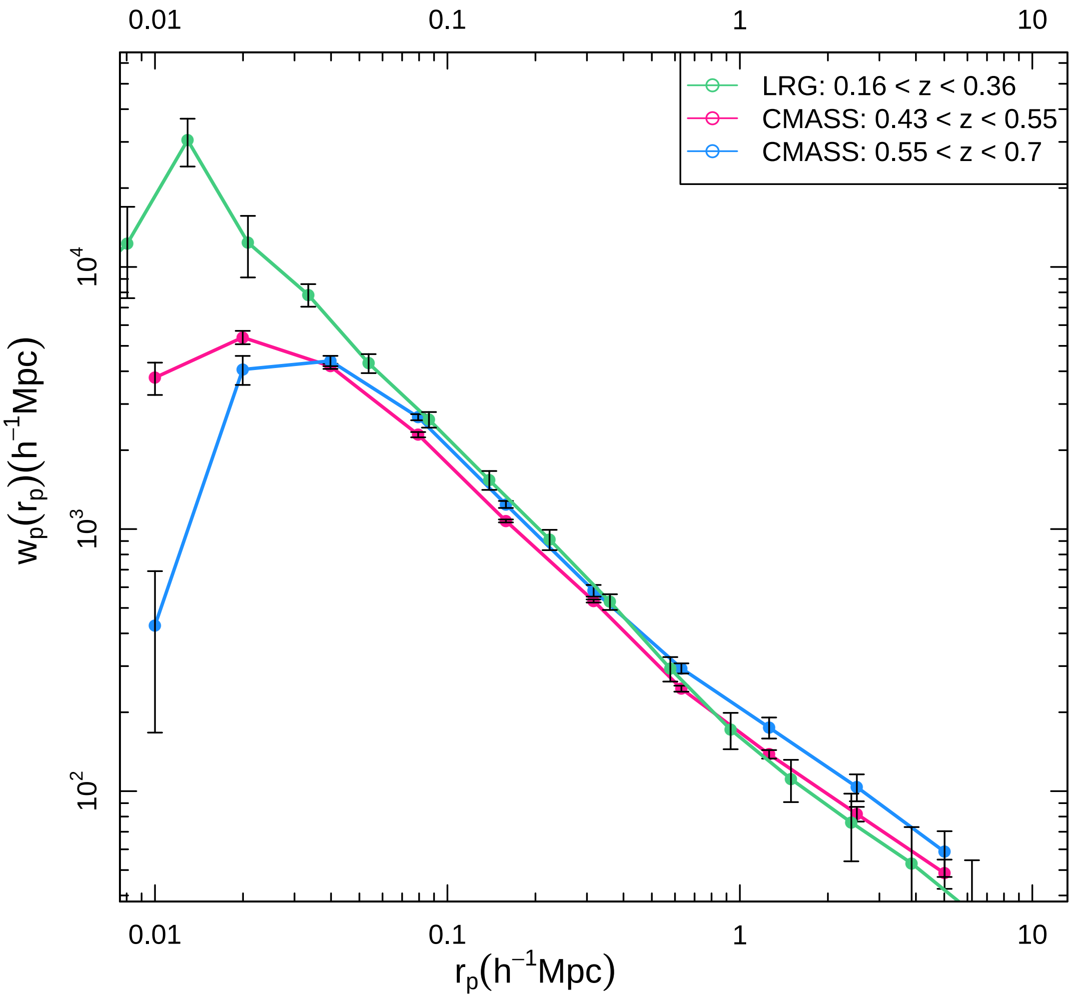

Figure 3.12 The projected cross correlation function w_{p} for SDSS-II Luminous Red Galaxies in green and two redshift bins of SDSS-III BOSS CMASS galaxies in blue and pink. We see for the first time a significant difference in the shape of clustering between the LRGs and the CMASS galaxies. We also see a smaller but still significant difference in the clustering between the two CMASS redshift bins.

about 5kpc. This translates to 1” on the sky– half the angular resolution of the SDSS survey. Additionally these half light radii are consistent with that of the LRGs, meaning that the deblending correction for CMASS galaxies would be similar to that for the LRGs. Based on these facts, we believe that the difference in slopes for our three measurements, the DR7 LRGs and the two redshift bins of CMASS galaxies, is not due to deblending.

3.4.1 Comparison to Theory Prediction

In order to better understand ρ_{gal} on very small scales, we compare our measurements of both CMASS galaxies and LRGs to mocks created using the HOD of Guo et al. (2014) and assuming that galaxies follow an NFW distribution inside their host dark matter halos. We use halos from the LasDamas N-body simulation evolved to a redshift of $z = 0.52$, consistent with the redshift of our second bin.

As compared to data we see a much more pronounced transition between the one and two halo terms, happening at 1 Mpc, that is not visible in the mock measurement. The measurements in Guo et al. (2012a) do show more of a transition than our measurement. We emphasize that the HOD we use was not fit to our data measurement, nor was it fit to an N-body simulation. We have compared the clustering results of this HOD with the values from White et al. (2011) and find very little difference in clustering. Another note is that the cosmology assumed in the LasDamas simulations differs from the one assumed in Guo et al. (2014). However, that will not affect the very small scale clustering enough to make the mock catalogue clustering measurements agree with the data measurement.

We see a clear difference between the clustering predicted from theory and what is actually observed in the data. The shape of the NFW model prediction is much more

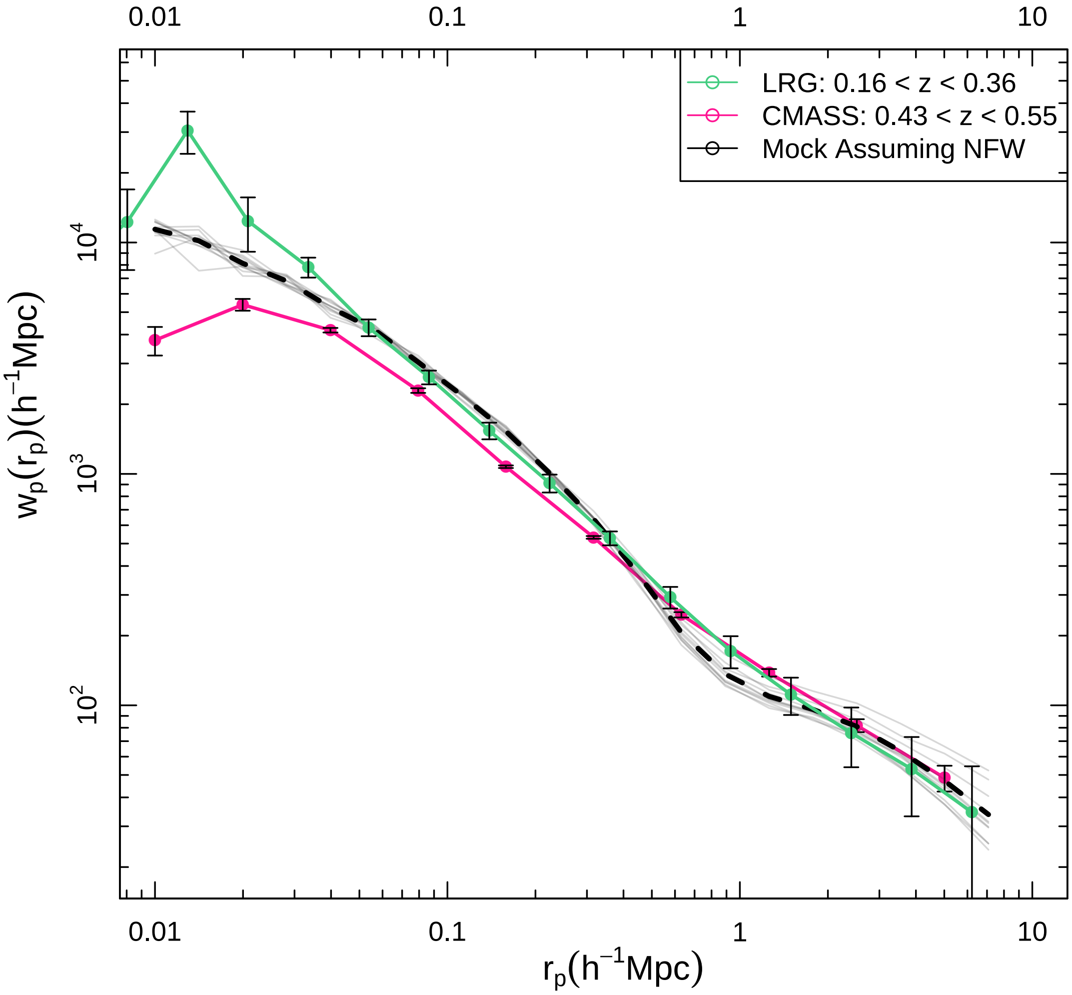

Figure 3.13 A comparison of the clustering of mock galaxy catalogues assuming an NFW profile with our data measurements. The two solid lines are same as in the previous figure. Each gray line is the measurement on an independent realizations and the heavy dotted black line is the mean of these measurements. NFW is unable to replicate the clustering of both the CMASS sample or the LRG clustering.

clustered at the smallest scales than the CMASS measurement, and does not exhibit the turnover that is observed. It is however less clustered than the LRG measurement, which is more in line with the findings in Chapter 2. We can theorize that the difference in shape is due to the actual ρ_{gal} of CMASS galaxies is much shallower than NFW. These systems are still assembling and the satellite galaxies have not yet fallen inward. Over time, satellites will fall inward causing the correlation function to steepen on small scales.

3.5 Summary & Discussion

We have presented the very small scale clustering of SDSS-II Data Release 7 Luminous Red Galaxies as well as SDSS-III BOSS CMASS galaxies. The clustering for the BOSS galaxies was measured in four separate redshift bins. We have validated both our code and method by successfully comparing to the previous results of M06 and Z05. We have also tested the cross correlation estimator on mocks. First, we compared the results of the cross correlation estimator to that of the LS estimator on mock galaxy catalogues without fiber collisions. We used a standard HOD present in the literature, dark matter halos from the LasDamas simulation and an assumption of an NFW profile for the galaxies to make our mocks. The cross correlation and LS estimator were virtually indistinguishable. Next, we tested how well the cross correlation estimator is at recovering the very small scale clustering in mocks with fiber collisions. In order to do this, we introduced fiber collisions in a similar way as is done by the data targeting algorithm. We do caution that we find a significantly fewer number of galaxies affected by fiber collisions in our mocks. This is expected because our mocks are not flux limited and do not contain any foreground stars or background quasars. We find that the cross correlation estimator does recover the very small scales clustering, and more effectively than the LS estimator.

The flattening in the very small scale clustering was observed in this work, something that has not been measured before. We find a significant evolution between the clustering of LRGs and CMASS galaxies on the very small scales, with the CMASS galaxies flattening out on small scales while the LRGs continue upwards in more of a power law like fashion. This small scale clustering behavior of LRGs has been measured before, but never compared to the galaxies thought to be their progenitors Tojeiro et al. (2012). We also see a trend, though less significant, between the clustering of CMASS galaxies in the redshift range of $0.43 < z < 0.55$ and $0.55 < z < 0.7$. It makes intuitive sense that there would be a difference in clustering given that the dynamical time between these bins is long enough for a satellite galaxy to infall, thus increasing the number of pairs at the smallest scales.

This downturn in clustering at early times implies a steepening of the galaxy distribution profile over cosmic time. As more satellite galaxies are accreted onto a system the number of pairs on very small scales increases, boosting the w_{p}. The dynamical time of a galaxy accreted at a redshift of $z = 0.55$, the median redshift of BOSS CMASS, is roughly $\tau_{dyn} \sim 1 Gyr$. Since the LRG sample has a median redshift of $z = 0.27$, there is enough time for an accreted satellite galaxy accreted sink into the inner part of a halo.

What is unclear however, and the focus of the follow up work, is whether this is a trend in redshift as stated above or the result of looking at a different class of galaxy. We also expect that different mass galaxies, with different number densities, occupy more or less rare density peaks and are such differently clustered. We see evidence for this at the large scale clustering of our two samples. In the two-halo term, started at separations of hundreds of kpc, the higher redshift bin is more clustered. This bin also has a lower number density, leading one to think that we are indeed looking at two different classes of galaxies, as opposed to the same class in two redshift bins.

The basic question we are asking is “Is the evolution in the downturn at the smallest scales due to comparing two different masses of galaxies or an evolution in the central-satellite pair separation of the same galaxies?” A simple way to check this would be adjust the number density of the lower redshift bin to that of the higher by only taking the brightest galaxies until the number densities were equal. This should fix the discrepancy at the largest scales, meaning that we are now comparing similarly rare density peaks in both redshift samples which results in the same large scale clustering. If we then still see the trend in downturn at the smallest scales, we can conclude that we are seeing a redshift evolution in the satellite pair separation. With some modeling, this result could be turned into a merger rate and an understanding of how both CMASS and LRGs are assembled.

Conversely, if once the large scale clustering is fixed there is no trend on the smallest scales we can conclude that the difference we see is mass related. This is also interesting because we do not expect to see a mass evolution in a sample that was specifically designed to be mass complete.

In conclusion, we have measured an evolution of the very small scale clustering of BOSS CMASS galaxies and SDSS-II LRGs. The very small scale clustering of BOSS CMASS galaxies shows shows an evolution in the turnover of very small scale clustering between $z = 0.7$ and $z = 0.5$. Both redshift ranges of BOSS CMASS galaxies exhibit significantly flatter very small scale clustering than the SDSS-II Luminous Red Galaxies at $z = 0.3$. Additionally we have shown that their clustering is most likely not fit by an NFW profile. Future work will focus on the physical explanation for this changing very small scale clustering, including modeling the changing distribution profile of massive satellite galaxies over cosmic time.

Chapter IV

CONCLUSIONS

Galaxy clustering on very small scales is driven by the spatial distribution of satellite galaxies– measuring the clustering means measuring where satellite galaxies are in their host dark matter halos. There is a direct connection between the distribution profile of satellites and the shape of clustering on scales less than one Megaparsec. This distribution profile has long been assumed to be the same as the host dark matter profile at these scales, but I have shown that this is not the case by measuring and modeling the very small angular clustering of Sloan Digital Sky Survey galaxies. Additionally, we see space for a trend with luminosity– a trend that has been observed before but neither with the precise accounting for observational systematics nor a more complete modeling of the galaxy distribution. It is not surprising that the satellite distribution profile is not necessarily consistent with that of the dark matter– where satellites are located in their host halos is dependent on a combination of complicated baryonic physics. More luminous satellite galaxies live a more turbulent life, which could explain why they have a steeper distribution profile. Modeling this trend required a large suite of cosmological N-body simulations of different volumes and a sophisticated massively parallel Bayesian fitting engine powered by petascale computing cluster. The inherent systematics and technical challenges in these measurements present opportunities to develop new techniques that can benefit the entire field, as opposed to just the very small slice of physics that it was developed for. I developed a technique that exploited regions on the sky with high redshift completeness in order to maximize the signal from the close pairs of galaxies that make up my measurement.

Once we turned our eyes to the higher redshift regime, we see evidence of a changing galaxy-halo connection with cosmic time. Higher redshift galaxy progenitors show a striking flattening in their clustering at very small scales, indicative of their systems having fewer satellite galaxies at earlier times. The steepening signal we see in the lower redshift galaxy sample is a signature of galaxies being gravitationally bound to more massive systems– eventually these satellites will fall to the center of their host dark matter halos and be accreted onto the central galaxy. Each redshift bin is a snapshot of the satellite galaxy distribution at that time, and combining these snapshots we have a better indication of how galaxies evolve. While it is unclear now whether these snapshots show the life cycle of the same type galaxy or the evolving halo occupation distribution of the same number density dark matter halo, this new measurement presents a vital building block in understand how the universe evolves.

E quindi uscimmo a riveder le stelle.

Dante Alighieri

REFERENCES

Abazajian, K. N., et al. 2009, ApJS, 182, 543

Aihara, H., et al. 2011, ApJS, 193, 29

Barnes, J., & Hut, P. 1986, Nature, 324, 446

Beifiori, A., et al. 2014, ApJ, 789, 92

Belokurov, V., et al. 2006, ApJ, 642, L137

Benson, A. J. 2010, Phys. Rep., 495, 33

Berlind, A. A., & Weinberg, D. H. 2002, ApJ, 575, 587

Berlind, A. A., et al. 2003, ApJ, 593, 1

—. 2006, ApJS, 167, 1

Bertschinger, E. 1998, ARA&A, 36, 599

Binney, J., & Tremaine, S. 1987, Galactic dynamics, ed. Binney, J. & Tremaine, S.

Blanton, M. R. 2006, ApJ, 648, 268

Blanton, M. R., Lin, H., Lupton, R. H., Maley, F. M., Young, N., Zehavi, I., & Loveday, J. 2003a, AJ, 125, 2276

—. 2003b, AJ, 125, 2276

Blanton, M. R., et al. 2003c, AJ, 125, 2348

—. 2005, AJ, 129, 2562

Bolton, A. S., et al. 2012, AJ, 144, 144

Bryan, G. L., & Norman, M. L. 1998, ApJ, 495, 80

Budavári, T., et al. 2003, ApJ, 595, 59

Bullock, J. S., Kolatt, T. S., Sigad, Y., Somerville, R. S., Kravtsov, A. V., Klypin, A. A., Primack, J. R., & Dekel, A. 2001, MNRAS, 321, 559

CDMS II Collaboration et al. 2010, Science, 327, 1619

Connolly, A. J., et al. 2001, in Mining the Sky, ed. A. J. Banday, S. Zaroubi, & M. Bartelmann, 323

Connolly, A. J., et al. 2002, ApJ, 579, 42

Cooray, A., & Sheth, R. 2002, Phys. Rep., 372, 1

Crocce, M., Cabré, A., & Gaztañaga, E. 2011, MNRAS, 414, 329

Davis, M., Efstathiou, G., Frenk, C. S., & White, S. D. M. 1985, ApJ, 292, 371

Dawson, K. S., et al. 2013, AJ, 145, 10

Del Popolo, A. 2012, MNRAS, 424, 38

Dodelson, S. 2003, Modern cosmology, ed. Dodelson, S.

Dunkley, J., et al. 2009, ApJS, 180, 306

Efstathiou, G., Bernstein, G., Tyson, J. A., Katz, N., & Guhathakurta, P. 1991, ApJ, 380, L47

Eisenstein, D. J. 2003, ApJ, 586, 718

Eisenstein, D. J., et al. 2001, AJ, 122, 2267

—. 2005, ApJ, 633, 560

—. 2011, AJ, 142, 72

Foreman-Mackey, D., Hogg, D. W., Lang, D., & Goodman, J. 2013, PASP, 125, 306

Fukugita, M., Ichikawa, T., Gunn, J. E., Doi, M., Shimasaku, K., & Schneider, D. P. 1996, AJ, 111, 1748

Fukushige, T., Kawai, A., & Makino, J. 2004, ApJ, 606, 625

Gao, L., Navarro, J. F., Cole, S., Frenk, C. S., White, S. D. M., Springel, V., Jenkins, A., & Neto, A. F. 2008, MNRAS, 387, 536

Geller, M. J., & Huchra, J. P. 1989, Science, 246, 897

Genel, S., et al. 2014, ArXiv e-prints

Gnedin, O. Y., Ceverino, D., Gnedin, N. Y., Klypin, A. A., Kravtsov, A. V., Levine, R., Nagai, D., & Yepes, G. 2011, ArXiv e-prints

Goodman, J., & Weare, J. 2010, Communications in Applied Mathematics and Computational Science, 5, 65

Groth, E. J., & Peebles, P. J. E. 1977, ApJ, 217, 385

Gunn, J. E., et al. 1998, AJ, 116, 3040

—. 2006, AJ, 131, 2332

Guo, H., Zehavi, I., & Zheng, Z. 2012a, ApJ, 756, 127

Guo, H., et al. 2014, ArXiv e-prints

Guo, Q., Cole, S., Eke, V., & Frenk, C. 2012b, MNRAS, 427, 428

Guth, A. H. 1981, Phys. Rev. D, 23, 347

Hamilton, A. J. S. 1998, in Astrophysics and Space Science Library, Vol. 231, The Evolving Universe, ed. D. Hamilton, 185–+

Hogg, D. W., et al. 2003, ApJ, 585, L5

Infante, L., et al. 2002, ApJ, 567, 155

Jenkins, A., Frenk, C. S., White, S. D. M., Colberg, J. M., Cole, S., Evrard, A. E., Couchman, H. M. P., & Yoshida, N. 2001, MNRAS, 321, 372

Jenkins, A., et al. 1998, ApJ, 499, 20

Jiang, T., Hogg, D. W., & Blanton, M. R. 2012, ApJ, 759, 140

Kaiser, N. 1987, MNRAS, 227, 1

Kazin, E. A., et al. 2010, ApJ, 710, 1444

Kerscher, M., Szapudi, I., & Szalay, A. S. 2000, ApJ, 535, L13

Kravtsov, A. V., Berlind, A. A., Wechsler, R. H., Klypin, A. A., Gottlöber, S., Allgood, B., & Primack, J. R. 2004, ApJ, 609, 35

Landy, S. D., & Szalay, A. S. 1993, ApJ, 412, 64

Linde, A. D. 1982, Physics Letters B, 108, 389

Ludlow, A. D., et al. 2013, MNRAS, 432, 1103

Maller, A. H., McIntosh, D. H., Katz, N., & Weinberg, M. D. 2005, ApJ, 619, 147

Mandelbaum, R., Seljak, U., & Hirata, C. M. 2008, , 8, 6

Maraston, C., Strömbäck, G., Thomas, D., Wake, D. A., & Nichol, R. C. 2009, MNRAS, 394, L107

Maraston, C., et al. 2013, MNRAS, 435, 2764

Martin, D. C., et al. 2005, ApJ, 619, L1

Masjedi, M., Hogg, D. W., & Blanton, M. R. 2008, ApJ, 679, 260

Masjedi, M., et al. 2006, ApJ, 644, 54

McBride, C. K., et al. 2014, In Prep

McCracken, H. J., Le Fèvre, O., Brodwin, M., Foucaud, S., Lilly, S. J., Crampton, D., & Mellier, Y. 2001, A&A, 376, 756

Merritt, D., Navarro, J. F., Ludlow, A., & Jenkins, A. 2005, ApJ, 624, L85

Mo, H., van den Bosch, F., & White, S. D. M. 2010, Galaxy Formation and Evolution, ed. Mo, H

Mo, H. J., & White, S. D. M. 1996, MNRAS, 282, 347

Narlikar, J. V., & Padmanabhan, T. 1991, ARA&A, 29, 325

Navarro, J. F., Frenk, C. S., & White, S. D. M. 1997, ApJ, 490, 493

Navarro, J. F., et al. 2004, MNRAS, 349, 1039

—. 2010, MNRAS, 402, 21

Newman, A. B., Treu, T., Ellis, R. S., & Sand, D. J. 2013, ApJ, 765, 25

Nierenberg, A. M., Auger, M. W., Treu, T., Marshall, P. J., & Fassnacht, C. D. 2011, ApJ, 731, 44

Nierenberg, A. M., Auger, M. W., Treu, T., Marshall, P. J., Fassnacht, C. D., & Busha, M. T. 2012, ApJ, 752, 99

Padmanabhan, N., et al. 2008, ApJ, 674, 1217

Padmanabhan, T. 1993, Structure Formation in the Universe, ed. Padmanabhan, T.

Parejko, J. K., et al. 2013, MNRAS, 429, 98

Peacock, J. A. 1999, Cosmological Physics, ed. Peacock, J. A.

Peacock, J. A., & Smith, R. E. 2000, MNRAS, 318, 1144

Peebles, P. J. E. 1980, The large-scale structure of the universe, ed. Peebles, P. J. E.

Press, W. H., & Schechter, P. 1974, ApJ, 187, 425

Pujol, A., et al. 2014, MNRAS, 438, 3205

Reed, D., Governato, F., Verde, L., Gardner, J., Quinn, T., Stadel, J., Merritt, D., & Lake, G. 2005, MNRAS, 357, 82

Reid, B., Padmanabhan, N., Tinker, J., Percival, W., Ho, S., & M., W. 2015, In Prep

Reid, B. A., Seo, H.-J., Leauthaud, A., Tinker, J. L., & White, M. 2014, ArXiv e-prints

Rubin, V. C., & Ford, Jr., W. K. 1970, ApJ, 159, 379

Ryden, B. 2003, Introduction to cosmology, ed. Ryden, B.

Sachs, R. K., & Wolfe, A. M. 1967, ApJ, 147, 73

Schlegel, D. J., Finkbeiner, D. P., & Davis, M. 1998, ApJ, 500, 525

Scoccimarro, R., Sheth, R. K., Hui, L., & Jain, B. 2001, ApJ, 546, 20

Scranton, R., et al. 2002, ApJ, 579, 48

Seljak, U. 2000, MNRAS, 318, 203

Seljak, U., & Warren, M. S. 2004, MNRAS, 355, 129

Smee, S. A., et al. 2013, AJ, 146, 32

Springel, V. 2005, MNRAS, 364, 1105

Strauss, M. A., et al. 2002, AJ, 124, 1810

Tal, T., Wake, D. A., & van Dokkum, P. G. 2012, ApJ, 751, L5

Tegmark, M., et al. 2004a, Phys. Rev. D, 69, 103501

—. 2004b, ApJ, 606, 702

The SDSS Collaboration et al. 2010, ArXiv e-prints

Tojeiro, R., et al. 2012, MNRAS, 424, 136

Totsuji, H., & Kihara, T. 1969, PASJ, 21, 221

Warren, M. S., Abazajian, K., Holz, D. E., & Teodoro, L. 2006, ApJ, 646, 881

Watson, D. F., Berlind, A. A., McBride, C. K., Hogg, D. W., & Jiang, T. 2012, ApJ, 749, 83

Watson, D. F., Berlind, A. A., McBride, C. K., & Masjedi, M. 2010, ApJ, 709, 115

Watson, D. F., Berlind, A. A., & Zentner, A. R. 2011, ApJ, 738, 22

Wechsler, R. H., Bullock, J. S., Primack, J. R., Kravtsov, A. V., & Dekel, A. 2002, ApJ, 568, 52

Weinberg, D. H., et al. 2007, in Bulletin of the AAS, Vol. 38, Bulletin of the AAS, 963–+

White, M., et al. 2011, ApJ, 728, 126

White, S. D. M., & Rees, M. J. 1978, MNRAS, 183, 341

York, D. G., et al. 2000, AJ, 120, 1579

Zehavi, I., et al. 2002, ApJ, 571, 172

—. 2004, ApJ, 608, 16

—. 2005a, ApJ, 621, 22

—. 2005b, ApJ, 630, 1

—. 2011, ApJ, 736, 59

Zel'dovich, Y. B. 1970, A&A, 5, 84

Zheng, Z., Coil, A. L., & Zehavi, I. 2007a, ApJ, 667, 760

—. 2007b, ApJ, 667, 760

Zheng, Z., et al. 2005, ApJ, 633, 791

Zwicky, F. 1937, ApJ, 86, 217

Appendix A

FIBER COLLISION INCOMPLETENESS

1.1 Cross Correlation Test

As we discussed in Chapter 2, our method of reducing the effects of fiber collisions consists of three parts. First, we restrict our samples to the plate overlap regions, where the vast majority of collided galaxies are recovered. Restricting our samples to the plate overlap regions guarantees that all cases of collision pairs are recovered. These are cases where there are only two galaxies within $55''$ of each other. Additionally, some cases of collision triplets are recovered. These are cases where one galaxy is within $55''$ of two other galaxies, but these others are not within $55''$ of each other. However, other cases of collision triplets are not recovered, such as when three galaxies are all within $55''$ of each other. In these cases, only two of the three galaxies in the triplet get measured redshifts. Naturally, higher order collision groups are also not fully recovered.

Second, for those few galaxies in collision triplets and higher multiplicity collision groups that do not get measured redshifts, we use the nearest-neighbor approximation and assign the collided galaxies the redshifts of their nearest neighbors on the sky. By studying close pairs that have been recovered in the plate overlap regions, we find that this correction works well roughly two-thirds of the time for SDSS Main galaxies. In other words, roughly two-thirds of the time, two galaxies that are closer than $55''$ on the sky are actually at the same redshift.

Third, we use the angular correlation function instead of the projected correlation function. This has the advantage that the angular separation of a given galaxy pair is completely unaffected by collisions, while the projected physical separation is not free of error. Since we are using angular instead of physical scales, a collided galaxy will only cause errors in our measurements if the nearest-neighbor redshift assignment causes the galaxy to drop out of or enter into the volume-limited sample in question. Errors in redshift that do not change a galaxy's membership in the sample (whether it was in or out) have zero effect on our measurements. Even when a collided galaxy's membership changes, there are really only three cases of collisions that can cause an error in our angular clustering measurements. The first occurs when the two galaxies straddle the redshift boundary of the volume-limited sample – i.e., one is inside the sample and the other is outside the sample – and the galaxy that is outside the sample did not get a measured redshift. The collided galaxy is then given the redshift of its neighbor and is thus brought into the sample. This error results in a small-scale pair that should not have been counted. The second failure mode occurs when the two collided galaxies straddle the luminosity limit of the sample, with the less distant galaxy being both outside the sample (i.e., below the luminosity limit) and the one that did not get a measured redshift. The collided galaxy is then moved to a larger distance and its calculated luminosity can now be high enough to bring it into the sample. This error also results in a small-scale pair that should not have been counted. The third and final failure mode occurs when the two collided galaxies are both in the sample, but the more distant one is close to the luminosity limit of the sample and does not get a measured redshift. When this galaxy is given the lower redshift of its neighbor, its calculated luminosity can make it drop out of the sample. This error results in a loss of a small-scale pair that should have been counted. All other cases of collisions result in no net gain or loss of a small-scale pair.

As a result of our methodology, only a very small fraction of all SDSS fiber collisions appear in our samples and only a small fraction of those that do appear actually cause errors on our measurements. Nevertheless, this small number of collided galaxies that do cause errors is not necessarily a negligible contribution to the number of galaxy pairs at very small scales. It is thus important to assess the magnitude of fiber collision errors in our analysis. The first test uses a cross-correlation test that is designed to maximize the fiber collision signal present in our samples. The second way is to approximate realistic mock galaxy catalogs that include fiber collisions so that we could directly test the extent to which our analysis method minimizes these errors. However, to be suitable for this purpose, the mock catalogs would have to cover the full flux-limited SDSS sample, making them a significant challenge to construct.

We first introduce a new low-redshift cut in each of our four volume-limited samples so that it does not overlap with any of the other samples. For example, we cut the $M_r <$ -21 sample at the outer redshift limit of the $M_r <$ -20 sample, we cut the $M_r <$ -20 sample at the outer redshift limit of the $M_r <$ -19 sample, and so forth. This results in four new volume-limited samples that have no spatial overlap with each other. We then measure the angular cross-correlation function between each of these new samples and the union of the other three new samples. In this way, each cross-correlation is measured between two samples that have no physical overlap and thus a minimal number of real physical correlated pairs, resulting in a measured cross-correlation that should be close to zero (it should actually be slightly higher than zero because there will be some real physical correlated pairs that straddle the redshift boundary between the samples). However, fiber

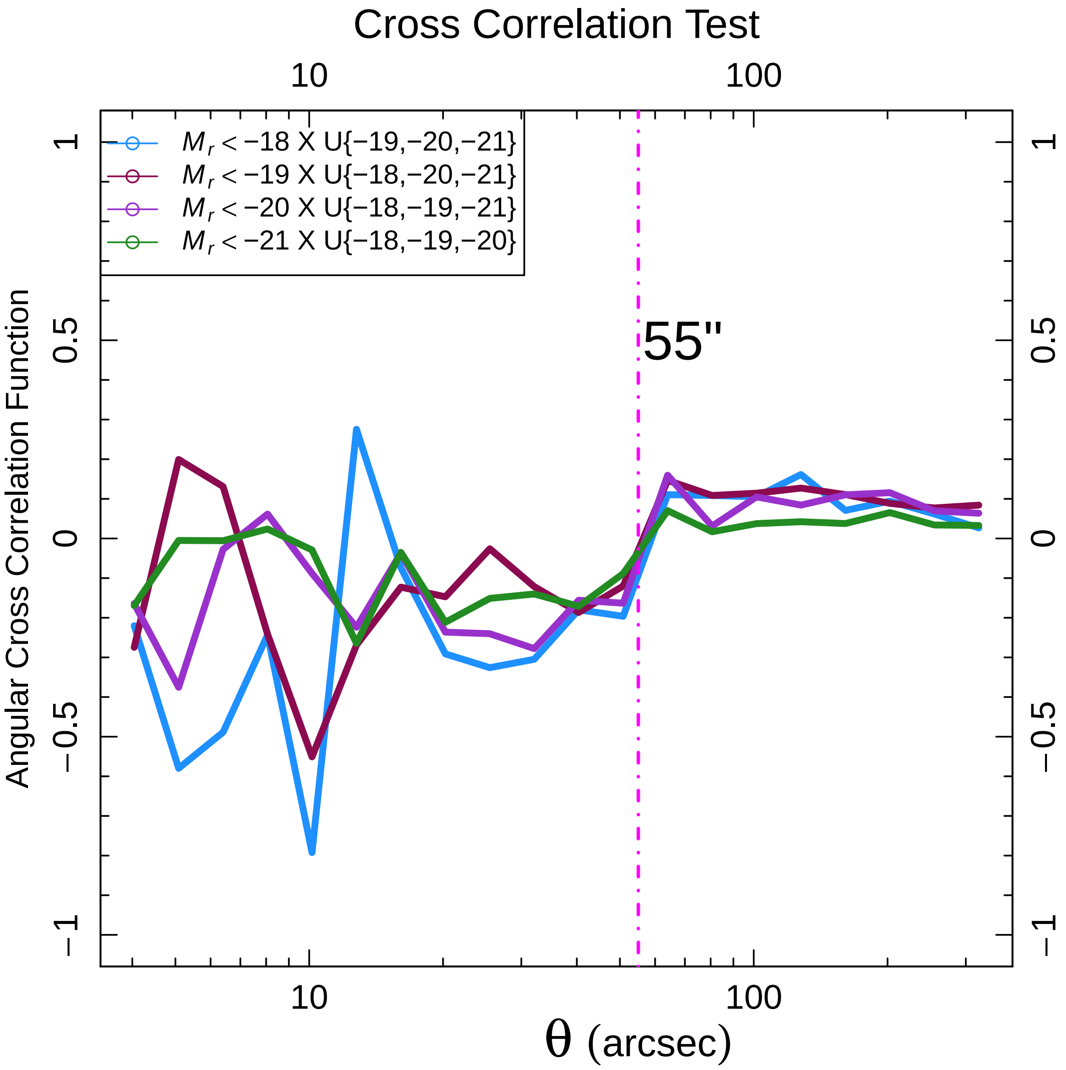

Figure 1.1 Cross-correlation test designed to probe the presence of fiber collision incompleteness in our volume-limited samples. The plot shows the angular cross-correlation function between sets of two samples that are designed to be spatially exclusive with each other. This is done by first introducing a low redshift cut in each of our four volume-limited samples so that no sample overlaps with another. We then measure the cross-correlation of each resulting sample with the union of the other three. With this setup there should be no real physical pairs in the cross-correlation except for erroneous pairs created due to fiber collisions. The 55″ collision scale is denoted by the vertical dashed line.

collision errors can cause a fake signal in this cross-correlation because these errors can move galaxies across sample boundaries, as we have discussed above. Furthermore, this signal should only appear below the collision scale of $55''$.

The results of this test are shown in Figure 1.1. The Figure shows exactly what we expect if fiber collision incompleteness is present in our samples. All four cross-correlations are slightly higher than zero for $\theta > 55''$, and they exhibit a sharp decrement below this scale. The transition occurs exactly where we expect it, between the two bins that straddle the collision scale. Moreover, the suppression of the cross-correlation function below this scale appears to be roughly scale-independent, at least as far as we can tell with the precision level of the measurements. Finally, the fiber collision signal is similar for all four samples, with perhaps some slight suggestion of a larger effect for less luminous samples.

The cross-correlation test proves definitively that our analysis methodology does not eliminate fiber collision incompleteness. This incompleteness is clearly present in the angular clustering of our samples. However, it is very difficult to translate the signal in this cross-correlation test into an estimate of the effect of fiber collisions on our auto-correlation measurements shown in Figures 2.3 and 2.7. The cross-correlation test maximizes the visible effect of fiber collisions in two ways. First, by removing any spatial overlap of the samples being cross-correlated, the unphysical pairs caused by fiber collisions become the entire signal, whereas in the auto-correlation measurements they are only a tiny fraction of the signal. Second, fiber collisions can only cause a deficit of pairs in this cross-correlation test, whereas they can both add or subtract pairs in the auto-correlation function, as we argued above. This adding and subtracting of pairs in the auto-correlation function could result in a smaller net effect due to fiber collisions. In the cross-correlation test, the only case of collisions that affects the measurement is when the two collided galaxies straddle the redshift boundary between the two samples. Regardless of which of these galaxies receives a redshift, the nearest-neighbor correction results in both galaxies ending up in the same sample, which removes the pair from the cross-correlation. This is why we see a deficit below the collision scale in Figure 1.1. The set of collisions affecting the cross-correlations is thus quite different from that affecting the auto-correlations, making it difficult to translate between the two.

We interpret the lack of any obvious visible features at the collision scale in our auto-correlation measurements as evidence that fiber collision errors in these measurements must be small relative to the real physical signal. Figure 1.1 demonstrates that these features should appear between the two bins that straddle $55''$, exactly as we expect. The various discontinuities seen in Figure 2.3 occur at other scales and are thus not likely caused by fiber collisions. The only exception to this is the slight discontinuity seen in the auto-correlation of the M_r <-19 sample. However, the magnitude of this discontinuity is consistent with the up and down fluctuations seen at other scales. Moreover, there does not appear to be any systematic enhancement or suppression at all scales less than $55''$, as is seen in the cross-correlation test. Errors due to fiber collisions are thus likely small relative to the real physical signals present in our auto-correlation measurements. Nevertheless, we emphasize that fiber collision errors are definitely present in our measurements at some level and it is not easy to asses their exact impact on our modeling results. Though we do not expect this impact to be large, it is prudent to treat our derived parameter values with some degree of caution.

1.2 Test on Mocks

It is important to understand how many galaxies pairs come from single pair collision groups, which will be recovered in the overlap, as opposed to higher order groups, which will not. We use mock galaxy catalogues to test the percentage of pairs that are recovered. Figure 1.2 shows how different sized collision groups affect $\omega(\theta)$. We measure $\omega(\theta)$ on a full mock sample without any fiber collisions. The group finder is then employed to introduce fiber collisions. Next, we allow all galaxies in pairs to have redshifts. This is equivalent to the entire sky being tiled twice. And finally, we allow all triplets to have redshifts which would be equivalent to the entire sky being tiled thrice. The overlap region recovers most of the counts that we are missing, with the double overlap sample recovering almost all. A very small percentage of counts are from groups with four or more collided galaxies. An important point is that the effect of fiber collisions goes out to higher separations than the fiber collision limit. We also test how the nearest neighbor correction affects the slope of $\omega(\theta)$ in the overlap region. The effect is small and does not effect our results due to the small number of pairs in triplet or higher collision groups.

Fiber collisions affect all pairs (or groups) of galaxies separated by less than its angular scale. Losing one galaxy to a fiber collision will also affect pairs on all scales. There are three types of galaxy-galaxy pairs in a measurement of $\omega(\theta)$ with a fiber collision correction:

1. DD_{true}, the physical galaxy pairs

2. DD_{recov} are correlated pairs of galaxies that have been recovered by a correction

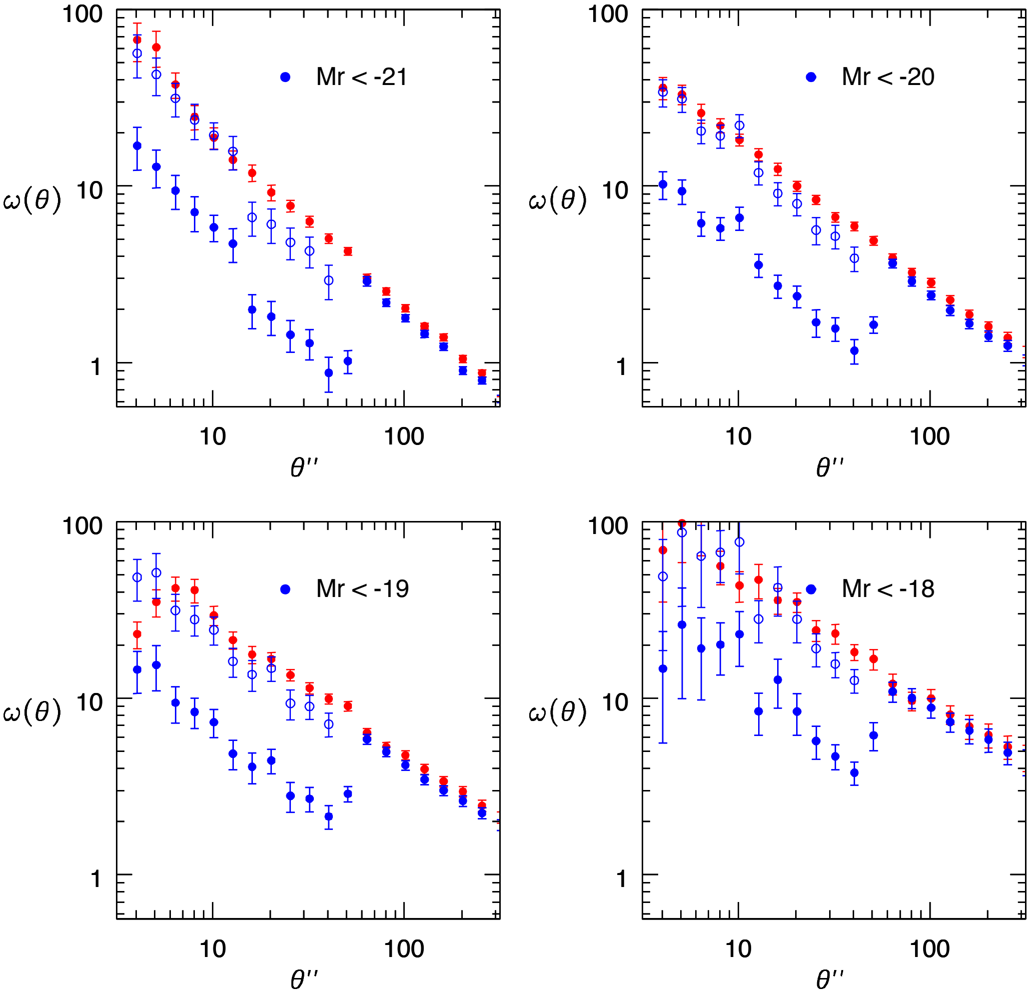

Figure 1.2 A comparison $\omega(\theta)$ for a sample of galaxies with measured redshifts (blue points) and a sample of galaxies using the 'nearest neighbor' correction. This correction assigns the redshift of the nearest neighbor galaxy to the galaxy that has collided with it. The drastic drop is at the fiber collisions limit and any pairs within that limit for the blue points are in overlap regions. The open points are the blue points when the missing area of the footprint is taken into account.

3. DD_{false} uncorrelated pairs of galaxies included in the samples by a correction

Figure 1.2 shows $\omega(\theta)$ for two samples of galaxies. The first sample, blue solid points, only has DD_{true} pairs less than the fiber collision limit. These pairs come from about one third of the SDSS footprint that was tiled twice. We account for this different in footprint by multiplying the blue points by the missing area, shown by the open blue points in Figure 1.2. We then apply a nearest neighbor correction to the missing redshifts shown by the red points in Fig 1.2. In this correction, a group finder is run on the galaxy target sample to maximize the number of galaxies that receive redshifts. If the corrected $\omega(\theta)$ did not include DD_{false} pairs, the red points would completely agree with the open blue points. This is not the case and the deviation between the two is a function of angular separation continuing further out than the fiber collision scale.

These reasons motivate our method of trimming the sample to the overlap region which recovers most of the missing counts while still using the nearest neighbor correction on the remaining galaxies. This is a compromise between having high enough number counts but being confident that the pairs we measure are correct.

Appendix B

SELECTION FOR SDSS-III BOSS CMASS GALAXIES AND LUMINOUS RED GALAXIES

```
#!/usr/bin/python
# Selection for BOSS CMASS Galaxies



import pyfits
import numpy as np
from astropy.cosmology import FlatLambdaCDM
cosmo=FlatLambdaCDM(H0=100,Om0=0.266)

## Read in the fits file
hdulist = pyfits.open('/hd0/Research/Clustering/Boss/dr11/dr11v2
    /cmass-dr11v2-N-Anderson.dat.fits')



## Uncomment the following line to view the info about the table
print hdulist.info()

```

```
## Read the tabular portion of the fits file into the variable '
    table'.  This assumes that the table of interest is located
    in extension 1
table = hdulist[1].data



ra=table.field('RA')
dec=table.field('DEC')
redshift=table.field('Z')
polygon=table.field('IPOLY')
extinction=table.field('EXTINCTION')
frac=table.field('FRACPSF')
exponential_flux=table.field('EXPFLUX')
dev_flux=table.field('DEVFLUX')

extinction_g=extinction[:,1]
extinction_r=extinction[:,2]
extinction_i=extinction[:,3]

fiberflux=table.field('FIBER2FLUX')
fiberflux_i= 22.5-2.5*np.log10(fiberflux[:,3]) - extinction_i
```

```
f_i_cmod=(1 - frac[:,3])*exponential_flux[:,3] + frac[:,3]*
    dev_flux[:,3]
i_cmod=22.5-2.5*np.log10(f_i_cmod) - extinction_i

modelflux=table.field('MODELFLUX')
modelflux_g= 22.5-2.5*np.log10( modelflux[:,1]) - extinction_g
modelflux_r= 22.5-2.5*np.log10( modelflux[:,2]) - extinction_r
modelflux_i= 22.5-2.5*np.log10( modelflux[:,3]) - extinction_i


d_perp=(modelflux_r-modelflux_i) - (modelflux_g-modelflux_r)/8.



i_cmod_cut= 19.86 + 1.6*(d_perp - 0.8)
d_perp_cut=0.55


weight_cp=table.field('WEIGHT_CP')
icollided=table.field('ICOLLIDED')

distance_modulus=cosmo.distmod(redshift)
```

```
Mag_i=i_cmod - distance_modulus - (-0.5)



array=np.column_stack((ra,dec,redshift,weight_cp,polygon,i_cmod,
    modelflux_g,modelflux_r,modelflux_i,fiberflux_i,
    distance_modulus,Mag_i))
dimensions=str(”ra␣dec␣redshift␣weight_cp␣polygon␣i_cmod␣
    modelflux_g␣modelflux_r␣modelflux_i␣fiberflux_i␣
    distance_modulus␣Mag_i”)



np.savetxt(’/hd0/Research/Clustering/Boss/dr11/dr11v2/dr11v2_all
    .out’,array,delimiter=’\t’,newline=’\n’,header=str(dimensions
    ),comments=’␣’)



#this really isn’t necessary but it’s here anyway

ids=np.where(( i_cmod > 17.5   ) &
           ( i_cmod < 19.9) &
           ( modelflux_r -  modelflux_i < 2 ) &
           ( d_perp > d_perp_cut) &
```

```
        ( fiberflux_i < 21.5 ) &
        ( i_cmod < i_cmod_cut))

array_filter=array[ids]

bin=(0.43,0.5,0.55,0.6,0.7)

np.savetxt('/hd0/Research/Clustering/Boss/dr11/dr11v2/
   dr11v2_imaging.txt',array_filter,delimiter='\t',newline='\n')



path=('/hd0/Research/Clustering/Boss/dr11/dr11v2/')


for x in range(0,4):
        array_list=array_filter[np.where((array_filter[:,2] >
          bin[x]) & (array_filter[:,2] < bin[x + 1]))]
        np.savetxt(str(path) + "bin" + str(x+1) + "_selection.
          txt",array_list,delimiter='\t',newline='\n')

```

```
array_output=array_filter[:,range(0,5)]


np.savetxt('/hd0/Research/Clustering/Boss/dr11/dr11v2/Di.dr11v2.
    out.selection',array_output,delimiter='\t',newline='\n')



for x in range(0,4):
        array_list=array_output[np.where((array_filter[:,2] >
            bin[x]) & (array_filter[:,2] < bin[x + 1]))]
        np.savetxt(str(path) + "bin" + str(x+1) + "
            _Ds_dr11v2_selection.out",array_list,delimiter='\t',
            newline='\n')
```

Selection for SDSS-II Luminous Red Galaxies

```
SELECT
  ra,
  dec,
  petroMag_r,
  extinction_r into mydb.lrg_dr7_primtarget_selection from
     photoObj
  WHERE ( PrimTarget=32 ) AND
    petroMag_r < 19.2 AND
    (psfMag_r - modelMag_r) > 0.3 AND
```

```
(r - i) - (g - r)/4.0 - 0.18 < 0.2 AND
(r - i) - (g - r)/4.0 - 0.18 > -0.2 AND
petroMag_r < 13.1 + (0.7*(g - r) + (1-0.7)*4.0*((r - i) -
    0.18))/0.3 AND
petroMag_r + 2.5*log10(2*3.14159*petroR50_r*petroR50_r) <
    24.2
```