Rosella: A mock catalogue of galaxy luminosities, colours and positions for cosmology

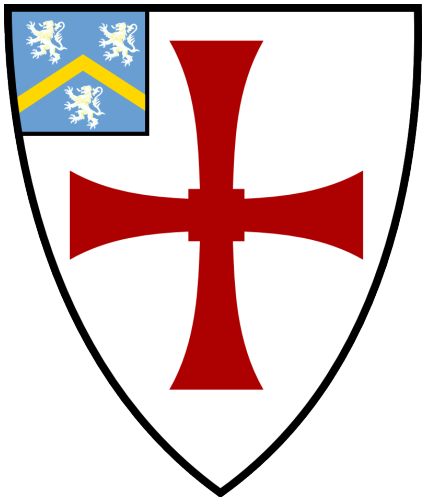
ALEKSANDRA SAFONOVA

Rosella: A mock catalogue of galaxy luminosities, colours and positions for cosmology

Sasha Safonova

A thesis presented for the degree of

Master of Science by Research



Institute for Computational Cosmology

Department of Physics

The University of Durham

United Kingdom

August 2019

Dedicated to

Margarita Safonova,

who devoted her life to the search for controlled thermonuclear fusion

and taught me that physics adds a game to any situation.

Rosella: A mock catalogue of galaxy luminosities, colours and positions for cosmology

Sasha Safonova

Abstract

The scientific exploitation of the Dark Energy Spectroscopic Instrument Bright Galaxy Survey (DESI BGS) data requires the construction of mocks with galaxy population properties closely mimicking those of the actual DESI BGS targets. We create a high fidelity mock galaxy catalogue that can be used to interpret the DESI BGS data, as well as meeting the precision required for the tuning of thousands of approximate DESI BGS mocks needed for cosmological analyses with DESI BGS data. The mock catalogue uses subhalo abundance matching (SHAM) with scatter to populate the P-Millennium N-body simulation with galaxies at the median DESI BGS redshift of ~0.2, using formation redshift information to assign ${}^{0.1}(g-r)$ rest-frame colours. The mock provides information about r-band absolute magnitudes, ${}^{0.1}(g-r)$ rest-frame colours, 3D positions and velocities of a complete sample of DESI BGS galaxies in a volume of (542 Mpc/h)^3. This P-Millennium DESI BGS mock catalogue is ideally suited to the tuning of approximate mocks unable to resolve subhalos that DESI BGS galaxies reside in, to test for systematics in analysis pipelines and to interpret (non-cosmological focused) DESI BGS analysis.

Supervisors: Peder Norberg and Shaun Cole

Acknowledgements

I owe an immense thanks to my research supervisors, Peder Norberg and Shaun Cole, for empowering me to create the first mock catalogue of my career. My supervisors have consistently emphasized the importance of my well-being over the course of my postgraduate studies and have gone out of their way to ensure that I had everything necessary for a productive and happy year at Durham. Time and again, they have provided guidance and access to computational resources at lightning speed. For every piece of writing that I submitted to them, my supervisors provided me with detailed and helpful feedback, and I owe my development as an effective researcher to their attentive advising. They never assumed that I possessed advanced knowledge and patiently walked me through concepts that were fundamental, but new to me.

I thank the brilliant researchers in our department who taught me about astrophysics outside my immediate field. John Helly, Alex Smith, and Miguel de Icaza-Lizaola have provided useful knowledge in many discussions. Thanks to Jaime Forero-Romero for creating a tutorial on fiber assignment for Rosella data.

I would like to extend special thanks to my parents Nikolay, Olga and Olena, grandparents Margarita, Aleksandr, Nina, and Viktor, my friends Nam, Linda and Netti, and fellow postgraduates who have shared in my excitement about research and extended information and reassurance that have empowered my work.

I thank the U.S.-U.K. Fulbright Commission for providing the funding I needed to enter the Master's program and spend a year in Durham. Finally, I thank the team behind the COSMA supercomputer and the P-Millennium simulation, without whom this project would not have been possible.

Contents

Declaration **vii**

List of Figures **viii**

List of Tables **x**

1 Introduction **1**

1.1 Challenges facing modern cosmology 1

1.2 The importance of galaxies to cosmic questions 5

1.2.1 Galaxy colour bimodality 6

1.3 Large-scale structure observations today 9

1.3.1 Measuring galaxy redshifts 9

1.3.2 A brief history of modern cosmological surveys 11

1.4 Connecting theory to observations in cosmology 12

1.4.1 The role of mock data 12

1.4.2 The computational gap 15

1.5 Populating dark matter-only simulations with galaxies 16

1.5.1 Biased Dark Matter 17

1.5.2 Halo occupation distribution (HOD) 17

1.5.3 Conditional luminosity function 19

1.5.4 Subhalo abundance matching (SHAM) 20

1.5.5 SHAM with galaxy colour assignment 23
1.5.6 Semi-analytic models (SAMs) 24
1.6 Aim of this research 25
1.7 Outline of this thesis 25

2 Methodology behind Rosella's construction **27**
2.1 The P-Millennium Simulation 27
2.2 The choice to apply SHAM to P-Millennium for DESI BGS 30
2.3 Algorithm for luminosity assignment 31
2.4 Tracing subhalo histories across P-Millennium snapshots 36
2.5 Definition of formation redshift 39
2.6 Algorithm for colour assignment 40
2.6.1 Luminosity-dependent galaxy colour distributions 41
2.6.2 Cumulative v_{peak}-dependent distribution of z_{form} 44
2.6.3 Colour assignment 45
2.7 Identifying central galaxies 48
2.8 Clustering 48
2.9 Summary of the methodology 49

3 Properties of the Rosella mock catalogue **50**
3.1 Choice of redshift for the mock catalogue 50
3.2 Galaxy luminosity function and Rosella resolution 51
3.3 Distribution of central and satellite galaxies 55
3.4 Galaxy clustering 59
3.4.1 Clustering as a function of luminosity 61
3.4.2 Clustering as a function of colour 61
3.5 Galaxy colour bimodality 64
3.6 Tuning the models of luminosity and colour assignment 64

4 Conclusion **71**
4.1 Extending this technology to new purposes 72

4.2 Deepening the approach to tuning the mock 73
4.3 Prospects of treating centrals and satellites separately 74
4.4 Comparison to hydrodynamic simulations 75
4.5 Access to developed code 76

Bibliography **77**

Appendix A **85**

Appendix B **87**

Declaration

The work in this thesis is based on research carried out at the Institute for Computational Cosmology, Department of Physics, University of Durham, England. No part of this thesis has been submitted elsewhere for any other degree or qualification, and it is the sole work of the author unless referenced to the contrary in the text.

List of Figures

1.1 The colour–magnitude relation of galaxies in the SDSS 8

1.2 Comparison of the galaxy distributions in spectroscopic surveys and mock catalogues 14

2.1 Illustration of SHAM without scatter 33

2.2 Illustration of scatter added to SHAM data 35

2.3 Scatter as a function of absolute magnitude. 36

2.4 Example histories of maximum circular velocities across the time of subhalo existence. 37

2.5 Example of a formation redshift search for a single subhalo. 40

2.6 Gaussian distributions of ${}^{0.1}(g-r)$ for red and blue galaxy populations 42

2.7 Cumulative distribution functions of ${}^{0.1}(g-r)$ 43

2.8 Expected fraction of blue galaxies and mean ${}^{0.1}(g-r)$ as a function of absolute magnitude. 45

2.9 Cumulative distribution functions of formation redshift in bins of $\mathrm{v}_{\mathrm{peak}}$. 46

2.10 Illustration of colour assignment for a single subhalo 47

3.1 Luminosity function of the mock catalogue. 52

3.2 SHAM absolute magnitudes with scatter. 53

3.3 Histograms of $\mathrm{v}_{\mathrm{peak}}$ in bins of M_r^h 54

3.4 Fraction of central galaxies in Rosella with and without scatter. 56

3.5 Distribution of central and satellite galaxies in halo mass bins. 57
3.6 Relation between subhalo v_{peak}, galaxy M_r^h, and halo mass for SHAM with no scatter 58
3.7 Relation between subhalo v_{peak}, galaxy M_r^h, and halo mass for SHAM with nominal scatter 59
3.8 Relation between subhalo v_{peak}, galaxy M_r^h, and halo mass for SHAM with a step scatter function. 60
3.9 Projected correlation function for luminosity threshold galaxy samples. . 62
3.10 Projected correlation function for red and blue galaxies. 63
3.11 Distribution of $^{0.1}(g-r)$ values among Rosella galaxies 65
3.12 Colour-magnitude diagram of Rosella galaxies 66
3.13 Projected correlation function for luminosity-limited samples for SHAM without scatter. 67
3.14 Projected correlation function for SHAM with a step scatter function . 68
3.15 Scatter step function based on Mccullagh et al. (2017). 69
3.16 Absolute magnitude-v_{peak} relation for SHAM with a smooth scatter function 69

B.1 *Rosella* the bird, our catalogue's eponym. 87

List of Tables

2.1 Selected cosmological parameters of the P-Millennium simulation. . . . 28

2.2 Subhalo properties provided in Rosella's subhalo history datasets. . . . 38

A.1 Descriptions of selected datasets available in P-Millennium DHALO output files. 85

A.2 Descriptions of selected datasets available in P-Millennium FoF and SUBFIND output files. 86

Chapter 1

Introduction

1.1 Challenges facing modern cosmology

What separates a scientific theory from speculation? The scientific method demands that a good scientific theory must make predictions that can be tested with real-world observations (Hawking, 1996). Theory and experiment go hand-in-hand, and, for centuries, humans have relied on this partnership to reach beyond uninformed intuition towards a more profound understanding of this world. It is remarkable, therefore, that today's most robust cosmological observations consistently contradict predictions made with theories that have stood many tests to date. The surprising discovery originates with the observation that the expansion of the universe is accelerating, a phenomenon first shown with supernovae in Perlmutter et al. (1999) and Riess et al. (1998) and that holds cosmologists' concentrated attention today. This discovery contradicts the scenarios that cosmologists had been predicting prior to 1998 based on the theory of general relativity for a matter-filled, flat universe. Some models, such as Efstathiou et al. (1990), had noted the need for a model that could account for cosmic acceleration based on analysis of existing data.

It was not long after the formulation of general relativity that cosmologists discovered the expansion of the universe. The early twentieth century brought about

the discovery of the redshifted* universe (Slipher, 1917; Hubble, 1929), in which all galaxies (disregarding local peculiar motion) are receding away from the Milky Way with the velocity $\vec{v}$:

$$\vec{v} = H(t)\ \vec{r} \tag{1.1}$$

$H(t)$ defines the rate at which that galaxy recedes as a function of redshift, and $\vec{r}$ is that galaxy's distance away from the observer (Lemaître, 1927; Hubble & Humason, 1934). Equation 1.1 and the parameter $H(t)$ have been named after one of the predominant extragalactic astronomers of the early twentieth century, as the Hubble Law and the Hubble Parameter, respectively.

$H(t)$ is one of the fundamental *cosmological parameters*, and it serves as a measure of the rate of expansion of the universe. Attempts to measure the Hubble parameter (and its present-day value, the Hubble constant H_0) began with the early works of Lemaître (1927) and Hubble (1929). Efforts to measure $H(t)$ continue, as we will discuss in section 1.3.

The Hubble law can be derived theoretically by making two assumptions that are central to contemporary cosmology, known by the collective name "the cosmological principle" and supported by observations of the cosmic microwave background (CMB) and the large-scale structure of the universe:

1. The universe is homogeneous: Observers at any randomly chosen points in space should see the same picture of the universe.

2. The universe is isotropic: To an observer in a given point in space, the universe should appear the same regardless of the direction of observation.

Supporting these assumptions has not been historically trivial. Homogeneity and isotropy, along with a static state, were the outcomes of solutions to Einstein's field equations (Einstein, 1917). These assumptions have been questioned, when, for example, Charlier (1922) pointed out that galaxies were unevenly clustered on

*We discuss redshifts in greater detail in section 1.3.1

the sky, which went against the assumption of homogeneity. Lemaître showed the static state result of Einstein's field equations to be unstable in 1927 and 1931.

The debates surrounding general relativity and the discovery of galaxies receding from the Milky Way demonstrated the importance of informing theoretical work in cosmology with observational details, such as systematic errors. The impact of experimental details on theory can be noted in Lemaître (1927) [translation mine]*:

> Some authors have sought to specify the relation between v and r and have obtained nothing more than a very weak correlation between these two quantities. The error in the determination of individual distances is of the same order of magnitude as the distance covered by the observations [...] All that this lack of observational precision allows us to do is assume v to be proportional to r and to try to avoid a systematic error in the determination of the v/r relation.

Theoretical efforts to draw a line between Einstein's relativistic equations and observations of galaxies outside our own led several scientists prior to Lemaître (1927) to abandon what is now known as Hubble Law, a concept so central to today's cosmological thinking that one might take it for an axiom. From the beginning of modern observational cosmology, thus, maintaining a connection between theory and observations has meant the difference between making and missing discoveries that define our field.

*The full comment on the value of *H(t)* in the original text in Lemaître (1927) reads:

> En ne donnant pas de poids aux observation, on trouverait 670 Km./sec à 1,16 $\times$ 10^6 parsecs, 575 Km./sec à 10^6 parsecs. Certains auteurs ont cherché à mettre en évidence la relation entre v et r et n'ont obtenu qu'une très faible corrélation entre ces deux grandeurs. L'erreur dans la détermination des distances individuelles est du même ordre de grandeur que l'intervalle que couvrent les observations et la vitesse propre des nébuleuses (en toute direction) est grande (300 Km./sec. d'après Strömberg), il semble donc que ces résultats négatifs ne sont ni pour ni contre l'interprétation relativistique de l'effet Doppler. Tout ce que l'imprécision des observations permet de faire est de supposer v proportionnel à r et d'essayer déviter une erreur systématique dans la détermination du rapport v/r. Cf. Lundmark. The determination of the curvature of space time in de Sitter's world M. N., vol. 84, p. 747, 1924, et Strömberg, *l.c.*

The interplay between theory and observations plays a major role in investigations of the puzzling phenomenon at the centre of today's research in cosmology: the acceleration of cosmic expansion. Understanding cosmic acceleration on a profound level could shed light on the relationship between gravity and the quantum vacuum, the possible existence of extra spatial dimensions, or the nature of quantum gravity (Weinberg et al., 2013).

While cosmologists have been producing a suite of models that could account for accelerated expansion (e.g. Martel et al., 1998; Zlatev et al., 1999; Caldwell & Linder, 2005), practical observations are necessary for us to select the most viable theories and refine them. Without experiments, the potential new physics that lies behind accelerated expansion stays outside our reach.

A major caveat, however, complicates cosmological observations: Modern cosmological models dictate that only about 5% of the energy in the universe is made of matter that interacts with radiation and gravity (Planck Collaboration et al., 2018). The commonly used model ΛCDM suggests that cold dark matter contains about 1/4 of the energy density in the universe, and a cosmological constant, Λ, accounts for the remaining 7/10 of the energy density in the universe (Turner et al., 1984; Davis et al., 1985; Efstathiou et al., 1990; Riess et al., 1998; Perlmutter et al., 1999; Planck Collaboration et al., 2018).

Cosmologists, thus, face a world, 95% of which cannot be directly observed. Understanding the nature of these "unobservable" components of the universe – dark matter and the cosmological constant Λ, also known by the name "dark energy"* – poses a set of challenges but presents a range of scientific possibilities.

Statistics offers a gateway towards information that cannot be directly observed. The power of statistics lies in uncovering phenomena that are too subtle or unobservable on the level of individual objects, like galaxies or supernovae. With

*The term "dark energy" was introduced into the literature in Huterer & Turner (1999). It can refer to the component of the universe which drives its accelerated expansion in a variety of models, including but not limited to the cosmological constant Λ.

the power of statistics, however, cosmologists create a wide range of approaches to thinking about the universe – approaches that include, for example, mapping dark matter with weak gravitational lensing and locating the early universe's fingerprint, baryon acoustic oscillations (BAO).

Understanding the nature of cosmic acceleration and dark matter offers great scientific potential, which has inspired a range of observational and theoretic efforts. Modern cosmological surveys aimed at probing this phenomenon propose to do so by measuring the expansion history and growth of large-scale structure (Weinberg et al., 2013). The observational projects that aim to understand cosmic acceleration and the nature of the "dark" components of the universe promise to gather unprecedented amounts of data. Thanks to the rise in large, collaborative surveys, cosmology has entered the golden age of statistics.

The unparalleled amount of information that cosmologists are beginning to gather, however, has pointed out limitations in the computational and data processing power currently available to the scientific community. In order to study the components of the universe that are invisible to the eye but filled with scientific potential, we must face three major hurdles: gathering data, processing the data, and, finally, making sense of what we have observed.

1.2 The importance of galaxies to cosmic questions

The large-scale cosmic structures that make up the present-day universe grew out of fluctuations in the distribution of matter that were created during inflation, a period that lasted only a fraction of the first second of the universe's existence (Guth & Pi, 1982). Over time, density fluctuations grew under the influence of gravity. Initially, gas was mixed with dark matter. As the universe evolved, dark matter halos formed, and gas collapsed into their centres. Eventually, the collapsed gas cooled into discs, where stars could form into protogalaxies (e.g. Cole et al., 2000; Wechsler & Tinker, 2018).

Dark matter serves as the three-dimensional fabric on which galaxies form, evolve, and interact. Conversely, galaxies can serve as visible tracers of the location of dark matter, a possibility that becomes useful in cosmological surveys.

The main unit of the connection between dark matter and visible matter is a dark matter halo, which accounts for the majority of the mass of any gravitationally bound region of space that contains one or more galaxies. The galaxies occupying a halo exist in substructures of the main halo, called dark matter subhalos. The assumption that every galaxy has a distinct subhalo serves as the premise behind the subhalo abundance matching (SHAM) method of populating N-body simulations with galaxies, which we discuss in section 1.5.4 (e.g. Vale & Ostriker, 2004; Kravtsov et al., 2004; Conroy et al., 2006; Chaves-Montero et al., 2016).

Galaxies evolve by forming stars and merging with other galaxies as a result of the mergers of their host dark matter halos (e.g. Wechsler & Tinker, 2018). One, consequently, expects the spatial distribution, intrinsic properties and evolution of galaxies to be interlaced with those of their host dark matter halos.

The matter power spectrum indicates that cosmic structure forms hierarchically, with smallest objects forming first and growing as time passes. The formation history of dark matter halos can schematically be described by "merger trees", which play an important role in theories of galaxy formation built around hierarchical structure formation (e.g. Lacey & Cole, 1993; Mo et al., 2010).

Galaxies are home to the universe's visible, luminous baryonic matter. In ΛCDM, dark matter is arranged in halos, and galaxies are located within them. Exploiting the connection between these halos and the galaxies that they contain can help us answer today's most pressing problems in cosmology.

1.2.1 Galaxy colour bimodality

Galaxy colours, as measured by the difference between magnitudes in two filters, reflect the composition of galaxies' stellar populations. This relationship between

galaxy colours and their stellar populations is reflected in their morphology (e.g. Morgan & Mayall, 1957; Chester & Roberts, 1964; Roberts & Haynes, 1994). As a result, galaxies can be classified as red or blue, which naturally relates with their star formation and the evolution of their metal content (e.g. Mo et al., 2010).

Local-universe observations have shown that galaxies present a bimodal colour distribution (e.g. Strateva et al., 2001; Baldry et al., 2006; Cassata et al., 2008). This bimodality appears to be independent of the environment in which galaxies reside (e.g. Hogg et al., 2004; Baldry et al., 2006). Reproducing this bimodality in simulated data, and understanding the environment's role in it are important goals for galaxy formation and evolution research (e.g. Trayford et al., 2015; Nelson et al., 2018).

The bimodality of the galaxy population has been demonstrated observationally, for instance in the colour–magnitude relation of SDSS galaxies in figure 1.1. The distribution in figure 1.1 comprises two galaxy populations that can be distinguished by eye. These populations are termed the red sequence (located in the higher ${}^{0.1}(g-r)$ range on the colour-magnitude diagram) and the blue sequence (distributed in the range of lower values of ${}^{0.1}(g-r)$). It should be noted that the red and blue populations physically form a continuum with two apparent groups rather than discrete bins of galaxies.

Red-sequence galaxies dominate the bright range (smaller ${}^{0.1}M_r$ - 5 log h values) of the colour-magnitude diagram. Conversely, the majority of galaxies on the fainter end of this diagram are blue. This is related to early-type galaxies dominating the bright end of the galaxy luminosity function, and the majority of late-type galaxies occupying the faint end of the luminosity function (e.g. Mo et al., 2010).

If we consider figure 1.1's contours that roughly correspond to the red and blue sequences separately, we can see that in both cases, the brighter a galaxy is, the redder we can expect its colour to be. Studies of how much of this effect is caused by stellar metallicity, quenching, or other phenomena are ongoing (for a recent

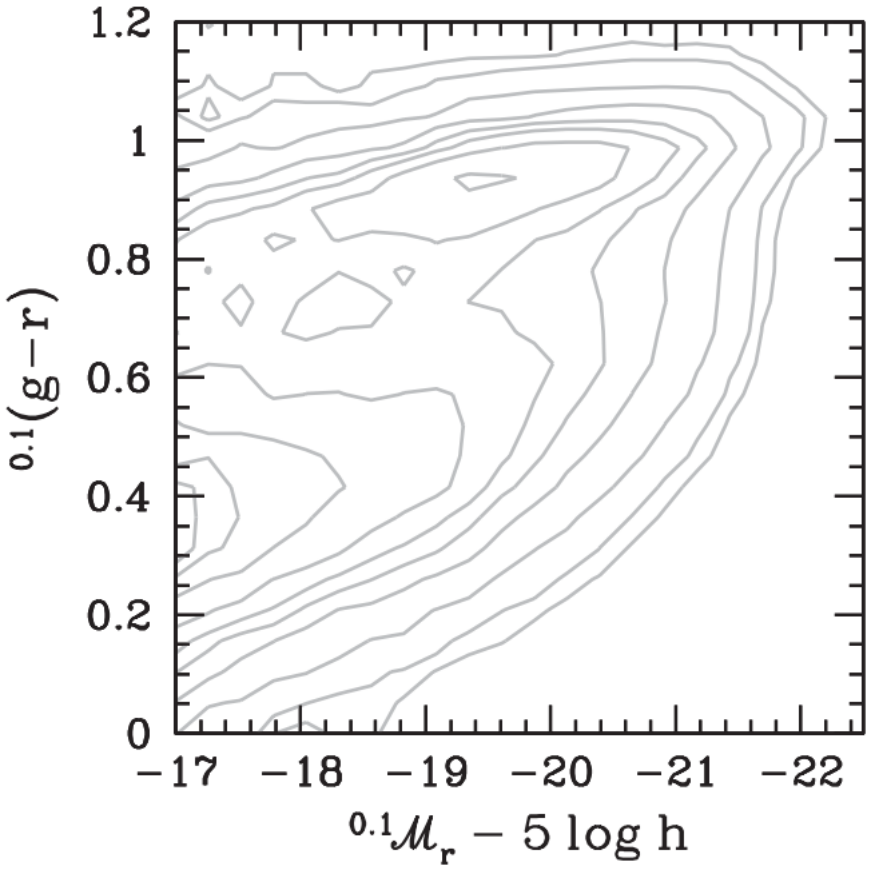


Figure 1.1. The colour–magnitude relation of about 365,000 galaxies in the SDSS, reproduced from figure 2.7 in Mo et al. (2010). The vertical axis shows galaxy rest-frame colour, defined as the difference of a given galaxy's absolute magnitudes in g and r bands, k-corrected to redshift 0.1. The horizontal axis is galaxy absolute magnitude in the r band, k-corrected to redshift 0.1. Here, each galaxy has been weighted by V_{max} to account for Malmquist bias. The colour–magnitude distribution shows two apparent galaxy populations, which are termed the red sequence (grouped around higher $^{0.1}(g-r)$ values) and the blue sequence (distributed over a range of lower values of $^{0.1}(g-r)$). This figure serves as evidence of galaxy bimodality (see section 1.2.1 for discussion).

example, see van Dokkum et al., 2015).

1.3 Large-scale structure observations today

The observations in studies of the large-scale structure of the universe today aim to meet a variety of scientific goals. Some of these aims include

- Test predictions made from ΛCDM and other models of physical cosmology
- Measure fundamental *cosmological parameters*, a set of constants that describe the state and history of the universe in the mainstream cosmological model
- Map the three-dimensional structure of the universe on large scales and track its evolution over the course of cosmic history (e.g. Postman et al., 2012)
- Deepen our understanding of galaxy formation
- Test the predictions made by general relativity and its alternatives, collectively known as "modified gravity" models.

A defining feature of many modern surveys that aim to fulfill any of the aims above is the requirement for a large amount of data. Modern cosmological surveys deal with phenomena that are imperceptible on the level of individual objects, such as stars, galaxies, or quasars. In order to answer the questions open in our field, surveys must provide data sets large enough to lead to statistically significant results, for which the systematics are very well understood and controlled.

1.3.1 Measuring galaxy redshifts

Galaxy redshifts are a fundamental part of observational cosmology. Redshifts enable us to go from a flat chart of galaxy positions on the sky to a three-dimensional map of galaxies in space and time. They allow us to measure the velocities with

which galaxies recede from ours at low redshifts and, given adequate precision, to study the relative velocities of galaxies bound together in clusters and groups.

Redshift z is calculated from light's observed wavelength λ_{obs} and the one emitted, λ_{emit}. The scale factor of the universe, a, is defined in terms of redshift:

$$1 + z \equiv \frac{\lambda_{\text{obs}}}{\lambda_{\text{emit}}} = \frac{1}{a} \tag{1.2}$$

The connection to the scale factor of the universe makes redshift a proxy for the expansion of the cosmos. Redshift's connection to the shift in observed wavelength gives us a gauge of the velocities with which galaxies move away from us and amongst each other. With the Hubble law (equation 1.1), we can translate the recession velocity provided by a redshift to a distance from an observer at small values of z. A deceptively simple number, redshift is an astrophysical powerhouse, carrying information that makes a large fraction of cosmological studies possible.

With that said, redshift uncertainties set a limit on their scientific power. The technique used for calculating a redshift dictates its precision. There are two major groups of redshift measurements: spectroscopic and photometric.

Photometric methods approximate galaxy redshifts using flux measurements in a combination of optical filters (for a review, see Salvato et al., 2019, and references within). Compared to their spectroscopic counterparts, photometric redshifts (known also by the name "photo-z") demand significantly less observational time and rely on numerical methods in lieu of a spectroscope. Their low resource requirements make photo-z a popular choice for providing the redshifts of galaxies that have been imaged optically, but not spectroscopically, as well as galaxies that are simply too faint for spectroscopy. Photo-z catalogues have been created for a wide range of cosmological surveys, including the Dark Energy Survey (DES) (Banerji et al., 2008; Sánchez et al., 2014), COSMOS (Ilbert et al., 2009), CLASH (Jouvel et al., 2014), and ALHAMBRA (Molino et al., 2014), to name just a few.

The variety of methods behind the development of these photo-z catalogues leads some datasets to have smaller uncertainties on redshift compared to others.

Spectra have been used since the earliest detections of galaxy velocities with respect to the Milky Way, such as in Slipher (1917). Spectroscopic redshifts, when available, generally come with small uncertainties*, which makes them an attractive asset for analyses of relatively small-scale phenomena (the non-linear regime) like cluster dynamics or small-scale redshift-space distortions. On all scales, the significant reduction in redshift uncertainties afforded by spectroscopy brings down uncertainty in every measurement that utilises galaxy redshifts.

1.3.2 A brief history of modern cosmological surveys

The rise of surveys driven by demands for large data sets pre-dates ΛCDM and the discovery of the accelerated expansion of the universe. The CfA survey†(Tonry & Davis, 1979; Davis et al., 1982; Davis & Huchra, 1982; Huchra et al., 1983; Davis & Peebles, 1983) assembled a map of the local universe catalogue with spectroscopic redshifts of about 2 thousand galaxies. Their aim, according to Huchra et al. (1983), was to

> assemble complete radial velocity information for well-defined samples of galaxies for use in statistical correlation analyses, the study of the local luminosity density and luminosity function, and the identification of bound aggregates of galaxies.

The key difference between CfA and previous observational efforts was the emphasis on driving the selection of targets for the catalogue with the statistical analyses ahead, as opposed to driving decisions about statistical analysis based on data after

*The level of uncertainty in spectroscopic redshifts depends on signal-to-noise levels, as well as the wavelength resolution of spectra. Consequently, the size of error bars on spectroscopic redshifts varies from survey to survey and is guided by each survey's scientific requirements and resources.

†The CfA survey was named after the Harvard-Smithsonian Center for Astrophysics, the academic home of the researchers behind the project.

its collection. CfA revealed the inhomogeneous distribution of galaxies on scales that were not sufficiently large for the cosmological principle to be applicaple. CfA data also demonstrated the existence of voids, the low-density complements of galaxy groups and clusters. A slice from the follow-up survey, CfA2, is shown in the inner top quadrant of figure 1.2, and showcases the enormous cosmic structure that was discovered in these early cosmological surveys.

Since the days of CfA, cosmological surveys have striven to increase the number of galaxy redshifts in the astrophysical arsenal. From CfA's 2 thousand, the Las Campanas Redshift Survey brought in over 26 thousand galaxy spectra (Shectman et al., 1996). The 2dF Galaxy Redshift Survey expanded the field with approximately 250000 spectra (Colless et al., 2001). SDSS and BOSS have collected millions of spectra. The Dark Energy Spectroscopic Instrument (DESI) aims to collect 35 million spectra over the course of five years (DESI Collaboration, 2016).

1.4 Connecting theory to observations in cosmology

In order to compare theoretical predictions to observed quantities, we must create a medium that renders both sides of scientific thought – theory and experiment – directly comparable. In the context of cosmology and the large-scale structure of the universe, that medium is a mock catalogue. Such a catalogue serves as a container of data about the quantities we would collect if we were observers in a simulated universe. These quantities might include the masses of galaxies or their brightnesses (in single or multiple bands), galaxy positions, velocities, redshifts, spectra, object type and more.

1.4.1 The role of mock data

To be a useful connector of theory to observations, mock data must provide quantities that resemble the observations against which it will be compared. The quantities should satisfy two major requirements:

1. The mock quantities must be statistically equivalent to real quantities on the level of individual objects.

2. The large-scale structure described by the mock data, as well as its summary statistics, should closely resemble what we observe on the sky. If our simulations and mock data were produced from a model that perfectly represented the Universe, the mock data we create from simulations should be indistinguishable from observed data if we examined both side-by-side. This level of statistical resemblance enables cosmologists to make comparisons between theory and observations at high levels of accuracy.

Figure 1.2 offers an illustration of mock data that satisfies the requirements outlined above (Springel et al., 2006). Red points in figure 1.2 stand for the positions of mock galaxies (in coordinates of angle and redshift), and blue points are observations. The comparison here is made between spectroscopic surveys, SDSS, CfA2 and 2dFGRS, and mock catalogues generated from the Millennium simulation. The figure shows mock catalogues' powerful way of representing the large scale structure of the universe. It visually demonstrates how well the theoretical model behind a cosmological simulation fits the data. Other comparative analyses of mock data, such as clustering comparisons, are necessary for a complete picture of the mock's representation of reality.

Mock catalogues can be used to develop and test the analysis tools intended for completed and upcoming surveys because a mock's cosmology is known a priori. The value of a number of parameters of interest can be measured directly in a mock, without the assumptions that are necessary in analyses of real data. Cosmological surveys also require mocks for testing observational strategies and quantifying biases (e.g. Smith et al., 2017).

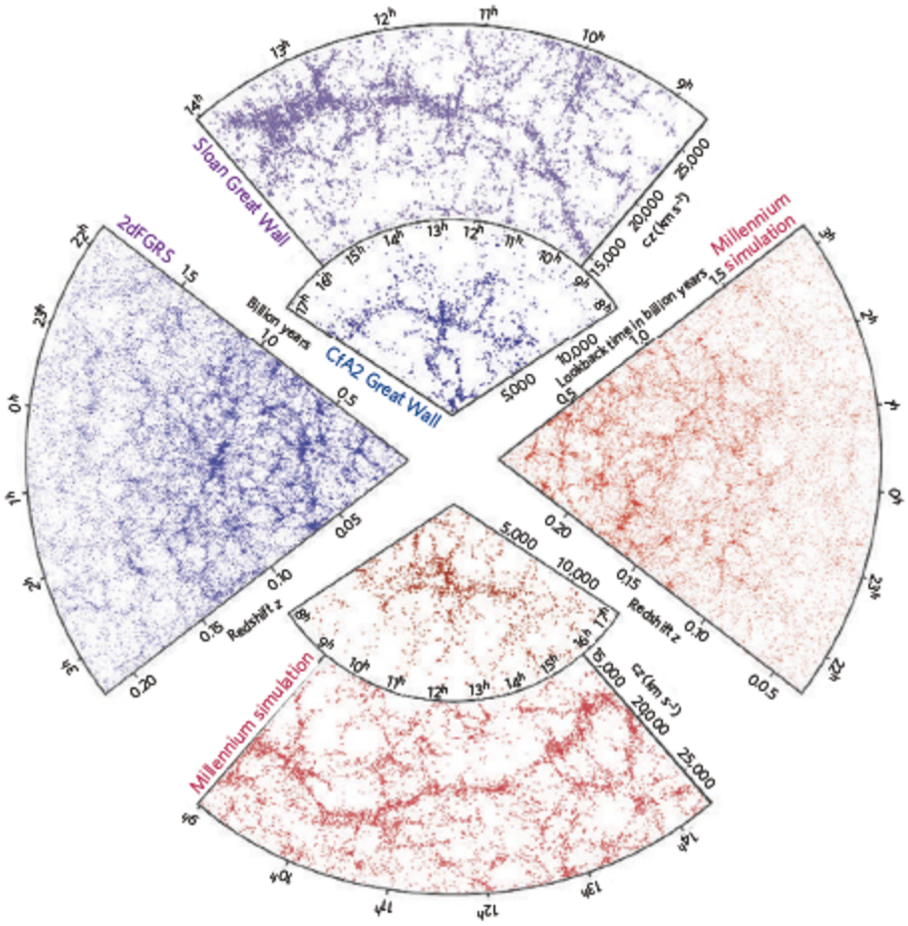


Figure 1.2. Comparison of the galaxy distributions in spectroscopic surveys and mock catalogues, reproduced from Springel et al. (2006). Blue points (top and left quadrants of the chart) represent galaxies with spectra observed in subregions of the SDSS, CfA2 and 2dFGRS surveys. The red points (in the right and bottom quadrants of the chart) represent mock galaxy data, created by populating the Millennium simulation with galaxies using a semi-analytic approach (see section 1.5.6 for a discussion). The slice in the top part of the chart shows the "Great Wall", a large-scale structure that comprises over 10 thousand galaxies, as it is observed in data from SDSS. The smaller top plot shows a structure whose scale dominated the CfA survey's data, the "CfA2 Great Wall", also known as the "CfA Stick Man". Below, the slices with red points show examples of "mock surveys", which were chosen to have large structures with similar properties to the real survey to demonstrate the striking resemblance of real observations that mock catalogues can supply.

1.4.2 The computational gap

Modern cosmological surveys such as eBOSS (Blanton et al., 2017; Dawson et al., 2013), DESI (DESI Collaboration, 2016, 2018), and LSST (Ivezić et al., 2008), require simulations that cover volumes that exceed 100 [Gpc/h]^3 in a multitude of realisations. Such great volumes are motivated by a combination of the scientific questions that the surveys attempt to tackle, as well as the systematics that accompany real-world observations.

For instance, for the analysis of systematics for BAO measurements, volumes of the order of $200\ h^{-3}$ Gpc^3 are necessary (DESI Collaboration, 2018). The simulations tailored for such measurements should cover volumes that are at least ten times greater than the volumes required to carry out the necessary measurements in order to limit the level of theoretic systematics (DESI Collaboration, 2018).

Ideally, these simulations would represent theory in a form that resembles observational data by solving the equations that describe the physics of baryons and dark matter across cosmic time. Simulations that solve equations that describe baryonic and dark matter simultaneously form a class called hydrodynamic simulations.

Complex simulations that account for the intricate physics that drives the Universe, however, are computationally expensive. The cost of simulating detailed physics that accounts for baryons in a volume that cosmological surveys require renders such simulations infeasible. Currently available hydrodynamical simulations cover volumes that are much smaller than what is required for cosmological surveys' needs. Examples of cosmological hydrodynamical simulations include:

- EAGLE covers volumes of 25, 50 and 100 [Mpc per side] with varying resolution levels (Crain et al., 2015)
- IllustrisTNG offers a [100 Mpc]^3 volume and a lower resolution [300 Mpc]^3 box (presented in Naiman et al., 2018; Nelson et al., 2018; Pillepich et al., 2018, and others)

- Massive Black II is a box of [100 h^{-1} Mpc] per side (Khandai et al., 2015).

While insufficient in volume, hydrodynamical simulations offer the potential for direct simulation of physical details behind galaxy formation and evolution. This property makes this class of simulations useful for the informing methods that produce realistic galaxy populations more quickly and at lower computational cost.

It is possible to approximate the distribution of baryonic matter by statistically populating halos in gravity-only simulations with galaxies. If they are implemented with sufficient precision, such statistical methods may be able to produce the cosmological-scale mock data that modern surveys require. We discuss a selection of such methods in section 1.5.

1.5 Populating dark matter-only simulations with galaxies

One way to circumvent the computational expense of running a full hydrodynamical cosmological simulation is to consider a dark matter-only N-body simulation, in which the equations of gravity only are solved, substantially bringing down computational costs. The simulation is then "populated" with galaxies following some algorithm, resulting in a catalogue of galaxies with properties and distribution that should be expected in a universe like the one that the N-body simulation represents. We discuss a selection of such algorithms for populating N-body simulations next. We open with a description of the statistical class of methods for populating N-body simulations (sections 1.5.1, 1.5.2 and 1.5.3). A discussion of statistical-empirical approaches follows in sections 1.5.4 and 1.5.5. Finally, we will describe physical methods (section 1.5.6).

1.5.1 Biased Dark Matter

"Biased dark matter" refers to a class of methods of selecting dark matter particles that are labelled as hosts of galaxies in N-body simulations (e.g. Cole et al., 1998; White et al., 2014). Biased dark matter uses the distribution of dark matter mass to calculate the locations of galaxies in simulations, based on the ansatz that galaxies are biased tracers of the mass distribution in the real universe (e.g. Cole et al., 1998).

The advantage offered by this class of methods is its low mass resolution requirements. If the locations of dark matter halos are estimated by calculating local mass density, as done, for example, in the quick particle-mesh method of White et al. (2014), there is no need to run a halo locating code to create a mock catalogue. The popularity of these methods is rising again due to the high CPU requirements for simulations used in covariance matrix analyses (e.g. Klypin & Prada, 2018).

1.5.2 Halo occupation distribution (HOD)

HOD is a method for connecting dark matter halos and galaxies (e.g. Benson et al., 2000; Peacock & Smith, 2000; Berlind & Weinberg, 2002; Berlind et al., 2003). The model aims to encapsulate the relationship between the spatial distributions of galaxies and of halos, with the expectation that dark matter halos host galaxies. The intended outcome of an HOD is a way to populate N-body simulation of dark matter particles with galaxies, with the final product enabling a direct comparison of observational data from galaxy surveys and the theory represented by simulations.

The standard HOD assumes that the probability of the presence of a galaxy at a certain position in space depends entirely on the mass of its host dark matter halo. This basic assumption is based on theoretical models suggesting that halo occupation is statistically independent of the halo's large scale environment at fixed halo mass (Berlind et al., 2003, and references within). Were the assumption true,

the HOD and population of dark matter halos from an N-body simulation could provide a complete background for galaxy clustering analyses on a wide range of scales (Berlind et al., 2003).

Conventionally, HOD models set the probability that N galaxies occupy a halo at a given mass above a given luminosity threshold with a parametrised functional form (e.g. Berlind & Weinberg, 2002; Cooray & Sheth, 2002). The product is a function that describes <N>, the expected number of galaxies occupying a halo at a given mass. To apply this relation when populating an N-body simulation, one needs to include the probability distribution providing this mean – often, a Poisson distribution with mean <N>. Significant literature exists on HOD's sub-Poisson behaviour in some regimes.

The problematic nature of the assumption of halo mass's role as a sole determining factor behind the distribution of galaxies had been discussed in published literature (for example, in Zentner et al., 2014). One situation where the validity of the assumption behind basic HOD breaks down is assembly bias* (e.g. Zehavi et al., 2019). A number of modifications to standard HOD have been proposed in attempts to remedy this problem while keeping the benefits of HOD, including computational speed and its ability to work in absence of resolved subhalos (e.g. Wechsler et al., 2006; Hearin et al., 2016).

One approach to address the assembly bias issues associated with a mass-dependent HOD is to add a second parameter, such as proposed in, for example, the decorated HOD method (Hearin et al., 2016). Research on the appropriate secondary quantity that addresses assembly bias while maintaining the simplicity of a model that depends only on halo mass is ongoing. Contenders include, but are not limited to, halo concentration (e.g. Paranjape et al., 2015; Hearin et al., 2016), halo spin (e.g. Hearin et al., 2016), or halo formation time (e.g. Hearin et al., 2016).

Alternatively, it may be possible to find a quantity that replaces halo mass in HOD

*Galaxy assembly bias is the phenomenon that results in galaxy properties, at fixed halo mass, showing a dependence on a secondary property, for example, halo concentration.

and addresses the issues associated with mass-dependent HOD, such as v_{max}*. However, no clear candidate has been found to address these problems completely yet (e.g. Zehavi et al., 2019).

1.5.3 Conditional luminosity function

Conditional luminosity functions (CLF) (e.g. Yang et al., 2003; Cooray, 2006; Yang et al., 2008) describe the full distribution of galaxy luminosities for a given halo mass. Typically, this is accomplished by calculating functions that separately describe the distribution of the luminosities of central and satellite galaxies (e.g. Wechsler & Tinker, 2018). A fundamental component for CLF is the relation between central galaxy luminosity and the mass of the halo to which that galaxy is gravitationally bound (Cooray, 2005b, 2006). Cooray (2006) provides a review of the mathematical details behind CLF.

There is not one way to calculate the CLF; one may obtain these distributions from measurements of galaxy clusters (e.g. Lin et al., 2004; Weinmann et al., 2006; Yang et al., 2008, 2009), from galaxy–mass cross-correlation through galaxy–galaxy lensing studies as a function of the galaxy luminosity (e.g. Sheldon et al., 2004; Mandelbaum et al., 2005) or from models of galaxy clustering and abundance (e.g. Yang et al., 2003; Cooray, 2006).

A CLF captures information that determines how the distribution of galaxies on large scales is related to the distribution of dark matter. This function can be used to populate an N-body simulation with galaxies to create a mock catalogue (e.g. Mo et al., 2004), which can subsequently be used to compare theory to observations statistically. An empirical model for the CLF, when combined with the halo mass function, describes the galaxy luminosity function (LF); this empirical model recovers the galaxy luminosity function outlined in Schechter (1976), with a

*v_{max} is the maximum circular velocity achieved by dark matter particles in a halo at a fixed simulation snapshot. We discuss this quantity in the context of a discussion of subhalo abundance matching in section 1.5.4.

characteristic luminosity L_* and α, the power-law slope of the LF for faint galaxies (e.g. Cooray, 2005b,a).

It has been suggested that one can extend CLF's empirical modelling approach to study statistics of galaxy types, such as, for example, the environmental dependence of galaxy colour bimodality (e.g. Cooray, 2005b, 2006).

1.5.4 Subhalo abundance matching (SHAM)

Subhalo abundance matching (SHAM) is a method of populating dark matter subhalos with galaxies by matching the cumulative abundance functions of a dark matter halo property (commonly, subhalo mass or circular velocity) to the luminosity function or a similar cumulative distribution function of a galactic property. Early work involving abundance functions included the assignment of individual galaxies to halos ranked by mass in an effort to perform model-independent studies of clustering in absence of baryonic physics (Wechsler et al., 1998). Halo/subhalo abundance matching was then established as a halo occupation model operating under the assumption that there is a one-to-one, monotonic relation between a host halo's dark matter mass and the luminosity of the galaxy residing in it (Vale & Ostriker, 2004; Kravtsov et al., 2004).

SHAM offers the advantage of using a cosmological model's predictive power for the number and properties of subhalos, as well as their relation to their host halos while requiring few, if any, parameters (Reddick et al., 2013). Cosmological simulations that resolve subhalos alleviate the need for assumptions about the occupation number and distribution of halo substructures, which are necessary for other models, such as HOD.

Implementations of SHAM have been shown to reproduce observed quantities that include the two-point correlation function (Conroy et al., 2006; Reddick et al., 2013; Lehmann et al., 2017, to name just a few), three-point statistics (e.g. Tasitsiomi

et al., 2004; Marín et al., 2008), galaxy–galaxy lensing (e.g. Tasitsiomi et al., 2004), and the Tully–Fisher relation (e.g. Desmond & Wechsler, 2015).

The original form of SHAM has been expanded to include the possibility of assigning galaxy luminosities or stellar masses by matching with the circular velocities that a given subhalo's dark matter particles achieve at various points in that subhalo's lifetime (for an example, see Conroy et al., 2006), $\mathrm{v_{circ}}$:

$$\mathrm{v_{circ}}(z) \equiv \max\left[\sqrt{\frac{\mathrm{G\ M}(z, < r)}{r}}\right], \tag{1.3}$$

where z is redshift, G is the gravitational constant, r is the halo radius in physical units at which circular velocity is measured, and $M(< r)$ is the halo mass enclosed within that radius. A variety of works have proposed that $\mathrm{v_{circ}}$ measured at various times in a subhalo's lifetime may be appropriate connectors of host subhalos to galaxies. Common proxies connecting dark matter subhalos to galaxies in SHAM include:

1. $\mathrm{v_{max}}$, the maximum $\mathrm{v_{circ}}$ of a subhalo at present time. Present time may mean redshift 0 or the redshift at which a galaxy is being assigned to the subhalo, for instance if SHAM is used to create a lightcone catalogue. $\mathrm{v_{max}}$ is used as a proxy quantity for SHAM in Conroy et al. (2006); Masaki et al. (2013a); Yamamoto et al. (2015); Guo et al. (2016); Chaves-Montero et al. (2016), among others.

2. $\mathrm{v_{infall}}$ or $\mathrm{v_{acc}}$, the $\mathrm{v_{max}}$ at the last time that the subhalo was central (e.g. Masaki et al., 2013a; Guo et al., 2016; Chaves-Montero et al., 2016).

3. $\mathrm{v_{peak}}$, the maximum $\mathrm{v_{circ}}$ reached by a subhalo's dark matter particles over the entire course of its existence (Reddick et al., 2013; Chaves-Montero et al., 2016; Mccullagh et al., 2017, among others).

4. $\mathrm{v_{relax}}$, the peak $\mathrm{v_{max}}$ reached during periods or subhalo relaxation, proposed as a novel SHAM quantity in Chaves-Montero et al. (2016).

5. Various definitions of subhalo mass, e.g. M_{max}, the maximum dark matter mass that a subhalo reaches during periods when it is a central subhalo. M_{max} is used, for instance, in Wetzel et al. (2013), to assign stellar masses to galaxies.

There is no consensus about the adequate dark matter quantity that can be used in SHAM (Wechsler & Tinker, 2018). However, a number of studies have raised evidence of some SHAM proxies offering more advantages than others.

A major issue that cosmologists are looking to address in the development of methods for generating mock data is assembly bias.

v_{max} characterizes the depth of gravitational potential and at fixed halo mass, v_{max} is directly related to halo concentration (e.g. Conroy et al., 2006; Zehavi et al., 2019). As halo concentration has been suggested to be a quantity that can track galaxy assembly bias, it offers the potential to lift the systematic effects of galaxy assembly bias in mock data. However, v_{max} describes the present state of a subhalo, which may miss some of the historical information contained in, for example, v_{peak}. Chaves-Montero et al. (2016) offers one comparison of the qualities that v_{circ}-related SHAM proxies impart on mock data.

If there is a perfect correlation between the quantities that SHAM connects, for instance between subhalo v_{max} and galaxy luminosity, then the no-scatter, one-to-one monotonic approach to subhalo abundance matching would fully describe the galaxy-halo connection for these two quantities. However, as shown in, for example, Chaves-Montero et al. (2016) via a comparison with a hydrodynamical simulation, the correlation between the subhalo and galaxy properties used in SHAM is not perfect. Hence, there is a need to add scatter when creating realistic mock catalogues with SHAM. The amount of scatter necessary should be informed by observable quantities, such as galaxy clustering, since the dark matter subhalo quantities used as proxies in SHAM are not directly observable in the real world.

A variety of approaches to add scatter to SHAM have also been developed. SHAM scatter methods include:

- Sampling a probability distribution (Guo et al., 2016; Chaves-Montero et al., 2016)
  - Fitting a parametrised model to a hydrodynamic simulation and sampling the resulting likelihood (Chaves-Montero et al., 2016)
  - Adding scatter to SHAM-style assignment of galaxy colours (Yamamoto et al., 2015; Masaki et al., 2013a)
- A deconvolution method (Reddick et al., 2013)
- Shuffling with a fixed scattering magnitude, used in Mccullagh et al. (2017), as well as the method described in chapter 2 of this work.

While addressing the issue of an imperfect correlation between galaxy luminosity, stellar mass, and the properties of their host subhalos, methods of adding scatter to SHAM datasets face the challenge of reproducing observational data, such as the luminosity function. Reproducing these observables successfully enables us to preserve the empirical premise behind SHAM.

1.5.5 SHAM with galaxy colour assignment

A number of methods that have built upon original abundance matching assign colours to galaxies in gravity-only simulations based on (sub-)halo age or environment (Hearin & Watson, 2013; Masaki et al., 2013b; Hearin et al., 2014; Yamamoto et al., 2015).

A common approach to assigning galaxy colours in a SHAM-like paradigm matches subhalos' directly simulated (sub-)halo property, such as v_{max} or v_{peak}, and a secondary (sub-)halo property that serves as a proxy for its age (see Masaki et al., 2013b; Kulier & Ostriker, 2015; Yamamoto et al., 2015). This is the so-called "age

model" of the dark matter halo-based prediction of galaxy colour. The approach is based on the notion that older galaxies should contain older, and, consequently, redder, stellar populations. Thus, if galaxy colour can be used as a measure of stellar population age when we analyse observations, we should be able to reverse the process and assign colours to simulated galaxies based on the ages of their subhalos.

A competing approach to assigning galaxy colours in an abundance matching-type process centres around the local environment of the galaxies (Masaki et al., 2013b). In this method, colour assignments are made to galaxies based on the local dark matter density around subhalos in which they live. This method is based on findings that the mass density profiles of early-type galaxies are higher than late-type profiles at z∼0–0.1 in several magnitude bins. The evidence originated from galaxy-galaxy lensing analyses (see Masaki et al., 2013b, and references within).

1.5.6 Semi-analytic models (SAMs)

Semi-analytic models of galaxy formation (e.g. White & Frenk, 1991; Kauffmann et al., 1993; Somerville & Primack, 1999; Cole et al., 2000; Baugh, 2006; Gonzalez-Perez et al., 2014; Croton et al., 2016; Lacey et al., 2016; Baugh et al., 2019) aim to model basic processes of galaxy formation by approximating the various physical processes with analytic prescriptions. This class of models treats baryonic matter in a post-processing stage of simulations, thus taking advantage of the computational efficiency of gravity-only simulations.

The prescriptions are traced through the histories of dark matter halos, often via merger trees extracted from N-body simulations. SAMs infer the properties of galaxies through differential equations that describe the physics of galaxy formation, informed with dark matter halo properties, such as their mass, size, spin, substructure, and merger history (see Croton et al., 2016, and references within).

Although these models are significantly less computationally expensive than hydro-

dynamical simulations, they distinguish themselves from other galaxy population schemes, such as HOD and SHAM, by requiring a large number of parameters. Constraining the parameters is a challenge for the semi-analytic approach. Like HOD and SHAM, SAMs make assumptions that need to be continually tested with observations and hydrodynamical simulations.

1.6 Aim of this research

The ultimate goal of this research is to produce a mock galaxy catalogue that closely mimics data that will be observed in DESI's Bright Galaxy Survey (DESI Collaboration, 2016). The resulting mock, Rosella, will provide the rest-frame r-band absolute magnitudes, rest-frame ${}^{0.1}(g-r)$ colours, 3D positions and velocities for galaxies inhabiting a volume of approximately (542 Mpc/h)^3. We will achieve this goal by performing SHAM on the P-Millennium N-body simulation. We evaluate the closeness of the match between our mock and real data by comparing the luminosity- and colour-dependent clustering of our mock's galaxies against previously published clustering of similar galaxy populations in observations and existing mock catalogues.

1.7 Outline of this thesis

In this thesis, the methods used for the creation and tuning of the mock catalogue Rosella will be presented. We will compare the mock's data to observations and existing mocks. We also discuss the appropriate future uses for the mock catalogue created here. The general structure is as follows:

- Chapter 2 describes the N-body simulation from which we built our mock. We describe our methods for assigning luminosities and colours to the galaxies in the catalogue. We go into detail of the age-centric approach to assigning galaxy colours.

- Chapter 3 describes the results of our work. We demonstrate the work that went into tuning the free parameters in our mock. We then demonstrate the catalogue's performance against a few metrics, including a comparison of the galaxy luminosity function and clustering to existing literature.

- Chapter 4 discusses the results presented in chapter 3. We make recommendations about the scenarios where this mock can be useful in its present form, and discuss ways in which this catalogue can be extended to future uses.

Chapter 2

Methodology behind Rosella's construction

The Rosella mock catalogue described here uses SHAM to populate the P-Millennium N-body simulation (described in section 2.1) with galaxies. Our approach provides rest-frame r-band luminosities and ${}^{0.1}(g-r)$ colours assigned with algorithms described in sections 2.3 and 2.6, as well as positions and velocities from P-Millennium.

2.1 The P-Millennium Simulation

The Planck Millennium N-body simulation (hereafter P-Millennium) is a high-resolution dark matter-only simulation of a 800 Mpc periodic box (Baugh et al., 2019). It is part of the 'Millennium' series of dark matter-only simulations of large-scale structure formation in cosmologically representative volumes carried out by the Virgo Consortium.

P-Millennium is run using cosmological parameters given by the best-fit cold dark matter (CDM) model to the first-year Planck cosmic microwave background data and measurements of large-scale structure in the spatial distribution of galaxies (Planck Collaboration et al., 2014). The analysis of the final Planck dataset has

introduced little change to these cosmological parameters (Planck Collaboration et al., 2018). See Table 2.1 for a summary of the specifications of the P-Millennium run.

Table 2.1. Selected cosmological parameters of the P-Millennium simulation. Note that the cosmology used corresponds to a flat universe. The first column lists the cosmological parameters, while the second column lists their values used in P-Millennium. The parameters are given in the following order: (1) $\Omega_{\rm M}$, present-day matter density in units of the critical energy density of the universe, (2) $\Omega_{\rm b}$, the baryon density parameter, (3) Ω_Λ, the energy density of the cosmological constant, Λ, (4) $\rm n_{spec}$, the spectral index of the primordial density fluctuations, (5) h, the reduced Hubble parameter, $h = \rm H_0$ / (100 km $\rm s^{-1}$ $\rm Mpc^{-1}$), (6) σ_8, the normalisation of the density fluctuations at the present day, (7) $\rm N_p$, the number of particles, (8) $\rm L_{box}$, the simulation box length, (9) $\rm M_p$, the mass of individual particles in the simulation, and (10) $\rm M_h$, the minimum mass of a resolved halo, corresponding to 20 particles. For a comparison with other simulations in the Millennium suite, see Baugh et al. (2019).

Parameter name	Value in P-Millennium
Ω_M	0.307
Ω_b	0.0483
Ω_Λ	0.693
$n_{\rm spec}$	0.9611
h	0.6777
σ_8	0.8288
$\rm N_p$	5040^3
$\rm L_{box}$ [h^{-1} Mpc]	542.16
$\rm M_p$ [h^{-1} $\rm M_\odot$]	1.06×10^8
$\rm M_h$ [h^{-1} $\rm M_\odot$]	2.12×10^9

P-Millennium follows the evolution of the matter distribution in a volume that is larger than that of the original Millennium Run (Springel, 2005) by a factor of $\times$ 1.43, after taking into account the slightly different Hubble parameters assumed in the two simulations (Guo et al., 2013; Baugh et al., 2019).

The mass resolution of P-Millennium, at 1.06×10^8 $\rm M_\odot h^{-1}$ per particle, and with 5040^3 particles representing the matter distribution, place P-Millennium at an intermediate resolution between the Millennium Simulation I of Springel et al. (2005) and the Millennium Simulation II run described in Boylan- Kolchin et al. (2009) (for a detailed comparison, see Baugh et al., 2019). The lowest mass for a resolved halo in P-Millennium is 2.12×10^9 $\rm M_\odot h^{-1}$. This makes the simulation an

appropriate choice for SHAM, since the simulation's mass resolution lets SUBFIND (Springel et al., 2001) resolve dark matter halo substructures, subhalos – a central component for creating a mock using SHAM (see section 1.5.4 for a discussion).

The lower boundary on the mass of a resolved subhalo in P-Millennium allows us to create a mock with a faint limit on absolute magnitude that reaches beyond the minimum luminosity cutoffs offered in other mock catalogues. For example, the *Buzzard* catalogue, presented in DeRose et al. (2019), creates a reference mock that models the galaxy distribution down to roughly M_r^h *$= -18.2$, with the caveat that a SHAM catalogue built on an N-body a simulation box of size 400 h^{-1} Mpc with 2048^3 particles is not strictly complete down to $M_r^h$of –18.2. As discussed in section 3.2, Rosella has the potential to reach galaxy absolute magnitudes down to about $M_r^h = -15.5$ (depending on the scatter implemented), should such an absolute magnitude cutoff be chosen, thanks to P-Millennium's resolution.

The initial conditions were generated at redshift 127 using second order Lagrangian perturbation theory, described in Jenkins (2010, 2013). The simulation was run on 4096 processors of the COSMA-4 supercomputer at Durham University, using a reduced memory version of the gravity-only code GADGET (Springel, 2005), taking approximately 20 TB of Rapid Access Memory (RAM). The N-body simulation and halo finding were run concurrently and accounted for 3/4 and 1/4 of the project's total 7 million CPU hours, respectively (Baugh et al., 2019). Halo and subhalo finding was completed using the SUBFIND code (Springel et al., 2001). Dark matter halo merger trees were constructed from the SUBFIND (Springel et al., 2001) subhalos using the DHALOS algorithm described in Jiang et al. (2014). Halo data amounts to approximately 0.5 Tb per snapshot, with files increasing in size for later snapshots compared to earlier ones, as expected with a hierarchical structure formation scenario.

*We define r-band absolute magnitude dependent on h and k-corrected to $z \sim 0.1$ as $M_r^h \equiv {}^{0.1}M_r$ - 5 log h, where h is the dimensionless constant given as $H_0 = 100\ h$ km s^{-1} Mpc^{-1}.

2.2 The choice to apply SHAM to P-Millennium for DESI BGS

The high mass resolution of P-Millennium makes it possible for us to apply SHAM to populate its subhalos with galaxies and create a mock catalogue tailored with the scientific requirements of DESI's Bright Galaxy Survey (BGS) in mind. The choice to use SHAM offers a few advantages over other methods.

A mock catalogue tailored for the needs of BGS already exists: it is a lightcone mock constructed with an application of HOD to the Millennium-XXL (MXXL) simulation (Smith et al., 2017). However, the HOD catalogue has some limitations that can be addressed by applying SHAM to P-Millennium. These include a need for data, the clustering of which better matches that of observations for faint samples, due to the presence of halos that fall below the MXXL halo mass resolution limit.

Additionally, SHAM is ideal for the analysis of groups and clusters for which BGS data may be used in the future and for which HOD models are not complex enough. For example, it is not clear whether the mitigation techniques planned for DESI can recover statistics affected by assembly bias. Mitigation techniques planned for the data analysis of DESI include pairwise inverse probability (PIP) weighting to enable the recovery of 2-point statistics. It is yet to be shown, however, that techniques like PIP weighting can be used in more complex statistical applications, such as group finding. Mock data that includes assembly bias provides a higher level of complexity that one expects to exist in real data. Hence, the algorithmic tests and science interpretation analysis for BGS will benefit from mock data that includes halo assembly bias.

Halo assembly bias describes the phenomenon that dark matter halo clustering depends on properties besides halo mass, including but not limited to formation time, concentration and spin (e.g. Gao et al., 2005; Wechsler et al., 2006; Gao &

White, 2007). For a given halo mass, clustering is stronger in dark matter halos that form at earlier times. The dependence of clustering on halo formation time increases with decreasing halo mass (Gao et al., 2005).

This presents a problem for halo occupation models that assume the independence of the distribution and properties of galaxies from their environment beyond halo mass (Gao et al., 2005). Abundance matching on subhalo quantities that include information about their history, such as peak circular velocity v_{peak} or satellite subhalo accretion mass M_{acc}, may lift part of this assumption of distribution-environment independence in the galaxy-halo occupation relation.

By incorporating a proxy that implicitly accounts for subhalo assembly history, v_{peak}, a SHAM catalogue can be more informative when investigating the effects of assembly bias on observational data and computing statistics that may be affected by it, compared to an HOD mock.

Implementing SHAM is relatively quick compared to a physical method, such as a SAM. Additionally, it can be arbitrarily tuned to reproduce certain statistics, as it includes an empirical component in its methodology.

2.3 Algorithm for luminosity assignment

We assign luminosity values to galaxies in our mock catalogue by assuming that every dark matter subhalo located in P-Millennium with the SUBFIND code (Springel et al., 2001) that satisfies a minimal condition on v_{peak} hosts a galaxy. We assume that galaxy luminosities correlate with the quantity v_{peak} .

v_{peak} is the central quantity that allows us to connect dark matter subhalos in our N-body simulation to the galaxies in our mock catalogue, and it is defined as the peak value of highest circular velocity reached by a subhalo's particles across the simulation snapshots in which that subhalo is found (see section 1.5.4 for further details). Maximum circular velocity here is the maximum value of v_{circ} across the

particles belonging to a subhalo at a given time, with v_{circ} defined as

$$v_{circ}(r, z) = \sqrt{\frac{GM(z, < r)}{r}} \tag{2.1}$$

r here is the physical distance between the particle and the centre of the subhalo, z is redshift, G is the gravitational constant, and M(z, $<r$) is the mass enclosed within the radius r, at redshift z.

v_{peak}, by construction, includes information about a subhalo's formation history. Since the effects of galaxy assembly bias have been seen in properties that are influenced by the assembly histories of their subhalos, using v_{peak} allows us to implicitly account for assembly bias to a greater degree than is possible with a present-day SHAM proxy, like v_{max} (e.g. Chaves-Montero et al., 2016). Thus, for example when we populate satellite subhalos with galaxies, v_{peak} allows us to account for the historical values of that subhalo's v_{max}, thus mitigating the influence of effects like dark matter mass stripping as a consequence of mergers. There has been evidence of subhalos with higher v_{peak} values tending to have higher concentration and earlier formation times, which are some of the properties associated with assembly bias (see Xu & Zheng, 2018, and reference therein).

To compute v_{peak}, we compile the histories of v_{max} values for individual subhalos and pick the highest v_{max} value. The process of tracing the aforementioned histories is described in section 2.4, and the finding of a single subhalo's v_{peak} is illustrated in figure 2.5.

We assume that subhalo v_{peak} follows a monotonic relation with galaxy absolute magnitude in r band, M_r^h. In the first step of luminosity assignment, we operate under the assumption that the relation between magnitude and v_{peak} are one-to-one, but that assumption is no longer applicable once we add scatter to the mock data. For the first, no-scatter, stage of our algorithm, the assumed relation between r-band magnitude and v_{peak} can be expressed as:

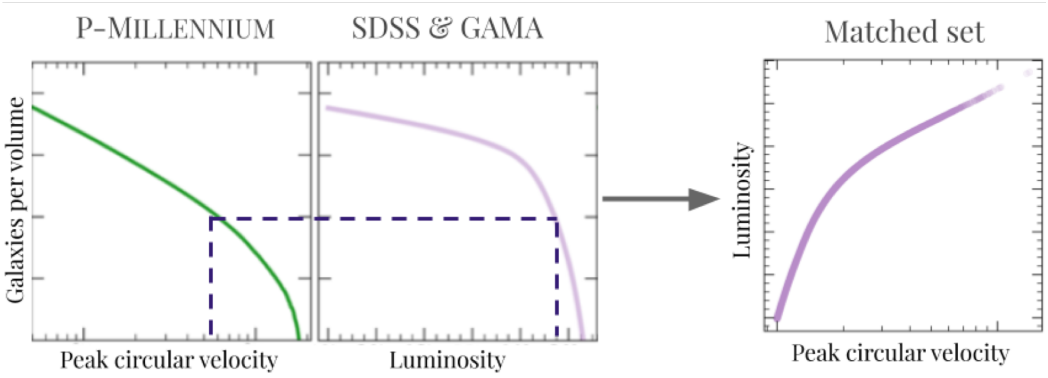


Figure 2.1. Illustration of assigning luminosities to galaxies with SHAM without scatter. The first two panels show abundance relations: left panel shows the abundance of subhalos as a function of their v_{circ} in our N-body simulation, and the central panel shows the abundance of galaxies as a function of their brightness (luminosity function). For a given subhalo with a known v_{circ}, we can follow the dashed line to match its abundance in a simulation to a luminosity value that has the same abundance in observations. Repeating this matching for a set of subhalos produces a set of points that form a tight line, as seen on the right panel.

$$n_g(< M_r^h) = n_h(> v_{peak}) \tag{2.2}$$

Here, n_g is the number of galaxies of a given M_r^h or brighter, and n_h is the number of subhalos of a given v_{peak} or higher. That is, if we consider a subhalo with a given v_{peak}, its abundance is the number of subhalos with the same value or greater, divided by the simulation volume. Let us suppose that a given subhalo's v_{peak} abundance is α $[h\ \mathrm{Mpc}^{-1}]^3$. Then the magnitude that we assign to that subhalo's galaxy should have the same abundance, as determined from the cumulative galaxy luminosity function (LF): the number of galaxies of the assigned magnitude should equal α $[\mathrm{Mpc}\ h^{-1}]^{-3}$.

We follow a number of specific steps to assign magnitude values to the galaxies in our sample:

1. Get the evolving r-band luminosity function using the SDSS r-band LF (Blanton et al., 2003) and the Galaxy and Mass Assembly survey (GAMA) r-band LF (Loveday et al., 2012). The combined smooth LF used here is the one also

used in Smith et al. (2017) for the development of a lightcone mock catalogue. We call this set of reference data the 'target luminosity function', as it is the LF that we aim to replicate in our mock.

2. Perform SHAM with zero scatter using the observed luminosity function with equation 2.2, the monotonic relation between luminosity and v_{peak}. Chaves-Montero et al. (2016) and Mccullagh et al. (2017) also used this relation as the basis of their SHAM assignments. Figure 2.1 offers an illustration of the process.

3. Add luminosity-dependent scatter using the method, the fundamentals of which are described in Mccullagh et al. (2017)*. This approach uses a scatter magnitude ($\sigma(M_r^h)$, also called ΔM_r^h in Mccullagh et al. (2017)) to produce results that are illustrated in figure 2.2. The steps for adding scatter for a given sample are:

   a) Starting with the array of SHAM-assigned galaxy magnitudes without scatter, M_r^h, create a new array, $M_r^{h\prime}$. Draw the $M_r^{h\prime}$ value for a given subhalo from a Gaussian distribution with mean equal to that subhalo's value in M_r^h and luminosity-dependent standard deviation $\sigma(M_r^h)$, which we propose to be given by the expression

   $$\sigma(\mathrm{M}_r^h) = \alpha + \beta \tanh(\mathrm{M}_r^h - M_{r,\mathrm{ref}}^h), \tag{2.3}$$

   where α, β and η are the free parameters that we have tuned with the help of clustering analysis (see section 3.6). The values that we use to create the Rosella catalogue presented in this work are:

   $\alpha = 0.75$; $\beta = 0.45$; $M_{r,\mathrm{ref}}^h$ =- 20

   b) Find the ordering that will rank order the new array of scattered magnitudes ($M_r^{h\prime}$), and sort the array of (original) galaxy magnitudes (M_r^h) by that ordering;

*This method effectively shuffles the ranks while maintaining the originally assigned set of luminosities. Hence, it doesn't perturb the cumulative luminosity function, and no deconvolution is necessary unlike other methods of adding scatter to SHAM data.

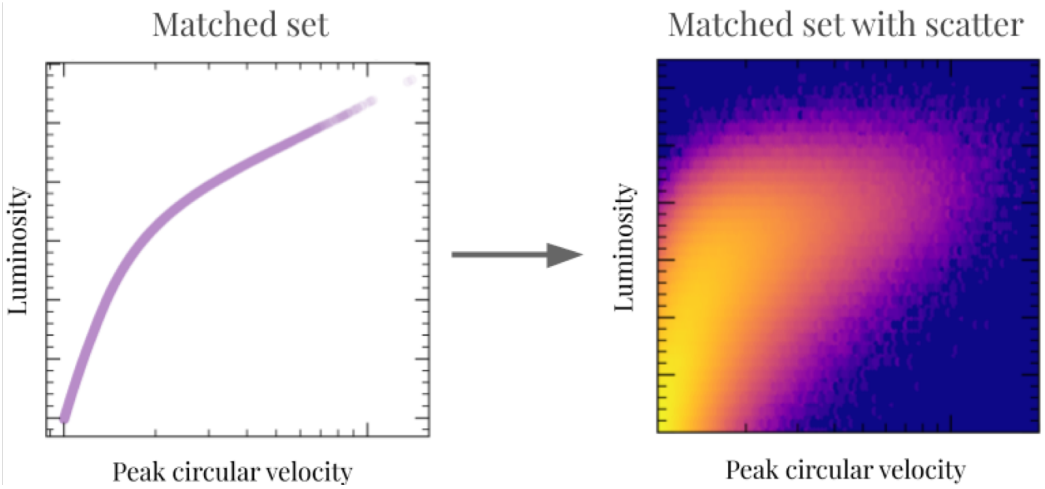


Figure 2.2. Illustration of scatter added to SHAM data. Beginning with the one-to-one matching procedure illustrated in figure 2.1 to produce the left panel, we add scatter to the luminosity-v_{circ} data set. The data with the added scatter no longer follows a line in luminosity-v_{circ} space. The right panel shows the logarithmic density of data points in luminosity-v_{circ} space after the addition of scatter. The method used here preserves the luminosity function of the no-scatter counterpart of this SHAM data set.

c) Rank order the v_{peak} values of the subhalos;

d) Assign the original values of galaxy magnitudes (M_r^h) to the subhalos such that the subhalo with the highest v_{peak} gets assigned the galaxy magnitude with the brightest $M_r^{h\prime}$.

The resulting set of M_r^h values is provided in the Rosella mock and also used to assign ${}^{0.1}(g-r)$ colours using the algorithm described in section 2.6.

In this work, we limit the analysis to subhalos with $v_{peak} \geq 50$ km/s, a choice motivated by the P-Millennium resolution, where subhalos with v_{peak} greater than 50 km/s are well-resolved.

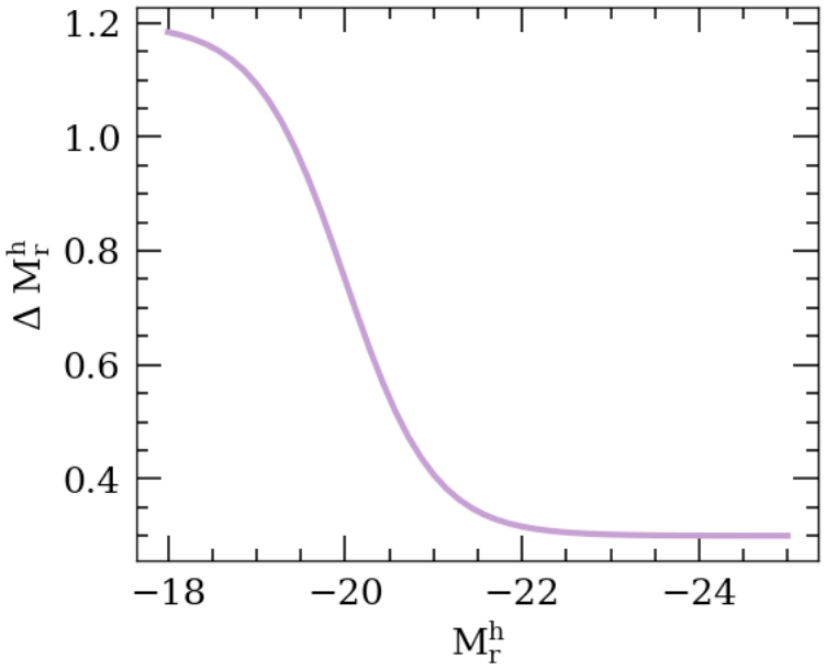


Figure 2.3. Scatter as a function of absolute magnitude. On the y-axis, we see the scatter that is added to the luminosity assignment during the "shuffle" part of our SHAM algorithm. This function follows the form outlined in equation 2.3.

2.4 Tracing subhalo histories across P-Millennium snapshots

To calculate v_{peak}, as well as a proxy for a subhalo's age, z_{form}, which we describe in section 2.5, we compile the histories of v_{max} values that individual P-Millennium subhalos reach over the course of the simulation. This is not a trivial task, since the files containing the necessary information (DHALO tree files) are not sorted by snapshot. Identifying a given subhalo's progenitor requires finding a DHALO file entry that lists the value of that subhalo's `nodeIndex`* under the `descendantIndex`* column. If multiple progenitors are found, we select the one with `isMainProgenitor`* = 1. To complete one subhalo's history, we repeat this progenitor tracking procedure until we reach a subhalo without a progenitor that can be found in DHALO files.

*See tables A.1 and A.2 for descriptions of the properties provided in the P-Millennium output catalogues relevant to Rosella's methodology.

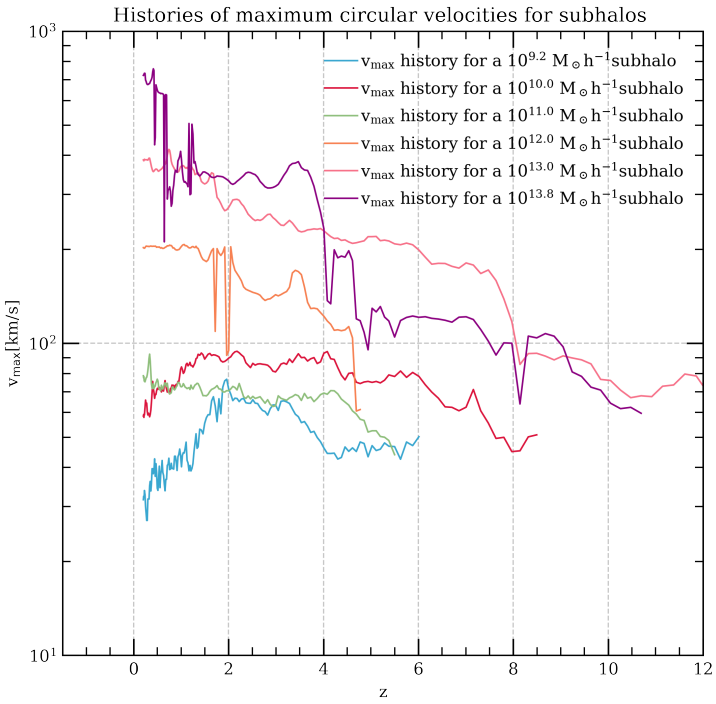


Figure 2.4. Example histories of maximum circular velocities across the time of subhalo existence. The vertical axis shows the v_{max} values of individual subhalos at given redshifts z. Each line tracks the v_{max} history of a subhalo of mass indicated in the legend. Although the subhalo masses demonstrated here are relatively evenly spaced apart, these examples were chosen blindly for the sake of illustrating the shapes that v_{max} history curves may take. The subhalos chosen here are all present in the P-Millennium SUBFIND catalogue at $z = 0.0$

Figure 2.4 shows a few examples of the resulting tracked subhalo histories. Note that this plot shows example v_{max} histories for subhalos picked from the catalogue of subhalos at $z = 0.0$, the redshift at which we compiled the first catalogue of subhalo histories during the initial development phase of our project. We then generated another full dictionary of subhalo histories for subhalos found at the P-Millennium snapshot corresponding to $z \sim 0.2$ (we discuss our choice of choice for Rosella in section 3.1). The need to compile a full dictionary of subhalo histories separately for redshifts 0.0 and 0.2 is due to some subhalos present in the $z \sim 0.2$ snapshot no longer being present at $z \sim 0.0$, due to, e.g. merger events. Conversely, some subhalos found in the subhalo history file for $z \sim 0.0$ are not yet present at redshift 0.2, rendering a $z \sim 0.2$ subhalo history database inappropriate for generating a

reference mock at $z \sim 0.0$.

Consequently, should the Rosella technology be used to create mock catalogues at other snapshots, one would need to create a separate subhalo history database for every additional snapshot. While the process of compiling a database of subhalo histories is time intensive (about 2 hours is required for each of the 1024 P-Millennium files to compile the information in the COSMA supercomputer), the process only needs to be executed once for every snapshot of interest. The code for this procedure, along with the code used to complete the rest of the Rosella methodology, is stored in a private repository on github.com/safonova/pmillennium-sham.

Table 2.2. Subhalo properties provided in Rosella's subhalo history datasets.

Dictionary key	Description
`nodeIndex_at_zero`	the `nodeIndex` of the subhalo at the snapshot at which the subhalo history database has been completed
`vmax`	Python dictionary of v_{max} values in physical km/s for that subhalo; keys: snapshots
`vpeak`	v_{peak} value for the given subhalo in physical km/s
`tree_file_number`	number identifying the DHALO output file (out of 1024) containing the subhalo's information
`node_indices`	`nodeIndex` values for the subhalo and its progenitors; keys: snapshots
`number_of_particles`	number of particles in the subhalo and its progenitors across their history; keys: snapshots
`idx_in_tree_file`	indices pointing to the elements in the DHALO file that contain the subhalo and its progenitors; keys: snapshots

The subhalo history catalogues we have compiled contain properties outlined in table 2.2. A subhalo history file is stored as a Python pickle file, containing a Python dictionary. To access the DHALO history of a given subhalo in the subhalo history database at snapshot 230 (which corresponds to $z \sim 0.2$), one simply needs to load the pickled dictionary and use the subhalo's `nodeIndex` at snapshot 230 as the key.

We have confirmed that the subhalo histories we have compiled represent their subhalos correctly by comparing the v_{peak} values that we get from the maximum of the history curve, like the one we show in figure 2.5, to v_{peak} values provided

in DHALO trees. The comparison has shown that v_{peak} values extracted from our subhalo v_{max} history dictionaries match those in DHALO catalogues. Thus, we have confidence that the dictionaries we have compiled represent the subhalos in P-Millennium well.

2.5 Definition of formation redshift

In order to assign colours to Rosella galaxies, we compute a subhalo's "formation redshift", z_{form}, which serves as a proxy for a subhalo's age. The choice to connect galaxy colours to the ages of their host subhalos stems from the idea that older subhalos are likely to have older and, consequently, redder stellar populations (e.g. Mo et al., 2010; Hearin, 2015). We compute an individual subhalo's z_{form} based on the criterion that z_{form} corresponds to the redshift at which a subhalo's v_{max} reaches v_{form}:

$$v_{form} = f \times v_{peak}, \tag{2.4}$$

provided that a subhalo's v_{form} is reached before its v_{peak}. Here, f is a free parameter that we have chosen to set at 0.75 to create Rosella and the results presented in this work. While 0.75 is the only value of f that we tested in this work, it would be possible to adjust this value in order to tune the mock data produced with the model presented here. Expression 2.4 is inspired by the works of Masaki et al. (2013b) and Yamamoto et al. (2015); however, those papers work with v_{max} instead of v_{peak}, and their v_{form} is formulated in terms of v_{max}. Nonetheless, the v_{form} in Masaki et al. (2013b) and Yamamoto et al. (2015) has a similar underlying structure to the criterion that serves as a proxy for subhalo age in our methodology.

Finding the value of v_{form} is achieved by searching the v_{max} history of a given subhalo, illustrated, for example, in figure 2.5. To ensure that v_{form} always meets the requirement that it must be reached before v_{peak} is reached, we only consider the subhalo history entries that include redshifts higher than the redshift at which a subhalo reaches its v_{peak}. If the v_{max} values in that subhalo's progenitor history

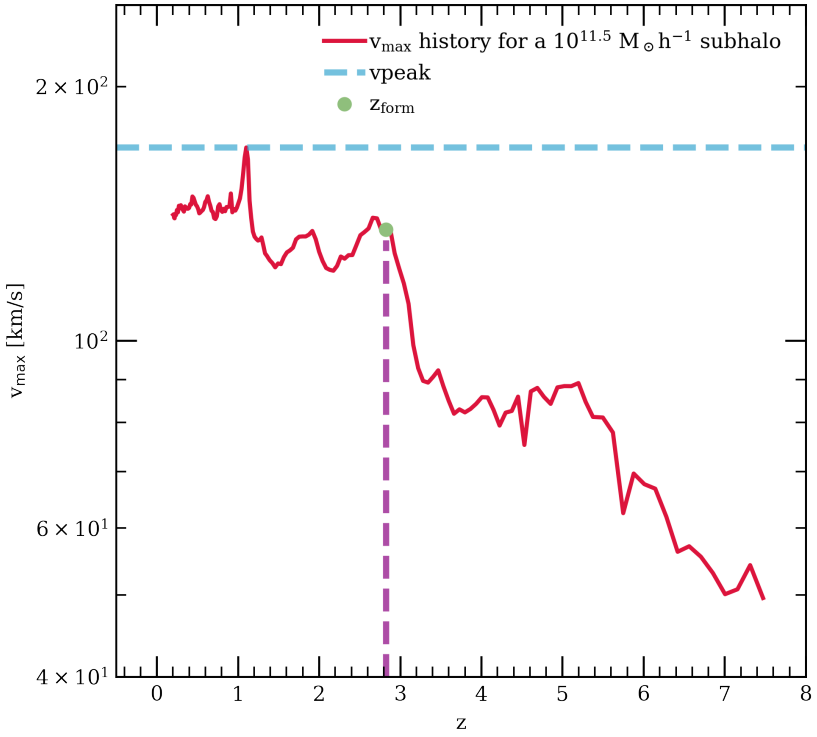


Figure 2.5. Example of a formation redshift search for a single subhalo. The vertical axis shows a subhalo's v_{max}, and redshift values corresponding to P-Millennium snapshots are plotted on the horizontal axis. The example here shows the v_{max} history of a blindly chosen subhalo with particle mass around $10^{11.5}\ M_{\odot}h^{-1}$. The horizontal dashed line indicates v_{peak}, taken as the maximum value in the v_{max} history of the subhalo. The red dot and vertical dashed line demonstrate the value of z_{form} for this subhalo, according to the criterion in equation 2.4, with f set to 0.75.

reach 0.75 times its v_{peak} or lower, we interpolate to find that subhalo's z_{form}. If, however, the history of the subhalo and its progenitors terminates before v_{form} is reached, z_{form} is set to the redshift corresponding to the last snapshot at which the subhalo is found.

2.6 Algorithm for colour assignment

The procedure for the assignment of ${}^{0.1}(g-r)$ colours to Rosella galaxies comprises multiple steps and is built around two notions. Galaxy colour bimodality analyses (for example, the colour-magnitude diagram of SDSS galaxies shown in figure 1.1)

show that brighter galaxies tend to be redder across both blue and red populations of galaxies. Thus, we begin colour assignment by calculating a cumulative distribution function of ${}^{0.1}(g-r)$, conditional on M_r^h, individually for each galaxy. We describe this procedure in section 2.6.1.

The other component of our colour assignment procedure builds upon the correlation between galaxy colour and ages (e.g. Mo et al., 2010; Hearin, 2015). In section 2.6.2, we compute the cumulative distributions of $z_{\rm form}$ (defined in section 2.5) for subhalo populations limited by $\rm v_{peak}$.

In section 2.6.3, we describe the procedure for finding each subhalo's position on the $\rm v_{peak}$-dependent distribution of $z_{\rm form}$ and translating it to a ${}^{0.1}(g-r)$ value for the galaxy residing in it.

2.6.1 Luminosity-dependent galaxy colour distributions

To build upon the connection between luminosity and colour, we begin by calculating two separate probability distribution functions of ${}^{0.1}(g-r)$ values for galaxies in blue and red galaxy populations at a given M_r^h. We assume these distribution functions to follow Gaussian form. Figure 2.6 shows examples of these distributions for a selection of absolute magnitude values at redshift 0.

The mean and rms values in the Gaussians are based on formulations of ${}^{0.1}(g-r)$ colour with redshift evolution in Smith et al. (2017), which were built upon SDSS and GAMA data, as well as earlier colour assignment work (Skibba & Sheth, 2009). The formulations for the mean and standard deviation of ${}^{0.1}(g-r)$ dependent on the absolute magnitude M_r^h and redshift z – in red and blue galaxy populations – are:

$$\begin{aligned} <^{0.1}(g-r)|M_r^h, z>_{\rm blue} = & 0.62 - 0.11(M_r^h + 20) - \\ & 0.25(\min[z, 0.4] - 0.1) \end{aligned} \tag{2.5}$$

$$\begin{aligned} {\rm rms}({}^{0.1}(g-r)|M_r^h, z)_{\rm blue} = & 0.12 - 0.02(M_r^h + 20) - \\ & 0.2(\min[z, 0.4] - 0.1) \end{aligned} \tag{2.6}$$

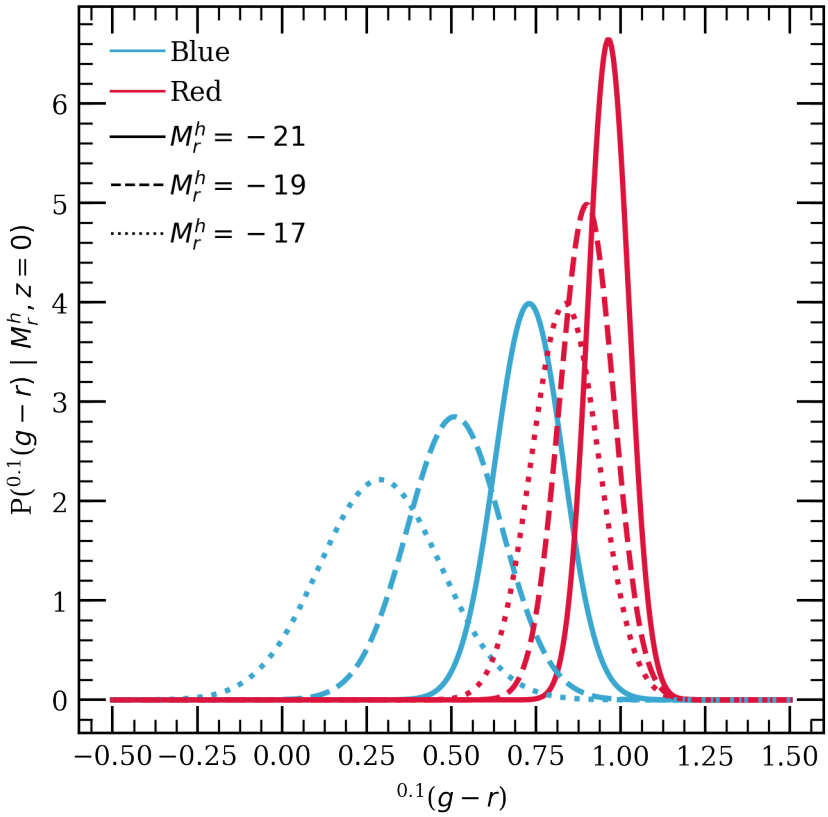


Figure 2.6. Gaussian distributions of $^{0.1}(g-r)$ for red (red lines) and blue (blue lines) galaxy populations, with mean and standard deviation given in equations 2.5 and 2.6 for blue galaxies and equations 2.7 and 2.8 for red galaxies. All lines are computed for $z = 0$. The line styles correspond to the M_r^h values for which the Gaussians are calculated, as indicated in the legend.

$$<^{0.1}(g-r)|M_r^h, z>_{\rm red} = 0.932 - 0.032(M_r^h + 20) - 0.18(\min[z, 0.4] - 0.1) \quad (2.7)$$

$$\mathrm{rms}(^{0.1}(g-r)|M_r^h, z)_{\rm red} = 0.07 - 0.01(M_r^h + 20) - 0.2(\min[z, 0.4] - 0.1) + 0.1(\min[z, 0.4] - 0.1)^2 \quad (2.8)$$

We calculate the $^{0.1}(g-r)$ cumulative distribution function (cdf), using the M_r^h value for every galaxy separately. Figure 2.7 shows examples of colour cdfs for a selection of M_r^h values. A $^{0.1}(g-r)$ cumulative distribution function is given by:

$$\mathrm{cdf}_{M_r^h} = f_{\rm blue}(M_r^h) * \mathrm{G}(M_r^h, z)_{\rm blue} + (1 - f_{\rm blue}(M_r^h)) * \mathrm{G}(M_r^h, z)_{\rm red}, \quad (2.9)$$

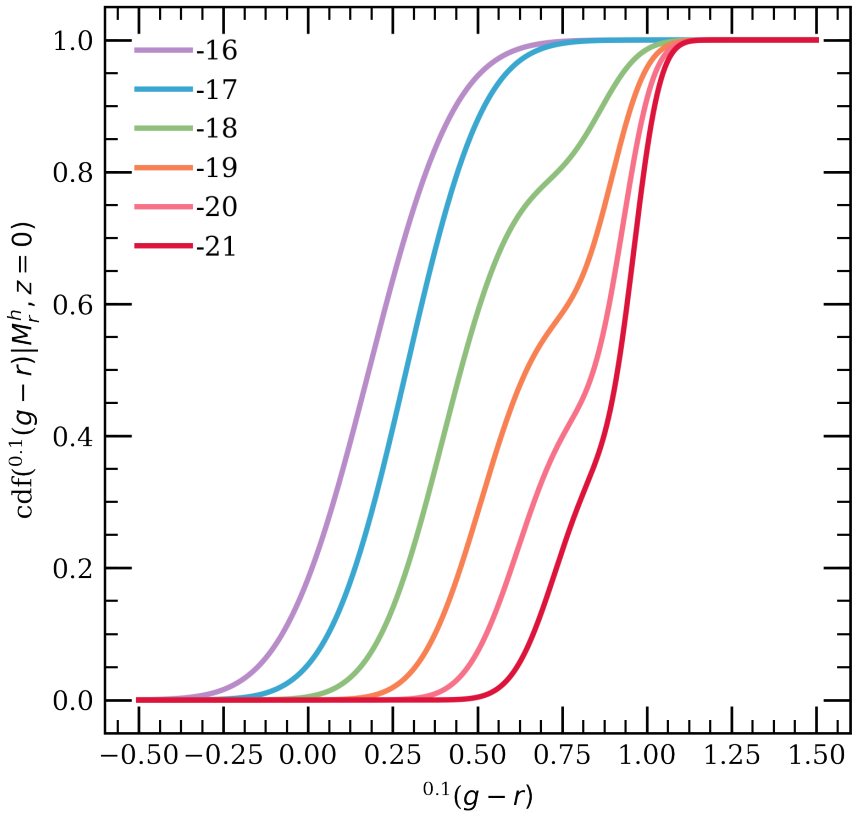


Figure 2.7. Cumulative distribution functions of $^{0.1}(g-r)$ for a selection of M_r^h values, as indicated in the legend. The functional form of these distributions is given in equation 2.9.

which is the sum of two Gaussians with mean and dispersions given as functions of M_r^h and redshift z, defined in equations 2.5, 2.6, 2.7 and 2.8. The M_r^h cdf is normalized to integrate to unity in such a way that the colour cdf for blue galaxies constitutes $f_{\rm blue}$ of the cdf. $f_{\rm blue}$ here is the expected fraction of blue galaxies in the total galaxy population at a fixed M_r^h. The rest of the cdf comes from the probability distribution of $^{0.1}(g-r)$ value in the red galaxy population at a given M_r^h.

Using the sum of two Gaussians that describe red and blue galaxy populations, we accommodate the fact that galaxy bimodality forms a continuum, and allow scatter in the colour assignment in Rosella to be dictated by the absolute magnitudes we assign during the SHAM M_r^h assignment phase.

The expected fraction of blue galaxies in a population at a given M_r^h is given as

$$f_{\text{blue}} = \begin{cases} \min \left\{ \begin{array}{c} 0.4 + 0.2 * (M_r^h + 20.) \\ \frac{1}{1+\exp(-(M_r^h+20.5))} \end{array} \right\}, & \text{if } M_r^h > -18 \\ \max \left\{ \begin{array}{c} 0 \\ 0.4 + 0.2 * (M_r^h + 20.) \\ 0.46 + 0.07 * (M_r^h + 20.) \end{array} \right\}, & \text{if } M_r^h \leq -18 \end{cases} \tag{2.10}$$

This formulation ensures that the fraction of blue galaxies always stays between 0 and 1. The sigmoid expression on the faint side of this function ensures that the size of the population of red galaxies slowly tapers off instead of meeting a sharp cutoff at a fixed magnitude, which makes our model different from the prescription in Smith et al. (2017). Figure 2.8 visually demonstrates the relationship between f_{blue}, the mean value of ${}^{0.1}(g-r)$, and M_r^h that is captured in equations 2.10, 2.5, 2.6, 2.7, and 2.8 mathematically.

2.6.2 Cumulative $\mathbf{v_{peak}}$-dependent distribution of $z_{\mathbf{form}}$

We calculate the cumulative distribution functions of subhalos in P-Millennium with respect to formation redshift z_{form}, in bins of v_{peak}. That is, for a given subhalo, we know its v_{peak} value from P-Millennium. We know its z_{form} by calculating it from the history of its v_{max}, with the aid of the algorithm outlined in section 2.5.

Examples of such z_{form} distribution functions are provided in figure 2.9. In general, the cumulative distribution functions' midpoints tend to move to lower values of redshift for subhalos that fall in bins of higher v_{peak} values. Because of the clear trend, during the assignment of ${}^{0.1}(g-r)$ to Rosella galaxies, we interpolate between the cdfs that we calculate in bins of v_{peak}, so that the colour assignment for each galaxy is operating with a z_{form} cdf that is tailored to its host subhalo's v_{peak}.

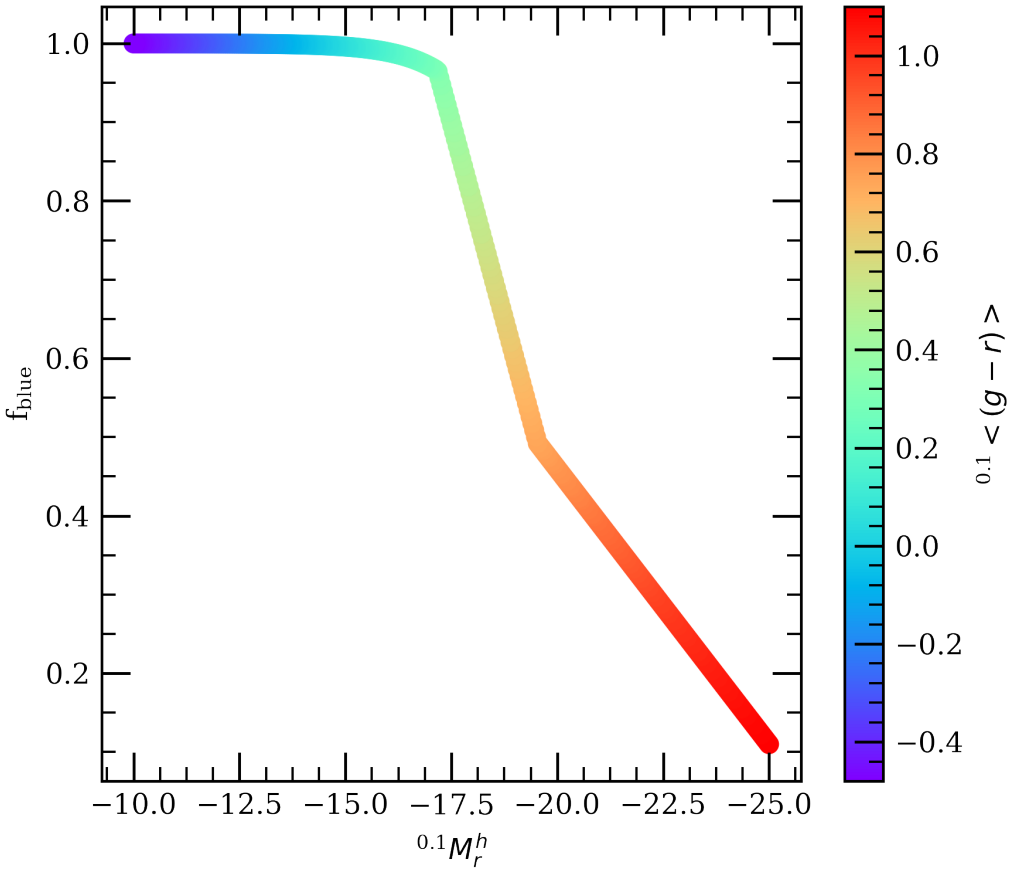


Figure 2.8. Expected fraction of blue galaxies $f_{\rm blue}$ as a function of M_r^h at redshfit 0, colour-coded by mean $^{0.1}(g-r)$. The vertical axis shows the expected fraction of blue galaxies for a population at a given M_r^h (provided in equation 2.10). The curve is colour-coded by the mean expected $^{0.1}(g-r)$ value for a given M_r^h, $<^{0.1}(g-r)>= f_{\rm blue}* <^{0.1}(g-r)>_{\rm blue} +(1-f_{\rm blue})* <^{0.1}(g-r)>_{\rm red}$.

2.6.3 Colour assignment

Next, we connect the cumulative distribution functions of $z_{\rm form}$ to those of $^{0.1}(g-r)$, as illustrated in figure 2.10. In figure 2.10, the $z_{\rm form}$ cumulative distribution function (cdf) for a single subhalo is given by the orange curve, and is conditional on its $\rm v_{peak}$. We trace each subhalo's $z_{\rm form}$ position on the $\rm v_{peak}$-dependent $z_{\rm form}$ cdf to find its "abundance" value. In figure 2.10, the green dashed line going from the top, $z_{\rm form}$, axis down to the orange curve shows us that abundance value, which always falls between 0 and 1 by construction.

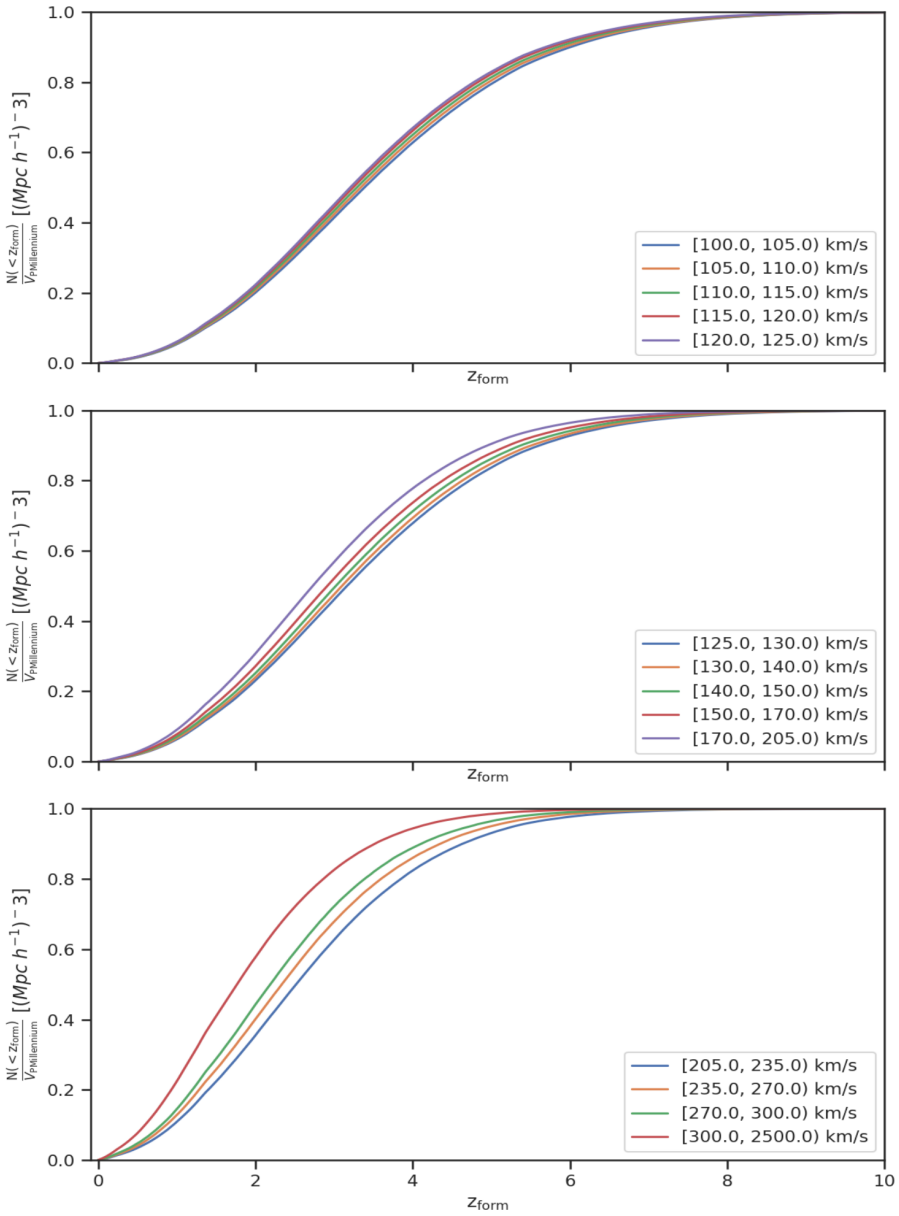


Figure 2.9. Example cumulative distribution functions of formation redshift in bins of v_{peak}. The vertical axes display the number of subhalos with a given z_{form} or lower, normalised so that the maximum of this function is 1. The horizontal axis is the formation redshift z_{form}. Curves are colour coded by bins of v_{peak}. Note that median z_{form} values decrease with increasing median v_{peak} values. This plot was made at the snapshot corresponding to $z \sim 0.0$ during the testing and development of the Rosella colour assignment methodology. The cumulative distribution functions used in the assignment of colours in the final catalogue are separated into finer bins of v_{peak} and cover the range of v_{peak} values between 50 km/s and 2010 km/s at $z \sim 0.2$. During colour assignment, we interpolate between these curves to find an appropriate z_{form} cdf for each subhalo.

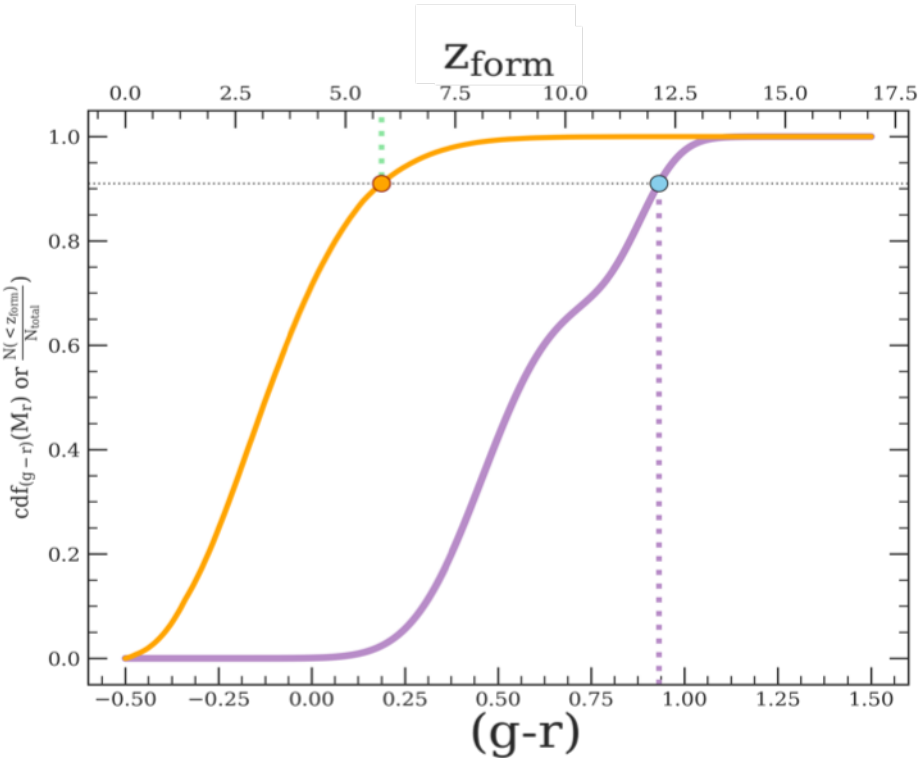


Figure 2.10. Illustration of colour assignment for a single subhalo. The top horizontal axis shows formation redshift, $z_{\rm form}$; the bottom horizontal axis shows ${}^{0.1}(g-r)$. The vertical axis shows cumulative distribution function (cdf) values for the orange and violet curves. The orange curve is the cdf of subhalo $z_{\rm form}$ values for a given $\rm v_{peak}$. 0 on this curve means that no subhalos of the given $\rm v_{peak}$ should be expected to have that or lower $z_{\rm form}$. 1 on the orange curve signifies that all subhalos of the given $\rm v_{peak}$ should be expected to have lower $z_{\rm form}$ values. The orange curve is computed by interpolating between $z_{\rm form}$ cdf curves calculated in bins of $\rm v_{peak}$ (see figure 2.9 for examples). The violet curve is the cdf of ${}^{0.1}(g-r)$ colour computed as the sum of two Gaussian cdfs, normalised by the expected fraction of blue galaxies at a given absolute magnitude (see equation 2.10). During colour assignment, we begin with a subhalo of known $\rm v_{peak}$ (which gives us the orange $z_{\rm form}$ cdf curve) and an M_r^h magnitude assigned to that subhalo's galaxy with SHAM (which we then use to create the violet ${}^{0.1}(g-r)$ cdf curve). Having computed the subhalo's $z_{\rm form}$, we trace the green dashed line from the $z_{\rm form}$ to the orange dot, which provides us with a cdf value. We then trace that value to find the cyan point on the violet colour cdf. The cyan point indicates the ${}^{0.1}(g-r)$ value we assign to the Rosella galaxy residing in this subhalo.

Once we have computed the cumulative distribution function of ${}^{0.1}(g-r)$ for a given subhalo, we find the value of ${}^{0.1}(g-r)$ whose cdf value matches that of this subhalo's $z_{\rm form}$, based on the subhalo's $\rm v_{peak}$-dependent $z_{\rm form}$ cdf. The resulting ${}^{0.1}(g-r)$ value is then assigned to the galaxy residing in the given subhalo and added to Rosella.

2.7 Identifying central galaxies

In order to define which galaxies are central, and which satellite, we create a column in Rosella, "is_central_in_halo". The column corresponds to whether a galaxy lives in a subhalo that is located at the position of the dark matter particle with the lowest energy (i.e. the most bound particle of the halo). Thus, if "is_central_in_halo" is set to 1, a galaxy is located at the position of the most gravitationally bound particle in the dark matter halo identified in P-Millennium by a Friends-of-Friends algorithm. This quantity was created after we noticed that values in the "isFoFCentre" column in P-Millennium DHALO catalogues did not seem to be a good indicator of subhalos hosting central or satellite galaxies, as it implied that only 0.2% of the galaxies in Rosella were centrals.

2.8 Clustering

We use the publicly available code corrfunc (https://github.com/manodeep/Corrfunc, Sinha & Garrison, 2017; Sinha & Garrison, 2019) to calculate the clustering results presented in this work. To calculate the projected correlation function, $\rm w_p(r_p)$, for our mock and for samples we used for tuning free parameters (see section 3.6), we used corrfunc's Theory routine for a periodic simulation box. We calculate $\rm w_p(r_p)$ with a maximum separation between galaxies along the Z dimension set to 80 h^{-1} Mpc.

2.9 Summary of the methodology

This chapter presents a method for creating a mock catalogue of galaxies. We provide the positions and velocities for these galaxies based on position and velocities of dark matter subhalos in the P-Millennium N-body simulation. Our method for assigning absolute magnitudes to galaxies builds on subhalo abundance matching (SHAM), under the initial assumption of a correlated monotonic relation between galaxy M_r^h and the $\mathrm{v_{peak}}$ of its host subhalo. We add scatter to the SHAM absolute magnitudes via shuffling, i.e. by using a functional form that preserves the input galaxy luminosity function and that varies depending on the value of M_r^h assigned with SHAM without scatter.

The colour assignment in our methodology, presented in section 2.6, is conditional to the luminosity assignment of section 2.3, as well as the cumulative distributions of the z_{form}, our proxy for subhalo age, which are conditional on subhalo $\mathrm{v_{peak}}$. The algorithm for computing z_{form} for subhalos is discussed in section 2.5 and relies on a database of subhalo progenitor histories. The latter is obtained using a method for tracing subhalo histories (section 2.4).

A major component of the assessment of the quality of the Rosella mock will be done through clustering analyses and presented in chapter 3. In that chapter, we also consider the fates of central and satellite galaxies in Rosella, and section 2.7 describes our method for identifying whether or not a galaxy is central to its host halo.

Chapter 3

Properties of the Rosella mock catalogue

In this chapter, we examine the data produced using the methodology introduced in chapter 2. In section 3.2, we open with a discussion of the properties of the luminosity function of the galaxies in Rosella and discuss the brightness limits that it can potentially reach. Section 3.3 describes the impact that our model of luminosity scatter has on the distribution of central and satellite galaxies in the mock. Section 3.4 includes a discussion of the clustering in our mock, with a comparison to previously published observational and simulated data. First, we consider luminosity-limited samples, then clustering data for red and blue galaxy populations. We explore the presence of galaxy colour bimodality in Rosella in section 3.5. Finally, we provide a short discussion of the process of tuning the parameters in our mock's methodology in section 3.6.

3.1 Choice of redshift for the mock catalogue

When creating Rosella, we have the option of creating a mock catalogue at any of the 269 snapshots available in P-Millennium. While the highest-redshift snapshots would not be a reasonable choice for this galaxy catalogue, we did consider a few

options at low redshifts. The contenders for the redshift at which we build Rosella included $z \sim 0.0$, $z \sim 0.1$, and $z \sim 0.2$. We performed the initial development of the code behind this mock and preliminary tests on the P-Millennium snapshot that corresponds to $z \sim 0.0$. We considered creating the mock at $z \sim 0.1$ due to the fact that this redshift closely matches the mean redshift of the SDSS survey. The empirical data with which we calibrated our mock is based on SDSS and GAMA, and, consequently, we provide absolute magnitudes and colours k-corrected to $z \sim 0.1$.

We have chosen to create this implementation of Rosella at $z \sim 0.2$. The choice is motivated by the needs of the DESI Bright Galaxy Survey (DESI BGS). The DESI survey will begin to gather data in the near future. BGS will take the spectra of relatively bright galaxies during bright observing time. Consequently, its selection of target galaxies places the median redshift for future BGS observations at $z \sim 0.2$. Rosella will be useful as a reference mock for BGS, for fulfilling tasks that include analysing survey biases and calibrating approximate mocks that meet the volume and abundance requirements of the experiment (DESI Collaboration, 2018).

3.2 Galaxy luminosity function and Rosella resolution

By construction, the implementation of SHAM used here is expected to reproduce its target luminosity function. Figure 3.1 demonstrates that the luminosity function produced in our mock exactly matches the cumulative galaxy luminosity function (LF) based on SDSS (Blanton et al., 2003) and GAMA (Loveday et al., 2012) data provided in Smith et al. (2017).

Note that figure 3.1 shows the cumulative galaxy luminosity function down to $M_r^h = -10$, which is the extent of the LF that we use during SHAM assignment, including the addition of scatter. Before scatter, the abundance value of the $M_r^h = -10$ limit allow us to assign absolute magnitudes to subhalos with $\mathrm{v_{peak}}$ of ~50 km/s and

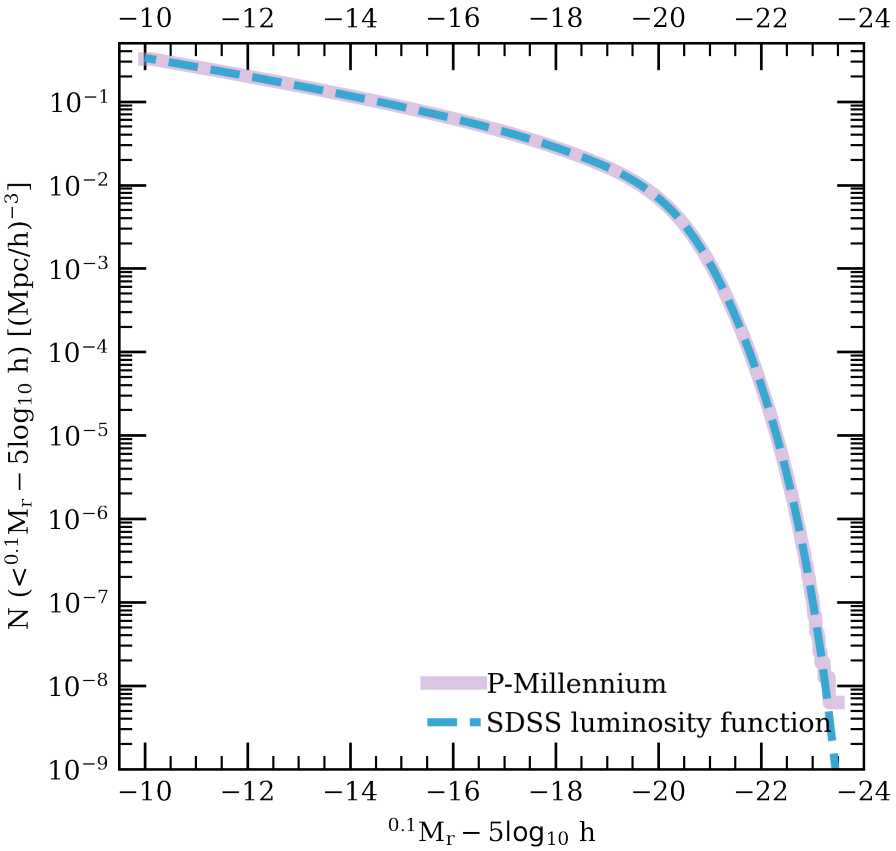


Figure 3.1. The r-band cumulative luminosity function. The function for galaxies in the mock catalogue is plotted in a solid violet line. The blue dashed line is the target luminosity function based on SDSS and GAMA observations, taken from the fit provided in Smith et al. (2017).

greater. We then take a sample of galaxies with a minimum magnitude cut-off to construct the final mock catalogue.

To illustrate the necessity for a minimum absolute magnitude cut when creating the final galaxy sample after including scatter, we turn our attention to figure 3.2.

Figure 3.2 shows SHAM absolute magnitudes before (white curve) and after (hexbin colour map) scatter has been added to M_r^h data. When we add scatter to SHAM data, a portion of the data that is fainter than a certain M_r^h limit reaches a wall on the lower v_{peak} side, rendering the sample incomplete.

The histograms in figure 3.3 offer another way to look at the absolute magnitude-

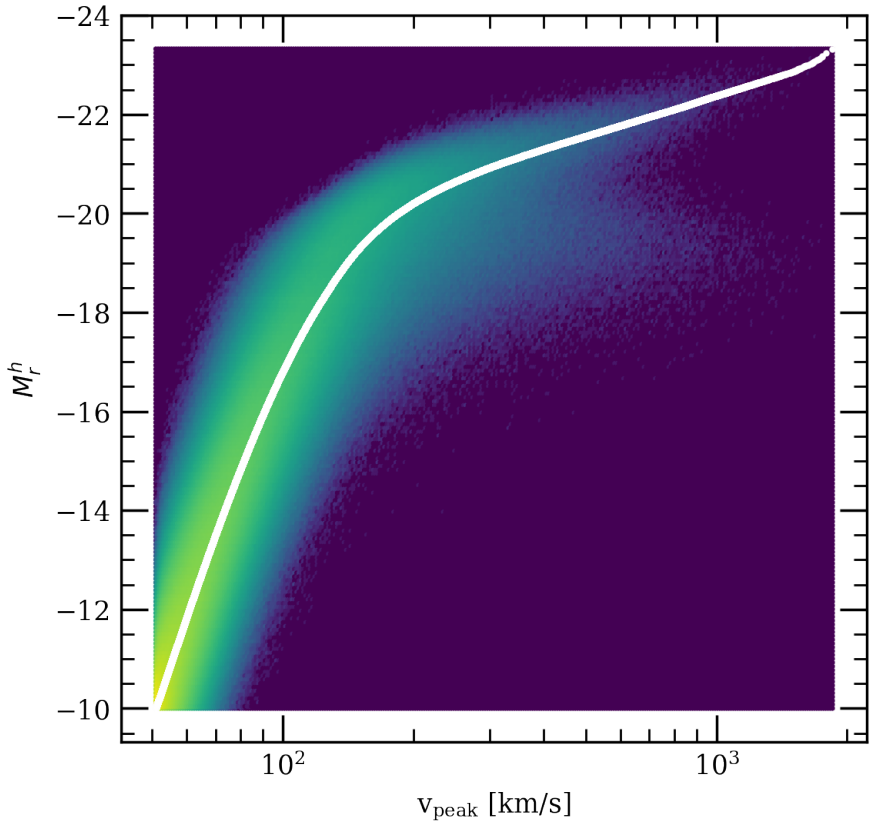


Figure 3.2. Hexbin map of SHAM absolute magnitudes with scatter. The colour indicates the number of galaxies per hexagonal bin of given M_r^h and v_{peak} values, plotted on a logarithmic scale, with dark blue indicating bins with zero galaxies and lime-green indicating bins with the most galaxies. The white line plotted on top of the hexbin map shows the M_r^h values assigned to P-Millennium subhalos before the addition of scatter.

dependent completeness of our scattered SHAM mock. When we plot the distribution of v_{peak} values among subhalos that host galaxies in bins of M_r^h, after the addition of scatter, we see that the distributions of samples with galaxies brighter than $M_r^h = -15.5$ are relatively smooth and taper off before reaching the lower limit of v_{peak}~50 km/s. The sample of galaxies in the $-13 \leq M_r^h < -15.5$ bin, however, is cut off at v_{peak}~50 km/s, rendering the sample of galaxies fainter than M_r^h~15.5 incomplete.

Thus, when built with SHAM constructed from a luminosity function that extends to M_r^h of -10 on the faint end, Rosella is able to reach magnitudes down to about

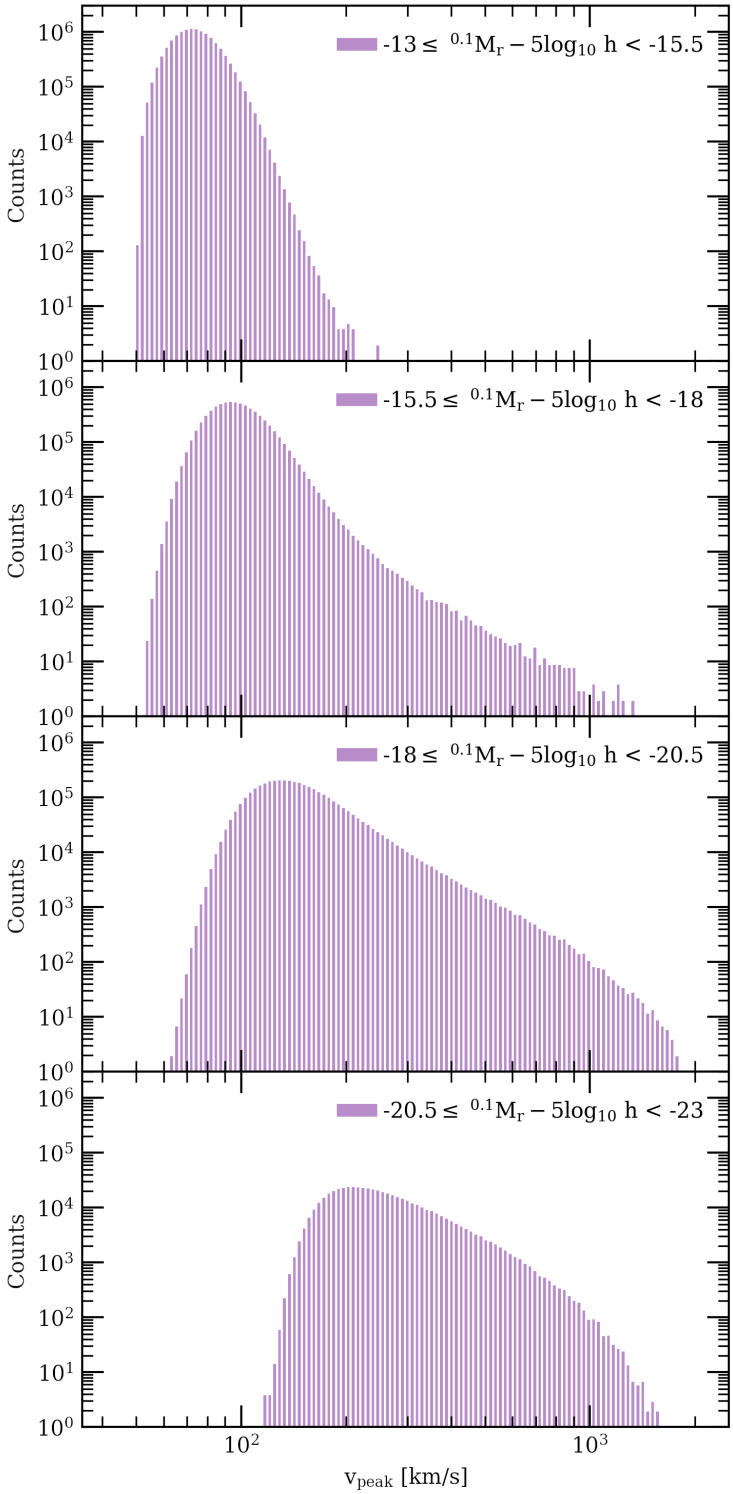


Figure 3.3. Histograms of v_{peak} in bins of M_r^h, created by drawing M_r^h-limited samples, as indicated in the legend, from the full set of galaxies whose range of absolute magnitudes reaches $M_r^h = -10$, depicted in figure 3.2. The histograms in all four panels are calculated over the same set of 125 bins that cover the range 45 km/s $< v_{peak} <$ 2500 km/s.

$M_r^h = -15.5$, thanks to P-Millennium's mass resolution. The lower limit on the absolute magnitude that produces a complete sample of galaxies may vary if one were to add scatter that follows a functional form different from equation 2.3 or has a different set of parameters.

In this work, we present results for a catalogue of galaxies with a faint cut-off set at $M_r^h = -18$, since such a brightness limit is sufficient for the scientific needs of the DESI Bright Galaxy Survey (BGS) at $z \sim 0.2$, which is the redshift chosen for our mock. It should be noted, however, that creating a mock that extends to fainter magnitudes than those presented here is possible with the methodology presented in this work.

3.3 Distribution of central and satellite galaxies

The number of satellite galaxies as a fraction of the total galaxy population in Rosella appears to vary with halo mass. Figure 3.4 shows the general trend, with galaxies that are assigned brighter absolute magnitudes displaying higher fractions of central galaxies. In both the cases of SHAM samples with and without scatter, the trend in the ratio of central galaxies to the total galaxy population tapers off to an almost constant rate of about 60% between $M_r^h = -20$ and $M_r^h = -18$. The scattered sample of SHAM, however, exhibits a lower fraction of satellites compared to the no-scatter sample on the bright end of the catalogue. This is the result of the scattering process moving the magnitudes of galaxies that start out in central subhalos to satellite subhalos.

This trend can be further examined in figure 3.5, which shows the normalised distributions of central and satellite galaxies in bins of host halo mass (defined by the $M_{200,\,\mathrm{mean}}$ mass of the host Friends-of-Friends halos) of 0.5 dex width. There, we see that the no-scatter SHAM sample (bottom panel) exhibits a clear and expected trend of the peak of the distribution of centrals in the catalogue moving to a brighter magnitude with increasing halo mass.

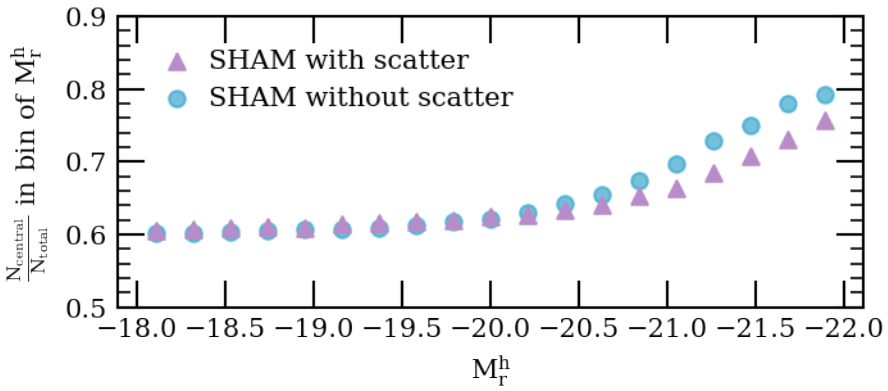


Figure 3.4. Fraction of central galaxies in Rosella with and without scatter as a function of r-band magnitude. The triangles show the fractions of central galaxies in a SHAM sample without scatter. The circles correspond to the fractions of central galaxies present in Rosella with scatter implemented with equation 2.3.

There is a correlation between the v_{peak} values of some, but not all, subhalos and the masses of the halos in which they reside. Figure 3.6 shows that at high halo mass, a separate population of subhalos emerges at highest values of v_{peak}. While the population containing the majority of subhalos illustrated in figure 3.6 stretches to about 600 km/s, regardless of host halo mass, the second population exhibits a halo mass-dependent increase in v_{peak}. One might interpret this as a sign of the correlation between halo mass and the v_{peak} of halos' central subhalos, which is supported by the distribution of central galaxies in figure 3.5. This correlation directly translates into galaxy absolute magnitudes in SHAM without scatter by construction, which is apparent in figure 3.6, which is colour-coded by M_r^h.

The relation between between subhalo v_{peak}, galaxy M_r^h, and halo mass in Rosella is affected by the scatter we add to the mock data. Figure 3.7 shows that for mock data with scatter added using the method outlined in section 2.3, the population of subhalos with high v_{peak} values that correlate with halo mass is still present. The population of galaxies limited to M_r^h of –18 and brighter, once we have added scatter, includes subhalos that reach lower values of v_{peak}, compared to the no-scatter case. Even though it is apparent that fainter galaxies have now been mixed into the high-v_{peak} population, the relation between v_{peak}, M_r^h and halo mass

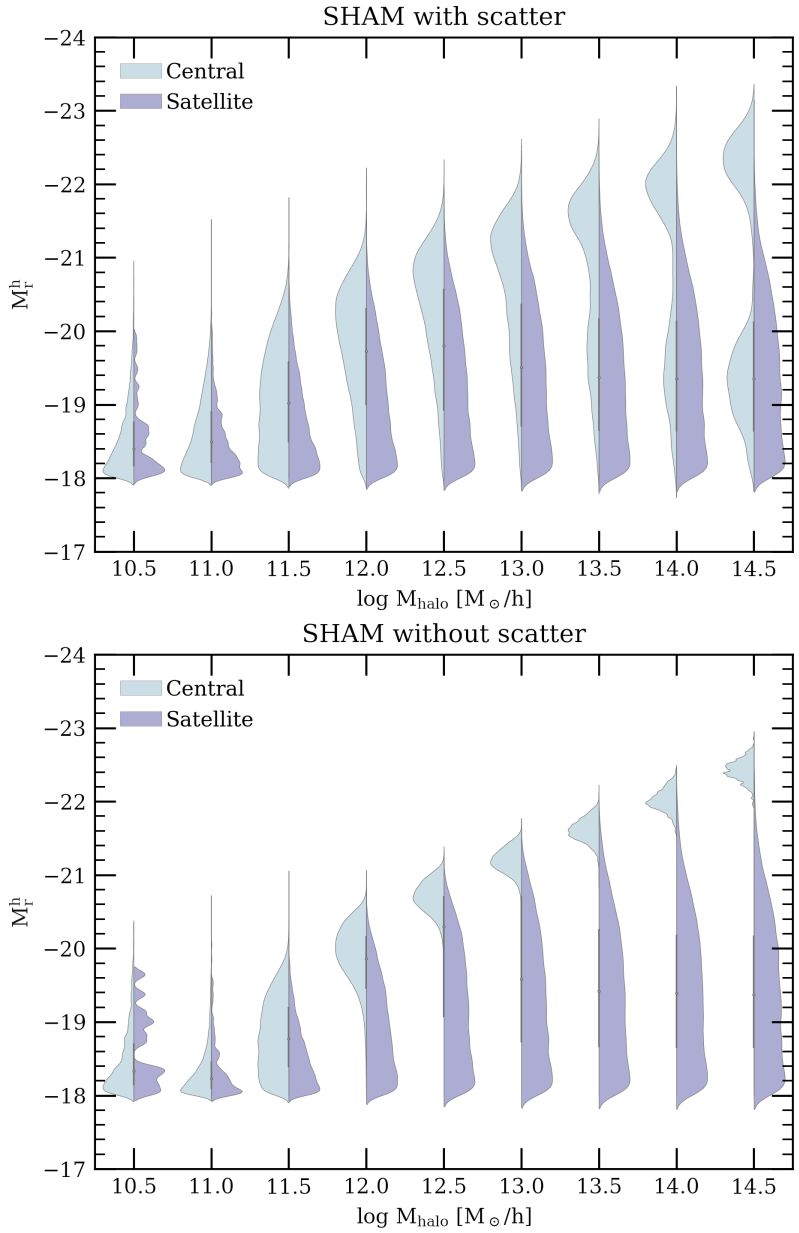


Figure 3.5. Distribution of central and satellite galaxies in halo mass bins for a sample of Rosella galaxies with $M_r^h < -18$. The vertical axis shows M_r^h, and the horizontal axis represents the masses of halos in which galaxies are located (in terms of 200 times the mean density of the universe). The top panel shows the distributions of satellite (light blue) and central (violet) galaxies with respect to their M_r^h values in bins of halo mass in Rosella with scatter described in section 2.3. The bottom panel shows analogous distributions for a SHAM catalogue with no scatter. Kernel smoothing has been applied to these violin histograms, which creates the incorrect appearance of data existing for magnitudes fainter than M_r^h of -18. The plots are normalised in a way that lets all histograms have the same width to draw our attention to the distribution of galaxies along the M_r^h axis, and not to the relative sizes of these populations.

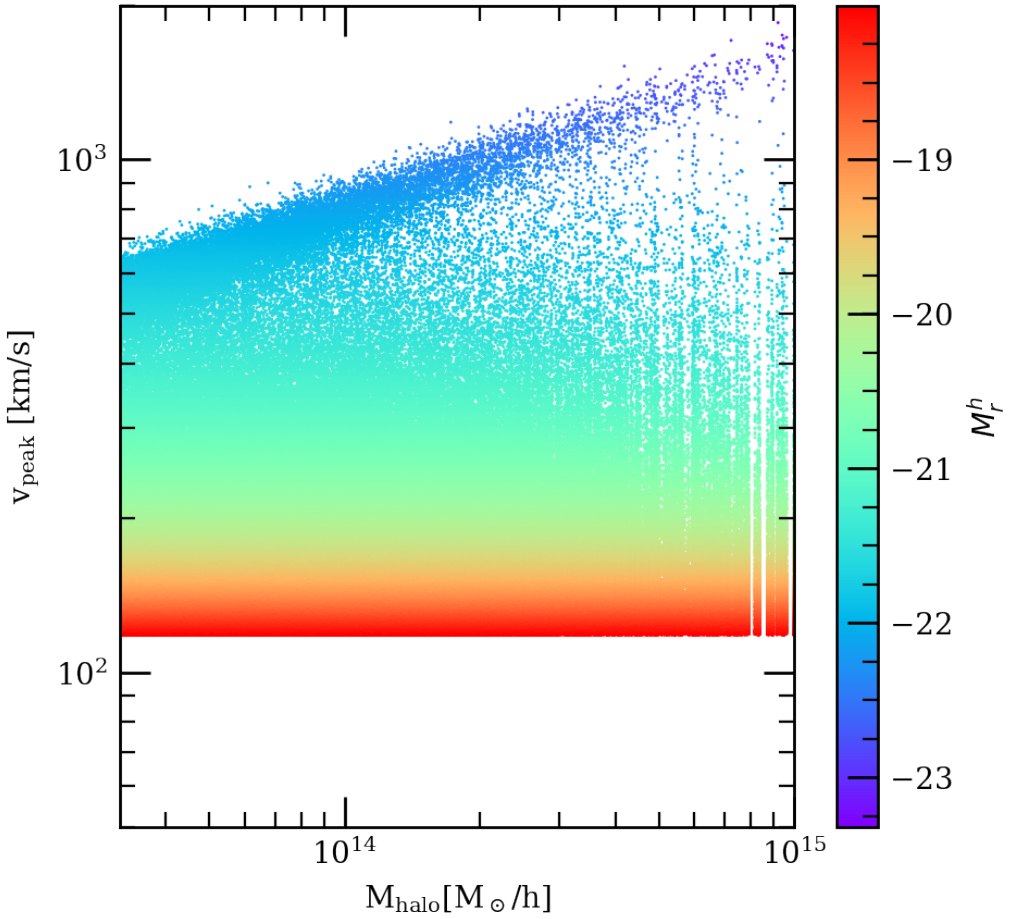


Figure 3.6. Relation between subhalo v_{peak}, galaxy M_r^h, and halo mass for SHAM with no scatter. Each point corresponds to a single galaxy and is colour-coded according to its luminosity. The vertical axes shows the v_{peak} of the subhalos that host these galaxies. The horizontal axis shows the $M_{200,\,mean}$ masses of the Friends-of-Friends halos that contain the aforementioned subhalos.

retains the M_r^h gradient trend that is clear in the no-scatter case of figure 3.6. We still see that the tilted high-v_{peak} population contains a significant amount of galaxies that fall on the brightest end of the galaxy luminosity function for the no-scatter mock.

In figure 3.8, we present information of the same nature as figures 3.7 and 3.6, but for a scatter that follows the step function shown in figure 3.15 instead of a sigmoid to illustrate the effect that various forms of scatter can have on the shape of the dataset. While the tilted high-v_{peak} population that contains central galaxies in

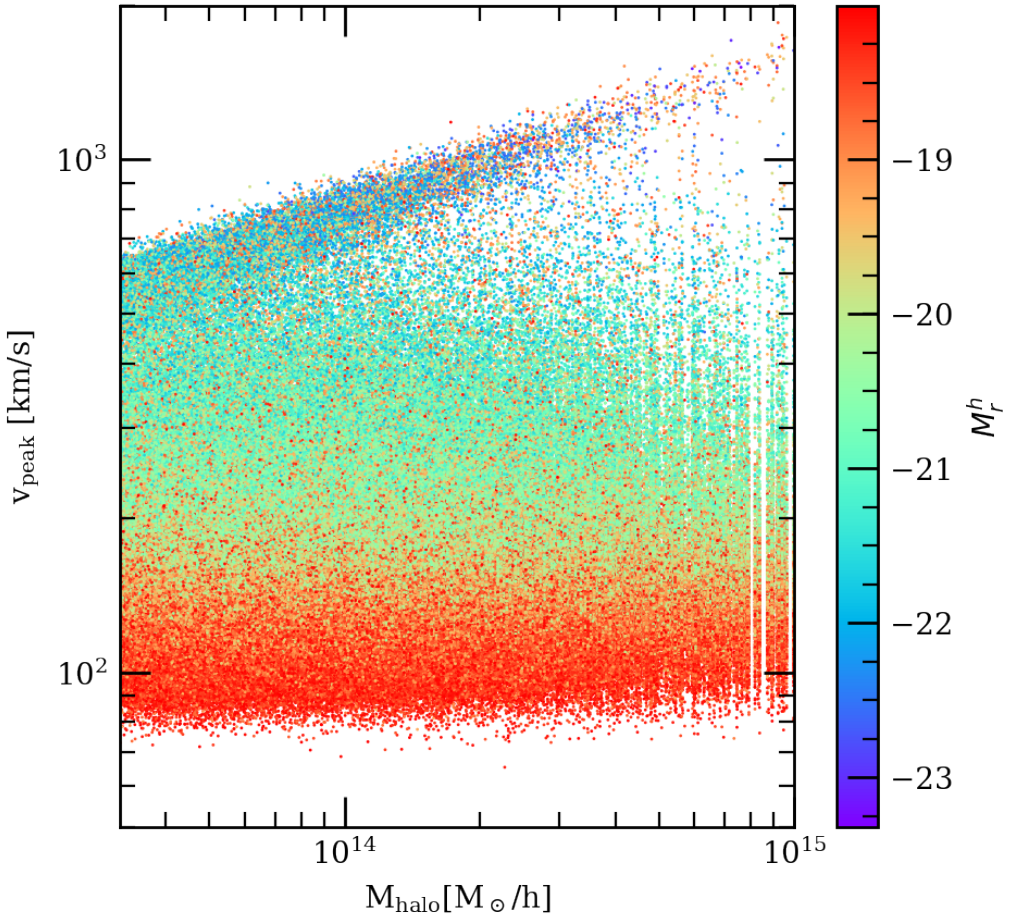


Figure 3.7. Relation between subhalo v_{peak}, galaxy M_r^h, and halo mass for SHAM with nominal scatter. Same as figure 3.6, but with nominal scatter given in equation 2.3 and shown as a solid line in figure 2.3.

the no-scatter data is still present, we can see in figure 3.8 that its colour (which corresponds to the M_r^h values of individual galaxies) shows no apparent difference from the body of galaxies that reside in subhalos whose v_{peak} values do not correlate with halo mass.

3.4 Galaxy clustering

Studies of the clustering of early- and late-type galaxies, classified by spectral type, offer observational evidence of the dependence of the strength of galaxy clustering on morphology and luminosity. Observational evidence points to a trend in

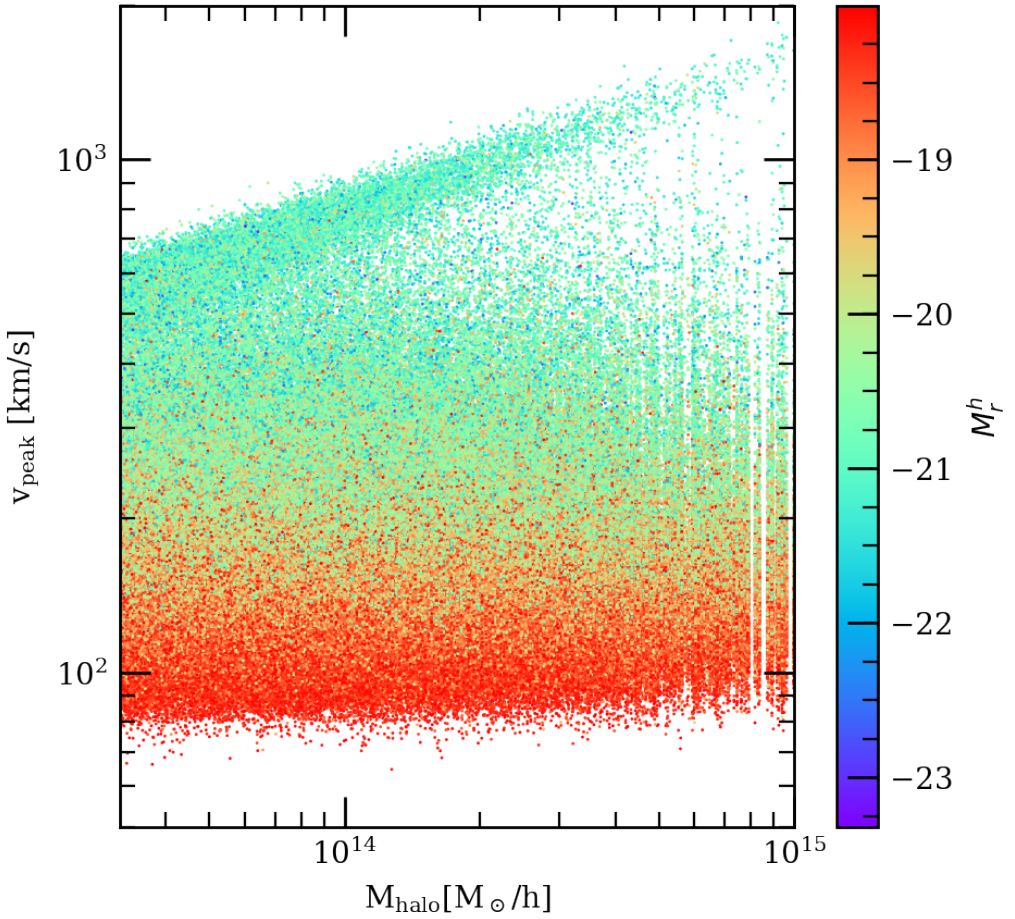


Figure 3.8. Relation between subhalo v_{peak}, galaxy M_r^h, and halo mass for SHAM with a step scatter function shown in figure 3.15. Same as figure 3.6, but with a step scatter function to illustrate the effects of a different level of scatter on mock data.

spatial correlation function, where brighter galaxies are more clustered than their fainter counterparts (e.g. Norberg et al., 2001; Zehavi et al., 2005, and references therein). Early studies of this phenomenon considered red and blue galaxies classified by spectral type, and observed that galaxies that belong to the "early type" population, which has been shown to be dominated by red and quenched galaxies, is more clustered than the "late type" population (e.g. Norberg et al., 2001; Norberg & 2dFGRS Team, 2002; Zehavi et al., 2005, and references therein). The relatively high clustering of more luminous, redder galaxies, has led the luminous red galaxies (LRG) population to be a popular target sample for galaxy surveys

that aim to study the large scale structure of the universe (e.g. Eisenstein et al., 2005a,b).

The luminosities and colours assigned to our high-fidelity mock offer a possibility of comparing the colour- and luminosity-dependent correlation functions to the trends observed in past surveys.

3.4.1 Clustering as a function of luminosity

Projected correlation functions of galaxies in Rosella are shown by the bold curves in figure 3.9 for different luminosity threshold samples at $z \sim 0.2$. In the figure, we show the projected two point correlation functions (2PCF) calculated using the publicly available code corrfunc (Sinha & Garrison, 2017; Sinha & Garrison, 2019).

The samples presented here show the projected 2PCF in samples of galaxies with a faint limit on absolute magnitude (luminosity threshold). The sample cut-off limits have been chosen to make it possible to directly compare the clustering results of Rosella data to those of the HOD mock presented in Smith et al. (2017) and of the SDSS data presented in Zehavi et al. (2011).

While Rosella's projected 2PCF fits the SDSS data quite well on scales greater than 1 h^{-1} Mpc, all but the two faintest samples exhibit clustering that appears to be slightly too high on small scales. We suspect that this might be a result of our SHAM model treating satellite and central galaxies in the same manner.

We have conducted the luminosity-dependent clustering analysis for a variety of models of scatter during the process of tuning our mock, presented in section 3.6.

3.4.2 Clustering as a function of colour

Figure 3.10 shows the projected correlation function of Rosella galaxies separately for red and blue galaxy populations in bins of absolute magnitude. The same figure shows a comparison of our data to those presented in Smith et al. (2017) and Zehavi

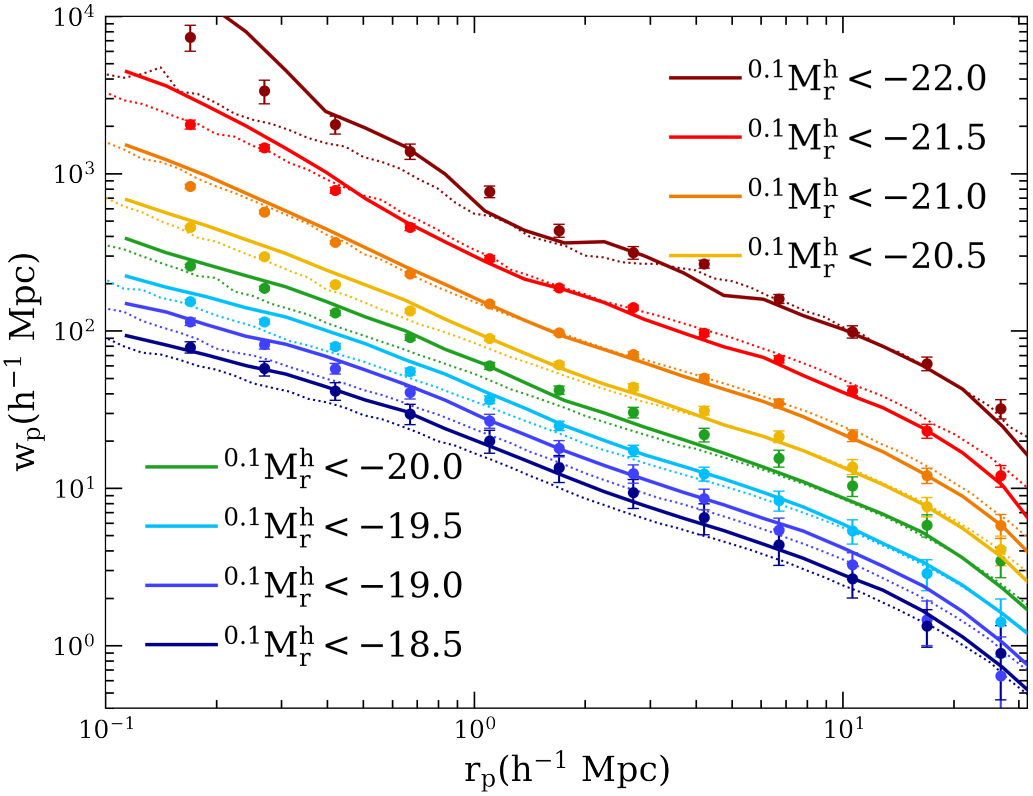


Figure 3.9. Projected correlation function for luminosity threshold galaxy samples. The solid lines show the clustering results of Rosella. The solid points with error bars represent SDSS data, as presented in Zehavi et al. (2011). The dotted lines show the projected correlation functions from the Millennium-XXL mock catalogue in Smith et al. (2017). The results for each sample have been offset by successive intervals of 0.15 dex, starting at the $M_r^h < 20.5$ sample.

et al. (2011), where red and blue samples are defined using the same colour cut as this work's, given by equation 3.1.

For the purposes of analysis, the nominal separation between "red" and "blue" galaxies is given as

$$^{0.1}(g-r)_{\rm cut} = 0.21 - 0.03M_r^h \tag{3.1}$$

Galaxies whose $^{0.1}(g-r)$ values are greater than this $^{0.1}(g-r)_{\rm cut}$ are classified as "red", while the others are "blue". It should be noted that this expression, first introduced in Zehavi et al. (2005), does not account for the fact that there is a continuum in galaxy colours, and instead serves as a tool for comparing colour-dependent clustering among different samples.

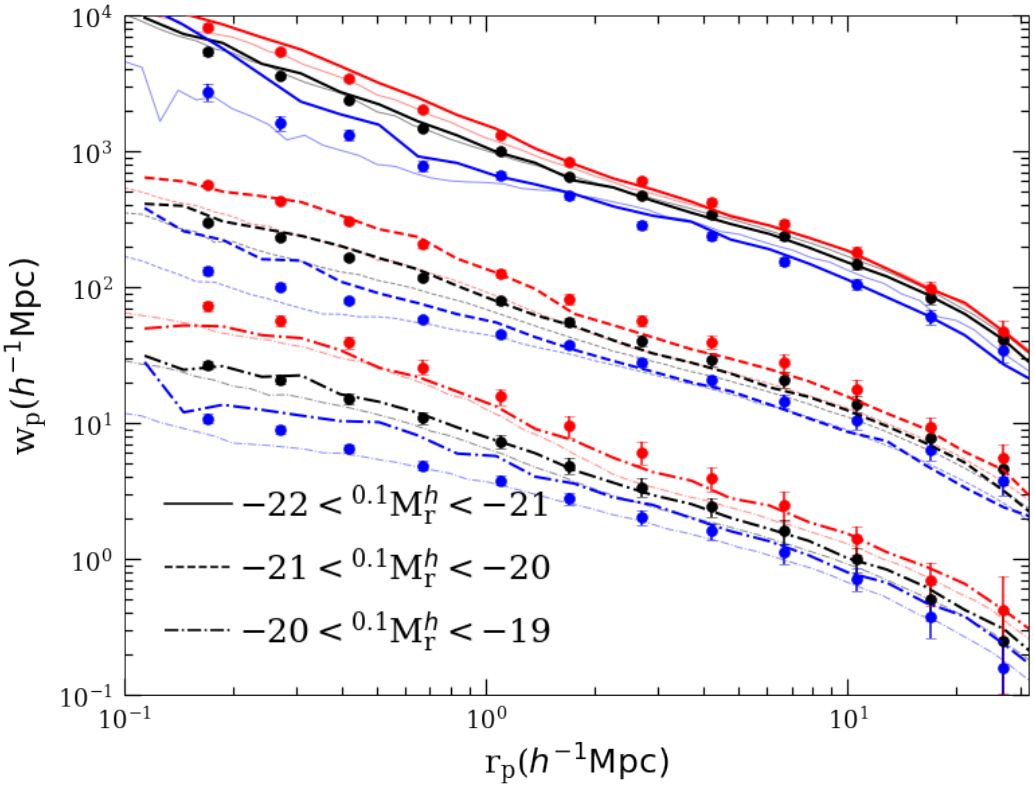


Figure 3.10. Projected correlation function for red and blue galaxies in absolute magnitude-limited bins. The clustering of Rosella galaxies is presented in bold lines. Clustering of low-redshift galaxies from the Millennium-XXL catalogue in Smith et al. (2017) is plotted in faint lines. The styles of lines for both Rosella and Millennium-XXL data correspond to absolute magnitude-limited bins, as indicated in the legend. Points with error bars correspond to the analysis of volume-limited samples of SDSS data in Zehavi et al. (2011). The clustering of all galaxies in a sample is shown in black. Red and blue galaxy populations, defined by equation 3.1, are plotted in red and blue colours, respectively. Magnitude-limited samples are offset from the $-21 < M_r^h < -20$ sample by 1 dex for clarity.

For SDSS, the clustering of the red galaxy population is stronger than that of the blue galaxy population. This effect is likely associated with the presence of red elliptical galaxies, which are more likely to reside in the more strongly biased massive halos (e.g. Eisenstein et al., 2005a). As the samples get fainter, the strength of the colour dependence evidently increases for both the observational data and the galaxies presented in Rosella.

3.5 Galaxy colour bimodality

Figure 3.11 shows histograms of ${}^{0.1}(g-r)$ colour values in Rosella. These histograms show a pattern that we expect to see from previously published distributions of ${}^{0.1}(g-r)$ colours in luminosity-limited samples of galaxies, such as in Smith et al. (2017). The histograms generally show two major peaks. The peaks on the red side, i.e. higher ${}^{0.1}(g-r)$, are more prominent in histograms of more luminous populations, and the peaks of the blue population in the lower ${}^{0.1}(g-r)$ range are more dominant in the fainter samples. Across all samples, the locations of the two peaks move towards the redder end with increasing luminosity, as expected from existing literature, for instance the colour-magnitude diagram of SDSS galaxies shown in figure 1.1.

The histograms in figure 3.11 show a small feature on the blue end; this is not a real physical feature, but rather an artefact of the current implementation of the Rosella code. The colour-magnitude diagram in figure 3.12 excludes this population of galaxies to make the red and blue galaxy populations more apparent once hexbin normalisation is applied.

The diagram in figure 3.12 shows red and blue galaxy populations that are akin to those shown in SDSS data in figure 1.1. The prominence of the red galaxy population tapers off gradually with fainter magnitudes, thanks to the sigmoid feature in the expected fraction of blue galaxies, expressed in equation 2.10.

3.6 Tuning the models of luminosity and colour assignment

The model behind the Rosella mock includes several formulations that may be tuned to find a fit that meets scientific requirements, including:

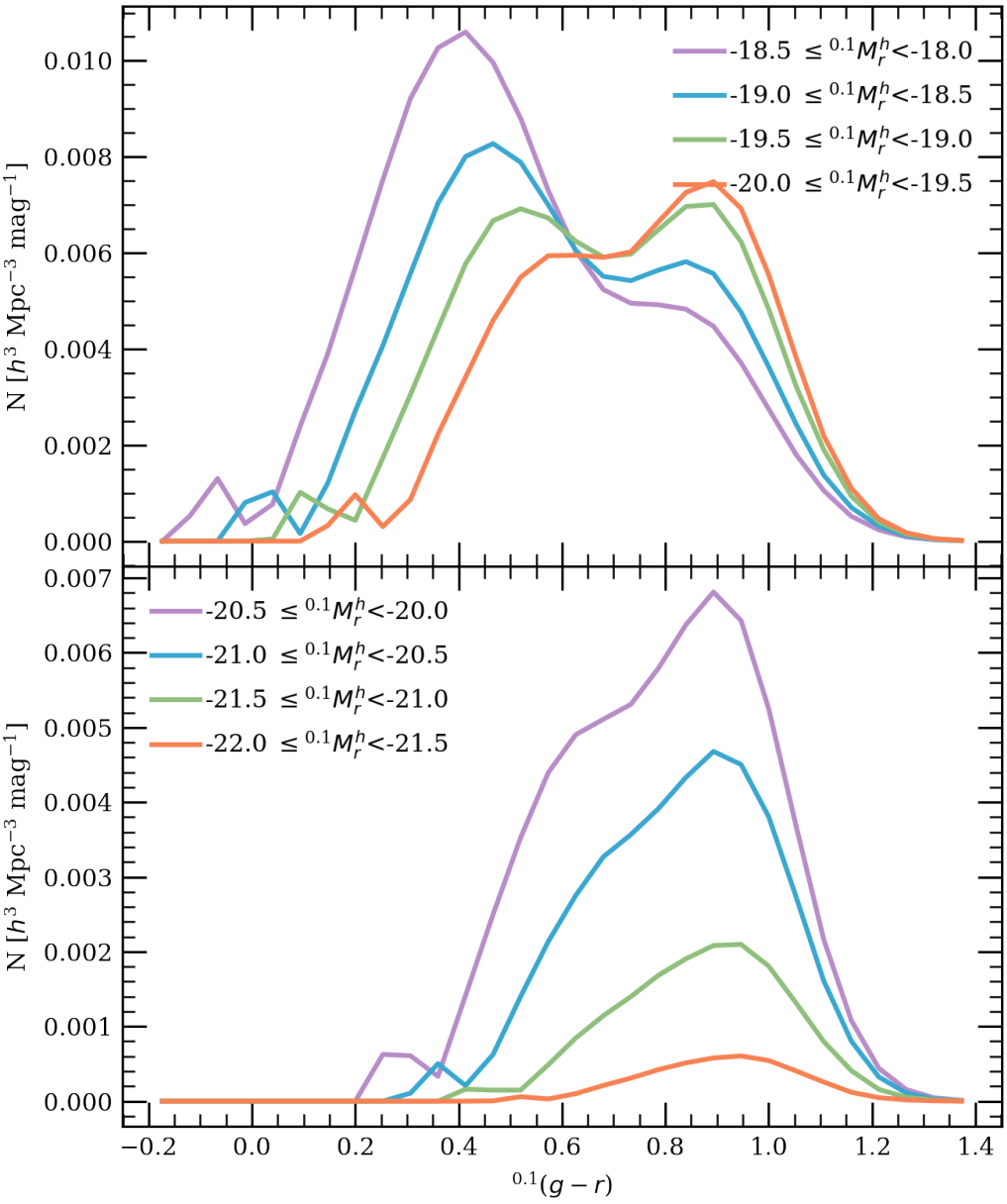


Figure 3.11. Distribution of $^{0.1}(g-r)$ values among Rosella galaxies. Line colours correspond to different ranges of M_r^h values, as shown in the legend. The diagram is split into two panels for clarity.

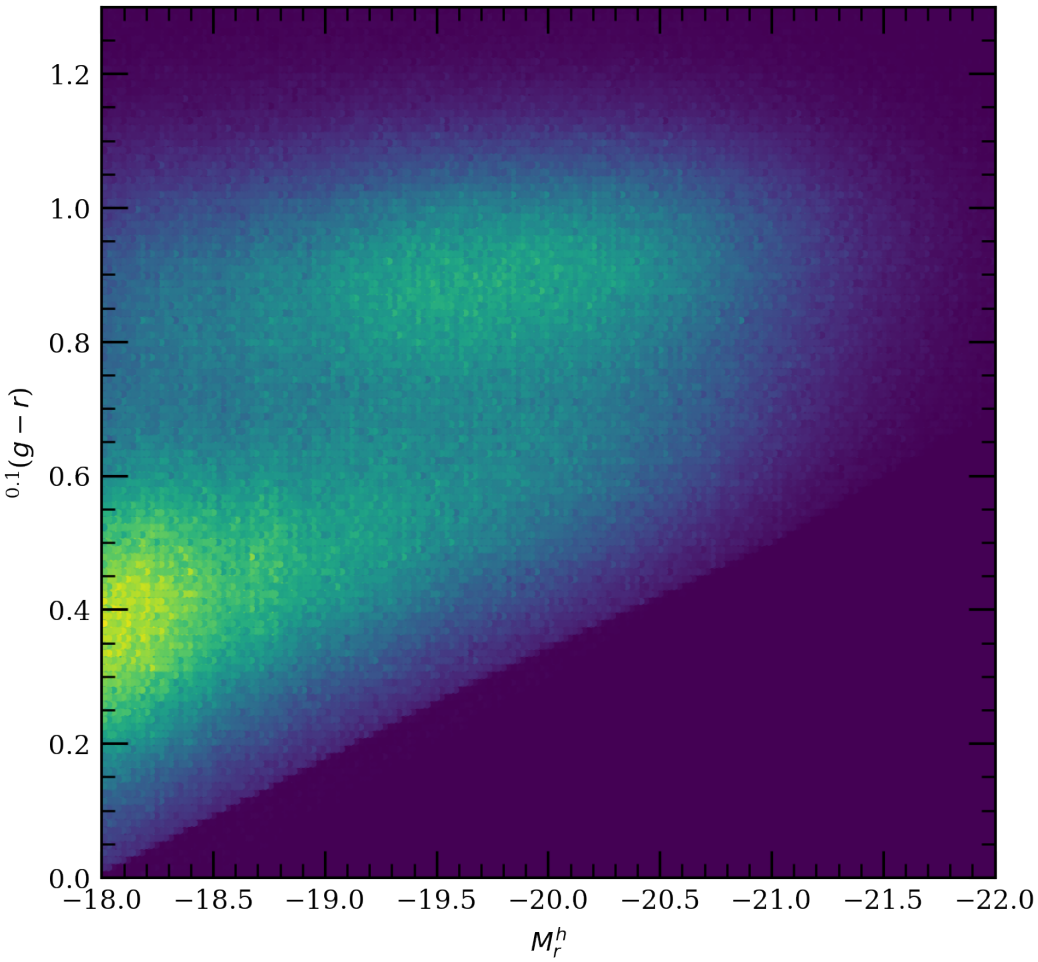


Figure 3.12. Colour-magnitude diagram of Rosella galaxies as a hexbin map. The map shows the density of galaxies in hexagonal bins of $^{0.1}(g-r)$ and M_r^h values.

- The functional model of the scatter added to SHAM M_r^h values (see equation 2.3);
- The free parameters that we use in the scatter function;
- The definition of $z_{\rm form}$(described in section 2.5);
- The fraction of $\rm v_{peak}$ that defines $\rm v_{form}$ and, by extension, $z_{\rm form}$ (see equation 2.4).

The model also depends on the choice of the subhalo property (i.e $\rm v_{peak}$) and the galaxy property (here M_r^h) in the definition of SHAM, represented by equation 2.2.

To find parameters that achieve a satisfactory level of fit to the data, we implemented variations of the scatter function and assessed the mock data that resulted from the various ways to add scatter with luminosity-dependent clustering analysis. In future, it would be desirable to perform a more detailed tuning of the definition of z_{form} to attempt to improve the colour-dependent clustering presented in section 3.4.2.

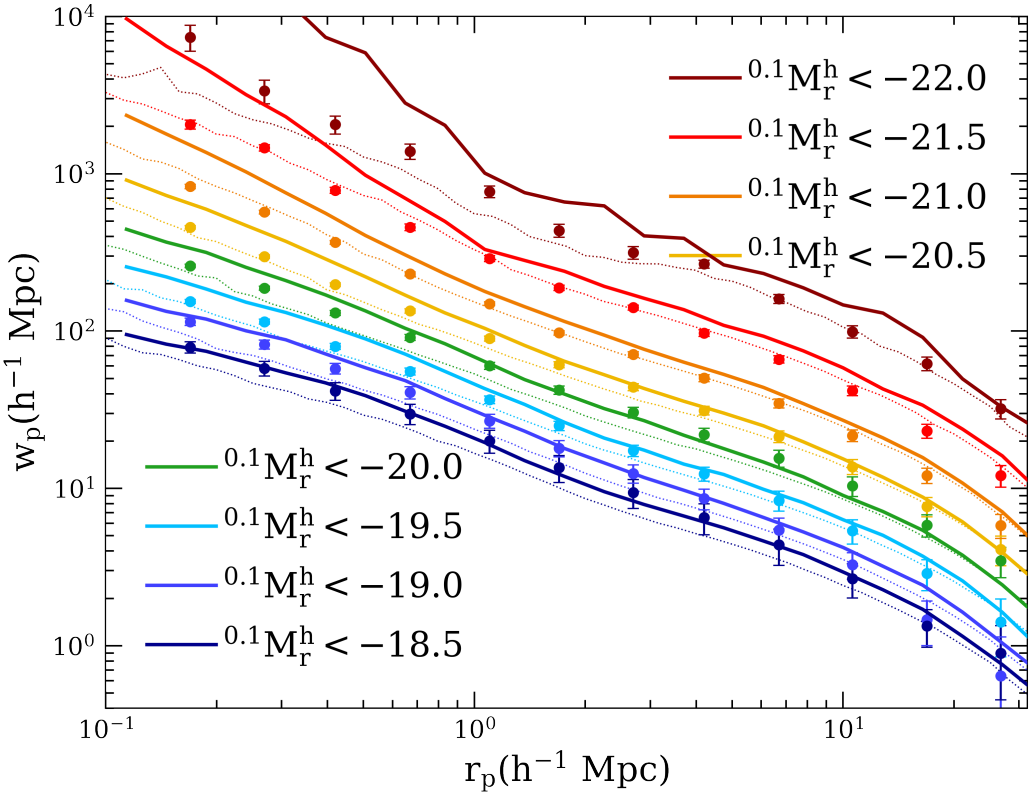


Figure 3.13. Projected correlation function for luminosity-limited samples for SHAM without scatter. See figure 3.9 for a description of line styles. The clustering of the model without scatter is too strong for nearly all samples.

To set a baseline for the luminosity-dependent clustering analysis, we begin by looking at figure 3.13. The data presented in it is analogous to the projected correlation function for luminosity-limited samples shown in figure 3.9, except figure 3.13 shows the projected 2PCF for SHAM *without scatter*. We can see that the correlation function for data without scatter falls above SDSS data in all but the faintest bin. In addition, the slope of the correlation function on small scales

appears to deviate from that of the observations by a noticeable amount.

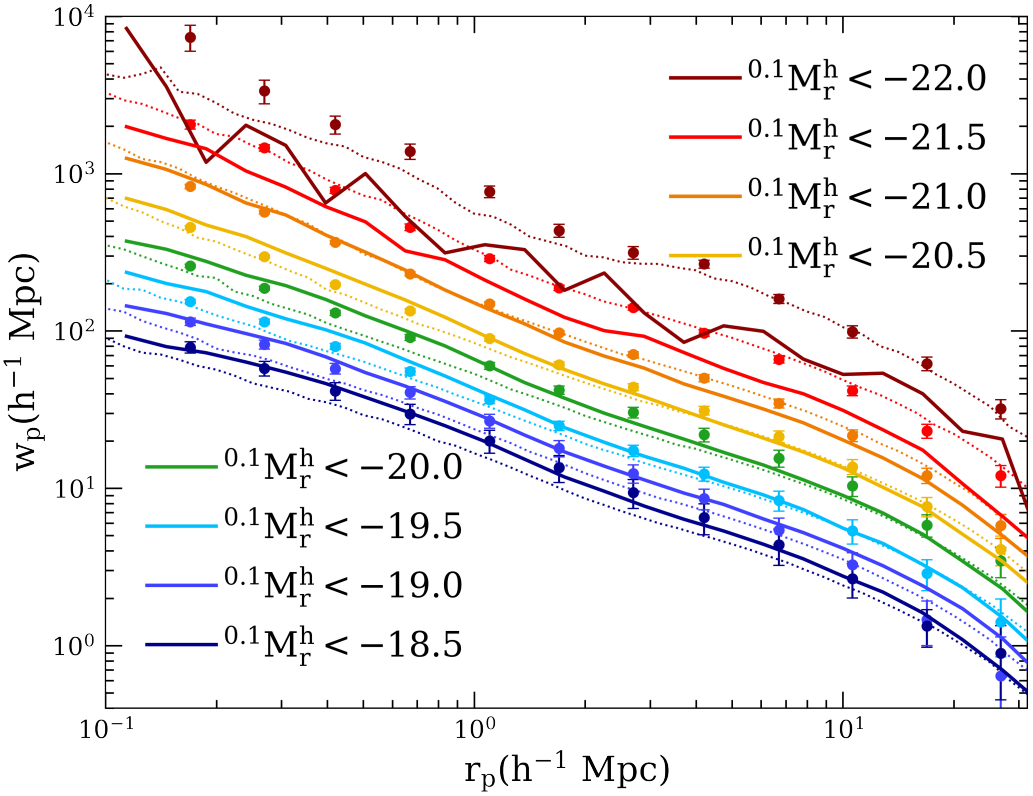


Figure 3.14. Projected correlation function for SHAM with a step scatter function based on Mccullagh et al. (2017) (shown in figure 3.15). See figure 3.9 for a description of line styles.

We might then look at figure 3.14, the clustering results produced by SHAM with scatter that follows a step function depicted in figure 3.15. Unlike the no-scatter case, this dataset produces galaxies whose clustering falls below observations in the brighter bins.

The observations have inspired us to create a scatter function that depends on galaxy luminosity, as such a dependence appeared to bring the projected 2PCF down from the values in figure 3.13 to those in figure 3.14. We also chose to write the scatter function in the form of the sigmoid shown in 2.3, which smooths out the distribution of absolute magnitudes with respect to subhalo v_{peak}, as seen in figure 3.16.

The maximum of the smooth luminosity-dependent scatter lies at about the same

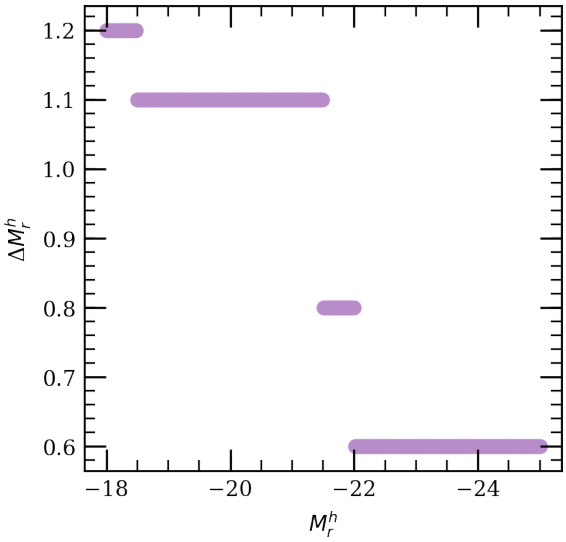


Figure 3.15. Scatter step function following the prescription of Mccullagh et al. (2017) for the amount of scatter added during the "shuffle" part of SHAM with scatter (step 3a in section 2.3).

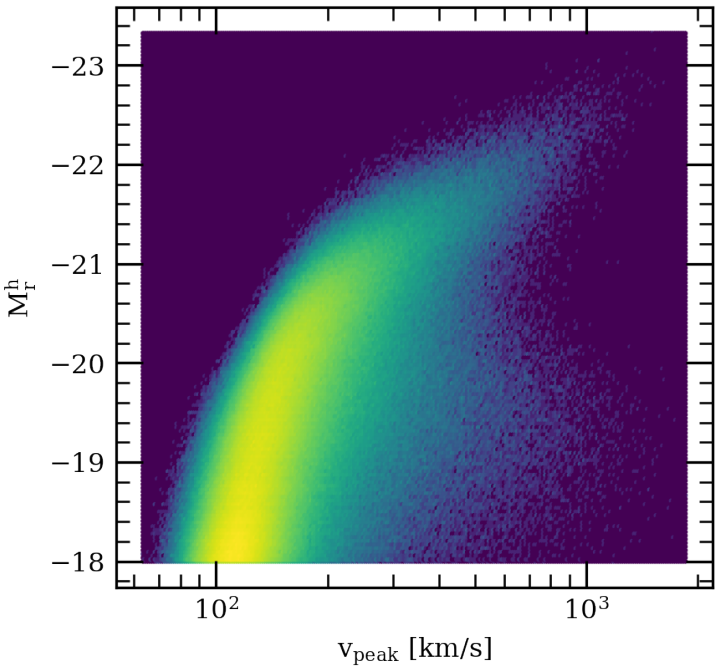


Figure 3.16. Hexbin map of the absolute magnitude-v_{peak} relation for SHAM with a smooth scatter function shown in figure 2.3. The colour of the hexagonal bins indicates the number of galaxies that fall in a given M_r^h-v_{peak} bin. The density is plotted on a logarithmic scale and includes the Rosella mock that covers galaxies brighter than M_r^h of –18, with fainter galaxies omitted here.

level of 1.2 as the maximum of the step function shown in figure 3.15. However, at the brightest magnitudes, our functional form of scatter decreases to a smaller amplitude compared to that shown in figure 3.15. The combination of luminosity dependence, a high amplitude on the brighter end and a low amplitude of scatter on the lower end produces the Rosella scattered data whose projected 2PCF is shown in figure 3.9.

Chapter 4

Conclusion

Dark energy, the component of the energy density of the Universe that is driving the present-day acceleration of cosmic expansion, is of great interest to modern cosmologists. Modern cosmological surveys, such as the Dark Energy Spectroscopic Instrument (DESI) (DESI Collaboration, 2016), aim to deepen our understanding of this phenomenon by creating three-dimensional maps of the large-scale structure of the Universe.

To prepare for cosmological surveys, realistic mock galaxy data is necessary. Mock galaxy catalogues can be used, for example, to guide survey strategy, develop analysis techniques, understand systematic effects that impact the statistics that the surveys aim to measure, and evaluate theoretical models via comparisons with data once it is collected.

This thesis describes the development of Rosella: a mock galaxy catalogue tailored for DESI's Bright Galaxy Survey. The mock has been designed as a reference catalogue that can be used as a training dataset for the creation of a multitude of future mocks that cover greater volumes on the way to meet the need for thousands of mocks that sample volumes of about 200 h^{-3} Gpc^3 (DESI Collaboration, 2018).

In this work, we have developed a method for the assignment of absolute magnitudes to galaxies with a novel formulation for adding scatter to SHAM data. We have also described our method for assigning ${}^{0.1}(g-r)$ colours to galaxies, building

on a foundation of SHAM and an assumption of a correlated relationship between dark matter subhalo age and galaxy colour.

We have demonstrated the properties of the galaxies in Rosella via analyses of luminosity- and colour-dependent clustering, as well as the effect that adding scatter to SHAM data has on the distribution of central and satellites, as well as their clustering.

In the following sections, we conclude this thesis by discussing the possible future directions that this work might take.

4.1 Extending this technology to new purposes

The catalogue presented here has been developed with the needs of the DESI Bright Galaxy Survey (BGS) in mind. Another existing mock, also tailored for BGS, provides lightcone data by populating the Millennium-XXL N-body simulation using an HOD method (Smith et al., 2017). However, the Millennium-XXL mock data has some limitations that we have aimed to address with Rosella. While we do not construct a lightcone catalogue in this work, Rosella in its present state provides a reference mock at a single redshift. If extended to more simulation snapshots and combined with a method for creating lightcone data, the technology presented here can form a basis for a lightcone catalogue that addresses some of the shortcomings of the data set presented in Smith et al. (2017). The limitations of the HOD catalogue in Smith et al. (2017) are primarily connected to the fact that the Bright Galaxy Survey will observe some galaxies that reside in halos not resolved in Millennium-XXL.

While the results presented here correspond to a catalogue that extends down to $M_r^h = -18$, it is possible to expand Rosella down to include galaxies as faint as $M_r^h = -15.5$ with the current implementation of scatter in SHAM absolute magnitudes. We discuss the reasoning behind this possibility in section 3.2.

The need for Rosella is motivated in part by the difficulty of matching clustering data with an HOD mock, such as the MXXL catalogue (Smith et al., 2017). It is promising to see evidence that Rosella's galaxy clustering is capable of fitting observations to a greater degree than the results of Smith et al. (2017), as can be seen in figure 3.9. More work can be done, however, to improve the colour-dependent clustering of Rosella data, as well as the overall 2-point correlation function fit on the small scales.

4.2 Deepening the approach to tuning the mock

In the future, it will be desirable to complement the catalogue tuning presented in this work. This can be accomplished through extending the methods of assessment of the quality of our mock data, as well as by extending the comparison data set. Tuning to SDSS data, as done in this work, could be limited by the SDSS Great Wall, the large cosmic structure that can be seen in figure 1.2, which makes some of the SDSS datasets unrepresentative via a non-negligible influence on some projected correlation function statistics (e.g. Zehavi et al., 2011). The observational data to which Rosella's statistics are compared can also be extended to other datasets, including DESI's BGS data, which will be collected in the near future.

The assessment methodology can be extended with statistical evaluations of the goodness of the fit of the Rosella mock data to observations. The choice of the proper evaluation tool depends on the scientific application for Rosella.

The data presented in section 3.4.2 demonstrates that the clustering of the blue population of galaxies in Rosella consistently exceeds that of SDSS data on scales smaller than 1 h^{-1} Mpc. The clustering of the red population of Rosella galaxies, on the other hand, shows a tendency to be weaker than observations on small scales for the faintest samples. Further tuning of the mock, primarily in the way that we compute subhalo z_{form} values, could potentially improve Rosella's fit to data.

The analysis of colour-dependent clustering can be extended from the bimodal red-blue population analysis, as it is presented in section 3.5, with an assessment of the colour-dependent clustering of Rosella galaxies in finer bins of ${}^{0.1}(g-r)$, as done in, for example, Zehavi et al. (2011).

It would also be useful to conduct an analysis of the luminosity- and colour-dependent halo occupation distribution of Rosella. This analysis can then provide another tool for comparing our catalogue to other datasets, and can also be used to inform future methods for populating N-body simulations to create mock galaxy data.

4.3 Prospects of treating centrals and satellites separately

As seen in figures 3.6, 3.7, and 3.8, it is apparent that the form of the scattering function we use during SHAM assignment of M_r^h affects the distribution of the brightnesses of central and satellite galaxies. We see a clear trend in the no-scatter panel of figure 3.5: a tight group of galaxies resides in central subhalos, distributed in an approximately Gaussian manner, the mean M_r^h of which increases with increasing FoF halo mass. The M_r^h distribution in the population of galaxies in satellite subhalos, on other hand, appears to show little to no dependence on FoF halo mass.

When we add scatter with the algorithm described in equation 2.3, the picture changes somewhat. The distribution of M_r^h values assigned to galaxies living in central subhalos becomes bimodal. This has to do with the shifting of the absolute magnitudes initially assigned to satellite subhalos, as well as subhalos that reside in FoF halos that are less massive, into the galaxies that reside in the central subhalos, which form a tight group on the high end of the $\mathrm{v}_{\mathrm{peak}}$ distribution.

The development of this bimodal distribution of central galaxies in the distribution

of galaxies across the range of M_r^h in Rosella is a result of the way we have chosen to implement scatter in our implementation of SHAM. During the absolute magnitude assignment procedure, we treat all subhalos – central and satellite – in the same manner. This leads to the development of the bimodal distribution of central galaxies seen in the top panel of 3.5. We suspect that this may also be the culprit behind the systematically high small-scale clustering of the brighter samples of Rosella, as seen in figure 3.9.

One possibility that can be explored in the future is treating central and satellite subhalos as distinct populations in the scatter algorithm. This might mean simply executing the scatter algorithm described in section 2.3 separately for subhalos identified as central or satellite using the procedure in section 2.7. Alternatively, this implementation might benefit from creating separate parametrisations of the absolute magnitude perturbation equation 2.3 (this approach has been proposed as a modification of SHAM and goes under the name SCAM in Guo et al., 2016). The decision about the appropriate direction for the treatment of satellite and central subhalos would need to be informed with clustering comparisons and similar assessments.

4.4 Comparison to hydrodynamic simulations

It might be interesting to compare the distribution of colours and absolute magnitudes to those in a hydrodynamic simulation, such as, for instance, EAGLE (Crain et al., 2015). A comparison to a hydrodynamic simulation rather than observational data would provide the benefit of knowing the input cosmology of both the Rosella mock data and that of the hydrodynamic simulation. It would also mean that biases associated with observational data, such as Malmquist bias, the absorption of certain wavelengths of light by dust, or the surface brightness limitations of optical telescopes, among others, are not influencing the comparison.

A comparison with a hydrodynamic simulation, however, would need to be con-

ducted while keeping in mind the differences in relative volume. As mentioned in section 1.4.2, the currently available volumes of cosmological hydrodynamic simulations tend to be around [100 Mpc]^3, with the exception of the lower-resolution IllustrisTNG box of 300 Mpc per side (e.g. Pillepich et al., 2018).

These volumes are much smaller than the P-Millennium box of 800 Mpc per side. Since a major portion of our assessment of Rosella builds on clustering analyses, the volume difference between hydrodynamic simulations and Rosella would impact the level to which clustering comparisons can be reliably made.

A comparison of SHAM results to a hydrodynamic simulation has been carried out in Chaves-Montero et al. (2016). However, that paper tuned its sample with data from a hydrodynamic simulation, instead of observational data like we do here.

4.5 Access to developed code

The code developed for the work presented here is stored in a private git repository on www.github.com. Should the code be useful for future use, be it to implement the adjustments to the free parameters in Rosella, or to extend Rosella's application beyond its current one, the code may be shared with the author's permission.

One possible use of this mock is a comparison to the target assignments, using a process called fibre assignment. This can be accomplished with the help of fibre assignment routines developed for DESI and available on the NERSC supercomputer, as well as on www.github.com. A tutorial* with fibre assignment on proto-Rosella data has been created and made available on github by Jaime Forero-Romero.

*https://github.com/forero/quickfiberassignmock/blob/master/QuickFiberAssignMock.ipynb

Bibliography

Baldry I. K., Balogh M. L., Bower R. G., Glazebrook K., Nichol R. C., Bamford S. P., Budavari T., 2006, MNRAS, 373, 469

Banerji M., Abdalla F. B., Lahav O., Lin H., 2008, MNRAS, 386, 1219

Baugh C. M., 2006, Reports on Progress in Physics, 69, 3101

Baugh C. M. et al., 2019, MNRAS, 483, 4922

Benson A. J., Cole S., Frenk C. S., Baugh C. M., Lacey C. G., 2000, MNRAS, 311, 793

Berlind A. A., Weinberg D. H., 2002, ApJ, 575, 1

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

Blanton M. et al., 2017, AJ, 154, 28

Blanton M. R. et al., 2003, ApJ, 592, 819

Caldwell R. R., Linder E. V., 2005, Phys. Rev. Lett., 95, 141301

Cassata P. et al., 2008, A&A, 483, L39

Charlier C. V. L., 1922, Meddelanden fran Lunds Astronomiska Observatorium Serie I, 98, 1

Chaves-Montero J., Angulo R. E., Schaye J., Schaller M., Crain R. A., Furlong M., Theuns T., 2016, MNRAS, 460, 3110

Chester C., Roberts M. S., 1964, AJ, 69, 635

Cole S., Hatton S., Weinberg D. H., Frenk C. S., 1998, MNRAS, 300, 945

Cole S., Lacey C. G., Baugh C. M., Frenk C. S., 2000, MNRAS, 319, 168

Colless M. et al., 2001, MNRAS, 328, 1039

Conroy C., Wechsler R. H., Kravtsov A. V., 2006, ApJ, 647, 201

Cooray A., 2005a, MNRAS, 364, 303

Cooray A., 2005b, MNRAS, 363, 337

Cooray A., 2006, MNRAS, 365, 842

Cooray A., Sheth R., 2002, 372, 1

Crain R. A. et al., 2015, MNRAS, 450, 1937

Croton D. J. et al., 2016, ApJS, 222, 22

Davis M., Huchra J., 1982, ApJ, 254, 437

Davis M., Peebles P. J. E., 1983, ApJ, 267, 465

Davis M., Huchra J., Latham D. W., Tonry J., 1982, ApJ, 253, 423

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

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

DeRose J. et al., 2019, arXiv e-prints, arXiv:1901.02401

DESI Collaboration, 2016, DESI Final Design Report Part I: Science,Targeting, and Survey Design

DESI Collaboration, 2018, DESI Cosmological Simulations Requirements Document. Tech. Rep. v1.0, Dark Energy Spectroscopic Instrument

Desmond H., Wechsler R. H., 2015, MNRAS, 454, 322

Efstathiou G., Sutherland W. J., Maddox S. J., 1990, Nature, 348, 705

Einstein A., 1917, Sitzungsber. Preuss. Akad. Wiss. Berlin (Math. Phys.), 142

Eisenstein D. J. et al., 2005a, ApJ, 633, 560

Eisenstein D. J., Blanton M., Zehavi I., Bahcall N., Brinkmann J., Loveday J., Meiksin A., Schneider D., 2005b, ApJ, 619, 178

Gao L., White S. D., 2007, MNRASLetters, 377

Gao L., Springel V., White S. D. M., 2005, MNRASLetters, 363, 66

Gonzalez-Perez V., Lacey C. G., Baugh C. M., Lagos C. D. P., Helly J., Campbell D. J. R., Mitchell P. D., 2014, MNRAS, 439, 264

Guo H. et al., 2016, MNRAS

Guo Q., White S., Angulo R. E., Henriques B., Lemson G., Boylan-Kolchin M., Thomas P., Short C., 2013, MNRAS, 428, 1351

Guth A. H., Pi S. Y., 1982, Phys. Rev. Lett., 49, 1110

Hawking S., 1996, A Brief History of Time. Bantam Books

Hearin A. P., 2015, MNRAS, 451, L45

Hearin A. P., Watson D. F., 2013, MNRAS, 435, 1313

Hearin A. P., Watson D. F., Becker M. R., Reyes R., Berlind A. A., Zentner A. R., 2014, MNRAS, 444, 729

Hearin A. P., Zentner A. R., Van Den Bosch F. C., Campbell D., Tollerud E., 2016, MNRAS, 460, 2552

Hogg D. W. et al., 2004, ApJ, 601, L29

Hubble E., 1929, PNAS, 15, 168

Hubble E., Humason M. L., 1934, Contributions from the Mount Wilson Observatory, 3, 85

Huchra J., Davis M., Latham D., 1983, A survey of galaxy redshifts. IV: The data

Huterer D., Turner M. S., 1999, Phys. Rev. D, 60, 081301

Ilbert O. et al., 2009, ApJ, 690, 1236

Ivezić Ž. et al., 2008, arXiv e-prints

Jenkins A., 2010, MNRAS, 403, 1859

Jenkins A., 2013, MNRAS, 434, 2094

Jiang L., Helly J. C., Cole S., Frenk C. S., 2014, MNRAS, 440, 2115

Jouvel S. et al., 2014, A&A, 562, A86

Kauffmann G., White S. D. M., Guiderdoni B., 1993, MNRAS, 264, 201

Khandai N., Di Matteo T., Croft R., Wilkins S., Feng Y., Tucker E., DeGraf C., Liu M. S., 2015, MNRAS, 450, 1349

Klypin A., Prada F., 2018, arXiv e-prints, arXiv:1809.03637

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

Kulier A., Ostriker J. P., 2015, MNRAS, 452, 4013

Lacey C. G., Cole S., 1993, MNRAS

Lacey C. G. et al., 2016, MNRAS, 462, 3854

Lehmann B. V., Mao Y. Y., Becker M. R., Skillman S. W., Wechsler R. H., 2017, ApJ, 834, 1

Lemaître G., 1927, Annales de la Sociètè Scientifique de Bruxelles, 47, 49

Lemaître G., 1931, MNRAS, 91, 483

Lin Y. T., Mohr J. J., Stanford S. A., 2004, ApJ, 610, 745

Loveday J. et al., 2012, MNRAS, 420, 1239

Mandelbaum R., Tasitsiomi A., Seljak U., Kravtsov A. V., Wechsler R. H., 2005, MNRAS, 362, 1451

Marín F. A., Wechsler R. H., Frieman J. A., Nichol R. C., 2008, ApJ, 672, 849

Martel H., Shapiro P. R., Weinberg S., 1998, ApJ, 492, 29

Masaki S., Hikage C., Takada M., Spergel D. N., Sugiyama N., 2013a, MNRAS, 433, 3506

Masaki S., Lin Y. T., Yoshida N., 2013b, MNRAS, 436, 2286

Mccullagh N., Norberg P., Cole S., Gonzalez-perez V., Baugh C., Helly J., 2017

Mo H., van den Bosch F. C., White S., 2010, Galaxy Formation and Evolution

Mo H. J., Yang X., van den Bosch F. C., Jing Y. P., 2004, MNRAS, 349, 205

Molino A. et al., 2014, MNRAS, 441, 2891

Morgan W. W., Mayall N. U., 1957, Publications of the Astronomical Society of the Pacific, 69, 291

Naiman J. P. et al., 2018, MNRAS, 477, 1206

Nelson D. et al., 2018, MNRAS, 475, 624

Norberg P., 2dFGRS Team, 2002, ASP Conf. Ser., 283, 47

Norberg P. et al., 2001, 44, 0

Paranjape A., Kovač K., Hartley W. G., Pahwa I., 2015, MNRAS, 454, 3030

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

Perlmutter S. et al., 1999, ApJ, 517, 565

Pillepich A. et al., 2018, MNRAS, 475, 648

Planck Collaboration, P. A. R. Ade, N. Aghanim, C. Armitage-Caplan, M. Arnaud, M. Ashdown, et al., 2014, A&A, 571

Planck Collaboration et al., 2018, arXiv e-prints

Postman M. et al., 2012, ApJS, 199, 25

Reddick R. M., Wechsler R. H., Tinker J. L., Behroozi P. S., Park M., 2013, ApJ, 771, 30

Riess A. G. et al., 1998, AJ, 116, 1009

Roberts M. S., Haynes M. P., 1994, ARA&A, 32, 115

Salvato M., Ilbert O., Hoyle B., 2019, Nature Astronomy, 3, 212

Sánchez C. et al., 2014, MNRAS, 445, 1482

Schechter P., 1976, ApJ, 203, 297

Shectman S. A., Landy S. D., Oemler A., Tucker D. L., Lin H., Kirshner R. P., Schechter P. L., 1996, ApJ, 470, 172

Sheldon E. S. et al., 2004, AJ, 127, 2544

Sinha M., Garrison L., 2017, Corrfunc: Blazing fast correlation functions on the CPU. Astrophysics Source Code Library

Sinha M., Garrison L., 2019, in A. Majumdar, R. Arora, eds, Software Challenges to Exascale Computing. Springer Singapore, Singapore, pp. 3–20

Skibba R. A., Sheth R. K., 2009, MNRAS, 392, 1080

Slipher V. M., 1917, Proceedings of the American Philosophical Society, 56, 403

Smith A., Cole S., Baugh C., Zheng Z., Angulo R., Norberg P., Zehavi I., 2017, MNRAS, 470, 4646

Somerville R. S., Primack J. R., 1999, MNRAS, 310, 1087

Springel V., 2005, MNRAS, 364, 1105

Springel V., White S. D. M., Tormen G., Kauffmann G., 2001, MNRAS, 328, 726

Springel V., Frenk C. S., White S. D. M., 2006, Nature, 440

Strateva I. et al., 2001, AJ, 122, 1861

Tasitsiomi A., Kravtsov A. V., Wechsler R. H., Primack J. R., 2004, ApJ, 614, 533

Tonry J., Davis M., 1979, AJ, 84, 1511

Trayford J. W. et al., 2015, MNRAS, 452, 2879

Turner M. S., Steigman G., Krauss L. M., 1984, Phys. Rev. Lett., 52, 2090

Vale A., Ostriker J. P., 2004, MNRAS, 353, 189

van Dokkum P. G. et al., 2015, ApJ, 813, 23

Wechsler R. H., Tinker J. L., 2018, Annual Reviews of Astronomy and Astrophysics, 1

Wechsler R. H., Gross M. A. K., Primack J. R., Blumenthal G. R., Dekel A., 1998, ApJ, 506, 19

Wechsler R. H., Zentner A. R., Bullock J. S., Kravtsov A. V., Allgood B., 2006, ApJ, 652, 71

Weinberg D. H., Mortonson M. J., Eisenstein D. J., Hirata C., Riess A. G., Rozo E., 2013, Phys. Rep., 530, 87

Weinmann S. M., van den Bosch F. C., Yang X., Mo H. J., 2006, MNRAS, 366, 2

Wetzel A. R., Tinker J. L., Conroy C., van den Bosch F. C., 2013, MNRAS, 432, 336

White M., Tinker J. L., McBride C. K., 2014, MNRAS, 437, 2594

White S. D. M., Frenk C. S., 1991, ApJ, 379, 52

Xu X., Zheng Z., 2018, arXiv e-prints, arXiv:1812.11210

Yamamoto M., Masaki S., Hikage C., 2015, 000

Yang X., Mo H. J., van den Bosch F. C., 2003, MNRAS, 339, 1057

Yang X., Mo H. J., van den Bosch F. C., 2008, ApJ, 676, 248

Yang X., Mo H. J., van den Bosch F. C., 2009, ApJ, 695, 900

Zehavi I., Kerby S. E., Contreras S., Jiménez E., Padilla N., Baugh C. M., 2019, arXiv e-prints, arXiv:1907.05424

Zehavi I. et al., 2005, ApJ, 630, 1

Zehavi I. et al., 2011, ApJ, 736, 59

Zentner A. R., Hearin A. P., van den Bosch F. C., 2014, MNRAS, 443, 3044

Zlatev I., Wang L., Steinhardt P. J., 1999, Phys. Rev. Lett., 82, 896

Appendix A

P-Millennium output comes as separate files containing Friends-of-Friends (FoF) halo and SUBFIND subhalo data, as well as DHALO files containing information about merger trees (as well as particle files for a selection of snapshots, although the particle data is not used in this dissertation). Below, we provide the definitions of relevant quantities that are included in P-Millennium output files[†].

Table A.1. Descriptions of selected datasets available in P-Millennium DHALO output files, which are relevant for the development of the Rosella mock catalogue.

Property name	Description
`nodeIndex`	unique identifier for a subhalo at a given snapshot
`snapshotNumber`	the snapshot at which the subhalo exists
`redshift`	the redshift at which the subhalo exists
`descendantIndex`	`nodeIndex` of the descendant of the subhalo
`descendantSnapshot`	snapshot at which the descendant exists
`isMainProgenitor`	indicates if this subhalo is its descendant's main progenitor (1) or not (0)
`isFoFCentre`	indicates if a subhalo is the most massive in the FoF group (1) or not (0)
`particleNumber`	number of particles in the subhalo
`position`	Cartesian subhalo coordinates, given by SUBFIND [comoving Mpc/h]
`velocity`	peculiar velocity of the subhalo [km/s]
`maximumCircularVelocity`	v_{max} [physical km/s]
`mainBranchMaximumVmax`	v_{peak} [physical km/s]

[†]The information presented here is sourced from the P-Millennium page in the Virgo Data Centre wiki portal: https://wiki-virgo.esc.rzg.mpg.de/projects/pmillennium/start

Table A.2. Descriptions of selected datasets available in P-Millennium FoF and SUBFIND output files, which are relevant for the development of the Rosella mock catalogue.

Property name	Description
`Halo_M_Mean200`	Mass within overdensity 200 times the mean for each FoF group $[10^{10} M_{\odot}/h]$
`Halo_M_Crit200`	Mass within overdensity 200 times critical for each FoF group $[10^{10} M_{\odot}/h]$
`Nsubs`	Number of Subfind groups in each FoF group
`SubLen`	Number of particles in each Subfind group
`SubGrNr`	Identification number for each Subfind group's parent FoF group
`SubNr`	Identification number for each Subfind group in this file
`SubPos`	Position of a particle with the lowest potential energy in a subhalo [comoving Mpc/h]
`SubVel`	Peculiar velocity of each subhalo [physical km/s]
`SubCofM`	Subhalo centre of mass [comoving Mpc/h]
`SubVmax`	v_{max} [physical km/s]
`SubRVmax`	Radius at which v_{max} is found [physical km/s]
`SubHalfMass`	Subhalo's half mass radius [physical km/s]

Appendix B

The mock catalogue presented here is named Rosella, after the species *Rosella* (*Platycercus*, shown in figure B.1). The motivation behind this choice is symbolic. This bird's red and green feathers signify the galaxy luminosity and colours that our mock provides; the spotted feathers symbolize dark matter subhalos, a central element of our methodology. *Rosella*'s blue plumage reminds us of our hope for future peaceful collaborations, both in the academic realm and outside it.

Figure B.1. *Rosella* the bird, our catalogue's eponym. Image courtesy of damselfly58 on Flickr.

Colophon

This thesis is based on a template developed by Matthew Townson and Andrew Reeves. It was typeset with LaTeX 2_ε. It was created using the *memoir* package, maintained by Lars Madsen, with the *madsen* chapter style. The font used is Latin Modern, derived from fonts designed by Donald E. Kunith.