UNDERSTANDING THE GALAXY-HALO CONNECTION THROUGH GALAXY GROUP CATALOGUES

By

Victor Calderon Arrivillaga

Dissertation

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

in

Physics

August 9th, 2019

Nashville, Tennessee

Approved:

Andreas Berlind, Ph.D.

Kelly Holley-Bockelmann, Ph.D.

Robert J. Scherrer, Ph.D.

Keivan G. Stassun, Ph.D.

Sait A. Umar, Ph.D.

Dedication

I dedicate this thesis to

My Family,
For their endless support and encouragement throughout the years,
for always making sure I stay grounded,
and for always being there for me! Muchas gracias!

and

My Friends
For always being there when I needed, for all your support and
assistance during this journey, and for making my time in
Nashville a memorable one!

Acknowledgment

First and foremost, I am thankful for all the support and supervision of my advisor, Dr. Andreas Berlind. This work would not have been possible without his constant encouragement, his mentorship and guidance throughout all these years. I am grateful to have been given the chance to present my research, meet new colleagues, and meet new destinations during my time at Vanderbilt. I am also grateful for all of the "1-minute" conversations with him during all the stages of graduate school. They have definitely made an impact in my professional and personal life.

I am also grateful to the members of my committee for taking time from their busy schedules to help me along this journey. I am also grateful to my collaborators for all of their helpful suggestions, discussion, and advice throughout these years.

I would like to thank my friends at the 9th floor for their support and friendship. My time at Vanderbilt would definitely not have been the same without them. I am also grateful for the Vanderbilt Astronomy & Physics Department for always helping me along the way, and for always making sure my time as a graduate career was always pleasant.

I would also like to thank Maria Fernanda Senoain for being there for me, especially during those times that I was writing the papers. For all the study breaks, for her patience and constant encouragement during these year, and for her companionship that has kept me going, especially when finishing this dissertation.

Finally, I would also like to thank my family for their endless support and constant encouragement throughout these years. I am extremely grateful for them always being there for me when I needed them, and for keeping me grounded all these years. They have truly make a difference in my life, and for that I'm am extremely grateful.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction

The Universe is comprised of structures on all scales, ranging from stars and planets to galaxies and clusters. These structures form part of a much bigger web-like structure built with systems of galaxies, filaments, walls, and cosmic voids between galaxies. The Universe is comprised of Large-Scale structure, in which galaxies form, evolve, and eventually die. Theoretical and observational research over the last half of a century have led to a consistent model of the Universe that includes a mysterious dark energy that is driving the accelerated expansion of the Universe, and dark matter that cannot be directly observed with telescopes. The theoretical frameworks of large-scale structure and galaxy formation and evolution are being tested by the new generation of galaxy surveys, and they will keep being challenged by the upcoming generation of telescopes and instruments, that will shed light into the evolutionary paths of galaxies and the Universe.

Galaxies constitute the primary objects in the observed Universe, and act as the building blocks of Large-Scale Structures in the Universe. By studying the evolutionary paths of galaxies, and their connection to their environments, we can help constrain cosmological parameters and provide a much clearer picture of the physical and statistical connection between the luminous matter in the Universe (galaxies) and the dark matter in the Universe. This relation is commonly referred to as the galaxy-halo connection. Understanding this relation is crucial for constraining cosmological parameters and probing the distribution and properties of dark matter in the Universe.

In this dissertation, I present some different statistical analyses that aim at analyzing various aspects of the galaxy-halo connection. As I will elaborate on in this document, these studies can provide us with a better understanding of the connection between galaxies and their environments, as well as a glimpse into the formation and evolution of large-scale structure of the Universe.

1.1 An Expanding Universe in the ΛCMD Cosmological Model

The accepted cosmological theory of the origin of the Universe revolves around the concept of the 'Big Bang', in which the Universe began in a hot, dense and nearly isotropic state some 13.7 billion years ago, and expanded exponentially moments after through inflation. Under this paradigm, the origin of cosmic structure is thought be caused by quantum mechanical fluctuations in the early Universe, which froze in the Cosmic Microwave Background (CMB) by inflation, and led to perturbations in the density field of the Universe. These anisotropies can be observed in the CMB as temperature variations, which are on the order of $\delta T/T \propto 10^{-5}$. The Universe has been expanding ever since, but at a lower rate. The structure that we see through a telescope is the direct result of the Big Bang, the inflationary period, and rapid expansion of the Universe.

Since late 1920's, with the advent of Friedmann's equations about the geometry of the Universe, researchers across the globe have tried to test the idea that the Universe is accelerating. In 1929, Edwin Hubble discovered a relation that relates the velocity of a galaxy to its inferred distance to use. This relation is known as 'Hubble's Law' and is indicative for an expansion of the Universe. However, it was until 1998 that Saul Perlmutter and Adam Riess were awarded the Nobel Prize in Physics for the discovery that the expansion of the Universe is accelerating by observations of Type Ia supernovaes (Riess et al., 1998; Perlmutter et al., 1999). Since then, our best cosmological models include a component responsible for the cosmic expansion.

Observational facts in the last few years have led towards the development of a cosmological model referred to the Λ cold dark matter (ΛCMD) model. This model states that the Universe is comprised of baryonic matter, and the dark sector. The dark sector is composed of two different components that dictate the ultimate faith of the Universe, namely: 1) 'cold dark matter', a type of non-relativistic particle that most likely interacts with baryonic only through gravity and does not produce or reflect any electromagnetic radiation, and 2) 'dark energy', which acts as negative pressure causing the accelerating expansion

of the Universe. This model is currently our best cosmological model for the formation and evolution of cosmic structure in the Universe (Wechsler & Tinker, 2018). This model indicates that the total energy density of the Universe today is comprised of about 70% dark energy, 26% dark matter, and only 4% baryonic matter with great accuracy (Planck Collaboration, 2016).

In general, based on various observations, the Universe started with the Big Bang. Within moments after the Big Bang, the Universe rapidly expanded exponentially through inflation. During the inflationary period, quantum mechanical fluctuations acted as precursors for the growth of structures in the Universe, and these were frozen in the CMB. These perturbations of the density field led to the growth of the structure in the Universe, including galaxy clusters, galaxies, filaments, walls, and more. After inflation, the Universe kept expanding, but at a slower rate, which allowed for nucleosynthesis to take place. This eventually led to the formation of stars, galaxies, and large-scale structure we see today.

1.2 Large-Scale Structure of the Universe

Large-Scale Structure (LSS) in the Universe is the result of the evolution of the density perturbations in the initial density field of the Universe. These perturbations have been amplified at a grand scale since then through gravitational forces, producing vast amount of dense clumps of dark matter, which would eventually become the home for galaxies. Today, the current ΛCMD paradigm is our best cosmological model for the formation and evolution of the cosmic structure in the Universe (Wechsler & Tinker, 2018). It predicts that all galaxies form and evolve within gravitationally bound structures of dark matter, commonly referred to as *dark matter haloes*. A halo refers to a gravitationally bound structure with overdensity of ~ 200 times the mean density of the Universe. These overdensed regions form part of a much larger, web-like extragalactic structure, also known as the *cosmic web*.

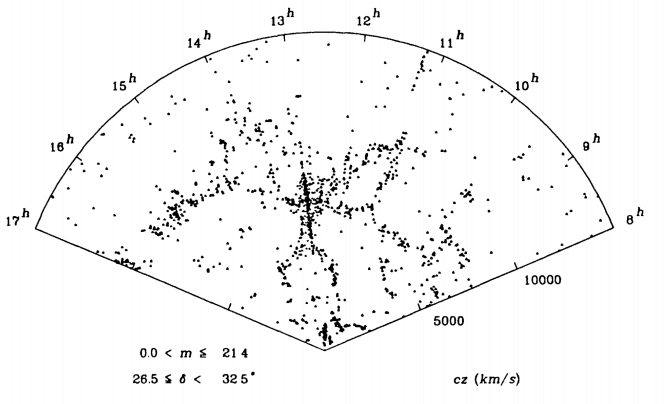

Figure 1.1: The initial galaxy map of the Center for Astrophysics (CfA) Redshift Survey (Huchra, 1988), showing the distributions of galaxies on the sky. The Earth is located at the point of the slice. Black dots correspond to the location of galaxies on the sky.

1.3 Observing the Universe

With the advent of spectroscopic redshift surveys in the early 1980's, we can now map the locations of galaxies on the sky with great accuracy. Galaxy redshift surveys, such as Center for Astrophysics (CfA) galaxy redshift survey (Huchra, 1988), the Two Degree Field Galaxy Redshift Survey (Colless et al., 2001), the Two Micron All Sky Redshift Survey (Skrutskie et al., 2006), and in particular, the Sloan Digital Sky Survey (SDSS; York, 2000), have observed and quantified the cosmic web, and have provided us with reliable spectroscopic information about the location of galaxies, clusters, and more. Figure 1.1 shows the initial galaxy map of the CfA Redshift Survey, and it shows the distribution of galaxies on the sky. This figure illustrates different types of environments, in which galaxies reside, e.g. filaments, sheets, cosmic voids between galaxies, and more. This work and more recent ones have tried to understand the connection between galaxies and their corresponding environments.

1.3.1 Sloan Digital Sky Survey

Many of the analyses presented in this dissertation are based on data collected by the Sloan Digital Sky Survey (York, 2000, hereafter SDSS). SDSS is one of the most ambitious, impressive, and influential surveys of the last two decades. It has completely revolutionized our understating of galaxy formation and evolution, growth of cosmic structure, and the demographics of galaxies. Additionally, it has also contributed to our understanding of the galaxy-halo connection by providing exact and reliable spectroscopic measurements of millions of galaxies. SDSS started in 2000, followed by SDSS-II in 2005, SDSS-III in 2008 (Eisenstein et al., 2011), and SDSS-IV in 2014 (Blanton et al., 2017). SDSS collected its data with a dedicated 2.5-meter telescope (Gunn et al., 2006), camera (Gunn et al., 1998), filters (Doi et al., 2010), and spectrograph (Smee et al., 2013) at the Apache Point Observatory. Each object passes through one column of 5 different CCDs that correspond to 5 different filters, arranged in 6 columns, with a total of 30 different CCDs. The SDSS filters cover from the ultraviolet to the near-infrared part of the light spectrum, and are denoted as '*u g r i z*' filters. Overall, the original SDSS covered a total of 8000 sq. degrees on the sky. The original as well as its two consequent extensions (SDSS-II and SDSS-III) have been instrumental in our understanding of galaxy formation and evolution, as they have measured more than 4 million spectra in total (Alam et al., 2015).

1.3.2 Redshift-Space Distortions

One of the most important information about a galaxy is its distance to us, since that can provide us put better constraints on cosmological models. In redshift galaxy surveys, distance to galaxies are inferred from their spectra under the assumptions that they are only being affected by cosmic expansion ('Hubble flow'). However, redshift of galaxies are not only a product of cosmic expansion, but it is also affected by the individual motions of galaxies, i.e. the peculiar motion. The relative motion of galaxies adds and extra redshift to that caused by the Hubble flow, and will potentially affect our inferred distances to galaxies.

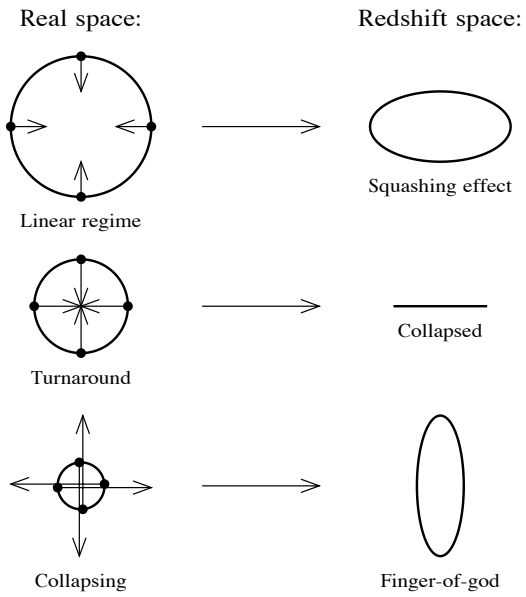

Figure 1.2: Schematic of the physical mechanisms and observed effects that induce redshift-space distortions. The direction of the arrows correspond to the directions of the peculiar velocity vectors (Hamilton, 1998).

The measured redshift of galaxies is the result of a combination of the Hubble flow and the peculiar velocity of galaxies. This results in galaxies being misplaced in redshift space, to which we refer as *redshift-space distortions* (RSD). Equation 1.1 corresponds to the redshift of galaxy, z, that an observer on Earth would measure, so that

$$cz = H_0 d + v_p \tag{1.1}$$

where z corresponds to the measured redshift of the galaxy, H_0 to the Hubble constant, d to the distance to the galaxy, and v_p to the peculiar velocity of a given galaxy.

Figure 1.2 shows the observed effects from different physical process that may induced RsD. At relatively small scales, the velocity dispersion of galaxies within their own halo can make closer galaxies look apart, and vice versa. This effect forms elongated shapes in redshift-space, and it is commonly referred to as "Fingers of God" (Hamilton, 1998). At

larger scales, the bulk motion of groups of galaxies towards a more massive cluster, can be observed as a compression of the galaxy distribution in redshift-space. This type of distortions is commonly referred to as the "Kaiser effect" (Kaiser, 1987).

RSD are one of the biggest systematic errors when determining distances to galaxies in redshift-space. They inadvertently cause errors when determining the cluster membership of galaxies, and can affect our statistical inferences about the galaxy-halo connection. For this reason, it is important to be able to model them and understand the effect that they have on our measurements.

1.3.3 Galaxy group catalogs

The ΛCMD cosmological model predicts that all galaxies form and evolve in dark matter halos. The physical and statistical connection between galaxies and their host haloes can provided us with an insight into how galaxies form and evolve, help us infer and constrain cosmological parameters, as well as help us probe the distribution of dark matter in the Universe. For these reasons, it is important to be able to identify galaxies from the same halo as single galaxy groups.

It is common practice to apply a group-finding algorithm to large galaxy redshift surveys, in order to assign galaxies to groups and construct a 'galaxy group catalog'. The resulting group catalog can be used to study various aspects of the galaxy-halo connection, including the relationship between galaxy properties and those of their host haloes, mechanisms that drive galaxy quenching within groups, or the impact that group environment may have on the morphology of galaxies. These analyses try to answer some of the most fundamental questions of galaxy formation and evolution, and they rely on a proper group membership assignment for of galaxies. In this dissertation, I show how RSD can induce errors in the statistical inferences of various aspects of the galaxy-halo connection, and perform a robust and comprehensive study of these effects.

1.4 N-body Simulations and Mock Catalogs

1.4.1 Overview of N-body simulations

The physics behind non-linear structure formation is very complex and cannot be fully explained through analytical solutions. Modern astrophysics relies on numerical simulations of the Universe that are able to reproduce the history of structure growth, incorporate various physical process that may have an impact on galaxy formation and evolution. N-body simulations also allow to trace the distribution of dark matter in the Universe and trace back the evolution of the density field at various redshifts. They have become indispensable for explaining and recreating various types of observations. However, they are highly time consuming and computationally expensive to run. The adoption of this technique has changed dramatically with the advent of supercomputers and better algorithms, as now we are able to run very complex cosmological simulations with billions of particles in an appropriate time.

There are different flavours of cosmological simulations. For example, (Vogelsberger et al., 2014) introduced a set of hydrodynamical simulations that were representative of the observable Universe, and were able to simulate physical processes that are relevant to galaxy formation and evolution, as well as simulate the distribution of dark matter in the Universe. Similarly, semi-analytic models consists of combining the results of N-body simulations with simple physical prescriptions to estimate the distribution of galaxies. N-body simulations constitute the third flavour of cosmological simulations, in which dark matter particles are laid down smoothly in a simulation box. Each particle, which is in actuality corresponds to a certain mass, is then perturbed and initial velocities are given based on perturbation theory from a given cosmological model. This approach provides us with the set of initial conditions at very high redshifts, and as 'time' progresses in the simulation, gravitational forces act on every particle of the simulation. At each time step, gravitational forces between particles are calculated and each particle is moved based on its velocity components and total force applied to it. This step is repeated over and over

until the simulation has reached the desired redshift.

For the case of N-body simulations, once the simulation is complete, one can identify the particles belonging to dark matter haloes of a given overdensity with the help of halo-finding algorithm. Once haloes are identified, dark matter haloes are populated with mock galaxies by specifying an Halo Occupation Distribution (Berlind et al., 2003), which provides a framework for describing the number, spatial and velocity distributions of galaxies within a dark matter halo of a certain mass. This procedure would lead to the construction of mock galaxy catalogs which provides with a way to test and compare theoretical models and ideas to observed galaxy distributions.

Throughout this dissertation, I make use of mock galaxy catalogs to make statistical inferences about the galaxy-halo connection. Without them, the majority of analyses and results would not have been possible to perform.

1.4.2 LasDamas Simulations

Throughout this dissertation, I present analyses about various aspects of the galaxy-halo connection. The main cosmological simulation that I use for these projects is the *Large Suite of Dark Matter Simulations* (LasDamas)[1]. LasDamas is a suite of many cosmological N-body simulations with the same cosmology but different initial conditions, that trace the evolution of dark matter in the Universe. The dark matter haloes are found by applying a Friend-of-Friends algorithm (Davis et al., 1985) with linking length of 0.2 times the mean inter-particle separation. LasDamas provides multiple realizations of the Universe with a common cosmological model, which is ideal for statistical studies, such as the ones presented in this dissertation.

1.5 Summary

Throughout this work, I make use of the myriad of data from the Sloan Digital Sky Survey to explore different aspects of the galaxy-halo connection. Moreover, I make use

[1] http://lss.phy.vanderbilt.edu/lasdamas/

of a suite of mock galaxy catalogs from N-body simulations to statistically determine the importance of each result. Moreover, I use galaxy and group galaxy catalogs to make inferences about how galaxies correlate with neighboring galaxies, and about the stellar content of galaxy groups. In chapter 2, I analyze the stellar content of galaxy groups with group catalogs from SDSS, and present a set of value-added galaxy (group) catalogues for three different galaxy samples. In chapter 3, I investigate a very important aspect of the galaxy-halo connection, namely "galactic conformity". I also make use of realistic mock galaxy catalogs to statistically claim the first robust detection of galactic conformity at large scales. Chapter 4 introduces a method of estimated the masses of galaxies' host dark matter by employing information about the galaxies and their corresponding group environment. A short conclusion is in chapter 5

Chapter 2

PROBING THE STELLAR CONTENT OF GALAXY GROUPS WITH VALUE-ADDED GROUP CATALOGUES IN THE SDSS DR7

The following work will be submitted to the Monthly Notices of the Royal Astronomical Society Journal and is reprinted below in its entirety

Probing the Stellar Content of Galaxy Groups with Value-Added Group Catalogues in the SDSS DR7

Victor F. Calderon[1], Andreas A. Berlind[1], Manodeep Sinha[2]

[1] Department of Physics and Astronomy, Vanderbilt University, Nashville, TN 37235

[2] Centre for Astrophysics and Supercomputing, Swinburne University of Technology, Hawthorn, Victoria 3122, Australia

2.1 Abstract

We investigate the ability to confidently make use of galaxy group catalogs to explore various aspects of the galaxy-halo connection, including the stellar-to-halo mass relation (SHMR) of galaxies. Moreover, we explore the role that group mass has determining galaxy quenching as function of stellar mass for central and satellite galaxies. We determine that group-finding errors do not affect the sSFR $- M_\star$ of central and satellite galaxies, and that central and satellite galaxies follow the same relation of sSFR $- M_\star$ at fixed group mass. Additionally, we compute a correction factor to recover the *true* SHMR of central galaxies as a function of group mass. To test the feasibility of group catalog derived statistics, we perform a robust analysis of the impact by group-finding systematic errors on group mass assignment and galaxy type identification in the SDSS DR7. We conclude that central galaxies are correctly identified as central galaxies $80-90\%$ of the time by group-finding, regardless of group richness. However, satellite galaxies are correctly identified as satellite

galaxies $60-70\%$ of the time. Finally, we present and make available sets of galaxy group catalogs for three volume-limited samples of SDSS DR7.

2.2 Introduction

Galaxies are gregarious by nature, and they can be found in different types of environments. Bright galaxies typically reside in large groups of galaxies or clusters, surrounded by less luminous neighbours. Interactions within the group environment may have an effect on the observational properties of galaxies, such as morphology, dynamics, star formation histories, among others. Characterizing the relation between galaxy properties and their group environment can shed light into how galaxies form and evolve. Today, the current Λ cold dark matter (ΛCDM) paradigm is our best cosmological model for the formation and evolution of the cosmic structure in the Universe (Wechsler & Tinker, 2018). It predicts that all galaxies form and evolve within gravitationally bound structures of dark matter (DM), commonly referred to as *dark matter haloes*[1]. The physical and statistical connection between the luminous matter in the Universe (galaxies) and the DM in haloes, is commonly referred to as the '*galaxy-halo*' connection, and it is crucial for constraining cosmological parameters and probe the distribution and properties of DM in the Universe.

Given the hierarchical nature of structure formation, and the tendency of luminous galaxies to reside in groups and clusters surrounded by less luminous neighbors, we expect galaxy groups to constitute a fundamental physical scale important for galaxy formation and evolution Campbell et al. (2015, ; hereafter C15). This idea motivates the study of galaxies in groups to better understand the galaxy-halo connection. Galaxy groups and clusters can be identified through various methods. Traditionally, galaxy clusters were first detected as overdensities of galaxies in the visible spectrum (e.g. Abell, 1958; Zwicky et al., 1968). Since then, galaxy systems can be identified as overdensities of red galaxies in both the visible and IR spectrum (e.g. Gladders & Yee, 2005; Hao et al., 2010; Ascaso

[1] Throughout this paper, we use the term "halo" to refer to gravitationally bound structure with overdensity $\rho/\bar{\rho} \sim 200$, so an occupied halo may host a single luminous galaxy, a group of galaxies, or a cluster.

et al., 2012). These can also be detected as extended X-ray sources (e.g. Rosati et al., 2002; Vikhlinin et al., 2009), and by their signature in the cosmic microwave background (e.g. Marriage et al., 2011; Staniszewski et al., 2009; Ade et al., 2015). With the onset of spectroscopic redshift surveys in the early 1980's, systems of galaxies can be selected based on the closeness of galaxies in redshift space, while minimizing the challenges associated with projection effects in determining galaxy group membership. This leads to the construction of group galaxy catalogs. Many of these analyses have adopted the widely-used Friends-of-Friends percolation algorithm to put galaxies into groups and compile group galaxy catalogs. The FoF algorithm puts into a single group all galaxies linked in pairs based on the separation on the sky and along the line-of-sight direction. Most notably, numerous group galaxy catalogs have been constructed for different spectroscopic redshift surveys, including the Center for Astrophysics Redshift Survey (CFA; Geller & Huchra, 1983), the Las Campanas Survey (Tucker et al., 1997), the Two Degree Field Galaxy Redshift Survey (2dFGRS; Merchán & Zandivarez, 2002; Eke et al., 2004; Yang et al., 2005; Einasto et al., 2007) the high-redshift DEEP2 survey (Gerke et al., 2005), the Two Micron All Sky Redshift Survey (Crook et al., 2007; Lavaux & Hudson, 2011), and in particular, the Sloan Digital Sky Survey (e.g. Goto, 2005; Berlind et al., 2006; Yang et al., 2007).

The Sloan Digital Sky Survey (York, 2000, ; hereafter SDSS) has been crucial for the study of galaxy properties and their environments by providing one of the largest samples of galaxies with spectroscopic information, along with detailed information on galaxy properties. SDSS has been widely used to analyse various aspects of galaxy demographics, and other aspects of the galaxy-halo connection. For example, Zehavi et al. (2011) analyzed the luminosity and color dependence of galaxy clustering in SDSS, and found that at fixed luminosity, the redshift-space correlation function of red galaxies exhibited stronger "finger-of-God" distortions than that of blue galaxies, while blue galaxies show stronger large-scale, coherent flow distortions. They also found a shallow, low-amplitude correlation function for the bluest galaxies in the sample, while the clustering of "green

valley" galaxies is intermediate between that of blue and red galaxies, with the reddest galaxies having a steeper correlation function. Similarly, Martinez et al. (2006) analysed the $u-r$ colour distributions for several galaxy samples in groups from SDSS Data Release 4 (McCarthy et al., 2006), and found that the fraction of galaxies in the red sequence is an increasing function of group mass. Additionally, they found that the fraction of red galaxies monotonically increases with decreasing redshift, implying a much stronger evolution of galaxies in groups than in the field.

It is clear that galaxy properties correlate strongly with environment in the local Universe. A colour-magnitude diagram of galaxies shows a bimodality in colour in the local Universe that persists out to larger redshifts (Bell et al., 2004). The bimodal distribution of galaxy colours is the result of the diverse star formation efficiencies of galaxies, dividing galaxies into a star-forming blue cloud and a more quenched red sequence. The origin of this relation is not well understood (C15), yet galaxies in dense environments exhibit an enhanced quenched fraction relative to that of galaxies residing in more isolated environments (Dressler, 1980; Postman & Geller, 1984; Kauffmann et al., 2004). However, it is not clear what drives this relation, and whether or not there exists a causal relationship between galaxy properties and environment.

Within the framework of galaxy groups and DM haloes, it is customary to describe galaxies as either 'central' galaxies or 'satellite' galaxies. Central galaxies are commonly referred to those galaxies located at the deepest point of the gravitational potential of a DM halo, and they are usually associated to the most massive or most luminous galaxies in the halos. Satellite galaxies are those galaxies that are not central galaxies, and are associated to DM subhaloes. Hence, a halo is comprised of a single central galaxy and zero or more satellite galaxies. Central galaxies and satellite galaxies undergo different physical processes that ultimately affect their galaxy properties. This criterion is motived by the idea that central galaxies grow in mass, and brightness by galactic cannibalism (Dubinski, 1998; Cooray & Milosavljević, 2005), while satellite galaxies experience a series of events

that strip them from their mass and inhibit star formation (Balogh et al., 2000; Grebel et al., 2003). Distinguishing between central and satellite galaxies allows for the study of the significance of various physical processes to the galaxy-halo connection.

Galaxy group catalogs serve as bridges between theory and observations of galaxies, as they aim to represent the *true* group membership of galaxies, and can be used to explore the multivariate distribution of properties of haloes and galaxies that form within them. Galaxy group catalogs have been used to measure galaxy property correlations beyond halo mass and galaxy type designations. For example, Zhang & Yang (2019) studied the dependency of intrinsic properties on the size of galaxies. In their analysis, they used a galaxy group catalog constructed by Yang et al. (2007) to distinguish between central and satellite galaxies, and determine if late-type galaxies exhibited a different trend in galaxy compared to that of early-type galaxies. Similarly, galaxy group catalogs have been used to better understand different aspects of the galaxy-halo connection, such as correlations between quenching properties of galaxies and those of neighboring galaxies (Weinmann et al., 2006; Tinker et al., 2018; Lim et al., 2017; Treyer et al., 2017; Calderon et al., 2018), prediction of galaxy halo masses (Calderon & Berlind, 2019), exploration of thermal energy contents in the intergalactic medium (Lim et al., 2018), among others.

The main goal of a group-finding algorithm is to correctly determine the group membership of galaxies in a galaxy sample. In an ideal scenario, a perfect group-finding algorithm would be able to classify galaxies which occupy a common halo as members of the same group. Unfortunately, due to peculiar motions of galaxies, it is not possible to perfectly determine the group membership of galaxies in redshift-space. This results in a two different scenarios, whereby galaxies from distinct haloes are assigned to the same group, or member galaxies of haloes are split into multiple groups. It is important to properly characterize the group-finding errors and determine the impact that these have on inferences about galaxy properties as function of halo properties. C15 investigated the use of group catalogs to recover colour-dependent halo occupation statistics, and analysed the impact

of group-finding errors from three different group-finders on the recovery of galaxy and group properties as a function of halo properties. Similarly, Lim et al. (2017) applied a halo-based group finder to four large redshift surveys and quantified the performance of the group-finder at halo mass assignment and group membership identification. In this paper, we release various sets of galaxy group catalogs for different volume-limited samples of SDSS, and quantify the ability of the Berlind et al. (2006) group-finding algorithm at recovering galaxy and group properties as a function of halo properties. Additionally, we make use of the group galaxy catalogs to explore the stellar-to-halo mass relation and the role of group mass in galaxy quenching.

This paper is organised as follows. In §2.3, we describe the observational (§2.3.1 and §2.3.2) and simulated data (§2.3.3) used in this work. In §2.4, we introduce the group-finding algorithm, as well as set of galaxy group catalogs. We then discuss and quantify the errors group finders make in determining group mass and galaxy type §2.5. In §2.6, we make use of the galaxy group catalogs to explore the stellar-to-halo mass relation of central galaxies (§2.6.1), and the role of mass in galaxy quenching (§2.6.2). We conclude with a discussion of our results and a summary in §2.7.

2.3 Data and Methods

In this section, we introduce the datasets used throughout this analysis, and discuss how the various galaxy catalogs have been constructed. in §2.3.1, we present the characteristics of the three volume-limited galaxy samples used in this paper. In §2.3.2, we present the formalism used when assigning stellar masses and star formation rates to galaxies in the three samples. Additionally, §2.3.3 summarizes the set of synthetic galaxy catalogues corresponding to the three volume-limited samples from §2.3.3, including details about the simulation used to create the synthetic catalogs (§2.3.3.1), the methods of assigning luminosities and stellar masses to synthetic galaxies (§2.3.3.2), and finally, the geometrical cuts employed in order to obtain realistic mock galaxy catalogues of the Universe (§2.3.3.3).

Table 2.1: Volume-limited Samples

Name	$M_r^{\rm lim}$	$z_{\rm min}$	$z_{\rm max}$	$N_{\rm gal}$	$\bar{n}_g$ ($h^3{\rm Mpc}^{-3}$)
Mr19-SDSS	-19	0.02	0.67	90,893	0.01503
Mr20-SDSS	-20	0.02	0.106	144,943	0.00593
Mr21-SDSS	-21	0.02	0.165	96,400	0.00104

Note. — The table shows the absolute r-bandmagnitude and redshift limits, the total number of galaxies and the number density of galaxies in each of the galaxy samples.

2.3.1 Sloan Digital Sky Survey

For this analysis, we use data from the Sloan Digital Sky survey (hereafter SDSS; York, 2000). SDSS collected its data with a dedicated 2.5-meter telescope (Gunn et al., 2006), camera (Gunn et al., 1998), filters (Doi et al., 2010), and spectrograph (Smee et al., 2013). We construct our galaxy sample from the `large-scale structure` sample of the NYU Value-Added Galaxy Catalogue (`NYU-VAGC`; Blanton et al., 2005), based on the spectroscopic sample in Data Release 7 (SDSS DR7; Abazajian et al., 2009). The main spectroscopic galaxy sample is approximately complete down to an apparent r-band Petrosian magnitude limit of $m_r = 17.77$. However, we have cut our sample back to $m_r = 17.6$ so it is complete down to that magnitude limit across the sky. Galaxy absolute magnitudes are k-corrected (Blanton et al., 2003) to rest-frame magnitudes at redshift $z = 0.1$.

We construct three volume-limited samples that contain all galaxies brighter than r-band absolute magnitudes M_r = -19, -20, and -21, and from this point forward, we will refer to these galaxy samples as Mr19-SDSS, Mr20-SDSS, and Mr21-SDSS, respectively. Table 2.1 summarizes the r-band absolute magnitude and redshift limits, the total number of galaxies, and the galaxy number density of each of the three volume-limited samples. These samples also include the right ascension, declination, redshift, Sérsic, and $(g-r)$ colour for each galaxy.

2.3.2 Stellar Masses and Star Formation Rates

To each galaxy in the three volume-limited galaxy samples, we assign a stellar mass and star formation rate (SFR) using the MPA Value-Added Catalogue DR7 (hereafter, MPA-JHU) [2]. This catalog includes, among many other parameters, stellar masses based on fits to photometry using Kauffmann et al. (2003) and Salim et al. (2007), and star formation rates based on Brinchmann et al. (2004). We cross-match the galaxies of the `NYU-VAGC` to those in the MPA-JHU catalog using their MJD, plate ID, and fiber ID. A total of 5.84% (7.14%), 6.72% (8.63%), and 8.16% (9.85%) of galaxies did not have corresponding values of SFR ($M_\star$) in the `Mr19-SDSS`, `Mr20-SDSS`, and `Mr21-SDSS` samples, respectively.

We follow the formalism presented in Bell et al. (2003, hereafter B03) to assign stellar masses to those galaxies, for which we were unable to find a corresponding stellar mass in the MPA-JHU catalog. Each galaxy has an accompanying flag that indicates if its stellar mass was extracted from the MPA-JHU catalog or was calculated using the B03 formalism. We discuss this further in §2.4.2.2.

Additionally, we explore the stellar mass at which the galaxy catalogs are complete, and only show the result for the `Mr19-SDSS` sample for brevity. Figure 2.1 presents the stellar mass completeness of `Mr19-SDSS` as a function of galaxy stellar mass. The orange and green dots and contours in the top panel correspond to the r-band absolute magnitudes and stellar masses of galaxies in the `Mr18-SDSS` and `Mr19-SDSS` galaxy samples, respectively. The dashed gray vertical line corresponds to the stellar mass, at which 95% of galaxies from `Mr18-SDSS` are brighter than the `Mr19-SDSS` r-band absolute magnitude limit of $M_r < -19$. The bottom panel shows the sample completeness level in galaxy stellar mass bins of 0.4 dex. In both panels, the dashed gray vertical line corresponds to the stellar mass at which 95% of galaxies in the `Mr18-SDSS` sample are brighter than the r-band absolute magnitude limit of `Mr19-SDSS`. In Figure 2.1, we observe that the `Mr19-SDSS` galaxy sample is complete at a stellar mass of of $\log M_\star \geq 10.6\ h^{-1} M_\odot$. which corresponds to $\sim 29.7\%$ of the

[2] http://www.mpa-garching.mpg.de/SDSS/DR7

total number of galaxies in the sample. Similarly, we compute the same statistics for the Mr20-SDSS and Mr21-SDSS galaxy samples, and conclude that these samples are stellar mass complete at log $M_\star$ of 11.0 and 11.4, respectively.

2.3.3 Mock Catalogues

To assess the performance of the group-finding algorithm, it is important to understand the systematic and statistical errors involved during the group assignment process. We use a set of realistic mock galaxy catalogs that have the exact same geometry as the SDSS volume-limited samples in §2.3.1. For the purpose of this paper, we will use the set of mock catalogs to estimate the accuracy and effectiveness of the group-finder, and evaluate the errors involved during the group-finding process, as described in 2.5.

In the following subsections, we present the suite of simulations used to produce the set of realistic galaxy catalogs, along with the methodology used to populate dark matter haloes with galaxies (§2.3.3.1); the framework used to assign luminosities, stellar masses and specific star formation rates to mock galaxies (2.3.3.2). Finally, 2.3.3.3 discusses the geometrical and redshift cuts we make to produce realistic mock galaxy catalogues that resemble SDSS.

2.3.3.1 Numerical Simulations

We construct a set of mock galaxy catalogs from the *Large Suite of Dark Matter Simulations*[3] (McBride et al., 2009), a suite of 50 cosmological N-body simulations per galaxy sample, that trace the evolution of dark matter in the Universe and have sufficient volume and mass resolution to properly model each of the galaxy samples in §2.3.1. The dark matter (DM) haloes are found by applying a Friends-of-Friends algorithm (FoF; Davis et al., 1985) using a linking length of 0.2 times the mean inter-particle separation. The total mass of the DM halo is the sum of all of the contributing DM particles. The suite assumes the same cosmology as the one in the Warren et al. (2006) halo mass function. Throughout

[3] http://lss.phy.vanderbilt.edu/lasdamas/

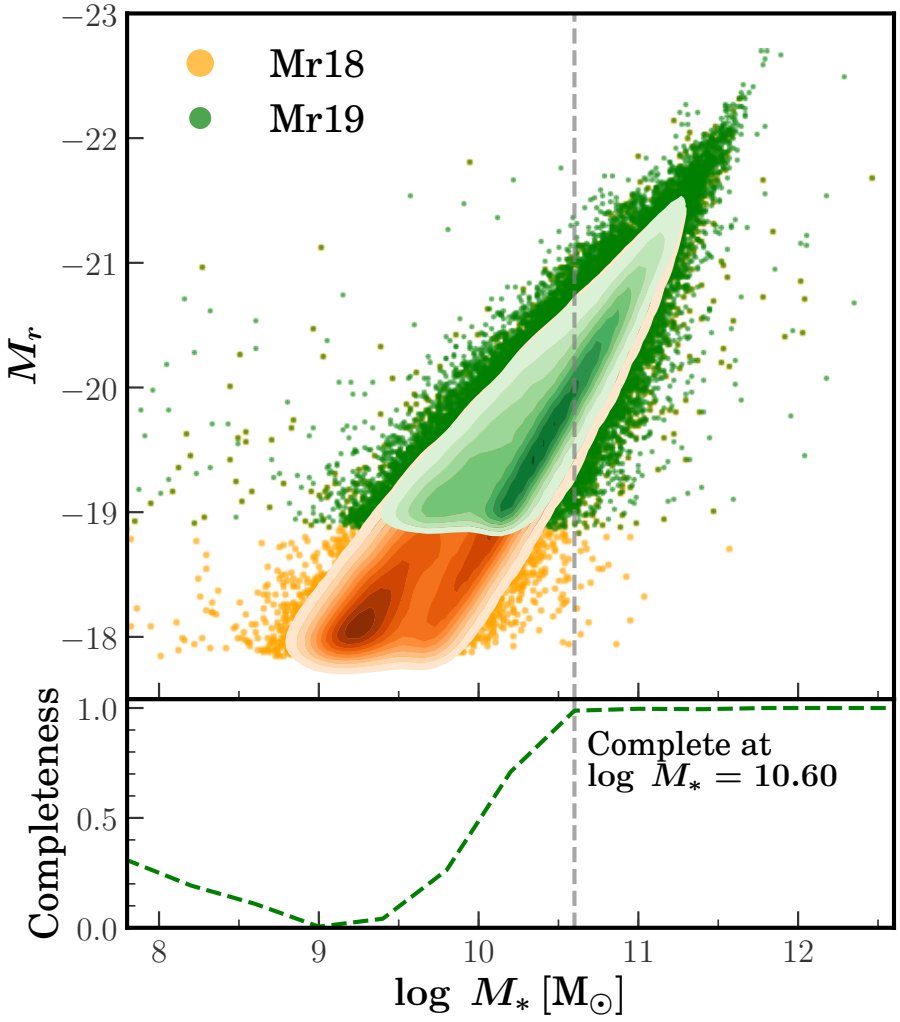

Figure 2.1: Sample completeness as a function of galaxy stellar mass for the Mr19-SDSS sample. *Top panel*: Orange and green dots and contours correspond to *r*-band absolute magnitudes and stellar masses of galaxies in the Mr18-SDSS and Mr19-SDSS galaxy samples, respectively. The dashed gray vertical line corresponds to the stellar mass, at which 95% of galaxies in Mr18-SDSS have brighter *r*-band absolute magnitudes than the Mr19-SDSS luminosity limit of $M_r < -19$. *Bottom panel*: Completeness level as a function of galaxy stellar mass. The dashed green line corresponds to the fraction of galaxies with brighter *r*-bandabsolute magnitudes than the luminosity limit of Mr19-SDSS, in stellar mass bins of 0.4 dex. The dashed gray vertical line corresponds to the stellar mass at which 95% of galaxies in Mr18-SDSS have brighter *r*-band absolute magnitudes the the Mr19-SDSS luminosity limit of $M_r < -19$.

this paper, we assume a cosmology of $\Omega_m = 1 - \Omega_\Lambda = 0.25$, $\Omega_{m,b} = 0.04$, $h = H_0/$ (100 km s^{-1} Mpc^{-1}) = 1, $\sigma_8 = 0.8$, and $n_s = 1.0$.

We used an Halo Occupation Distribution (HOD; Berlind & Weinberg, 2002) model to populate the DM haloes with central and satellite galaxies, whose numbers as a function of halo mass were chosen to reproduce the number density, $n_{\rm gal}$, and the projected 2-point correlation function, $w_p\left(r_p\right)$ of the `Mr19-SDSS`, `Mr20-SDSS`, and `Mr21-SDSS` samples. Each central galaxy was placed at the minimum of the halo gravitational potential and was assigned the mean velocity of the halo. Satellite galaxies were assigned the positions and velocities of randomly chosen DM particles within the halo in the galaxy sample. The next step is to assign realistic galaxy properties to each of the mock galaxies in the samples, including stellar masses, specific star formation rates, and luminosities.

2.3.3.2 Luminosities, stellar masses, and star formation rates

For this project, the mock galaxy catalogs come in two different *flavors*. The first version of the mock catalogs includes *luminosities* for each mock galaxy, while the second version includes *stellar masses* instead. We will refer to these two different versions of mock catalogs as `mr-set` and `mstar-set` , respectively.

1. For the `mr-set` mock galaxy catalogues, we adopt the formalism of the *conditional luminosity function* (CLF; Yang et al., 2003; Van Den Bosch et al., 2003) that specifies functional forms for the luminosity distributions of central and satellite galaxies as a function of halo mass. Specifically, we use the Cacciato et al. (2009, hereafter C09) version of the CLF, but modified slightly to match our adopted cosmological model (van den Bosch, private communication). This methodology allows us to create a link between the distribution of DM haloes and that of the residing galaxies, while also differentiating between central and satellite galaxies (c.f. Eq. 32-39 in C09). The values of the parameters used in the analysis are $a_1 = 0.501$, $a_2 = 2.106$, $b_0 = -0.766$, $b_1 = 1.008$, $b_2 = -0.094$, $\sigma_c = 0.142$, $\gamma_1 = 3.273$, $\gamma_2 = 0.255$, $\log M_1 = 11.070$, $\log M_2 = 14.280$, and $\log L_0 = 9.935$.

We refer the reader to C09 for further discussion on the different variables used in CLF. We then abundance match the luminosities obtained from the CLF to the r-band absolute magnitude in the SDSS galaxy samples. As a result, our mock catalogs have the same exact luminosity function as the SDSS data.

2. For the `mstar-set` mock catalogues, we adopt the formalism of the *conditional stellar mass function* (CSMF) presented in Moster et al. (2010, hereafter M10), which provides the functional forms for the stellar mass distributions of central and satellite galaxies as a function of halo mass (c.f. Eq. 7-15 in M10). The values used in this model are taken from Table 3 $\sigma_m = 0$. The values of the parameters used in the analysis are $\log M_{1c} = 11.9347$, $(m_c/M)_0 = 0.0267$, $\beta_c = 1.0059$, $\gamma_c = 0.5611$, $\log M_2 = 11.9652$, $\sigma_\infty = 0.0569$, $\sigma_1 = 0.1204$, $\xi = 6.3020$, $\log M_{1s} = 12.1988$, $(m_s/M)_0 = 0.0186$, $\beta_s = 0.7817$, $\gamma_s = 0.7334$, $\log \Phi_0 = -11.1622$, $\lambda = 0.8285$, $\log M_3 = 12.5730$, $\alpha_\infty = -1.3740$, $\alpha_1 = -0.0309$, $\zeta = 4.3629$. We refer the reader to M10 for further discussion on each of the model parameters. Similarly to `mr-set` , we abundance match the stellar masses obtained from the CSMF to the stellar masses in each galaxy sample. As a result, the mock catalogs have the same exact stellar mass function as the SDSS data.

We assign specific star formation rates, $(g-r)$ colours and Sérsic indices to mock galaxies by first adopting the formalism presented in Zu & Mandelbaum (2016, hereafter Z16), and then sampling from the original distributions of sSFR, $(g-r)$ colour and Sérsic indices of the three SDSS galaxy samples. Specifically, we adopt the '*halo*' quenching model from Z16, which assumes that halo mass is the sole driver of galaxy quenching. According to that model, the red/quenched fraction of central and satellite galaxies is given by

$$f_{\rm cen}^{\rm red}(M_{\rm h}) = 1 - \exp\left[-(M_{\rm h}/M_{\rm h}^{\rm qc})^{\mu^{\rm c}}\right] \tag{2.1}$$

and

$$f_{\rm sat}^{\rm red}(M_{\rm h}) = 1 - \exp\left[-(M_{\rm h}/M_{\rm h}^{\rm qs})^{\mu^{\rm s}}\right], \tag{2.2}$$

where $M_{\rm h}^{\rm qc}$, $M_{\rm h}^{\rm qs}$, $\mu^{\rm c}$, and $\mu^{\rm s}$ are parameters of the model that Z16 fit to the observed clustering and galaxy-galaxy lensing measurements of red and blue galaxies in the SDSS. We assign each of our mock galaxies a probability of being quenched from equations (2.1) and (2.2) and we randomly designate it as 'active' or 'passive' consistent with that probability (e.g., if $f_{\rm sat}^{\rm red} = 0.8$ for a particular mock satellite galaxy, we give it an 80% chance of being labeled 'passive'). To assign realistic values of sSFR, $(g-r)$ colour, and Sérsic index to mock galaxies, we divide the observed distributions of these properties of `Mr19-SDSS` into 'active' and 'passive' distributions by making cuts at $\log_{10}{\rm sSFR} = -11$, $(g-r)_{cut} = 0.75$ and $n_{cut} = 3$ for sSFR, $(g-r)$ colour, and Sérsic index, respectively. For example, to assign sSFRvalues to mock galaxies from `Mr19-Mock` in the `mr-set` , we do the following. For each mock galaxy, we randomly draw a sSFRvalue from the active or passive distribution, depending on the designation that the mock galaxy has received. Moreover, we do this in a way that preserves the sSFR-luminosity distribution. For example, if a mock galaxy has been labeled 'active', we randomly select a real active galaxy from `Mr19-SDSS` that has a similar luminosity as the mock galaxy, and we assign its sSFRtot he mock galaxy. As a result of this procedure, the final joint sSFR-luminosity distribution of mock galaxies closely resembles the one for `Mr19-SDSS`. We repeat this procedure for `Mr20-SDSS` and `Mr21-SDSS`. For the case of `mstar-set` , we perform these steps with stellar masses instead of luminosities.

2.3.3.3 Geometrical and Redshift Cuts

As the final step, we construct volume-limited galaxy redshift survey catalogs from simulation boxes. First, we place a virtual observer at the center of the box and define the right ascension (RA) and declination (DEC) for each galaxy with respect to the virtual

observer. Then, for every mock galaxy, we compute the angular coordinates and redshift, including the effect due to line-of-sight peculiar velocities, also referred to as *redshift-space distortions*. The final result is a set of volume-limited galaxy catalogs in redshift-space, with the exact same geometry as that of SDSS. We construct sets of 100, 94, and 100 volume-limited realistic mock galaxy catalogs for the `Mr19`, `Mr20-SDSS`, and `Mr21-SDSS` samples, respectively. The resulting mock galaxy catalogs for the various galaxy samples are publicly available online[4].

2.4 Group-Finding Algorithm and Group Catalog

In this section, we summarize the motivation for using group finders, and the different aspects of the group finder that we use throughout the project (§2.4.1); the set of group (galaxy) catalogues constructed from the `NYU-VAGC` and MPA-JHU catalogs after the group finding process (§2.4.2); and the set of mock group catalogs and their corresponding content for each of the three volume-limited samples (§2.4.3).

2.4.1 Group-finding algorithm

We identify galaxy groups using the Berlind et al. (2006, hereafter `berlind-fof`) group-finding algorithm. This is a Friends-of-Friends (FoF; Huchra & Geller, 1982) algorithm that links galaxies recursively to other galaxies that are within a cylindrical linking volume around the galaxy. The FoF algorithm assumes no geometry for the resulting groups, but it encloses galaxies within an isodensity surface that is closely related to the st of chosen linking lengths. For a pair of galaxies i and j separated by an angular distance θ_{ij}, the projected separation $D_{\perp,ij}$, and the line-of-sight separation, $D_{\parallel,ij}$, are given by

$$D_{\perp,ij} = (c/H_0)(z_i + z_j)\ \sin(\theta_{ij}/2) \tag{2.3}$$

$$D_{\parallel,ij} = (c/H_0)|z_i - z_j| \tag{2.4}$$

[4] http://vpac00.phy.vanderbilt.edu/~caldervf/Group_Catalogue_Websites/

, where z_i and z_j correspond to the redshifts of the galaxies i and j, respectively. The galaxies are linked if

$$D_{\perp,ij} \leq b_{\perp}\bar{n}_g \tag{2.5}$$

$$D_{\parallel,ij} \leq b_{\parallel}\bar{n}_g \tag{2.6}$$

where $\bar{n}_g$ is the mean density of galaxies in the sample, and $b_{\perp}$ and $b_{\parallel}$ are the projected and line-of-sight linking lengths in units of the mean inter-galaxy separation, respectively. For a chosen set of linking lengths (one linking length in real-space and two linking lengths in redshift-space), the FoF algorithm produces a unique group catalogue. The projected and line-of-sight linking lengths used in this analysis are $b_{\perp} = 0.14$ and $b_{\parallel} = 0.75$ in units of the mean inter-galaxy separation, respectively. This choice of linking lengths was optimized by Berlind et al. (2006) to identify galaxy systems that live within the same DM halo, and the performance of the algorithm is expected to be *slightly* inferior for smaller groups with 10 or less member galaxies.

2.4.2 SDSS Group Catalogs

In this subsection, we introduce the set of **Group and Cluster** and **Member galaxies** catalogs for the three SDSS volume-limited samples, and describe in detail the information attached to each of the different catalogs.

2.4.2.1 Group and Cluster Catalogue

We apply the `berlind-fof` algorithm to the three volume-limited samples described in §2.3 using the sets of linking lengths from §2.4.1. For each galaxy sample we produce a set of group and group member catalogs, each containing information about the galaxy groups and member galaxies. The fractions of singletons or isolated galaxies are 41.18%, 45.19%, and 54.85% for the `Mr19-SDSS`, `Mr20-SDSS`, and `Mr21-SDSS` samples, respectively. The fractions of galaxies grouped in pairs are 17.49%, 18.74% and 19.82%. The remaining

41.33%, 36.07%, and 25.33% of galaxies are in groups of three or more members. The Mr19-SDSS, Mr20-SDSS, and Mr21-SDSS samples contain a total of 6439, 10124, and 5712 groups with richness $N \geq 3$, respectively.

The '*Group and Cluster*' catalog for the three galaxy samples include the following information:

Group and Cluster catalog

1 **Group ID**: This number corresponds to the ID of the galaxy group in the catalog.

2 **Group richness, $\mathbf{N_{gal}}$**: It indicates the total number of galaxies in the group.

3–5 **ra, dec, $\mathbf{c\bar{z}}$**: For each group, we calculate an unweighted group centroid, which consists of a group right ascension, declination, and mean velocity ($c\bar{z}$). RA and DEC are given in units of degrees, and $c\bar{z}$ in units of $\mathrm{km\,s^{-1}}$.

6 **Velocity dispersion, $\sigma_{\mathbf{v}}$**: We compute a group one-dimensional velocity dispersion given by

$$\sigma_v = \frac{1}{1+\bar{z}} \sqrt{\frac{1}{N-1} \sum_{i=1}^{N} (cz_i - c\bar{z})^2} \tag{2.7}$$

where 'N' is the total number of galaxies in the group, '$c\bar{z}$ ' is the mean velocity of the group, and '$c\bar{z}_i$' is the velocity of the ith member galaxy in the group.

7–8 **Absolute magnitudes, $\mathbf{M_{g,y}}$ and $\mathbf{M_{r,y}}$**: Total luminosity of the group. This parameters is the sum of the luminosities of each of the member galaxies. We compute the total group absolute magnitude in the g-band and r-band

$$M_{x,y} = -2.5 \log_{10} \left(\sum_{i=1}^{N} 10^{-0.4 M_{0.1}{}_{x,i}} \right) \tag{2.8}$$

where 'x' corresponds to the colour band (r-band or g-band), 'y' to the absolute r-band

magnitude limit of the volume-limit sample, and 'N' to the number of member galaxies in the group

9 $\mathbf{r_{edge}}$: Perpendicular distance to the group center from the survey edge.

10 **Projected radius, $\mathbf{R_{\perp,rms}}$**: Projected *rms* radius of the galaxy group. This variable is calculated as follows

$$R_{\perp,rms} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} r_i^2} \tag{2.9}$$

where 'r_i' corresponds to the projected distance between the i-th member galaxy of the group and the group centroid.

11 **Total stellar mass, log $\mathrm{M}_{\star,\mathbf{G}}$**: Logarithmic 10-based total stellar mass of the group. This quantity is the sum of the stellar mass of the individual member galaxies of the group, without making distinctions between stellar masses from the MPA-JHU or B03 catalogs.

12 **Total specific star formation rate, log $\mathrm{sSFR}_{\mathbf{G}}$**: Logarithmic 10-base total specific star formation rate of the group, sSFR. For each group, the total specific star formation rate computes as

$$\mathrm{sSFR}_G = \frac{\mathrm{SFR}_G}{M_{\star,G}} = \frac{\sum_{i=1}^{N} SFR_i}{\sum_{i=1}^{N} M_{\star,i}} \tag{2.10}$$

where 'N' refers to the number of member galaxies in the group with *measured* star formation rates (SFR's), and stellar mass '$M_\star$'. We *only* include a value for sSFR_G if more than 50% of the member galaxies in the group have a measured SFR from the MPA-JHU catalog. Otherwise, a value of 'nan ' gets assigned to sSFR_G instead.

13 **Group mass, M_{HAM}**: Logarithmic 10-base estimated mass via `HAM` method. We estimate the total mass of the group via *abundance matching*. This method assumes a monotonically increasing relation between the group *r*-band total luminosity, $M_{r,y}$, and the dark matter halo mass. We adopt the Warren et al. (2006) mass function for this purpose.

In Table 2.2, we present an excerpt of the structure of the '*Group and Cluster*' catalog for the `Mr19-SDSS` sample. The rest of the table and the tables for `Mr20-SDSS` and `Mr21-SDSS` are available at the url in 2.

2.4.2.2 Member Galaxies of Groups Catalog

We produce a separate set of 'Member Galaxies of Groups' catalogs for each of the three volume-limited samples. These catalogs include information about the member galaxies and their galaxy groups. For each galaxy, we include the following information

The '*Member Galaxies of Groups*' catalog for the three galaxy samples include the following information:

Member Galaxies of Groups catalog

1 **Galaxy ID**: Galaxy ID in the `NYU-VAGC` catalog. This number corresponds to the index of the galaxy in the list of properties from the `NYU-VAGC` catalog.

2–4 **Angular coordinates and velocity, ra, dec, cz**: Angular coordinates and velocity of the galaxy. The (J2000) right ascension and declination are given in units of degrees, and '*cz*' is given in units of $\mathrm{km\,s^{-1}}$.

5–6 **Absolute magnitudes, $M_{0.1,g}$ and $M_{0.1,r}$**: Absolute magnitudes of the galaxy in the *g*-band and *r*-band. These magnitudes have already been *k*-corrected to $z = 0.1$

7 **Sérsic index**: This parameter provides an insight into the *morphology* of the galaxy, as derived by the MPA-JHU catalog.

8 **Fiber collision flag, fibcol**: A value of 'x' other than '-1' indicates that the galaxy collided with the x-th galaxy in `NYU-VAGC` due to fiber collisions. A value of '-1' corresponds to an uncollided galaxy.

9 **Distance to survey edge, $\mathbf{r_{edge}}$**: Perpendicular distance of the galaxy from the survey edge. This quantity is given in units of h^{-1}Mpc .

10 **Galaxy stellar mass, log $\mathbf{M_\star}$**: Stellar mass of the galaxy, either from the MPA-JHU catalog or calculated using the B03 formalism. The stellar mass is determined as discussed in §2.3.2, and it is in units of $M_\odot$.

11 **Stellar mass flag, $\mathbf{flag_\star}$**: Stellar mass flag. It designates the source of the galaxy's stellar mass value. A value of '1' corresponds to stellar masses from the MPA-JHU catalog, while a value of '0' corresponds to stellar masses derived from the B03 formalism.

12 **Specific star formation rate, sSFR**: Logarithmic 10-base specific star formation rate of the galaxy. This value is calculated by dividing the star formation rate (SFR) of the galaxy by the galaxy's stellar mass, ($M_\star$). A galaxy is assigned a value of 'nan ' if it did not have a corresponding SFR value in the `NYU-VAGC` catalog. sSFRis given in units of yr^{-1}.

13 **Group ID**: ID of the galaxy group, to which the galaxy belongs. This variable is computed by the `berlind-fof` algorithm.

14 **Group galaxy type, $\mathbf{Type_G}$**: Group type of the galaxy. We denote a value of '1' to *group central* galaxies, and a value of '0' to *group satellite* galaxies. As mentioned in §2.2, central and satellite galaxies undergo different evolutionary paths. For this reason, it is important to make the distinction between the two types of galaxies. After determining the group membership of each galaxy, we designate the galaxy type based on the galaxy's stellar mass or absolute magnitude. For the case of `mr-set` , we des-

ignate the brightest galaxy of the group in the *r*-band as the *group central*, while the rest of galaxies are identified as *group satellites*. Hence, a galaxy group is composed of one bright group central and a number of group satellites. For the `mstar-set` , we identify the group central as the most massive galaxy in the group. This criterion is motivated by the idea that central galaxies grow in mass and brightness by galactic cannibalism (Dubinski, 1998; Cooray & Milosavljević, 2005), while satellite galaxies experience a series of events that strip them from their mass and inhibit star formation (e.g ram-pressure stripping and tidal stripping).

In Table 2.3, we present an excerpt of the 'Member Galaxies of Groups' catalog for the `Mr19`-SDSS sample. The rest of the table and the tables for `Mr20`-SDSS and `Mr21`-SDSS are available at the url in 2. Next, we discuss the mock group and galaxy catalogs that we produce after running the `berlind-fof` on the mock galaxy catalogs from §2.3.3.

Table 2.2: Group and Cluster Catalogue for the Mr19-SDSS sample

GroupID	N	RA (deg)	DEC (deg)	$c\bar{z}$ (km s^{-1})	σ_v (km s^{-1})	$M_{g,19}$	$M_{r,19}$	r_{edge} (h^{-1} Mpc)	$R_{\perp,rms}$ (h^{-1} Mpc)	log $M_{*,G}$ ($M_\odot$)	log sSFR$_G$ (yr^{-1})	log $M_{g,ab}$
Mr19-SDSS												
0	1	38.049133	0.224026	16195.24	0.0	-19.11	-20.03	1.825	0.0	10.55	-11.62	11.99
1	7	38.326144	0.042179	16203.43	60.13	-21.67	-22.51	1.605	0.431	11.6	-10.87	13.39
2	2	55.978967	0.526755	12075.56	126.14	-20.86	-21.61	0.77	0.185	11.08	-10.87	12.59
3	1	55.977487	0.459108	11119.14	0.0	-18.9	-19.35	0.733	0.0	9.77	-10.23	11.62
4	1	36.881847	0.877409	12220.45	0.0	-18.47	-18.95	0.353	0.0	9.67	-9.8	11.6
5	1	37.895509	1.003827	16314.83	0.0	-19.04	-19.95	0.125	0.0	10.61	-11.98	12.04
6	3	38.037866	0.928994	6513.03	212.85	-20.07	-20.82	0.135	0.034	10.46	-9.99	11.91
7	2	40.571982	1.028086	13777.96	102.14	-19.82	-20.49	0.056	0.022	10.72	-10.41	12.15
8	3	41.335543	0.924811	7378.73	93.02	-20.3	-21.16	0.158	0.023	11.19	-11.06	12.76
9	1	42.216541	0.989533	8376.85	0.0	-20.25	-21.22	0.087	0.0	11.23	-12.05	12.82

Note—The rest of the table and the tables for Mr20-SDSS and Mr21-SDSS can be found in the electronic version of the MNRAS, or at http://vpac00.phy.vanderbilt.edu/~caldervf/Group_Catalogue_Websites/

Table 2.3: Member Galaxies of Groups and Clusters for Sample Mr19-SDSS

ID	RA (deg)	DEC (deg)	cz (km s^{-1})	$M_{0.1_g}$	$M_{0.1_r}$	Sersic	fibcol	r_{edge} (h^{-1} Mpc)	log M_* ($M_\odot$)	flag$_B$	log sSFR (yr^{-1})	Group ID	Type$_G$
Mr19-SDSS													
749	38.049133	0.224026	16195.24	-19.11	-20.03	5.903	-1	1.825	10.552	1.0	-11.621	0	1
750	38.352526	0.212491	16134.08	-18.35	-19.19	3.637	-1	1.664	10.17	1.0	-11.105	1	0
1759858	38.29581	0.067402	16123.53	-19.72	-20.36	1.958	-1	1.846	10.429	1.0	-10.156	1	0
751	38.363598	0.210654	16203.85	-20.11	-20.89	3.315	-1	1.674	10.912	1.0	-10.735	1	0
768168	38.231888	-0.114193	16192.7	-18.55	-19.46	4.372	-1	1.413	10.369	1.0	-11.515	1	0
1759897	38.557012	0.188478	16265.25	-18.93	-19.71	5.903	-1	1.822	10.45	1.0	-10.48	1	0
768173	38.249256	-0.137101	16207.11	-18.82	-19.69	4.586	-1	1.417	10.197	1.0	-10.378	1	0
768169	38.232506	-0.131238	16298.07	-20.49	-21.46	3.572	-1	1.398	11.318	1.0	-11.941	1	1
1121	55.989529	0.437805	11982.79	-19.4	-20.33	4.489	-1	0.829	10.713	1.0	-11.353	2	0
1127	55.968565	0.614348	12168.36	-20.53	-21.21	3.31	-1	0.711	10.837	1.0	-10.687	2	1
1129	55.977487	0.459108	11119.14	-18.9	-19.35	0.987	-1	0.733	9.774	1.0	-10.227	3	1
1796	36.881847	0.877409	12220.45	-18.47	-18.95	1.51	-1	0.353	9.675	1.0	-9.804	4	1
1885	37.895509	1.003827	16314.83	-19.04	-19.95	3.219	-1	0.125	10.609	1.0	-11.977	5	1
1887	37.998277	0.910002	6546.46	-19.33	-20.11	5.903	-1	0.16	9.61	1.0	-9.243	6	0
1889	37.929902	0.904399	6280.79	-18.62	-19.21	1.484	-1	0.155	10.009	1.0	-10.271	6	0
1898	38.177512	0.970528	6711.87	-18.48	-19.35	3.821	-1	0.091	10.159	1.0	-11.424	6	1
2067	40.530732	1.018938	13853.5	-18.69	-19.47	2.527	-1	0.077	10.555	1.0	-10.825	7	1
2073	40.613687	1.037334	13702.42	-19.34	-19.95	2.764	-1	0.034	10.221	1.0	-10.048	7	0
2117	41.280586	0.950152	7346.99	-18.77	-19.34	2.981	-1	0.127	10.219	1.0	-10.629	8	0
2126	41.364842	0.914361	7303.35	-19.69	-20.66	2.772	-1	0.17	11.083	1.0	-11.14	8	1
2125	41.360894	0.910134	7485.87	-18.47	-19.33	4.515	-1	0.178	10.244	1.0	-11.311	8	0
2186	42.216541	0.989533	8376.85	-20.25	-21.22	2.752	-1	0.087	11.23	1.0	-12.052	9	1
2242	42.785181	1.031012	18097.97	-18.99	-19.43	1.141	-1	0.059	9.87	1.0	-9.72	10	1

Note—The rest of the table and the tables for Mr20-SDSS and Mr21-SDSS can be found in the electronic version of the ApJ, or at http://vpac00.phy.vanderbilt.edu/~caldervf/Group_Catalogue_Websites/

2.4.3 Mock Group Catalogs

We are interested in producing a set of mock catalogs analogous to the ones presented in §2.4.2. We apply the `berlind-fof` group-finder to the set of mock galaxy catalogs from §2.3.3, and produce corresponding sets of *Mock Group Cluster Catalog* and *Mock Member Galaxies of Groups and Clusters* catalogs for the `Mr`19-SDSS, `Mr`20-SDSS, and `Mr`21-SDSS samples.

The format of 'Mock Group Cluster Catalog' is similar to the one presented in §2.4.2.1, with the exceptions of M_g, r_{edge} and log $sSFR_G$ of each group. The final format of this version of 'Mock Group Cluster Catalog' is shown in Table 2.4. For each group, we include the following information:

Mock Group and Cluster catalog

1–9 These parameters are calculated in the same way as those presented in Table 2.2

10 **Group's 'true' mass, log $\mathbf{M_{halo}}$**: 'True' group mass. For a galaxy group, we wish to know the *true*, realistic mass that represents the underlying DM distribution of the halo. We estimate this mass for galaxy groups based on the contributions of number of galaxies from each of the DM halos that contribute galaxies to the group to the overall number of galaxies in the group. We quantify these contributions by determining the '*pointing fraction*' of group-halo pairs, $f_{h,g}$, using Equation 2.11,

$$f_{h,g} = \frac{N_c^2}{N_h \times N_g} \tag{2.11}$$

where 'N_c' corresponds to the number of galaxies shared between the galaxy group and the halo; 'N_h', to the total number of galaxies in a given halo; and 'N_g', to the total number of galaxies in the group. The halo, whose $f_{h,g}$ is the largest out of all the group-halo pairs will be identified as the halo that is most representative of the group, i.e. the group is mostly comprised of galaxies from this DM halo. We assign the mass

Table 2.4: Mock Group Catalogue Parameters

Column	Param.	Description	Unit
1	Group ID ..	ID of the galaxy group	-
2	N	Group richness	-
3	RA	Right ascension of group centroid	deg
4	Dec	Declination of group centroid	deg
5	$c\bar{z}$	Mean velocity of group	$\mathrm{km}s^{-1}$
6	σ_v	One-dimensional line-of-sight velocity dispersion of group	$\mathrm{km}s^{-1}$
7	$M_{r,y}$	Total r-bandabsolute magnitude	-
8	$R_{\perp,rms}$	Rms projected distance	h^{-1}Mpc
9	$M_{\rm HAM}$	Estimated group mass via abundance matching	$h^{-1}\mathrm{M}_\odot$
10	$M_{\rm halo}$	Estimated *true* group mass	$h^{-1}\mathrm{M}_\odot$

Note. — This table summarizes the content of the Group and Cluster catalogues for the mock catalogues. These catalogues have a similar format as those in Table 2.2. These catalogues are available for the Mr19, Mr20, and Mr21 mock galaxy samples.

of this halo t the group, and refer to this mass as the group's *halo mass*, $M_{\rm halo}$. In the case where there are two halos with equal $f_{h,g}$values, we randomly choose one of the halos and assign its mass $M_{\rm halo}$. Each galaxy group in the mock galaxy catalog has both a $M_{\rm halo}$ and $M_{\rm HAM}$ masses, while groups in the group catalogs only have $M_{\rm HAM}$ masses. In the case of a *perfect* group finder, both masses, $M_{\rm HAM}$ and $M_{\rm halo}$, would be very similar. This is not the case due to group-finding errors

Additionally, we construct a 'Mock Group Member Galaxies' catalog that includes the true positions, velocities, halo and group membership of mock galaxies for the Mr19-Mock, Mr20-Mock, and Mr21-Mock samples. Table 2.5 shows the format of this catalog. For each mock galaxy, we include the following information

Mock Member Galaxies of Groups catalog

1–2 Angular coordinates, **ra, dec**: Angular coordinates of the galaxy. The (J2000) right ascension and declination are given in unts of degrees.

3 **cz_{obs}**: Line-of-sight component of the galaxy with respect to the observer. This component includes the redshift-space distortions due to the peculiar velocity of the galaxy. This quantity is given in units of $\mathrm{km\,s^{-1}}$.

4 **cz_{true}**: Line-of-sight velocity component of the galaxy, without the effects of redshift-space distortions. This provides a measure of the exact location of the galaxy with respect to the observer. This quantity is given in units of $\mathrm{km\,s^{-1}}$.

5 **$cz_{\perp}$**: Tangential velocity component of the galaxy's peculiar velocity. This quantity is connected to the absolute value of the peculiar velocity of the galaxy, v_p, as follows:

$$v_p^2 = (cz_{\rm obs} - cz_{\rm true})^2 + cz_{\perp}^2 \tag{2.12}$$

This quantity is given in units of $\mathrm{km\,s^{-1}}$.

6 **Absolute Magnitude, $M_{0.1,r}$**: r-band absolute magnitude of the galaxy. This value is assigned using the ranking of the luminosities from the CLF and the absolute magnitudes from SDSS DR7, as described in §2.3.3.2. In the case of `mstar-set` , stellar mass is included instead of $M_{0.1,r}$.

7 **Halo ID**: True halo membership. This variable indicates the ID of the DM halo, to which the galaxy belongs.

8 **Halo richness, $N_{gal,\,h}$**: Total number of galaxies in the galaxy's host DM halo. Some galaxies may not be present in a catalog due to geometrical and/or redshift cuts made to the sample.

9 **Halo galaxy type, $Type_{halo}$**: Type of the galaxy in the *halo*. This parameter indicates if the galaxy is a central or satellite in the halo. As discussed in §2.3.3.1, a halo can only have one central galaxy, but the number of satellite galaxies can range from zero to multiple satellite galaxies, depending on the mass of the halo. We denote a value of

Table 2.5: Mock Group Member Galaxies Catalogue Parameters

Column	Param.	Description	Unit
1	RA	(J2000) Right Ascension	deg
2	Dec	(J2000) Declination	deg
3	$cz_{\rm obs}$	Line-of-sight observed velocity	$\rm km\,s^{-1}$
4	$cz_{\rm true}$	True line-of-sight velocity	$\rm km\,s^{-1}$
5	$cz_{\perp}$	Tangential velocity component	$\rm km\,s^{-1}$
6	$M_{0.1_r}$	r-band absolute magnitude	-
7	HaloID	True halo membership	-
8	$\rm N_h$	Number of galaxies in halo	-
9	Halo type . . .	Galaxy type in the halo	-
10	GroupID . . .	Group membership of the galaxy	-
11	$\rm Type_G$	Galaxy type in group	-

Note. — This table summarizes the content of the Group and Cluster catalogues for the mock catalogues. These catalogues are available for the `Mr19`, `Mr20`, and `Mr21` mock galaxy samples.

'1' to *halo central* galaxies, and a value of '0' to *group satellite* galaxies.

10 **(g – r) colour of galaxy**: The difference between the absolute magnitudes in the g-band and r-band. This variable was assigned to mock galaxies in a manner similar to that of sSFR.

11 **Galaxy morphology**: This parameter gives an insight into the morphology of the galaxy. This variable was assigned to mock galaxies in a similar fashion as sSFRand $(g-r)$ colours.

Finally, we construct a set of 'perfect' mock group catalogs based on the the mock group galaxy catalogs. The idealized versions of group catalogs are comprised of groups that perfectly recover the group membership of galaxies, i.e. a group is comprised of galaxies from the same DM halo. For each of these groups, we recompute group properties such as total stellar mass and r-band luminosity, specific star formation rate, velocity dispersion, among others, using the new set of galaxies. This approach allows for the study of 'perfect', idealized systems of galaxies, as they do not, by construction, include any group-finding

errors. This set of catalogs will prove to be useful when determining the impact that group-finding errors have on different metrics that quantify the efficiency and performance of the `berlind-fof` group-finder.

The reader is directed to the URL in 2 to obtain copies of both the SDSS (mock) group and galaxy catalogs for the rest of `Mr19`, `Mr20`, and `Mr21` samples.

2.5 Group-Finding Errors

A group finder can suffer of different failure modes that can ultimately change the properties of the overall group population. In this section, we identify different ways a group-finder can fail at identifying galaxy groups (§2.5.1); the failures involves in determining galaxy types within a group, along with the metrics used to evaluate the overall performance of the group finder (§2.5.2); and the effects that group-finding errors have in the estimation of group mass (§2.5.3).

2.5.1 Merging and Fragmentation

The goal of a group-finder is to correctly identify the galaxies from the same DM halo. In an ideal scenario, a 'perfect' group-finder would be able to identify and group those galaxies from the same DM halo, while distinguishing among those galaxies from distinct halos. Such algorithm would produce a set of *perfect* group catalogs, both in real- and redshift-space. Unfortunately, one of the main challenges that a group-finder faces is to correctly identify the group membership of galaxies. Figure 2.3 shows the schematic of the second largest galaxy group as defined by the `berlind-fof` group-finder in the `Mr19`-SDSS galaxy catalog. Each panel corresponds to the two-dimensional projection of the group in Cartesian coordinates centered at the group centroid in redshift-space. The black cross corresponds to the coordinates of the group centroid. Each point corresponds to the location of member galaxies in the group, with the size of the point being representative of the galaxy's stellar mass, i.e. larger dots correspond to galaxies with larger stellar masses than galaxies with smaller dots. Each galaxy is also color-coded by its specific star formation

rate, with bluer colours corresponding to more active galaxies than redder galaxies. Finally, the position of the group central galaxy is shown by the black circle.

The X-Y projection is analogous to how the galaxy group would look on the sky to an observer, while the Z-coordinate is parallel to the line-of-sight direction of the observer. In redshift-space, galaxy groups suffer from redshift-space distortions, as they appear elongated along the line-of-sight direction due to the peculiar motions of the galaxies, as they move within the group itself. This *finger-of-god* effect distorts the relative positions and velocities of galaxies, as some galaxies may appear to be closer to the observer than in reality, and vice-versa. By visually inspecting the galaxy group in Figure 2.3, this group portraits the effects of redshift-space distortions, as it is possibly the result of galaxies from distinct halos being merged into a single group. We use the mock catalogs from §2.3.3 to investigate this further.

Throughout this analysis, we adopt the terminology presented in Duarte & Mamon (2014) and define two failure modes in the identification of group membership of galaxies, and refer to these as 'fragmentation' and 'merging' of DM haloes. A halo has been *fragmented* if its member galaxies have been assigned to multiple galaxy groups. On the contrary, a halo has been *merged* if its galaxies have been assigned to a group that is comprised of galaxies from multiple haloes. These concepts are illustrated in Figure 2.2, with 'fragmentation' in top panel and 'merging' in the bottom panel. Solid circles corresponds to the boundaries of DM haloes, within which coloured the points correspond to galaxies that truly reside in the haloes. Galaxies from the same halo share the same colour. Dashed circles on the right correspond to the boundaries of galaxies groups, as identified by the group finder. Additionally, this figure includes the *pointing fractions*, $f_{h,g}$, for each halo-group pair, as described in Equation 2.11.

We are interested in exploring the impact that group-finding errors have on the galaxy assignment to groups. In a complementary fashion to Figure 2.3, we explore how these errors affect the group population in mock catalogs. Figure 2.4 shows the schematic of

the largest galaxy group in the Mr19-Mock galaxy sample across all mock galaxy catalogs. The panels in the figure are similar to the ones from Fig. 2.3. The black cross corresponds to the coordinates of the group centroid as defined the `berlind-fof` group-finder. Each dot corresponds to the Cartesian coordinates of galaxies, with each galaxy being color-coded by its *halo* membership, i.e. galaxies from the same DM halo share the same color. Additionally, member galaxies of the galaxy group have an additional black edge. The colored circles in the $x-y$ panel show the haloes that contribute with galaxies to the group, with each circle's radius denoting the virial radius of the halo. Finally, the location of the group central galaxy is depicted by the black star within the black circle.

The group in Fig. 2.4 is comprised of galaxies from 116 different DM haloes and a total of 528 galaxies. This group is an example of multiple haloes being merged into a single galaxy group. This aspect is an artifact of the FoF algorithm, and it can lead to other cascading effects, as discussed in §2.5.2.

2.5.2 Galaxy type designation

The second challenge during the group-finding process pertains to the correct classification of group central and satellite galaxies. As mentioned in §2.4, we define the group central galaxy as being the brightest or most massive galaxy in the group, depending on the type of group catalog. Group-finders that rely solely on the position and velocity information of galaxies can incorrectly place galaxies into groups due to 'merging' and 'fragmentation' errors, and these lead to incorrect designations of the galaxy type of member galaxies. They also increase the likelihood for halo centrals to be misidentified as group satellites, and vice-versa. Ultimately, the misclassification of galaxy type lead to error in the overall group mass estimation, as discussed in the next section.

The impact of group-finding errors on the classification of galaxy types can be measured by quantifying the following metrics of a group galaxy catalog:

- **Purity**: Fraction of group central (satellite) galaxies that are also central (satellite)

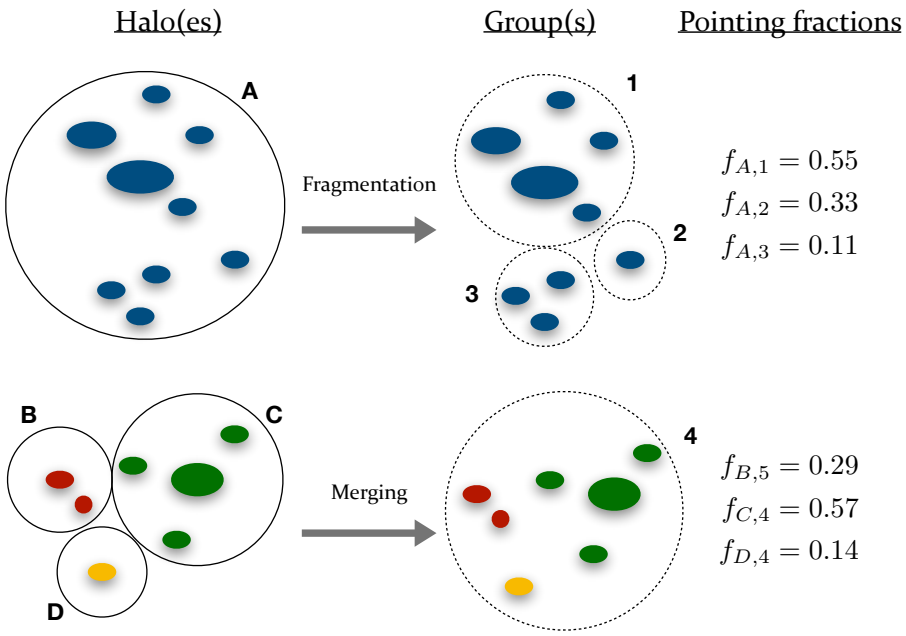

Figure 2.2: Illustration of 'fragmentation' (top) and 'merging' (bottom) of DM haloes. Solid circles on the left correspond to the boundaries of DM haloes, within which the coloured points refer to galaxies that truly reside in the haloes. Each point is colour-coded based on its host halo. Dashed circles on the right correspond to the boundary of galaxy groups, as identified by the group-finder. The relative size of each circle can be interpreted as the group/halo mass of the system. Additionally, we compute the pointing fractions for each halo-group pairs, as described in Equation 2.11.

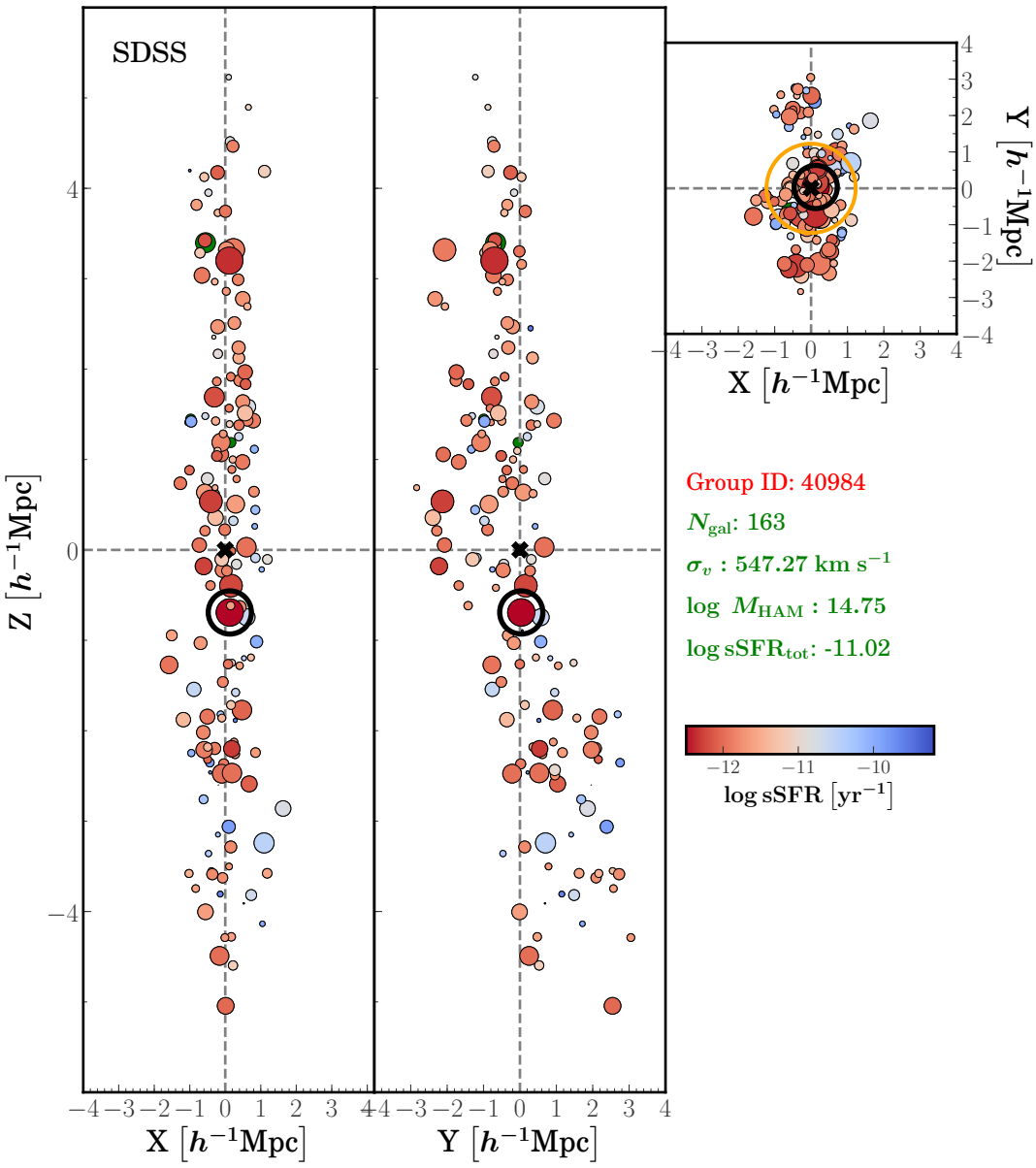

Figure 2.3: Schematic of the second largest galaxy group as defined by the `berlind-fof` group-finder in the `Mr19-SDSS` galaxy catalog. Each panel corresponds to a two-dimensional projection of the group in Cartesian coordinates centered at the centroid of the galaxy group. The black cross corresponds to the coordinates of the group centroid. Additionally, each point corresponds to the locations of member galaxies in the group, with the size of the point being representative of the amount of stellar mass in each galaxy, i.e. smaller dots correspond to less massive galaxies, while larger dots correspond to more massive member galaxies in terms of stellar mass content. Each galaxy is color-coded based on its specific star formation rate, with bluer colors corresponding to more active galaxies than more redder galaxies. Finally, the central galaxy is depicted by the black circle.

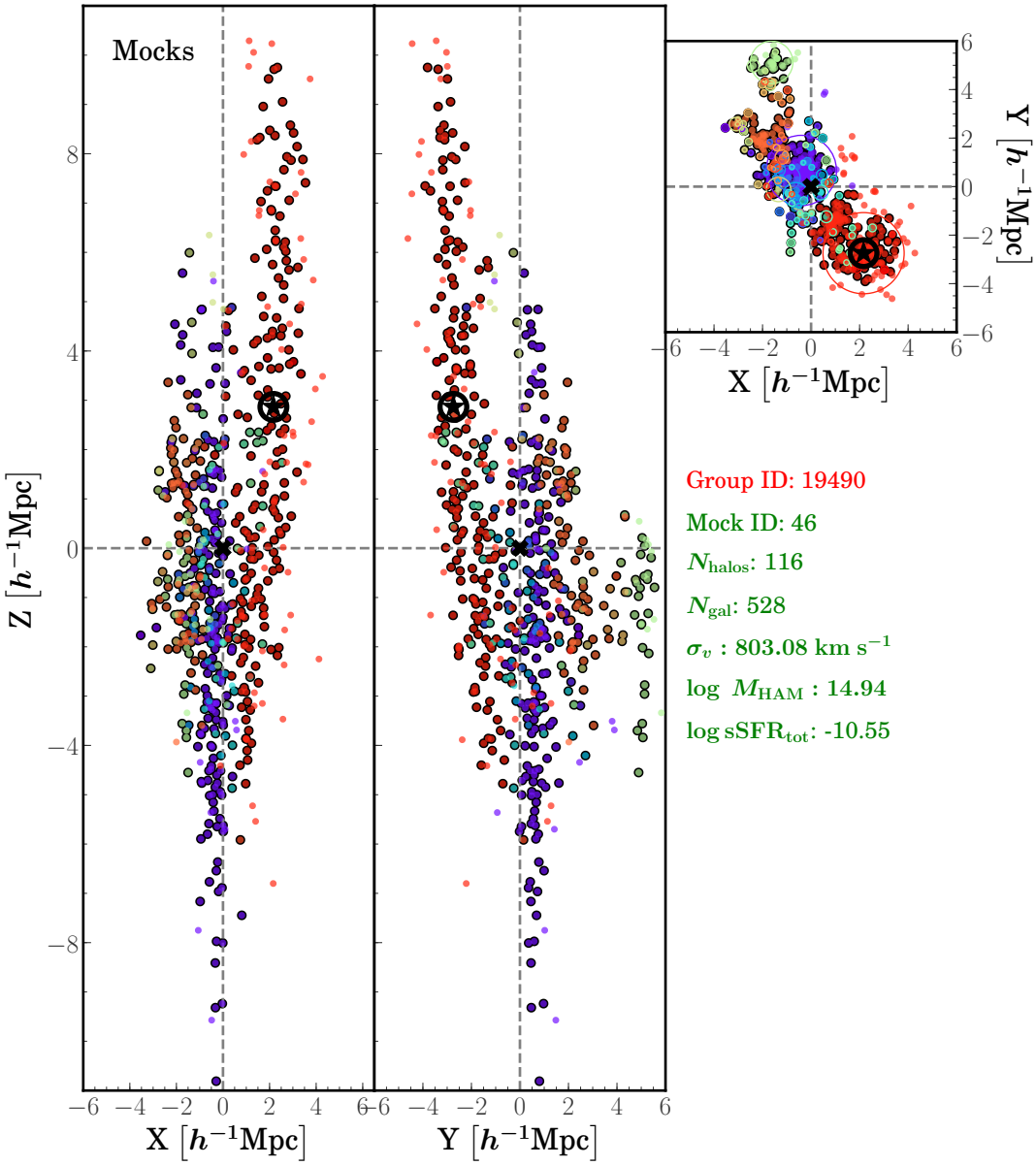

Figure 2.4: Schematic of the largest galaxy group found by the `berlind-fof` group-finder across all galaxy catalogs in the `Mr19-Mock` galaxy sample. The panels in this figure are similar to those in Fig 2.3. The black cross corresponds to the coordinates of the group centroid as defined by the group-finder. Each dot corresponds to the Cartesian coordinates of galaxies, with each galaxy being color-coded by its *halo* membership, i.e. galaxies from the same dark matter halo share the same color. Additionally, member galaxies of the galaxy group have an additional black edge. In the $x-y$ panel, colored circles show the halos that contribute with galaxies to the group, with each circle's radius denoting the virial radius of the halo. Finally, the central galaxy is depicted by the black star within the black circle.

galaxies in their host DM halo.

- **Completeness**: Fraction of halo central (satellite) galaxies that are also central (satellite) galaxies in their galaxy group.

We adopt the nomenclature presented in Campbell et al. (2015) to describe purity and completeness by defining the completeness of central and satellite galaxies as

$$C_{\rm cen} = \frac{N_{\rm cen|cen}}{N_{\rm cen|cen} + N_{\rm sat|cen}} \tag{2.13}$$

and

$$C_{\rm sat} = \frac{N_{\rm sat|sat}}{N_{\rm sat|sat} + N_{\rm cen|sat}} \tag{2.14}$$

where, for example $N_{\rm cen|sat}$ refers to the number of halo satellite galaxies that have been identified as group central galaxies. Similarly, purity of central and satellite galaxies are defined as

$$P_{\rm cen} = \frac{N_{\rm cen|cen}}{N_{\rm cen|cen} + N_{\rm cen|sat}} \tag{2.15}$$

and

$$P_{\rm sat} = \frac{N_{\rm sat|sat}}{N_{\rm sat|sat} + N_{\rm sat|cen}} \tag{2.16}$$

To clarify further, the total number of galaxies and number of galaxy groups in the catalog are defined by

$$N_{\rm gal} = N_{\rm cen|cen} + N_{\rm cen|sat} + N_{\rm sat|cen} + N_{\rm sat|sat} \tag{2.17}$$

$$N_{\rm groups} = N_{\rm cen|sat} + N_{\rm cen|cen} \tag{2.18}$$

These two metrics provide an insight into how well a group-finder is performing, and about its efficiency at classifying central and satellite galaxies in a group. Figure 2.5 shows the purity and completeness levels for central and satellite galaxies in the `Mr19-Mock` galaxy sample. In the top panel, the light blue (red) bars correspond to the fractions of group central (satellite) galaxies that are also central (satellite) galaxies in their corresponding DM haloes, as a function of galaxy group richness. Similarly, the dark-shaded blue (red) bars correspond to the fraction of group central (satellite) galaxies that are also halo central (satellite) galaxies *and* reside in DM haloes of similar sizes as their corresponding groups. In the bottom panel, the light blue (red) bars correspond to the fractions of halo central (satellite) galaxies that have been correctly classified as group central (satellite) galaxies, in bins of halo richness. Similarly, the dark-shaded blue (red) bars correspond to the fractions of halo central (satellite) galaxies that have been correctly classified as group central (satellite) galaxies in their group, and reside in galaxy groups of similar sizes as their host DM haloes.

Figure 2.5 shows prominent results about the inner-workings of the `berlind-fof` group-finder. From this figure, we notice that halo central galaxies are being correctly classified as group centrals $80-90\%$ of the time, regardless of the number of galaxies of the galaxy group. Halo satellites exhibit a similar trend as that of halo centrals, as they are being correctly classified as group satellites at a similar rate. However, this fraction gets smaller with increasing halo richness and when controlling for group and halo sizes, as satellites tend to be correctly identified as group satellites $60-70\%$ of the time for halos with nine or more galaxies. Additionally, we notice that in poor groups with as much as 4 galaxies, $\sim 50\%$ of group satellite galaxies are truly halo satellites. This fraction improves for groups with more galaxies, reaching purity levels of up to $\sim 84\%$. Group centrals tend to be true halo centrals about 90% of the times, regardless of the group richness. This fraction becomes smaller when taking the group and halo richness into account, reaching purity levels of up $\sim 60\%$ for groups with three and 8 galaxies. This fraction increases

up to purity levels of $\sim 83\%$ for group centrals that are truly central galaxies and reside in galaxy groups of similar size as their host haloes. This figure shows that the `berlind-fof` group-finder is efficient but not perfect at correctly identifying central and satellite galaxies within a group environment.

Additionally, the results of this analysis are in agreement with those by Campbell et al. (2015), in which they report similar trends of purity and completeness for central and satellite galaxies. In their work, they analyze these metrics for three different group-finding algorithms as functions of group and halo mass. The results in this work differ from Campbell et al. (2015) in that they use a single luminosity-based group galaxy catalog, and compute these metrics as functions of mass. However, in this analysis we utilize all of the mock catalogs to compute purity and completeness metrics as a function group and halo richnesses. We will discuss this further in §2.7.

2.5.3 Halo mass estimates

The third challenge during the group-finding process deals with correctly estimating the masses of galaxy groups, and how these are affected by group-finding errors, such as the ones discussed in previous sections. As mentioned in §2.4.2 and §2.4.3, group masses are determined by assuming a monotonically increasing relation between the group total stellar mass or luminosity and a mass function. However, group-finding errors, such as fragmentation and merging of haloes, can lead to a false estimate of the group mass. For example, in the case of fragmentation, in which galaxies from a single halo are being distributed among various groups, the total stellar mass or luminosity of the group is lower than expected, resulting in an underestimate of the total mass of the group. The opposite is true for 'merging', in which galaxies from distinct haloes are assigned to a single group. In this case, the total stellar mass and luminosity of the group is greater than expected when no merging takes place, thus leading to an overestimate of group mass.

To understand this further, we explore the effects that group-finding errors, such as

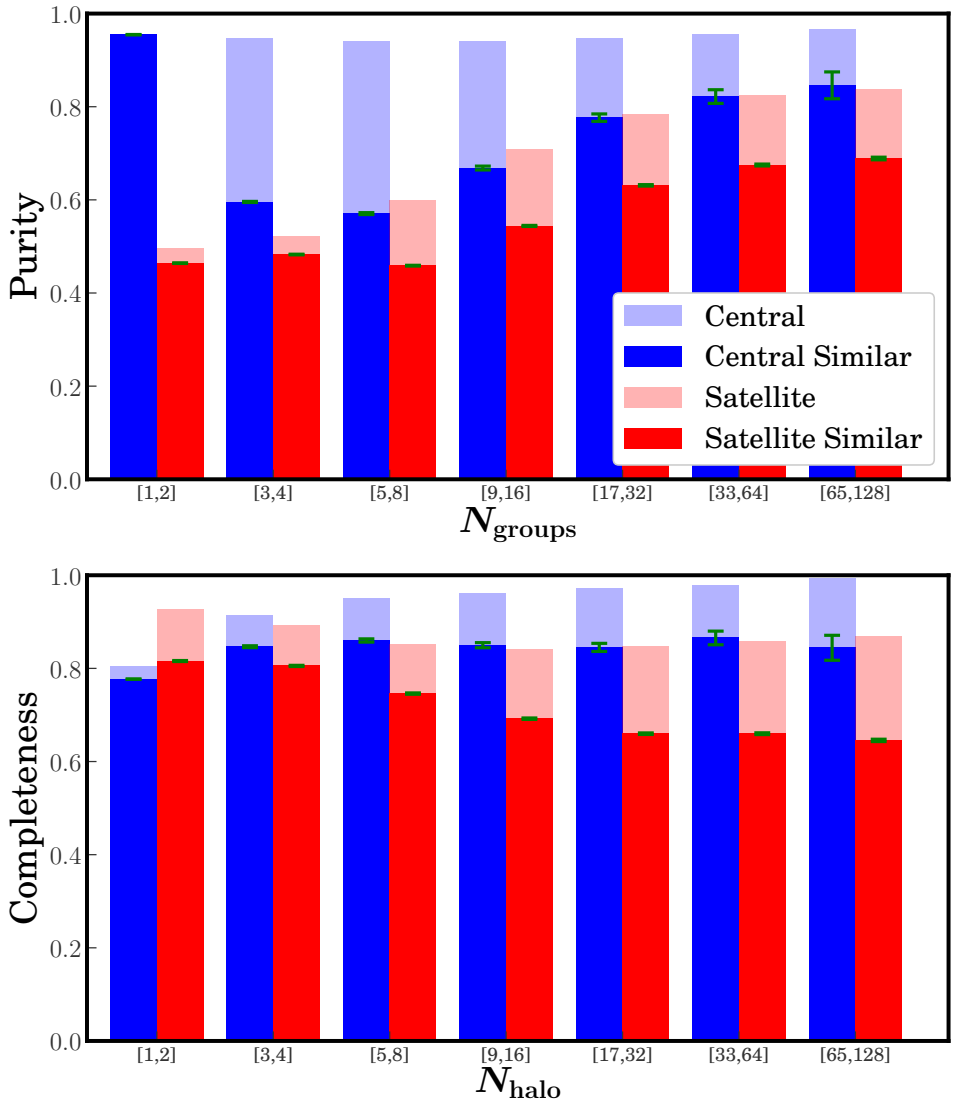

Figure 2.5: Purity and completeness of central and satellite galaxies in the Mr19-Mock galaxy sample. *Top panel*: The light blue (red) bars correspond to the fractions of group central (satellite) galaxies that are also central (satellite) galaxies in their corresponding DM halo, as function galaxy group richness. Similarly, the dark-shaded blue (red) bars correspond to the fraction of group central (satellite) galaxies that are 1) also central (satellite) galaxies in their DM halo, and 2) reside in DM halos of similar size as their corresponding galaxy group. All fractions are given in terms of galaxy group richness. *Bottom panel*: In this panel, the light blue (red) bars correspond to the fractions of halo central (satellite) galaxies that are classified as group central (satellite) galaxies, as function of halo richness. Similarly, the dark-shaded blue (red) bars show the fractions of halo central (satellite) galaxies that 1) are classified as group central (satellite) galaxies, and 2) reside in galaxy groups of similar size as their parent DM halo. All fraction in this panel are given in terms of number of galaxies in a given halo.

fragmentation and merging, have on the final assessment of group mass. We define two sets of galaxy groups based on the 'pointing fractions' (see Equation 2.11), and refer to these as *good* and *bad* matches. In the case of merging and fragmentation, haloes can contribute with galaxies to multiple groups, while a group can be comprised of galaxies from multiple haloes (see Fig. 2.2). However, we are interested in determining which group is mostly representative of a given halo, and vice versa. For example, Fig. 2.2 shows the case where a halo 'A' is being fragmented into 3 different groups. However, this halo *points* only to the first galaxy group, as most of its galaxies get assigned to this group. Similarly, the first group *points* to halo 'A', as most of its galaxies are true members of this halo. A 'good match' corresponds to the group-halo pair, in which both the halo and group point to each other. Any other group is considered a catastrophic failure or 'bad match'. Catastrophic failures are indicative of *fragmentation*, and they account for $\sim 3.4\%$ of galaxy groups in `Mr19`-Mock sample. Additionally, we make use of the set of *perfect* galaxy groups in mock catalogs (§2.4.3) to determine how well group masses are being recovered, when no group-finding errors are involved.

Figure 2.6 presents the comparison of groups masses as determined by `HAM` , $M_{\rm HAM}$, to those determined by the *pointing* method (§2.4.3), $M_{\rm halo}$, for three types of galaxy groups in `Mr19`-Mock, i.e. 'good matches' (left), 'bad matches' (centre), and 'perfect groups' (right). The yellow, solid lines and errorbars correspond to the median and standard deviation of $M_{\rm halo}$ in bins of $M_{\rm HAM}$. The blue shading shows the frequency of galaxy groups in two-dimensional bins, where the number of groups in each bin has been normalized by the value for the bin containing the most galaxies. The dashed black lines show the one-to-one relation between $M_{\rm HAM}$ and $M_{\rm halo}$. Additionally, in the bottom panel, the yellow line corresponds to the scatter in $M_{\rm halo}$ as a function of $M_{\rm HAM}$. By comparing $M_{\rm halo}$ and $M_{\rm HAM}$, one can determine the level of merging and fragmentation induces by the group-finder.

Figure 2.6 shows prominent results, as *good* and *bad* matches constitute two distinct populations when comparing $M_{\rm HAM}$ and $M_{\rm halo}$ masses. In the case of 'good' matches',

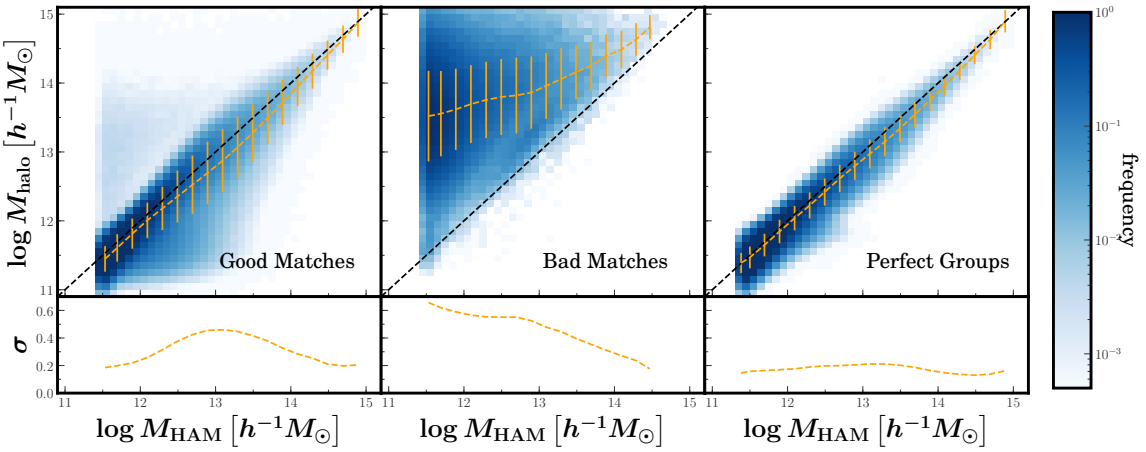

Figure 2.6: Galaxy group masses for Mr19-Mock galaxy groups calculated from HAM compared to those determined using the *pointing* method, for three different kinds of galaxy groups, i.e. good matches (left), bad matches (centre), and perfect groups (right) *Top panels*: The x-axis shows mass estimates for galaxy groups through HAM , and y-axis shows the mass estimates of galaxy groups through the *pointing* formalism. The blue shading shows the frequency of galaxy groups in two-dimensional bins, where the number of groups in each bin is normalized by the value for the bin containing the most galaxies. Yellow solid lines and errorbars correspond to the mean and standard deviation of $M_{\rm halo}$ in bins of $M_{\rm HAM}$. The dashed black lines show the one-to-one relation between mass estimates. *Bottom panels*: Scatter in $M_{\rm halo}$ as a function of estimated group mass, $M_{\rm HAM}$. Yellow solid lines correspond to the standard deviation of $M_{\rm halo}$ as function of $M_{\rm HAM}$.

the median relation is in close agreement to the one-to-one relation, indicating that $M_{\rm HAM}$ is able to recover the true, underlying DM halo mass with minimal error. Additionally, the scatter in $M_{\rm halo}$ peaks at $M_{\rm HAM} \approx 10^{13}\ h^{-1}{\rm M}_{\odot}$, reaching values of up to 0.5 dex. The scatter is smaller at smaller and larger $M_{\rm HAM}$, with values as low as 0.2 dex for $M_{\rm HAM} \sim 10^{11.5}\ h^{-1}{\rm M}_{\odot}$ and $M_{\rm HAM} \sim 10^{15}\ h^{-1}{\rm M}_{\odot}$. This indicates that mid-sized galaxy groups are more prone to be affected by group-finding errors than low-mass and high-mass systems, and suggests that *merging* of haloes takes place within this mass range much more frequently than at low- and high-mass regimes. On the other hand, 'bad matches' exhibit a different relation. This population is the result of fragmentation, in which member galaxies of haloes are split into multiple groups. This group-finding error causes $M_{\rm HAM}$ estimates to be much lower than in reality. The median relation of $M_{\rm halo}$ is much larger than that of 'good' matches, and includes a much larger scatter in $M_{\rm halo}$. The scatter in $M_{\rm halo}$ is large for small groups with $M_{\rm HAM} \approx 10^{11.5}\ h^{-1}{\rm M}_{\odot}$, and it decreases with increasing $M_{\rm HAM}$, reaching values down to 0.2 dex in $M_{\rm halo}$ for groups with $M_{\rm HAM} \approx 10^{14.5}\ h^{-1}{\rm M}_{\odot}$. As mentioned earlier, this population only constitutes $\sim 3.4\%$ of the galaxy groups in Mr19-Mock. Lastly, 'perfect groups' show a relatively small but steady scatter in $M_{\rm HAM}$ of ~ 0.2 dex across $M_{\rm HAM}$ masses. This result suggests that group-finding errors tend to inflate group mass estimates to levels of up to ~ 0.2 dex from the true, underlying scatter between $M_{\rm HAM}$ and $M_{\rm halo}$. Moreover, mid-sized groups have mass estimates that are on average 50% higher than the *true* group mass, while both low mass systems and massive clusters have unbiased mass estimates. We obtain similar results for the Mr20-Mock and Mr21-Mock samples.

2.6 Stellar Content of Group Centrals and Group Satellites

In this section, we explore the stellar content of galaxy groups in SDSS and verify our results carefully taking into account any effects induces by group-finding errors. In §2.6.1, we characterize the stellar-to-halo mass relation of central and satellite galaxies as function of group mass, and predict a correction factor to account for group-finding errors. In §2.6.2,

we explore the role of group mass in galaxy quenching by carefully analyzing this relation for central and satellite galaxies in different environments.

2.6.1 Stellar-to-Halo mass relation

One of the most important aspects of the galaxy-halo connection relates to the correlation between galaxy properties and those of haloes. Most empirical models of galaxy formation relate galaxy properties to properties of their host DM haloes, with larger haloes hosting larger galaxies with relatively low scatter in the stellar-to-halo mass relation (SHMR) More et al. (2009); Yang et al. (2009); Leauthaud et al. (2012); Reddick et al. (2013); Watson & Conroy (2013); Tinker et al. (2013); Gu et al. (2016); Behroozi et al. (2018). Hence, it has become common to explore how average galaxy growth depends on the average growth of haloes (see Wechsler & Tinker (2018) for a review).

In this section, we are interested in exploring how the SHMR is affected by group-finding errors, and by much does it change from the *true*, underlying SHMR. To test the fidelity of the inferred SHMR from group galaxy catalogs, we use mock galaxy group catalogs from Mr19-Mock, and compare it to the *perfect* mock group catalogs. This comparison allows us to calculate a correction factor that can be applied to Mr19-Mock to account for group-finding errors and recover an idealized SHMR of central galaxies. Figure 2.7 shows the SHMR of central and satellite galaxies as a function of group or halo mass. in Mr19-Mock group catalogs. In the top panel, the dashed, blue line and the shaded contours correspond to the median relation of stellar mass, $M_\star$, and the 1σ, 2σ, and 3σ ranges of $M_\star$ for group centrals as functions of group mass in the Mr19-Mock sample. Similarly, the dotted, cyan line shows the median relation of 'true' central galaxies as a function of *halo* mass, in the perfect version of the Mr19-Mock group catalogs. Additionally, we add the Moster et al. (2010) and Behroozi et al. (2013) SHMR of central galaxies as functions of halo mass, for comparison purposes. In the middle panel, we are showing a correction factor of $\log M_\star$ for mock group central galaxies, $\Delta \log(M_\star)_{\rm med}$, as a function of group mass. The blue,

dashed line shows the correction factor, by which one would modify the SHMR of 'group' central galaxies to remove the effects of group-finding errors. Finally, in the bottom panel, we show correction factor for the scatter in galaxy stellar mass, $\Delta\sigma(\log M_\star)$, for central galaxies as a function of group mass. The red, orange, and cyan lines correspond to the correction factors of the 1σ, 2σ, and 3σ scatter in SHMR of central galaxies as functions of group mass, respectively. This relations show the amount of correction one would need to apply to the scatter in SHMR for group central galaxies in order to remove any effects induced by group-finding errors.

Figure 2.7 shows prominent results, as it shows that the SHMR of group centrals is not different by much to that of *true* central galaxies, i.e. group-finding errors do not affect the SHMR of central galaxies drastically, yet group-finding errors do influence this relation slightly. The first panel of this figure compares the SHMR of group centrals to that of halo central galaxies. The result of such comparison is that these two exhibit a similar trend with increasing group/halo mass, and are in agreement with the Moster et al. (2010) relation, and not so much with Behroozi et al. (2010). Moreover, the second panel of this figure shows the correction factor needed to recover the idealized SHMR for central galaxies as function of group mass. This quantity is the result of taking difference between the logarithmic 10-base median SHMR relation of true halo centrals in the 'perfect' groups to that of group centrals in the `Mr19`-Mock sample. For example, if the correction factor, at a given group mass, were to have a value of $\Delta\log(M_\star)_{\rm med} = 0.1$, this would mean that the median SHMR of group centrals at that given group mass would be adjusted by increasing it by 0.1, in order to remove the effects of group-finding errors. This panel shows prominent results. It shows that the median SHMR relations of group central galaxies and true central galaxies are only affected slightly by group-finding errors. Typically, $\Delta\log(M_\star)_{\rm med}$ ranges from -0.05 dex to 0.05 dex in $\log M_\star$, with lower-mass group of $M_{\rm HAM} \sim 10^{11.6}\,h^{-1}{\rm M}_\odot$ reaching values of up to $\Delta\log(M_\star)_{\rm med} \approx 0.2$ dex, while higher-mass groups of $M_{\rm HAM} \sim 10^{14.4}\,h^{-1}{\rm M}_\odot$ reaching $\Delta\log(M_\star)_{\rm med} \approx 0.05$ dex. Similarly, the third panel indicates by much the scatter of the

SHMR of group centrals as function of group mass would need to increase or decrease in order to exclude any effects from group-finding errors. For example, if $\Delta\sigma(\log M_\star) = -0.2$ at a given group mass, one would need to reduce the width of the scatter by subtracting 0.2 dex from the median SHMR of group centrals, in order to remove effects induced by group-finding errors. This panel shows that the SHMR of group centrals in groups with $M_{\rm HAM} \lessapprox 10^{12}\,h^{-1}{\rm M}_\odot$ typically would need to increase by factors of up to 0.6 dex in $\log M_\star$ in order to recover the idealized SHMR of centrals. On the other hand, the SHMR of central galaxies that live in groups with $M_{\rm HAM} \gtrapprox 10^{12}\,h^{-1}{\rm M}_\odot$ would need to reduce their scatter by factors of up to $\Delta\sigma(\log M_\star) \approx -0.5$ dex.

After evaluating the results from Fig.2.7 and determining that only a *small* correction is needed on the SHMR to account for group-finding errors, we feel confident to apply such correction to the SHMR of central galaxies in SDSS. Figure 2.8 show the SHMR of group centrals and group satellites in the `Mr19-SDSS`. This figure is similar in fashion to the top panel of Fig. 2.7 with a few exceptions. The solid, blue line shows the median SHMR of group central galaxies in SDSS as a function of group mass, $M_{\rm HAM}$. The shaded contours show the 1σ, 2σ, and 3σ ranges of $\log M_\star$ for group central galaxies. The dotted, blue line shows the result of applying the correction factor, $\Delta\log(M_\star)_{\rm med}$, from Fig. 2.7 to the median SHMR relation of group centrals in SDSS. For comparison purposes, the green and dark green dashed lines correspond to the SHMR of central galaxies from Moster et al. (2010) and Behroozi et al. (2010) as functions of halo mass. Lastly, the red dots refer to the SHMR of group satellites.

We find our SHMR results from Fig. 2.8 to be in agreement with those from Behroozi et al. (2010) across $M_{\rm HAM}$masses. However, they differ from the Moster et al. (2010) at masses larger than $M_{\rm HAM} \geq 10^{13}\,h^{-1}{\rm M}_\odot$. This novel approach leverages the use of mock group catalogs to understand the errors induced by the group-finding process, and aims at recovering SHMR of central galaxies while removing any systematic offset induced by group-finding errors.

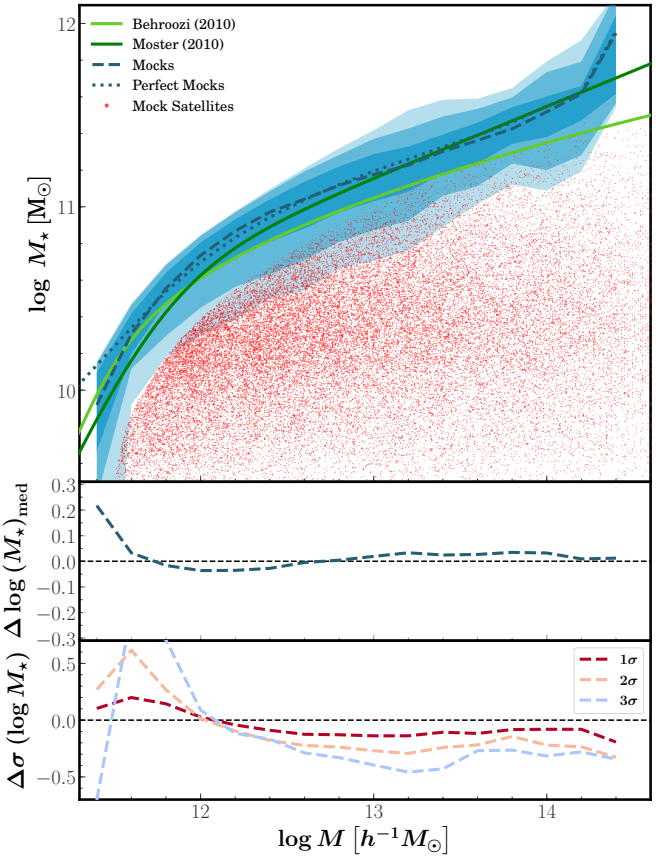

Figure 2.7: Stellar-to-Halo mass relation (SHMR) of central and satellite galaxies in the Mr19-Mock galaxy sample. *Top panel*: Galaxy stellar mass of central and satellite galaxies as a function of mass. The dashed, blue line corresponds to the median relation of stellar mass, $M_\star$, of *group* central galaxies as a function of group mass in the Mr19-Mock group catalogs. The shaded contours show the 1σ, 2σ, and 3σ ranges of $M_\star$ for group centrals in Mr19-Mock. Similarly, the dotted, cyan line corresponds to the median relation of *halo* central galaxies as a function of 'halo' mass, in the perfect version of the Mr19-Mock group catalogs. Additionally, we plot the Moster et al. (2010) and Behroozi et al. (2010) SHMR relations of central galaxies as functions of halo mass, for comparison purposes. Finally, the red dots refer to the SHMR of group satellite galaxies in the Mr19-Mock sample. *Middle panel*: Correction factor of $\log M_\star$ for mock group central galaxies, $\Delta\log(M_\star)_{\rm med}$, as a function of group mass. The dashed, blue line refers to the correction factor of SHMR in bins of group mass. This relation shows by much one needs to modify the SHMR of central galaxies to remove the effects of group-finding errors. *Bottom panel*: Correction factor of the scatter in galaxy stellar mass, $\Delta\sigma(\log M_\star)$, for central galaxies as a function of group mass. The red, orange, and cyan lines correspond to the correction factors of the 1σ, 2σ, and 3σ scatter in SHMR of central galaxies as functions of group mass, respectively. These lines show the factor, by which the scatter would need to change in order to remove the effects of group-finding errors in the SHMR of central galaxies.

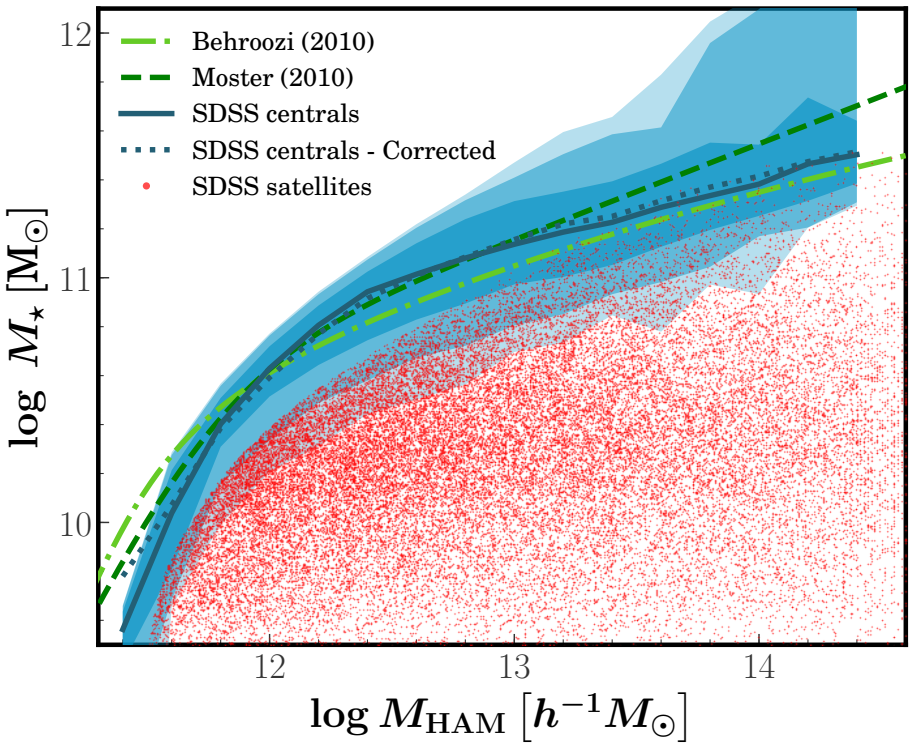

Figure 2.8: Similar in fashion to the top panel of Fig. 2.7, with the exception that the central and satellite galaxies corresponds to the Mr19-SDSS sample. The solid, blue line shows the median relation of the SHMR of group central galaxies in SDSS as a function of group estimated mass, $M_{\rm HAM}$. The shaded contours correspond to the 1σ, 2σ, and 3σ ranges of $M_\star$ for group central galaxies. The dotted, blue line shows the *corrected* version of the median relation of the SHMR of group centrals, after applying the correction factor, $\Delta\log(M_\star)_{\rm med}$ from Fig. 2.7. The green and dark green lines correspond to the Moster et al. (2010) and Behroozi et al. (2010) SHMR of central galaxies as function of halo mass for comparison purposes. Finally, the red dots refer to the SHMR of group satellites in Mr19-SDSS.

2.6.2 Unraveling the role of group mass in galaxy quenching

It is clear that galaxy properties tend to correlate with galaxy environment. As noted in §2.2, a galaxy sample can be divided into a star-forming blue cloud and a more quenched red sequence. The origin of this relation is still unclear, yet galaxies in more dense environments exhibit an enhance quenched fraction relative to that of galaxies in more isolated environments. In this work, we are interested in understanding exploring the relationship between the specific star formation rate (sSFR) of galaxies and galaxy stellar mass as function of different types of environments, i.e. as a function of group mass. Moreover, we investigate how group-finding errors affect these relations, and whether or not central and satellite galaxies exhibit similar trends.

To explore this further, we first quantify to what degree group-finding errors can affect the sSFR – $M_\star$ relation of central and satellite galaxies as function of group mass. Figure 2.9 shows the galaxy specific star formation rate as a function of galaxy stellar mass for group centrals (top row) and group satellites (bottom row) in the `Mr19`-Mock galaxy sample. Each column corresponds to bins of group mass, $M_{\rm HAM}$. Each panel is divided into active (top of the panel) and passive (bottom of the panel) galaxies, with division at log sSFR = 11. The blue (red) solid lines and errorbars correspond to the median and standard deviation of log sSFR of group central (satellite) galaxies in bins of stellar mass, $M_\star$. For comparison purposes, we show the median relations for group central and group satellite galaxies in each of the panels as dashed blue and red lines, respectively. Similarly, Figure 2.10 the sSFR – $M_\star$ relation of galaxies, with the exception that top and bottom rows correspond to the *true* central and satellite galaxies in in the 'perfect' version of `Mr19`-Mock sample, respectively. Each column corresponds to a bin in group mass, $M_{\rm HAM}$.

Figure 2.9 and 2.10 show interesting results. By comparing each of the panels on both figures, we can determine to which degree group-finding errors affect the sSFR – $M_\star$ relations for central and satellite galaxies as function of group mass. At fixed group mass and galaxy type, we conclude that group-finding errors do not severely affect the median

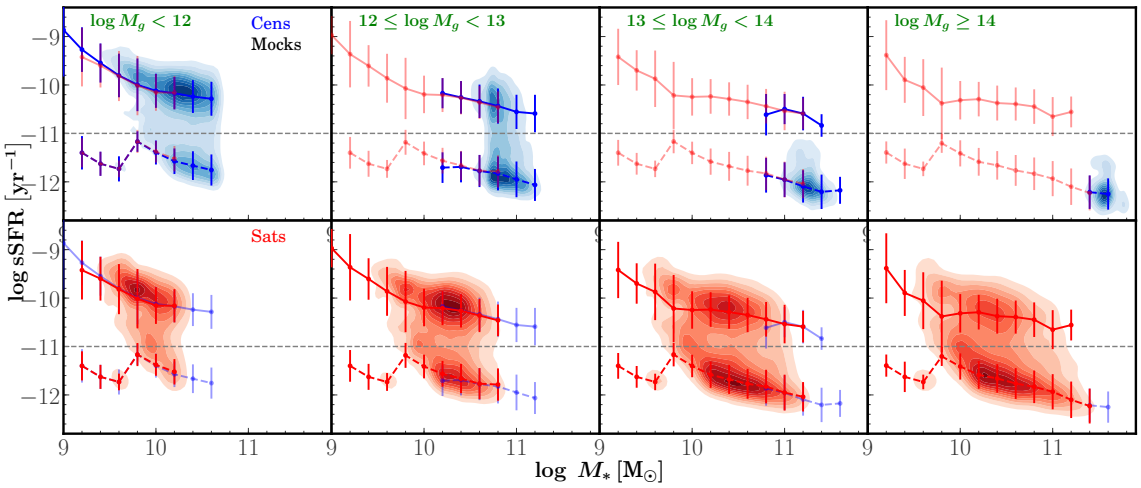

Figure 2.9: Galaxy specific star formation rate as a function of galaxy stellar mass for group centrals (top row) and group satellite (bottom row) in the Mr19-Mock galaxy sample. Each column corresponds to a bin of group mass, $M_{\rm HAM}$, as listed in each panel. Each panel is divided into active (top) and passive (bottom) galaxies, with division at log sSFR = −11. Blue (red) solid lines and errorbars correspond to the median and standard deviations of log sSFR of group central (satellite) galaxies in bins of $M_\star$. We show the lines for group central and group satellite galaxies in *each* of the panels to make it easier to compare the relations among galaxy types.

sSFR – $M_\star$ relation of galaxies. This result implies that at fixed group mass and galaxy type, the sSFR – $M_\star$ relations of active and passive galaxies are unbiased to group-finding errors, and one can confidently use the `berlind-fof` group-finder to identify central and satellite galaxies and characterize the sSFR – $M_\star$ relation.

The results from Figs. 2.9 and 2.10 are encouraging results, and provide us with the confidence of applying the `berlind-fof` to SDSS and characterize the sSFR – $M_\star$ relation of central and satellite galaxies as a function of group mass. Figure 2.11 shows the sSFR – $M_\star$ relation for group centrals (top row) and group satellite (bottom row) galaxies in the Mr19-SDSS galaxy sample. This figure is similar in fashion to Fig. 2.9. This figure shows two main trends. First, most group centrals become quenched in groups of $M_{\rm HAM} \geq 10^{13}\,h^{-1}{\rm M}_\odot$, while the quenching halo mass scale is higher for group satellites. Second, at fixed group mass and galaxy quenching state (active or passive), group central

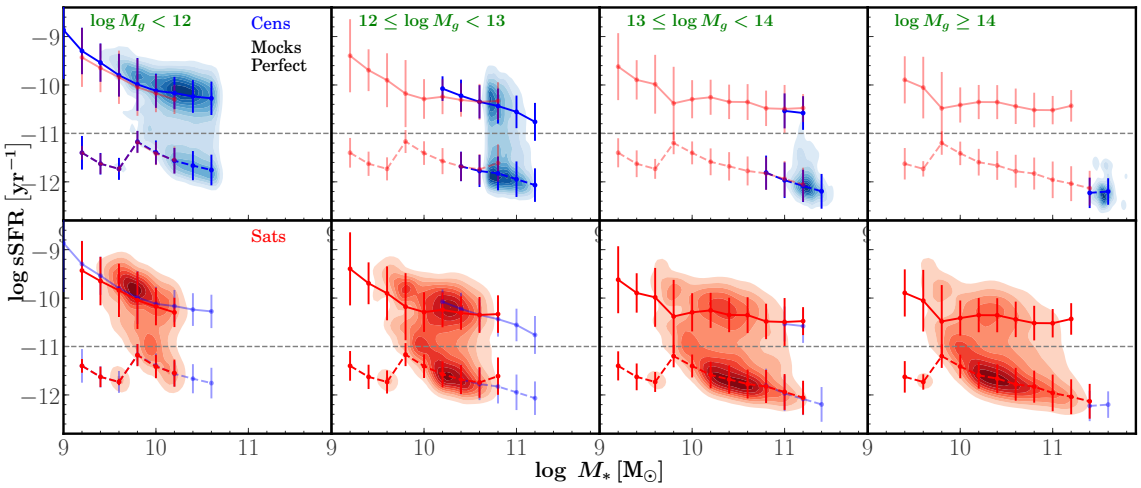

Figure 2.10: Similar to Fig. 2.9, except that the galaxy sample corresponds to the *perfect* version of Mr19-Mock. This figure tries to recover the log sSFR – $M_\star$ relation of galaxies if no group-finding errors were involved.

and group satellite galaxies exhibit similar trends with $M_\star$. This suggests that, for example, active centrals and active satellites share a similar trend with increasing $M_\star$. This relation persists with increasing group mass. These results are in agreement with other analyses that have found that galaxies in denser environments exhibit an enhanced quenched fraction relative to that of galaxies in more isolated, less dense environments (Dressler, 1980; Postman & Geller, 1984; Kauffmann et al., 2004). Finally, we conclude that group-finding error do not have impact on the sSFR – $M_\star$ relation of active and passive galaxies, and one can confidently employ such algorithm to further characterise this relation.

2.7 Summary and Discussion

In this paper, we investigate the ability to confidently make use of galaxy group catalogs to explore different aspects of the galaxy-halo connection, including the stellar-to-halo mass relation (SHMR) of central and satellite galaxies. Moreover, we explore the role that group mass plays in determining the quenching state of galaxies as a function of galaxy stellar mass. We are motivated to conduct a comprehensive and robust study of the impact that group-finding systematic errors have on group mass assignment, galaxy type determination,

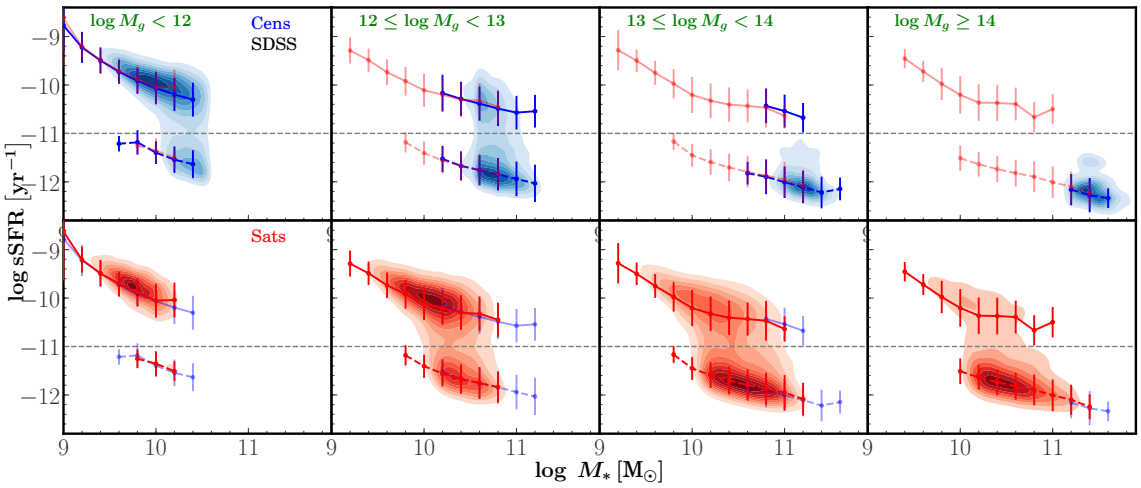

Figure 2.11: Similar to Fig. 2.9, except that the galaxy sample used corresponds to the Mr19-SDSS galaxy sample.

and characterization of the SHMR of galaxies, in order to implement the use of group catalogs to further constrain the relationship between galaxies and their host haloes.

Our best cosmological model of structure formation and evolution predicts galaxies to reside in dark matter (DM) haloes, with the tendency of luminous galaxies to reside in groups and clusters surrounded by less luminous neighbors. This motivates the usage of groups to better understand galaxy formation and evolution. These systems can be identified through various methods, and can identified as overdensities of galaxies or as extended X-ray sources. With the advent of spectroscopic surveys, galaxy groups can now be identified by the closeness of their member galaxies with minimum errors associated with projected effects with the use of group-finding algorithms. An ideal group-finder would identify member galaxies from the same halo while distinguishing galaxies from distinct haloes. In reality, these group-finders are limited to work with observations, and it is not possible to perfectly recover the group membership of galaxies, resulting in 'group-finding' errors. These error may impact inferred group statistics, and result in wrong statistical inferences about the galaxy and group populations. In this analysis, we perform a comprehensive analysis of the effect of group-finding errors on several group-related statistics,

such as purity and completeness of the sample, and group mass assignment, among others, on three volume-limited galaxy samples of SDSS DR7. Additionally, we test the group-finding algorithm on multiple types of realistic mock catalogs to test the fidelity of our results.

The main results of our work are as follows

- We construct three volume-limited samples that contain all galaxies brighter than r-band absolute magnitudes M_r = -19, -20, and -21 from SDSS Data Release 7. We refer to these catalogs as `Mr19-SDSS`, `Mr20-SDSS`, and `Mr21-SDSS`, respectively. We assign stellar masses ($M_\star$) and star formation rates (SFR) to galaxies from the MPA-JHU Value Added Catalogue DR7 (MPA-JHU). We perform this for every galaxy of the three volume-limited galaxy samples. Additionally, we construct analogous versions of the SDSS volume-limited samples using realistic mock galaxy catalogs from a cosmological N-body simulation that traces the evolution of DM in the Universe. Finally, we apply the the Berlind et al. (2006) (hereafter `berlind-fof`) group-finding algorithm to `Mr19-SDSS`, `Mr20-SDSS`, and `Mr21-SDSS` and their corresponding mock catalogs, and construct galaxy group catalogs for each sample. We refer to these catalogs as `Mr19-Mock`, `Mr20-Mock`, and `Mr21-SDSS`. Lastly, we construct 'perfect' versions of the mock group catalogs that include *no* redshift-space distortions and do not suffer from group-finding errors.

- We test the efficiency of `berlind-fof` to correctly distinguish between central and satellite galaxies in a galaxy sample. We compute the 'purity' and 'completeness' metrics for the `Mr19-Mock` sample, and conclude that central galaxies are correctly identified as central galaxies within groups $80-90\%$ of the time, regardless of galaxy group richness. Satellite galaxies are correctly identified $60-70\%$ of the time by `berlind-fof`. Moreover, in poor groups with as much as 4 galaxies, $\sim 50\%$ of group satellite galaxies are truly satellites in their host haloes. This fraction improves for groups with more galaxies, reaching purity levels of satellites up to $\sim 84\%$. Group centrals tend to be central galaxies in their host haloes $\sim 90\%$ of the times, and this fraction is unbiased to group richness.

These results are in agreement with C15, in which they report similar trends of purity and completeness for central and satellite galaxies.

- We test the proficiency of estimating group masses when making use of group galaxy catalogs. We classify groups from Mr19-Mock into 'good' and 'bad' matches, and compare these against 'perfect' galaxy groups to determine how well group masses are being recovered for galaxy groups. We find that abundance matched mass, $M_{\rm HAM}$, are good estimates of the true, underlying DM halo mass. The scatter in $M_{\rm HAM}$ peaks at $M_{\rm HAM} \approx 10^{13}\,h^{-1}{\rm M}_{\odot}$, reaching values of up to 0.5 dex. This scatter gets smaller at smaller and larger $M_{\rm HAM}$. This indicates that mid-sized galaxy groups are more prone to be affected by group-finding errors than low-mass and high-mass systems, and it suggests that merging of haloes takes place within this mass range much more frequently than at other mass regimes.

- After carefully characterizing how group-finding errors affected inferred galaxy group statistics, we explore the SHMR of galaxies and compute a correction factor that removes the effects of group-finding errors. We compare the SHMR of central galaxies in galaxy groups from Mr19-Mock to that of 'true' central galaxies in the 'perfect' group catalogs from Mr19-Mock. This lead to a correction factor of the median SHMR relation of central galaxies as a function of group mass, $M_{\rm HAM}$. We determine that only a small correction is needed for the SHMR to account for group-finding errors. We later applied this correction factor to the SHMR of group centrals in SDSS, and compare the *corrected* relation of that of previous empirical models of the SHMR of central galaxies.

- Finally, we explore the dependence of group mass on the quenching state of galaxies as a function of galaxy stellar mass, sSFR $- M_{\star}$, and explore how it is affected by group-finding errors. We test the impact of group-finding errors on this relation by comparing active and passive galaxies in bins of group mass from Mr19-Mock and the perfect group catalogs, and find that group-finding errors do not severely affect the median sSFR –

$M_\star$ relations of central and satellite galaxies as functions of group mass. Moreover, we explore the sSFR – $M_\star$ relation in Mr19-SDSS. We find that most group central galaxies become quenched in groups of of $M_{\rm HAM} \geq 10^{13}\,h^{-1}{\rm M}_\odot$, while the quenching halo mass scale is higher for group satellite. We also conclude that, at fixed group mass and galaxy quenching state (active and passive), central and satellite galaxies exhibit similar trends with $M_\star$, and follow the same median sSFR – $M_\star$ relation. This relation persists with increasing group mass.

- In this paper, we release various sets of galaxy group catalogs for different volume-limited SDSS galaxy samples, including the set of realistic mock group catalogs. We make these catalogs available for download.

These results demonstrate the feasibility of using group galaxy catalogs to explore aspects of the galaxy-halo connection. Moreover, this analysis provides a robust and comprehensive examination of the impact that group-finding errors have on inferred galaxy group statistics, and validates the use of galaxy group catalogs to further explore various aspects of the galaxy-halo connection. To conclude, group-finding algorithms suffer from systematic errors that may have an impact on the the overall group membership of galaxies, and can induce systematic offsets to inferred statistics. However, galaxy group catalogs have proven to be extremely useful when examining various aspects of the galaxy-halo connection.

Chapter 3

SMALL- AND LARGE-SCALE GALACTIC CONFORMITY IN SDSS DR7

The following work has been accepted by the Monthly Notices of the Royal Astronomical Society Journal (Calderon et al. 2018) and is reprinted below in its entirety

Small- and Large-Scale Galactic Conformity in SDSS DR7

Victor F. Calderon[1], Andreas A. Berlind[1], Manodeep Sinha[2]

[1] Department of Physics and Astronomy, Vanderbilt University, Nashville, TN 37235

[2] Centre for Astrophysics and Supercomputing, Swinburne University of Technology, Hawthorn, Victoria 3122, Australia

[3]ARC Centre of Excellence for All Sky Astrophysics in 3 Dimensions (ASTRO 3D)

3.1 Abstract

Galactic conformity is the phenomenon whereby galaxy properties exhibit excess correlations across distance than that expected if these properties only depended on halo mass. We perform a comprehensive study of conformity at low redshift using a galaxy group catalogue from the SDSS DR7 and their satellites (1-halo), and between central galaxies in separate haloes (2-halo). We use the quenched fractions and the marked correlation function (MCF), to probe for conformity in three galaxy properties, $(g-r)$ colour, specific star formation rate (sSFR), and morphology. We assess the statistical significance of conformity signals with a suite of mock galaxy catalogues that have no built-in conformity, but contain the same group-finding and mass assignment errors as the real data. In the case of 1-halo conformity, quenched fractions show strong signals at all group masses. However, these signals are equally strong in mock catalogues, indicating that the conformity signal is spurious and likely entirely caused by group-finding systematics, calling into question previous claims of 1-halo conformity detection. The MCF reveals a significant detection of

radial segregation within massive groups, but no evidence of conformity. In the case of 2-halo conformity, quenched fractions show no significant evidence of conformity in colour or sSFRonce compared with mock catalogues, but a clear signal using morphology. In contrast, the MCF reveals a small, yet highly significant signal for all three properties in low mass groups and scales of $0.8-4h^{-1}$Mpc, possibly representing the first robust detection of 2-halo conformity.

3.2 Introduction

Characterizing the relation between the properties of galaxies and their host dark matter (DM) haloes – referred to as the "galaxy-halo" connection – has emerged as a powerful tool to constrain theories of galaxy formation with statistical measurements in galaxy surveys. The phenomenon called "*galactic conformity*" is a subtle feature of this galaxy-halo connection, whereby galaxy properties are spatially correlated even *at fixed halo mass*. Specifically, several studies have claimed to detect a correlation between quenching properties of galaxies, such as morphology, gas content, star formation rate, neutral hydrogen content, and broad-band colour, and those of neighbouring galaxies (Weinmann et al., 2006; Ann et al., 2008; Ross & Brunner, 2009; Kauffmann et al., 2010; Prescott et al., 2011; Wang & White, 2012; Kauffmann et al., 2013; Knobel et al., 2015; Hartley et al., 2015; Wang et al., 2015; Kawinwanichakij et al., 2016; Berti et al., 2017; Zu & Mandelbaum, 2018). This effect of "galactic conformity" exists over two distance regimes, both between central and satellite galaxies within the same halo, and between galaxies separated by several virial radii of their haloes. We refer to these regimes as "1-halo" and "2-halo" conformity, respectively (Hearin et al., 2015). 2-halo conformity is closely linked to "halo assembly bias" or "secondary bias" (e.g., Gao et al., 2005; Wechsler et al., 2005; Salcedo et al., 2017), whereby the clustering of haloes depends on secondary properties, like age, at fixed mass, and "galaxy assembly bias" (e.g., Croton et al., 2007), whereby galaxies inherit this clustering when their observed properties correlate with these secondary halo properties.

Assembly bias provides a natural explanation for 2-halo conformity (Hearin et al., 2015, 2016).

Conformity detections are notoriously difficult to make because it is hard to be confident that measurements are truly being made at fixed halo mass and also to know whether a given galaxy pair lives in the same halo or not. At the present time, there are several detection claims of 1-halo conformity at both low and high redshifts. These studies have looked at correlations between galaxy properties of the central galaxies and their respective satellite galaxies. Some have used isolation criteria to distinguish between centrals and satellites, while others have used group galaxy catalogues to do this. However, the impact of systematic errors on these results has not been quantified. In the 2-halo regime, conformity has not yet been detected, as a couple recent works showed convincingly that past detections were entirely caused by selection biases. The current state of affairs for both 1-halo and 2-halo galactic conformity is still inconclusive and it is thus important to investigate this further.

The term "galactic conformity" was first coined by Weinmann et al. (2006, hereafter W06) after finding a correlation between the colours and star formation rates (SFR) of central and satellite galaxies in common Yang et al. (2005) galaxy groups of similar mass at low-redshifts, i.e. $z < 0.05$, in SDSS (York, 2000) DR2 (Abazajian et al., 2004). Specifically, W06 found that in galaxy groups of similar mass, quenched satellite galaxies occur more frequently around quenched central galaxies than around star-forming central galaxies. Controlling for halo mass is of critical importance in conformity studies because the SFRs of both centrals and satellites decrease with halo mass, which can naturally induce a conformity-like signal. W06 attempted to control for halo mass by adopting bins in total group luminosity.

Several subsequent studies also found correlations in SFR and other properties between central and satellite galaxies, using different methods for distinguishing between centrals and satellites and different ways of controlling for mass. Ann et al. (2008) used isola-

tion criteria, rather than a group catalogue, to identify centrals and satellites in SDSS DR5 (AdelmanâĂŘMcCarthy et al., 2007). They found that early-type satellite galaxies tend to reside in the vicinity of early-type central galaxies, and argue that this conformity in morphology is likely due to hydrodynamic and radiative influence of central galaxies on satellite galaxies, in addition to tidal effects. They attempted to control for mass by restricting their analysis to central galaxies in a limited range of luminosity. Wang & White (2012) also used isolation criteria to study correlations between isolated bright primary galaxies in SDSS DR7 (Abazajian et al., 2009) and nearby secondary galaxies (i.e., satellites) in SDSS DR8 (Aihara et al., 2011). They found that the colour distribution of satellites is redder for red primaries than for blue primaries of the same stellar mass. This is a similar 1-halo conformity trend in colour as found by W06, except that Wang & White (2012) control for central galaxy stellar mass. In addition, Wang & White (2012) compared their results to the Guo et al. (2011) semi-analytic model (SAM). They found that the SAM predicted a similar conformity signal as the SDSS. However, when they re-analysed the SAM controlling for halo mass instead of central galaxy stellar mass, they found a substantially reduced signal. This implies that a large portion of their observed SDSS conformity signal could be due to halo mass differences between red and blue galaxies at fixed stellar mass. Phillips et al. (2014a,b) also used isolation criteria to study the SFR of $\sim 0.1L^*$ satellites around isolated $\sim L^*$ central galaxies in the local Universe using SDSS DR7. They found that satellites of quiescent primaries are more than twice as likely to be quenched than similar mass satellites of star forming primaries. Unlike other studies, these authors control for the stellar mass of satellites, rather than centrals. This might seem risky since satellite galaxy stellar mass is not expected to correlate strongly with halo mass. However, the authors compare the velocity distributions of satellites around star forming and quiescent primaries and they conclude that the difference in halo mass between the two samples is not large enough to account for the conformity signal they observe. Finally, Knobel et al. (2015) used the Yang et al. (2012) group catalogue in SDSS DR7 to study the degree of central-satellite confor-

mity, controlling for several combinations of properties, including total group stellar mass. They confirmed that satellites of quenched central galaxies are more likely to be quenched than those of active central galaxies.

Ross & Brunner (2009) found evidence of 1-halo conformity using a completely different approach. They used a Halo Occupation Distribution (HOD; Berlind & Weinberg, 2002) model to fit the clustering of photometric samples in SDSS DR5. They found that they could only simultaneously match the clustering of all, early- and late-type galaxies with a model that segregates early- and late-type galaxies into separate haloes as much as possible. This is similar in spirit to the previous work of Zehavi et al. (2005) who modelled the cross-correlation function between red and blue galaxies in SDSS DR2, though that study concluded that red and blue galaxies are well mixed within their haloes. Zehavi et al. (2010) revisited this issue using SDSS DR7 and found significant evidence of colour segregation into different haloes, but the degree of segregation was much less than that found by Ross & Brunner (2009).

There have also been studies that have claimed a detection of 1-halo conformity at higher redshift. Hartley et al. (2015) used isolation criteria in the UKIDSS (Lawrence et al., 2007) Ultra Deep Survey DR8, to explore the redshift evolution of the correlation between the SFR of central galaxies and satellite galaxies at intermediate to high redshifts ($0.4 < z < 1.9$). They confirmed that passive satellites tend to be preferentially located around passive central galaxies, and showed that the trend persists to at least $z \sim 2$ without any significant evolution. Kawinwanichakij et al. (2016) carried out a similar analysis and identified central and satellite galaxies in the range of $0.3 < z < 2.5$ by combining imaging from three deep near-infrared-selected surveys ZFOURGE/CANDELS, UDS, and UltraVISTA (McCracken et al., 2012) and deriving accurate photometric redshifts. They found that, at similar central stellar mass, satellites of quiescent central galaxies are more likely to be quenched compared to satellites of star-forming central galaxies. This conformity signal is only significant at $0.6 < z < 1.6$, and becomes weaker at both lower and higher

redshifts. Kawinwanichakij et al. (2016) argue that their detection is unlikely to arise from any difference in halo mass between star-forming and quiescent centrals. To check this they allowed for star-forming centrals to have a stellar mass of up to 0.2 dex higher than quiescent centrals and found that, though the conformity signal weakens, it does not vanish. Most recently, Berti et al. (2017) used isolation criteria in the spectroscopic PRIMUS Survey (Coil et al., 2011; Cool et al., 2013) to look for conformity at $0.2 < z < 1.0$. After matching the stellar mass and redshift distributions of star-forming and quenched centrals, Berti et al. (2017) claimed a 3σ detection of a $\sim 5\%$ excess of star-forming neighbours around star-forming central galaxies on scales of 0-1 Mpc. This conformity signal is substantially weaker than the W06 signal observed in SDSS at $z \lesssim 0.05$. Berti et al. (2017) also reported on a 2-halo conformity detection, albeit with weaker statistical significance.

In the 2-halo regime, Kauffmann et al. (2013, hereafter K13) claimed a detection of conformity using a volume-limited sample of galaxies with redshifts $z < 0.03$ from the SDSS DR7. They adopted isolation criteria to identify central galaxies and studied the median specific SFR of neighbouring galaxies as a function of different properties of the centrals. K13 found that the SFR of neighbours correlates with that of centrals, even up to 4 Mpc, a distance that is well outside the virial radius of the primary galaxy's halo. This 2-halo conformity signal is present for low stellar mass galaxies, with massive galaxies only exhibiting a 1-halo conformity signal. The K13 result was intriguing and motivated a number of theoretical studies to explain it. However, a pair of recent studies have shown convincingly that the result in K13 is entirely due to selection bias. Tinker et al. (2018, hereafter T17) reproduced the result of K13 and then used a group finding algorithm and a mock catalogue to show that the majority of the 2-halo conformity signal comes from a subset of satellite galaxies that were mis-identified as primaries in the galaxy sample. After removing this small fraction of satellite galaxies, T17 detect no statistically significant 2-halo conformity in galaxy star formation rates. Sin et al. (2017, hereafter S17) carried out a similar analysis, and argued that the isolation criteria in K13 could potentially in-

clude low-mass central galaxies in the vicinity of massive systems, and that the large-scale conformity signal is likely a short-range effect coming from massive haloes. In addition to the misclassification of satellite galaxies as central galaxies in the isolation criteria, S17 argued that a weighting in favour of central galaxies in very high-density regions, and the use of medians to characterize the bimodal distribution of sSFRcould potentially amplify the large-scale conformity signal seen in K13.

Zu & Mandelbaum (2018) came to a similar conclusion about the lack of 2-halo conformity by finding that conformity measurements in SDSS DR7 are consistent with predictions from the `iHOD` *halo-quenching* model (Zu & Mandelbaum, 2015, 2016), in which galaxy colours depend *only* on halo mass. This suggests that all conformity signals are simply due to the combination of the environmental dependence of the halo mass function combined with the strong correlation between galaxy colours and halo mass. In other words, no galaxy assembly bias or other environmental quenching mechanisms are required to explain 2-halo conformity signals.

On the theoretical side, there have been several studies looking at both 1-halo and 2-halo conformity. Paranjape et al. (2015) called into question the conformity signal measured by K13 at a projected distance of $\lesssim 4$ Mpc by generating mock catalogues with varying levels of built-in galactic conformity, and comparing these to SDSS galaxies in the Yang et al. (2007) galaxy group catalogue. They argued against the K13 result being evidence of galaxy and halo assembly bias. Paranjape et al. (2015) also argued that only at very large separations, ($\gtrsim 8$ Mpc), does 2-halo conformity, driven by the assembly bias of small haloes, manifest distinctly. They suggest that the observed conformity at $\lesssim 4$ Mpc is simply due to central galaxies of similar stellar mass residing in haloes of different masses. Other papers have tried to explore the origin of galactic conformity. Hearin et al. (2016) studied the correlation between the mass accretion rates of nearby haloes as a potential physical origin for 2-halo galactic conformity. They found that pairs of host haloes have correlated assembly histories, despite being separated from each other by distances greater than thirty

virial radii at the present day. They presented halo accretion conformity as a plausible mechanism driving 2-halo conformity in SFR. Moreover, they argued that galactic conformity is related to large-scale tidal fields, and predicted that 2-halo conformity should generically weaken at higher redshift and vanish to undetectable levels by $z \sim 1$. In this context, the 2-halo galactic conformity signal in Berti et al. (2017) is consistent with the Hearin et al. (2016) prediction and Berti et al. (2017) state that their detection of galactic conformity is thus likely indicative of assembly bias and arises from large-scale tidal fields. Additionally, Bray et al. (2015) investigated the role of assembly bias in producing galactic conformity in the Illustris (Vogelsberger et al., 2014) simulation, and argued to have found 2-halo conformity in the red fraction of galaxies. They found that, at fixed stellar mass, the red fraction of galaxies around redder neighbour galaxies is higher than it is around bluer galaxies and this effect persists out to distances of 10 Mpc. They concluded by saying that the predicted amplitude of the conformity signal depends on the projection effects, stacking techniques, and the criteria used for selecting central galaxies. Lacerna et al. (2018) used three semi- analytic models to study the correlations between sSFRof central galaxies and neighbour galaxies out to scales of several Mpc. They predicted a strong 1-halo galactic conformity signal when the selection of primary galaxies was based on an isolation criterion in real space, and claimed a significant 2-halo conformity signal as far as ~ 5 Mpc. However, the overall signal of galactic conformity decreased when satellites that had been misclassified as central galaxies were removed in the selection of primary galaxies. The authors concluded that the SAMs used in the analysis do not show galactic conformity for the 2-halo regime.

Galactic conformity remains a debated topic, and it is unclear if all previous detection claims are valid. The work of Campbell et al. (2015) exposed the dangers of using group catalogues to study 1-halo conformity. They showed that group finders do a good job at recovering galactic conformity, but they also tend to introduce weak conformity when none is present in the data. This calls into question previous claims, such as the one by W06. More

recently, T17 and S17 challenged the measurement of 2-halo conformity made by K13 by showing that their isolation criteria were not sufficiently robust. These conflicting results open the door for improvements in the measurements of 1-halo and 2-halo conformity. In this paper we investigate both regimes using a galaxy group catalogue from the SDSS DR7. Our analysis contains three main improvements over previous works. First, we study three observed properties of galaxies: $(g-r)$ colour, sSFR, and Sérsic index. Second, we use a new statistic, the marked correlation function, $\mathcal{M}(r_p)$, in addition to the previously used quenched fractions. $\mathcal{M}(r_p)$ is ideally suited for conformity studies and is a more sensitive probe of weak conformity signals. Third, we use a suite of 100 mock galaxy catalogues to quantify the statistical significance of our results. The mock catalogues do not have any built-in conformity, but they are affected by the same systematic errors as the SDSS data. By comparing our SDSS measurements to the distribution of mock measurements, we can quantify the probability that whatever signal we detect could have arisen from a model with no conformity.

This paper is organized as follows. In §3.3, we describe the observational and simulated data used in this work, as well as the main analysis methods. In §3.4, we present a detailed examination of galactic conformity, distinguishing between 1-halo (§3.4.1) and 2-halo (§3.4.2). We summarize our results and discuss their implications in §3.5. The Python codes and the catalogues used in this project will be made publicly available on Github [1] upon publication of this paper.

3.3 Data and Methods

In this section, we present the datasets used throughout this analysis, and introduce the main statistical methods that we use to search for conformity signals. In §3.3.1 we briefly describe the SDSS galaxy sample that we use, along with the parameters that are included in this catalogue. In §3.3.2 we summarize how we identify galaxy groups and estimate their masses. In §3.3.3 we describe in detail the mock catalogues that we use throughout

[1] https://github.com/vcalderon2009/SDSS_Conformity_Analysis

the paper. Finally, we describe the two main statistical methods for probing conformity in §3.3.4 and §3.3.5.

3.3.1 SDSS Galaxy Sample

For this analysis, we use data from the Sloan Digital Sky Survey. SDSS collected its data with a dedicated 2.5-meter telescope (Gunn et al., 2006), camera (Gunn et al., 1998), filters (Doi et al., 2010), and spectrograph (Smee et al., 2013). We construct our galaxy sample from the `large-scale structure` sample of the NYU Value-Added Galaxy Catalogue (NYU-VAGC; Blanton et al., 2005), based on the spectroscopic sample in Data Release 7 (SDSS DR7; Abazajian et al., 2009). The main spectroscopic galaxy sample is approximately complete down to an apparent r-band Petrosian magnitude limit of $m_r = -17.77$. However, we have cut our sample back to $m_r = -17.6$ so it is complete down to that magnitude limit across the sky. Galaxy absolute magnitudes are k-corrected (Blanton et al., 2003) to rest-frame magnitudes at redshift $z = 0.1$.

We construct a volume-limited galaxy sample that contains all galaxies brighter than $M_r = -19$, and we refer to this sample as `Mr19-SDSS`. The redshift limits of the sample are $z_{\rm min} = 0.02$ and $z_{\rm max} = 0.067$ and it contains 90,893 galaxies with a number density of $n_{\rm gal} = 0.01503 h^3 {\rm Mpc}^{-3}$. The sample includes the right ascension, declination, redshift, Sérsic index, and $(g-r)$ colour for each galaxy.

To each galaxy, we assign a star formation rate (SFR) using the MPA Value-Added Catalogue DR7 [2]. This catalogue includes, among many other parameters, stellar masses based on fits to the photometry using Kauffmann et al. (2003) and Salim et al. (2007), and star formation rates based on Brinchmann et al. (2004). We cross-match the galaxies of the NYU-VAGC catalogue to those in the MPA-JHU catalogue using their MJD, plate ID, and fibre ID. A total of 5.85% of galaxies in the sample did not have corresponding values of SFR and we remove them from the sample. This leaves a sample of 85,578 galaxies. For each of these galaxies, we divide its SFR by its stellar mass to get a specific star formation

[2] http://www.mpa-garching.mpg.de/SDSS/DR7

rate sSFR.

sSFRand $(g-r)$ colour are highly correlated galaxy properties with the main difference coming from dust attenuation that moves some intrinsically star forming galaxies onto the red sequence. However, we have chosen to use both galaxy properties in this analysis in order to facilitate the comparison of our work to previous claims of conformity detection.

3.3.2 Group Finding Algorithm and Mass Assignment

We identify galaxy groups using the Berlind et al. (2006) group-finding algorithm. This is a Friends-of-Friends (FoF; Huchra & Geller, 1982) algorithm that links galaxies recursively to other galaxies that are within a cylindrical linking volume. The projected and line-of-sight linking lengths are $b_{\perp} = 0.14$ and $b_{\parallel} = 0.75$ in units of the mean inter-galaxy separation. This choice of linking lengths was optimized by Berlind et al. (2006) to identify galaxy systems that live within the same dark matter halo. In each group, we define the most luminous galaxy (in the r-band) to be the 'central' galaxy. The rest of the galaxies are defined as 'satellite' galaxies.

We estimate the total masses of the groups via *abundance matching*, using total group luminosity as a proxy for mass. Specifically, we assume that the total group r-band luminosity $L_{\rm group}$ increases monotonically with halo mass $M_{\rm h}$, and we assign masses to groups by matching the cumulative space densities of groups and haloes:

$$n_{\rm group}(> L_{\rm group}) = n_{\rm halo}(> M_{\rm h}). \tag{3.1}$$

To calculate the space densities of haloes, we adopt the Warren et al. (2006) halo mass function assuming a cosmological model with $\Omega_m = 1 - \Omega_{\Lambda} = 0.25$, $\Omega_b = 0.04$, $h \equiv H_0/(100\ {\rm km\ s^{-1}\ Mpc^{-1}}) = 0.7$, $\sigma_8 = 0.8$, and $n_s = 1.0$. We refer to these abundance matched masses as *group masses*, $M_{\rm group}$.

3.3.3 Mock Galaxy Catalogues

To quantify the statistical significance of any conformity signal that we measure using our SDSS groups, it is necessary to compare to a null model (i.e., no intrinsic conformity) that incorporates the same systematic errors as our measurements and also includes robust error distributions. For this purpose, we construct a suite of 100 realistic mock catalogues that are based on the *Large Suite of Dark Matter Simulations* (LasDamas) project[3] (McBride et al., 2009).

We start with a set of 50 cosmological N-body simulations that trace the evolution of dark matter in the Universe and have sufficient volume and mass resolution to properly model the Mr19-SDSS sample. These simulations assumed the same cosmological model described at the end of §3.3.2. Dark matter haloes were identified with a FoF algorithm (Davis et al., 1985) using a linking length of 0.2 times the mean inter-particle separation. We used an HOD model to populate the DM haloes with central and satellite galaxies, whose numbers as a function of halo mass were chosen to reproduce the number density, $n_{\rm gal}$, and the projected 2-point correlation function, $w_p\left(r_p\right)$, of the Mr19-SDSS sample. Each central galaxy was placed at the minimum of the halo gravitational potential and was assigned the mean velocity of the halo. Satellite galaxies were assigned the positions and velocities of randomly chosen dark matter particles within the halo. Within each simulation box, we applied redshift space distortions and then we carved out two independent volumes that precisely mimic the geometry of our Mr19-SDSS sample. This procedure yields 100 independent mock catalogues from the 50 simulation boxes.

To assign a luminosity to each mock galaxy, we adopt the formalism of the *conditional luminosity function* (CLF; Yang et al., 2003; Van Den Bosch et al., 2003) that specifies functional forms for the luminosity distributions of central and satellite galaxies as a function of halo mass. Specifically, we use the Cacciato et al. (2009) version of the CLF, but modified slightly to match our adopted cosmological model (Van den Bosch, private com-

[3] http://lss.phy.vanderbilt.edu/lasdamas/

munication). We then abundance match the luminosities obtained from the CLF to the r-band absolute magnitudes in Mr19-SDSS. As a result, our mock catalogues have the same exact luminosity function as the SDSS data.

We assign specific star formation rates, $(g-r)$ colours and Sérsic indices to mock galaxies by first adopting the formalism presented in Zu & Mandelbaum (2016, hereafter Z16), and then sampling from the original distributions of sSFR, $(g-r)$ colour and Sérsic indices of Mr19-SDSS. Specifically, we adopt the *'halo'* quenching model from Z16, which assumes that halo mass is the sole driver of galaxy quenching. According to that model, the red/quenched fraction of central and satellite galaxies is given by

$$f_{\rm cen}^{\rm red}(M_{\rm h}) = 1 - \exp\left[-(M_{\rm h}/M_{\rm h}^{\rm qc})^{\mu^{\rm c}}\right] \tag{3.2}$$

and

$$f_{\rm sat}^{\rm red}(M_{\rm h}) = 1 - \exp\left[-(M_{\rm h}/M_{\rm h}^{\rm qs})^{\mu^{\rm s}}\right], \tag{3.3}$$

where $M_{\rm h}^{\rm qc}$, $M_{\rm h}^{\rm qs}$, $\mu^{\rm c}$, and $\mu^{\rm s}$ are parameters of the model that Z16 fit to the observed clustering and galaxy-galaxy lensing measurements of red and blue galaxies in the SDSS. We assign each of our mock galaxies a probability of being quenched from equations (3.2) and (3.3) and we randomly designate it as 'active' or 'passive' consistent with that probability (e.g., if $f_{\rm sat}^{\rm red} = 0.8$ for a particular mock satellite galaxy, we give it an 80% chance of being labelled 'passive'). To assign realistic values of sSFR, $(g-r)$ colour, and Sérsic index to mock galaxies, we divide the observed distributions of these properties of Mr19-SDSS into 'active' and 'passive' distributions by making cuts at $\log_{10}{\rm sSFR} = -11$, $(g-r)_{cut} = 0.75$ and $n_{cut} = 3$ for sSFR, $(g-r)$ colour, and Sérsic index, respectively. For example, to assign sSFRvalues to mock galaxies, we do the following. For each mock galaxy, we randomly draw a sSFRvalue from the active or passive distribution, depending on the designation that the mock galaxy has received. Moreover, we do this in a way that preserves the joint

sSFR-luminosity distribution. For example, if a mock galaxy has been labelled 'active', we randomly select a real active galaxy from Mr19-SDSS that has a similar luminosity as the mock galaxy, and we assign its sSFR to the mock galaxy. As a result of this procedure, the final joint sSFR-luminosity distribution of mock galaxies closely resembles the one for Mr19-SDSS. However, the model contains *no intrinsic* 1-halo or 2-halo conformity because the galaxy sSFR values *only* depend on halo mass. We apply this same procedure to assign $(g-r)$ colours and Sérsic indices to each mock galaxy in order to preserve the joint distributions of these galaxy properties with luminosity as seen the Mr19-SDSS sample.

After constructing our 100 mock catalogues, we run the group-finding algorithm on each one to produce a corresponding group catalogue. We then label each mock galaxy as 'central' or 'satellite' and estimate total group masses by following the same methodology as in §3.3.2. The end result is a set of mock catalogues that do not have built-in galactic conformity in sSFR, $(g-r)$ colour, or Sérsic index, but suffer from the same kinds of systematics as the SDSS data, i.e. group-finding errors that lead to central-satellite misclassification and errors in the estimated group masses.

3.3.4 Quenched Fraction Difference $\Delta f_{\rm q}$

Previous studies of conformity have mostly focused on measuring the fractions of quenched neighbour galaxies around active and passive primary galaxies, either as a function of group mass or as a function of distance (e.g., W06, K13). Following these studies, we also consider quenched fractions of neighbour galaxies, focusing on the *difference* between the fraction for passive primaries and that for active primaries. Moreover, we use three different galaxy properties to search for conformity: $(g-r)$ colour, sSFR, and Sérsic index. The cuts we use to designate galaxies as red, passive or early type are $(g-r) > 0.75$, $\log\ \mathrm{sSFR} < -11$, and $n > 3$, respectively. These are the same cuts we discuss in §3.3.3.

To explain this better, let us consider the specific case of probing 1-halo conformity in galaxy colour. We measure the fraction of red satellite galaxies around red centrals,

$P(\mathrm{sat} = \mathrm{red} \mid \mathrm{cen} = \mathrm{red})$, and the fraction of red satellite galaxies around blue centrals, $P(\mathrm{sat} = \mathrm{red} \mid \mathrm{cen} = \mathrm{blue})$. We then determine the *difference* between these two fractions, which we refer to as Δf_{red}. A conformity signal is then the case of $|\Delta f_{\mathrm{red}}| > 0$. We define similar quantities using sSFRand morphology. The three quenched fraction differences that we measure are thus

$$\Delta f_{\mathrm{red}} = P(\mathrm{sat} = \mathrm{red} \mid \mathrm{cen} = \mathrm{red}) - P(\mathrm{sat} = \mathrm{red} \mid \mathrm{cen} = \mathrm{blue}) \tag{3.4}$$

$$\Delta f_{\mathrm{passive}} = P(\mathrm{sat} = \mathrm{passive} \mid \mathrm{cen} = \mathrm{passive}) - P(\mathrm{sat} = \mathrm{passive} \mid \mathrm{cen} = \mathrm{active}) \tag{3.5}$$

$$\Delta f_{\mathrm{early}} = P(\mathrm{sat} = \mathrm{early} \mid \mathrm{cen} = \mathrm{early}) - P(\mathrm{sat} = \mathrm{early} \mid \mathrm{cen} = \mathrm{late}) \tag{3.6}$$

Finally, as a way to control for halo mass, we measure these fractions in bins of M_{group}. In the mock catalogues, we follow the same procedure to calculate Δf_{red}, $\Delta f_{\mathrm{passive}}$ and $\Delta f_{\mathrm{early}}$. For convenience, we refer to all three of these quantities as "quenched" fraction differences, Δf_{q}, recognizing that Sérsic index is a measure of galaxy morphology and not star formation activity.

In the case of 2-halo conformity, we use the same formalism of equations (3.4)−(3.6), with the difference that we only consider pairs of central galaxies with line-of-sight separations of $\pi_{\mathrm{max}} < 20\,h^{-1}$Mpc and we calculate the fractions in bins of projected separation within each M_{group} bin. For each central-central galaxy, we designate one to be the primary and the other to be the secondary and we calculate the difference between the quenched fractions of secondary galaxies that are associated with quenched primaries and those that are associated with active primaries. Each galaxy pair contributes twice to the calculation of Δf_{q} because both galaxies get a turn at being considered the primary galaxy. For example, suppose there is a pair of galaxies, one red and one blue, that are

both centrals in groups of similar mass. When the blue galaxy is the primary, the pair will contribute positively to the fraction $P(\text{secondary = red} \mid \text{primary = blue})$. On the other hand, when the red galaxy is the primary, the pair will contribute negatively to the fraction $P(\text{secondary = red} \mid \text{primary = red})$. Therefore, red-red and blue-blue pairs act to increase $\Delta f_{\rm red}$, while red-blue pairs do the opposite. $\Delta f_{\rm red}$ essentially measures the excess number of similar pairs (i.e., red-red or blue-blue) over what one would expect if the population of red and blue galaxies were randomly mixed. The value of $\Delta f_{\rm q}$ ranges from +1 where all pairs are similar, to −1 where pairs are as different as possible.

3.3.5 Marked Correlation Function $\mathcal{M}(r_p)$

Galactic conformity is essentially a correlation between the properties of galaxies across distance. In the case of 1-halo conformity, we care about the correlation between properties of central galaxies and satellites within the same halo. In the case of 2-halo conformity, we look for a correlation between properties of central galaxies in separate haloes. The "*marked correlation function*" is an ideal tool for quantifying correlations across scale and it has been used successfully to probe the environmental dependence of galaxy properties (Beisbart & Kerscher, 2000; Sheth et al., 2005; Skibba et al., 2006; Martinez et al., 2010).

The marked statistic $\mathcal{M}(r_p)$ provides a measure of the clustering of galaxy properties, or "marks". In this paper, we analyse the marked statistics for $(g-r)$ colour, specific star formation rate (sSFR), and Sérsic index n in bins of group mass $M_{\rm group}$. We adopt the formalism presented in Sheth et al. (2005) and Skibba et al. (2006) for defining $\mathcal{M}(r_p)$

$$\mathcal{M}(r_p) = \frac{1+W(r_p)}{1+\xi(r_p)} \equiv \frac{WW}{DD} \tag{3.7}$$

where $\xi(r_p)$ is the usual two-point correlation function with pairs summed in bins of projected separation r_p, and $W(r_p)$ is the same except that galaxy pairs are weighted by the product of their marks. The estimator used in equation (3.7) can also be written as WW/DD, where DD is the raw number of galaxy pairs separated by r_p and WW is the weighted num-

ber of pairs. Defining the statistic as a ratio in this way is advantageous because, unlike the correlation function, it can be estimated without explicitly constructing a random catalogue, but, like the correlation function, it accounts for edge effects so one does not need to worry about the geometry of the survey (Sheth et al., 2005).

The marked statistic is essentially a measurement of the correlation coefficient between the marks of galaxies, as a function of projected separation. Though it is similar in spirit and goal to the quenched fraction difference statistic described in the previous section, the marked correlation function contains more information because it uses the full values of galaxy properties (e.g., colour) instead of just a binary classification (e.g., red or blue). There is thus reason to hope that $\mathcal{M}(r_p)$ is a more sensitive probe of galactic conformity than the usual quenched fractions.

3.4 Galactic Conformity Results

In this section, we present the results of the galactic conformity analysis of SDSS DR7. In §3.4.1, we investigate 1-halo conformity by looking at both quenched fraction differences, $\Delta f_{\rm q}$, as a function of group mass (§3.4.1.1), and the mark correlation function, $\mathcal{M}(r_p)$, as a function of projected separation (§3.4.1.2). In §3.4.2, we investigate 2-halo conformity, also using $\Delta f_{\rm q}$ (§3.4.2.1) and $\mathcal{M}(r_p)$ (§3.4.2.2).

3.4.1 1-halo Conformity

3.4.1.1 Quenched Fractions and 1-halo Conformity

We first study 1-halo conformity using the quenched fraction difference statistic defined in § 3.3.4 as a function of group mass. This is very similar to the original method that W06 used to detect 1-halo conformity. Specifically, we create six $M_{\rm group}$ bins of width 0.4 dex in the range $\log M_{\rm group}$: 11.6–14.0. Within each bin of group mass, we make a list of all satellite galaxies that are in groups with a red central and a second list of all satellites in groups with a blue central. We then calculate the red fraction of satellites in each list and take the difference $\Delta f_{\rm red}$. We repeat this process using sSFRand Sérsic index to calculate

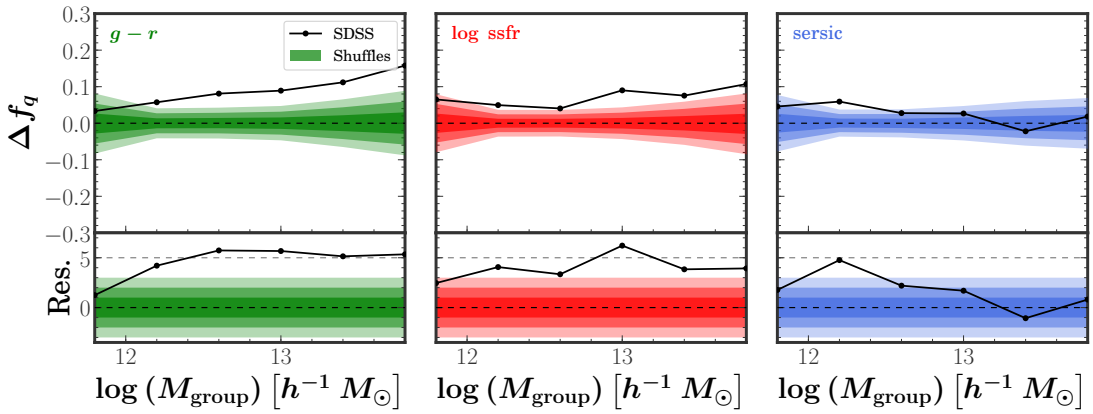

Figure 3.1: Difference of fractions, Δf, of red (left), passive (centre), and early-type (right) satellites as function of estimated group mass, $M_{\rm group}$, where the difference is measured between groups with red and blue, passive and active, early-type and late-type central galaxies, as measured in the Mr19-SDSS sample. *Top panels*: The solid black lines correspond to the Δf of each galaxy property. The shaded contours show the 1σ, 2σ, and 3σ ranges of Δf calculated from many realizations in which the values of the galaxy properties are randomly shuffled, thus erasing any trace of 1-halo galactic conformity. *Bottom panels*: Normalised residuals of Δf of each galaxy property with respect to the shuffled realizations. The solid black lines show the difference between Δf and the mean of the shuffles, divided by the standard deviation of Δf for the shuffles. The shaded contours show the 1σ, 2σ, and 3σ ranges of the shuffled scenario in this normalised space.

$\Delta f_{\rm passive}$ and $\Delta f_{\rm early}$. When using these quenched fraction differences, a conformity signal corresponds to values that are not zero, i.e., $\left|\Delta f_{\rm q}\right| > 0$.

To determine the statistical significance of any conformity signal, we use a random shuffling method to eliminate any intrinsic conformity or correlation in the sample at the group level. Specifically, we shuffle the properties (colour, sSFR, Sérsic index) of all central and satellite galaxies within each group mass bin. Each central galaxy swaps properties with a randomly selected central galaxy from a different group of similar mass, and each satellite swaps properties with a randomly selected satellite from a group of similar mass. This procedure preserves the distributions of central and satellite properties as a function of group mass, but it explicitly erases any correlation between the properties of centrals and their satellites within any single group. The shuffling thus completely erases any 1-halo conformity signal that may exist in the data. We repeat this shuffling process a total of 1000

times (using different random seeds) and we re-measure the quenched fraction differences each time. The resulting distribution of $\Delta f_{\rm q,shuffle}$ values thus allows us to quantify the probability that any measured conformity signal could be a statistical fluke. We find that the distribution of shuffle values is consistent with being Gaussian and so we use the standard deviation of the shuffled values to calculate the 1σ, 2σ, and 3σ ranges of the distribution of $\Delta f_{\rm q,shuffle}$. We adopt the 3σ level as our detection threshold.

For each measurement of $\Delta f_{\rm q}$ on the un-shuffled data, we calculate the residual with respect to the shuffled data as

$$\mathrm{Res} = \frac{\Delta f_{\rm q} - \overline{\Delta f_{\rm q,shuffle}}}{\sigma_{\rm q,shuffle}} \tag{3.8}$$

where $\overline{\Delta f_{\rm q,shuffle}}$ is the mean of the 1000 shuffles and $\sigma_{\rm q,shuffle}$ is their standard deviation.

Figure 3.1 presents our main results of probing 1-halo conformity using quenched fraction differences. The black lines in the top three panels show the $\Delta f_{\rm q}$ for $(g-r)$ colour, sSFR, and Sérsic index, as measured in the `Mr19-SDSS` sample. The shaded contours show the 1σ, 2σ, and 3σ ranges of $\Delta f_{\rm shuffle}$ for the *shuffle* cases of each galaxy property. The bottom panels show the residuals of each galaxy property with respect to the shuffles, as defined in equation (3.8). Figure 3.1 shows prominent conformity signals in the quenched fraction differences for $(g-r)$ colour and sSFRat large group masses, while for morphology the signal only appears at low group mass. Specifically, the conformity signal in colour rises with mass from $\Delta f_{\rm red}$=0.06 to 0.14 and is at the $4-6\sigma$ level of statistical significance for masses above $10^{12}\,h^{-1}\mathrm{M}_{\odot}$. In the case of sSFR, the signal is lower, rising from $\Delta f_{\rm passive}$=0.05 to 0.1 and is at the $3-4\sigma$ level, except for a 6σ peak at $\sim 10^{13}\,h^{-1}\mathrm{M}_{\odot}$. Finally, in the case of Sérsic index, the signal is only significant for groups of mass $\sim 10^{12.2}\,h^{-1}\mathrm{M}_{\odot}$, where $\Delta f_{\rm early}$=0.07 and has a statistical significance of $5-6\sigma$. These results are in agreement with the results shown in W06, who also find a significant difference in the red fraction at high masses and a slightly weaker signal when using sSFR.

We have found statistically significant correlations between the properties of central and satellite galaxies within groups in Mr19-SDSS by comparing to the distribution of shuffled measurements, where any correlations between the properties of centrals and satellites have been erased. However, this does not mean that we have detected 1-halo galactic conformity, which is a correlation at fixed *halo* mass. Grouping errors that cause misidentification of centrals and satellites as well as errors in the estimated group mass $M_{\rm group}$ could be responsible for inducing a conformity-like signal (Campbell et al., 2015). To test this, we need to compare our measurements to mock catalogues that contain no built-in conformity, but are analysed in the same way as the SDSS data.

We apply the same procedure described above to the set of mock catalogues described in §3.3.3. The goal is to determine if the signal revealed in Figure 3.1 remains statistically significant when compared to the distribution of $\Delta f_{\rm q,mock}$ measurements from the 100 mock catalogues with no conformity built-in. We find that the distribution of 100 values of $\Delta f_{\rm q,mock}$ is approximately Gaussian and so we use their standard deviation to estimate the 1σ, 2σ, and 3σ ranges of the distribution. As we did previously for the shuffles, for each measurement of $\Delta f_{\rm q}$ on the SDSS data, we calculate the residual with respect to the mocks as

$$\mathrm{Res} = \frac{\Delta f_{\rm q} - \overline{\Delta f_{\rm q,mock}}}{\sigma_{\rm q,mock}} \tag{3.9}$$

where $\overline{\Delta f_{\rm q,mock}}$ is the mean of the 100 mocks and $\sigma_{\rm q,mock}$ is their standard deviation.

Figure 3.2 is analogous to Figure 3.1, except that the shaded contours now show the distribution of mocks rather than shuffles. The black lines in the top panels show the $\Delta f_{\rm q}$ for each galaxy property as measured in the Mr19-SDSS sample and are thus identical to the black lines in the three top panels of Figure 3.1. The shaded contours show the 1σ, 2σ, and 3σ ranges of $\Delta f_{\rm q,mock}$ values of the mock catalogues. The bottom panels show the residuals with respect to the mocks, as defined in equation (3.9). The prominent conformity

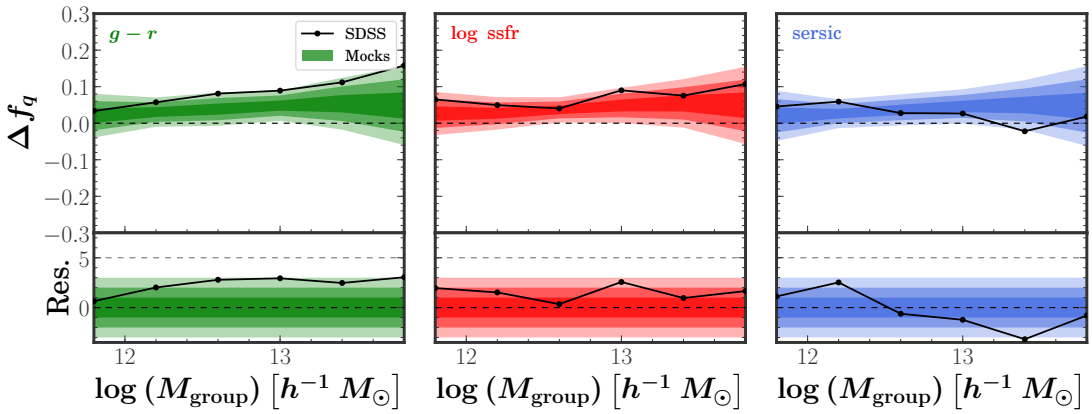

Figure 3.2: Similar to Figure 3.1, except that the $\Delta f_{\rm q}$ of Mr19-SDSS are compared to the distributions of measurements from mock catalogues rather than randomly shuffled data. *Top panels*: The solid black lines correspond to the $\Delta f_{\rm q}$ in Mr19-SDSS. The shaded contours show the 1σ, 2σ, and 3σ ranges of $\Delta f_{\rm q}$ calculated from 100 mock catalogues with no built-in conformity. *Bottom panel*: Normalised residuals of $\Delta f_{\rm q}$ with respect to the mock catalogues. The solid black lines show the difference between $\Delta f_{\rm q}$ for Mr19-SDSS and the mean of the mocks, divided by the standard deviation of the mocks. The shaded contours show the 1σ, 2σ, and 3σ ranges of the mocks in this normalised space.

signals that we found previously disappear when compared against the mock catalogues. This is because the whole shaded bands are no longer centred at $\Delta f_{\rm q} = 0$, but have shifted up significantly. In other words, the mock catalogues with no built-in conformity have an average quenched fraction difference of 0.02 to 0.05 for the three galaxy properties, depending on group mass. These *spurious* conformity signals must be due to grouping errors – either in misidentification of centrals and satellites, or in estimation of $M_{\rm group}$. We have examined the contributions to the induced signal from both of these factors by constructing versions of our mock catalogues that do not contain these errors. We find that most of the induced signal comes from errors in assigning $M_{\rm group}$, consistent with Campbell et al. (2015).

Figure 3.2 shows that the statistical significance of the 1-halo conformity signal in the quenched fractions of the Mr19-SDSS sample drops from $5-6\sigma$ when using the shuffles to $2.5-3\sigma$ when using the mocks. Consequently, we can no longer claim a significant detection of 1-halo galactic conformity. This result illustrates the importance of using mock

catalogues to compute the null model (i.e., no conformity case) in any conformity analysis. Moreover, it is necessary to use a large suite of mock catalogues to properly specify the distribution of the null model. A few of our 100 mock catalogues do not display spurious conformity signals and so if we had only used one mock that happened to lack any conformity signals, we would have come to the wrong conclusion about the significance of our conformity detection. Our result calls into question previous claims of 1-halo conformity detections, especially from papers that used similar group-based methods as ours, including the original detection by W06.

3.4.1.2 $\mathcal{M}(r_p)$ for 1-halo Conformity

We now move to the second statistic that we are using to probe galactic conformity, the "marked correlation function", $\mathcal{M}(r_p)$. Since the $\mathcal{M}(r_p)$ can be more sensitive than binary statistics, and can potentially uncover the scale dependence of any correlations (see discussion in §3.3.5), the $\mathcal{M}(r_p)$ is well-suited to exploring the correlations between central and satellite galaxies.

We evaluate $\mathcal{M}(r_p)$ for the three galaxy properties, i.e. $(g-r)$ colour, sSFR, and Sérsic index, in different bins of $M_{\rm group}$. Each galaxy pair is comprised of a central galaxy and a satellite galaxy of the same galaxy group, and the projected distance, r_p, is the distance between the two member galaxies. We then take the product of the 'marks' of the two galaxies and average this over all pairs in bins of r_p. The mark for each galaxy is just the value of its property (e.g., colour) normalised by the mean value over the whole population of similar galaxies. We do this in two ways. First, we normalise using the mean of all central or satellite galaxies in the same bin of $M_{\rm group}$. For example, the colour of each central (satellite) galaxy is divided by the mean colour of all central (satellite) galaxies that live in similar mass groups. $\mathcal{M}(r_p)$ then measures the correlation coefficient between the normalised colours of central and satellite galaxies. Since this measurement is done in r_p bins, it is sensitive to radial gradients in the properties of satellite galaxies within groups,

typically referred to as *segregation*. For example, if groups contain colour segregation in the sense that satellite galaxies in the central regions of groups tend to be redder than satellite galaxies in the outskirts of groups, then $\mathcal{M}(r_p)$ will be larger than unity in bins of small r_p and less than unity in bins of large r_p. Such a radial segregation effect will masquerade as a 1-halo conformity signal. To account for this, we do a second normalization where the properties of satellite galaxies are normalised by the mean values of all satellites that live in the same bin of both $M_{\rm group}$ and r_p. Measured in this way, $\mathcal{M}(r_p)$ is not sensitive to radial segregation and so values different from unity are direct indications of conformity.

To assess the statistical significance of a conformity signal while at the same time avoiding any biases due to grouping errors, we now only compare the results of the $\mathcal{M}(r_p)$ of `Mr19-SDSS` to those of the mock catalogues and not to those from the shuffling technique. By making this type of comparison, we avoid systematic errors that might masquerade as conformity signals. For example, it may be the case that galaxies that live in the outskirts of large groups are more likely to have been mis-assigned to their group than galaxies in the central regions of groups. These "satellites" may actually be centrals in much smaller neighbouring haloes that were incorrectly merged into the large groups. Since these low-mass centrals are likely to be bluer in colour than actual satellites of the large group, this error will masquerade as a radial colour gradient within groups. Such an effect may represent itself as an anti-correlation at large 1-halo scales. This type of systematic error will be present in the mocks as well and so we can account for the role of grouping errors by comparing our measurements to mock catalogues that contain no built-in conformity or segregation, but are analyzed in the same way as the SDSS data

Like we did for the quenched fraction differences in §3.4.1.1, we analyse the 100 mock catalogues in the same way as we analyse the `Mr19-SDSS` data. Specifically, we compute $\mathcal{M}(r_p)$ of each galaxy property, i.e. sSFR, $(g-r)$ colour, and Sérsic index, on the mocks after first normalizing each galaxy property the two different ways (in bins of $M_{\rm group}$ and in bins for both $M_{\rm group}$ and r_p). We use the standard deviation of $\mathcal{M}(r_p)$ values to estimate the

1σ, 2σ, and 3σ ranges of the distributions for each galaxy property. We then determine the statistical significance of the result by calculating the residuals of the SDSS measurements with respect to mocks as

$$\mathrm{Res} = \frac{\mathcal{M}(r_p) - \overline{\mathcal{M}(r_p)_{\mathrm{mock}}}}{\sigma_{\mathrm{mock}}}. \tag{3.10}$$

This is similar to the residuals in equation (3.9).

Figure 3.3 shows $\mathcal{M}(r_p)$ of $(g-r)$ colour (left), sSFR(centre), and Sérsic index (right), as a function of projected distance, r_p , with each row corresponding to a bin of M_{group}. In this figure, we only show bins with $M_{\mathrm{group}} > 10^{12.4}\,h^{-1}\mathrm{M}_{\odot}$ since these exhibited the largest signals in the quenched fraction difference statistic for colour and sSFR, as shown in Figure 3.1. In the top part of each panel, the black, solid line corresponds to the $\mathcal{M}(r_p)$ of SDSS galaxies, when properties are normalised within bins of r_p in order to remove the effects of radial segregation. For comparison, the grey dashed line corresponds to the case when the segregation effect is included, i.e., the contributions for the $\mathcal{M}(r_p)$ results are coming from both galactic conformity and the segregation effect. The shaded regions correspond to the 1σ, 2σ, and 3σ ranges of the distributions of $\mathcal{M}(r_p)$ values for mock catalogues. However, these results are analysed by normalizing properties within bins of r_p , so only the black, solid lines can be compared to the shaded regions. We do not show the results that correspond to the grey, dashed lines. The bottom part of each panel shows the residuals of each $\mathcal{M}(r_p)$ with respect to the mock catalogues, as defined in equation (3.10). In this case, the black solid lines and grey dashed lines are each computed using their corresponding set of mock results.

In Figure 3.3 the shaded regions for $(g-r)$ colour, sSFR, and Sérsic index are not centred at $\mathcal{M}(r_p)$ =1, indicating the effect of group errors. The strength of both radial segregation and conformity signals in the SDSS are weak when compared to mock catalogues containing neither effect. First we examine the case where we normalise galaxy properties

by their mean values in bins of $M_{\rm group}$, making $\mathcal{M}(r_p)$ sensitive to both conformity and segregation (dashed grey lines). We do detect significant radial segregation (dashed grey lines) for colour and sSFRat scales smaller than $0.2\,h^{-1}$Mpc in the case of massive groups, in the sense that satellite galaxies close to the centres of their groups tend to be more quenched (and thus more similar to their central galaxies) than satellite galaxies farther out. We do not find such correlations for the Sérsic index at those scales. Next we examine the case where radial segregation is removed (solid black lines). The 1-halo conformity signal hovers near the 3σ level for a wide range of small scales for colour and sSFR. However, the signal is not strong enough for us to claim a conformity detection. In summary, neither the quenched fractions nor the marked correlation function reveal any statistically significant 1-halo conformity signal after controlling for group errors for the cases of $(g-r)$ colour, sSFR, and Sérsic index.

3.4.2 2-halo Conformity

We next study 2-halo conformity, which is the correlation of properties for galaxies that live in separate haloes. As we discussed in §3.2, a detection of 2-halo conformity in sSFRwas claimed by K13 for low-mass central galaxies out to scales of 4 Mpc. This claim has been challenged by T17 and S17 who reproduced the result of K13 and showed that the conformity signal is mainly driven by contamination in the isolation criterion to select the sample of central galaxies. After removing a small fraction of satellite galaxies that were misclassified as centrals from the primary sample, only a weak conformity signal remains out to projected distances of 2 Mpc.

3.4.2.1 Quenched fractions for 2-halo Conformity

We first analyse 2-halo conformity using the quenched fraction difference statistic, which is similar to what was used by some of these previous works. We use a sample composed of only central galaxies as classified by our group-finder, which are the most luminous galaxies in the r-band within their respective groups. Then, using the same group

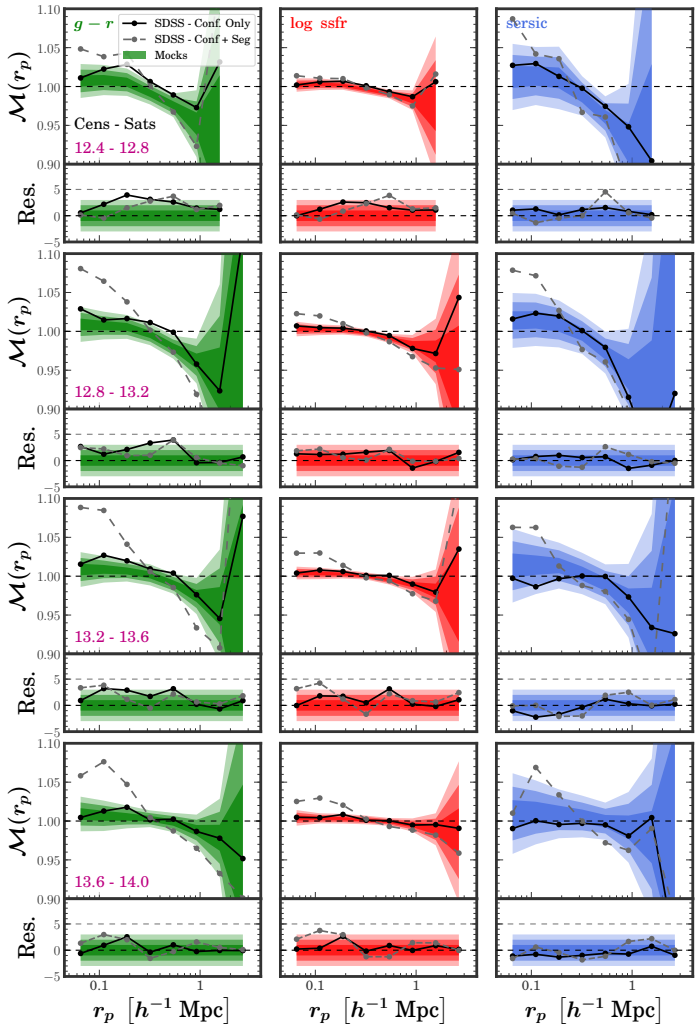

Figure 3.3: Marked correlation function, $\mathcal{M}(r_p)$, of $(g-r)$ colour (left), sSFR(centre), and Sérsic index (right), as a function of projected distance r_p for central-satellite galaxy pairs within the same galaxy groups in Mr19-SDSS and mock catalogues. Each row corresponds to a bin of group mass, $M_{\rm group}$, as listed in the left panels. *Top panels*: The solid black lines correspond to the case where the marks have been normalised to remove the effects of radial segregation, while the dashed grey lines include segregation. The shaded contours show the 1σ, 2σ, and 3σ ranges of $\mathcal{M}(r_p)$ calculated from 100 mock catalogues with no built-in conformity or radial segregation. These mock results can only be compared to the solid black lines. We do not show the mock results that correspond to the dashed grey lines. *Bottom panels*: Normalised residuals of $\mathcal{M}(r_p)$ with respect to the mock catalogues. The lines show the difference between $\mathcal{M}(r_p)$ of the SDSS and the mean of the mocks, divided by the standard deviation of $\mathcal{M}(r_p)$ for the mocks. The solid black and dashed grey lines correspond to the cases where effects of radial segregation are removed and included, respectively. The shaded contours show the 1σ, 2σ, and 3σ ranges of the mocks in this normalised space.

mass bins as before, we compute the quenched fraction difference statistic, as described in §3.3.4, in bins of projected separation r_p and only counting galaxy pairs within a line-of-sight separation of $\pi_{\rm max} = 20\,h^{-1}$Mpc. For example, to calculate $\Delta f_{\rm red}$ for the smallest group mass bin we consider, we first list all the central galaxies in groups with $\log M_{\rm group}$: 11.6–12.0, then find all pairs of these galaxies that have line-of-sight separations less than $\pi_{\rm max}$, and place them in logarithmic bins of r_p. Each radial bin now contains a set of central-central galaxy pairs where one of the galaxies is designated as “primary" and the other as “secondary" (each pair is counted twice so that both galaxies have a turn at being primary). We then make one list of pairs where the primary is red and another where it is blue. For each list we then calculate the fraction of pairs where the secondary is red (i.e., the “quenched fraction") and we take the difference between these two fractions. We repeat this procedure for all group mass bins and for sSFRand Sérsic index. As before, we assess the statistical significance of conformity signals by comparing with our set of 100 mock catalogues that contain no intrinsic conformity, but do contain the same types of systematic errors that affect the SDSS analysis.

Figure 3.4 presents our main results of probing 2-halo conformity using quenched fraction differences. The three columns show results for $(g-r)$ colour (left column), sSFR(middle column), and Sérsic index (right column), as measured in the `Mr19-SDSS` sample and mock catalogues. Each row corresponds to a bin of $M_{\rm group}$, as listed in the left column of plots. We focus on the four lowest-mass bins since K13 found 2-halo conformity signals at these masses. The black lines in the top portions of each panel show the $\Delta f_{\rm q}$ as a function of projected separation r_p , while the shaded contours show the 1σ, 2σ, and 3σ ranges of $\Delta f_{\rm q,mock}$ for the 100 mock catalogues of each galaxy property. The bottom panels show the residuals of each galaxy property with respect to the mock catalogues, as defined in equation (3.9).

Figure 3.4 does not reveal any 2-halo conformity signals for most group masses and scales for $(g-r)$ colour and sSFR. Sérsic index exhibits a prominent 2-halo conformity

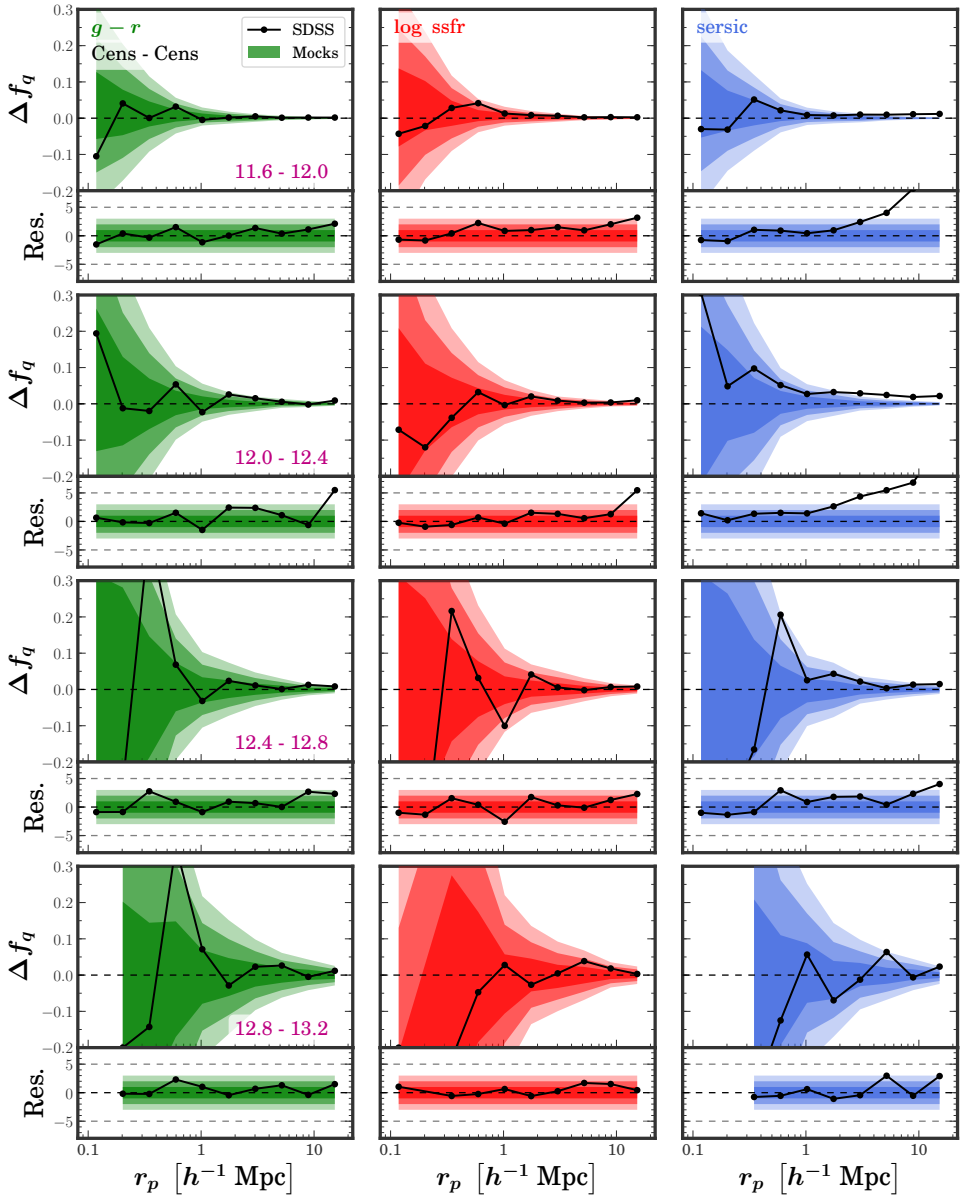

Figure 3.4: Difference of fractions, Δf, of red (left), passive (centre), and early-type (right) secondary central galaxies as a function of their projected distance, r_p , from primary central galaxies in groups of similar mass, where the difference is measured between primary galaxies that are red and blue, passive and active, early-type and late-type, respectively. Each row corresponds to a bin of group mass, $M_{\rm group}$, as listed in the left panels. *Top panels*: The solid black lines correspond to the Δf of each galaxy property in Mr19-SDSS. The shaded contours show the 1σ, 2σ, and 3σ ranges of Δf calculated from 100 mock catalogues with no built-in conformity. *Bottom panels*: Normalised residuals of Δf with respect to the mock catalogues. The solid black lines show the difference between Δf for Mr19-SDSS and the mean of the mocks, divided by the standard deviation of Δf for the mocks. The shaded contours show the 1σ, 2σ, and 3σ ranges of the mocks in this normalised space.

signal for the two lowest-mass bins, i.e., for group masses of $\log M_{\rm group}$ 11.6–12.4 and at scales of $r_p > 3\,h^{-1}$Mpc. This large Sérsic index signal is caused by the fact that SDSS central galaxies in groups of $\log M_{\rm group}$ 11.6–12.4 exhibit a small $\Delta f_{\rm early}$=1–2% that is constant with scale, while the scatter among the mock catalogues reduces with scale, resulting in a strongly increasing significance of the conformity signal. This figure also shows that the mock results are perfectly centred at $\Delta f_{\rm q}$=0, which means that group errors do not seem to impact 2-halo conformity measurements nearly as much as they did in the 1-halo case.

3.4.2.2 $\mathcal{M}(r_p)$ for 2-halo Conformity

We next study 2-halo conformity using the marked correlation function $\mathcal{M}(r_p)$. We perform a similar analysis as the 1-halo case presented in §3.4.1.2, except that now we only consider pairs of central galaxies from different groups of similar mass. As with the quenched fraction difference case, we count all central-central pairs with a line-of-sight separation less than $\pi_{\rm max} = 20h^{-1}$Mpc and place them in logarithmic bins of projected distance r_p . We then compute $\mathcal{M}(r_p)$ for our three galaxy properties after normalizing them by their mean values within $M_{\rm group}$ bins.

To assess the statistical significance of our results and investigate the impact of grouping errors and mass assignment, we compare our SDSS results with measurements on our 100 mock catalogues that contain no built-in 2-halo conformity. Once again, we use the standard deviation of mock $\mathcal{M}(r_p)$ values to estimate the 1σ, 2σ, and 3σ ranges of the mock distribution. We then calculate the residuals of the SDSS measurements with respect to mocks as in equation (3.10).

Figure 3.5 shows the $\mathcal{M}(r_p)$ of $(g-r)$ colour (left), sSFR(middle), and Sérsic index (right), as a function of projected distance, r_p , with each row corresponding to a bin of $M_{\rm group}$. They layout is similar to that in the previous figures. The figure reveals weak, but highly significant 2-halo conformity signals for all three properties in low mass haloes. In the lowest mass bin, $\log M_{\rm group}$: 11.6–12.0, these signals reach a significance as high

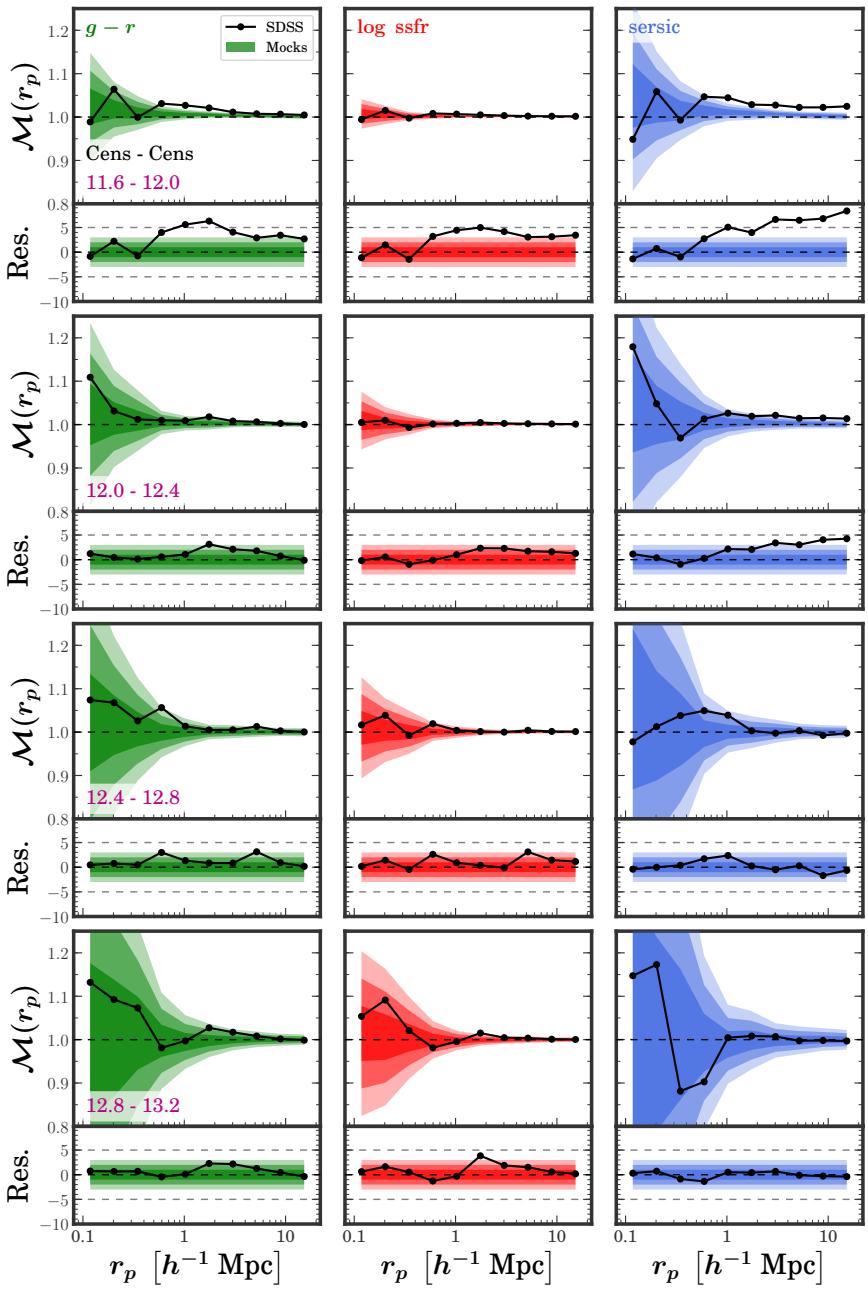

Figure 3.5: Mark correlation function, $\mathcal{M}(r_p)$, of $(g-r)$ colour (left), sSFR(centre), and Sérsic index (right), as a function of projected distance r_p for central-central galaxy pairs within separate galaxy groups in the Mr19-SDSS sample and mock catalogues. Each row corresponds to a bin of group mass, $M_{\rm group}$, as listed in the left panels. *Top panels*: The solid black lines show results for SDSS, while the shaded contours show the 1σ, 2σ, and 3σ ranges of $\mathcal{M}(r_p)$ calculated from 100 mock catalogues with no built in 2-halo conformity. *Bottom panels*: Normalised residuals of $\mathcal{M}(r_p)$ with respect to the mock catalogues. The solid black lines show the difference between $\mathcal{M}(r_p)$ of the SDSS and the mean of the mocks, divided by the standard deviation of $\mathcal{M}(r_p)$ for the mocks. The shaded contours show the 1σ, 2σ, and 3σ ranges of the mocks in this normalised space.

as 7σ. In the case of $(g-r)$ colour, the signal reaches as high as $\mathcal{M}(r_p)$ =1.02-1.03 and then declines with scale, while the statistical significance peaks at scales $r_p : 0.6-4h^{-1}$Mpc and hovers at the 3σ level out to $r_p \sim 10h^{-1}$Mpc before dropping at larger scales. There is no significant large-scale conformity signal in more massive group bins. sSFRbehaves the same way, except that the conformity signal is much weaker (yet equally signifiant), peaking at $\mathcal{M}(r_p)$ value of less than 1.007. In the case of Sérsic index, the signal is also as high as $\mathcal{M}(r_p)$ =1.02-1.03, but, unlike with colour, it keeps this constant amplitude out to the largest scales we consider. As a result, the statistical significance of the conformity signal keeps rising with scale because the scatter in the mock distribution decreases with scale. In the next mass bin, $\log M_{\rm group}$: 12.0-12.4, the conformity signals almost disappear, but are still significant for Sérsic index. There are no 2-halo conformity signals in the higher group mass bins.

These results are similar to what we found using the quenched fraction difference statistic, where Sérsic index displayed the strongest 2-halo conformity signal but only for the low mass groups. However, the marked correlation function is a more sensitive statistic for detecting 2-halo conformity as demonstrated by the much higher statistical significance of the weak observed signals in the case of color and sSFR. Where we found no strong evidence of 2-halo conformity using Δf in Figure 3.4, we find strong such evidence using $\mathcal{M}(r_p)$ in Figure 3.5. The marked correlation function is clearly a more sensitive probe of 2-halo conformity than quenched fractions, and gives us a better handle on 2-halo conformity signals for colour and sSFR. In summary, we have found low amplitude, but highly significant 2-halo conformity signals for $(g-r)$ colour and sSFRout to $4h^{-1}$Mpc and an intriguing signal in Sérsic index out to the largest scales that we probe.

3.5 Summary and Discussion

In this paper, we study galactic conformity, which is the phenomenon that galaxy properties, such as colour or morphology, may exhibit correlations across distance, beyond what

would be expected if these properties only depended on halo mass. At small scales, this "1-halo conformity" is seen as a correlation between the properties of satellite galaxies with those of the central galaxy whose halo they inhabit. At large scales, "2-halo conformity" is seen as a correlation between central galaxies in haloes that are well separated from each other. In both cases, it is important to control for halo mass in order to ensure that any detected correlations are not simply due to the well-established correlations between galaxy properties and halo mass, as well as the correlation between halo mass and larger-scale environment. We are motivated to perform a comprehensive study of conformity because recent works have exposed systematic problems with previous claims of conformity detection at $z = 0$, calling into question whether conformity has actually been detected. In the 1-halo regime, the original detection came from W06 using a group catalogue to designate central and satellite galaxies and to control for halo mass. However, Campbell (2015) used a mock catalogue to show that errors in group-finding and group mass assignment can lead to a spurious 1-halo conformity signal when none is actually present. In the 2-halo regime, K13 detected conformity out to 4 Mpc using isolation criteria to avoid including satellite galaxies. However, T17 and S17 showed that this result was most likely due to insufficiently stringent isolation criteria and that the detected conformity signal arose from a small number of satellite galaxies that were misidentified as centrals.

We investigate both 1-halo and 2-halo conformity using a galaxy group catalogue from the SDSS DR7. Our analysis contains three main improvements over previous works. First, we study three observed properties of galaxies: $(g-r)$ colour, sSFR, and Sérsic index. Second, we use a new statistic, the marked correlation function, $\mathcal{M}(r_p)$, in addition to the previously used quenched fractions. $\mathcal{M}(r_p)$ is ideally suited for conformity studies and is a more sensitive probe of weak conformity signals. Third, we use a suite of 100 mock galaxy catalogues to quantify the statistical significance of our results. These mock catalogues have the same clustering and same distributions of "observed" properties as the SDSS data (luminosity, $(g-r)$ colour, sSFR, and Sérsic index), and we analyse them in exactly the

same way (i.e., same group-finding algorithm, same way of assigning group masses, etc). The mock catalogues do not have any built-in conformity, but they are affected by the same systematic errors as the SDSS data. By comparing our SDSS measurements to the distribution of mock measurements, we can quantify the probability that whatever signal we detect could have arisen from a model with no conformity.

The main results of our work are as follows.

- When measuring the difference between quenched fractions of satellite galaxies around quenched vs. non-quenched centrals, we detect a strong 1-halo conformity signal at all group masses, which is strongest for $(g-r)$ colour, somewhat weaker for sSFR, and only significant at low masses for Sérsic index. These results are in perfect agreement with the results of W06. However, when we compare the $(g-r)$ colour, sSFR, and Sérsic index results to measurements made on our mock catalogues, we find that they are also in perfect agreement. Since the mock catalogues contain no built-in conformity, this strongly suggests that the conformity signal we detected is a result of systematic errors in the group mass estimation and in central/satellite mis-assignment. This calls into question the validity of the W06 detection, as well as other 1-halo conformity detections at $z = 0$ that use group catalogues.

- The marked correlation function, $\mathcal{M}(r_p)$, calculated with central-satellite galaxy pairs is sensitive to the radial segregation of satellite galaxy properties within groups. Using the 1-halo $\mathcal{M}(r_p)$, we find significant radial segregation for colour and sSFRat scales smaller than $r_p < 0.2\,h^{-1}$Mpc in the case of groups more massive than $\log M_{\rm group} > 13$. We do not find such a signal for Sérsic index. We thus claim a detection of radial segregation in $(g-r)$ color and sSFR.

- After removing the effect of radial segregation from $\mathcal{M}(r_p)$ by properly renormalising galaxy properties, the amplitude of $\mathcal{M}(r_p)$ reduces and the conformity signal mostly vanishes. We thus do not detect 1-halo conformity using the $\mathcal{M}(r_p)$ statistic in any of the

three galaxy properties.

- Studying the quenched fraction difference statistic as a function of projected scale for central-central galaxy pairs in groups of similar mass reveals no 2-halo conformity signal for $(g-r)$ colour or sSFR. However, we find a highly significant 2-halo conformity signal for Sérsic index in low mass groups of $\log M_{\rm group} < 12.4$. This signal is constant with scale and thus increases in statistical significance with scale. The mock measurements of the three galaxy properties indicate that group errors do not strongly affect our 2-halo quenched fractions, and that the detection of 2-halo conformity in Sérsic index is likely robust.

- The $\mathcal{M}(r_p)$ of central-central galaxy pairs proves to be a more sensitive probe of conformity than quenched fractions. We find a low amplitude, yet highly significant signal in all three galaxy properties for group masses below $\log M_{\rm group} = 12$. For $(g-r)$ colour and sSFR, the signal is strongest at scales of $r_p : 0.6-4h^{-1}$Mpc and hovers at the 3σ level out to $r_p \sim 10h^{-1}$Mpc before dropping at larger scales. For Sérsic index, the 2-halo conformity signal increases in significance with scale. There is no significant large-scale conformity signal in more massive groups. Our detection is unlikely caused by group errors and thus represents robust 2-halo conformity detections in colour, sSFR, and Sérsic index for central-central galaxy pairs at low masses.

These results demonstrate the importance of using mock galaxy catalogues in any study of galactic conformity. Comparing our SDSS measurements with the distribution of mock measurements allows us to test the null model (i.e., no conformity) in a way that includes systematic errors in group-finding or mass estimation. Without the mock catalogues, we would have claimed a strong detection of 1-halo conformity. Instead, we are driven to the conclusion that the 1-halo signal is not real. This result calls into question whether any study has actually detected 1-halo conformity in the SDSS data. The one caveat to these conclusions is that they only hold to the extent that our mock catalogues faithfully represent

the real universe. If, for example, the correlation between sSFRand halo mass in the mocks is not as strong as it should be, then the impact of group mass errors on the conformity signal will not be accurate.

In the case of 2-halo conformity, we do not find any statistically significant signals when looking at quenched fractions using colour or sSFR. We thus agree with the claim in T17, that the K13 result must have suffered from errors in the isolation criteria used. On the other hand, we show that the marked correlation function is more sensitive to the underlying weak signal and displays a clear conformity trend, even when compared against the mock catalogues. This measurement may thus represent the first robust detection of 2-halo conformity to-date. Our finding that 2-halo conformity is strongest when considering galaxy Sérsic index is curious and merits further study. Overall, to understand the physical origin of these conformity signals, it will be necessary to model them in detail, which we leave for future work.

Chapter 4

PREDICTION OF GALAXY HALO MASSES IN SDSS DR7 VIA A MACHINE LEARNING APPROACH

The following work has been submitted to the Monthly Notices of the Royal Astronomical Society Journal and is reprinted below in its entirety

Prediction of galaxy halo masses in SDSS DR7 via a machine learning approach

Victor F. Calderon[1], Andreas A. Berlind[1]

[1] Department of Physics and Astronomy, Vanderbilt University, Nashville, TN 37235

4.1 Abstract

We present a machine learning (ML) approach for the prediction of galaxies' dark matter halo masses that achieves an improved performance over conventional methods. We train three ML algorithms (`XGBoost`, Random Forests, and neural network) to predict halo masses using a set of synthetic galaxy catalogues that are built by populating dark matter haloes in N-body simulations with galaxies, and that match both the clustering and the joint-distributions of properties of galaxies in the Sloan Digital Sky Survey (SDSS). We explore the correlation of different galaxy- and group-related properties with halo mass, and extract the set of nine features that contribute the most to the prediction of halo mass. We find that mass predictions from the ML algorithms are more accurate than those from halo abundance matching (`HAM`) or dynamical mass (`DYN`) estimates. Since the danger of this approach is that our training data might not accurately represent the real Universe, we explore the effect of testing the model on synthetic catalogues built with different assumptions than the ones used in the training phase. We test a variety of models with different ways of populating dark matter haloes, such as adding velocity bias for satellite galaxies. We

determine that, though training and testing on different data can lead to systematic errors in predicted masses, the ML approach still yields substantially better masses than either `HAM` or `DYN`. Finally, we apply the trained model to a galaxy and group catalogue from the SDSS DR7 and present the resulting halo masses.

4.2 Introduction

The practice of grouping galaxies observed in a galaxy catalogue into galaxy groups and clusters has been utilised extensively in astrophysics and cosmology, since the pioneering work of George Abell and Fritz Zwicky (Abell, 1958; Zwicky et al., 1968), who constructed cluster catalogues from the Palomar Observatory Sky Survey (POSS) using local galaxy surface number densities. Galaxy clusters represent the largest primordial density perturbations to have formed by now, and typically contain tens to hundreds of galaxies embedded within a common dark matter halo[1], thus tracing the high mass tail of the halo mass function. As a result, clusters constitute one of the most powerful cosmological probes and measurements of their abundance can be used to constrain cosmological parameters (e.g., Voit, 2005; Allen et al., 2011; Kravtsov & Borgani, 2012; Weinberg et al., 2013; Mantz et al., 2014). Additionally, our current understanding of galaxy formation and evolution revolves around the idea that all galaxies are formed and live within dark matter haloes. Therefore, galaxy groups and clusters, if identified correctly, can be used to study the *galaxy-halo* connection and thus how galaxies form and evolve within dark matter haloes. Whether we wish to use galaxy groups as probes of cosmology or galaxy formation, determining their masses accurately and robustly has proven to be a difficult task.

Galaxy groups and clusters can be identified in various ways. Originally, clusters were first detected as overdensities of galaxies in broad-band images in the visible spectrum (e.g. Abell, 1958; Zwicky et al., 1968). Since then, clusters have mainly been identified as

[1]Throughout this paper, we use the term "halo" to refer to a gravitationally bound structure with overdensity $\rho/\bar{\rho} \sim 200$, so an occupied halo may host a single luminous galaxy, a group of galaxies, or a cluster.

overdensities of red galaxies in visible and IR bands (e.g. Gladders & Yee, 2005; Hao et al., 2010; Ascaso et al., 2012), as extended X-ray sources (e.g. Rosati et al., 2002; Vikhlinin et al., 2009), or by their signature in the cosmic microwave background (e.g. Marriage et al., 2011; Staniszewski et al., 2009; Ade et al., 2015). Since the early 1980's and with the onset of redshift surveys, groups of galaxies have also been selected based on the closeness of galaxies in redshift space using three-dimensional algorithms. Many of these analyses have adopted the widely-used Friends-of-Friends percolation algorithm (Geller & Huchra, 1983) to place galaxies into groups and thus compile group catalogues. This algorithm links galaxies in pairs based on their separation along the line-of-sight or on the sky and places all linked galaxies into a single group. Numerous group galaxy catalogues have been constructed in this way for different redshift surveys, including the Center for Astrophysics Redshift Survey (CfA; Geller & Huchra, 1983), the Las Campanas Survey (Tucker et al., 1997), the Two Degree Field Galaxy Redshift Survey (2dFGRS; Merchán & Zandivarez, 2002; Eke et al., 2004; Yang et al., 2005; Tago et al., 2006; Einasto et al., 2007), the high-redshift DEEP2 survey (Gerke et al., 2005), the Two Micron All Sky Redshift Survey (Crook et al., 2007), and the Sloan Digital Sky Survey (e.g., Goto, 2005; Berlind et al., 2006).

Once galaxy groups and clusters are identified, mass measurements are needed to map observable properties to the underlying masses of dark matter haloes. Traditionally, there are two main methods to assign masses to galaxy groups and clusters that are built from galaxy redshift surveys, i.e., Halo Abundance Matching (hereafter `HAM`; e.g., Kravtsov et al., 2004; Tasitsiomi et al., 2004; Vale & Ostriker, 2004; Conroy et al., 2006) and dynamical mass estimates (hereafter `DYN`; e.g., Teague et al., 1990; Colless & Dunn, 1996; Fadda et al., 1996; Carlberg et al., 1997; Girardi et al., 1998; Brodwin et al., 2010; Rines et al., 2010; Sifón et al., 2013; Ruel et al., 2014). The `HAM` method assumes a monotonic relation between a theoretical mass-like quantity related to dark matter haloes and another observable quantity related to galaxies. This approach is simple yet powerful, wherein matching

cumulative number densities of galaxies and haloes yields an implicit relationship between the theoretical quantity and the observational quantity (Hearin & Watson, 2013). `HAM` is typically used to connect galaxies to both host haloes and subhaloes, but in this context we refer to a variant of the method that connects galaxy groups to host haloes alone. For example, Yang et al. (2007) applied a halo-based group-finder (Yang et al., 2005) to the 2dFGRS and assigned halo masses to galaxy groups based on characteristic luminosity and characteristic stellar mass. Lim et al. (2017) extended this approach and applied a modified version of the same algorithm to multiple large redshift surveys. Calderon et al. (2018) applied the Berlind et al. (2006) algorithm to the SDSS and used `HAM` to estimate halo masses, based on the integrated luminosity of the groups. Moffett et al. (2015) did the same for the REsolved Spectroscopy of a Local VolumE (RESOLVE; Eckert et al., 2015) and the Environmental COntext catalog (ECO; Moffett et al., 2015). On the other hand, `DYN` estimates of clusters use the line-of-sight velocity dispersion of galaxies within clusters, together with measurements of their size, as dynamical tracers of the underlying gravitational potential. These estimates make use of variants of the virial theorem to estimate group masses.

Each of these approaches are not perfect, and may include possible biases or systematic errors in their mass estimates that may influence the final results. Old et al. (2014) performed an extensive comparison between various galaxy-based cluster mass estimation techniques that use position, velocities, and colours of galaxies to quantify the scatter, systematic biases and completeness of cluster masses derived from a diverse set of 25 galaxy-based methods. They found that abundance-matching and richness-based methods provide the best results, with some estimates being under- and overestimated by a factor greater than ten. Wojtak et al. (2018) studied these results further and found that contamination in cluster membership can affect the mass estimates greatly, with all methods either overestimating or underestimating the final cluster masses when applied to contaminated or incomplete galaxy samples, respectively. Additionally, Armitage et al. (2018) used the `C-EAGLE` galaxy clusters sample (Barnes et al., 2017) to quantify the bias and scatter of

three mass estimators, and found no significant bias, but a large scatter when comparing estimated to true masses. For the case of `HAM` , Campbell et al. (2015) compared three different FoF-based group-finding algorithms by applying them to a realistic mock galaxy catalogue where the halo masses are known. They found that estimating group masses using `HAM` is limited by the intrinsic scatter in the relation between the observed quantity and the halo mass. They also show that errors in the group-finding process can cause catastrophic errors in estimated halo mass.

These previous works have demonstrated that galaxy groups and clusters identified in redshift surveys have mass estimates that are prone to large statistical and systematic errors, mostly due to failures of the group finding algorithms. These methods for estimating mass use one or two properties of groups, such as total luminosity in the case of `HAM`, or velocity dispersion and radius in the case of `DYN`. However, there are many additional properties of groups that should contain information about halo mass, such as colours and star formation rates, full density and velocity profiles, large scale environments, etc. This suggests the opportunity to apply nonparametric algorithms to analyse the abundant data at our disposal. There has been a significant increase in recent years in the number of studies applying machine learning (ML) techniques to astronomy. One of the most important applications of ML in astronomy is the classification of various objects, e.g. transient events (Mahabal et al., 2008) and galaxy morphology (Banerji et al., 2010). Other applications include the determination of photometric redshifts of galaxies from a set of broadband filters (Ball et al., 2007; Gerdes et al., 2010), the assignment of dark matter haloes to generate synthetic catalogues from N-body simulations (Xu et al., 2013; Kamdar et al., 2016a,b) and the study of the structure of the Milky Way (Riccio et al., 2016). Relevant to this work, ML has also been used to improve galaxy cluster dynamical mass measurements by employing the entire line-of-sight velocity PDF information of galaxy clusters (Ntampaka et al., 2015, 2016). More recently, ML algorithms have been used to measure cluster masses using a combination of dynamical and X-ray data (Armitage et al., 2019), and more complex

algorithms have been employed to estimate the masses of galaxy clusters using synthetic X-ray images from cosmological simulations (Ntampaka et al., 2018). However, these studies were restricted to the massive cluster regime.

In this paper, we explore the possibility of employing ML techniques to estimate the halo masses of galaxies in a wide range of mass. We adopt observed properties of both the galaxies and their groups to act as features and we train the ML algorithms on synthetic data. This paper is organised as follows. In §4.3, we describe the observational and simulated data used in this work (§4.3.1), introduce the set of *features* used in this analysis (§4.3.2), and present the main set of ML algorithms that we use (§4.3.3). In §4.4, we provide the main analysis of *feature selection* (§4.4.1), and present our main results of mass estimates (§4.4.2). In §4.5 we also present a detailed examination of how mass estimates may vary depending on the choice of HOD parameters (§4.5.1), velocity bias, $\sigma_{v,b}$ (§4.5.2), or scatter in the mass-to-light ratio of central galaxies (§4.5.3). In §4.6, we apply our trained algorithms to SDSS, and present the resulting galaxy catalogue with various estimates of halo mass. We summarise our results and discuss their implications in §4.7. The Python code and catalogues used in this project will be made publicly available on Github[2] upon publication of this paper.

4.3 Data and Methods

In this section, we present the datasets used throughout this analysis, and introduce the main ML algorithms and statistical methods that we use to estimate the halo masses of galaxies. In §4.3.1, we briefly describe the SDSS galaxy sample and synthetic galaxy catalogues that we use, along with the parameters that are included in these catalogues. In §4.3.2, we introduce the different *features* that we use for training our ML predictors, and provide a guide on how these are calculated. Finally, in §4.3.3 we provide a brief overview of the different algorithms that we use in this analysis, as well as the default tuning parameters used by each algorithm.

[2] https://github.com/vcalderon2009/SDSS_Groups_ML

4.3.1 SDSS Galaxy Sample and Mock Galaxy Catalogues

For this analysis, we make use of a modified version of the galaxy and group galaxy catalogues used in Calderon et al. (2018). We will provide a brief description of the galaxy sample used, and also an overview of the synthetic galaxy and group galaxy catalogues used in this analysis.

4.3.1.1 SDSS Galaxy Sample

For this analysis, we use data from the Sloan Digital Sky Survey (hereafter SDSS; York, 2000). SDSS collected its data with a dedicated 2.5-meter telescope (Gunn et al., 2006), camera (Gunn et al., 1998), filters (Doi et al., 2010), and spectrograph (Smee et al., 2013). We construct our galaxy sample from the `large-scale structure` sample of the NYU Value-Added Galaxy Catalogue (NYU-VAGC; Blanton et al., 2005), based on the spectroscopic sample in Data Release 7 (SDSS DR7; Abazajian et al., 2009). The main spectroscopic galaxy sample is approximately complete down to an apparent r-band Petrosian magnitude limit of $m_r = 17.77$. However, we have cut our sample back to $m_r = 17.6$ so it is complete down to that magnitude limit across the sky. Galaxy absolute magnitudes are k-corrected (Blanton et al., 2003) to rest-frame magnitudes at redshift $z = 0.1$.

We construct a volume a volume-limited galaxy sample that contains all galaxies more luminous than $M_r = -19$, and we refer to this sample as `Mr19-SDSS`. The redshift limits of the sample are $z_{\rm min} = 0.02$ and $z_{\rm max} = 0.067$ and it contains 90,893 galaxies with a number density of $n_{\rm gal} = 0.01503\ h^3{\rm Mpc}^{-3}$. The sample includes the right ascension, declination, redshift, and $(g-r)$ colour for each galaxy.

To each galaxy, we assign a star formation rate (SFR) using the MPA-JHU Value-Added Catalogue DR7[3]. This catalogue includes, among many other parameters, stellar masses based on fits to the photometry using Kauffmann et al. (2003) and Salim et al. (2007), and star formation rates based on Brinchmann et al. (2004). We cross-match the galaxies of the

[3] https://wwwmpa.mpa-garching.mpg.de/SDSS/DR7/

NYU-VAGC to those in the MPA-JHU catalogue using their MJD, plate ID, and fibre ID. A total of 5.65% of galaxies in the sample did not have corresponding values of SFR and were removed from the main sample. This leaves a sample of 85,578 galaxies. For each of these galaxies, we divide its SFR by its stellar mass to get specific star formation rates, sSFR.

Ultimately, we identify galaxy groups using the Berlind et al. (2006) group-finding algorithm. This is a Friends-of-Friends (FoF; Huchra & Geller, 1982) algorithm that links galaxies recursively to other galaxies that are within a cylindrical linking volume. The projected and line-of-sight linking lengths are $b_\perp = 0.14$ and $b_\parallel = 0.75$ in units of the mean inter-galaxy separation, respectively. This choice of linking lengths was optimised by Berlind et al. (2006) to identify galaxy systems that live within the same dark matter halo. In each group, we define the most luminous galaxy (in the *r*-band) to be the 'central' galaxy. The rest of the galaxies are defined as 'satellite' galaxies.

In previous works, we have estimated the total masses of the groups via *abundance matching*, using total group luminosity as a proxy for mass. Specifically, we assume that the total group *r*-band luminosity $L_{\rm group}$ increases monotonically with halo mass $M_{\rm h}$, and we assign masses to groups by matching the cumulative space densities of groups and haloes:

$$n_{\rm group}(> L_{\rm group}) = n_{\rm halo}(> M_{\rm h}). \tag{4.1}$$

To calculate the space densities of haloes, we adopt the Warren et al. (2006) halo mass function assuming a cosmological model with $\Omega_m = 1 - \Omega_\Lambda = 0.25$, $\Omega_b = 0.04$, $h \equiv H_0/$ (100 km s^{-1} Mpc^{-1}) $= 0.7$, $\sigma_8 = 0.8$, and $n_s = 1.0$. We refer to these abundance matched masses as *group masses*, $M_{\rm group}$. In this paper, we also use a dynamical mass estimate for each group, as well as other group properties, which are described in §4.3.2.

4.3.1.2 Mock Galaxy Catalogues

In order to make proper predictions of the halo masses of galaxies, we need a training dataset where the halo mass of each galaxy is known. This necessitates that we use mock, rather than real data. However, the accuracy of our predictions hinges on the degree to which the mock data are truly representative of the observable Universe. Therefore, the mock dataset must not only contain the same observable properties that we will use as features in the SDSS data, it should also faithfully reproduce the true correlations between these properties and halo mass. At a minimum, the training data should be able to accurately reproduce the observed clustering of galaxies and the joint distributions of "observed" galaxy properties.

For this project, we use a suite of 10 realistic synthetic galaxy and group galaxy catalogues similar to Calderon et al. (2018), with the one exception that we use a different definition when identifying dark matter haloes, i.e. we use a *spherical-overdensity* (SO) definition as opposed to the *Friends-of-Friends* (FoF) halo definition used in Calderon et al. (2018). These synthetic catalogues are based on the *Large Suite of Dark Matter Simulation* (`LasDamas`) project[4] (McBride et al., 2009), and have the same clustering and same distributions of "observed" properties as the SDSS data (luminosity, $(g-r)$ colour, sSFR, and Sérsic index). We use an Halo Occupation Distribution (HOD; Berlind & Weinberg, 2002) model to populate the DM haloes with central and satellite galaxies, whose numbers as a function of halo mass were chosen to reproduce the number density, $n_{\rm gal}$, projected 2-point correlation function, $w_p\left(r_p\right)$, and group multiplicity function, $n(N)$, of the `Mr19`-SDSS sample. Specifically, we use the best-fit HOD values of Sinha et al. (2018) for the case of the `Mr19`-SDSS sample, the 'LasDamas' cosmology, the 'Mvir' halo definition, and the 'PCA' option.

Once galaxies are placed in haloes, we assign luminosities and colours using modified versions of the Conditional Luminosity Function (CLF; Yang et al., 2003) framework and

[4] http://lss.phy.vanderbilt.edu/lasdamas/

the Zu & Mandelbaum (2016) halo-quenching model. This approach yields luminosity and colour distributions as well as luminosity- and colour-dependent clustering that are in agreement with SDSS measurements. The resulting mock catalogues have been analysed in exactly the same way as the SDSS data (i.e. same group-finding algorithm, same method of assigning group masses, etc). In their final version, the catalogues contain information on various galaxy-related properties (e.g., sSFR, Sérsic index, $(g-r)$ colour, luminosity) and group-related properties (e.g., group richness, groups' total r-band absolute magnitudes, velocity dispersion within the groups, etc).

For a more detailed explanation of what went into producing the set of mock catalogues used in this analysis, we refer the reader to §2.3 of Calderon et al. (2018).

4.3.2 Galaxy properties as *features*

As part of our analysis, we must make a decision on which *features* to use when training the ML algorithms to predict the masses of galaxies' dark matter halos. The set of features that we use includes properties of the galaxy in question as well as properties of the group to which the galaxy belongs. All features can be observed and measured in the SDSS. Here we provide a list of the features that we consider initially with a description of how each is computed. Later on we reduce this to a shorter list using a feature selection algorithm.

Galaxy-related features

1 **Distance to group's centre**: This feature refers to how far a galaxy is from the centre of its corresponding galaxy group. This variable is given in units of of h^{-1}Mpc , but it is calculated in three-dimensional space so it is dominated by the velocity component of the galaxy's position. The centre of the group is computed as the centroid of the group's member galaxy positions.

2 **Absolute Magnitude**: r-band absolute magnitude of the galaxy, k-corrected to $z = 0.1$.

3 **Specific star formation rate of the galaxy**, **sSFR**: Logarithmic value of the specific

star formation rate of the galaxy. As mentioned in §4.3.1.2 and in Calderon et al. (2018), in our mock catalogues these sSFRvalues were assigned using the Zu & Mandelbaum (2016) *halo-quenching* model, and matched to the distribution of sSFR values in SDSS DR7 through abundance matching.

4 **Group galaxy type**: The galaxy type of the galaxy, in terms of its galaxy group. We denote a value of "1" if the galaxy is a *group central*, and a "0" if the galaxy is a *group satellite*. After determining the group membership of each galaxy, we designate the brightest galaxy of the group in the *r*-band as the *group central*, while the rest of galaxies are identified as *group satellites*. Hence, a galaxy group is composed of one bright group central and a number of group satellites. This criterion is motivated by the idea that central galaxies grow in mass and brightness by galactic cannibalism (Dubinski, 1998; Cooray & Milosavljević, 2005), while satellite galaxies experience a series of events that strip them from their mass and inhibit star formation (e.g. ram-pressure stripping and tidal stripping).

5 **(g – r) colour of galaxy**: The difference between the absolute magnitudes in the *g*-band and *r*-band, after these have been *k*-corrected to $z = 0.1$. In our mock catalogues, galaxy colours were assigned in a manner similar to that of sSFR.

Group-related features

6 **Luminosity of brightest galaxy**: *r*-band absolute magnitude value of the brightest galaxy in the group that the galaxy in question belongs to. This absolute magnitude is the same as that of the group central galaxy, according to our designation of *group centrals* and *group satellites*.

7 **Luminosity ratio**: Ratio between the *r*-band luminosity of the brightest and second brightest galaxies in the group.

8 **Total luminosity**, $\mathbf{M_{r,tot}}$: The total *r*-band luminosity of the group is the sum of the *r*-

band luminosities of its member galaxies. We compute the total group r-band absolute magnitudes as

$$M_{r,\mathrm{tot}} = -2.5 \log_{10} \left(\sum_{i=1}^{N} 10^{-0.4 M_{0.1_{r,i}}} \right), \tag{4.2}$$

where 'N' corresponds to the number of member galaxies in the group, and '$M_{0.1_{r,i}}$' to the k-corrected r-band absolute magnitude of the i-th galaxy in the galaxy group. The resulting variable is the groups' total r-band absolute magnitude, $M_{r,\mathrm{tot}}$.

9 **Total specific star formation rate**, $\mathbf{sSFR_G}$: Logarithmic value of the total specific star formation rate of the group. For each group, the total specific star formation rate is calculated as:

$$\mathrm{sSFR}_G = \frac{\mathrm{SFR}_G}{M_{*,G}} = \frac{\sum_{i=1}^{N} \mathrm{SFR}_i}{\sum_{i=1}^{N} M_{*,i}}, \tag{4.3}$$

where 'N' refers to the number of member galaxies in the galaxy group, '$M_{*,i}$' and 'SFR_i' to the stellar mass and star formation rate of the i-th galaxy in the galaxy group.

10 **Shape**: The shape of the group is calculated by first computing the eigenvalues of the group's moment of inertia tensor, and then by taking the ratio between the values of the largest and second largest eigenvalues. This ratio is what we designate as the *group shape* feature.

11 **Richness**: Richness is the total number of galaxies in the galaxy group. A galaxy group can be composed of a single galaxy, or many galaxies.

12 **Projected rms radius, $\mathbf{R_{\perp,rms}}$**: Projected *rms* radius of the group. It is given by

$$R_{\perp,rms} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} r_i^2}, \tag{4.4}$$

where r_i is the projected distance between each member galaxy and the group centroid. This variable is only computed for galaxy groups with two or more member galaxies. For groups with just one member galaxy, we assign a value of '0' to $\mathbf{R_{\perp,rms}}$.

13 **Maximum projected radius**, $\mathbf{r_{tot}}$: The total radius of the galaxy group corresponds to the projected distance between the centre of the galaxy group and and the most distant member galaxy of the group.

14 **Median projected radius**, $\mathbf{r_{med}}$: The median radius of the galaxy group is the median distance between the centre of the group and the group's member galaxies.

15 **Total velocity Dispersion**, $\sigma_{\mathbf{v}}$: We compute a group one-dimensional velocity dispersion given by

$$\sigma_v = \frac{1}{1+\bar{z}}\sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(cz_i - c\bar{z})^2}, \tag{4.5}$$

where N is the total number of galaxies in the group, $c\bar{z}$ is the mean velocity of the group, and $c\bar{z}_i$ is the velocity of each member galaxy. This variable is only computed for galaxy groups with two or more member galaxies. For groups with just one member galaxy, we assign a value of '0' to $\sigma_{\mathbf{v}}$.

16 **Velocity dispersion within $\mathbf{r_{med}}$**: Similar to $\sigma_{\mathbf{v}}$. We compute a one-dimensional velocity dispersion of the galaxies that are within $\mathbf{r}_{\rm med}$ with Equation 4.5, but only using galaxies within the designated radius from the centre of the galaxy group.

17 **Abundance-matched mass, $M_{\mathbf{group}}$**: We estimate the total mass of the group via *abun-*

dance matching. This method assumes a monotonically increasing relationship between the group total luminosity, $M_{r,\rm tot}$, and the dark matter halo mass. We adopt the Warren et al. (2006) mass function for this purpose.

18 **Dynamical mass**: We follow the prescription from Girardi et al. (1998) for estimating the group dynamical mass, using $\sigma_{\mathbf{v}}$and $\mathbf{R}_{\perp,\mathbf{rms}}$ as follows

$$M_{\rm dyn} = A \times \frac{3\pi}{2} \frac{\sigma_v^2 R_{\perp,rms}}{G}, \tag{4.6}$$

where G is the gravitational constant. A is a fudge factor that we use to remove any systematic offset between the dynamical mass estimate and the true halo mass in the cluster mass regime. Based on tests with our mock catalogs, we set this fudge factor to a value of '1.04'. With this value of A, the above equation recovers the correct halo mass for a massive halo in the ideal case where the radius and velocity dispersion of the halo are known perfectly.

19 **Distance to closest cluster**: Distance to the closest cluster of galaxies that is at least a factor of *10* times more massive than the host group of the galaxy in question. Masses are measured using halo abundance matching and the distance is in units of h^{-1}Mpc and is calculated in three-dimensional space. If no such cluster of galaxies is to be found, we assign a value of '0' to this variable.

This list of features contains spectro-photometric properties of the galaxies, sizes and velocity dispersions of their groups, two halo mass estimates (one derived from spectro-photometric properties, i.e., `HAM` , and one derived from group size and velocity dispersion, i.e., `DYN`), a group morphological parameter, and a large-scale environmental metric. All of these features are expected to contain information about halo mass.

4.3.3 Machine Learning Algorithms

Machine learning is an inventive field in computer science, with a variety of different applications in a number of areas. As mentioned in §4.2, ML algorithms are able to *learn* non-parametric relationships between some input *data* and an expected output, without having to explicitly provide an analytic prescription. In the case of *supervised* learning, which is the type of ML used in this paper, a training dataset $(\mathbf{X}, \mathbf{y})$ is provided, and the ML algorithms try to *learn* the mapping $\mathbf{F}(\mathbf{X} \rightarrow \mathbf{y})$ between the set of features, $\mathbf{X}$, and the expected output, $\mathbf{y}$. Once the algorithm is trained, it is tested on a different 'test' dataset in order to quantify how well it works. Ultimately, the goal is to apply the algorithm to an application dataset where $\mathbf{y}$ is not known.

For our study, we test the performance of 3 different flavours of ML algorithms in order to see which algorithm can provide us with the best prediction for the halo masses of galaxies. We use the *Random Forest* an *Neural Network* algorithms from the python package `scikit-learn`[5] (Pedregosa et al., 2012), as well as the `XGBoost` algorithm [6].

4.3.3.1 Random Forest

One of the ML algorithms that we use in this analysis is *Random Forests* (hereafter `RF`; Breiman, 2001). A random forest is an ensemble learning technique that builds upon a collection of tree-structured classifiers, also known as *decision trees*. For the purpose of this analysis, we implement `RF` for regression rather than for classification, and decision trees are to be referred as *regression trees* in this context. `RF` makes use of the *bagging* method, in which it generates n samples from the dataset, trains each sample individually and averages all of the predictions at the end. For a more comprehensive account of this technique, the reader is referred to Breiman et al. (1984). We implement the `scikit-learn` version of `RF`, `RandomForestRegressor`, with its default settings.

[5] http://scikit-learn.org/
[6] https://xgboost.readthedocs.io/

4.3.3.2 XGBoost

`XGBoost` (Chen et al., 2006) is part of the family of *boosting* algorithms, which makes use of the *boosting* method. In Boosting, unlike in *Bagging*, the algorithm generates n random samples for training with replacement over weighted data. Each of these regression trees are referred to as *weak learners*, and they each get assigned weights based on the accuracy of their predictions. After these weak learners are trained, the weighted averages of each of their estimates are used to compute the final predictions. The combination of weak learners is referred to as *strong learners*. For a more in-depth discussion of `XGBoost` and its different features, the reader is referred to the online documentation [6].

4.3.3.3 Neural network

The last ML algorithm used in this analysis is the simplest type of a neural network (`NN`), i.e. the *Multi-Layer Perceptron* (`MLP`). A `MLP` is a model with interconnected information processing units, often referred to as *neurons*, that learns the mapping $\mathbf{F}(\mathbf{X} \rightarrow \mathbf{y})$ given a training set $(\mathbf{X}, \mathbf{y})$, with $\mathbf{X}$ being the input features and $\mathbf{y}$ the target elements to predict. We implement the `scikit-learn` version of a 3-layer `MLP` with each layer containing 100 *neurons*. We refer the user to the `scikit-learn` documentation[5] for a more comprehensive account of this method.

4.4 Training and Testing ML algorithms

In this section, we present results from the training and testing of the three ML algorithms for predicting the halo masses of galaxies in SDSS DR7. Moreover, we compare these predictions to the more traditional estimates from halo abundance matching (`HAM`) and dynamical mass measurements (`DYN`). In §4.4.1, we present the set of features that contribute the most to the overall prediction of halo mass in order to reduce the dimensionality of our feature space in further training. In §4.4.2, we present results from the training and testing phases of each of the three ML algorithms using our synthetic catalogues of the Universe. The mock catalogues used in the training and testing phases are built using the

same HOD model and thus represent the overly optimistic scenario in which the training data perfectly represents the real universe. Results in this section thus serve as a proof of concept that ML is a feasible method of determining the halo masses of galaxies. We explore the more realistic case that the training data is drawn from a different underlying model than the real universe in §4.5.

4.4.1 Feature Selection

In §4.3.2 we presented a list of 19 properties of galaxies and their groups that may contain useful information about halo mass. In this section we analyse the predictive power of these features in order to eliminate ones that are not as useful and thus reduce the overall number of features that we will use as inputs to the ML algorithms. This is conventionally referred to as *feature selection*, and it plays an important role into the training process of a ML algorithm. Reducing the dimensionality of the feature space is desirable because it reduces the computational cost of ML algorithms and can also improve their predictive performance.

Before we determine the importance of each feature for the prediction of halo mass, we first explore the amount of correlation among the different features from §4.3.2. Figure 4.1 presents the correlation matrix of these 19 features as measured from our mock galaxy catalogues. The matrix shows the correlation coefficient between each pair of features, with red and blue shadings corresponding to positive or negative correlation, respectively. The matrix also includes halo mass in the first column and thus reveals how much each feature is correlated with the quantity we are trying to predict. Figure 4.1 shows that almost all 19 of our features exhibit correlations with halo mass. Additionally, many of the features are highly correlated with each other, as expected, and are thus unlikely to contain independent information about halo mass.

To quantify the importance of each feature for the purpose of feature selection, we use the native feature importance calculation within the `RF` and `XGBoost` algorithms (the `NN`

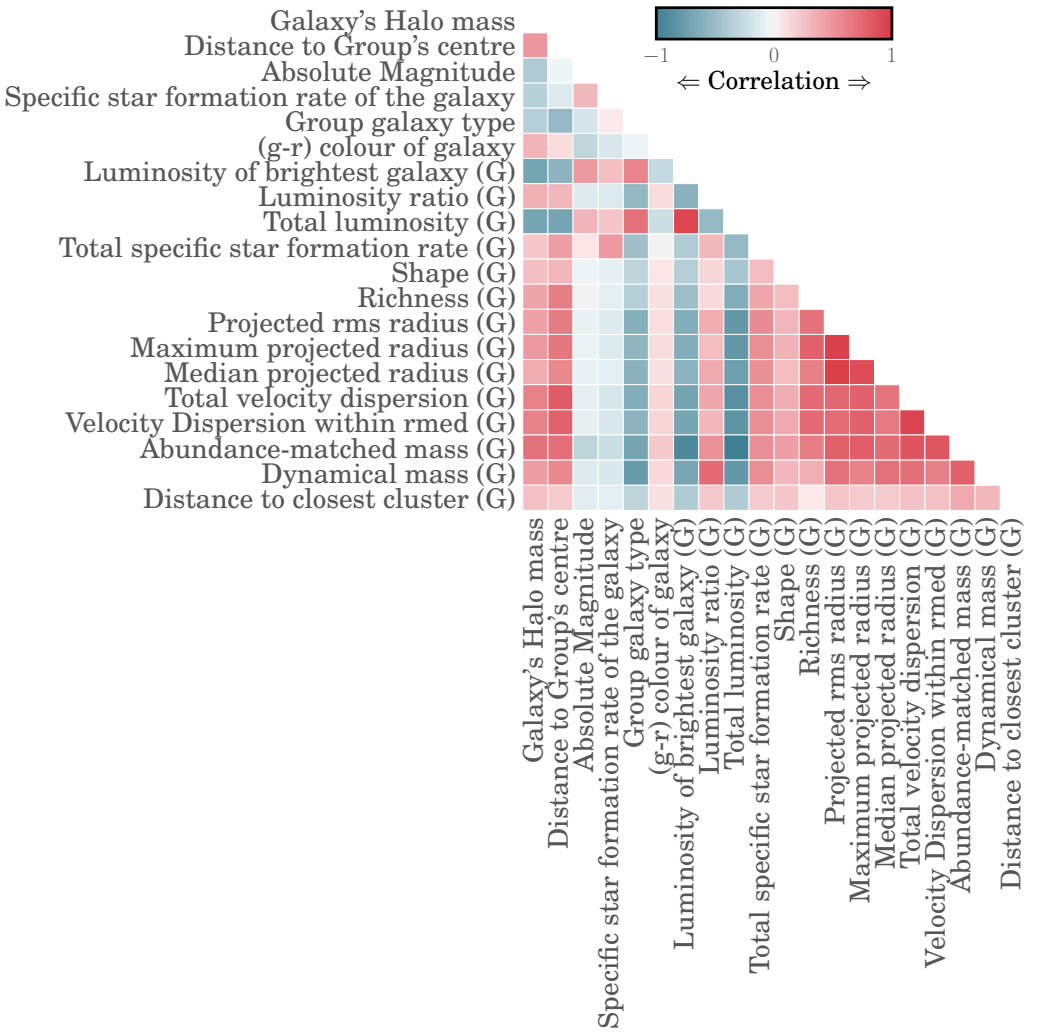

Figure 4.1: Correlation matrix of the galaxy- and group-related features presented in §4.3.2, computed using our mock galaxy catalogues. The figure shows how correlated the features are with each other, with red and blue shadings corresponding to positive and negative correlations, respectively. Additionally, the first column displays the degree of correlation of each feature with halo mass, which is the quantity we wish to predict. This figure conveys the point that the mass of the dark matter halo is strongly correlated with almost all of the features that we consider for training the different ML algorithms.

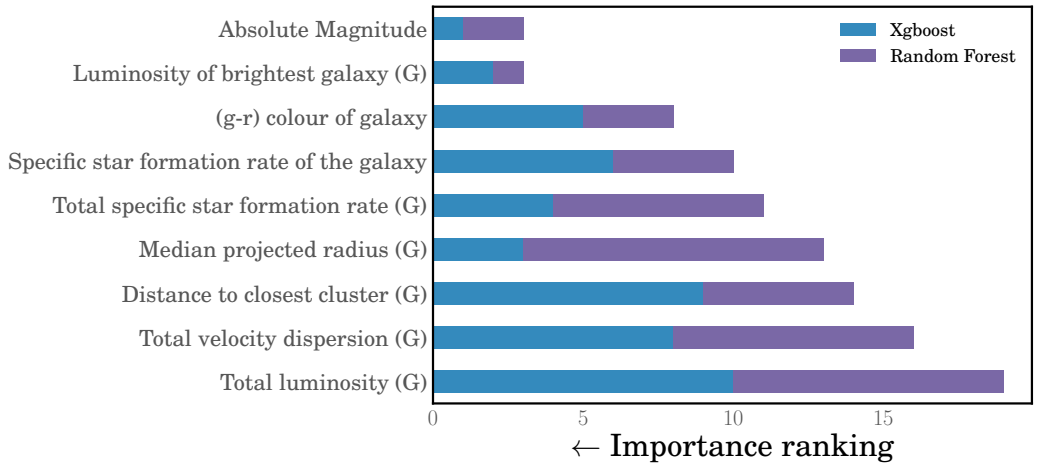

Figure 4.2: Feature importance for the top nine features used when predicting the mass of a galaxy's host dark matter halo, as calculated by the `XGBoost` (blue bars) and `RF` (purple bars) ML algorithms. The length of each bar indicates its importance rank, with shorter bars corresponding to more important features.

algorithm does not compute such a statistic). In general, these algorithms estimate the importance of a feature by calculating how much it is used to make key decisions with their decision trees. Each feature gets an importance score allowing us to compare them to each other and rank them. Though later on we will split our 10 mock catalogues into training and testing subsets, for the purpose of feature selection we use them all to train the `RF` and `XGBoost` algorithms. Each algorithm then produces a ranked list of the 19 features in order of their importance, as discussed above. Though the two algorithms differ in their detailed ranking of features, they are generally consistent and are almost in perfect agreement on which features land in the top nine (out of 19). The remaining set of features do not contribute much to the overall prediction of halo mass and so we focus on these nine features moving forward.

Figure 4.2 shows the feature importance ranks for these top nine features for both the `XGBoost` and `RF` algorithms. In the case of each feature, the length of the blue and purple bar indicates its importance rank as calculated by `XGBoost` and `RF` , respectively, with shorter bars corresponding to more important features. We estimate the overall importance of each feature by adding its two ranks (the combined length of the blue and purple bars)

and we order the features in Figure 4.2 according to this overall score. The figure shows that the luminosity of the galaxy itself and the luminosity of the brightest galaxy in the galaxy's group are the overall most useful features in predicting halo mass, while the total group luminosity is the least useful from this set of top nine features.

We select these top nine features that contribute the most to the prediction of halo mass as our final set of features moving forward.

Final set of features

1 Galaxy's r-band absolute magnitude

2 Luminosity of the *brightest* galaxy in the group

3 Galaxy's $(g-r)$ colour

4 Galaxy's specific star formation rate

5 Group's total specific star formation rate

6 Group's median projected radius

7 Distance to the closest cluster

8 Group's total velocity dispersion

9 Group's total r-band absolute magnitude

For the rest of the analysis in this paper, we will exclusively use this set of features to train the various ML algorithms and evaluate their performance at correctly predicting halo masses.

4.4.2 Training and Testing

Now that we have a final list of nine input features, we can proceed to the training and testing of the ML algorithms. We start with our set of 10 mock galaxy catalogues, each of which has the same volume and approximate number density as the `Mr19-SDSS` sample. Combined, these catalogues contain a total of 758,528 mock galaxies. For each galaxy we

have values for the nine input features as well as the target halo mass. We also have the traditional HAM and DYN mass measurements to compare against.

We split the mock data into *training* and *testing* sets. The training set consists of 8 of the 10 catalogues, while the testing set consists of the remaining 2. We will use the testing set to evaluate how well the trained algorithms perform. It is important to perform this evaluation on an independent set of data from the training set in order to guard against the problem of over-fitting. Sometimes ML analyses also use a third, *validation*, dataset for the purpose of tuning the hyper-parameters of a given ML algorithm. However, in this paper we choose to adopt the default values of hyper-parameters and thus we do not need to add a validation step to our workflow.

After training the three ML algorithms to predict the dark matter halo masses of mock galaxies in the training set, we apply these trained algorithms to the testing data and get a list of predicted masses, $M_{\rm pred}$, for these galaxies. We then compare these predictions against the true halo masses, $M_{\rm true}$, and compute the fractional difference between their logarithmic values as

$$\Delta f = 100 \times \left[\frac{\log M_{\rm pred}}{\log M_{\rm true}} - 1 \right]. \tag{4.7}$$

Each galaxy in the testing set gets three values of Δf (one for each ML algorithm), which are essentially the fractional errors in the ML predictions. Note that these are errors in the *logarithm* of halo mass. A value of Δf =5% thus corresponds to a fractional error in mass of $\sim 250 - 400\%$ for the mass range we consider here. For comparison, we also calculate Δf using the HAM and DYN masses in place of $M_{\rm pred}$. This will allow us to examine how well the ML algorithms perform relative to traditional methods for estimating halo mass.

Figure 4.3 presents results for Δf, as a function of predicted mass, for different methods of estimating the halo masses of galaxies. The solid, coloured lines correspond to the mean fractional difference of galaxies in bins of $M_{\rm pred}$, while the shaded regions represent the

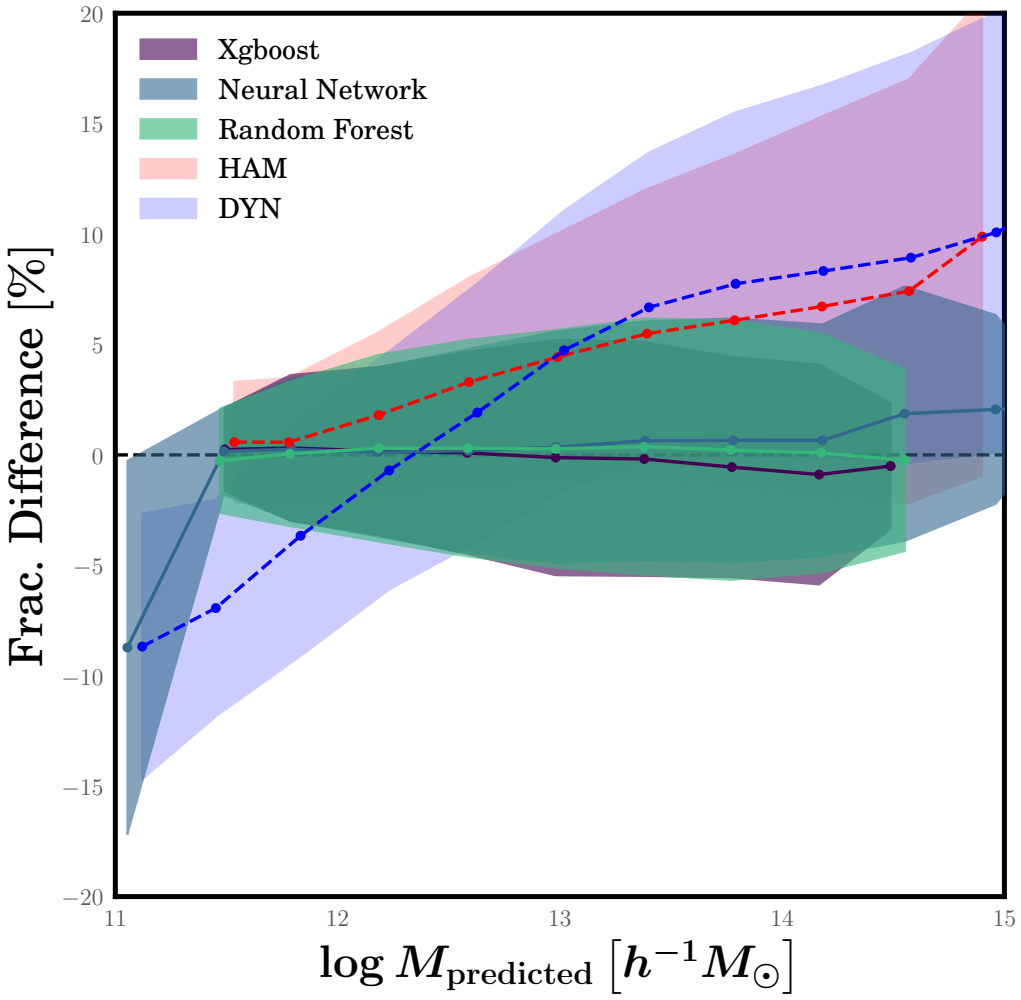

Figure 4.3: Fractional difference between *predicted* and *true* logarithmic halo mass for galaxies, as a function of predicted halo mass, for different methods of estimating the dark matter halo mass of a galaxy. Results are shown for a testing set of mock galaxies, for which their true masses are known. The solid, coloured lines correspond to the mean fractional difference of each method, while the shaded regions represent the 1σ ranges. This figure compares the predictions of halo mass made by the three different ML algorithms to the estimates from conventional methods, i.e. halo abundance matching (`HAM`) and dynamical mass estimates (`DYN`).

1σ ranges of Δf. We show predictions made by the `XGBoost`, `RF`, and `NN` algorithms, and compare these to the mass estimates obtained from `HAM` and `DYN`. Figure 4.3 shows promising results, in that all three ML algorithms are performing significantly better at predicting the mass of a galaxy's host halo than either `HAM` and `DYN`. Specifically, `HAM` yields halo masses that are unbiased on average at low masses and have a 1σ error of $\sim 3\%$, but it systematically overestimates masses on average at high masses, reaching a systematic error as high as $\Delta f \sim 10\%$ in the cluster regime. Moreover, the scatter grows to $\sim 10\%$ in this regime as well. `DYN` exhibits even worse performance since it has similar poor performance for large masses, but also does badly at low masses, systematically underestimating masses on average as much as $\Delta f \sim 10\%$. In contrast, the three ML algorithms yield predicted masses that are unbiased on average at all masses and have a 1σ scatter in Δf of $\sim 3-5\%$.

To understand the poor performance of the `HAM` and `DYN` methods, it is important to consider that we are not evaluating the ability of these methods to correctly estimate the halo masses of galaxy *groups*, but rather of individual *galaxies*. Grouping errors made by the group-finding algorithm can thus cause catastrophic errors in the halo masses of galaxies that have been incorrectly grouped. For example, if the group-finder incorrectly merges together a few galaxies that live in small haloes with the galaxies of a large halo to yield a single massive galaxy group, both `HAM` and `DYN` will estimate a large halo mass for this group and, thus, for all its members. The error in this estimate will be small for the galaxies that actually belong to the large halo, but will be enormous for the galaxies that were mistakenly grouped. It is these catastrophic errors that drive both methods to overestimate the masses of galaxies on average in the high mass regime in Figure 4.3. At low masses, where most galaxies live in $N = 1$ groups, `HAM` does a good job at recovering the mass because galaxy luminosity correlates strongly with mass. `DYN`, however, does poorly because dynamical measurements are very unreliable for systems with a small number of galaxies. The ML algorithms have the advantage that they use additional information that can help fix some of the problems caused by grouping errors. In the example above, the

colours of incorrectly grouped galaxies are likely different from those of actual satellite galaxies in massive halos and the ML algorithms exploit this to distinguish between the two. An exciting possibility that arises from this is that the halo masses predicted by ML could be used to improve the group-finding itself since galaxies whose predicted masses are much smaller than the groups they've been assigned to could be removed from them. We return to this point in the final section.

Another way to quantify the effectiveness of these algorithms at predicting halo masses is to determine the percentile discrepancy between the true and predicted halo masses across a big range of $M_{\rm pred}$ masses. To compute this statistic, we first determine the absolute value of the log-difference between predicted and true halo mass, and rank-order them from smallest to largest. We then determine the discrepancy that corresponds to the 68% of galaxies that are best predicted. This statistic is given by the following equation:

$$(\Delta \log M)_{68} = \mathcal{P}_{68}\Big(\big|\log M_{\rm pred} - \log M_{\rm true}\big| \Big). \tag{4.8}$$

In other words, 68% of galaxies have their masses predicted with an error less than $(\Delta \log M)_{68}$. We split the test sample into a *low-mass* and *high-mass* galaxy sample. Galaxies with $\log M_{\rm pred} \leq 12.5$ are assigned to the *low-mass* sample, while those with $\log M_{\rm pred} > 12.5$ are assigned to the *high-mass* sample. For each sample, we compute $(\Delta \log M)_{68}$ for each ML algorithm, and compare them to those for `HAM` and `DYN`. This statistic shows how well each method is at estimating the halo masses in these two mass regimes.

Figure 4.4 presents the results for the typical mass error $(\Delta \log M)_{68}$. Horizontal bars show values for the three ML algorithms, while solid and dashed vertical lines show results for the `HAM` and `DYN` methods, respectively, for comparison. In all cases, results for galaxies with low predicted masses are shown in red, while results for galaxies with high predicted masses are shown in blue. Figure 4.4 shows clearly that the three ML algorithms exhibit similar performance and they significantly outperform traditional methods in most cases.

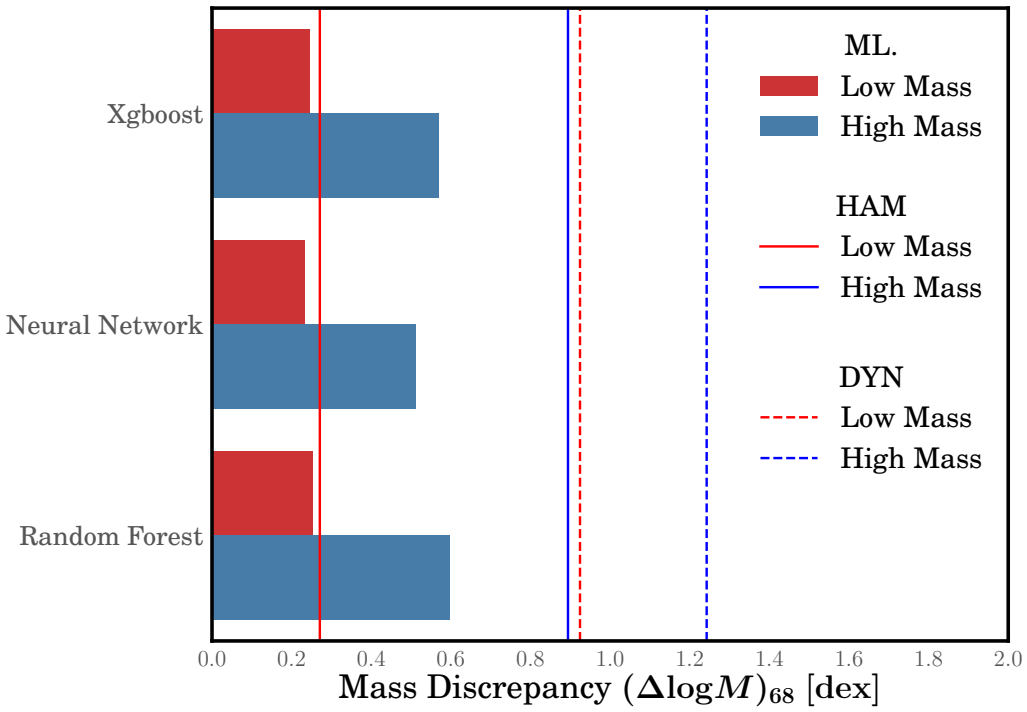

Figure 4.4: Mass discrepancies, $(\Delta \log M)_{68}$, for the three ML algorithms, as compared to those of `HAM` and `DYN` methods, when splitting the galaxy sample into *low-mass* and *high-mass* samples. The quantity $(\Delta \log M)_{68}$ is the 68% prediction error in the log of halo mass, meaning that 68% of galaxies are predicted better than this. The horizontal bars show this typical error for the ML algorithms while the solid and dashed vertical lines correspond to the `HAM` and `DYN` methods, respectively. In all cases, results for galaxies with $\log M_{\rm pred} \leq 12.5$ are shown in red, while results for galaxies with $\log M_{\rm pred} > 12.5$ are shown in blue. The three ML algorithms exhibit similar performance and are significantly better than traditional methods, especially in the high mass regime.

HAM does well at low masses, but at high masses its error is $\sim 50-60\%$ larger than ML methods. DYN does poorly in both mass regimes, with a typical error that is $2-4$ times larger than that for ML methods. More specifically, HAM is able to estimate halo masses to within a $(\Delta \log M)_{68} \approx 0.27$ dex and $(\Delta \log M)_{68} \approx 0.90$ dex for the low-mass and high-mass regimes, respectively. On the other hand, DYN can only recover halo masses to within $(\Delta \log M)_{68} \approx 0.92$ dex and $(\Delta \log M)_{68} \approx 1.25$ dex for the low-mass and high-mass regimes, respectively. The corresponding errors for the XGBoost, RF, and NN ML algorithms range from, $(\Delta \log M)_{68} \approx 0.23-0.25$ dex and $(\Delta \log M)_{68} \approx 0.51-0.60$ dex for the low-mass and high-mass samples, respectively.

In summary, we find that we are able to obtain better mass estimates for a galaxy's host halo by using ML methods in place of the more traditional mass estimators, such as HAM or DYN . This statement is true regardless of predicted mass, $M_{\rm pred}$. However, so far this statement only holds for the case in which the training and testing samples share the same underlying model that connects galaxies to dark matter halos. This is not likely to be true when we apply the trained models to real SDSS data. We address this issue in the next section.

4.5 Are Mock-Trained Models Universally Applicable?

The results shown in §4.4.2 support the notion that we can obtain better halo mass estimates for galaxies by employing ML algorithms instead of the traditional HAM or DYN methods. We evaluated the performance of the ML algorithms using a testing set of mock galaxy catalogues that are independent from the set that we used to train the models. In this context, "independent" means that they are constructed from cosmological N-body simulations that are independent realisations of the density field (i.e., have initial conditions with different random phases). However, the testing catalogues adopt the same prescription for populating dark matter halos with galaxies and assigning them observed properties like luminosity and colour. A better approach would be to test the ML algorithms using cata-

logues that were built with different such prescriptions, since the real universe is unlikely to perfectly conform to the assumptions made in the training phase. In this section, we test the impact of these assumptions in order to assess whether mock-trained models can be applied to the real universe.

4.5.1 Varying HOD models

The first step we make to build mock galaxy catalogues from a dark matter halo distribution is to populate the halos using a HOD model. This model specifies the number of central and satellite galaxies that are placed in each halo. The model is flexible and has five free parameters. We use the best-fit parameter values of Sinha et al. (2018), which ensure that the number density, clustering, and group statistics of our catalogues match those observed in the SDSS. This is the *fiducial* HOD model that we used to train and test our models in §4.4. To test how sensitive our results are to the HOD model of the testing sets, we now produce different versions of our two synthetic testing catalogues, each with different values for the five HOD parameters. We select the parameter sets from the Sinha et al. (2018) MCMC chain so that the resulting mock catalogues are still consistent with SDSS observations. We then run the previously trained ML algorithms on these new test mock catalogues to investigate how much performance we lose from modifying the HOD model in the testing phase.

Figure 4.5 shows the fractional difference between predicted and true halo mass, Δf, for these new test sets. The figure is similar to Figure 4.3, except that it only shows results for the `XGBoost` algorithm and it focuses on the different HOD models instead. Also shown are the `HAM` and `DYN` results for comparison, which are applied to the fiducial test catalogues. We have also done the same tests using the `RF` algorithm and obtained similar results. Figure 4.5 reveals that the performance of the ML algorithm degrades significantly at low masses when it is applied to testing catalogues with different HOD models. For predicted masses larger than $\gtrapprox 10^{12} h^{-1} M_{\odot}$ the effect is negligible and ML clearly outperforms the

HAM and DYN methods just as it did when tested on the fiducial model. However, for $M_{\rm pred} \lessapprox 10^{12} h^{-1} M_{\odot}$, the mean Δf is significantly biased for some of the HOD models, reaching values as high as 4%.

To understand why the ML algorithms degrade at low $M_{\rm pred}$, we take a close look at the HOD parameters of our models to see if there is a trend that explains why some models result in high Δf while others do not. We find a very strong correlation between Δf and $\sigma_{\log M}$, the scatter in halo mass at the luminosity limit of the sample. Test catalogues with high values of this scatter receive predicted masses that are systematically overestimated when trained using the fiducial model. The fiducial model adopts a value of $\sigma_{\log M} = 0.14$ (Sinha et al., 2018), while the most extreme HOD models we test have values of $0.5 - 0.9$. Increasing the scatter this much is equivalent to removing some central galaxies from larger halos and placing them in lower mass halos. However, their observed properties (e.g., luminosity and colour) don't change much because they are assigned in a way that perfectly recovers the observed distributions in the SDSS. For example, in our mock catalogues the faintest r-band absolute magnitudes for mock galaxies are always equal to -19 regardless of their halo mass, since that is the luminosity limit of our SDSS sample. As a result, ML algorithms trained on a catalogue where these faintest galaxies live in more massive haloes, but applied to a catalogue where they live in less massive halos, will learn an incorrect mapping between luminosity and halo mass and thus predict masses that are too high.

Figure 4.5 suggests that in the low mass regime, the HAM method can yield more reliable halo masses than the ML algorithms. However, this is not the case. The HAM result shown is only for the fiducial model and performs well at low mass. However, the HAM method applied to the other HOD models exhibits even worse performance than the ML algorithms. The reason for this is that catalogues built assuming a high $\sigma_{\log M}$ have their lowest luminosity galaxies living in lower mass haloes than they do in catalogues with a smaller scatter, but their number density is not correspondingly higher because not all haloes down to this mass are occupied. Since the HAM method uses abundances to assign mass, it will overpre-

dict these galaxies' masses. So even though ML does poorly when applied to high $\sigma_{\log M}$ datasets, it still outperforms HAM. Another thing to consider is that the ML algorithms only perform poorly when applied to very large values of $\sigma_{\log M} = 0.5 - 0.9$, which are likely inconsistent with observed data. The true amount of this scatter in the real universe is most likely close to ~ 0.2 where our trained ML algorithms perform quite well.

4.5.2 Varying Satellite Galaxy Velocity bias

In the previous section, we demonstrated the effect of varying the HOD parameters that control the number of central and satellite galaxies that occupy haloes as a function of mass. Now we investigate varying how we place these galaxies in their haloes when we construct test mock catalogues. Specifically, we study the effect of adding velocity bias to our mocks. In the fiducial model, satellite galaxies are assigned the positions and velocities of randomly selected dark matter particles within their haloes. However, it is possible that satellite galaxies have kinematics that are either hotter or colder than the underlying dark matter (e.g., Guo et al., 2015). This is referred to as *velocity bias*. We parameterise this bias as the ratio between the velocity dispersion of satellite galaxies, $\sigma_{v,\rm sat}$, within a halo and the velocity dispersion of dark matter, $\sigma_{v,\rm dm}$,

$$\sigma_{v,\rm sat} = f_{\rm vb} \times \sigma_{v,\rm dm}, \tag{4.9}$$

where $f_{\rm vb}$ is the velocity bias parameter, and we explore models with values between $f_{\rm vb}$ =0.9 and 1.1. We implement velocity bias into our mock catalogues simply by scaling satellite galaxies' assigned velocities by $f_{\rm vb}$. Velocity bias is important in this ML context because it directly affects dynamical measurements of group mass. A test mock catalogue with velocity bias will have a different relationship between group velocity dispersion and halo mass, which could cause errors in the predicted mass since velocity dispersion is a feature used by the ML algorithms. In addition, velocity bias will change the size of small-scale redshift distortions in groups, which can affect grouping errors.

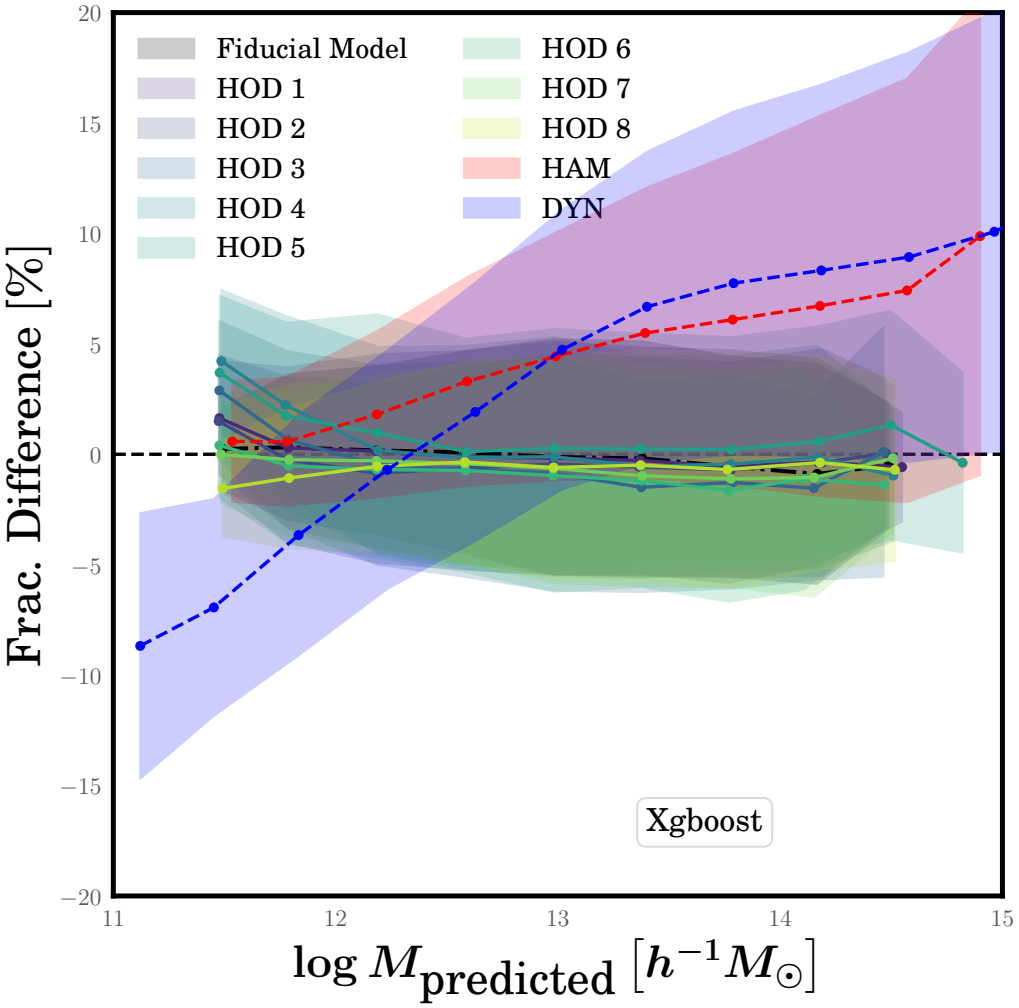

Figure 4.5: Fractional difference between *predicted* and *true* logarithmic halo mass for galaxies, as a function of predicted halo mass, for a variety of testing data sets that were constructed using different halo occupation models than what was used in the training phase. All models shown use the `XGBoost` ML algorithm. Lines and shaded regions have the same meaning as in Fig. 4.3. Results for `HAM` and `DYN` are also shown for comparison (for the fiducial model case).

To probe the effect of velocity bias on the performance of the ML algorithms, we construct a few sets of the two testing mock catalogues, each time adopting the fiducial HOD model, but adding an amount of velocity bias between $f_{\rm vb}$ =0.9 and 1.1. We then apply our previously trained ML algorithms to these new test sets. Figure 4.6 shows the fractional difference Δf for these test cases compared, as always, to the `HAM` and `DYN` methods. We only show results for the `XGBoost` algorithm, but the other algorithms exhibit similar behaviour. The figure shows clearly that the performance of ML is almost entirely unaffected by velocity bias. This is to say that, regardless of the choice of $f_{\rm vb}$ in the testing catalogues, the predictions of halo mass made by ML algorithms that were trained on the fiducial model are not biased by this choice of parameters.

4.5.3 Varying the Luminosity-Mass relation

Having explored the impact of training ML models on data sets that assume incorrect relationships between the numbers and velocities of galaxies with halo mass, we now turn to assumptions about the mass-luminosity relation. This is potentially important since our feature selection procedure showed that a galaxy's luminosity and the luminosity of the brightest galaxy in its group are the two most important features for predicting halo mass. In our mock catalogues, we assign luminosities to galaxies using the *Conditional Luminosity Function* (CLF) formalism of Cacciato et al. (2009). Within the CLF model, the main parameter that controls the strength of the correlation between the mass of a halo and the luminosity of its central galaxy is $\sigma_{\log L}$, which is the scatter in the log of luminosity of central galaxies at fixed halo mass.[7] In the fiducial model that we used to train the ML algorithms, the value of this scatter is $\sigma_{\log L}$=0.142. To investigate the effect of applying the algorithms to data with different correlation between halo mass and luminosity, we construct sets of our two test catalogues that assume different values of $\sigma_{\log L}$, ranging from 0.1 to 0.3.

Figure 4.7 shows the fractional difference Δf for these test cases. As before, we only

[7] In Cacciato et al. (2009) this parameter was called σ_c.

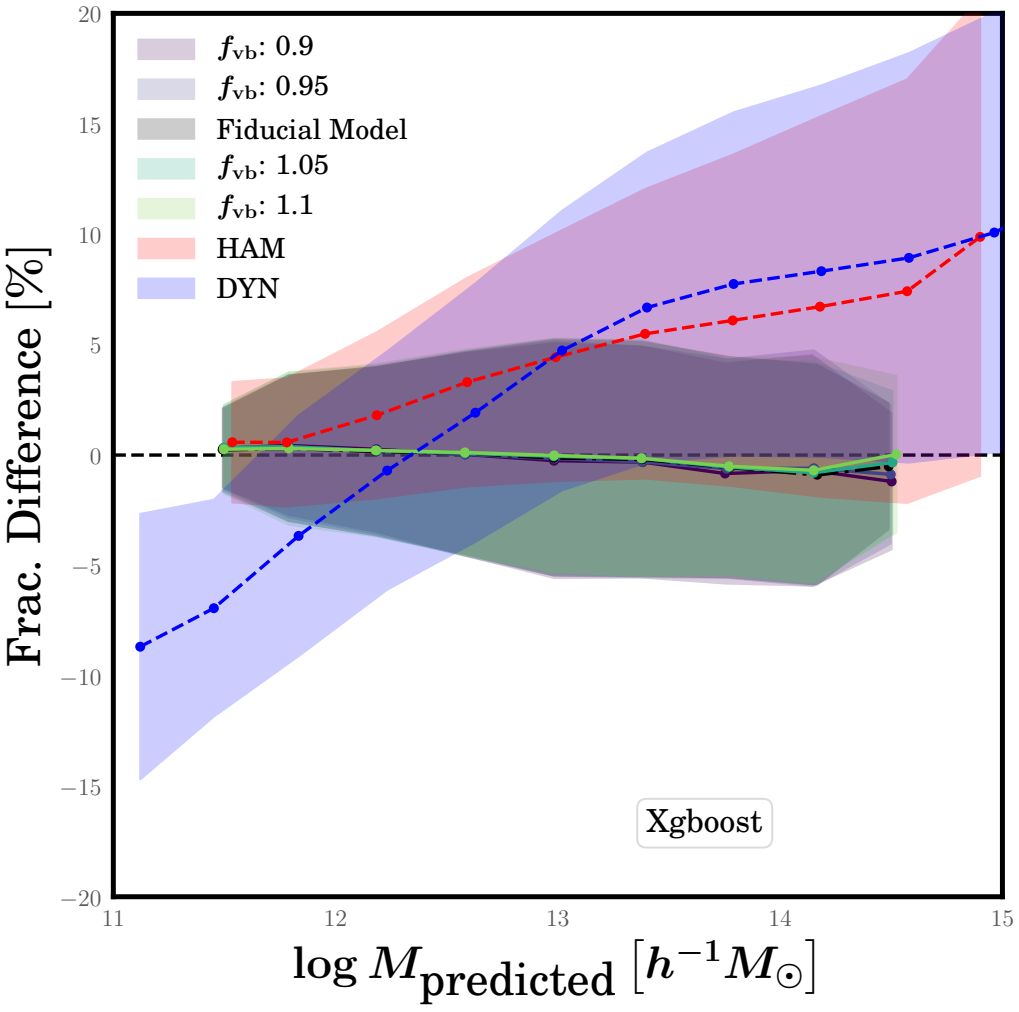

Figure 4.6: Similar to Fig. 4.5, except that the various testing data-sets now share the same set of halo occupation parameters as the training data, but cover a wide range of different values for satellite galaxy velocity bias, $f_{\rm vb}$.

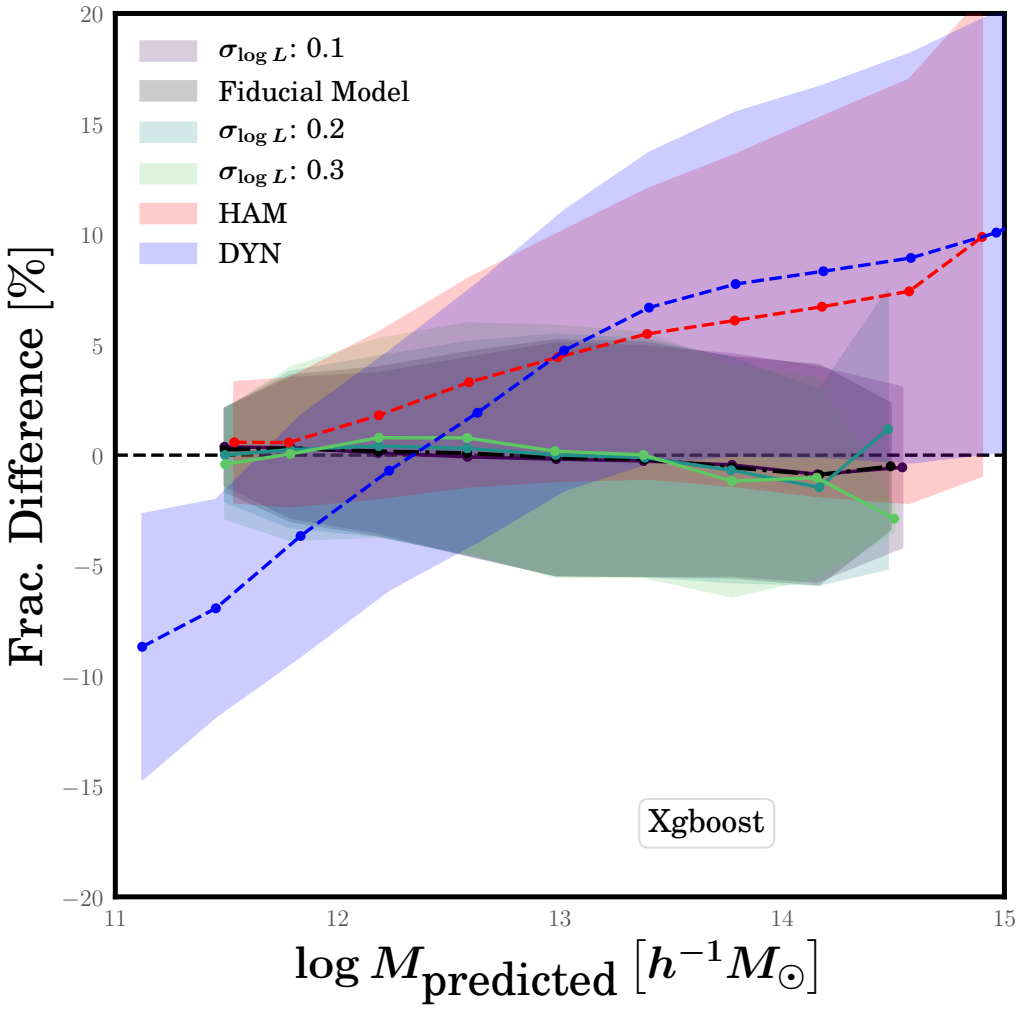

Figure 4.7: Similar to Figs. 4.5 and 4.6, except that the various testing data-sets now share the same set of halo occupation parameters as the training data and have no velocity bias, but cover a range of different values for the assumed scatter in the luminosity-mass relation for central galaxies, $\sigma_{\log L}$.

show results for the `XGBoost` algorithm and we include the results for `HAM` and `DYN` for comparison. The figure shows that the performance of ML algorithms is not affected much by the assumed value of $\sigma_{\log L}$. This is reassuring and implies that our halo mass predictions are not sensitive to the detailed form of the mass-luminosity relation.

4.6 Application to SDSS Galaxies

In §4.4 and §4.5, we showed how machine learning algorithms, such as `XGBoost`, `RF`, and `NN`, can be used to predict the mass of a galaxy's host halo with a higher accuracy on average than more conventional mass estimators, such as `HAM` and `DYN`. The next logical step is to choose the best of these algorithms and apply the trained model to *real* observed data. All three ML algorithms that we have explored perform very similarly so we choose `XGBoost` to be our algorithm of choice because it is faster than `RF` and `NN`. We apply the `XGBoost` model that we trained and tested on mock catalogues to the `Mr19`-SDSS catalogue, using the nine features described in §4.4.1 as inputs to the model. The model outputs a predicted halo mass, $M_{\rm pred}$, for each SDSS galaxy. We produce a final catalogue that includes the set of nine features for each galaxy in the sample, our value for $M_{\rm pred}$, and the `HAM` and `DYN` group mass estimates. The catalogue is available for download. [8]

Figure 4.8 shows the relationship between $M_{\rm pred}$ for SDSS galaxies and the masses from the `HAM` and `DYN` methods. The figure shows the two-dimensional histogram (blue shaded pixels) as well as the mean and standard deviation of $M_{\rm pred}$ in bins of $M_{\rm HAM}$ and $M_{\rm dyn}$ (yellow lines and error bars). In the case of `HAM`, Figure 4.8 shows that the masses predicted by `XGBoost` tend to be lower, on average, than those determined by `HAM` for all but the lowest $M_{\rm HAM}$ masses. This is in agreement with Figure 4.3, which showed that the masses determined by `HAM` tend to have larger Δf's than the $M_{\rm pred}$'s by `XGBoost` for $M_{\rm pred} > 10^{12}\,h^{-1}{\rm M}_{\odot}$. In the case of `DYN`, the `XGBoost` predicted masses are larger, on average, than those determined by `DYN` at small dynamical masses, but smaller for $M_{\rm dyn}$ larger than $M_{\rm dyn} > 10^{12}\,h^{-1}{\rm M}_{\odot}$. This is also in agreement with what we expect based on Figure 4.3. The qualitative agreement between these results from SDSS and what we found in our mock catalogue is encouraging.

Our tests with mock catalogues suggest that these predicted halo masses for SDSS galaxies may be significantly more accurate than those estimated using `HAM` or `DYN` meth-

[8] http://lss.phy.vanderbilt.edu/groups/ML_Catalogues/

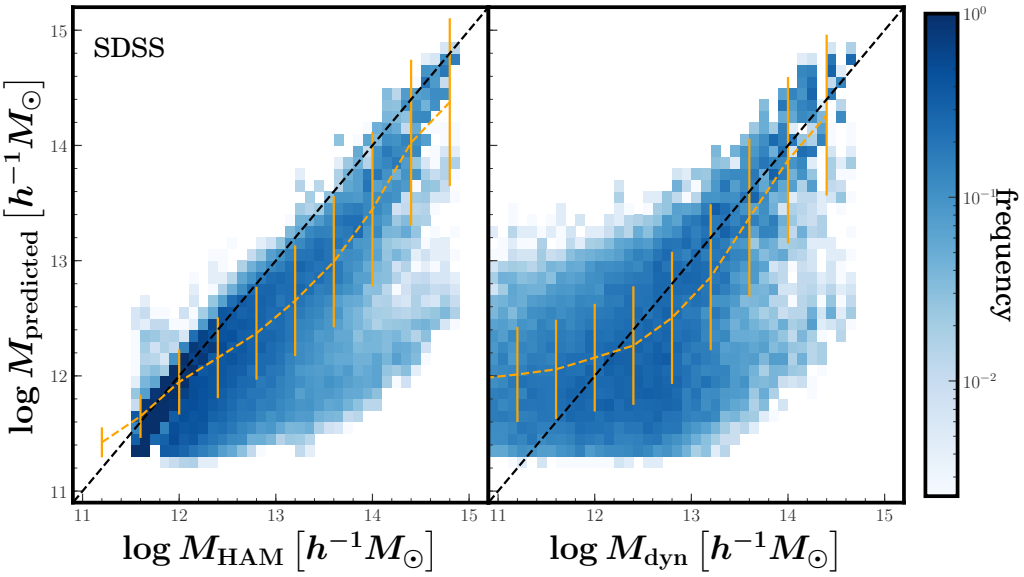

Figure 4.8: Galaxy halo masses for SDSS galaxies predicted by ML compared to traditional methods. The y-axis shows galaxy mass predictions from the `XGBoost` algorithm that was trained on mock catalogues. The x-axis shows mass estimates for galaxies through `HAM` (*left panel*) and `DYN` (*right panel*). The blue shading shows the frequency of galaxies in two-dimensional bins, where the number of galaxies in each bin is normalised by the value for the bin containing the most galaxies. Yellow solid lines and errorbars correspond to the mean and standard deviation of $M_{\rm pred}$ in bins of $M_{\rm HAM}$ or $M_{\rm dyn}$. The dashed black lines show the one-to-one relation between mass estimates.

ods, especially at large masses. Naturally, the worry with using these masses is the possibility that the real universe does not look like our training mock data in some critical way and that the predicted SDSS masses thus contain a large systematic error. Though this is certainly possible, it is not likely because the mock catalogues were constructed to have several statistical properties that are in agreement with the SDSS data. Moreover, `HAM` and `DYN` masses are known to have large systematic errors. We thus feel fairly confident that our ML halo masses are the best available measurements for galaxy halo environments in the SDSS and are safe to use.

4.7 Summary and Discussion

In this paper, we estimate halo masses of galaxies by employing machine learning (ML) techniques, and we compare these to results by other, more traditional, mass estima-

tion techniques, such as *Halo Abundance Matching* (HAM) and *Dynamical Mass Estimates* (DYN). We are motivated to explore ML because of limitations in these traditional methods and because we expect that we can obtain more precise halo mass estimates if we use information from all the galaxy properties that correlate with mass, such as luminosities, colours, group dynamics, and large-scale environments.

We investigate three ML algorithms: XGBoost, Random Forest (RF), and neural networks (NN). Each of the algorithms is trained on synthetic mock galaxy catalogues to predict the masses of galaxies' host halos, using a set of *features* selected from both galaxy- and group-related properties. The mock catalogues were constructed to have the same clustering and same distribution of *observed* properties as the SDSS data, such as luminosity, $(g-r)$ colour, and sSFR. The final set of nine features that we use (§4.4.1) are chosen based on their *feature importance* towards the overall prediction of halo mass, i.e., how much each feature contributes to the overall prediction of halo mass. To quantify the performance of the ML algorithms, we test them using an independent set of mock catalogues and we compare them to the HAM and DYN methods. We probe to what extent the trained ML models can be universally applied by testing them on data that have different properties from the training data. Specifically, we investigate variations in the halo occupation distribution (HOD), velocity bias for satellite galaxies, and the mass-luminosity relation for central galaxies. Finally, we apply our mock-trained XGBoost model to the Mr19-SDSS galaxy sample and produce a SDSS catalogue that contains predicted halo masses, as well as the nine features used and the HAM and DYN masses.

The main results of our work are as follows:

(i) We determine the set of nine features (out of the 19 features from §4.4.1) that contribute the most to the prediction of a galaxy's host halo mass. Among the set of nine features, we find that the two strongest features are the r-band absolute magnitude of the galaxy and the absolute magnitude of the brightest galaxy in the group to which the galaxy belongs. Following these are the $(g-r)$ colour and specific star formation rate of the galaxy and the

group as a whole, the size and velocity dispersion of the group, and the galaxy's distance to the nearest cluster.

(ii) We find that HAM and DYN overestimate halo masses on average for large $M_{\rm pred}$, reaching average fractional errors in $\log M$ as high as 10% at the highest masses. This is due to group-finding errors that misclassify some galaxies as satellites and thus assign them too large halo masses. At low $M_{\rm pred}$ HAM works well, but DYN underestimates galaxies' halo masses. In contrast, the ML algorithms all predict halo masses that are unbiased, on average, across the whole range of masses probed. To quantify the typical error in predicted halo mass, we calculate the quantity $(\Delta \log M)_{68}$, where 68% of galaxies have their masses predicted with an error less than this. The three trained ML models have values for this typical mass error of $0.23-0.25$ dex and $0.51-0.60$ dex for values of $M_{\rm pred}$ smaller or greater than $10^{12.5}\,h^{-1}{\rm M}_{\odot}$, respectively. On the other hand, HAM yields typical halo mass errors of 0.27 dex and 0.90 dex for the low-mass and high-mass regimes, respectively, while DYN can only recover halo masses to 0.92 dex and 1.25 dex for low and high masses.

(iii) When tested against mock data built with different assumptions than the training data, ML models mostly perform well. Results are insensitive to the presence of satellite galaxy velocity bias or the amount of scatter in the mass-luminosity relation for central galaxies. When we vary the relation between halo mass and occupation number, there is no effect at large masses, but predicted masses can be over-estimated in the low mass regime. However, ML predictions still outperform HAM and DYN

(iv) Predicted XGBoost halo masses for galaxies in the Mr19-SDSS sample are similar to HAM masses, but higher than DYN masses in the low mass regime, but smaller, on average, than HAM or DYN masses in the high mass regime. This is in qualitative agreement with our testing results on mock catalogues.

These results demonstrate the power of using ML algorithms to infer the *true* underlying

mass of a galaxy's dark matter halo. Spectrophotometric properties of galaxies and their groups, dynamical properties of the groups, and large scale environments, all correlate with halo mass in different ways. It is thus not surprising that, when used jointly, they deliver tighter constraints on halo mass than any one method. Our results confirm this, especially at large masses, where methods like `HAM` and `DYN` suffer from the standard group-finding errors that mistakenly place some field galaxies into large groups.

The big caveat to these results is that they only hold to the extent that the mock catalogues used to train the ML algorithms match the real universe. We have taken care to make sure that our mock galaxies have distributions of observed properties and clustering that are consistent with those in the SDSS. However, we cannot guarantee that the correlations between these properties and halo mass are correct in the training data. Though our tests modifying the galaxy-halo connection are encouraging, we have not explored the whole possible space of mock catalogues. Readers are advised to use the SDSS predicted masses in §4.6 at their own discretion.

Perhaps the most interesting implication of this paper is the possibility that we can use ML approaches to eliminate some of the systematic issues with the group-finding process, such as merging of galaxies from different host haloes into the same group, or the splitting of galaxies from the same halo into several different galaxy groups. For example, galaxies in the same group that have very discrepant ML-predicted halo masses may have been incorrectly grouped together. We plan to explore this in future work.

Chapter 5

CONCLUSIONS

In this dissertation, I have presented several different but related analyses on various aspects of the galaxy-halo connection.

First, I constructed a set of volume-limited galaxy samples from the Sloan Digital Sky Survey (SDSS) Data Release 7 (DR7). Moreover, I presented a set of galaxy group catalogs for three different volume-limited samples from SDSS, along with analogous realistic mock galaxy group catalogs. These catalogs have been made publicly available for consumption. The suite of galaxy and group catalogues contain information about several different galaxy properties, as well as that related to galaxies' group environments. Additionally, I also investigated the impact that group-finding errors have on inferred statistics from galaxy group catalogs by performing a careful analysis using mock galaxy (group) catalogs. In this analyses I showed that group-finding errors do indeed have an impact of the estimation of group mass and galaxy type. Additionally, I also made use of this framework to study the stellar-halo mass relation of central galaxies, and computed a correction to SDSS that I determined through the use of mock catalogs. I also explored what the role of group mass is for determining the galaxy's quenching state as a function of galaxy stellar mass. I found that group-finding errors do not affect this relation significantly, and one can use these mocks to further constrain this relation.

Secondly, I performed a comprehensive study of "galactic conformity" at low redshift using a galaxy group catalog from SDSS DR7 and their satellites (1-halo), and between central galaxies in separate haloes (2-halo). I used two metrics to probe for conformity in three galaxy properties, $(g-r)$ colours, specific star formation rate sSFR, and morphology of galaxies. I also assessed the statistical significance of conformity signals with mock galaxy catalogs from LasDamas simulation, and was able to make the first robust detection

of 2-halo conformity.

At last, I presented a machine learning approach (ML) for the prediction of galaxies' dark matter halo masses that achieved an improved performance over conventional methods. I trained three different ML algorithms to predict halo masses using a set of realistic mock galaxy catalogs. I used these mock catalogs to explore how the choice of different model parameters affected the predicted masses, and found that the ML approach still yielded substantially better mass estimates than those of conventional methods, even when modifying our choices of model parameters. I ultimately applied the trained model to a galaxy group catalog from SDSS and presented the resulting halo masses.

In conclusion, I have explored various aspects of the galaxy-halo connection and determined that galaxy group catalogs are important tools that allow us to statistically measure how galaxies correlate with their host haloes. What I present in this dissertation are some examples of how group catalogs can efficiently characterize the galaxy-halo connection, and can provide us with a better insight into how galaxy properties depend of cosmic structure. Moreover, the next generation of astronomical surveys, such as the Large Synoptic Survey Telescope (LSST) and Euclid will provide an immense amount of data in the next decades. Analyzing these data will require novel approaches and techniques and will result in a thorough understanding of the Universe. It will also allow to better understand our place in the Universe.

REFERENCES

Abazajian K., Adelman-McCarthy J. K., Agüeros M. A., et al., 2004, The Astronomical Journal, 128, 1, 502

Abazajian K. N., Adelman-McCarthy J. K., Agüeros M. A., et al., 2009, The Astrophysical Journal Supplement Series, 182, 2, 543

Abell G. O., 1958, The Astrophysical Journal Supplement Series, 3, 211

Ade P. A. R., Aghanim N., Armitage-Caplan C., et al., 2015, Astronomy & Astrophysics, 581, A14

AdelmanâĂŘMcCarthy J. K., Agueros M. A., Allam S. S., et al., 2007, The Astrophysical Journal Supplement Series, 172, 2, 634

Aihara H., Allende Prieto C., An D., et al., 2011, The Astrophysical Journal Supplement Series, 193, 2, 29

Alam S., Albareti F. D., Prieto C. A., et al., 2015, The Astrophysical Journal Supplement Series, 219, 1, 12

Allen S. W., Evrard A. E., Mantz A. B., 2011, Annual Review of Astronomy and Astrophysics, 49, 1, 409

Ann H. B., Park C., Choi Y. Y., 2008, Monthly Notices of the Royal Astronomical Society, 389, 1, 86

Armitage T. J., Kay S. T., Barnes D. J., 2019, Monthly Notices of the Royal Astronomical Society, 484, 2, 1526

Armitage T. J., Kay S. T., Barnes D. J., Bahé Y. M., Vecchia C. D., 2018, MNRAS, 000, 1

Ascaso B., Wittman D., Benítez N., 2012, Monthly Notices of the Royal Astronomical Society, 420, 2, 1167

Ball N. M., Brunner R. J., Myers A. D., et al., 2007, The Astrophysical Journal, 663, 2, 774

Balogh M. L., Navarro J. F., Morris S. L., 2000, The Astrophysical Journal, 540, 1, 113

Banerji M., Lahav O., Lintott C. J., et al., 2010, Monthly Notices of the Royal Astronomical Society, 406, 1, 342

Barnes D. J., Kay S. T., Bahé Y. M., et al., 2017, Monthly Notices of the Royal Astronomical Society, 471, 1, 1088

Behroozi P., Wechsler R., Hearin A., Conroy C., 2018, The Economic History Review, 66, 3, 715

Behroozi P. S., Conroy C., Wechsler R. H., 2010, The Astrophysical Journal, 717, 1, 379

Behroozi P. S., Wechsler R. H., Wu H.-Y., 2013, The Astrophysical Journal, 762, 2, 109

Beisbart C., Kerscher M., 2000, The Astrophysical Journal, 545, 1, 6

Bell E. F., McIntosh D. H., Katz N., Weinberg M. D., 2003, The Astrophysical Journal Supplement Series, 149, 2, 289

Bell E. F., Wolf C., Meisenheimer K., et al., 2004, The Astrophysical Journal, 608, 2, 752

Berlind A. A., Frieman J., Weinberg D. H., et al., 2006, The Astrophysical Journal Supplement Series, 167, 1, 1

Berlind A. A., Weinberg D. H., 2002, The Astrophysical Journal, 575, 2, 587

Berlind A. A., Weinberg D. H., Benson A. J., et al., 2003, The Astrophysical Journal, 593, 1, 1

Berti A. M., Coil A. L., Behroozi P. S., et al., 2017, The Astrophysical Journal, 834, 1, 87

Blanton M. R., Bershady M. A., Abolfathi B., et al., 2017, The Astronomical Journal, 154, 1, 28

Blanton M. R., Hogg D. W., Brinkmann J., et al., 2003, The Astrophysical Journal, 592, 2, 819

Blanton M. R., Schlegel D. J., Strauss M. A., et al., 2005, The Astronomical Journal, 129, 6, 2562

Bray A. D., Pillepich A., Sales L. V., et al., 2015, Monthly Notices of the Royal Astronomical Society, 455, 1, 185

Breiman L., 2001, Machine Learning, 45, 1, 5

Breiman L., Friedman J., Stone C., Olshen R., 1984, Classification and Regression Trees, Taylor & Francis

Brinchmann J., Charlot S., White S. D. M., et al., 2004, Monthly Notices of the Royal Astronomical Society, 351, 4, 1151

Brodwin M., Ruel J., Ade P. A. R., et al., 2010, The Astrophysical Journal, 721, 1, 90

Cacciato M., Van Den Bosch F. C., More S., Li R., Mo H. J., Yang X., 2009, Monthly Notices of the Royal Astronomical Society, 394, 2, 929

Calderon V. F., Berlind A. A., 2019, submitted to MNRAS

Calderon V. F., Berlind A. A., Sinha M., 2018, Monthly Notices of the Royal Astronomical Society, 480, 2, 2031

Campbell D., van den Bosch F. C., Hearin A., et al., 2015, Monthly Notices of the Royal Astronomical Society, 452, 1, 444

Campbell L. E., 2015, Towards the Unbound Stellar Population in the Sloan Digital Sky Survey, Ph.D. thesis, The Vanderbilt University

Carlberg R. G., Yee H. K. C., Ellingson E., 1997, The Astrophysical Journal, 478, 2, 462

Chen J., Kravtsov A. V., Prada F., et al., 2006, The Astrophysical Journal, 647, 1, 86

Coil A. L., Blanton M. R., Burles S. M., et al., 2011, The Astrophysical Journal, 741, 1, 8

Colless M., Dalton G., Maddox S., et al., 2001, Monthly Notices of the Royal Astronomical Society, 328, 4, 1039

Colless M., Dunn A. M., 1996, The Astrophysical Journal, 458, 435

Conroy C., Wechsler R. H., Kravtsov A. V., 2006, The Astrophysical Journal, 647, 1, 201

Cool R. J., Moustakas J., Blanton M. R., et al., 2013, The Astrophysical Journal, 767, 2, 118

Cooray A., Milosavljević M., 2005, The Astrophysical Journal, 627, 2, L85

Crook A. C., Huchra J. P., Martimbeau N., Masters K. L., Jarrett T., Macri L. M., 2007, The Astrophysical Journal, 655, 2, 790

Croton D. J., Gao L., White S. D. M., 2007, Monthly Notices of the Royal Astronomical Society, 374, 4, 1303

Davis M., Efstathiou G., Frenk C. S. S., White S. D. M. D. M., 1985, The Astrophysical Journal, 292, 371

Doi M., Tanaka M., Fukugita M., et al., 2010, The Astronomical Journal, 139, 4, 1628

Dressler A., 1980, The Astrophysical Journal, 236, 351

Duarte M., Mamon G. A., 2014, Monthly Notices of the Royal Astronomical Society, 440, 2, 1763

Dubinski J., 1998, The Astrophysical Journal, 502, 1, 141

Eckert K. D., Kannappan S. J., Stark D. V., et al., 2015, Astrophysical Journal, 810, 2, 166

Einasto J., Einasto M., Tago E., et al., 2007, Astronomy and Astrophysics, 462, 2, 811

Eisenstein D. J., Weinberg D. H., Agol E., et al., 2011, The Astronomical Journal, 142, 3, 72

Eke V. R., Baugh C. M., Cole S., et al., 2004, Monthly Notices of the Royal Astronomical Society, 348, 3, 866

Fadda D., Girardi M., Iuricin G., Mardirossian F., Mezzetti M., 1996, The Astrophysical Journal, 473, 2, 670

Gao L., Springel V., White S. D. M., 2005, Monthly Notices of the Royal Astronomical Society: Letters, 363, 1, L66

Geller M. J. M., Huchra J. P., 1983, The Astrophysical Journal Supplement . . . , 52, 61

Gerdes D. W., Sypniewski A. J., McKay T. A., et al., 2010, The Astrophysical Journal, 715, 2, 823

Gerke B. F., Newman J. A., Davis M., et al., 2005, The Astrophysical Journal, 625, 1, 6

Girardi M., Giuricin G., Mardirossian F., Mezzetti M., Boschin W., 1998, The Astrophysical Journal, 505, 1, 74

Gladders M. D., Yee H. K. C., 2005, The Astrophysical Journal Supplement Series, 157, 1, 1

Goto T., 2005, Monthly Notices of the Royal Astronomical Society, 359, 4, 1415

Grebel E. K., Gallagher III J. S., Harbeck D., 2003, The Astronomical Journal, 125, 4, 1926

Gu M., Conroy C., Behroozi P., 2016, The Astrophysical Journal, 833, 1, 2

Gunn J. E., Carr M., Rockosi C., et al., 1998, The Astronomical Journal, 116, 6, 3040

Gunn J. E., Siegmund W. A., Mannery E. J., et al., 2006, The Astronomical Journal, 131, 4, 2332

Guo H., Zheng Z., Zehavi I., et al., 2015, Monthly Notices of the Royal Astronomical Society, 453, 4, 4369

Guo Q., White S., Boylan-Kolchin M., et al., 2011, Monthly Notices of the Royal Astronomical Society, 413, 1, 101

Hamilton A. J. S., 1998, in The Evolving Universe, 185–275, Kluwer Academic

Hao J., McKay T. A., Koester B. P., et al., 2010, Astrophysical Journal, Supplement Series, 191, 2, 254

Hartley W. G., Conselice C. J., Mortlock A., Foucaud S., Simpson C., 2015, Monthly Notices of the Royal Astronomical Society, 451, 2, 1613

Hearin A. P., Behroozi P. S., van den Bosch F. C., 2016, Monthly Notices of the Royal Astronomical Society, 461, 2, 2135

Hearin A. P., Watson D. F., 2013, Monthly Notices of the Royal Astronomical Society, 435, 2, 1313

Hearin A. P., Watson D. F., van den Bosch F. C., 2015, Monthly Notices of the Royal Astronomical Society, 452, 2, 1958

Huchra J. P., 1988, in The Minnesota lectures on Clusters of Galaxies and Large-Scale Structure, edited by J. M. Dickey, vol. 5 of Astronomical Society of the Pacific Conference Series, 41–70

Huchra J. P., Geller M. J., 1982, The Astrophysical Journal, 257, 423

Kaiser N., 1987, Monthly Notices of the Royal Astronomical Society, 227, 1, 1

Kamdar H. M., Turk M. J., Brunner R. J., 2016a, Monthly Notices of the Royal Astronomical Society, 455, 1, 642

Kamdar H. M., Turk M. J., Brunner R. J., 2016b, Monthly Notices of the Royal Astronomical Society, 457, 2, 1162

Kauffmann G., Heckman T. M., White S. D. M., et al., 2003, Monthly Notices of the Royal Astronomical Society, 341, 1, 33

Kauffmann G., Li C., Heckman T. M., 2010, Monthly Notices of the Royal Astronomical Society, 409, 2, 491

Kauffmann G., Li C., Zhang W., Weinmann S., 2013, Monthly Notices of the Royal Astronomical Society, 430, 2, 1447

Kauffmann G., White S. D. M., Heckman T. M., et al., 2004, Monthly Notices of the Royal Astronomical Society, 353, 3, 713

Kawinwanichakij L., Quadri R. F., Papovich C., et al., 2016, The Astrophysical Journal, 817, 1, 9

Knobel C., Lilly S. J., Woo J., Kovač K., 2015, The Astrophysical Journal, 800, 1, 24

Kravtsov A. V., Berlind A. A., Wechsler R. H., et al., 2004, The Astrophysical Journal, 609, 1, 35

Kravtsov A. V., Borgani S., 2012, Annual Review of Astronomy and Astrophysics, 50, 1, 353

Lacerna I., Contreras S., González R. E., Padilla N., Gonzalez-Perez V., 2018, Monthly Notices of the Royal Astronomical Society, 475, 1, 1177

Lavaux G., Hudson M. J., 2011, Monthly Notices of the Royal Astronomical Society, 416, 4, 2840

Lawrence A., Warren S. J., Almaini O., et al., 2007, Monthly Notices of the Royal Astronomical Society, 379, 4, 1599

Leauthaud A., Tinker J., Bundy K., et al., 2012, The Astrophysical Journal, 744, 2, 159

Lim S. H., Mo H. J., Lu Y., Wang H., Yang X., 2017, Monthly Notices of the Royal Astronomical Society, 470, 3, 2982

Lim S. H., Mo H. J., Wang H., Yang X., 2018, Monthly Notices of the Royal Astronomical Society, 480, 3, 4017

Mahabal A., Djorgovski S., Turmon M., et al., 2008, Astronomische Nachrichten, 329, 3, 288

Mantz A. B., Allen S. W., Morris R. G., et al., 2014, Monthly Notices of the Royal Astronomical Society, 440, 3, 2077

Marriage T. A., Acquaviva V., Ade P. A., et al., 2011, Astrophysical Journal, 737, 2, 61

Martinez H. J., O'Mill A. L., Lambas D. G., 2006, Monthly Notices of the Royal Astronomical Society, 372, 1, 253

Martinez V. J., Arnalte-Mur P., Stoyan D., 2010, Http://Arxiv.Org/Abs/1001.1294, 22, 6

McBride C., Berlind A., Scoccimarro R., et al., 2009, American Astronomical Society Meeting Abstracts #213, 41, 425.06

McCarthy J. K., Agueros M. A., Allam S. S., et al., 2006, The Astrophysical Journal Supplement Series, 162, 1, 38

McCracken H. J., Milvang-Jensen B., Dunlop J., et al., 2012, Astronomy & Astrophysics, 544, A156

Merchán M., Zandivarez A., 2002, Monthly Notices of the Royal Astronomical Society, 335, 1, 216

Moffett A. J., Kannappan S. J., Berlind A. A., et al., 2015, The Astrophysical Journal, 812, 2, 89

More S., van den Bosch F. C., Cacciato M., Mo H. J., Yang X., Li R., 2009, Monthly Notices of the Royal Astronomical Society, 392, 2, 801

Moster B. P., Somerville R. S., Maulbetsch C., et al., 2010, The Astrophysical Journal, 710, 2, 903

Ntampaka M., Trac H., Sutherland D. J., Battaglia N., Póczos B., Schneider J., 2015, The Astrophysical Journal, 803, 2, 50

Ntampaka M., Trac H., Sutherland D. J., Fromenteau S., Póczos B., Schneider J., 2016, The Astrophysical Journal, 831, 2, 135

Ntampaka M., ZuHone J., Eisenstein D., et al., 2018, arXiv e-prints

Old L., Skibba R. A., Pearce F. R., et al., 2014, Monthly Notices of the Royal Astronomical Society, 441, 2, 1513

Paranjape A., Kovač K., Hartley W. G., Pahwa I., 2015, Monthly Notices of the Royal Astronomical Society, 454, 3, 3030

Pedregosa F., Varoquaux G., Gramfort A., et al., 2012, Journal of Machine Learning Research, 12, 2825

Perlmutter S., Turner M. S., White M., 1999, Physical Review Letters, 83, 4, 670

Phillips J. I., Wheeler C., Boylan-Kolchin M., Bullock J. S., Cooper M. C., Tollerud E. J., 2014a, Monthly Notices of the Royal Astronomical Society, 437, 2, 1930

Phillips J. I., Wheeler C., Cooper M. C., Boylan-Kolchin M., Bullock J. S., Tollerud E., 2014b, Monthly Notices of the Royal Astronomical Society, 447, 1, 698

Planck Collaboration, 2016, Astronomy & Astrophysics, 13

Postman M., Geller M. J., 1984, The Astrophysical Journal, 281, 8, 95

Prescott M., Baldry I. K., James P. A., et al., 2011, Monthly Notices of the Royal Astronomical Society, 417, 2, 1374

Reddick R. M., Wechsler R. H., Tinker J. L., Behroozi P. S., 2013, The Astrophysical Journal, 771, 1, 30

Riccio G., Cavuoti S., Schisano E., et al., 2016, in Advances in Neural Networks, edited by S. Bassis, A. Esposito, F. C. Morabito, E. Pasero, 27–36, Springer International Publishing, Cham

Riess A. G., Filippenko A. V., Challis P., et al., 1998, The Astronomical Journal, 116, 3, 1009

Rines K., Geller M. J., Diaferio A., 2010, Astrophysical Journal Letters, 715, 2 PART 2, L180

Rosati P., Borgani S., Norman C., 2002, Annual Review of Astronomy and Astrophysics, 40, 1, 539

Ross A. J., Brunner R. J., 2009, Monthly Notices of the Royal Astronomical Society, 399, 2, 878

Ruel J., Bazin G., Bayliss M., et al., 2014, Astrophysical Journal, 792, 1, 45

Salcedo A. N., Maller A. H., Berlind A. A., et al., 2017, eprint arXiv:1708.08451

Salim S., Rich R. M., Charlot S., et al., 2007, The Astrophysical Journal Supplement Series, 173, 2, 267

Sheth R. K., Connolly A. J., Skibba R., 2005, ArXiv e-prints, 13, February, 13

Sifón C., Menanteau F., Hasselfield M., et al., 2013, The Astrophysical Journal, 772, 1, 25

Sin L. P. T., Lilly S. J., Henriques B. M. B., 2017, Monthly Notices of the Royal Astronomical Society, 471, 1, 1192

Sinha M., Berlind A. A., McBride C. K., Scoccimarro R., Piscionere J. A., Wibking B. D., 2018, Monthly Notices of the Royal Astronomical Society, 478, 1, 1042

Skibba R., Sheth R. K., Connolly A. J., Scranton R., 2006, Monthly Notices of the Royal Astronomical Society, 369, 1, 68

Skrutskie M. F., Cutri R. M., Stiening R., et al., 2006, The Astronomical Journal, 131, 2, 1163

Smee S. A., Gunn J. E., Uomoto A., et al., 2013, The Astronomical Journal, 146, 2, 32

Staniszewski Z., Ade P. A. R., Aird K. A., et al., 2009, Astrophysical Journal, 701, 1, 32

Tago E., Einasto J., Saar E., et al., 2006, Astronomische Nachrichten, 327, 4, 365

Tasitsiomi A., Kravtsov A. V., Wechsler R. H., Primack J. R., 2004, The Astrophysical Journal, 614, 2, 533

Teague P. F., Carter D., Gray P. M., 1990, The Astrophysical Journal Supplement Series, 72, 715

Tinker J. L., Hahn C., Mao Y.-Y., Wetzel A. R., Conroy C., 2018, Monthly Notices of the Royal Astronomical Society, 477, 1, 935

Tinker J. L., Leauthaud A., Bundy K., et al., 2013, The Astrophysical Journal, 778, 2, 93

Treyer M., Kraljic K., Arnouts S., et al., 2017, MNRAS, 000, 0

Tucker D. L., Hashimoto Y., Kirshner R. P., et al., 1997, Astronomische Nachrichten, 321, 2, 101

Vale A., Ostriker J. P., 2004, Monthly Notices of the Royal Astronomical Society, 353, 1, 189

Van Den Bosch F. C., Yang X., Mo H. J., 2003, Monthly Notices of the Royal Astronomical Society, 340, 3, 771

Vikhlinin A., Kravtsov A. V., Burenin R. A., et al., 2009, Astrophysical Journal, 692, 2, 1060

Vogelsberger M., Genel S., Springel V., et al., 2014, Monthly Notices of the Royal Astronomical Society, 444, 2, 1518

Voit G. M., 2005, Reviews of Modern Physics, 77, 1, 207

Wang J., Serra P., Józsa G. I. G., et al., 2015, Monthly Notices of the Royal Astronomical Society, 453, 3, 2400

Wang W., White S. D. M., 2012, Monthly Notices of the Royal Astronomical Society, 424, 4, 2574

Warren M. S., Abazajian K., Holz D. E., Teodoro L., 2006, The Astrophysical Journal, 646, 2, 881

Watson D. F., Conroy C., 2013, The Astrophysical Journal, 772, 2, 139

Wechsler R. H., Tinker J. L., 2018, Annu. Rev. Astron. Astrophys, AA, 1

Wechsler R. H., Zentner A. R., Bullock J. S., Kravtsov A. V., Allgood B., 2005, The Astrophysical Journal, 652, 1, 71

Weinberg D. H., Mortonson M. J., Eisenstein D. J., Hirata C., Riess A. G., Rozo E., 2013, Physics Reports, 530, 2, 87

Weinmann S. M., Van Den Bosch F. C., Yang X., Mo H. J., 2006, Monthly Notices of the Royal Astronomical Society, 366, 1, 2

Wojtak R., Old L., Mamon G. A., et al., 2018, Monthly Notices of the Royal Astronomical Society, 481, 1, 324

Xu X., Ho S., Trac H., Schneider J., Poczos B., Ntampaka M., 2013, Astrophysical Journal, 772, 2, 147

Yang X., Mo H. J., Bosch F. C. v. d., Zhang Y., Han J., 2012, The Astrophysical Journal, 752, 41

Yang X., Mo H. J., Van den Bosch F. C., 2003, Monthly Notices of the Royal Astronomical Society, 339, 4, 1057

Yang X., Mo H. J., van den Bosch F. C., 2009, The Astrophysical Journal, 693, 1, 830

Yang X., Mo H. J., van den Bosch F. C., Jing Y. P., 2005, Monthly Notices of the Royal Astronomical Society, 356, 4, 1293

Yang X., Mo H. J., van den Bosch F. C., Pasquali A., Li C., Barden M., 2007, The Astrophysical Journal, 671, 1, 153

York D. G., 2000, The Astronomical Journal, 120, 3, 1579

Zehavi I., Zheng Z., Weinberg D. H., et al., 2005, The Astrophysical Journal, 630, 1, 1

Zehavi I., Zheng Z., Weinberg D. H., et al., 2010, The Astrophysical Journal, 59, 1, 35

Zehavi I., Zheng Z., Weinberg D. H., et al., 2011, The Astrophysical Journal, 736, 1, 59

Zhang Y.-C., Yang X.-H., 2019, Research in Astronomy and Astrophysics, 19, 1, 006

Zu Y., Mandelbaum R., 2015, Monthly Notices of the Royal Astronomical Society, 454, 2, 1161

Zu Y., Mandelbaum R., 2016, Monthly Notices of the Royal Astronomical Society, 457, 4, 4360

Zu Y., Mandelbaum R., 2018, Monthly Notices of the Royal Astronomical Society, 476, 2, 1637

Zwicky F., Herzog E., Wild P., 1968, Catalogue of Galaxies and of Clusters of Galaxies, vol. 4, California Institute Of Technology