On Dark Matter Halos and Their Connection to Galaxies

by

Kuan Wang

Bachelor of Science, University of Science and Technology of China, 2015

Submitted to the Graduate Faculty of

the Kenneth P. Dietrich School of Arts and Sciences in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2021

UNIVERSITY OF PITTSBURGH

KENNETH P. DIETRICH SCHOOL OF ARTS AND SCIENCES

This dissertation was presented

by

Kuan Wang

It was defended on

July 8th 2021

and approved by

Andrew Zentner, Department of Physics and Astronomy, University of Pittsburgh

Brian Batell, Department of Physics and Astronomy, University of Pittsburgh

Arthur Kosowsky, Department of Physics and Astronomy, University of Pittsburgh

Rachel Mandelbaum, Department of Physics, Carnegie Mellon University

Jeffrey Newman, Department of Physics and Astronomy, University of Pittsburgh

On Dark Matter Halos and Their Connection to Galaxies

Kuan Wang, PhD

University of Pittsburgh, 2021

In the concordance model of modern cosmology, dark matter is five times as abundant as ordinary matter. While its nature remains one of the most challenging questions in today's physics, dark matter has been established as a defining factor in the large-scale structure. Visible galaxies form in the potential wells of dark matter density peaks, known as halos. Empirical galaxy–halo connection models, which reconstruct the observable components of the Universe from theory of the dark sector, are broadly used for their simplicity and effectiveness. The fundamental premise of empirical models is the statistical dependence of galaxy properties on halo properties, the latter of which are easily accessible through simulations.

With tremendous amounts of data being produced by new surveys, theoretical tools need also be further developed to exploit the full potential of data. In particular, small-scale observables, which require detailed knowledge of halos and the connection between galaxies and halos, are a promising source of information for constraining cosmology and galaxy physics. It is urgent and important in the new era of precision cosmology to improve models of these factors.

This thesis aims to improve our understanding of dark matter halo evolution and the dependence of galaxies on the halos in which they reside. In the first part, we investigate how the present-day halo structure emerges from the halo mass assembly history, and characterize respective contributions from pseudo-evolution and physical merger events. We uncover the significant impact of mergers on the evolution of halo structure, and recognize universal patterns in mergers. These findings will also shed light on the galaxy evolution in halos. In the second part, we test the validity of the simplifying assumptions adopted in galaxy–halo connection models. We identify the optimal combination of observable statistics that contain the most information on the galaxy–halo connection, and obtain observational constraints on the model using these statistics. We observationally confirm that the inclusion of galaxy count statistics significantly improves the constraining power, and find definitive evidence that the galaxy–halo connection depend on secondary halo properties besides mass. These results inform the physics of galaxy formation and evolution and cosmological inferences.

Table of Contents

List of Tables

List of Figures

Acknowledgements

First and foremost, I am grateful to my advisor, Andrew Zentner, for his guidance and support in the past six years. Andrew has been incredibly patient, understanding, and encouraging, for the entire time we have known each other. Despite my many faults, I have never heard one harsh word from him, and he went out of his way to be kind and caring. He is always optimistic and cheerful, and also exceptionally fun to work with. In working with Andrew, I have become a more confident and happy person. He also taught me how to be a good researcher, a good writer, and a good colleague.

I am equally indebted to Yao-Yuan Mao, who has been my mentor and close collaborator. Yao showed me not just what a postdoc should be, but also what a scientist should be, with his extraordinary academic standards and leadership capabilities. How he managed to be always accessible and reliable while accomplishing so many other things remains a mystery to me, and I wish I could be like him one day.

I thank my other collaborators, especially Frank van den Bosch, Johannes Lange, Andrew Hearin, Chad Schafer, Risa Wechsler, Hong Guo, and Elisa Chisari. They have provided me with much help, and their perspectives are a valuable source of scientific inspiration. Because of them, I will also do my best to help other, more junior scientists.

I thank my committee members, Brian Batell, Arthur Kosowsky, Rachel Mandelbaum, and Jeffrey Newman, as well as other faculty and staff members in our department, especially Brett Andrew, Carlos Badenes, Rachel Bezanson, Dan Boyanovsky, Bob Breze, Ayres Freitas, Gracie Gollinger, John Hillier, Leyla Hirschfield, Hrvoje Petek, and Michael Wood-Vasey, for giving me advice and assistance, and making the department a welcoming, friendly place.

I thank my past and current fellow students, especially Runyu Bi, Hongbo Cai, Christine Mazzola Daher, Biprateep Dey, Daniel Doucette, Cat Fielder, Yilun Guan, Nathan Herring, Yang Hu, Nafis Karim, Yasha Kaushal, Andi Li, Zehan Li, Hongkai Liu, Lorena Mezini, Alan Pearl, Daniel Perrefort, Kara Ponder, Troy Raen, David Setton, Hang Su, Juntong Su, Harsh Vashistha, Antonio Villareal, Kevin Wilk, Xinyi Wu, and Rongpu Zhou, for being good friends.

I thank my other friends, who have provided emotional support for me throughout the years,

especially Xiaoheng Cheng, Hongchang Gao, Xiaojiang Guo, Fei Huang, Mengqiao Li, Dazhuang Mao, Edinson Najera, Giacomo Nebbia, Chen Peng, Zhaoyang Qiu, Xiaoshan Sun, Miya Sylvester, Bin Wang, Qing Wang, Yunmiao Wang, Lian Xia, Yang Yang, Chuanping Yu, Jiachen Zhang, and Qiong Zhang.

I thank Cynthia Carlson, Teresa Blum, Jeanne, Bill, Emily, and Tim Griffith, among other people who have been kind to me and given me help when I was in need. Sometimes I find it hard to believe how much kindness I am met with, wherever I go. I will work hard to pass that kindness on to others as I live.

Finally, I would like to thank my grandparents and parents for their love. They raised me to be who I am today, and never asked for anything in return. They have given up the time we could have spent together, in support of my career, and have always stood behind me. I look forward to returning to China to be reunited with them.

Dedication

Dedicated to the loving memory of my grandfather, Dai Qifan, whom I lost on May 6th, 2017. His life was one of integrity and sacrifice, devoted to his country and family. His love was richer than his tongue, most tender, most unselfish, and unfailing from the day I was born. I shall always remember him with pride and love.

I. Introduction

A. Background

1. Lambda-Cold Dark Matter Cosmology

The past century has witnessed dramatic development in our understanding of the Universe[1]. In 1925, Edwin Hubble first showed that the Universe extends far beyond our Milky Way. In 1929, he found that galaxies recede from us with velocities linearly related to their distances, which served as the definitive evidence that the Universe is not static, but expanding [78]. The subsequent observation of the cosmic microwave background radiation and cosmic element abundances in the 1960s established the Hot Big Bang model as the standard model of cosmology. In 1981, Alan Guth further proposed the inflation scenario [69], which solved the horizon problem and the flatness problem in the model.

In the framework of the Big Bang model, the fate of the Universe is determined by its matter and energy content. Following Fritz Zwicky's study of the velocities of galaxies in the Coma Cluster in 1933 [233], observations of satellite galaxy kinematics and galaxy rotation curves convinced the community of the existence of dark matter in the 1970s. In the 1980s and 1990s, the cold dark matter (CDM) paradigm, where dark matter consists of massive exotic particles, together with the introduction of dark energy, which drives the expansion of space, became the most widely accepted description of the Universe by successfully predicting numerous observational phenomena. While the fundamental nature of either dark matter or dark energy remains unknown, observation shows that they are the major components of the Universe. Ref. [142] inferred that dark energy makes up approximately 68.7% of the total energy density, dark matter approximately 26.4%, whereas ordinary baryonic matter only contributes approximately 4.9%.

This concordance model is known as the "ΛCDM" model, where Λ represents dark energy, which acts as a cosmological constant. The analyses in this thesis are based on the ΛCDM model.

[1] For a more complete account of the history of cosmology, see, for example, Ref. [124]

2. Large Scale Structure and Dark Matter Halos

While the Universe is isotropic and homogeneous in general, inflation leaves the density field with small perturbations. These initial perturbations evolve linearly, until overdensities break away from the expansion of the Universe and collapse under gravity, and underdensities become voids. The power spectrum of matter is a decreasing function of scale, and nonlinear collapse happens on smaller scales first. Matter collapses triaxially to form sheets, filaments, and clusters; smaller objects then merge to form larger ones hierarchically.

Dark matter halos are dense clusters of dark matter that are approximately virialized [39]. The region that is approximately virialized typically has a density that is several hundred times the mean matter density of the Universe, by which halo boundaries are often defined. One most commonly adopted definition is the virial boundary [27], which depends on the cosmology and evolves with time. In the hierarchical model, most of the mass in the Universe today is contained in halos, rendering them the basic units for understanding the large-scale matter distribution. Halos trace the underlying density field, and their clustering is a probe of cosmology and the large-scale structure. However, halos only trace the field in a biased manner, in the sense that they cluster differently from dark matter, and the bias needs to be understood in order to interpret the clustering statistics.

The clustering strength of halos has a strong dependence on halo mass [83, 123], with more massive halos clustered more strongly together, i.e, having more bias. On the other hand, halo clustering is also dependent on other properties of halos [205, 60, 104], though the effect is much weaker than the mass dependence. The most commonly studied halo properties include concentration, spin, age, etc. Halos found in simulations can be described by an approximately universal mass profile — the Navarro, Frenk and White (NFW) [130, 131, 132] profile, where the density $\rho(r) \propto r^{-1}$ in the inner halo and $\rho(r) \propto r^{-3}$ in the outer halo. The concentration parameter characterizes the scale of the dense core relative to the entire halo. Halo spin quantifies the rotation of halos. Halo age is a measure of when halos form, and has multiple definitions due to the fact that halo formation is an extend process rather than an instantaneous event. Ref. [61] used the term *halo assembly bias* to refer to the age dependence of halo clustering, but because many other halo properties are also correlated with the assembly history of halos, this term is historically used to refer to all secondary dependences of clustering strength on these halo properties. Ref. [118]

advocated referring to these dependences as the *secondary halo biases* instead.

3. Halos as Hosts to Galaxies

In the potential wells of dark matter halos, gas cools and condenses to form galaxies [208]. Galaxies reside in the halos where they form and evolve with them. In addition to the central galaxy at the center of its potential well, a massive halo can host multiple satellite galaxies. These satellite galaxies are associated with subhalos that orbit within the host, which are accreted through mergers. We interpret observations by associating galaxies with halos, which trace the large-scale structure.

As halos dominate the immediate environments of galaxies that they host, their properties also dominate the number and properties of these galaxies. The connection between galaxies and halos is of great interest to us, as it is at the intersection of cosmological theory and direct observables from data. Galaxies trace halos, and hence the underlying matter field. While dark matter halos cannot be directly observed, light signals from galaxies come in great abundance, and galaxies are the major source of information on the large-scale structure of the Universe. However, the manner in which galaxies populate and trace halos is nontrivial, and we only have imperfect knowledge of it. To make cosmological inference from galaxies, the galaxy–halo connection must be properly modeled and marginalized over. Also, knowledge of the statistical connection between galaxies and their host halos informs how galaxies form and evolve in their environments.

The galaxy–halo connection is understood in terms of the dependence of galaxy number and properties on halo properties. Traditionally the primary halo property that is considered to determine galaxy occupation is halo mass, where more massive halos host more galaxies. Other properties that quantify the sizes of halos, such as the maximum circular velocity, which characterizes the depth of the potential well, can also serve as the primary property. However, as galaxies co-evolve with halos, it is natural to speculate that their properties depend on the evolution process of halos, and other halo properties that reflect this evolution. The secondary dependence of galaxy properties on halo properties other than mass is termed the *galaxy assembly bias*, which has not yet been definitively detected in data. Galaxy assembly bias is a major subject of this thesis, and we will look further into it in later chapters. In particular, we will attempt to answer the question

of whether or not the properties of galaxies depend upon properties of halos other than mass.

B. General Methodology

Due to the complexity of cosmology and astrophysics, the wide range of scales that are involved, and the stochasticity in physical processes, a large part of modern cosmology research is of numerical, empirical, and statistical nature. Mock universes are created through numerical simulation, empirical models are built to reconstruct more detailed aspects, and predictions are compared against observational data in a statistical manner. In this section, I give a brief introduction of each part of this process, focusing on the methods related to the galaxy–halo connection.

1. Numerical Simulations

In studying dark matter halos and galaxies, numerical simulations of different levels of complexity are used (see Ref. [197] for a recent review). There are two main categories of cosmological simulations: N-body simulations that simulate the gravitational evolution of structure, and hydrodynamical simulations that further simulate the baryonic physics of galaxy formation and evolution. Simulations start from initial conditions at a high redshift, and are evolved to the present day, with the expansion of space dictated by dark energy.

N-body simulations (e.g., Ref. [171, 169, 87]) are often referred to as dark matter-only simulations, because they treat all mass like dark matter and consider gravitational interactions only. As dark matter constitutes the majority of the mass in the Universe, N-body simulations capture the main features of the large-scale structure. N-body simulations have the obvious merit of relatively low computational costs. However, with computational resources that are currently available, there is still a balance between the simulation volume and resolution[2], and different choices suit different needs. High-resolution simulations can be used to study the detailed substructure and evolution of halos, whereas large-volume simulations provide better datasets for studying large-scale statistics.

[2] A recent series of papers [103, 135] explores the possibility of using machine learning approaches to rapidly construct high-resolution simulations from large-volume, low resolution ones.

Hydrodynamical simulations (e.g., Ref. [196, 160]), on the other hand, take into account various aspects of baryonic physics, besides gravity. The baryonic physics considered include gas cooling, star formation, supernovae feedback, active galactic nuclei feedback, magnetic fields, dust, etc. Hydrodynamical simulations are computationally more expensive, and inevitably depend on many simplifying assumptions, but provide a more realistic reconstruction of the Universe, and richer mocks of galaxies. Comparison between hydrodynamical simulations and their dark matter-only counterparts inform the connection between galaxies and halos.

2. Empirical Models

Because of the great complexity and uncertainty of the full physics of galaxy formation and evolution, it is often difficult and costly to directly generate galaxies in mock universes. Instead, the easy acquisition of halo catalogs from N-body simulations gives advantage to the empirical approach that assign galaxies to halos based on halo properties[3]. The empirical approach is based on the fact that galaxy properties have a statistical dependence on halo properties. The choice of different statistical dependences leads to different flavors of empirical galaxy–halo connections (see [203] for a recent review). The commonly used models include:

(1) Subhalo abundance matching (SHAM) [184, 36], where the rank order of one primary galaxy property (such as luminosity) is matched to that of one primary halo property (such as maximum circular velocity), with a possible scatter.

(2) Halo occupation distribution (HOD) [92, 230], where the number of central and satellite galaxies in a halo are drawn from distributions with means determined by the halo's mass. Ref. [74] developed an extension to the model (the decorated HOD) that quantifies the effect of secondary halo properties on the occupation. The work on galaxy–halo connection in this thesis is based on the HOD and decorated HOD models, and a more detailed description is available in Chapter III.

(3) Conditional luminosity function (CLF) [215, 188], where not only the number of galaxies, but also their luminosity distribution, is modeled from halo mass.

Despite their simplicity, empirical models have been able to explain observations to a large extent, and provide physical insight into galaxy formation and evolution in halos, enabling straight-

[3]Besides hydrodynamical simulations and empirical models, there are also semi-analytical models, which analytically model physical processes along the merging history of dark matter halos.

forward tests of the outcome of hydrodynamical simulations. In cosmological interpretation of data, galaxy physics can be marginalized over through empirical models, which circumvents the need to understand the numerous astrophysical details.

3. Observational Tests

Galaxy surveys cover sections of the sky, and observe galaxies in great quantities (e.g., Refs. [134, 108, 101, 2]). Spectroscopic surveys measure photon fluxes at different wavelengths, and infer the redshifts of galaxies from the spectra; the redshifts are then used for estimating their distances from us. Photometric surveys measure light in broader bands, and are thus able to detect fainter objects, though the finer spectral features are not observed, and redshift estimates are much less accurate.

Statistics of galaxy spatial distribution can be extracted from the survey catalogs, and used to inform the galaxy–halo connection. Theoretical predictions and real data are compared through these summary statistics. Different statistics are measured for galaxy populations selected by certain criteria, and reflect different aspects of the galaxy distribution. Some of the statistics often used to constrain galaxy–halo connection models are:

(1) Number densities of galaxies. The number density of a galaxy population reflects that of the halo population in which they reside, which also constrains cosmological structure formation.

(2) Two-point correlation functions. A Gaussian field can be fully specified by its two-point function, which is the Fourier transform of the power spectrum. Even in the present-day field that deviates from Gaussianity, the two-point function still incorporates most of the field's information. The two-point function is the standard statistic used in most works for analyzing galaxy clustering.

(3) Weak lensing (galaxy–galaxy lensing). The extent to which light from background galaxies is bent by the matter around lens galaxies reflects the mass profiles of the halos of the lens galaxies, and directly probes the galaxy–matter correlation. However, lensing suffers from numerous systematics, which results in low signal-to-noise ratios.

(4) Count statistics or group statistics. Counts-in-cells statistics are in effect cheaper alternatives of galaxy counts in individual halos (i.e., groups). These statistics probe higher-order information of the field. Unlike the two-point function or weak lensing, which are measured in the average

sense, the full probability distribution of counts are analyzed. Coupled with the distribution of halo properties, they can provide much insight into the galaxy occupation of halos.

(5) Void statistics. Instead of the other statistics that focus on overdensities, void statistics put more weight on underdense regions of the Universe, and also embodies higher-order information of the field.

(6) Satellite kinematics. The kinematics of satellite galaxies in halos directly measure the potential wells of halos, but this approach is very much limited by our ability to categorize galaxies as satellites and to associate them with centrals.

There are other summary statistics that can be measured from the galaxy distribution and used to constrain the galaxy–halo connection, each with its own merits and limits.

C. Structure of Thesis

This thesis has three major components, below I outline the motivation and findings of each.

- **Chapter II** — The galaxy–halo connection model is based on the intertwined evolution of halo and galaxy properties. Halo concentration is a key halo property that characterizes the structure of a halo; it is used in various theoretical and observational analyses, and is often treated as a proxy for the halo assembly history. In this study, we examine the connection between halo concentrations and their mass assembly histories with the Dark Sky Simulations. Upon finding that traditional definitions of the formation time inevitably leave a considerable amount of the scatter in concentration unexplained, we further investigate the details of halo mass assembly. We stack and compare merging events between halos of similar masses (major mergers), and find that they induce violent responses in the concentration. We observe remarkably universal shapes and dynamical timescales in these responses, which can be associated with the orbital dynamics of mergers. These effects are significant in scale compared to the scatter in concentrations, and last for large fractions of the age of the Universe. We also examine mergers between halos of significantly different masses (minor mergers) and showed that they are similar but less dramatic. We demonstrate that the cumulative effect of major mergers and frequent minor mergers leads to an irreducible scatter in concentration at fixed halo mass and

fixed formation time, which is present even in halos with assembly histories that are typically considered quiescent. These findings have profound consequences for semi-analytical and analytical models of halo structure and galaxy formation, which depend on the mechanism of halo profile evolution. These results also impact the interpretation of observations that rely on the concentration–mass or concentration–formation time relations, such as strong lensing, weak lensing, and satellite kinematics.

- **Chapter III** — The empirical galaxy–halo connection is widely used to interpret observed data in cosmological analyses and inform the physics of galaxy formation. It is therefore crucial to make the correct assumptions in these models. While halo mass is the main factor in determining how galaxies populate halos, evidence suggests that the galaxy–halo connection depends on other halo properties. This effect, termed galaxy assembly bias, is a source of significant systematic error in cosmological analyses. To conclusively detect or reject its presence in the Universe, tighter constraints on galaxy assembly bias are required. In this study, we develop techniques to better constrain the parameterized strength of galaxy assembly bias in an empirical model which populates halos in N-body simulations with galaxies. We find that the conventional combination of observables — the projected two-point correlation function $w_{\rm p}(r_{\rm p})$ and the galaxy-galaxy lensing signal $\Delta\Sigma(r_{\rm p})$ — have largely common information contents and do not complement each other. Therefore, we augment the set of observables with higher-order statistics including the void probability function VPF(r), the counts-in-cylinders statistic $P(N_{\rm CIC})$, and two novel count statistics that probe the immediate environments of halos. As an improvement upon the previous state of the art, we conduct a comprehensive study of the auto- and cross-covariance of all of the candidate statistics, accounting for various sources of uncertainty. With this full covariance matrix and a careful treatment of the stochasticity in the dependence of statistics on model parameters, we make a forecast of the constraining power on assembly bias parameters from different combinations of statistics. We find that each count statistic significantly outperforms $\Delta\Sigma(r_{\rm p})$ in complementing $w_{\rm p}(r_{\rm p})$, as they encode the higher-order information of the field and cut through the constraints from the two-point function $w_{\rm p}(r_{\rm p})$ in the parameter space. We therefore advocate the combined use of the two-point function and count statistics as a probe for galaxy assembly bias. This observable set has the potential to lead to the definitive detection of galaxy assembly bias.

- **Chapter IV** — We apply our findings from the previous study to data from the Sloan Digital Sky Survey (SDSS). To validate the use of a cubic simulation for fitting data measurement, we build light cone mocks and cubic mocks with the same underlying galaxy–halo connection. In the light cone mocks we implement the observational effects present in the SDSS catalog, including geometry and fiber collision. We validate our measuring algorithms by testing that the cubic box yields unbiased estimates of the light cone statistics that mimic SDSS observation. We then make our own measurements of the complementary pair of observables — the projected two-point function and the counts-in-cylinders statistic — in the SDSS catalog, and estimate both the theoretical and observational covariances. We fit an empirical model that incorporates galaxy assembly bias to our measurements using Bayesian inference methods. We get tighter constraints on the galaxy–halo connection, in particular the galaxy assembly bias effect, than preceding studies. We find definitive evidence for galaxy assembly bias in some samples. These findings will in turn improve both the cosmological constraints from galaxy data and the physical models of galaxy formation and evolution in halos.

  Conclusions and implications are described in **Chapter V**.

II. Halo Assembly and Halo Structure

This chapter is originally published as: Wang, K., Mao, Y.-Y., Zentner, A. R., Lange, J. U., van den Bosch, F. C., Wechsler, R. H. (2020), Monthly Notices of the Royal Astronomical Society, 498, 4450.

Minor modifications have been made to the text. The inclusion of this article in this dissertation is in compliance with the copyright policies of the journal.

The concentration parameter is a key characteristic of a dark matter halo that conveniently connects the halo's present-day structure with its assembly history. Using "Dark Sky", a suite of cosmological N-body simulations, we investigate how halo concentration evolves with time and emerges from the mass assembly history. We also explore the origin of the scatter in the relation between concentration and assembly history. We show that the evolution of halo concentration has two primary modes: (1) smooth increase due to pseudo-evolution; and (2) intense responses to physical merger events. Merger events induce lasting and substantial changes in halo structures, and we observe a universal response in the concentration parameter. We argue that merger events are a major contributor to the uncertainty in halo concentration at fixed halo mass and formation time. In fact, even haloes that are typically classified as having quiescent formation histories experience multiple minor mergers. These minor mergers drive small deviations from pseudo-evolution, which cause fluctuations in the concentration parameters and result in effectively irreducible scatter in the relation between concentration and assembly history. Hence, caution should be taken when using present-day halo concentration parameter as a proxy for the halo assembly history, especially if the recent merger history is unknown.

A. Introduction

In the concordance, ΛCDM cosmological model [91, 141, 143, 2], the formation of galaxies and clusters proceeds hierarchically: smaller dark matter haloes are the first to collapse and these haloes grow larger through mergers. Dark matter haloes form around peaks in the initial density

field, and gas cools and condenses to form galaxies within the potential wells provided by these haloes [208, 23]. The formation and evolution of haloes and galaxies are thus inextricably linked. A key goal of developing a modern theory of structure formation has thus been to understand the detailed connection between galaxy properties and the structure and assembly histories of the dark matter haloes in which they form.

Contemporary computational hardware and algorithms enable large-volume, high-resolution, gravity-only, N-body simulations of structure formation, as well as the rapid analysis of these simulations [171, 88, 154, 87, 145, 76, 46]. Consequently, simulations have largely replaced analytic models [147, 24, 162, 16, 39] as the primary framework for the interpretation of large-scale structure measurements. In these analyses, dark matter haloes are considered the basic units of nonlinear structure and observable statistics are computed by associating galaxies with haloes using some physically-motivated, empirical model (see Ref. [203] for a recent review). Therefore, an understanding of halo structure is necessary in order to interpret observations and to test models of galaxy formation, cosmology, and/or the nature of the dark matter.

The most commonly accepted model for the density profiles of haloes is the two-parameter profile defined by Navarro, Frenk, and White [130, 131, 132] (NFW hereafter),

$$\rho_{\rm NFW}(r) = \frac{\rho_{\rm s}}{(r/r_{\rm s})(1+r/r_{\rm s})^2}, \tag{1}$$

where $\rho_{\rm s}$ is the inner scale density, and $r_{\rm s}$ is the scale radius, which characterizes the transition from $\rho(r) \propto r^{-1}$ in the inner halo to $\rho(r) \propto r^{-3}$ in the outer halo. Though refinements to the NFW profile have been suggested [125, 55, 59, 129], the NFW profile successfully describes the general structure of haloes found in simulations and has become the *de facto* standard halo profile.

It is now customary to quantify the relative concentration of a halo's mass toward its center using the *concentration parameter*:

$$c_{\rm vir} = R_{\rm vir}/r_{\rm s}, \tag{2}$$

where $R_{\rm vir}$ is the halo's virial radius. NFW discovered that the concentration parameter is a decreasing function of halo mass. This is known as the concentration–mass relation, which has since been extensively studied [28, 204, 113, 146, 111, 42, 51, 87, 33].

In addition to establishing the *de facto* standard density profile, NFW suggested a relationship between halo concentrations and halo mass assembly histories, and this was quickly seized upon in

subsequent studies. For example, Ref. [159] argued that violent relaxation, induced by the rapidly-fluctuating gravitational potentials present during halo mergers, rearranges halo structure leading to a nearly universal mass profile. Based on the framework first proposed by Ref. [132], Ref. [28] quantitatively modeled halo concentration by relating it with an epoch of initial halo collapse that sets the initial inner halo density. Ref. [204] (W02 hereafter) found a general functional form of the mass assembly history (see also [189, 175, 119, 211, 191, 40]), and established a tight correlation between halo concentration and the characteristic formation epoch, a_c, at which $\mathrm{d}\log M/\mathrm{d}\log a$ falls below a specified value of S (W02 took $S = 4.1$ for their primary results). Later works found that, on average, the halo mass assembly history can be roughly divided into an early phase of fast accretion that builds up the potential well, and a late phase of slow accretion that adds mass without significantly changing the potential well [228, 105, 227]. In this scenario, the fast accretion phase sets an approximately universal initial concentration, while the concentration only grows slowly during the slow accretion phase. Moving beyond the one-parameter description characterized by the concentration parameter, Ref. [109] studied the entire halo mass profile, and interpreted it in terms of the entire halo assembly history, demonstrating a link between the two.

The physical nature of halo mass growth was further studied by Ref. [49], who distinguished "physical evolution" from "pseudo-evolution," which refers to the increase in halo mass resulting from the dilution of the background density rather than the coherent infall of matter associated with mergers. The virial radius of a halo, $R_{\rm vir}$, is typically defined as the radius of the spherical region within which the average density is some multiple (the exact value depends upon the specific analysis) of the mean density or critical density of the universe. As the universe expands and the reference density dilutes, halo radii and halo masses grow even in the absence of any physical mass accretion onto the halo. Pseudo-evolution increases halo radii, so it also proportionally increases halo concentrations. In the majority of models proposed to explain the relation between concentration and mass, and/or the relation between concentration and formation time, the scale radii of haloes were assumed to be set during an initial stage of rapid mass acquisition. After this initial phase, scale radii were typically assumed to be fixed or to evolve only slowly. In these models, concentrations subsequently increase as haloes slowly acquire mass via mergers, smooth accretion[1],

[1]In the present work we consider smooth physical accretion as the limit of minor mergers and do not treat it separately.

or pseudo-evolution, all of which increase $R_{\rm vir}$ while $r_{\rm s}$ is assumed to remain approximately fixed. Differentiating between mass growth modes has had an important role in interpreting the evolution of the concentration. Ref. [199], for instance, separated mergers that affect inner regions of haloes from "diffuse" accretion during which the inner regions remain stable; this later effect in fact includes pseudo-evolution. However, the assumption of a stable inner region and constant scale radius would only hold if the halo has a perfectly quiescent assembly history. Ref. [105], for example, found that the slow accretion phase is still dominated by minor mergers, which, as we will show, can impact the scale radius.

The scatter around the mean relation between concentrations and mass assembly histories, and the origin of such scatter, have also been of considerable interest. W02 demonstrated that a large part of the scatter in the concentration–mass relation can be attributed to different formation times at a fixed mass, but the remaining scatter in the relation of concentration and formation time prompts further investigation. W02 and Ref. [113] found that the scatter in the concentration is reduced when the haloes with recent mergers are excluded from the sample, which suggests that mergers contribute to this scatter. Ref. [112] found that haloes identified when they are substantially out of equilibrium, primarily due to mergers, experience oscillations in their concentrations. Ref. [99] observed similar behavior in their phase-space analyses of haloes during post-merger relaxation. This could result in a scatter in the concentration, depending on the time of measurement. It is also natural to expect that, beyond the identification of a single proxy for the formation time of a halo, the various details of mass assembly histories play a part in shaping halo structure. Ref. [133], for instance, found evidence suggesting that halo concentration depends not only on the mass assembly history of the halo, but also on the mass assembly histories of the haloes that merged to form the final halo. A more recent study by Ref. [153] demonstrated that halo concentrations are sensitive to both the smoothness of the merger history and the order in which mergers happen, by generating versions of the same halo with different assembly histories (see also [156]). Ref. [81] developed a model for predicting scale radii and hence concentrations, which takes into account the entire structure of the merger tree, and were able to better capture the scatter than previous models [110, 15].

In this study, we seek a detailed understanding of the relationship between the mass assembly histories of haloes and their concentrations. We perform a systematic search to identify character-

istics of the mass assembly history that can effectively predict present-day concentration. Various summary statistics of the mass assembly history are highly correlated with the present-day concentration. In this work, we explore different ways to represent the mass assembly history to further optimize such correlations. We then study the evolution of the concentration parameter, and investigate how pseudo-evolution and merger events impact the evolution of concentration, both for individual haloes and statistical samples. We study how mergers contribute to the scatter in the relation between concentration and mass assembly history.

This chapter is organized as follows. In Section II.B, we describe the simulations we use and specify the selection criteria for our samples. In Section II.C, we report the correlation between halo concentration and mass assembly history that we find in our samples. The separate roles of pseudo-evolution and physical growth in the evolution of halo concentration and halo scale radius are examined in Section II.D. We discuss our findings and draw conclusions in Section II.E.

B. Simulation and Sample Selection

1. Simulation

In this work, we use the Dark Sky Simulations, a suite of cosmological, gravity-only simulations [169]. The Dark Sky Simulations are run with the 2HOT code [201], adopting a flat cosmology with $h = 0.688$, $\Omega_{\rm m} = 0.295$, $n_s = 0.968$, and $\sigma_8 = 0.834$. We use two of the Dark Sky Simulations: `ds14_b` and `ds14_i`. The `ds14_b` box has a volume of $(1\,h^{-1}\,\mathrm{Gpc})^3$, with 10240^3 particles; however, the halo catalogs and merger trees that we use are generated with a downsampled version[2] of `ds14_b` that has only $10240^3/32 \simeq 3225^3$ particles, with an effective mass resolution of $2.44 \times 10^9\,h^{-1}\,\mathrm{M}_\odot$. The `ds14_i` box has a volume of $(400\,h^{-1}\,\mathrm{Mpc})^3$, with 4096^3 particles, and hence a mass resolution of $7.63 \times 10^7\,h^{-1}\,\mathrm{M}_\odot$.

Both simulations have outputs at 99 epochs:

$a = \{0.06, 0.065, ..., 0.09, 0.095, 0.1, 0.11, 0.12, ..., 0.99, 1\}$.

The halo catalogs are generated by the ROCKSTAR halo finder [13], using the virial definition

[2]Unfortunately, halo catalogs and merger trees of the full `ds14_b` simulation are not available due to computational infeasibility.

as the halo boundary, corresponding to a spherical overdensity of $\Delta_{\rm crit}$, which takes the value of 100.46 at $a = 1$ in this cosmology, with respect to the critical density [27]. Throughout this study, we use $M_{\rm vir}$ as the halo mass and $c_{\rm vir}$ as the halo concentration, and we will omit the subscript "vir" in places for brevity.

ROCKSTAR identifies haloes in the six-dimensional phase-space, utilizing both position and velocity information. This algorithm greatly improves performance in distinguishing subhaloes and tracking merger events, compared with friends-of-friends algorithms that are based solely on dark matter particle positions. Subhaloes are haloes with centers that fall within the the virial radius of a larger halo, while haloes with centers that do not lie within the virial radius of any larger halo are referred to as host haloes. In ROCKSTAR, the scale radius, $r_{\rm s}$, is directly fitted using a χ^2 fit of the NFW profile. The particles associated with a halo are divided into up to 50 radial equal-mass bins, with a minimum of 15 particles per bin, and radial bins that are smaller than 3 times the force resolution scale are assigned a low weight in the estimation of χ^2, to suppress resolution effects at small scales. The concentration is calculated from $c = R/r_{\rm s}$, where R is the halo radius. As most halo finders do, ROCKSTAR fits the radially averaged profile, and includes substructures in the fit. It is reasonable to expect that the results of our analyses would be different if substructures were removed from the profile. A lower bound is enforced on the fitted concentration, $c \geq 1$. We have tested that our conclusions do not rely on the fitting scheme, and are not affected qualitatively when $V_{\rm max}/V_{\rm vir}$ is used as a proxy for concentration [89].

The merger history is analyzed using the CONSISTENT TREES merger tree builder [14]. At each merger event, we refer to the merging halo that shares the most particles with the resulting halo, as the *main progenitor halo.* Merger trees are constructed by tracing the evolution of a halo from today backward in time. The *main branch* of the halo merger tree follows the main progenitor halo at each merger event. We refer the interested reader to Ref. [13] and Ref. [14] for details.

2. Sample Selection

a. Present-day Mass Samples

We first study three host halo samples defined by present-day virial mass, around $10^{12}h^{-1}\,{\rm M}_\odot$, $10^{13}h^{-1}\,{\rm M}_\odot$, and $10^{14}h^{-1}\,{\rm M}_\odot$ respectively. The details of the selection are listed in Table II.1. We

Present-day Mass Samples			
Box	$M_{\rm min}$	$M_{\rm max}$	Sample size
$400\,h^{-1}$ Mpc	$10^{12}h^{-1}\,{\rm M}_\odot$	$1.1\times10^{12}h^{-1}\,{\rm M}_\odot$	21099
$400\,h^{-1}$ Mpc	$10^{13}h^{-1}\,{\rm M}_\odot$	$2.0\times10^{13}h^{-1}\,{\rm M}_\odot$	14543
$1\,h^{-1}$ Gpc	$10^{14}h^{-1}\,{\rm M}_\odot$	$3.8\times10^{15}h^{-1}\,{\rm M}_\odot$	25438

Table II.1: Three present-day mass halo samples. We list the simulation box from which each sample is selected, the lower and upper bounds of virial mass, and the resulting sample sizes.

choose to select the halo samples from different simulation boxes because the mass resolution of the $1h^{-1}$ Gpc simulation does not suffice to resolve the internal structures of lower-mass haloes at early times, while the number of cluster-size haloes in the $400h^{-1}$ Mpc simulation is relatively limited. Hereafter we will refer to these three samples as the $10^{12}h^{-1}\,{\rm M}_\odot$, $10^{13}h^{-1}\,{\rm M}_\odot$, and $10^{14}h^{-1}\,{\rm M}_\odot$ samples.

b. Major Merger Samples

To examine major merger events and the impacts of these mergers on halo structure, we identify the haloes that undergo major mergers in their main branches at the time step preceding $a = 0.33, 0.50, 0.67$ and 0.80, corresponding to redshifts of $z = 2, 1, 0.5$ and 0.25 respectively[3]. Our sample selection is based on the major mergers identified by CONSISTENT TREES, which defines major mergers as mergers in which the ratio of the masses of the progenitors exceeds $1/3$. We compare our major merger samples with a control group of haloes selected randomly from the simulation. For all these samples, we require the haloes to be host haloes today and at the time of the major merger. We further require each halo in our samples to have mass above $4\times10^{10}h^{-1}{\rm M}_\odot$ since $z = 2$ to circumvent the effect of mass resolution[4]. All the samples are selected from the $400\,h^{-1}$ Mpc box. Each halo can belong to multiple samples; a halo that undergoes major mergers

[3]The time of merger is defined as the first snapshot in which the center of the smaller progenitor has entered the virial boundary of the main progenitor and the smaller progenitor has become a subhalo.

[4]We have tested that this resolution requirement does not affect our results.

Major Merger Samples	
Sample	Sample size
$a_{\rm MM} = 0.33$	58241
$a_{\rm MM} = 0.50$	17784
$a_{\rm MM} = 0.67$	10426
$a_{\rm MM} = 0.80$	7091
Random	95087

Table II.2: Sample size of each major merger sample and the random sample. The parameter $a_{\rm MM}$ denotes the scale factor of the universe when the major merger occurred. The samples are not mutually exclusive.

at more than one snapshot of interest will be included in multiple major merger samples, and the random sample can include haloes that are in the major merger samples. The size of each sample is listed in Table II.2.

C. Relation between Concentration and Mass Assembly History

In this section, we revisit the connection between halo concentration and halo mass assembly history using the Dark Sky Simulations, exploring several aspects of halo mass assembly histories. In all cases, we study samples of haloes within a narrow range of contemporary mass and further control for any mass-dependent effects within each sample. This implies that these results also characterize the correlations between present-day scale radii and halo mass assembly histories because haloes of fixed mass have identical virial radii.

1. Correlation with Mass at a Specific Time

There have been several attempts to summarize the mass assembly history with a single parameter that correlates strongly with the present-day concentration [28, 204, 228, 41]. Two common choices are the halo half-mass scale, $a_{1/2}$, which is the epoch at which the halo first assembled half of its present-day mass, and the W02 formation time, a_c, which serves as an estimate of the end of the early phase of rapid mass accretion by the halo. These attempts were relatively successful, suggesting that much of what determines contemporary halo concentration can be summarized with one quantity and that there may exist a "key stage" in a halo's assembly history that substantially impacts the halo's internal structure.

This motivates us to conduct a systematic, empirical search for the stage of mass assembly that is most correlated with the present-day concentration c. We quantify mass assembly histories in two ways: (1) by the epoch $a(m)$ at which a fraction m of the present-day halo mass is first assembled (for example, the half-mass scale $a_{1/2} = a(m = 0.5)$); and (2) by the relative mass fraction $m(a) = M_{\rm vir}(a)/M_{\rm vir}(a = 1)$, which is the mass of the halo at time a in units of its contemporary mass.

For the purposes of this study we choose to represent the mass assembly histories using the mass of main progenitors as a function of scale factors. However, there are multiple other characterizations of the formation history that we have not explored, for example, the collapsed mass history [132, 59, 112], and the transition between rapid and slow accretion phases [228].

Both concentration and mass assembly history are known to correlate with present-day mass. While we work with mass-selected halo samples, we further mitigate any correlations induced by the mass dependence of the relative mass fraction and concentration as follows [118]. We divide each mass-selected halo sample shown in Table II.1 into narrow bins. Within each of these bins, we assign each halo a mark, $\mathcal{M}(x)$, where x is the property of interest. Either $x = c$ or $x = m(a)$ in our present discussion. $\mathcal{M}(x)$ is the percentile rank among all of x within the bin. For example, $\mathcal{M}(x)$ ranges between $\mathcal{M}(x) = 0$, for the halo with the lowest value of x in the bin, and $\mathcal{M}(x) = 1$, for the halo with the highest value of x in the bin. Each of our three mass-selected samples corresponds to a range of halo masses given in Table II.1. We divide the $10^{12}\,h^{-1}\,{\rm M}_\odot$ and $10^{13}\,h^{-1}\,{\rm M}_\odot$ samples into 20 logarithmically-spaced mass bins and the $10^{14}\,h^{-1}\,{\rm M}_\odot$ sample into 30 logarithmically-spaced

bins.

We study correlations with the concentration mark $\mathcal{M}(c)$, for the two forms of mass assembly history, $\mathcal{M}(a(m))$ and $\mathcal{M}(m(a))$, as defined in the preceding paragraph. Specifically, we compute $\mathcal{M}(m(a))$ for the 99 values of a that correspond to the 99 available snapshots of the simulations, and $\mathcal{M}(a(m))$ for $m = \{0.01, 0.02, 0.03, ..., 0.99, 1\}$. We then calculate the Spearman rank-order correlation, ρ, between these marks of the assembly history and $\mathcal{M}(c)$.

The Spearman rank-order correlations between $\mathcal{M}(a(m))$ and $\mathcal{M}(c)$ as a function of the fraction m are shown in the top panel of Fig. II.1. The lines of different colors represent the different mass samples, as labeled in the same panel. The correlation coefficients are negative for all the values of m, and we show them in absolute values. This is in accordance with previous understanding that haloes are likely to be more concentrated if they assembled their masses at smaller scale factors. For all three samples, the correlations at large mass fractions are smaller than those at both medium and small fractions. The $a(m)$'s defined at a range of medium mass fractions ($0.3 \lesssim m \lesssim 0.7$) contain similar and relatively high levels of information about the present-day concentration. This also explains the comparable effectiveness of various definitions of formation time in previous literature. On the other hand, it is obvious from the figure that the time at which the main progenitor of a halo gains a low fraction of its final mass (e.g., 4% as in Ref. [227]) is not as informative as medium mass fractions, such as the commonly used half-mass scale, $a_{1/2}$.

The Spearman correlations between $\mathcal{M}(m(a))$ and $\mathcal{M}(c)$ as a function of a are shown as solid lines in the middle panel of Fig. II.1. The positive correlation at all times before $a = 1$ is also consistent with earlier-forming haloes being more concentrated. The correlations for all three samples are relatively low at early and late epochs, and peak between $a \approx 0.3$ and $a \approx 0.7$, depending upon halo mass. By construction, $m(a = 1) = 1$ in all cases, so all correlations converge to 0 at $a = 1$. The peak of the correlation curve, which indicates the epoch at which the relative mass fraction $m(a)$ is best correlated with concentration, occurs later for more massive haloes. This is consistent with the tendency of more massive haloes to form later, so that if there is an important epoch in the evolution of a halo that influences its internal structure, it too occurs later for more massive haloes.

The significant correlation between the concentration and the two characterizations of mass assembly history is in broad accordance with previous studies that identify formation epochs of

haloes that influence halo concentration. However, notice that the correlation curves in both the top and the middle panels peak at $\rho \lesssim 0.7$, suggesting that factors in addition to the mass of a halo at a particular time contribute to contemporary halo concentration. We will investigate this further below.

2. Correlation with Mass Change at a Specific Time

The values of mass fraction, $m(a)$, at successive time steps are strongly correlated with each other, and the resulting correlation coefficients in the top panel of Fig. II.1 are not independent. To resolve the relative importance of *instantaneous* mass growth at different epochs, we repeat the analysis in Section II.C.1 for the increment in mass fraction between adjacent snapshots, $\Delta m(a)$, instead of $m(a)$, with $\Delta m(a_i) = m(a_{i+1}) - m(a_i)$.

The correlations between instantaneous mass acquisition, $\mathcal{M}(\Delta m(a))$, and concentration, $\mathcal{M}(c)$, are shown in the bottom panel of Fig. II.1. It is evident that earlier growth is positively correlated with concentration (with the exception of the very earliest snapshots at which time the haloes are poorly resolved), while later growth exhibits anti-correlation. Similar to Section II.C.1, earlier growth is less informative for more massive haloes. Moreover, ρ peaks at lower values for more massive haloes indicating that their early assembly histories generally have less information on concentration compared to haloes of lower mass.

The peaks of the correlation curves in all panels of Fig. II.1 are broad. This indicates that a wide variety of times during the formation of a halo provide similar amounts of information on contemporary halo concentration. This is likely why a variety of halo formation time measures, such as $a_{1/2}$ and a_c, show similar levels of correlation with present-day halo concentration. The breadth of the peaks in Fig. II.1 further suggests that one cannot choose a single definition of the formation time that will dramatically outperform a variety of other reasonable choices. The distillation of the mass assembly history into a single parameter inevitably leads to a significant loss of information.

At late times, the correlation between concentration and mass increase becomes negative and reaches a minimum at $a \approx 0.83$ for all three mass samples. This is suggestive that the same dynamical process has caused this behavior. In Section II.D below, we identify this dynamical process to

be mergers. Mergers account for the anticorrelation in general, the stronger anticorrelation between $\mathcal{M}(c)$ and $\mathcal{M}(\Delta m(a))$ for more massive haloes, and the uneven feature at $a \approx 0.9$.

3. Linear Regression of Mass Assembly History to Predict Concentration

Efforts to explain concentration with mass assembly history are often focused on singling out a formation time that best represents the mass assembly history. However, we have shown in the previous subsections that multiple epochs in the mass assembly history contain similar amounts of information on concentration, which disfavors a single definition of formation time for this purpose. In order to integrate information on concentration from the full mass assembly histories, we perform an ordinary least squares linear regression with $\boldsymbol{a} = \{a(m = 0.01), \ldots, a(m = 0.99), a(m = 1)\}$ as the predictor variables and $\mathcal{M}(c)$ as the outcome variable, by fitting for a set of linear coefficients that minimizes $\sum_n \left(A_0 + \sum_i A_i a_i - \mathcal{M}(c) \right)^2$, where $\sum_n$ is the sum over all haloes, $\sum_i$ is the sum over all values of mass fraction m at which $a(m)$'s are defined, and A_0 and A_i are the linear coefficients. Similarly, we fit a set of linear coefficients, B_0 and B_i, with $\boldsymbol{m} = \{m(a = 0.06), m(a = 0.065), ..., m(a = 0.99), m(a = 1)\}$ as the predictor variables, that minimizes $\sum_n \left(B_0 + \sum_i B_i m_i - \mathcal{M}(c) \right)^2$, but here $\sum_i$ is the sum over all snapshots (i.e. over all values of a). For the present study, we elect to perform a simple linear regression, and refrain from more sophisticated forms of regression, because mass assembly histories of individual haloes are both volatile and noisy and these properties introduce the possibility of unphysical overfitting. For this reason, more complex regression methods warrant further, dedicated study.

We compare the results of the linear regression to the results of the previous section as follows. We determine the set of coefficients, A_0 and A_i, that gives the linear combination of the elements of $\boldsymbol{a}$ that is the most strongly correlated with $\mathcal{M}(c)$. We repeat the process for the elements of $\boldsymbol{m}$ to obtain the optimal coefficients B_0 and B_i. We then calculate the Spearman's correlations between $\mathcal{M}(c)$ and the marks corresponding to the resulting linear combinations.

The correlation coefficients for the optimal linear combinations of the two characterizations of mass assembly histories are shown in the top and middle panel of Fig. II.1 respectively, as horizontal dashed lines. For both the set of $a(m)$'s and $m(a)$'s, and at all three masses, even the optimal linear combinations leave much of the dependence of concentration on mass assembly

history unexplained, though they exhibit moderate improvements upon the best performing single parameters. Performing the linear regression with $\log \boldsymbol{m}$ instead of $\boldsymbol{m}$ yields similar, but slightly weaker correlations.

For comparison, we measure the mark correlation between the concentration and the formation time defined as the epoch at which the mass of the main progenitor first reaches the characteristic mass enclosed within the scale radius at the present-day, $M_s = M(r < r_s)$. We find that the level of correlation for this formation time is similar to the optimal linear combination of $m(a)$'s for all three of our mass samples. We further note that this definition of the formation time is not independent of the concentration measured at the present-day, as it requires knowledge of r_s, and therefore is different from the proxies of the assembly history that we have employed so far. This comparison suggests that our optimal linear combination captures most of the information in different forms of the formation history on the main branch.

In Section II.D, we explore the combined effect of merger events happening at different times on halo structure. We show that this combined effect cannot be described linearly.

D. Concentration from Pseudo-evolution and Mergers

In the previous section, we attempted to explain contemporary halo concentrations using halo mass assembly histories. The incomplete success of this endeavor prompts further inquiry into additional factors in the evolution of haloes that may influence halo density profiles. We expect the density profile of a halo to be largely determined by the halo's prior mass assembly history, independent of the redshift at which the halo is observed. We therefore extend our investigation to the study of the full evolution of halo concentrations, and search for connections between the behavior of halo concentrations and events in halo mass assembly histories.

In this section, we study the evolution of halo concentration c, and halo scale radius $r_{\rm s}$, both during quiescent periods of halo pseudo-evolution and during merger events. We find that halo structure undergoes significant changes in response to major, and even minor, mergers in a manner that is qualitatively universal. We propose a physical explanation for the response features that we observe. We further propose that the scale radii and concentrations of haloes result from

pseudo-evolution punctuated by marked fluctuations associated with merger activity.

1. The Pseudo-evolution of Halo Mass and Concentration

Pseudo-evolution refers to the fact that halo masses, virial radii, and concentrations all evolve even in the absence of merger activity or changes to scale radii [49]. This is because haloes are traditionally defined to be regions with a mean density larger than ∼50–100 times the critical density (or ∼200–350 times the background density). As cosmological expansion dilutes the universe, the size of the region above the density threshold increases even in the absence of any coherent, inward flow of mass. Consequently, $c = R_{\rm vir}/r_{\rm s}$ grows because $r_{\rm s}$ remains approximately constant in the absence of significant merger activity.

In the left column of Fig. II.2, we show the pseudo-evolution of halo mass, concentration, virial radius, and scale radius between $a = 0.2$ and $a = 1$, calculated using the Colossus software package [50], and assuming NFW profiles. Each panel depicts halo properties evolved both forward and backward from an initial point of $a = 0.4$ assuming pure pseudo-evolution. The pseudo-evolution is, itself, a function of halo concentration and we show halo pseudo-evolution for three different initial concentration values, $c_{0.4} \equiv c(a = 0.4) = 5, 10, 20$, in each panel. The *top panel* shows the pseudo-evolved mass normalized by the mass at $a = 0.4$, $M(a)/M_{0.4}$, which is only a function of the concentration, independent of halo mass. In the *middle panel*, we show the pseudo-evolution of concentration for the three values of $c_{0.4}$ separately. The evolution of $c(a)$ under pseudo-evolution is also independent of mass. The bottom panel shows the evolution of the scale radius $r_{\rm s}(a)/r_{\rm s,0.4}$ and virial radius $R_{\rm vir}(a)/R_{\rm vir,0.4}$ in physical units where $r_{\rm s,0.4}$ is the scale radius evaluated at $a = 0.4$ and likewise for $R_{\rm vir,0.4}$. These ratios are also independent of mass, and since the physical $r_{\rm s}$ remains constant under pseudo-evolution, the ratio $r_{\rm s}(a)/r_{\rm s,0.4} = 1$ independent of a. In all of the panels, the lines are labeled by the corresponding $c_{0.4}$ values. The left panels of Fig. II.2 show that pseudo-evolution contributes substantially to the evolution of halo size and concentration in the absence of any physical mass inflow or accretion. In the following subsections, we study the effect of physical accretion, which includes all the merger activities beyond pseudo-evolution.

2. Case Study: The Co-evolution of Halo Mass, Concentration, and Scale Radius During Mergers

In the case of pure pseudo-evolution, the evolution of halo mass, virial radius, and concentration from some initial state can be predicted. Significant deviations from the predictions of pseudo-evolution can likely be attributed to the physical inflow of mass across the virial boundary of the halo. To investigate how deviations from pseudo-evolution affect halo structure, we begin with a case study.

In the right column of Fig. II.2, we show the evolution of mass, concentration, virial radius, and scale radius for an individual halo. We neglect the evolution before $a = 0.2$, which is relatively poorly resolved. For consistency, we use the same quantities as in the left column, i.e., the concentration, as well as the mass and radii normalized by the values at $a = 0.4$, but note that the ranges of the y-axes are different. In the middle panel, $c = 1$, the lower boundary of fitted halo concentration, is marked by the horizontal dashed line. Major mergers in the mass assembly history of this halo are marked by gray, vertical, dashed lines.

Notice in the top panel that this particular halo undergoes no major mergers between $a \approx 0.25$ and $a \approx 0.7$. During this relatively quiescent period in the halo's mass accretion history, the halo's mass evolution is quite close to that predicted by pseudo-evolution, which we show with a dashed line for comparison. As in the left column of Fig. II.2, the pseudo-evolution is computed from $a = 0.4$. In the middle and bottom panels, the pseudo-evolution of the concentration and the scale radius during this period are also shown as dashed lines. The evolution of both the concentration and the scale radius during the period between major mergers is relatively mild. Comparing the actual evolution of these properties to the predictions of pseudo-evolution reveals non-negligible differences. Furthermore, decreases in the actual evolution of halo concentration seem to be visually associated with small deviations in the mass assembly history that are not identified as major mergers. This suggests that even small amounts of physical mass accretion can lead to significant deviations from the pseudo-evolution of concentration and scale radius. This, in turn, suggests that the scatter in the profiles of a population of haloes may be caused by small differences in mass assembly histories.

Focus now on the major merger events in Fig. II.2. Prominent features can be observed

in the temporal neighborhood of each major merger event. Concentration decreases rapidly and significantly to a minimum at approximately the time of the major merger. Subsequent to the merger, concentration immediately increases, decreases again, and then stabilizes. After stabilizing, there is a long period of secular increase of halo concentration. The change in concentrations due to major mergers is large compared with the scale of the overall concentration evolution throughout the entire history of the halo. Meanwhile, the scale radius follows the same trend but in the opposite sense, as is expected.

In the bottom panel, it is obvious that the change in $R_{\rm vir}$ is much less dramatic and much simpler than that in $r_{\rm s}$ in response to major mergers. $R_{\rm vir}$ increases due to both pseudo-evolution and the physical increase in mass, while $r_{\rm s}$ remains constant unless the inner profile is impacted. Concentration is the ratio $c = R_{\rm vir}/r_{\rm s}$. As is now apparent, discussing this ratio complicates our discussion unnecessarily, because the two radii have very different mechanisms of evolution. $R_{\rm vir}$ evolves rather modestly and in approximate correspondence with predictions from pseudo-evolution along with mass increases due to mergers. Large changes in concentration are induced by the large deviations in $r_{\rm s}$ brought about by mergers. We will therefore focus on the scale radius $r_{\rm s}$ instead of the concentration for the rest of this subsection.

We take the major merger at $a = 0.71$ as an example to discuss the common features, and interpret the response in $r_{\rm s}$ with the dynamical processes that occur during the major merger event. The two progenitors of this major merger are examined in Fig. II.3, which illustrates the orbit of the incoming progenitor around the main progenitor, as well as the mass evolution of the incoming and main progenitor haloes. During the major merger, the incoming progenitor loses mass to the main progenitor before being completely disrupted. Without significant physical mass growth, the scale radius only varies slowly, which can be seen in the period prior to the major merger at $a = 0.71$ in Fig. II.2, where the mass growth of the halo is mainly due to pseudo-evolution.

Notable deviations from the pseudo-evolution of the halo scale radius can be seen as the incoming halo traverses the main progenitor halo. As the merger begins, the halo's scale radius departs from its original evolution, and quickly increases, approaching the physical boundary $r_{\rm s} = R_{\rm vir}$, which suggests essential deviation from an NFW profile. This is due to the incoming progenitor entering the virial boundary of the main progenitor, shown in the upper part of the upper right panel in Fig. II.3, placing a relatively large amount of mass in the periphery of the main halo

and rendering the outer profile shallower. The shallower spherically-averaged density profile yields a larger scale radius. Later, as the incoming halo approaches the center of the main halo, the scale radius falls because mass is then inordinately concentrated near the halo center. The scale radius increases again as the merging halo moves outward from the center of the main halo on its orbit. Compared to the secular evolution in scale radius seen during quiescent periods, this evolution of halo scale radius is rapid, occurring over approximately one halo crossing time.

The merger concludes with the incoming halo spiraling inward toward the center of the main halo due to dynamical friction. As this happens, the scale radius once again increases. The incoming object is gradually disrupted, and $r_{\rm s}$ resumes secular evolution. The recovery after the major merger at $a = 0.71$ is interrupted by a later major merger that follows at $a = 0.94$; however, the recovery process can be observed in Fig. II.2 after the major merger at $a = 0.26$.

With this example, we have shown that during a major merger event, $r_{\rm s}$ experiences consequential changes, that can be attributed to the dynamical processes of the progenitors. Our interpretation is in agreement with Ref. [112], who also observed the oscillations in a halo and related them with the crossings of the merging object before virialization. These changes are extended in time, motivating an investigation of the time scales that are involved in the next section.

3. Universality of Response

In the previous subsection, we followed the co-evolution of mass, concentration, and scale radius of one halo, focusing on the dynamical processes associated with major mergers that drive the evolution of halo scale radius. Based on this case study, we argued that halo scale radii respond to mergers in an oscillatory manner and that the oscillations are due to orbital evolution. Accordingly, it is natural to study the evolution of haloes due to mergers with time measured in units of the local dynamical time, the time required to orbit across an equilibrium dynamical system, in our case a halo. We adopt the definition of the dynamical time

$$\tau_{\rm dyn} = \sqrt{\frac{3\pi}{16G\bar{\rho}}}, \tag{3}$$

where G is the gravitational constant and $\bar{\rho}$ is the mean density of the system, which we choose to be the virial density of haloes. With a given cosmology, the dynamical time $\tau_{\rm dyn}$ is dependent on

the scale factor a through $\bar{\rho}$. In the cosmology adopted by the Dark Sky Simulations, the dynamical time scales as

$$\tau_{\rm dyn} \approx 3.15(1+z)^{-3/2}\,{\rm Gyr}. \tag{4}$$

Following Ref. [80], we then define a new quantity T, which measures the time between two epochs in units of the dynamical time, as

$$T(a; a_{\rm ref}) = \int_{t(a_{\rm ref})}^{t(a)} \frac{dt}{\tau_{\rm dyn}(t)}, \tag{5}$$

where $t(a)$ is the age of the Universe corresponding to the scale factor a, and $a_{\rm ref}$ is the reference epoch.

To study the general behavior of major mergers, we select haloes from the simulation that undergo major mergers along the main branch, independently of their masses. The major merger times we select are $a_{\rm MM} = 0.33, 0.50, 0.67, 0.80$, corresponding to $z = 2, 1, 0.5$, and 0.25 (see Table II.2).

We stack each $a_{\rm MM}$ group and examine the median evolution to reduce noise. In Fig. II.4, we show the median response of the concentration and scale radius in logarithmic scale, normalized by their respective values at $a_{\rm MM}$. Time is measured both in terms of the scale factor and in terms of the number of dynamical times with respect to the time of merger.

In both columns, we observe the orbital features discussed in Section II.D.2, demonstrating that the dynamical processes shown in Fig. II.3 are universal, and that the incoming progenitor goes through one orbit on average before being disrupted (see also[190]). However, only in the right column, where time is measured in units of dynamical times, are the responses from the different $a_{\rm MM}$ groups aligned, going through the oscillations with a remarkably universal dynamical timescale. This further confirms the connection between the concentration and scale radius evolution and the dynamical processes during major mergers, as well as the universality of this mechanism when scaled with dynamical time.

4. All Merger Activity

We have shown that the evolution of halo structure in response to major mergers have common features, with universal timescales measured in units of dynamical times. The amplitude of the change in r_s due to major mergers is large compared with the average scale of change over the entire history, and also much larger than that of the halo radius evolution, causing large fluctuations in halo concentration as well. However, major mergers are relatively rare events. The average numbers of major mergers between $a = 0.25$ and $a = 1$ for a halo are 1.14, 1.51 and 2.00 for the $10^{12}h^{-1}\,M_\odot$, $10^{13}h^{-1}\,M_\odot$ and $10^{14}h^{-1}\,M_\odot$ samples respectively. Minor mergers with smaller ratios between progenitor masses happen much more frequently, and dominate the physical mass growth beyond pseudo-evolution. As major mergers are the extreme cases of merger events, it is reasonable to expect that minor mergers have similar but less dramatic effects.

To examine the response to all merger activity, we search for instances of minor merger events in the random catalog described in Section II.B.2.b. As the Consistent Trees code identifies major mergers only, we define minor mergers based on the rate of fractional mass increase between adjacent snapshots. We calculate the rate of fractional mass increase as

$$\Gamma(a_i) = \frac{\Delta M(a_i)/M(a_i)}{T(a_{i+1}; a_i)}, \tag{6}$$

where $\Delta M(a_i) = M(a_{i+1}) - M(a_i)$ is the mass increase between the adjacent snapshots, and $T(a_{i+1}; a_i)$ is the corresponding time interval in units of dynamical times. The rate of fractional mass increase, $\Gamma(a_i)$, is a dimensionless quantity. The time interval $T(a_{i+1}; a_i)$ for a fixed scale factor interval $a_{i+1} - a_i$ decreases as the Universe evolves, and drops below 0.2 by $a = 0.33$, the first merger epoch we consider. When selecting minor mergers, we consider the same epochs as for the major merger samples, $a = 0.33, 0.50, 0.67, 0.80$. The mean values of $\Gamma(a_{\rm MM})$ for the major merger samples are 2.77, 2.77, 2.96 and 3.43 for the four major merger times respectively. At each of these time steps, we select our minor merger sample to have values of Γ between 1.0 and 1.5. We also require that there are no major mergers within $\pm 0.25\tau_{\rm dyn}$ around the minor mergers[5], to exclude mass increase associated with major mergers.

In Fig. II.5, we compare the haloes that undergo these minor mergers against the major merger samples and the randomly-selected halo sample. The median evolution of each sample is plotted

[5]The mass increase associated with a major merger occurs over approximately $\pm 0.25\tau_{\rm dyn}$ around the time of merger.

in terms of both concentration and scale radius, as functions of time. In the bottom x-axes, time is represented as the number of dynamical times since the first available snapshot, while the corresponding scale factor is labeled on the top x-axes. The major merger samples are shown in the left column, and the lines are color coded according to the time of merger, marked by vertical dashed lines of the same colors. The solid black curve shows the evolution of the random sample for comparison. Similarly, the right column shows the minor mergers that happen at the same time steps.

It is obvious from Fig. II.5 that minor mergers indeed cause qualitatively similar responses in the halo structure. The magnitudes of these features, though smaller than those of the major merger response, are still significant compared with the scale of overall evolution throughout cosmic time. This shows that all mergers, major or minor, involve similar dynamical processes, with the effect of expanding the inner profile and suppressing concentration during an extended period. We also note that the haloes that undergo mergers have lower concentrations than the random sample of haloes even after several dynamical times, and we have tested that this difference in concentration cannot be accounted for by the difference in their mass distributions. This could be due to the fact that mergers are correlated, perhaps due to environmental dependences, or that mergers have a persistent effect on the internal structures of haloes, or a combination thereof. Determining the nature of this effect is worthy of a distinct study in its own right. The fluctuations in the scale radii and concentrations following minor mergers, which happen frequently for most haloes, are also a likely source of spread in the present-day values of halo internal properties, which we investigate in Section II.D.5.

5. Irreducible Scatter Due to Stochastic Mergers

With our improved understanding of mergers, we examine the role that these events play in producing the present-day concentrations and scale radii of halo samples with fixed masses. We have shown in Fig. II.5 that the impact on $r_{\rm s}$ from major mergers and even minor mergers is significant compared with the scale of the overall $r_{\rm s}$ evolution in the entire history, and persists over a considerable amount of time (several dynamical times, meaning several Gyr). Therefore, we expect that the cumulative response of a halo to the merger events in its mass assembly history is

crucial to the determination of the scale radius and concentration of the halo. Moreover, both the relative sizes of mergers and the temporal distribution of these mergers are important in determining present-day concentration and scale radius.

In the top panel of Fig. II.6, we show the scatter in $\log c$ for each present-day mass sample, and the remaining scatter after further dividing the samples into quintiles by $a_{1/2}$. There is a scatter of approximately 0.1–0.2 dex in concentration with populations of haloes with both mass and $a_{1/2}$ fixed. The scatter increases with later half-mass scales, as there are more recent merger events for these haloes. This scatter originates from the variety of possible paths of mass assembly. In the middle panel, we examine the same samples, but exclude haloes that undergo major mergers since their half-mass scales. The resulting scatter in $\log c$ decreases in every sample, which is consistent with our conclusion that major mergers contribute to the uncertainty in today's concentration. The decrease is not as significant as one might naively expect from the large fluctuations in concentration caused by major mergers, because major mergers are rare events and impact a small fraction of the population. In the bottom panel, we further exclude all haloes that have stepwise mass increases with $\Gamma \geq 1.0$ since $a_{1/2}$, and the scatter is indeed further reduced. That a more stringent restriction on mergers further reduces scatter in concentrations strongly suggests that it is the mergers themselves that drive a significant portion of the scatter. The dependence of the scatter on the half-mass scale is largely removed by excluding these mass increase events, which confirms that different frequencies of mergers are the cause of this dependence. It is reasonable to expect that when even more stringent limits are put on the mass increase rate, the scatter will be further reduced; however, we are unable to test this explicitly due to limited sample sizes. As a supplement to Fig. II.6, in Appendix A we show the dependence of the concentration–mass relation on the half-mass scale.

It is tempting to synthesize the present-day concentration from the full mass assembly history, by superposing the effect of each merger event upon pseudo-evolution. However, we show in Fig. II.7 that even small deviations from pseudo-evolution in mass can cause large fluctuations in concentration. This sensitivity of the concentration to small mergers and the stochastic nature of mergers make it virtually impossible to predict the concentration of an individual halo from its formation history without some uncertainty. In Fig. II.7, we select the five haloes from the random catalog that have the most quiescent mass assembly histories in the last five dynamical times. We do this by minimizing the deviation of the mass assembly history from pure pseudo-evolution. We

calculate the forward pseudo-evolution of mass from the halo properties at $T(a; a_{\rm ref} = 1) = -5$, which is marked by the vertical dotted dark blue lines in Fig. II.7, and quantify the deviation in mass evolution by $\sum |M_{\rm hist}/M_{\rm pseudo} - 1|$, where $M_{\rm hist}$ is the actual evolution, $M_{\rm pseudo}$ is the forward pseudo-evolution for each halo, and the sum is taken over the available snapshots in the last five dynamical times. The dark blue lines in the figure show the logarithmic deviation in mass. For these haloes with quiescent mass assembly histories, we then compare between the actual and pseudo-evolution of concentration during the last two dynamical times, since $T(a; a_{\rm ref} = 1) = -2$ (vertical dashed pink lines), to exclude the effect of mass evolution before the controlled period. The comparison of concentrations is shown as pink lines in Fig. II.7. In the first panel, we also show the 68% range of the absolute deviation from both mass and concentration pseudo-evolutions for the entire random sample.

From the figure it is apparent that even selecting the most quiescent haloes in our sample, which are usually considered relaxed, does not greatly reduce fluctuations in concentration. This shows that even very minor mass accretion can affect halo structures. The fluctuations in concentration seen in Fig. II.7 could also be partly due to the finite number of snapshots available from the simulation, which leaves events that happen between the discrete snapshots undetected. We also notice that for some haloes (e.g., Halos 2 and 3), the concentration evolution has a general trend that deviates from the pseudo-evolution prediction. This is likely due to the oversimplified assumptions in the pseudo-evolution model, such as an NFW profile and an isolated halo, which may not hold true in simulations [48]. The environments of individual haloes and further details of mergers, such as the relative velocities of the progenitors, the exact orbit of the incoming object, the detailed density profiles of each progenitor, are beyond the scope of this work, and might also have caused part of the uncertainty that we observe.

At this point we reflect on the limited ability to predict halo concentration with a linear regression of the mass assembly history in Section II.C.3, and conclude that this is unsurprising, because a fixed set of linear coefficients is naturally incapable of describing a convolution of merger responses at different times. Also, in the top panel of Fig. II.1, the concentration of the cluster-size halo sample is less correlated with the step-wise mass assembly history than for the less massive samples, probably ascribable to its higher frequency of merger events. On the other hand, the higher frequency of mergers in the cluster-size sample causes its stronger anticorrelation between

the concentration and the mass increment at late times in the bottom panel of Fig. II.1. The oscillatory behavior in the bottom panel of Fig. II.1 has approximately the same timescales as the oscillation of the concentration and scale radius in Fig. II.4, and it is now apparent that it arises from merger responses.

We have demonstrated that mergers play a vital role in shaping the internal structures of haloes, and merger events that happen at different epochs trigger responses with nontrivial forms, preventing a simple description of the combined end result, and contribute to the scatter in the scale radius and hence concentration.

E. Discussion and Summary

In this study, we investigate the connection between halo concentration and halo mass assembly history using the halo catalogs and merger trees from the Dark Sky Simulations. In particular, we scrutinize the effect of mergers on the subsequent evolution of halo concentration. We summarize our primary results as follows:

- Conventionally defined halo formation times, such as the scale factor at which a halo reaches 50% of its contemporary mass, exhibit significant correlations with the present-day halo concentration. In fact, the same holds true for the broad range of mass fractions between approximately 30% and 70%. A linear combination of $a(m_i)$, where m_i's are different choices of mass fractions, correlates with present-day concentration better than any individual $a(m_i)$, but still does not fully account for the scatter in concentration at a fixed halo mass. The same conclusions apply when we use the mass fractions at different times, $m(a)$, instead of $a(m)$. For more details, see Fig. II.1.
- Major mergers induce dramatic changes to halo concentrations. These responses linger over a period of several dynamical times, corresponding to many Gyr. The evolution of concentration due to a merger can be associated with the orbital dynamics of the merger and is largely universal. Minor mergers have similar, but less dramatic effects on concentration compared with major mergers. In the absence of merger events, pseudo-evolution causes a gradual increase in halo

concentration and halo mass (Fig. II.2), in agreement with Ref. [49]. See Fig. II.3, Fig. II.4, and Fig. II.5.

- The cumulative effect of major mergers and frequent minor mergers leads to a scatter in concentration at fixed halo mass and fixed formation time (any conventional definition). At fixed halo mass, the scatter can be reduced from 0.2 dex to below 0.1 dex when we control for both formation time and merger events. Even minor mergers impart non-negligible alterations to concentrations. Haloes with quiescent mass assembly histories experience fewer fluctuations in concentration, but still with an irreducible scatter, due to unresolved small mergers. See Fig. II.6 and Fig. II.7.

In this work, we have developed a further understanding of the relation between halo concentrations and mass assembly histories. Our results show that the correlation strengths with concentration at multiple intermediate epochs of the assembly history are similar and relatively high, in accord with previously found concentration–formation time relations [204, 227]. Our findings support the use of the half-mass scale, $a_{1/2}$, as an effective definition of formation time, whereas a variety of similar formation time definitions would yield similar insight into concentrations. However, we also argue that such simple characterizations of the mass assembly history inevitably omit information on halo structure and leave a non-negligible residual scatter.

We find that merger events during the assembly of haloes contribute to the scatter in the concentration–formation time relation (at fixed halo mass), as was suggested by the results of, e.g., Refs. [204, 113], and [153]. We broaden the discussion of the impact of recent mergers on the measurement of concentration in Ref. [112], confirming their explanation of the features in the merger response with a case study of the orbital processes during a merger, and these fluctuations in concentration induced by mergers also lead to the non-monotonic relation between concentration and formation time observed by Ref. [112]. We recognize the significant effect of mergers on halo concentrations, which greatly exceeds the secular evolution during quiescent periods. The effect of mergers lasts for several Gyr (a few dynamical times). Our results also establish the universality of halo responses to merger events.

These results can have important implications for the interpretation of observations, as the observed density profiles of the dark component of clusters are systematically dependent on the merger history. The concentration–mass relation is broadly adopted for inferring concentrations

from mass measurements, comparing measured concentrations against theoretical predictions, or modeling other halo properties with concentrations [35, 54, 116, 19, 122, 102]. Based on our findings regarding the scatter around the mean relation due to mergers, we advise caution in the application of the concentration–mass or concentration–formation time relation without taking these effects into account.

Our study provides insight into secondary halo biases, commonly known as halo assembly bias [61, 104, 118], the dependence of halo clustering on halo properties other than mass. Ref. [205] first found that with fixed masses above the typical collapse mass, haloes with lower concentrations cluster more strongly than haloes with higher concentrations. Ref. [98], for example, found that the scale radii of haloes evolve very differently in regions of different environmental densities. Our findings suggest that these are primarily due to the suppression of concentration by merger events, which happen more frequently in denser environments. We expect similar coupling of the environmental preference of mergers and the impact of mergers on other secondary halo properties to be present.

Our analyses are performed at the halo level, which introduces a dependence on the halo finding algorithm. We limit our characterization of the mass assembly history to linear descriptions, and do not propose a mathematical model of the concentration. We are also unable to resolve all merger events, and further details, including the initial profiles, initial velocities, and trajectories of merging objects, are beyond the scope of this work. Each of these important issues merits further study. More sophisticated statistics or machine learning techniques might be more effective in extracting information on concentrations from assembly histories. Using explicit mathematical descriptions of concentration responses to mergers, together with a comprehensive demographic study of merger events with even higher mass and temporal resolutions is a possible way of improving predictions of concentrations. We are hopeful that such follow-up studies could greatly enhance our understanding of halo formation and structure.

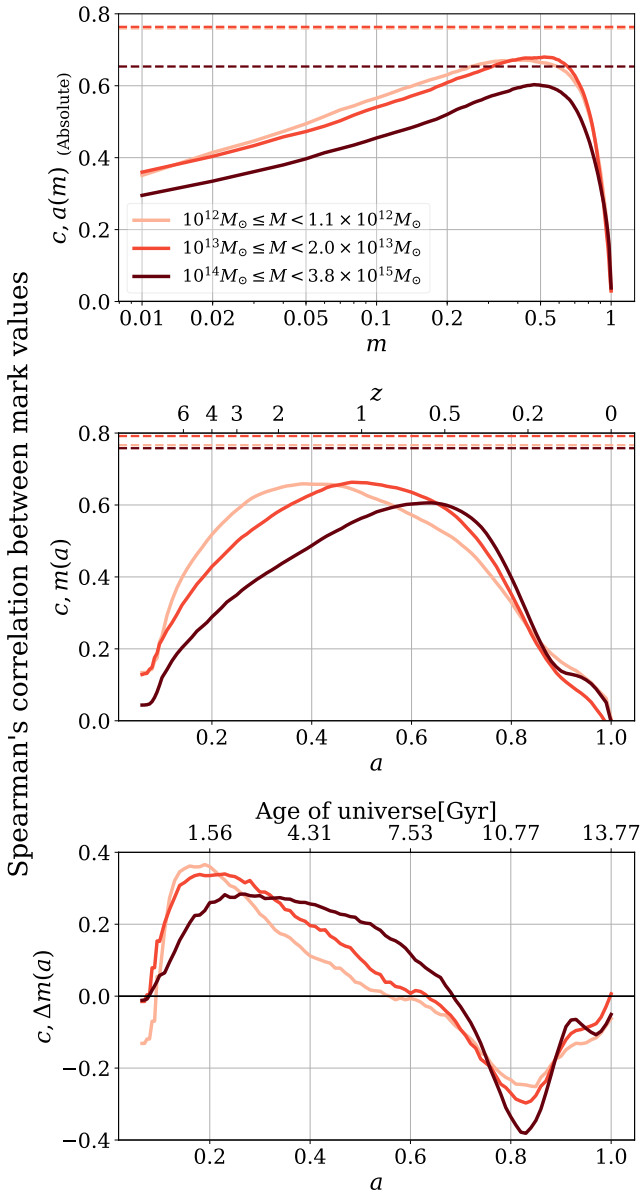
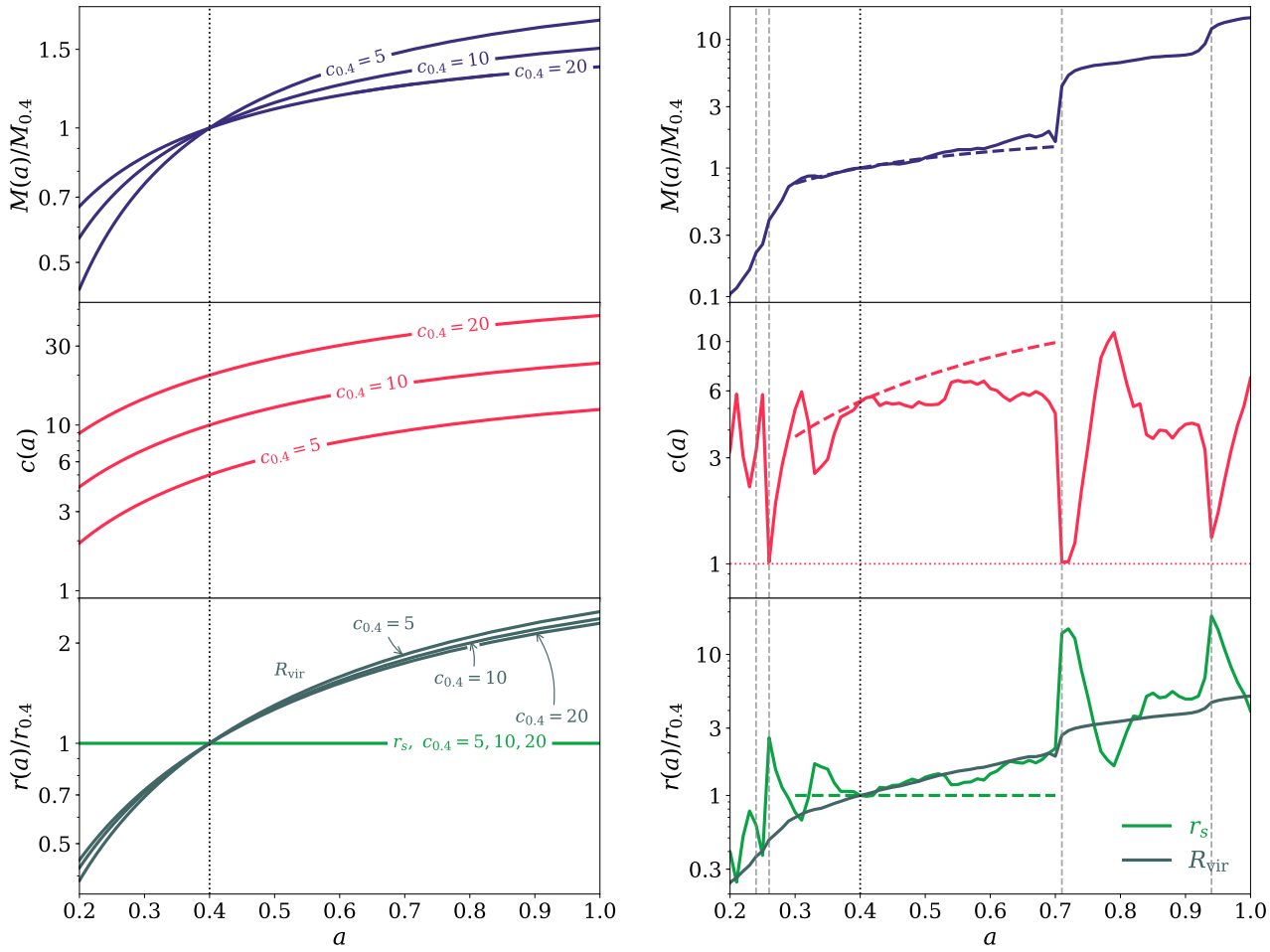


Figure II.1: Spearman's correlation between the mark values of haloes' present-day concentrations and mass assembly histories. In the top panel, mass assembly history is characterized by the epoch $a(m)$ at which a fraction m of the present-day mass has been first assembled. The *absolute values* of the otherwise negative correlation coefficients are shown in this panel, and m is shown in logarithmic scale. In the middle panel the mass assembly history is alternatively characterized by the fraction of the present-day mass, $m(a) = M(a)/M(a = 1)$, that has been assembled by the time of each a, and in the bottom panel by $\Delta m(a)$, the step-wise increment in m at each a.

Figure II.1: *(cont.)* The top x-axes of the middle and bottom panels show the corresponding redshift and age of the universe respectively. In all three panels, the different colors represent results for the different mass samples, as is labeled in the top panel. The solid lines show the Spearman's rank-order correlation coefficients between $\mathcal{M}(c)$ and $\mathcal{M}(a(m))$ (top panel), $\mathcal{M}(c)$ and $\mathcal{M}(m(a))$ (middle panel), or $\mathcal{M}(\Delta m(a))$ (bottom panel). Each horizontal dashed line in the top panel shows the Spearman's correlation between $\mathcal{M}(c)$ and the mark value of the optimal linear combination of $a(m)$'s for the corresponding mass sample, while the horizontal dashed lines in the middle panel indicate the Spearman's correlations between $\mathcal{M}(c)$ and the mark values of the optimal linear combinations of $m(a)$'s. These optimal linear combinations contain most but not all of the information about the present-day concentration.



(a) Pseudo-evolution

(b) Actual evolution

Figure II.2: (a) In the left column, we show the change in mass (top panel), concentration (middle panel), and virial radius and scale radius (bottom panel) due to pseudo-evolution, choosing $a = 0.4$ as the reference state, denoted with the subscript "0.4", and marked by vertical black dotted lines in the panels. In the bottom panel, the y-axis indicates $r(a)/r_{0.4}$, where r is either the virial radius $R_{\rm vir}$ or the scale radius $r_{\rm s}$. The change in mass and radii are plotted in terms of the ratio between the pseudo-evolved values and the values at $a = 0.4$. Each line of pseudo-evolution is labeled with the corresponding concentration at the reference point $a = 0.4$, as the ratios and concentration are only functions of the concentration, independent of halo mass. It can be observed that significant growth in both halo mass and halo concentration can be associated with pseudo-evolution.

Figure II.2: *(cont.)* (b) The right column is similar to the left column, but shows the actual evolution of an individual halo's mass, concentration, and virial radius and scale radius, as functions of the scale factor a. The vertical gray dashed lines mark the major mergers identified by Consistent Trees, and the horizontal dotted line in the middle panel indicates $c = 1$ to guide the eye. Besides the actual evolution, we also show the pseudo-evolution for comparison. In each panel, the dashed curve of the same color as the solid curve shows the corresponding quantity pseudo-evolved from the state at $a = 0.4$ (vertical black dotted line), between $a = 0.3$ and $a = 0.7$. In the bottom panel, only the pseudo-evolution of $r_{\rm s}$, which is a constant function of time, is shown, while the pseudo-evolution of $R_{\rm vir}$ is omitted for clarity.

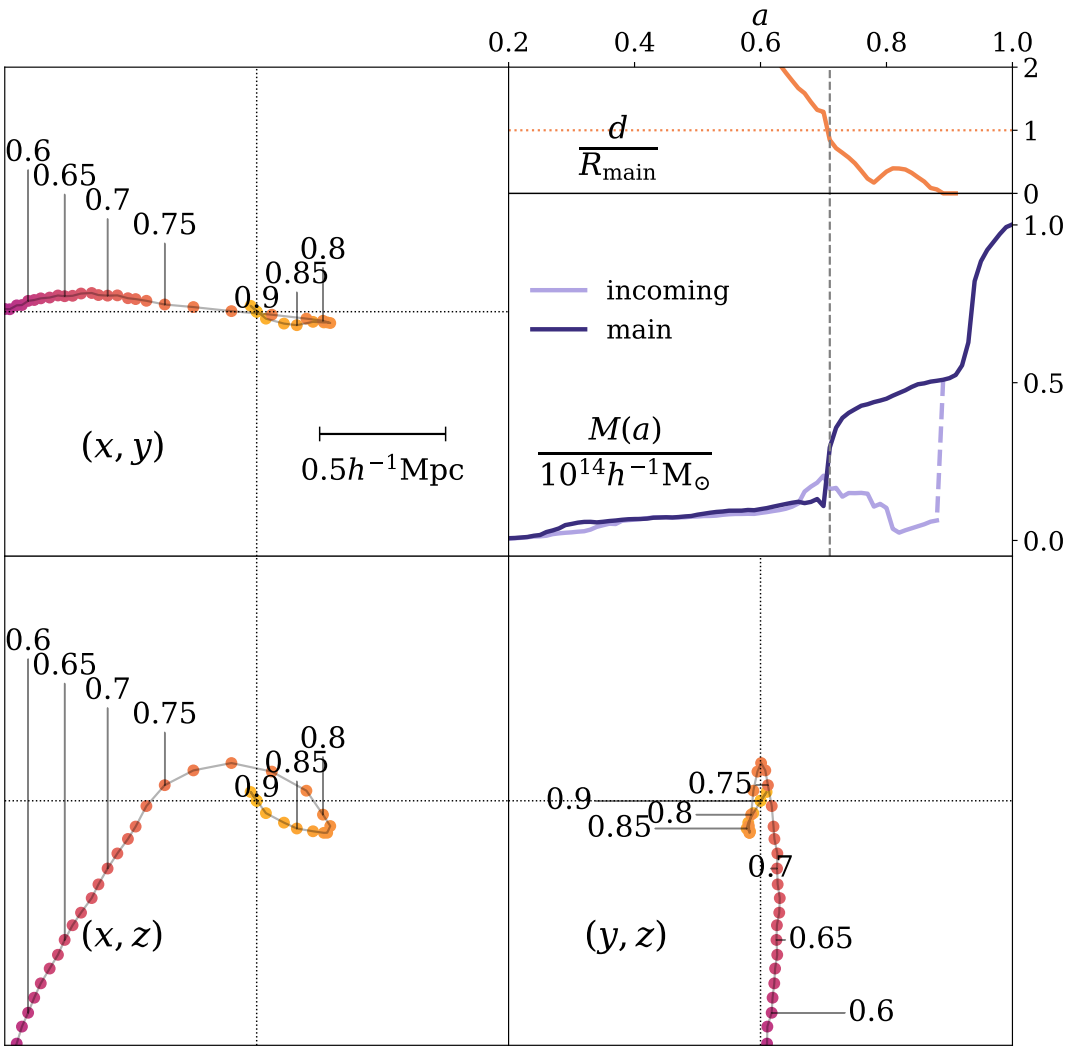


Figure II.3: In this figure, we show the process of the major merger that the halo in Fig. II.2 undergoes at $a = 0.71$. The top left and bottom panels show the orbit of the incoming progenitor around the main progenitor in the three projected planes respectively, displaying the comoving space from $-1h^{-1}$ Mpc to $1h^{-1}$ Mpc in each direction, and the comoving length scale of $0.5h^{-1}$ Mpc is shown in the upper left panel for visual clarity. Each point in an orbit represents the state in a different snapshot, color coded from dark to bright with the increase of time; the scale factor a is labeled at several points. The upper right panel shows the time span between $a = 0.2$ and $a = 1$, with the time of the major merger marked by a gray vertical dashed line.

Figure II.3: *(cont.)* The lower part of this panel tracks the mass changes of the main and incoming progenitors in units of $10^{14}h^{-1}\,\mathrm{M}_{\odot}$. The incoming halo's evolution ends when it is completely disrupted and is no longer identified as an object, and the transition is shown as a dashed line. The increase in mass in the main branch afterwards is due to another major merger that follows. The upper part of the same panel shows the evolution of d/R_{main}, where d is the distance between the centers of the two progenitors, and R_{main} is the virial radius of the main progenitor. The ratio d/R_{main} decreases below 1 at around a_{MM}, marked by the horizontal dotted line, and reaches 0 as the incoming object disappears.

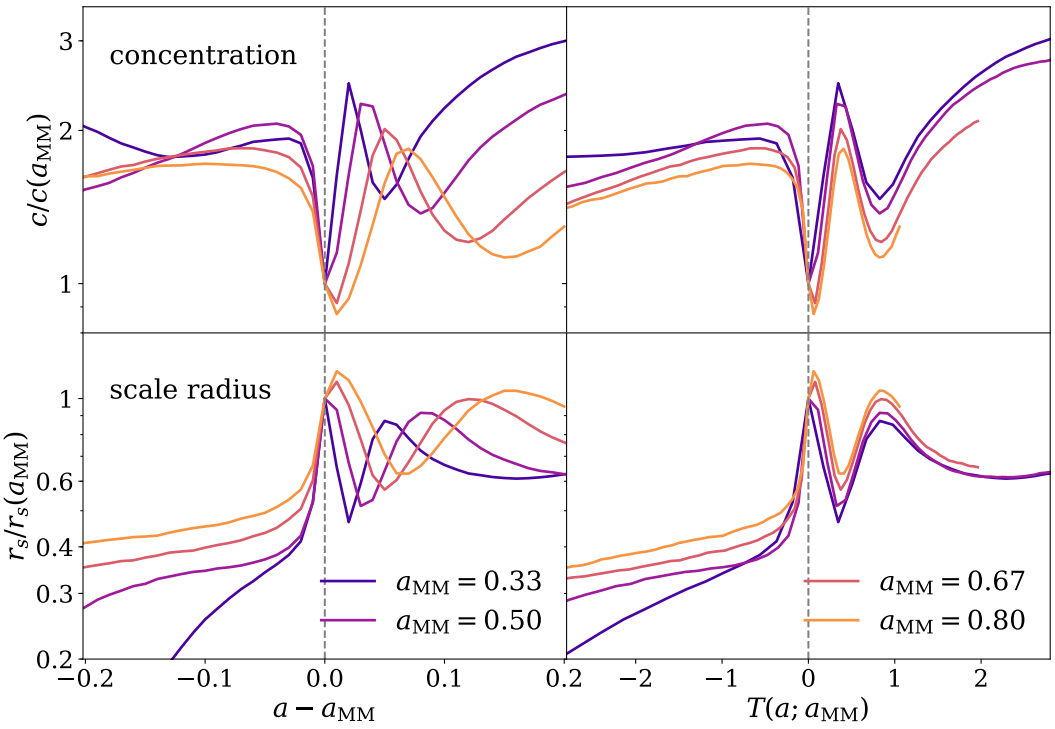


Figure II.4: Median response to major mergers that happen at different times. The top row displays the concentration, and the bottom row displays the scale radius, both in logarithmic scale and normalized by the value at the time of merger. In the left column, time is measured in terms of the scale factor a, shifted with respect to $a_{\rm MM}$, while in the right column, time is measured in units of dynamical times, with the merger time as the reference point. The groups are color coded by their respective $a_{\rm MM}$. The time of merger is marked by a vertical dashed line in each panel. In the right column, some of the lines are truncated due to the limited time range of the simulation. The response of haloes of different masses and major merger times are remarkably similar when scaled by dynamical time.

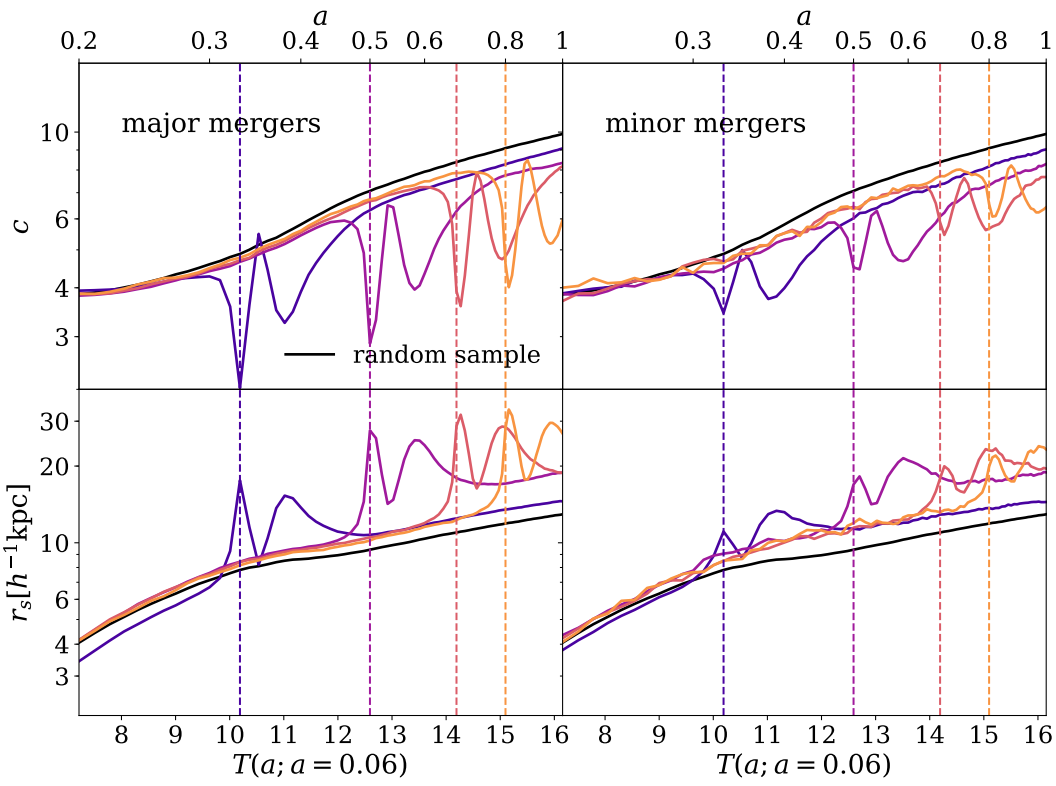


Figure II.5: This figure compares major mergers with minor mergers that happen at the same epochs, which are marked by vertical dashed lines, each with the same color as the corresponding evolution curve. Similar to Fig. II.4, the halo evolution is tracked in terms of concentration in the top row and scale radius in the bottom row. Time is measured in units of dynamical times, adopting $a_{\rm ref} = 0.06$, and the corresponding scale factor a is labeled at the top. The minor merger events are selected from the random catalog by their rates of fractional mass increase Γ, defined in Equation 6, $1.0 \leq \Gamma \leq 1.5$, and no major mergers within $\pm 0.25 \tau_{\rm dyn}$ of the time step of interest. The left column shows the median evolution of each major merger sample, and the right column shows those of the minor merger samples. In every panel, the solid black curve depicts the median evolution of the random sample for comparison.

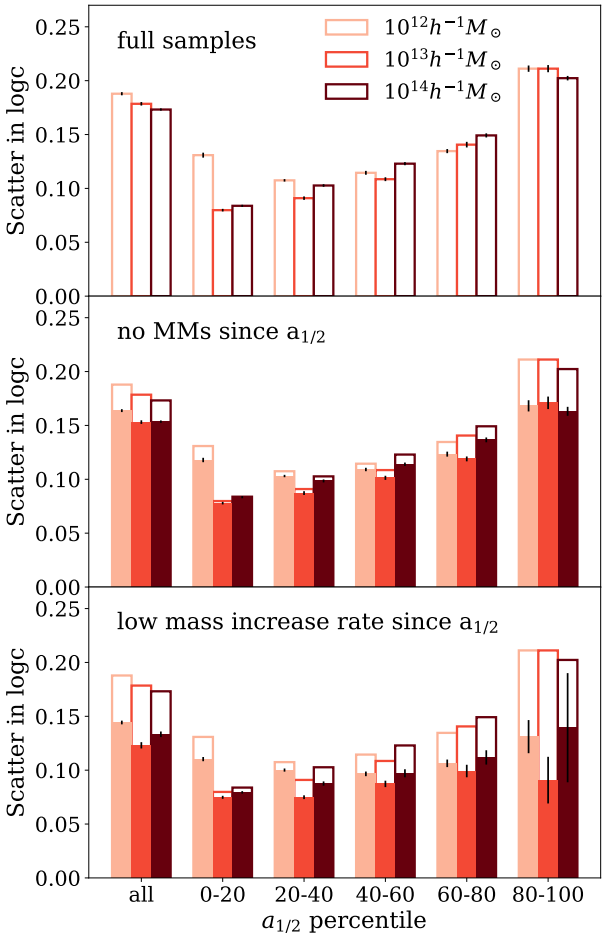


Figure II.6: In this figure, we show the logarithmic scatter in concentration for the present-day mass samples. The present-day mass is color coded and labeled in the bottom panel. In the top panel, the first group of bars shows the scatter for the entire samples, while the other groups are subsamples selected by their half-mass scale percentiles within each mass sample. The error bars are calculated using bootstrap resampling. The same bars are also shown in the two lower panels for visual guidance. The filled bars in the middle panel shows the scatter for the same samples, but excluding haloes that undergo major mergers after the half-mass scale. The filled bars in the bottom panel adopt a more stringent selection criterion, excluding haloes that have mass increase events with $\Gamma \geq 1.0$.

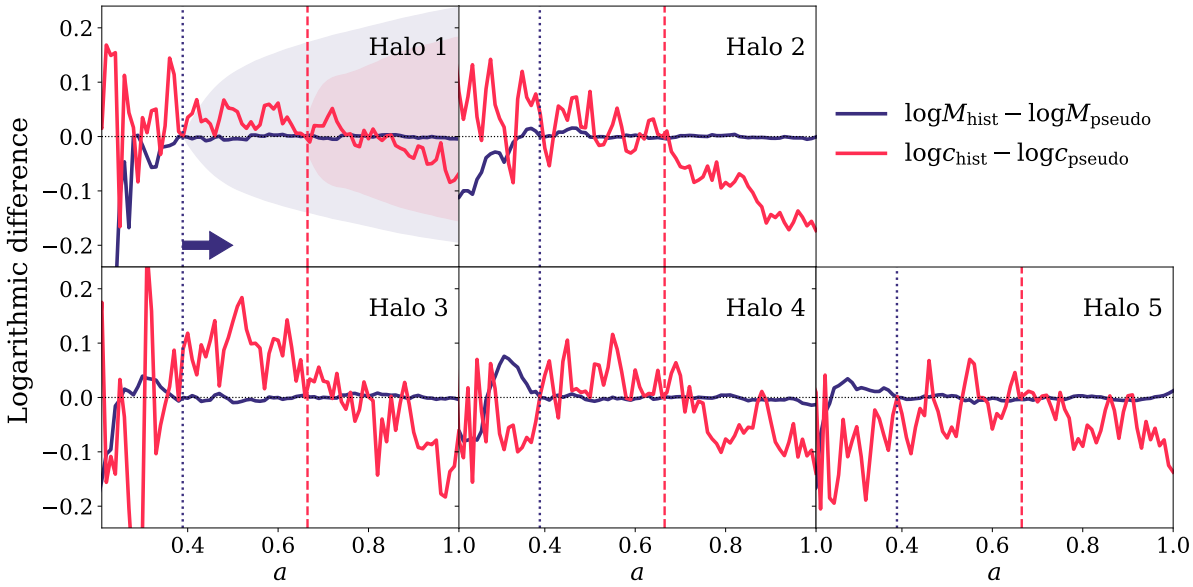


Figure II.7: Comparison of the actual evolution against pseudo-evolution for five individual haloes in the simulation. These haloes are selected to have the least deviation from pseudo-evolution in mass in the last five dynamical times, which is marked by the vertical dotted dark blue line in each panel, and the dark blue arrow in the first panel. The concentration is compared against the forward pseudo-evolution from two dynamical times before $a = 1$ (vertical dashed pink lines). We show the difference in logarithmic space between the pseudo-evolution of the mass and concentration. In the first panel, the shaded regions show the 68th percentile of the absolute deviation from pseudo-evolution as a function of time, for the entire random sample, in the time ranges of interest for the mass and concentration respectively.

III. Probes for Secondary Effects in Galaxy–Halo Connection

This chapter is originally published as: Wang, K., Mao, Y.-Y., Zentner, A. R., van den Bosch, F. C., Lange, J. U., Schafer, C. M., Villarreal, A. S., Hearin, A. P., Campbell, D. (2019), Monthly Notices of the Royal Astronomical Society, 488, 3541.

Minor modifications have been made to the text. The inclusion of this article in this dissertation is in compliance with the copyright policies of the journal.

In cosmological studies, we often exploit the statistical connection between galaxies and their haloes. Most models for this connection ignore the possibility that galaxy properties may be correlated with halo properties other than halo mass, a phenomenon known as galaxy assembly bias. And yet, it is known that such correlations can lead to systematic errors in the interpretation of survey data that are analyzed using traditional halo occupation models. At present, the degree to which galaxy assembly bias may be present in the real Universe, and the best strategies for constraining it remain uncertain. We study the ability of several observables to constrain galaxy assembly bias from redshift survey data using the decorated halo occupation distribution (dHOD), an empirical model of the galaxy–halo connection that incorporates assembly bias. We cover an expansive set of observables, including the projected two-point correlation function $w_{\rm p}(r_{\rm p})$, the galaxy–galaxy lensing signal $\Delta\Sigma(r_{\rm p})$, the void probability function VPF(r), the distributions of counts-in-cylinders $P(N_{\rm CIC})$, and counts-in-annuli $P(N_{\rm CIA})$, and the distribution of the ratio of counts in cylinders of different sizes $P(N_2/N_5)$. We find that despite the frequent use of the combination $w_{\rm p}(r_{\rm p})$ + $\Delta\Sigma(r_{\rm p})$ in interpreting galaxy data, the count statistics, $P(N_{\rm CIC})$ and $P(N_{\rm CIA})$, are generally more efficient in constraining galaxy assembly bias when combined with $w_{\rm p}(r_{\rm p})$. Constraints based upon $w_{\rm p}(r_{\rm p})$ and $\Delta\Sigma(r_{\rm p})$ share common degeneracy directions in the parameter space, while combinations of $w_{\rm p}(r_{\rm p})$ with the count statistics are more complementary. Therefore, we strongly suggest that count statistics should be used to complement the canonical observables in future studies of the galaxy–halo connection.

A. Introduction

In the concordance ΛCDM model of the Universe [91, 18, 141, 143, 2], galaxies reside in dark matter haloes [208, 23], which form around peaks in the primordial dark matter density field [8, 24, 166, 165, 222]. In practice, the abundance, clustering, and structure of dark matter haloes have been precisely documented by high-resolution, gravity-only N-body simulations of cosmological structure growth [132, 123, 177, 180]. Halo occupation models use empirical data to link galaxies to haloes in a statistical sense. They are useful because they provide a convenient means to compare observed galaxy clustering statistics with theoretical predictions without a complete theory of galaxy formation and evolution. This, in turn, is useful because one can use such models to test cosmological models using data on non-linear scales, and because empirical models distill the formidable amount of information available in survey data into a relatively simpler galaxy–halo relationship that can be used to inform models of galaxy formation and evolution. Ref. [203] provide a contemporary review of these models.

The term *assembly bias* has, unfortunately, taken on several related but distinct meanings in the literature. The clustering of dark matter haloes is a strong function of halo mass [83, 123], but it has become clear over the last decade that haloes cluster as a function of a number of other properties [61, 205, 60, 104, 118, 213]. Ref. [61] first studied the age-dependence of halo clustering, which led to the term *assembly bias*, but because many halo properties are correlated with formation history, the dependence of halo clustering on many other properties (e.g., concentration, spin, and so on) is often loosely referred to as *assembly bias* or *halo assembly bias* as well. Ref. [118] advocate referring to these dependences as *secondary biases*[1]. This nomenclature is clearer because the secondary biases do not necessarily have a clear origin in the correlations of halo properties, such as concentration, with conventional measures of halo formation history.

Assembly bias of observed galaxies as well as simulated halo populations has received significant attention in the recent literature regarding the analysis and interpretation of galaxy survey data. Assembly bias may challenge survey analyses because it may induce (1) systematic errors in the inferred galaxy–halo relationship inferred from survey data [224] and/or (2) biases in inferred cosmological parameters[45, 210, 120]. On the other hand, an unambiguous detection of assembly

[1]The *primary bias* is the strong dependence of halo clustering on halo mass.

bias in survey data may pave the way to a richer and more complete understanding of the connection between galaxies and their host dark matter haloes.

In the context of the interpretation of galaxy surveys, assembly bias can lead to the following possibility. Galaxies in a certain luminosity sample may form in haloes with a probability that depends upon not only the mass of the halo, but on any number of halo properties. If this happens, then the resultant clustering of any galaxy sample must be interpreted within the context of a model that incorporates not only the mass dependence of halo clustering, but also the secondary biases, as is done in e.g., Refs. [31, 223, 100, 155, 47]. This case is sometimes loosely referred to as *galaxy assembly bias.* Signals of galaxy assembly bias are also present in hydrodynamical simulations [214, 6, 26]. The issue facing survey data analysis is that the vast majority of studies treat survey data using models that assume that galaxies of a particular luminosity (or other galaxy properties that determine the sample selection) form within haloes with a probability that depends upon only the mass of the halo (and no other halo property). Thus, these analyses account only for the mass dependence of halo clustering [203].

The most widely used empirical models for interpreting survey data include the halo occupation distribution (HOD) [16] and the conditional luminosity function (CLF) [216]. Both of these models in the original (and standard) forms assume that galaxies of a particular type reside in haloes with a probability that depends only on the masses of the haloes. Both the HOD [230, 221, 68, 64, 97, 65, 223] and the CLF [187, 215, 38, 188, 30, 95] have been used successfully to interpret a variety of observational samples. On the other hand, subhalo abundance matching (AM or SHAM, e.g., [92, 175, 184, 181, 75]) has the power to naturally incorporate assembly bias, by matching galaxy properties to halo properties (e.g., $V_{\rm peak}$) that can assume varying values at a fixed halo mass. The success of SHAM in interpreting galaxy survey data [11, 12, 150, 100] is striking considering the simplicity of its assumptions.

In this paper, we will use an expanded form of the HOD, known as the decorated HOD (dHOD, [74]). The dHOD builds upon the traditional HOD by adding parameters that enable tunable levels of galaxy assembly bias. We limit our treatment to the HOD and dHOD for specificity and simplicity.

We explore the utility of several galaxy survey observables to constrain assembly bias within the context of simple dHOD [74] models. Aside from the overall galaxy number density $n_{\rm gal}$, which we

use as a basic constraining observable throughout our analyses, the observables that we explore are the projected galaxy correlation function $w_{\rm p}(r_{\rm p})$, the excess surface density inferred from galaxy–galaxy lensing $\Delta\Sigma(r_{\rm p})$, the void probability function VPF(r), galaxy counts-in-cylinders $P(N_{\rm CIC})$, galaxy counts-in-annuli $P(N_{\rm CIA})$, and the probability distribution of the ratio between $N_{\rm CIC}$ of different cylinder sizes $P(N_2/N_5)$. In this first study of the subject, we intentionally avoid utilizing satellite kinematics (SK), redshift space distortions (RSD), or other observables that require a detailed model of galaxy velocities relative to haloes. Treating such statistics requires additional modeling and additional assumptions that can greatly complicate such a study. For the observables we study, we examine and compare their effectiveness at constraining not only assembly bias within the context of a dHOD model, but nearly all HOD parameters.

The remainder of this chapter gives the details necessary to support the summary of our findings stated in the previous paragraph. In Section III.B, we describe the simulation that we use, the dHOD models within which we work, the observables we consider, and our approach to estimating parameter constraints. In Section III.C, we present our results in detail. We discuss our results in the context of the contemporary literature, draw broad conclusions, and propose future steps in Section III.D.

B. Methods

In this section, we give the details of our analysis procedures. This includes a discussion of the simulation that we use, the dHOD models that we explore, the observables that we consider, our methods for estimating parameter constraints, and our estimates of the covariance matrices used in our analyses.

1. Simulation

In order to mitigate the limitations of analytic estimates of clustering and lensing statistics [186], the calculations that we perform in this paper are based upon N-body simulations of the formation of structure in a concordance cosmological model. These simulations evolve dark matter

particles under the influence of gravity from initial over-densities in the early universe to the present day. In particular, the analysis in this work utilizes the Bolshoi Planck simulation[2] [87, 13, 14, 154]. Bolshoi Planck is a dark matter only simulation within a cubic box of length $250\,h^{-1}$ Mpc, which adopts values of cosmological parameters from Ref. [141], namely $\Omega_\Lambda = 0.693$, $\Omega_\mathrm{m} = 1 - \Omega_\Lambda = 0.307$, $\Omega_\mathrm{b} = 0.048$, h = 0.7, $n_\mathrm{s} = 0.96$, and $\sigma_8 = 0.82$. The simulation contains 2048^3 particles, implying a particle mass of $m_\mathrm{p} = 1.55 \times 10^8\,h^{-1}\,\mathrm{M}_\odot$.

We use the `bolplanck` halo catalog included with the `Halotools` software package[3] [71], which also provides an implementation of customizable dHOD models.

The `bolplanck` halo catalog was produced from the Bolshoi Planck simulation using the `ROCKSTAR` halo-finder [13]. To compute lensing observables, we use the particle catalog included with `Halotools` which contains 10^6 randomly-selected particles from the Bolshoi Planck volume, and make a downsampled catalog containing $\sim 10^5$ particles with an acceptance rate of 0.1 for runtime considerations. We have tested that the measured $\Delta\Sigma(r_\mathrm{p})$ is not sensitive to the downsampling, and the noise introduced in this process is accounted for in our covariance matrix (see Section III.B.5). The catalogs are included in `Halotools` version `halotools_v0p4`, which adopts the virial definition of haloes, and we work at $z = 0$, corresponding to an overdensity parameter $\Delta_\mathrm{vir} = 333$ with respect to the mean matter density of the Universe.

2. Halo Occupation Model

We describe the galaxy–halo connection using the HOD and the dHOD. Both of these models specify the probability for a halo of mass M_vir to host N_cen central galaxies and N_sat satellite galaxies above a certain threshold stellar mass, $P(N_\mathrm{cen}|M_\mathrm{vir})$ and $P(N_\mathrm{sat}|M_\mathrm{vir})$ respectively. Central and satellite galaxies are considered separately because central galaxies reside in the potential wells of host haloes while satellite galaxies are associated with subhaloes and experience different physics of formation and evolution. It is well known that subhaloes experience very distinct evolution from host haloes and thus have demographics that are distinct from host haloes [92, 229, 225]. Moreover, numerous observations, using many different approaches, have established that central galaxies and satellites have different properties and can be described as two distinct populations

[2] Available at www.cosmosim.org/cms/simulations/bolshoip

[3] halotools.readthedocs.io

[136, 53, 144, 220, 206, 185].

The central galaxy occupation is modeled as a Bernoulli random variable, which takes the value 1 with probability p and the value 0 with probability $1 - p$, with $p = \langle N_{\rm cen}|M_{\rm vir}\rangle$. Satellite galaxies follow a Poisson distribution with first moment $\langle N_{\rm sat}|M_{\rm vir}\rangle$.

In principle, the formalism we use in this work is the same as in Ref. [223] which, in turn, was chosen to mimic the analysis of Ref. [221]. However, we adopt the implementation of the model in `Halotools`, and introduce some subtle modifications which will be elaborated on in Section III.B.2.b.

a. Standard HOD

In the standard HOD, which does not account for any potential galaxy assembly bias, the mass of a halo solely determines the galaxy occupation. The mean central and satellite galaxy occupations vary with halo mass according to

$$\langle N_{\rm cen}|M_{\rm vir}\rangle = \frac{1}{2}\left(1 + {\rm erf}\left[\frac{\log(M_{\rm vir}) - \log(M_{\rm min})}{\sigma_{\log M}}\right]\right), \tag{7}$$

$$\langle N_{\rm sat}|M_{\rm vir}\rangle = \left(\frac{M_{\rm vir} - M_0}{M_1}\right)^{\alpha} \times \langle N_{\rm cen}|M_{\rm vir}\rangle, \tag{8}$$

[230], where $M_{\rm min}$ is the mass at which a halo has a 50% probability of hosting a central galaxy; $\sigma_{\log M}$ is a measure for the scatter in the stellar mass–halo mass relation that determines the steepness of the $\langle N_{\rm cen}|M_{\rm vir}\rangle$ transition from zero to unity; M_0 is the truncating mass, below which $\langle N_{\rm sat}|M_{\rm vir}\rangle = 0$; the mass M_1 indicates the halo mass at which there is, on average, one satellite[4] if a central is present; and, finally, α is the index of the satellite occupation power law. Note that Eq. (8) expresses the probability of having a satellite galaxy for a halo with mass $M_{\rm vir}$, after marginalizing over the central occupation. The first term on the right hand side indicates the mean satellite occupation in haloes with a central galaxy, while the second term modulates this occupation by the probability for a halo to contain such a central. Hence, the presence of a central boosts the probability for a halo to host satellite galaxies. Note, though, that for individual haloes a central galaxy is not strictly required for satellites to be present. Although this modulation with $\langle N_{\rm cen}|M_{\rm vir}\rangle$ is fairly common [230, 221, 223], we emphasize that it is not used by all authors.

[4] More accurately, this mass is $M_1 + M_0$, but M_0 is typically much smaller than M_1.

These specifications, along with the assumptions that the central galaxy HOD is a Bernoulli distribution and the satellite galaxy HOD is a Poisson distribution, suffice to specify fully the halo occupation statistics of dark matter haloes in a standard HOD model without assembly bias.

b. Decorated HOD

Galaxy assembly bias can be incorporated into the HOD formalism in any number of ways. For a secondary halo property x, (e.g., concentration, spin, etc.), one can specify a functional form for the probability distributions $P(N_{\rm cen}|M_{\rm vir}, x)$ and $P(N_{\rm sat}|M_{\rm vir}, x)$. In such a generalized HOD, the clustering of galaxies can be altered if halo clustering depends upon secondary property x. The decorated HOD (dHOD, [74]) is one way of incorporating assembly bias into the HOD formalism such that integrating the dHOD probability distributions over the secondary properties of interest yields the standard HOD.

In the present paper, we use a simple variation of the dHOD as an illustrative model. In particular, we divide haloes into two categories based upon secondary halo property x. Haloes with higher values of x are assigned distinct HODs compared to haloes with lower values of x, with a pivot value of $x_{\rm piv}$. This is the "discrete halo subpopulations" example discussed in Section 4.2 of Ref. [74] and used to analyze SDSS data [1] in Ref. [223]. To specify completely the dHOD, we assume that $P(N_{\rm cen}|M_{\rm vir}, x)$ is a Bernoulli distribution and that $P(N_{\rm sat}|M_{\rm vir}, x)$ is a Poisson distribution, but that these distributions have first moments of

$$\langle N_{\rm gal}|M_{\rm vir}, x > x_{\rm piv}\rangle = \langle N_{\rm gal}|M_{\rm vir}\rangle + \delta N_{\rm gal}, \tag{9}$$

$$\langle N_{\rm gal}|M_{\rm vir}, x \leq x_{\rm piv}\rangle = \langle N_{\rm gal}|M_{\rm vir}\rangle - \delta N_{\rm gal}, \tag{10}$$

where we use the notation $N_{\rm gal}$ because this modification applies equally well to both the central and satellite occupations. We choose $x_{\rm piv}$ to be the median value of x at a given halo mass, so that each population contains 50% of all the haloes. In this toy model, assembly bias manifests itself as a step function in the secondary property x, though we expect that any assembly bias realized in nature would be represented by a smooth function of x. This simple model is practical in the sense that current data are not sufficient to constrain more complex models [223]; however, richer models

of assembly bias can naturally be accommodated within the dHOD framework [74] and future data sets are likely to enable constraints on richer models.

The differences $\delta N_{\rm gal}$ above are characterized by two assembly bias parameters, $A_{\rm cen}$ and $A_{\rm sat}$, both constructed so that they range between 1 and -1, in addition to the five standard HOD parameters. A list of the 7 dHOD parameters can be found in Table III.1. Positive values of $A_{\rm gal}$ indicate a positive correlation between galaxy number and halo property x (i.e., haloes with $x > x_{\rm piv}$ contain more galaxies, on average, than those with $x < x_{\rm piv}$), while negative values represent anti-correlation. When $A_{\rm gal} = 0$, the model reduces to the traditional standard HOD. Note that $A_{\rm cen}$ and $A_{\rm sat}$ vary independently of one another and do not necessarily have the same sign. The stipulations that the occupation of a halo never be negative, and the requirement that

$$\langle N_{\rm gal}|M_{\rm vir}\rangle = \int \langle N_{\rm gal}|M_{\rm vir}, x\rangle \, P(x|M_{\rm vir}) \, {\rm d}x \, , \tag{11}$$

with $P(x|M_{\rm vir})$ the probability distribution for x given $M_{\rm vir}$, implies that

$$\delta N_{\rm cen} = A_{\rm cen} \min \left[\langle N_{\rm cen}|M_{\rm vir}\rangle, 1 - \langle N_{\rm cen}|M_{\rm vir}\rangle\right] , \tag{12}$$

$$\delta N_{\rm sat} = A_{\rm sat} \, \langle N_{\rm sat}|M_{\rm vir}\rangle . \tag{13}$$

It should be noted that, when populating a mock galaxy catalog using HOD or dHOD, the actual number of galaxies in each halo is a random variable: the number of central galaxies follows the Bernoulli distribution, and the number of satellite galaxies follows the Poisson distribution. Since we will be conducting a Fisher analysis, the random fluctuation in realizations can masquerade as a dependence of galaxy number density on (d)HOD parameters, and yield artificially tight constraints. Hence, we need to reduce the random fluctuation in realizations as much as possible so that a small change in one or more (d)HOD parameters results in a small change in the total number of galaxies. We achieve this by assigning to each halo two random variates, $p_{\rm cen}$ and $p_{\rm sat}$, both drawn from the uniform distribution $U(0, 1)$, independently from the (d)HOD parameter values. We then find the number that corresponds to these p-values in the cumulative distribution of a Bernoulli distribution (for $p_{\rm cen}$) or a Poisson distribution (for $p_{\rm sat}$). This minimizes the random fluctuations among realizations that only differ slightly in their corresponding (d)HOD parameters.

In the case of the dHOD, the mean number density of galaxies is strictly independent of the dHOD parameters $A_{\rm cen}$ and $A_{\rm sat}$, and so the problem of preserving the total number density from

one mock realization to another is particularly acute. In the dHOD, changes to $A_{\rm cen}$ or $A_{\rm sat}$ result in changes to the *mean* occupations of individual haloes, but should result in no change to the total number density. If the galaxy occupation for each halo is realized independently, then the total number of galaxies can vary from mock realization to another as $A_{\rm cen}$ and/or $A_{\rm sat}$ are varied. The result of such a variation would be to infer additional constraining power on $A_{\rm cen}$ and $A_{\rm sat}$ where there should be none. To mitigate this possibility, we slightly modified the dHOD implementation in `Halotools`[5] to ensure the total number density of galaxies is preserved among mock catalogs that differ only in their values of $A_{\rm cen}$ and $A_{\rm sat}$. We achieved this by conditioning the dHOD on the total number of galaxies before realizing the occupation of each individual halo. It should be noted that once the total number of central galaxies is fixed, the number of central galaxies in each halo would no longer be strictly a Bernoulli distribution. However, for satellite galaxies, both the total number of galaxies and the number of galaxies in an individual halo follow Poisson distributions.

In this work, we choose the NFW concentration parameter [132] as our secondary property [so $x = c_{\rm NFW}$ in Eqs. (9) and (10)] when studying constraints on the parameters $A_{\rm cen}$ and $A_{\rm sat}$. As has been shown in Refs. [45, 195], concentration only partially accounts for galaxy assembly bias, and other halo properties (e.g., halo age, spin, environment density) may also contribute to assembly bias. Nevertheless, we choose concentration for several physically motivated reasons. First, concentration is known to correlate with assembly history [204][6], and has the advantage that it can be measured in a single snapshot of a simulation. Second, the success of abundance matching suggests that the HODs realized by nature may, indeed, have some dependence upon halo structure [36, 150, 72, 224, 117]. Indeed, Ref. [100] showed that abundance matching in a manner that does not include any concentration dependence is excluded by galaxy clustering. Third, haloes are known to exhibit large concentration-dependent clustering in the mass range of interest to us ($\sim$ a few $\times 10^{12}\ h^{-1}\,{\rm M}_\odot$). Consequently, concentration-dependent clustering is an excellent test case with which to study methods to constrain assembly bias. Fourth, concentration-dependent clustering has already been studied in Ref. [74] for the dHOD and for observational samples by, e.g., Ref. [223], providing a baseline for comparison. Given the above reasons, we believe the concentration parameter is the most reasonable choice for this study, yet we note that our findings

[5]Our implementation is called `PreservingNgalHeavisideAssembias` in `Halotools`.

[6]Though this does not guarantee that concentration and assembly history metrics will lead to similar secondary biases [118].

may not be trivially generalized to the assembly bias induced by other secondary halo properties, as other properties may induce different assembly bias behaviors.

c. Spatial and Velocity Distribution

The detailed predictions of an empirical model depend not only on the model for halo occupation, but also upon the positions and velocities, relative to the host halo, that are assigned to the galaxies. We place the central galaxy at the halo center and the central galaxy inherits the host halo's peculiar velocity. Satellite galaxies are distributed within the virial radius of the host halo according to a spherically symmetric NFW profile characterized by the same concentration as the dark matter distribution. This assumption is supported by various works [193, 106, 192], though other authors find that the distribution of satellite galaxies are described by a concentration different from that of dark matter particles, depending on the satellite population [32, 127, 202, 174, 96]. The radial velocity distribution of satellite galaxies is modeled as a Gaussian distribution with the host halo velocity as the first moment and the solution of the isotropic Jeans equation for an NFW profile [86] as the second moment. We assume velocities to be isotropic, and draw the peculiar velocities in each Cartesian direction independently from this distribution. In practice, the statistics that we examine are quite insensitive to moderate alterations to the treatment of galaxy peculiar velocities (this is by design), though it would be interesting to explore statistics that are sensitive to peculiar velocities as a follow-up study. To examine the effect of alternative velocity models, we have tested the velocity bias model in Ref. [66] with $\alpha_c = 0.3$ and $\alpha_s = 1$, and find that the systematic change in our observables is negligible (within 1.5% in all cases).

3. Observables

In search of effective ways of utilizing existing and future galaxy surveys to constrain the dHOD, we consider a number of observables that are sensitive to halo occupation. In particular, while including the overall galaxy number density of the simulation volume, $n_{\rm gal}$, as a constraining observable in all of our analyses, we examine

I. the projected two-point correlation function, $w_{\rm p}(r_{\rm p})$;

II. the galaxy–galaxy lensing signal, $\Delta\Sigma(r_{\rm p})$;

III. the void probability function VPF(r);

IV. the distribution of counts-in-cylinders, $P(N_{\rm CIC})$;

V. the distribution of counts-in-annuli, $P(N_{\rm CIA})$ (analogous to counts-in-cylinders, but with an excised inner region);

VI. the distribution of the ratio of counts in cylinders of different sizes $P(N_2/N_5)$.

We discuss these observables in more detail in the remainder of this subsection.

We compute all observables numerically, by generating mock galaxy catalogs and subsequently measuring each observable from the mock catalog. This forward-modeling approach enables us to mitigate modeling uncertainty associated with analytic approaches to galaxy clustering and to incorporate possible systematic errors into our calculations. All observables are computed in redshift space, as they would be from observational data, namely, the coordinates of galaxies (x, y, z) are mapped onto $(x, y, z + v_z/aH(a))$. We show examples of the measured values of the observables and their uncertainties from jackknife subsampling (see Section III.B.5) in Fig. III.1 for our fiducial HOD models. Our fiducial models are taken from the fits of Ref. [223], the parameters of which are listed in Table III.1.

Each of the observables is binned in a particular manner. We have selected the binning scheme to ensure that our binning does not significantly degrade the constraining power of any individual observable. We do this by performing a series of analyses in which the bin sizes are reduced in each analysis. We choose bin sizes for each observable such that further refinement of the bins would not yield significant improvement in parameter constraints. We specify the range of the independent variable for each observable observable (for example, in the case of $w_{\rm p}(r_{\rm p})$, we take $0.1 \leq r_{\rm p}/h^{-1}\,{\rm Mpc} \leq 31.6$) and increase the number of bins until parameter constraints saturate. This process has been described in detail in ref. [73]. We find that the constraining power of all observables saturates at fewer than 30 bins, so we take 30 bins for all observables for simplicity. The binning scheme for which our main results are obtained is shown in Table III.2. [7]

[7]Our results are insensitive to the largest length scales included in our analysis because statistics on these scales are measured with relatively low signal-to-noise. We have verified that excluding the few largest bins of $w_{\rm p}(r_{\rm p})$ and $\Delta\Sigma(r_{\rm p})$ from our analyses results in negligible quantitative change in our constraints (typically below 1%, and as large as $\sim 2\%$ in the most extreme cases), and no qualitative change to our conclusions.

Fiducial HOD Parameters							
	$\log M_{\rm min}$	$\sigma_{\log M}$	α	$\log M_0$	$\log M_1$	$A_{\rm cen}$	$A_{\rm sat}$
M_r < -19.0	11.64	0.5119	1.040	10.25	12.80	0	0
M_r < -19.5	11.75	0.4458	1.116	11.29	13.06	0	0
M_r < -20.0	11.97	0.3485	1.144	11.31	13.29	0	0
M_r < -20.5	12.25	0.1854	1.197	11.20	13.59	0	0
M_r < -21.0	12.82	0.5595	1.337	11.96	13.99	0	0

Table III.1: In this table, we list the fiducial HOD parameters adopted for each luminosity threshold of galaxies, taken from the fits of Ref. [223]. Of the 5 standard HOD parameters, $M_{\rm min}$ is the mass at which a halo has a 50% probability of hosting a central galaxy; $\sigma_{\log M}$ determines the rate that $\langle N_{\rm cen}|M_{\rm vir}\rangle$ transitions from zero to unity; M_0 is the truncating mass, below which no satellite galaxies are allowed; M_1 is the halo mass at which the mean satellite number is unity; and α is the index of the satellite occupation power law. Besides the standard HOD parameters, we also allow $A_{\rm cen}$ and $A_{\rm sat}$ to vary, which control the amount of galaxy assembly bias for central and satellite galaxies respectively. In doing this we treat galaxy assembly bias as a deviation from the standard HOD model to be constrained.

Observable Bin Definition						
Bin	$w_{\rm p}(r_{\rm p})$ $r_{\rm p}[h^{-1}\,{\rm Mpc}]$	$\Delta\Sigma(r_{\rm p})$ $r_{\rm p}[h^{-1}\,{\rm Mpc}]$	VPF(r) $r[h^{-1}\,{\rm Mpc}]$	$P(N_{\rm CIC})$ $N_{\rm CIC}$	$P(N_{\rm CIA})$ $N_{\rm CIA}$	$P(N_2/N_5)$ N_2/N_5
1	*	0.11	1.00	{0}	{0}	[0.000,0.033)
2	0.11	0.13	1.08	{1}	{1}	[0.033,0.067)
3	0.14	0.16	1.17	{2}	{2}	[0.067,0.100)
4	0.17	0.20	1.27	{3}	{3}	[0.100,0.133)
5	0.20	0.24	1.37	{4}	{4}	[0.133,0.167)
6	0.25	0.29	1.49	{5}	{5}	[0.167,0.200)
7	0.30	0.35	1.61	{6}	{6}	[0.200,0.233)
8	0.37	0.43	1.74	{7}	{7}	[0.233,0.267)
9	0.45	0.52	1.89	{8}	{8}	[0.267,0.300)
10	0.54	0.62	2.04	{9}	{9}	[0.300,0.333)
11	0.66	0.76	2.21	[10,12)	[10,12)	[0.333,0.367)
12	0.81	0.92	2.40	[12,13)	[12,14)	[0.367,0.400)
13	0.99	1.11	2.59	[13,15)	[14,16)	[0.400,0.433)
14	1.20	1.35	2.81	[15,18)	[16,19)	[0.433,0.467)
15	1.47	1.63	3.04	[18,20)	[19,22)	[0.467,0.500)
16	1.79	1.98	3.29	[20,24)	[22,26)	[0.500,0.533)
17	2.18	2.39	3.56	[24,27)	[26,30)	[0.533,0.567)
18	2.66	2.90	3.86	[27,31)	[30,35)	[0.567,0.600)
19	3.24	3.51	4.18	[31,36)	[35,41)	[0.600,0.633)
20	3.95	4.26	4.52	[36,42)	[41,48)	[0.633,0.667)
21	4.82	5.16	4.89	[42,48)	[48,57)	[0.667,0.700)
22	5.88	6.25	5.30	[48,55)	[57,66)	[0.700,0.733)
23	7.17	7.57	5.74	[55,64)	[66,78)	[0.733,0.767)
24	8.75	9.17	6.21	[64,74)	[78,91)	[0.767,0.800)
25	10.6	11.11	6.72	[74,85)	[91,106)	[0.800,0.833)
26	13.0	13.46	7.28	[85,98)	[106,125)	[0.833,0.867)
27	15.8	16.30	7.89	[98,113)	[125,146)	[0.867,0.900)
28	19.3	19.75	8.53	[113,130)	[146,171)	[0.900,0.933)
29	23.6	23.93	9.24	[130,150)	[171,200)	[0.933,0.967)
30	28.7	28.99	10.00	[150, + ∞)	[200, + ∞)	[0.967,1.000]

Table III.2: Definition of bins for each observable are listed in this table, each measured in 30 bins.

Table III.2: *(cont.)*We show the values of bin centers for the bins in which $w_{\rm p}(r_{\rm p})$ and $\Delta\Sigma(r_{\rm p})$ are measured, the set of radii of spheres used for evaluating VPF(r), the intervals defining each bin in the histograms of counts-in-cylinders and annuli, and the ratio N_2/N_5. We use the same number of bins for each observable, such that our comparison of constraining power is not sensitive to bin number. * Note that the number density $n_{\rm gal}$ is listed as the first bin of $w_{\rm p}(r_{\rm p})$ in this table, but it is included in the analysis for all possible combinations of observables.

a. Projected Two-Point Correlation Function

The projected two-point correlation function, $w_{\rm p}(r_{\rm p})$, is a canonical observable that has been considered in numerous previous analyses to inform halo occupation [188, 231, 221]. It is defined by

$$w_{\rm p}(r_{\rm p}) = 2\int_0^{\pi_{\rm max}} d\pi\, \xi(r_{\rm p},\pi) \tag{14}$$

where $\xi(r_{\rm p},\pi)$ is the excess probability of finding galaxy pairs with projected and line-of-sight separations $r_{\rm p}$ and π, respectively. We estimate $w_{\rm p}(r_{\rm p})$ from our mock catalogs by counting galaxy pairs that have a projected separation in a bin of $r_{\rm p}$ within a perpendicular distance of $\pi_{\rm max}$ in redshift space. We choose $\pi_{\rm max} = 60h^{-1}$ Mpc, as is done by [221], according to whom this integration limit is large enough to include most correlated pairs and minimize the impact of the details of peculiar velocity models, yet sufficiently small to give a stable result by suppressing noise from very distant, uncorrelated pairs. We compute $w_{\rm p}(r_{\rm p})$ in 29 logarithmically spaced radial bins from $r_{\rm p} = 0.1\,h^{-1}$ Mpc to $r_{\rm p} = 31.6\,h^{-1}$ Mpc.

The projected two-point clustering of our fiducial models are shown for two luminosity threshold samples in the upper, left panel of Fig. III.1. The figure exhibits several well-known characteristics of galaxy clustering. First, brighter galaxies cluster more strongly. Second, the galaxy two-point correlation function can roughly be described as a power law, $w_{\rm p} = (r_{\rm p}/r_0)^{\alpha}$, with index $\alpha \approx -0.8$. Third, in more detail, the correlation function exhibits a small deviation from a power law near $r_{\rm p} \sim 2\,h^{-1}$ Mpc which is due to the transition from galaxy pairs that reside in distinct haloes (the "two-halo" term) on large scales ($r_{\rm p} \gtrsim 2h^{-1}$ Mpc) and pairs of galaxies that reside in a common halo (the "one-halo term") on scales $r_{\rm p} \lesssim 1h^{-1}$ Mpc.

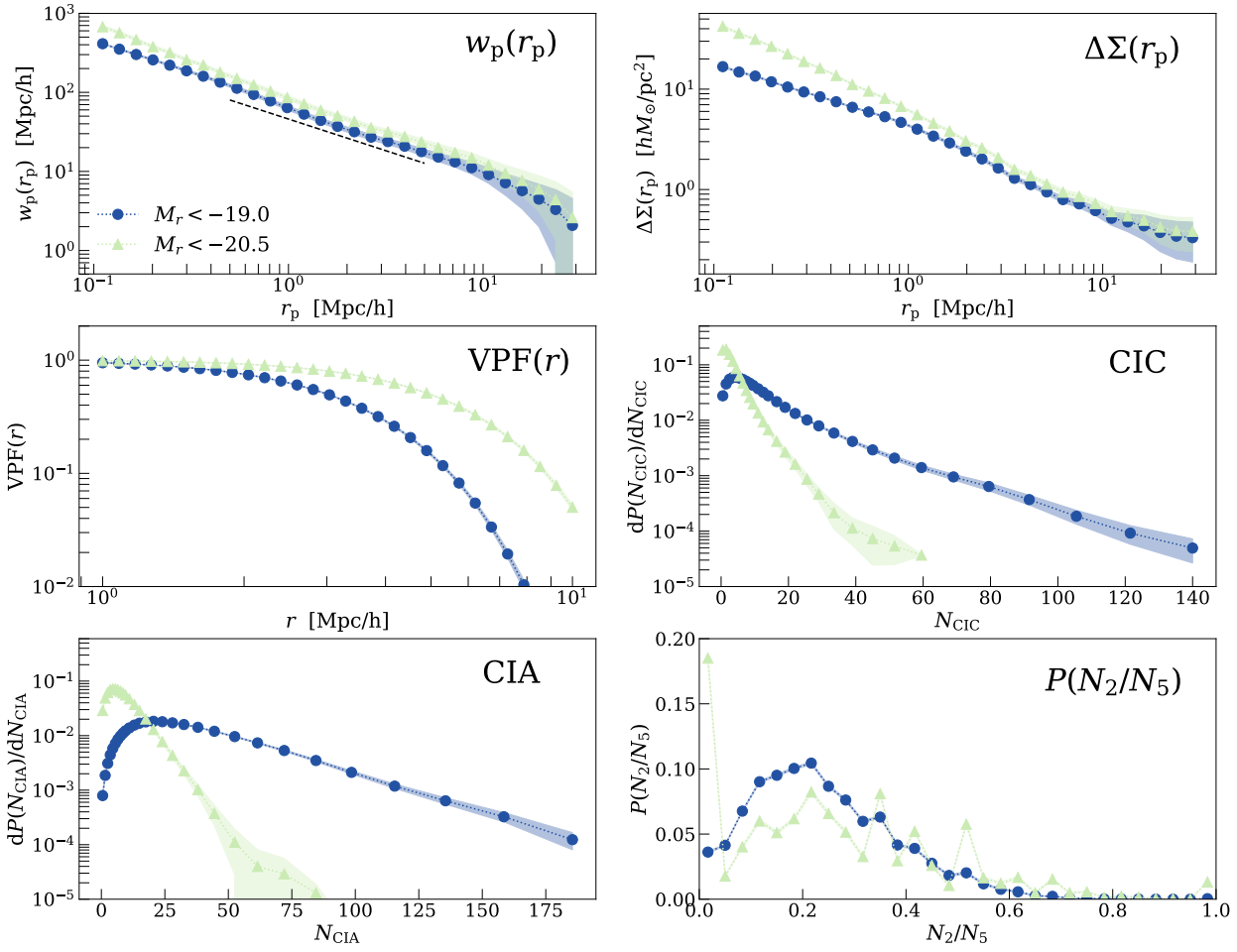


Figure III.1: Examples of the observables that we consider to constrain assembly bias. In each panel, we show examples of the observables. We show examples for our fiducial HOD parameters corresponding to two luminosity thresholds, as illustrated by the legend in the upper left panel. The observable values are shown as connected data points, while the colored bands show the uncertainty from jackknife subsampling, which we describe in detail in Section III.B.5.

Figure III.1: *(cont.)* The upper left panel shows the projected correlation function, $w_{\rm p}(r_{\rm p})$. For illustration, this panel also contains a dashed line illustrating the slope of a power law with $w_{\rm p}(r_{\rm p}) \propto r_{\rm p}^{-0.8}$ for comparison. The upper right panel shows the excess surface density about galaxies, $\Delta\Sigma(r_{\rm p})$, in the samples. The left panel in the middle row depicts the void probability function, VPF(r). The right panel in the middle row depicts the distribution of counts-in-cylinders (CIC). Notice that the lower luminosity sample has a much more significant tail to high companion counts than the higher luminosity sample. Similarly, the left panel of the bottom row depicts counts-in-annuli (CIA). Finally, the right panel of the bottom row shows the probability distribution of the ratio of cylinder counts on distinct scales, $P(N_2/N_5)$. Each panel is labeled by the observable shown.

b. Galaxy–Galaxy Weak Lensing

In addition to the projected two-point clustering, galaxy–galaxy weak lensing is another observable statistic that has been used by many previous authors to constrain halo occupation from observational data [115, 29, 194]. The canonical observable, $\Delta\Sigma(r_{\rm p})$, is the excess surface density of mass around galaxies projected along the line-of-sight, and averaged over all potential lens galaxies in the sample,

$$\Delta\Sigma(r_{\rm p}) = \bar{\Sigma}(< r_{\rm p}) - \Sigma(r_{\rm p}) \tag{15}$$

where $\Sigma(r_{\rm p})$ is the projected surface density evaluated at position $r_{\rm p}$ relative to the center of the lens galaxy, and $\bar{\Sigma}(< r_{\rm p})$ is the mean projected, two-dimensional, surface mass density within a projected distance or $r_{\rm p}$ from the lens galaxy. We compute $\Delta\Sigma(r_{\rm p})$ in 30 logarithmically-spaced radial bins from $r_{\rm p} = 0.1\,h^{-1}$ Mpc to $r_{\rm p} = 31.6\,h^{-1}$ Mpc. The simulations that we use are gravity-only N-body simulations, so our estimates of $\Delta\Sigma(r_{\rm p})$ include neither baryonic mass nor any influences of baryons on the dark matter distribution [157].

The galaxy–galaxy lensing signal in our fiducial catalogs is depicted in the upper, right panel of Fig. III.1. It is evident from this panel that galaxies in the higher-luminosity samples are more strongly correlated with mass, indicating the well-known fact that more luminous galaxies tend to reside in more massive dark matter haloes. The feature due to the transition between the one-halo

and two-halo terms in the galaxy matter correlation function is evident near $r_{\rm p} \sim 2\,h^{-1}$ Mpc as well.

c. Void Probability Function

We examine a number of options for observables in addition to the canonical $w_{\rm p}(r_{\rm p})$ and $\Delta\Sigma(r_{\rm p})$, among them the VPF. The VPF has been examined in previous studies on assembly bias with mixed conclusions [179, 224] and, in principle, depends upon all of the n-point functions [139]. To estimate VPF(r), we randomly place spheres of radius r throughout our simulation volume and enumerate the probability of the spheres containing zero galaxies (and thus being classified as voids) as

$$\mathrm{VPF}(r) = \frac{N_{\rm void}(r)}{N_{\rm sphere}(r)}, \tag{16}$$

where $N_{\rm sphere}$ is the total number of spheres that we use for the estimate ($N_{\rm sphere} = 10^5$ in this work) and $N_{\rm void}$ is the number of spheres that are found to enclose zero galaxies. We compute VPF(r) at 30 logarithmically-spaced radii from $r = 1\,h^{-1}$ Mpc to $r = 10\,h^{-1}$ Mpc. We remind the reader that these calculations are performed in redshift space, by mapping the coordinates of galaxies according to their line-of-sight velocities.

The VPF(r) of our fiducial models are depicted in the left, middle panel of Fig. III.1. As with $w_{\rm p}(r_{\rm p})$ and $\Delta\Sigma(r_{\rm p})$, several expected features of the VPF(r) are evident. The VPF drops from nearly unity on small scales to well below unity beyond a scale of $r \sim 10\,h^{-1}$ Mpc and voids are more likely for higher luminosity galaxy samples, due largely to their overall lower number density.

d. Counts-in-cylinders (CIC) Statistic

Galaxy counts, particularly counts of galaxies within cylindrical volumes in redshift space, have been studied for decades [58, 7, 4, 10, 34, 173, 85, 77, 84, 22, 9, 152, 17, 137, 63]. The average number of companions that a galaxy will have within a particular cylinder can be computed from the two-point correlation function; however, the distribution of counts-in-cylinders depends, at least in principle, upon all of the higher n-point functions [139] and can complement the two-point function as a study of the galaxy halo relationship.

We compute counts in cylinders (CIC) from our galaxy catalogs as follows. We center a cylinder of transverse radius $r_{\rm CIC}$ and depth $\pm\Delta v$ (in redshift space) on each galaxy in the sample

and count the number of companion galaxies that fall within the cylinder. This procedure enables us to estimate a probability distribution of companion number, $P(N_{\rm CIC})$, which is the probability that any galaxy has $N_{\rm CIC}$ companions within the cylinder. For the primary results that we present in this paper, we use $r_{\rm CIC} = 2h^{-1}\,{\rm Mpc}$ and a maximum relative velocity of $\Delta v = 1000\,{\rm km\,s^{-1}}$, corresponding to a half-length of $L = 10h^{-1}\,{\rm Mpc}$, assuming velocities are only due to the Hubble flow. We choose cylinders of a transverse radius $r_{\rm CIC}$ on the order of a few $h^{-1}\,{\rm Mpc}$ in order to include galaxy companions separated by a scale on which assembly bias is known to introduce a distinct feature in halo clustering [74, 223, 172]. We have experimented with a variety of alternative cylinder radii and depths, finding that our results remain qualitatively similar. When characterizing the PDF $P(N_{\rm CIC})$, we count the first few values of $N_{\rm CIC}$ individually and group larger $N_{\rm CIC}$ values into logarithmically spaced bins, as indicated in Table III.2.

Examples of our counts-in-cylinders distributions, $P(N_{\rm CIC})$, for two of our luminosity threshold samples, can be seen in the middle, right-hand panel of Fig. III.1. As is expected, the probability of having a large number of companions in a cylindrical cell increases dramatically with decreasing galaxy luminosity due to the higher number density of galaxies with lower luminosities.

e. Counts-in-annuli (CIA) Statistic

To complement counts-in-cylinders, we also examine counts of neighbor galaxies in annuli. The counts-in-annuli (CIA) enable one to get a sense of clustering as a function of scale and to compare smaller-scale, intra-halo clustering to larger-scale clustering. The statistic $P(N_{\rm CIA})$ is the probability that the number of companions within the annulus is equal to $N_{\rm CIA}$, analogous to $P(N_{\rm CIC})$. Unlike the counts-in-cylinders statistic, which roughly probes the halo-occupation statistics on the "one-halo" scale, $P(N_{\rm CIA})$ is a novel statistic, introduced here, and specifically designed to probe the immediate, supra-halo environments of galaxies.

As with CIC, we choose fixed dimensions for the annuli that we use and explore the constraining power of the distribution of counts around galaxies in our catalogs. Our annuli have inner radii of $r_{\rm inner} = 2\,h^{-1}\,{\rm Mpc}$ and outer radii of $r_{\rm outer} = 5\,h^{-1}\,{\rm Mpc}$. As with CIC, the annuli have a depth in the redshift dimension of $10\,h^{-1}\,{\rm Mpc}$, corresponding to a velocity difference of $\Delta v = 1000\,{\rm km\,s^{-1}}$. As with CIC, this geometry is chosen in order to probe the immediate environments of haloes,

particularly on scales where assembly bias has already been shown to induce a feature in galaxy clustering [74, 172]. We have experimented with moderately different annular dimensions and obtained qualitatively similar results in all cases. We group values of $N_{\rm CIA}$ in a similar way to $N_{\rm CIC}$, as detailed in Table III.2.

Examples of $P(N_{\rm CIA})$ for our fiducial catalogs are given in the lower, left-hand panel of Fig. III.1. The CIA distribution shares most of the qualitative features of the CIC distribution, though the counts are generally higher because the volumes of our annuli exceed the volumes of our cylinders by a factor of ~ 5.

f. Distribution of Cylinder Count Ratios

As a distinct way of characterising the clustering environments of galaxies, we also consider the distribution of the ratio of two cylinder counts. The first count is within a cylinder with a radius of $r_{\rm CIC} = 2\,h^{-1}$ Mpc and the second, larger cylinder has a radius $r_{\rm CIC} = 5\,h^{-1}$ Mpc for each galaxy. For both cylinders we adopt the same depth, $\Delta v = 1000\,{\rm km\,s^{-1}}$ as for the CIC and CIA statistics discussed above. For each galaxy in our catalogs, we compute the companion counts N_2 and N_5 within each of these cylinders, and take the probability distribution of the ratio of these two numbers as the statistic of interest (notice that the inner cylinder is the same cylinder used in our CIC calculations, so that $N_5 = N_2 + N_{\rm CIA}$. Similar to $N_{\rm CIA}$, this is a novel statistic to probe the large scale distribution of galaxies that, to the best of our knowledge, has not been utilized before. The intention of this statistic is to probe the relative clustering within a halo (the "one-halo term") to that in its immediate vicinity. We measure the probability distribution of this ratio, $P(N_2/N_5)$ in 30 linearly-spaced bins from 0 to 1.

Probability distributions of the cylinder count ratio are shown in the lower, right-hand panel of Fig. III.1. It is evident that the higher luminosity sample has a distribution with more discreteness noise, as a result of the low number density. The more luminous sample has a significantly higher probability of having low values of N_2/N_5 than the lower luminosity sample. This is due to the fact that satellite galaxies are increasingly rare in the higher-luminosity samples, so that N_2 is increasingly likely to be either small or zero in such samples compared to lower-luminosity samples (see the right, middle panel of Fig. III.1).

4. Fisher Analysis

We use a Fisher matrix analysis [25, 52] to forecast the constraining power of each of the observables described above and combinations thereof. Despite its approximate nature, we elect to use a Fisher matrix due to the computational expense of utilizing other techniques. Employing a technique that directly quantifies uncertainty by averaging the posteriors found from each mock catalog, either via Markov Chain Monte Carlo [62, 56] or Approximate Bayesian Computation [207], to perform the very large number of analyses that we undertake using mock catalogs is substantially more computationally intensive.

The Fisher matrix is a measure of the ideal amount of information that can be obtained from an experiment. The Fisher matrix is defined as

$$\mathbf{I} = \frac{\partial \mathbf{f}^{\mathrm{T}}}{\partial \mathbf{p}} \mathbf{C}^{-1} \frac{\partial \mathbf{f}}{\partial \mathbf{p}}, \tag{17}$$

where $\mathbf{f}$ is the set of observables, $\mathbf{C}$ is the covariance of the observables, and $\mathbf{p}$ is the parameter set to be constrained. The set of observables $\mathbf{f}$ includes all bins of each observable quantity (e.g., 29 bins of $w_{\mathrm{p}}(r_{\mathrm{p}})$, 30 bins of $\Delta\Sigma(r_{\mathrm{p}})$, etc.), while $\mathbf{p}$ represents the set of all model parameters. The notation $\partial\mathbf{f}/\partial\mathbf{p}$ represents the matrix of values constructed by differentiating each observable with respect to each of the individual parameters, so that the matrix element $\partial f_{\mathrm{i}}/\partial p_{\mathrm{j}}$ represents the derivative of the i^{th} observable with respect to the j^{th} model parameter. Both the derivatives and the covariance are evaluated at a single, *fiducial* point in the parameter space, which is assumed to be the true underlying model.

The expected 1σ error on any inferred parameter, marginalized over all other parameters, can be obtained by taking the square root of the corresponding diagonal term of the posterior covariance matrix Σ, which is the inverse of the Fisher matrix. Hence, the forecasted marginalized uncertainties in the parameters are

$$\sigma = \sqrt{\mathrm{Diag}(\Sigma)} = \sqrt{\mathrm{Diag}(\mathbf{I}^{-1})}. \tag{18}$$

In our study of the constraining power of various observables, we explore the 6 dimensional parameter space, spanned by α, $\log M_1$, $\sigma_{\log M}$, $\log M_{\mathrm{min}}$, A_{cen}, and A_{sat}. The parameter $\log M_0$ is part of both the standard HOD and the dHOD models. However, we set $\log M_0$ to its fiducial value

in our analyses and do not allow it to vary. We do this because $\log M_0$ is poorly constrained by these data [223].

The fiducial points about which we evaluate our Fisher matrices are given by the parameters listed in Table III.1, and differ for each luminosity threshold sample. In the next two subsections, we discuss the computation of the covariance matrix, $\mathbf{C}$, and the derivatives of the observables, $\partial\mathbf{f}/\partial\mathbf{p}$, respectively.

5. Covariance

In order to implement the Fisher approximation for the marginalized constraints on model parameters, we must compute a covariance matrix about the fiducial point in the parameter space. The covariance matrix that we calculate has three contributions. The first, and dominant, contribution is from sample variance (sometimes called "cosmic" variance in this context). We estimate the sample variance contribution using jackknife resampling of the simulation volume, while recognizing the caveat that jackknife resampling is known to underestimate covariances. We will refer to this component of the covariance as $\mathbf{C}_{\rm jackknife}$. The second contribution to the covariance matrix is due to the stochasticity of populating a simulation with galaxies drawn from the probability distribution functions of the (d)HOD. Multiple realizations of the same underlying model in identical, finite volumes will lead to mildly different predictions due to this stochasticity. We refer to this contribution to the covariance as $\mathbf{C}_{\rm realization}$. Third, we use a fixed set of randomly distributed centers of spheres in the calculation of VPF(r) as well as a fixed subsample of dark matter particles in the calculation of $\Delta\Sigma(r_{\rm p})$. These choices introduce a small contribution to the covariance that we denote $\mathbf{C}_{\rm random}$. The total covariance matrix that we use is the sum of each of these contributions

$$\mathbf{C}_{\rm total} = \mathbf{C}_{\rm jackknife} + \mathbf{C}_{\rm realization} + \mathbf{C}_{\rm random}. \tag{19}$$

As an example, Fig. III.2 depicts the covariance matrices $\mathbf{C}_{\rm jackknife}$, $\mathbf{C}_{\rm random}$, $\mathbf{C}_{\rm realization}$ and their linear combination $\mathbf{C}_{\rm total}$, for the $M_r < -19.0$ threshold sample, as normalized correlation matrices.

The contributions from $\mathbf{C}_{\rm realization}$ and $\mathbf{C}_{\rm random}$ are straightforward to compute. To estimate $\mathbf{C}_{\rm realization}$, we populate the halo catalog with the fiducial HOD multiple times, each time using a new random seed, and compute the covariance across the measurements from the resultant mock

galaxy catalogs. $\mathbf{C}_{\rm realization}$ is displayed in the upper right panel of Fig. III.2. For $\mathbf{C}_{\rm random}$, we repeat measurements of $\Delta\Sigma(r_{\rm p})$ and VPF(r) on the same mock galaxy catalog, but with different sets of particle subsamples and sphere centers, respectively, and calculate the covariance. Note that this matrix only consists of the two corresponding blocks, as shown in the lower left panel of Fig. III.2.

Finally, to estimate $\mathbf{C}_{\rm jackknife}$, which is shown in the upper left panel of Fig. III.2, we divide the simulation box into 10×10 cuboid cells, each of size $25\,h^{-1}$ Mpc $\times$ $25\,h^{-1}$ Mpc $\times$ $250\,h^{-1}$ Mpc. The long axes of each cuboid are the same as the length of the simulation volume and are assumed to lie along the line of sight. For each mock catalog, we construct three such sets of jackknife samples by choosing, in turn, the x, y, and z dimensions of the simulation cube as the line-of-sight direction. Our final covariances are the averages of the three covariances computed for each of the three projections. We construct this average to minimize the contributions from any significant variations that may, by chance, fall along any individual projection. For each set of jackknife samples coming from each of the three projections of the mock catalogs, we exclude individual jackknife cells in turn, and compute the jackknife contribution in the usual manner [82, 148, 182, 198].

For the purposes of computing jackknife covariances only, the mock catalogs that we use are *not* based on our fiducial HODs. Our jackknife covariance mock catalogs are based upon abundance matching with zero scatter [92]. We construct these catalogs by populating haloes that have the highest values of $V_{\rm peak}$, with galaxy number densities consistent with HOD realizations. This modification is necessary for the following reason. Each mock HOD-based catalog is a realization of the underlying HODs. Therefore, there is inherent stochasticity in the covariance matrix estimates. Moreover, in the HOD formalism, each luminosity threshold must be treated independently, which, in turn, means that the covariances in different threshold samples can fluctuate independently. This makes comparing covariances across luminosity thresholds challenging because to do this using the HOD approach requires marginalizing the stochasticity over a very large number of mock catalogs. The abundance matching approach that we have adopted allows us to circumvent this difficulty because there is no stochasticity in the mock catalogs. Therefore, the stochasticity associated with building mock catalogs does not contribute to our $\mathbf{C}_{\rm jackknife}$ estimates. This ensures that our $\mathbf{C}_{\rm jackknife}$ estimates vary smoothly with the luminosity threshold of the sample. We have found that this procedure reduces the noisiness of our forecasts, yet does not alter our qualitative results.

All six of our candidate observables are based on pair or neighbor counting, which reduces the choice of algorithm to determining which counts to exclude for each jackknife subsample. For $w_{\rm p}(r_{\rm p})$, we discard a pair if either or both of the galaxies reside in the excluded cell. For $\Delta\Sigma(r_{\rm p})$, we only calculate the dark matter density profile around galaxies that live outside the excluded cell. Note, though, that in doing so we include dark matter particles that lie in that cell. Excluding such particles would lead to anomalous density profiles that are not easily corrected because the subsampling procedure violates the periodicity of the simulation volume. Similarly, for VPF(r), we place random spheres about points outside of the excluded cell; however, for the purposes of determining whether or not a particular sphere is a void region, galaxies within the excluded cell are taken into account. And for the count statistics, $P(N_{\rm CIC})$, $P(N_{\rm CIA})$, and $P(N_2/N_5)$, we only center cylinders on galaxies outside the excluded cell, but include companion galaxies within the excluded cell in our counts.

Of the three contributions, $\mathbf{C}_{\rm jackknife}$ is the dominant component, and $\mathbf{C}_{\rm random}$ is negligibly small, suggesting that we have used sufficiently large samples of VPF centers and dark matter particle positions to render the noisiness induced by finite sampling of these distributions negligibly small.

Direct inversion of the covariance matrix $\mathbf{C}$ is problematic numerically. Briefly stated, the uncertainty in the covariance will lead to the smallest eigenvalues of $\mathbf{C}$ being dominated by noise. When a matrix is inverted, its eigenvalues are inverted, which implies that the small, noisy eigenvalues of $\mathbf{C}$ become the large and noisy eigenvalues of its inverse. The inverse hence becomes dominated by this noise. This problem is further compounded by the large differences among the matrix elements inherited from the differences between the natural scales of the different observables, leading to extremely large differences in the sizes of the matrix eigenvalues. For example, the natural scale of the two-point function, $w_{\rm p}(r_{\rm p})$, is $\sim 10^2$, whereas the natural scale of $P(N_{\rm CIC})$ is on the order of $\sim 10^{-2}$ (see Fig. III.1) and this difference leads to very different covariance matrix elements.

A common approach when faced with this problem is to truncate the smallest eigenvalues of $\mathbf{C}$, and calculate the Moore-Penrose pseudo-inversion [140]. We therefore normalize $\mathbf{C}$, and perform pseudo-inversion, excluding the eigenvalues of the correlation matrix smaller than 10^{-5} times the largest eigenvalue. Choosing other reasonable values of this cutoff does not substantially impact our results. In the future, when larger data sets are available, and the noise level of the covariance

is sufficiently low, pseudo-inversion may no longer be necessary.

Our covariance matrices have several noteworthy features. Firstly, from an inspection of Fig. III.2, we see strong self correlation between the bins of $w_{\rm p}(r_{\rm p})$, $\Delta\Sigma(r_{\rm p})$, and VPF(r) over a wide range of scales. On the contrary, $P(N_2/N_5)$ shows weak correlation among its bins and with other observables, as it measures the distribution of the dimensionless ratio N_2/N_5, and is insensitive to the cosmic variance of galaxy number density. Secondly, for $P(N_{\rm CIC})$ and $P(N_{\rm CIA})$, the probability of smaller counts and larger counts sum up to unity, and are anti-correlated by construction, producing the sign reversal in the corresponding matrix blocks. Additionally, when comparing $\mathbf{C}_{\rm jackknife}$ and $\mathbf{C}_{\rm realization}$, it is obvious that the observable values are more correlated among jackknife subsamples than stochastic realizations. The sign of correlation coefficients approximately coincide between the two contributions, with the exception of blocks involving VPF(r). In $\mathbf{C}_{\rm realization}$, VPF($r$) has a weak positive correlation with $w_{\rm p}(r_{\rm p})$ and $\Delta\Sigma(r_{\rm p})$ in most of the bins, while in $\mathbf{C}_{\rm jackknife}$, VPF($r$) is anti-correlated with $w_{\rm p}(r_{\rm p})$ at larger $r_{\rm p}$ and $\Delta\Sigma(r_{\rm p})$. These are non-trivial effects, as VPF(r) is dependent on multiple moments of the galaxy number density field. The jackknife subsamples probe different regions of the box, with denser regions corresponding to stronger galaxy–galaxy and galaxy–matter correlation as well as fewer voids, leading to the anti-correlation in $\mathbf{C}_{\rm jackknife}$. On the other hand, among different realizations, higher values of $w_{\rm p}(r_{\rm p})$ and $\Delta\Sigma(r_{\rm p})$ result not from higher galaxy number densities but when more galaxies are concentrated in clusters, allowing more voids to exist in the rest of the space, giving rise to a positive correlation with VPF(r). Covariances for other luminosity samples have qualitatively similar features.

When applying the same analyses to galaxy survey data, observational uncertainties need to be taken into account, here we discuss how our covariances compare to observational covariances from SDSS data. For statistics that depend only on galaxy distribution, e.g., all of our observables except $\Delta\Sigma(r_{\rm p})$, the observational uncertainties depend on the survey volume and target number density to first order, both of which we mimic in our analyses. In comparison with the covariances from SDSS, which has a similar volume for the $M_r < -20.0$ sample to the Bolshoi Planck simulation, indeed our jackknife covariance matrix for $w_{\rm p}(r_{\rm p})$ is comparable to the measurement in Ref. [221]. For the fainter samples, SDSS has smaller volumes and hence larger covariances than ours, and vice versa for the brighter samples. Similarly, we expect the other observables to have covariances comparable

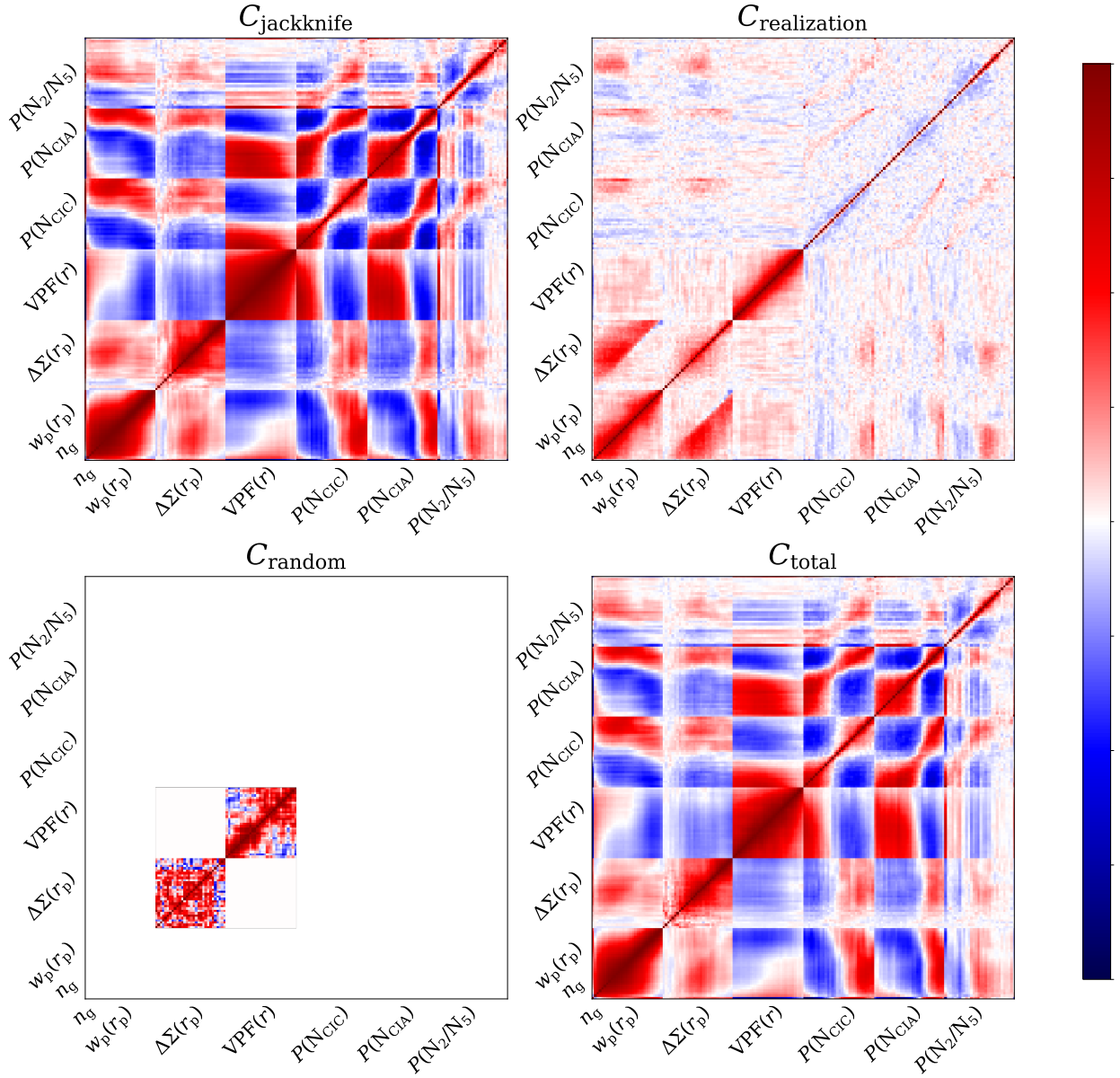


Figure III.2: In this figure, we show the total covariance matrix along with the three matrices that we sum in order to compute the total covariance matrix, each normalized to correlation matrices, for the $M_r < -19.0$ threshold sample.

Figure III.2: *(cont.)* $\mathbf{C}_{\rm realization}$ and $\mathbf{C}_{\rm random}$ are computed using our fiducial HOD parameters. The sample variance contribution, $\mathbf{C}_{\rm jackknife}$, is computed using a mock catalog based on abundance matching in order to eliminate the stochasticity associated with any individual mock HOD catalog. $\mathbf{C}_{\rm jackknife}$ is the major contribution, while $\mathbf{C}_{\rm realization}$ and $\mathbf{C}_{\rm random}$ are subdominant. Covariances for other thresholds are qualitatively similar. Blocks of these matrices corresponding to the bins of a specific type of observable (e.g., $w_{\rm p}(r_{\rm p})$ or $P(N_{\rm CIC})$) are labeled as such. Each such block contains 30 rows and columns corresponding to the 30 bins used for each observable.

to SDSS observation, with the exception of $\Delta\Sigma(r_{\rm p})$. Aside from the first-order comparability, there are other factors that need to be accounted for in observation. Fiber collision [138] affects pairs with small separations, i.e., the smallest bins of $w_{\rm p}(r_{\rm p})$ and $\Delta\Sigma(r_{\rm p})$, as well as $P(N_{\rm CIC})$ and therefore $P(N_2/N_5)$. This effect needs to be forward modeled in survey data analyses. However, $P(N_{\rm CIA})$ does not depend on the closest galaxy companions, and is robust to fiber collision. As we will show in Section III.C, $P(N_{\rm CIA})$ typically provides tight constraints comparable to $P(N_{\rm CIC})$. Some of the other second-order observational uncertainties are blending and saturation in target selection, redshift measurement errors, and geometric features, all of which we expect to be subdominant in this context.

On the other hand, the covariances for $\Delta\Sigma(r_{\rm p})$ measured for SDSS data [114] are significantly larger than our covariances. This is expected because the survey data is dominated by shape noise, while we neglect shape noise in our study, assuming infinite source densities. Taking non-zero shape noise into account will result in weaker constraints from lensing than those that we find here, and our forecasts therefore must be regarded as the upper limit of constraining power that can be achieved with $\Delta\Sigma(r_{\rm p})$.

6. Derivative Fitting

In order to compute the elements of the Fisher matrix, it is necessary to estimate the partial derivatives of the observables with respect to the parameters in the neighborhood of the fiducial point in the parameter space. We designated these derivatives as $\partial\mathbf{f}/\partial\mathbf{p}$ in Eq. (17) above. Assessing these

derivatives from realizations of the perturbed models is non-trivial due to the inherent stochasticity of using mock catalogs based upon the direct population of N-body simulations. Therefore, we give a detailed description of our approach to estimating derivatives in Appendix B.

C. Results

We perform Fisher matrix analyses in order to forecast the constraints on the dHOD model that can be extracted from combinations of the galaxy observables described in Section III.B.3. In Table III.1, we list the fiducial HOD parameters corresponding to 5 galaxy samples selected by luminosity. In this section, we present our primary results in terms of estimated posterior 1σ constraints for the four lower luminosity samples that we have studied. We exclude the brightest, $M_r < -21.0$ sample from our primary results because the results from this sample are subject to excessive statistical fluctuations due to the relatively small number of galaxies above this luminosity threshold within the volume of the Bolshoi Planck simulation. For completeness, the results from the $M_r < -21.0$ sample are included in our comprehensive list of results in Appendix C.

Before proceeding to our results, we note that the dominant contribution to our errors are from sample variance due to the finite volume of the Bolshoi Planck simulation (see Section III.B.5 above). The absolute constraints on parameters will decrease with increasing volume. As such, we focus on *relative* constraints on parameters from different combinations of observables, rather than on the absolute values. Our study requires that haloes be resolved with a very large number of particles which, in turn, stipulates the use of a high-resolution, relatively smaller volume simulation such as Bolshoi Planck.

1. Assessing the Complementarity of Observables

We begin with a discussion of our forecast constraints from *individual* observables and combinations of any *two* observables. We use the $M_r < -20.0$ sample as an example in Fig. III.3, to compare the constraints on $A_{\rm cen}$ and $A_{\rm sat}$ from the individual observables and all the possible combinations of two observables. We caution that each of the four panels has a different y-axis

range, which the reader must account for when comparing results among different panels.

We display constraints from the individual observables on $A_{\rm cen}$ in the upper left panel of Fig. III.3 and $A_{\rm sat}$ in the lower left panel of Fig. III.3. The columns are ordered according to increasing constraining power, and the filled circles indicate the observables from which the constraints were derived. For example, the leftmost column in the upper left panel of Fig. III.3 shows the constraint on $A_{\rm cen}$ derived from $P(N_{\rm CIC})$.

The constraints from individual observables, displayed in the left panels of Fig. III.3 convey several points. The observable $P(N_{\rm CIC})$ more strongly constrains the assembly bias of satellite galaxies than that of central galaxies. This is expected because $P(N_{\rm CIC})$ primarily probes the "one-halo term", and is sensitive to the satellite population that accompany centrals. For the brighter samples with higher satellite fractions, $P(N_{\rm CIA})$ is more dependent on the satellite population, and constrains $A_{\rm sat}$ more strongly, while its constraining power decreases for $A_{\rm cen}$. The constraints from $w_{\rm p}(r_{\rm p})$ are dominated by the smaller radial bins, which have higher signal to noise ratios than the measurements at larger scales. This causes the absolute constraints on $A_{\rm cen}$ to be weaker than those on $A_{\rm sat}$. However, with larger volumes, e.g., DESI [101], large-scale clustering will be measured with higher precision, enabling better constraints on $A_{\rm cen}$. We also find that VPF(r) gives strong constraints on $A_{\rm cen}$, but is extremely inefficient in constraining $A_{\rm sat}$. This can be explained by the fact that a single galaxy suffices to eliminate the possibility that a region could be a void. The vast majority of satellite galaxies reside in haloes where there are central galaxies, which already eliminate the void, therefore the void probability function is largely insensitive to the abundances of satellite galaxies. For this reason, VPF(r) is a poor probe of not only $A_{\rm sat}$, but all of the parameters that determine satellite populations (see Tables C- C in Appendix C).

The constraints from the combinations of two observables are shown in the right-hand panels of Fig. III.3. The columns are again ordered from least constraining to most constraining, and the observables used in each analysis are marked by filled circles. For example, the leftmost column in the upper right panel of Fig. III.3 shows constraints derived from the combination of $w_{\rm p}(r_{\rm p})$ and $\Delta\Sigma(r_{\rm p})$. The combination of $w_{\rm p}(r_{\rm p})$ and $\Delta\Sigma(r_{\rm p})$ has been used in a number of previous studies; however, we find that this is one of the least constraining of the combinations that we have considered for both $A_{\rm cen}$ and $A_{\rm sat}$. As we discuss further below, this is chiefly because $w_{\rm p}(r_{\rm p})$ and $\Delta\Sigma(r_{\rm p})$ share largely common degeneracies among the (d)HOD parameters, so that combining the

two does not result in significant decreases in marginalized constraints that can be expected when combining highly complementary data.

The primary result to be gleaned from Fig. III.3 is the overall efficacy of the count-based observables, i.e., $P(N_{\rm CIC})$ and $P(N_{\rm CIA})$, to complement either $w_{\rm p}(r_{\rm p})$ or $\Delta\Sigma(r_{\rm p})$ to constrain the galaxy assembly bias parameters. We find that when used in combination with count-based statistics, $w_{\rm p}(r_{\rm p})$ typically outperforms $\Delta\Sigma(r_{\rm p})$. Moreover, it is worth noting that our lensing covariance assumes an infinite density of lensing sources, so lensing constraints realized from a real survey analysis will be further diluted by shape noise contributions to the covariance, as we have discussed in Section III.B.5. Therefore, we suspect that this general result will be robust to actual survey analyses. While Fig. III.3 displays only constraints on $A_{\rm cen}$ and $A_{\rm sat}$ from a single luminosity threshold sample, we find that these qualitative results hold for all thresholds samples considered here (see Fig. III.4 and Appendix C).

2. Complementarity with Clustering

We now turn to a more detailed exploration of the complementarity of various observables with galaxy clustering, as quantified by $w_{\rm p}(r_{\rm p})$. Figure III.4 displays the constraining power of different combinations of observables on the two galaxy assembly bias parameters of our dHOD model: $A_{\rm cen}$, the central galaxy dHOD assembly bias parameter (top panel), and $A_{\rm sat}$, the satellite galaxy dHOD assembly bias parameter (bottom panel). We include in Fig III.4 constraints from $w_{\rm p}(r_{\rm p})$ individually along with constraints from combining $w_{\rm p}(r_{\rm p})$ with each of the other observables that we study. For completeness, we also show the constraints from all the observables combined as an illustration of the maximal constraining power that can be achieved using the complete set of observables considered in our study. The constraints in Fig. III.4 are depicted as bar plots, with the bars grouped by combination of observables. Bars of different colors within each group correspond to different luminosity threshold samples, as indicated. Finally, the heights of the bars represent the fully marginalized 1σ constraints, with smaller values corresponding to tighter, more restrictive constraints. Similar plots for the other dHOD parameters can be found in Appendix C.

Examining the bars in Fig. III.4, several general trends are apparent. Most prominently, similar to what we find for the $M_r < -20.0$ sample in Fig. III.3, for all the luminosity samples we study,

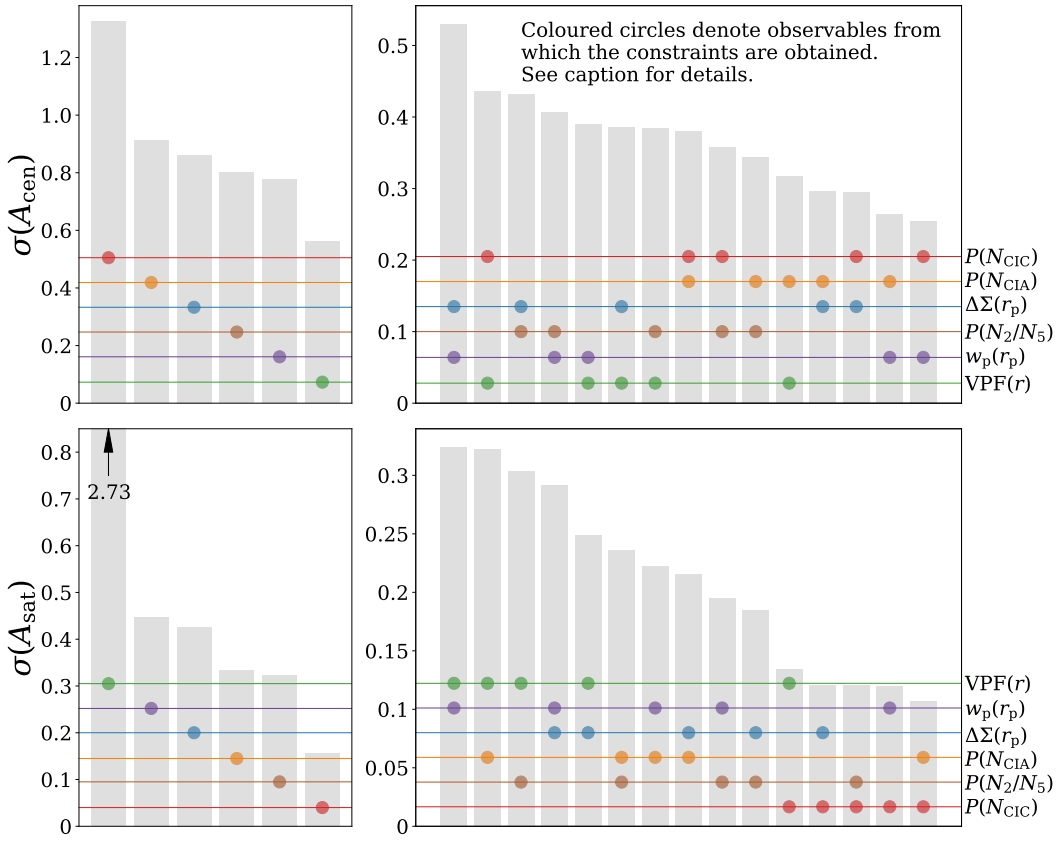


Figure III.3: In this figure, we compare the constraints on $A_{\rm cen}$ and $A_{\rm sat}$ from the individual observables (left-hand panels) and all the possible combinations of two observables (right-hand panels), using the $M_r < -20.0$ sample as an example. The top row shows the constraints on $A_{\rm cen}$ and the bottom row shows the constraints on $A_{\rm sat}$. Note that each of the four panels has a different y-axis range. In each panel, we arrange the columns from least constraining (at left) to most constraining (at right). The filled circles in different colors indicate the observables used to compute the constraints of the corresponding column, as labeled on the right. The relative heights of the circles are ordered by the constraining power from each *individual* observable on the relevant parameter, shown in the left panels. The absolute heights of each colored circle do not correspond to the absolute constraints from that individual observable. In the bottom left panel, the black arrow indicates that the individual constraint from VPF(r) on $A_{\rm sat}$, the value of which is shown below the arrow, greatly exceeds the range of the y-axis.

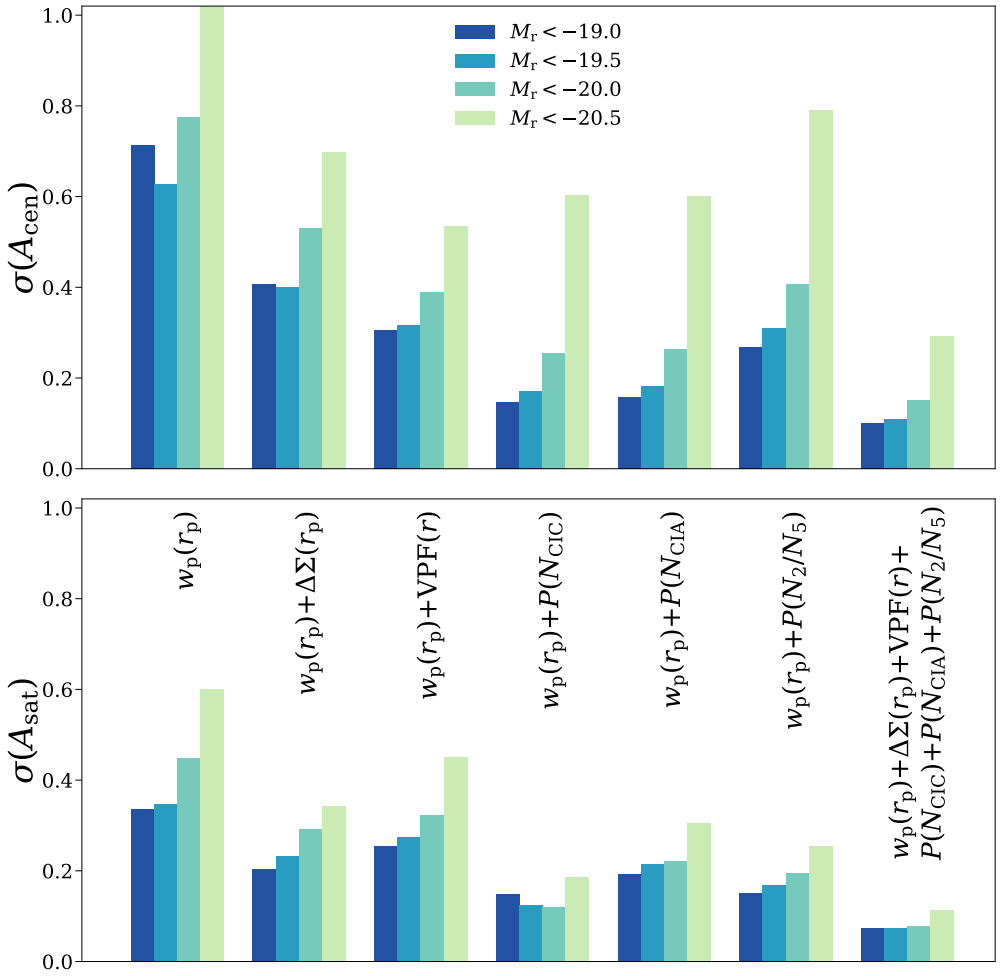


Figure III.4: This figure shows the marginalized 1σ constraint on $A_{\rm cen}$ (top panel), and $A_{\rm sat}$ (bottom panel), as grouped histograms. Each group of bars corresponds to a different combination of observables, and within each group, results for different luminosity thresholds are plotted in different colors, as detailed in the legend. We show $w_{\rm p}(r_{\rm p})$ individually, its combination with every other observable, and the combination of all 6 of our observables. In the upper panel, the constraint from $w_{\rm p}(r_{\rm p})$ for the $M_r < -20.5$ sample exceeds the range of the y-axis, and since $A_{\rm cen}$ is restricted to the range between -1 and 1, $A_{\rm cen}$ is unconstrained in this case.

when combined with $w_{\rm p}(r_{\rm p})$, $P(N_{\rm CIC})$ is generally more effective at constraining $A_{\rm cen}$ and $A_{\rm sat}$ than the commonly-used $\Delta\Sigma(r_{\rm p})$, despite our assumption of infinite lensing source density. In fact, in most cases the count statistics are the most effective observables to combine with clustering in an effort to constrain dHOD models of assembly bias. Furthermore, as is shown in Appendix C, this statement is typical of the constraints on most of the HOD parameters, especially for the fainter samples.

When comparing results for galaxy samples defined by different luminosity thresholds, the constraints are typically tighter for the fainter samples. This mainly reflects the fact that brighter samples have lower number densities, resulting in higher levels of noise. However, there are some exceptions. For example, the constraints on $A_{\rm cen}$ from $w_{\rm p}(r_{\rm p})$ alone are tighter for the $M_r < -19.5$ sample than for the $M_r < -19.0$ sample. In these cases, the degeneracies among different parameters depend on luminosity, such that after marginalization over all other parameters the noisier, high luminosity sample yields tighter constraints. The unmarginalized constraints are all monotonically increasing functions of luminosity threshold, as they must be.

3. Marginalized Two-Dimensional Constraints

In addition to fully marginalized constraints, it is interesting to examine parameter constraints in two-dimensional subspaces of the full parameter space. In Figure III.5, we plot the marginalized 1σ contours in each of the 2D projections of our 6-dimensional parameter space. As we utilize a Fisher matrix to estimate parameter constraints, all contours are elliptical and are centered around the fiducial point in the parameter space. Fig. III.5 corresponds to the $M_r < -20.0$ threshold sample, but the other samples have qualitatively similar features. Different contours correspond to different observables, or combinations thereof, as indicated, and we have highlighted the results for $w_{\rm p}(r_{\rm p}) + \Delta\Sigma(r_{\rm p})$ and $w_{\rm p}(r_{\rm p}) + P(N_{\rm CIC})$ using thicker contours. To avoid crowding, we use $P(N_{\rm CIC})$ as the representative case for the various count statistics. For comparison, the gray shaded ellipse shows the constraints derived from exploiting all of our observables simultaneously.

From Fig. III.5, it is apparent that in all projections, $w_{\rm p}(r_{\rm p}) + P(N_{\rm CIC})$ (thick red lines) is superior to the other combinations of observables, particularly the commonly-used combination of $w_{\rm p}(r_{\rm p}) + \Delta\Sigma(r_{\rm p})$. This result remains strictly true for the $M_r < -19.0$ and $M_r < -19.5$

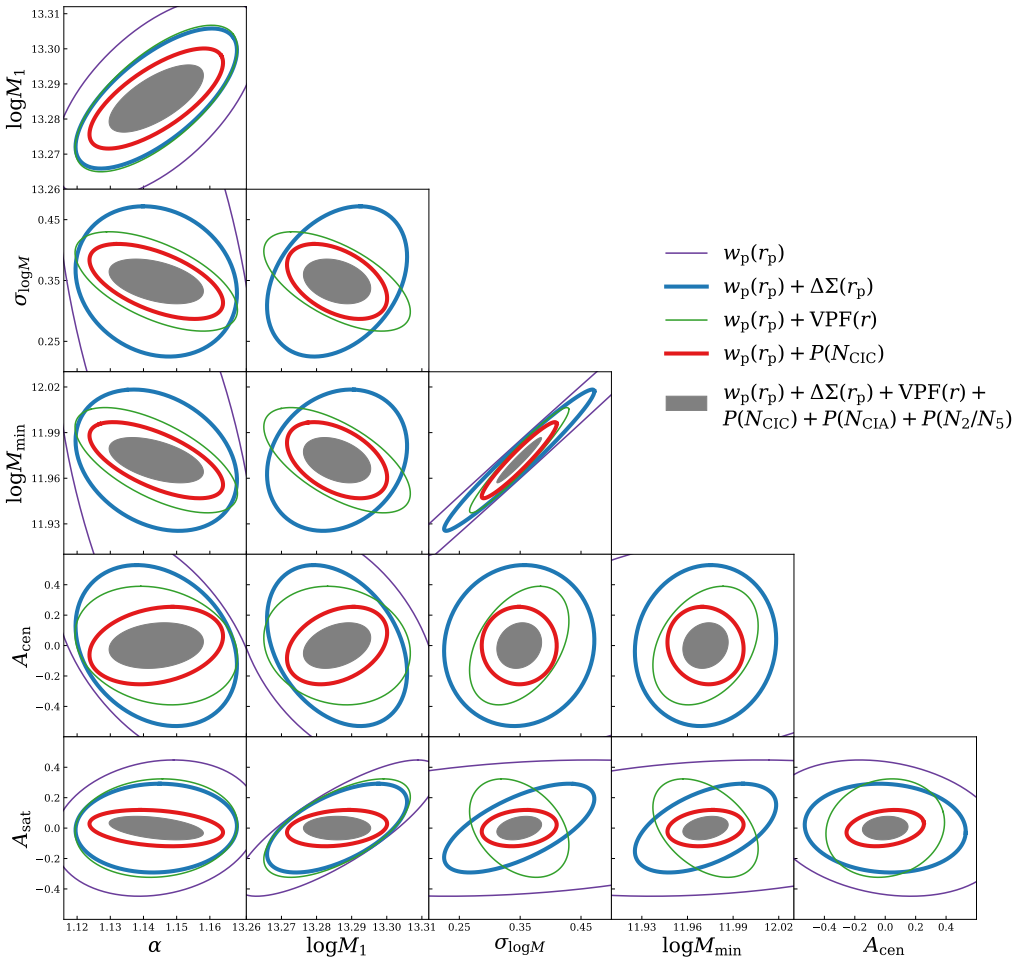


Figure III.5: The marginalized 1σ constraint contours on the dHOD parameters in each of the two-dimensional projections of the dHOD parameter space for the $M_r < -20.0$ sample, from the combinations of $w_p(r_p)$ with $\Delta\Sigma(r_p)$, VPF(r) and $P(N_{\mathrm{CIC}})$ respectively, as well as $w_p(r_p)$ individually. Each such combination is shown in a solid line, color coded consistently with Fig. III.3, as labeled in the legend. We highlight $w_p(r_p) + \Delta\Sigma(r_p)$ and $w_p(r_p) + P(N_{\mathrm{CIC}})$ using thick solid lines. The combination of all 6 observables is shown as shaded regions in gray, to indicate the maximal constraining power in our analyses, and for cross comparison with Fig. III.5. The $w_p(r_p)$ contours are not shown in some panels, because they exceed the range of the axes.

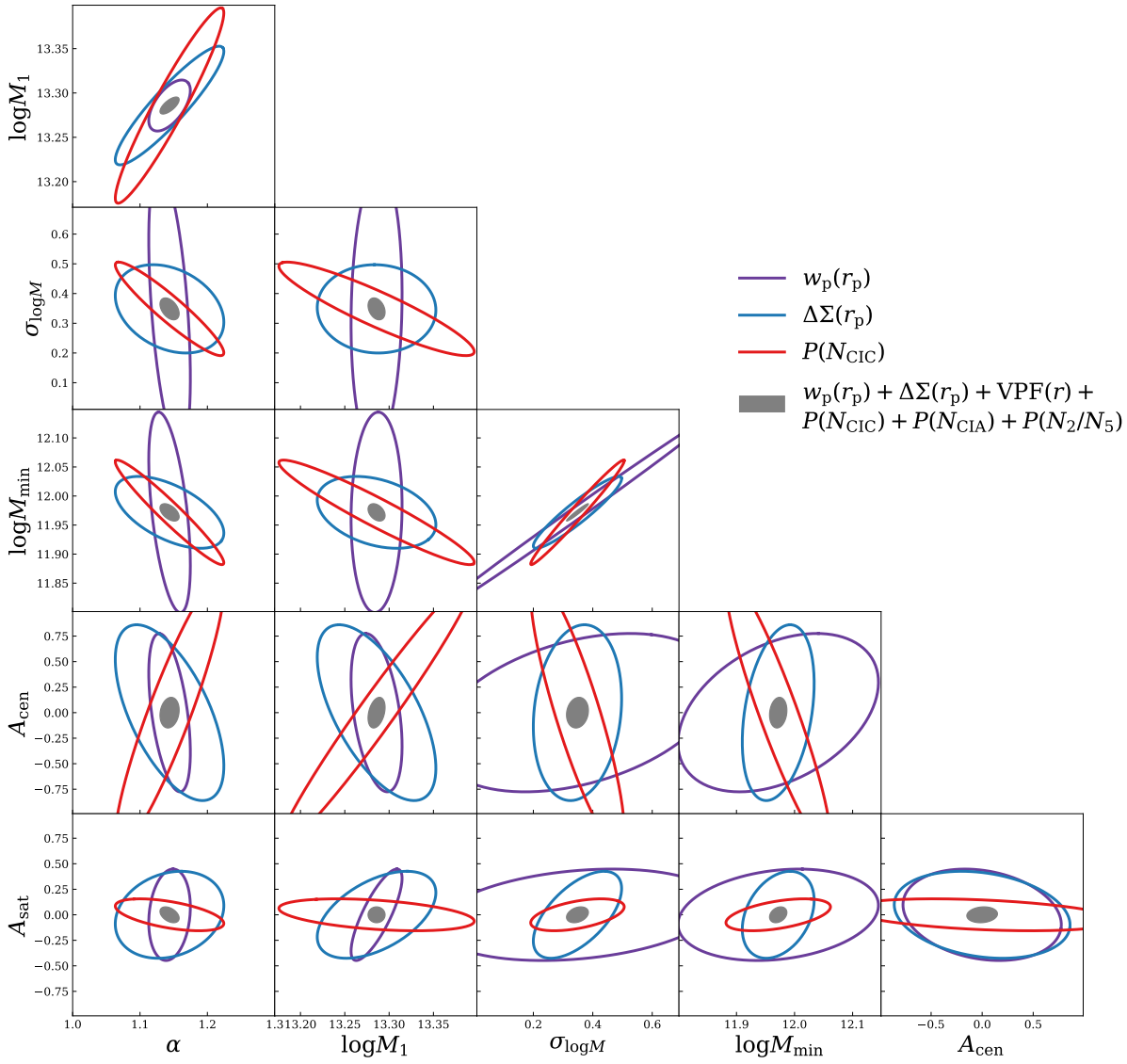


Figure III.6: Same as Fig. III.5, but for the individual observables. We only show $w_p(r_p)$, $\Delta\Sigma(r_p)$ and $P(N_{CIC})$ for clarity. The combination of all 6 observables is shown as shaded regions in gray, for cross comparison with Fig. III.5.

samples, while for the brighter, $M_r < -20.5$ sample, a few parameter combinations are more tightly constrained using parameter combinations other than $w_{\rm p}(r_{\rm p}) + P(N_{\rm CIC})$.

Figure III.6 displays confidence contours for our model parameters constrained by $w_{\rm p}(r_{\rm p})$, $\Delta\Sigma(r_{\rm p})$ and $P(N_{\rm CIC})$ individually (rather than combinations of observables). This visualization can aid in the qualitative understanding of our results. Combining $w_{\rm p}(r_{\rm p})$ with $\Delta\Sigma(r_{\rm p})$ yields limited improvement because both of these observables share similar degeneracy directions in multiple dimensions of the parameter space. The combination of $w_{\rm p}(r_{\rm p})$ with $P(N_{\rm CIC})$ is superior because these observables have largely complementary degeneracy directions in the parameter space and combining these observables leads to the simultaneous breaking of multiple degeneracies. Notice that constraints from $P(N_{\rm CIC})$ on any single parameter are not particularly restrictive; however, the constraints from $P(N_{\rm CIC})$ restrict parameter values to exceedingly narrow degeneracy regions, which, in turn, leads to significant improvements in constraining power when combined with $w_{\rm p}(r_{\rm p})$.

4. Constraints on Parameters of the Standard HOD

In the previous subsections, we focused on constraints on assembly bias parameters in the dHOD model and showed that counts-in-cylinders is an effective complement to the galaxy projected two-point function for diagnosing and constraining assembly bias. However, it is also interesting to study constraints on the standard HOD parameters in a standard HOD model that does not include assembly bias.

Figure III.7 depicts 2D marginalized, projected constraint contours on the standard HOD parameters from an analysis to constrain a standard HOD model using the various observables that we consider. It is clear that the complementarity of $P(N_{\rm CIC})$ extends to the parameters of the standard HOD, as the combination of $w_{\rm p}(r_{\rm p})$ with $P(N_{\rm CIC})$ outperforms the combination of $w_{\rm p}(r_{\rm p})$ with $\Delta\Sigma(r_{\rm p})$ in all projections. This strongly suggests that $P(N_{\rm CIC})$ is a favorable observable even in studies using standard HOD.

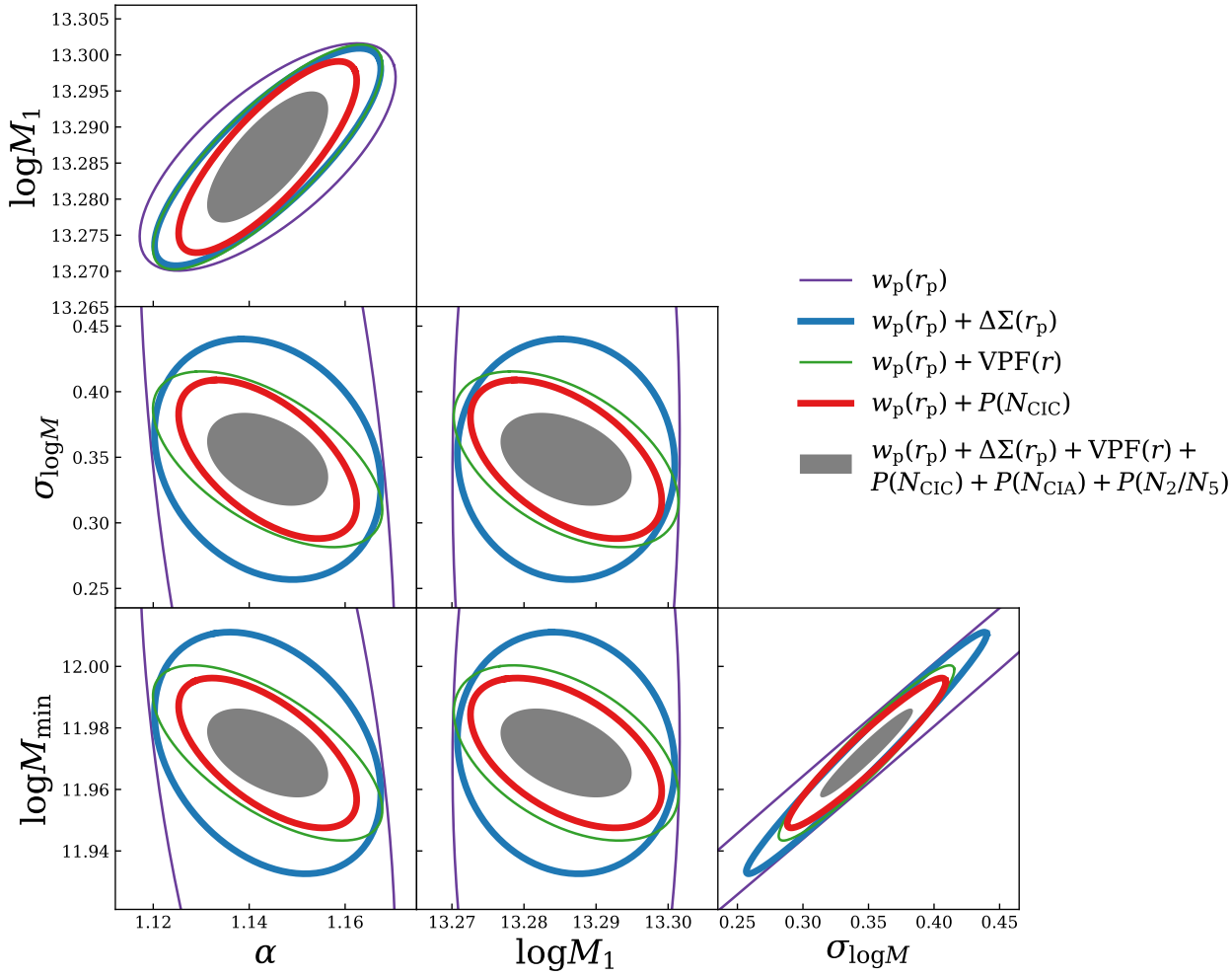


Figure III.7: Marginalized 1 sigma constraint contours in two-dimensional projections of the HOD parameter space from observable combinations. Same as Fig. III.5, but for analysis on the standard HOD model. In this figure we show representative results for the $M_r < -20.0$ luminosity threshold sample.

5. Limitations and Caveats

Our results are subject to several limitations and caveats that we describe in this subsection. First, all Fisher matrix analyses are based on a linear expansion of the likelihood: The observable–parameter relation is treated as a linear function of the observable on the parameter. This linearity is not an accurate model of the observable–parameter relation over the entirety of the relevant domain of parameters for all of the observable–parameter combinations that we explore. As a result, the derivative values $\partial \mathbf{f}/\partial \mathbf{p}$, and therefore the constraints, depend on the choice of the fiducial point in the parameter space. For the results presented above, the fiducial point is motivated by previous data analyses using a standard HOD, specifically, the study of Ref. [223]. The fiducial values of the assembly bias parameters, $A_{\rm cen}$ and $A_{\rm sat}$ are set to zero. We address this particular caveat in the following subsection, pointing out the dependence of our conclusions upon the fiducial location in the parameter space.

Second, the decorated HOD parameters $A_{\rm cen}$ and $A_{\rm sat}$ can only vary over the interval $[-1, 1]$ because the degree of galaxy assembly bias that is possible is limited (see Ref. [74] for details). The Fisher formalism assumes a multivariate Gaussian posterior distribution, so such hard boundaries on the parameter space can lead to gross violations of this assumption. In particular, any time that $\sigma_{A_{\rm cen}}$ or $\sigma_{A_{\rm sat}}$ approach unity (or even exceed it, see Appendix C), our estimates will not be a reliable, quantitative estimate of the constraining power of the observables. However, the *qualitative* comparisons among observables should not be impacted by this shortcoming and the Fisher matrix will still give a reliable ranking of the relative utility of different combinations of observables.

An additional caveat to our results is associated with the particular model that we explore. While we phrase our results qualitatively in terms of constraints on galaxy assembly bias, it is important to realize that our calculations pertain only to a specific model, namely the dHOD with a binary split on galaxy populations. It is possible that our conclusions would change significantly if a different halo occupation model is used. Examples of different models might include a standard HOD with an augmented set of parameters or a wholly different model for the galaxy–halo relationship, such as the conditional luminosity function [216] or a parameterized form of abundance matching [100]. There are a limited number of models that include tunable galaxy assembly bias based on halo

properties. Aside from the dHOD, Ref. [100] parametrizes the dependence of galaxy luminosity on halo concentration at a fixed halo mass in the abundance matching model with an interpolation scheme, and Refs. [178, 209, 121] allow for local density-dependent variations of the HOD. We limit our results to the dHOD model and assembly bias due to concentration, and relegate more comprehensive studies of galaxy assembly bias to future work.

Our constraints from the galaxy–galaxy weak lensing signal $\Delta\Sigma$ represent the upper limit of information that can be gained in a real data analysis, because we use a weak lensing covariance that includes *only* sample variance. This is equivalent to assuming an infinite background source galaxy density and thus a shape-noise-free measurement of the galaxy–galaxy lensing signal. Since the galaxy–galaxy lensing signal presented is optimistic, our primary qualitative result, namely that the combination of $w_{\rm p}(r_{\rm p})$+$P(N_{\rm CIC})$ is superior to combinations that include the galaxy–galaxy lensing signal, will not be affected when complete galaxy–galaxy lensing covariances are used.

As a final caveat, we emphasize that our work uses only a single simulation and thus, we work in the context of a single set of cosmological parameters. In particular, we consider the best-fit Planck cosmology used as the modeling framework in the Bolshoi Planck simulation. This limitation is difficult to circumvent at this time due to the need for simulations that are both large volume (to model clustering) and high resolution (to measure the internal properties of haloes). The computational costs of such simulations prohibit simultaneous explorations of assembly bias and cosmology within the scope of this paper. However, important steps are being taken in precisely this direction [226]. Nonetheless, the Planck constraints on cosmological parameters are quite restrictive [143, 142] and we do not expect modifications to the cosmological model to have a significant impact on our qualitative results.

6. Dependence on Fiducial Parameters

One of the caveats mentioned in the previous section is that Fisher analyses yield results that may depend upon the fiducial point in the parameter space about which the likelihood is expanded. Another way to say this is that the constraints depend upon the point in parameter space that corresponds to the *true* underlying model. In the results we presented above, we assumed that the true fiducial model corresponded to zero galaxy assembly bias ($A_{\rm cen} = A_{\rm sat} = 0$).

To examine how our results depend on this choice of the fiducial model, we now repeat our analysis for two alternative assumptions for the fiducial values of the galaxy assembly bias parameters.

In the first, we adopt a fiducial central galaxy assembly bias of $A_{\rm cen} = 0.5$, which is motivated by the recent analysis of galaxy clustering by Ref. [223], while keeping $A_{\rm sat} = 0$. In this case, we find results that are both qualitatively and quantitatively similar to our main model with $A_{\rm cen} = A_{\rm sat} = 0$.

We have also explored the dependence of our forecast constraints on the underlying amount of satellite assembly bias. To do so, we repeated our analyses with the fiducial satellite assembly bias parameter set to $A_{\rm sat} = -0.6$. This value of $A_{\rm sat}$ has several motivations. First, Ref. [223] showed that clustering of galaxies in the SDSS $M_r < -19.5$ threshold sample is consistent with significantly negative values of $A_{\rm sat}$. Furthermore, it is known that the abundance of dark matter subhaloes is anti-correlated with host halo concentration [225, 117, 79]. Since subhaloes are believed to host satellite galaxies, this anti-correlation implies a negative value for $A_{\rm sat}$.

The results of this experiment are shown in Fig. III.8. While the results for $A_{\rm cen}$ (upper panel) are similar to the case of our main model with $A_{\rm cen} = A_{\rm sat} = 0$ (i.e., $A_{\rm cen}$ is always best constrained by the combination of $w_{\rm p}(r_{\rm p})$ plus $P(N_{\rm CIC})$), the results for $A_{\rm sat}$ (lower panel) are notably different. In particular, the combination of $w_{\rm p}(r_{\rm p})$ and $\Delta\Sigma(r_{\rm p})$ now yields the *tightest* constraints on $A_{\rm sat}$, rather than the weakest. This very tight constraint stems from two things. First, $w_{\rm p}(r_{\rm p})$ on small scales ($\lesssim 1\,h^{-1}\,$Mpc) has a much stronger dependence on $A_{\rm sat}$ near $A_{\rm sat} = -0.6$. This improves constraints from $w_{\rm p}(r_{\rm p})$ alone, reducing parameter degeneracy. Second, the observables $w_{\rm p}(r_{\rm p})$ and $\Delta\Sigma(r_{\rm p})$ are more complementary to one another near $A_{\rm sat} = -0.6$ because the degeneracy directions selected by the $w_{\rm p}(r_{\rm p})$ constraints change their orientation slightly in the parameter space. These improvements jointly boost the constraining power of the combination of $w_{\rm p}(r_{\rm p})$ and $\Delta\Sigma(r_{\rm p})$. However, we must note that several observational and theoretical factors that will likely impact the constraining power of very-small-scale clustering statistics are not accounted for in our analysis. Chief among these omissions is our neglect of shape noise in galaxy–galaxy lensing covariances. Including shape noise is likely to reduce significantly the complementarity of $\Delta\Sigma(r_{\rm p})$ with other probes, including $w_{\rm p}(r_{\rm p})$. Modeling uncertainties, such as the choice of the radial distributions of satellite galaxies will also reduce the constraining power of observables on small-scales with similar result. Hence, the tight constraints on $A_{\rm sat}$ we observed here are likely to

be quite optimistic.

To summarize, the results that we have shown so far point toward a clear conclusion: Count-based galaxy clustering statistics, such as counts-in-cells distributions ($P(N_{\rm CIC})$), can be instrumental in constraining galaxy assembly bias, and are particularly powerful in constraining the galaxy assembly bias of central galaxies. For reference, we tabulate the forecast constraints on all dHOD parameters from all of the observable combinations we study, and for all five luminosity threshold samples in Appendix C.

D. Discussion and Conclusions

Constraining galaxy assembly bias is important to the study of the connection between galaxies and haloes and for extracting the maximum possible information on both galaxy evolution and cosmology from survey data. Numerous studies use galaxy clustering to constrain either the galaxy–halo connection or cosmology or both [70, 187, 215, 163, 164, 38, 188, 230, 128, 221, 5, 68, 64, 97, 30, 150, 65, 151, 43, 126, 158, 95, 44, 168, 212]. Several of these works combine clustering with either weak galaxy–galaxy lensing measurements or with measurements of redshift space distortions in order to constrain the galaxy–halo connection and/or cosmology, and this use of complementary variables is becoming increasingly common. However, these results may suffer from systematic bias when galaxy assembly bias is not properly included in the model [224]. To date, there are only a small number of studies constraining assembly bias with galaxy clustering data [100, 223, 183].

We have studied the ability of various galaxy clustering statistics to constrain assembly bias in an effort to determine which combination(s) of observables are most informative. In particular, we have estimated the relative constraining power of several spatial galaxy clustering statistics to constrain the assembly bias parameters of the decorated halo occupation distribution (dHOD). In this first study of its kind, we have restricted our attention to statistics that are not particularly sensitive to galaxy peculiar velocities (however, we do work in redshift space so our results are not completely immune to peculiar velocities). We have chosen to do this because including peculiar velocities (such as redshift space distortions) opens up many distinct modeling questions. Exploring

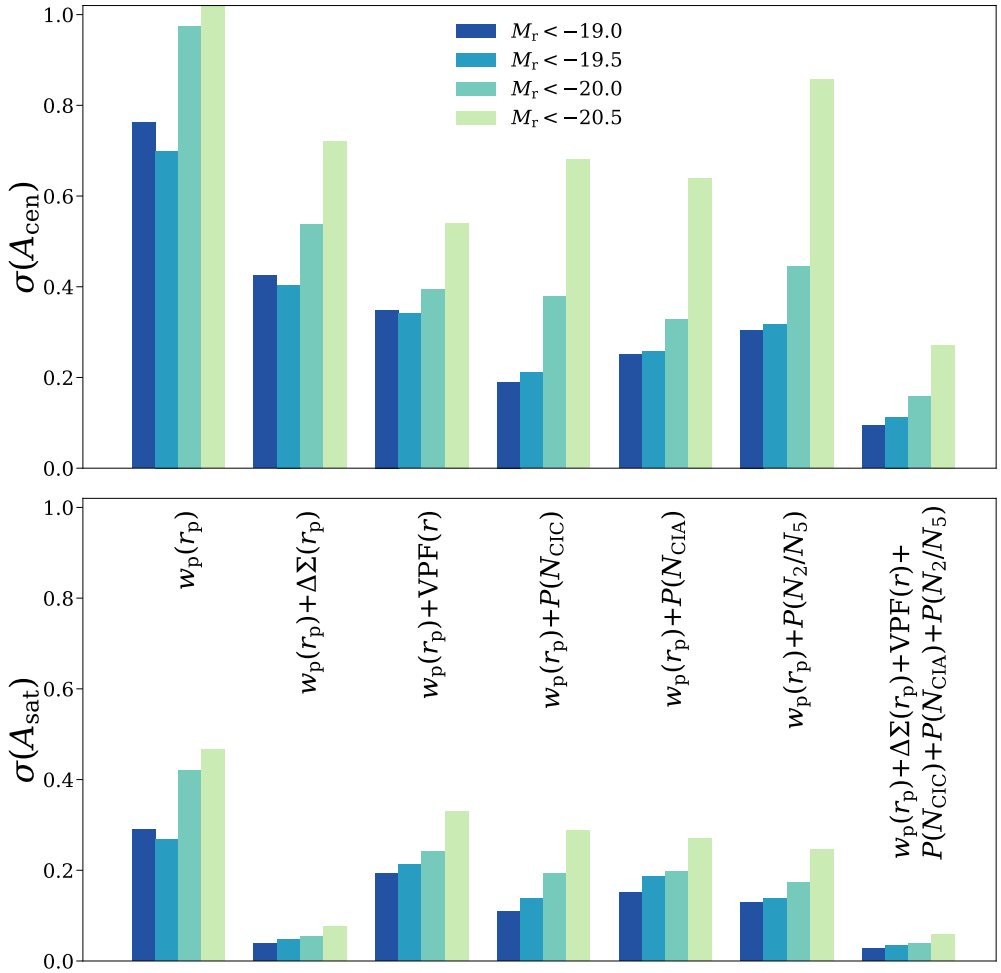


Figure III.8: Marginalized 1 sigma constraint on the assembly bias parameters for alternative fiducial points. Same as Fig. III.4, but with fiducial $A_{sat} = -0.6$.

redshift space distortions will be the subject of future work.

In general, we have found that the combination of $w_{\rm p}(r_{\rm p})$ with the counts-in-cylinders distribution ($P(N_{\rm CIC})$) is the most effective combination of two observables for constraining assembly bias within the dHOD modeling framework. This combination outperforms the commonly-used combination of $w_{\rm p}(r_{\rm p})$ and $\Delta\Sigma(r_{\rm p})$ in constraining central galaxy assembly bias by a factor of ~ 2 for lower luminosity samples, and to a slightly lesser extent in constraining satellite galaxy assembly bias. The primary reason for this is that $w_{\rm p}(r_{\rm p})$ and $\Delta\Sigma(r_{\rm p})$ share roughly common degeneracy directions in the dHOD parameter space, while the combination $w_{\rm p}(r_{\rm p})$ and $P(N_{\rm CIC})$ is much more complementary. This implies that the combination of $\Delta\Sigma(r_{\rm p})$+$P(N_{\rm CIC})$ is nearly as good as $w_{\rm p}(r_{\rm p})$+$P(N_{\rm CIC})$, which we have confirmed. The complementarity between $\Delta\Sigma(r_{\rm p})$ and $P(N_{\rm CIC})$ is not unexpected. The lensing signal traces the matter density contrast around galaxies while the counts-in-cylinders statistics probe the galaxy distribution profiles in approximately the same regions, and are therefore expected to complement each other in constraining cosmology as well as the connection between galaxies and the matter field [63, 57].

We have shown that the count statistics are also effective in constraining the standard HOD parameters, independent of whether the actual halo occupation statistic are affected by galaxy assembly bias or not. In addition, we have tested different fiducial models and find that when strong satellite assembly bias is present, the combination of $w_{\rm p}(r_{\rm p}) + \Delta\Sigma(r_{\rm p})$ actually provides the tightest constraints on $A_{\rm sat}$ among all sets of observables studied here. Therefore we caution that the preferred statistics may depend on the true, underlying relationship between galaxies and dark matter haloes, and the degree to which galaxy assembly bias is realized in nature.

Our results complement recent work studying the information that can be gained from higher-order statistics beyond the two-point function. In particular, in Ref. [217, 218] it was shown that the three-point function in the squeezed limit contains significant additional constraining power on HOD parameters that is complementary to $w_{\rm p}(r_{\rm p})$. This is consistent with our findings in the sense that the complementarity of counts statistics to $w_{\rm p}(r_{\rm p})$ derives precisely from the extraction of information in higher-order n-point moments of the density field. Considering the results in Ref. [218] together with our findings, statistics beyond two-point clustering and lensing should be seriously considered in future analyses of large-scale structure data that utilize models of the galaxy–halo connection.

In conclusion, our results strongly suggest that galaxy assembly bias may be significantly better constrained and/or better understood by employing simple counts statistics as measured from forthcoming and present-day data sets. There are numerous forthcoming data sets with which can be used to inform assembly bias and/or whose interpretation may be challenged by small levels of assembly bias. These include large redshift surveys, such as may be carried out by DESI [101] or WFIRST [170]. As we have already mentioned, interesting follow up work includes an exploration of velocity statistics, such as redshift space distortions, in redshift surveys. While our work relates specifically to redshift surveys, it would be interesting to explore possible avenues for studying assembly bias within photometric surveys, such as the DES [176, 2] and LSST [108]. It is our aim to study and deploy these statistics to constrain the galaxy–halo connection and to encourage others to do the same.

IV. Observational Constraints on Galaxy–Halo Connection

Following the work in Chapter III, we apply the findings to real data in this chapter. Namely, we measure the optimal set of statistics identified previously from observational data, and use them to constrain the halo occupation distribution (HOD) model of galaxy–halo connection, especially the galaxy assembly bias parameters.

This chapter is organized as follows. The observational data and cosmological simulation are described in Section IV.A. In Section IV.C, we detail the construction of light cone galaxy mocks, that are used for validating the algorithms with which we measure our observable statistics. We describe the statistics, our measuring algorithms, and the validation process in Section IV.D. In Section IV.B, we present measurements of the statistics from data. We conduct a Markov Chain Monte Carlo fit and obtain constraints on the galaxy–halo connection, in particular galaxy assembly bias, in Section IV.E. We discuss implications and draw conclusions in Section IV.F.

A. Data and Simulation

In this section, we describe the observational data and cosmological simulation that are used in this work.

1. Data and Sample Selection

In this work, we use galaxy data from the Sloan Digital Sky Survey Data Release 7 (SDSS DR7, [1]). In particular, we select our samples from the `bright0` catalog[1], with r-band apparent magnitudes $10.0 < m_r < 17.6$, in the NYU Value-Added Galaxy Catalog (NYU VAGC, [21]). The sample contains galaxies that fall within the survey window, with the bright star-contaminated areas masked out. We additionally discard sector areas (intersections of tile regions) with low fractions of galaxies that have spectroscopic redshift measurements, i.e., we require that the sector fraction

[1] `http://sdss.physics.nyu.edu/lss/dr72/bright/0/`

$f_{\rm sector} \geq 0.8$. We measure the resulting angular area of the data footprint to be approximately 7461 deg^2, and the total number of galaxies in the sample is 562620.

Due to the finite size of fibers used in the survey, no two targets on the same plate can be closer than 55", which results in a fraction of targeted galaxies not having a measured redshift (known as the fiber collision effect). These galaxies are assigned the redshifts of their nearest neighbors.

We select volume-limited, luminosity-threshold samples, based on r-band absolute magnitudes that are K-corrected [20] and passively evolved to the median redshift of the DR7 main galaxy sample, $z = 0.1$. The absolute magnitude values we list as M_r throughout this paper are in fact values of $M_r - 5\log h$ for $h = 1$, which are measured independently of h. We apply a universal lower limit of $z_{\rm min} = 0.02$ to all of our samples, and adopt the upper bounds of redshift for each luminosity threshold in Ref. [221]. Our sample selection is illustrated in Fig. IV.1, and the details are listed in Table IV.1.

2. Simulation

Our forward modeling analyses are based on the Small MultiDark Planck Simulation (SMDPL), which is a gravitational N-body simulation that belongs to the series of MultiDark simulations with Planck cosmology [87]. The Small MultiDark Planck Simulation has a cubic volume of side length 400 h^{-1} Mpc, which is comparable to the volume of our $M_r < -21.0$ data sample. The cosmological parameters adopted are $\Omega_\Lambda = 0.6929$, $\Omega_{\rm m} = 1 - \Omega_\Lambda = 0.3071$, $\Omega_{\rm b} = 0.0482$, $h = 0.6777$, $n_{\rm s} = 0.96$, and $\sigma_8 = 0.8228$. We assume this cosmology in our analyses throughout this paper. The simulation is evolved from $z_{\rm initial} = 120$, with 3840^3 particles, implying a particle mass resolution of $m_{\rm p} = 9.63 \times 10^7 h^{-1}\,{\rm M}_\odot$. We use halo catalogs extracted from the $z = 0.1$ snapshot using the `Rockstar` halo-finder [13], downloaded from `https://www.cosmosim.org`. We truncate the halo peak mass $M_{\rm peak}$ at $300 \times m_{\rm p} = 2.889 \times 10^{10} h^{-1}\,{\rm M}_\odot$, below which halos are extremely unlikely to host galaxies above our luminosity thresholds.

$M_{r,\max}$	$cz_{\max}\,[\mathrm{km\,s^{-1}}]$	N_{gal}
-21.0	47650	82263
-20.5	39700	130707
-20.0	31900	140149
-19.5	25450	131322
-19.0	19250	76442
-20.0^{*}	19250	29951
-19.5^{*}	19250	51007

Table IV.1: Volume-limited samples with luminosity threshold. All of the samples have minimum redshift $z_{\min} = 0.02$. Redshift upper bounds are listed in terms of their products with the speed of light, c. The galaxies in each sample satisfy the conditions $z_{\min} \leq z_{\mathrm{obs}} < z_{\max}$ and $M_r < M_{r,\max}$. The second part of the table lists alternatively selected samples for the $M_r < -20.0$ and $M_r < -19.5$ thresholds (marked with asterisks), excluding the cosmic structure known as the Sloan Great Wall.

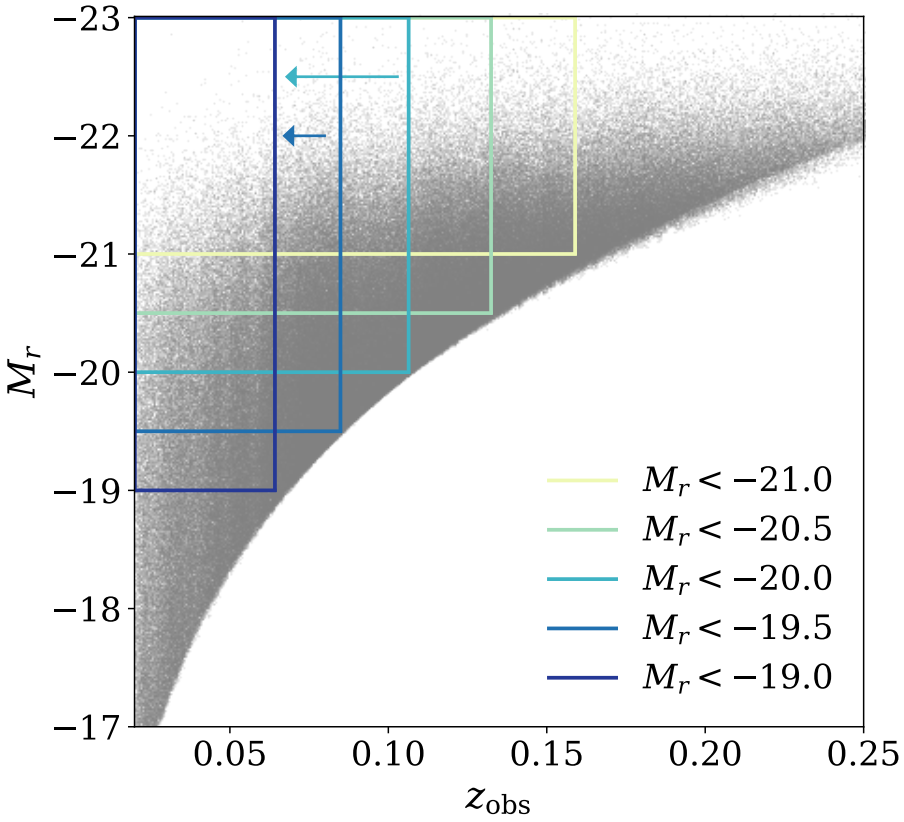


Figure IV.1: Illustration of our volume-limited, luminosity-threshold galaxy samples from the NYU VAGC `bright0` catalog. $z_{\rm obs}$ is the observed redshift, where fiber collided galaxies are assigned the redshifts of their nearest neighbors. M_r is the r-band absolute magnitude. The scatter points represent the galaxies from the catalog, and the colored boxes mark the selection criteria of the different samples, as are labeled in the figure. The arrows indicate the alternative $M_r < -20.0$ and $M_r < -19.5$ samples with lower redshift limits that exclude the Sloan Great Wall.

B. Data Measurement

In this section, we present statistic measurements from our SDSS galaxy samples.

1. Observable Statistics

We employ summary statistics to extract information from the spatial distribution of galaxies. In W19, we showed that the combination of the projected two-point correlation function $w_{\rm p}(r_{\rm p})$ and the counts-in-cylinders statistic $P(N_{\rm CIC})$ yields tight constraints on the galaxy–halo connection by breaking degeneracies in the model parameter space. In this work, we elect to measure these statistics along with the galaxy number density.

The projected two-point correlation function is defined by

$$w_{\rm p}(r_{\rm p}) = 2\int_0^{\pi_{\rm max}} d\pi\, \xi(r_{\rm p}, \pi), \tag{20}$$

where $\xi(r_{\rm p}, \pi)$ is the excess probability of finding galaxy pairs with projected and line-of-sight separations $r_{\rm p}$ and π, respectively. It is commonly used in previous works [188, 231, 221], to inform halo occupation. Considering the depth of the data samples and the size of the simulation, we choose $\pi_{\rm max} = 40h^{-1}$ Mpc, to include most correlated pairs and reduce the impact of peculiar velocities, while excluding very distant, uncorrelated pairs. We compute $w_{\rm p}(r_{\rm p})$ in 12 logarithmically spaced radial bins between $r_{\rm p} = 0.158h^{-1}$ Mpc and $r_{\rm p} = 39.81h^{-1}$ Mpc.

The counts-in-cylinders statistic is the probability distribution of the number of companions found in cylinders around galaxies. As was done in W19, we center a cylinder of transverse radius $r_{\rm CIC} = 2h^{-1}$ Mpc and line-of-sight half-length $L = 10h^{-1}$ Mpc on each galaxy in the sample, and count the number of companion galaxies that fall within the cylinder. We then estimate a probability distribution of companion number, $P(N_{\rm CIC})$, which is the probability that any galaxy has $N_{\rm CIC}$ companions within the cylinder. When characterizing $P(N_{\rm CIC})$, we bin $N_{\rm CIC}$ values linearly on the lower end and logarithmically on the higher end, as listed in Table IV.4 and Table IV.5.

2. Measurement Algorithm

For fiber collided galaxies, we adopt the nearest-neighbor corrected redshifts for both sample selection and statistic measurements. The angular and redshift separations between points are converted to transverse and line-of-sight separations according to the cosmological model that we adopt.

The number density $n_{\rm gal}$ is calculated by dividing the number of galaxies by the volume within the survey footprint and redshift range of each sample.

For $w_{\rm p}(r_{\rm p})$, we use the Landy-Szalay estimator [93],

$$\hat{\xi}_{\rm LS} = \frac{DD - 2DR + RR}{RR}, \tag{21}$$

where DD is the normalized galaxy–galaxy pair count, RR is the normalized random–random pair count, and DR is the normalized cross pair count between galaxies and randoms. Randoms are drawn from a uniform distribution in the survey volume, to account for the complicated geometry of the light cone. ξ is integrated along the line of sight to $\pi_{\rm max}$. The calculation is done using the CORRFUNC package [167].

For $P(N_{\rm CIC})$, we do not use every galaxy in the sample as a cylinder center, but impose limits such that the sampling of companions is sufficient in the neighborhood of each cylinder center. We define the angular completeness $f_{\rm AC}$ of a galaxy to be the fraction of the circular area with radius $r_{\rm CIC} = 2h^{-1}\,$Mpc around it that falls inside the survey footprint. For cylinder centers we require that the angular completeness around them to be above 0.9, and cylinders centered on them to be completely within the redshift ranges of the volume-limited samples. We then count companions in these cylinders, and upscale the count numbers by $1/f_{\rm AC}$, which results in non-integer counts. The counts are binned to yield the $P(N_{\rm CIC})$ statistic.

3. Covariance Estimation

We calculate the jackknife covariance, which provides an estimate of the uncertainty due to the finite volume of the survey. We use the method described in Ref. [232][2], to divide the survey

[2]The code is available at `https://github.com/rongpu/pixel_partition`

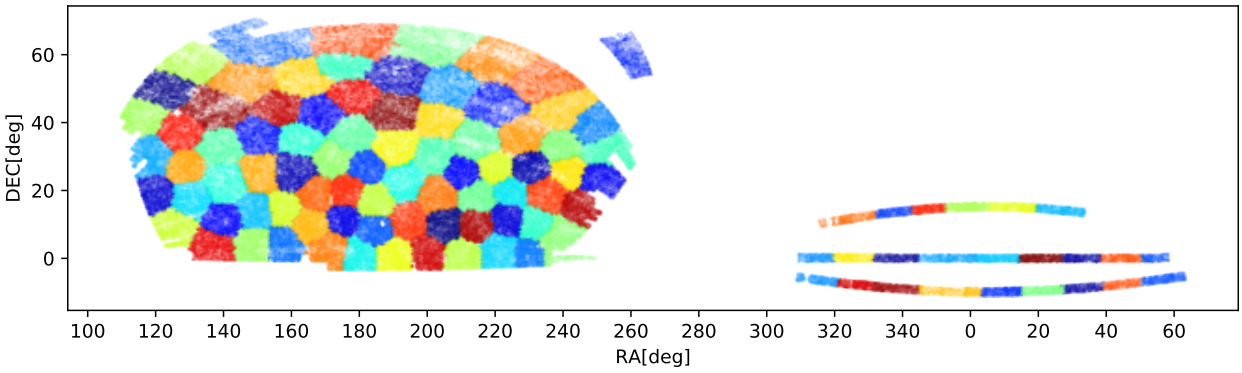


Figure IV.2: SDSS footprint divided into jackknife cells. Objects in different cell are plotted in different colors. All the cells are similar in size, except some in the three stripes away from the main footprint, which are smaller than average.

footprint into subareas of similar sizes. We show a map of the cells in Fig. IV.2, where galaxies in different cell are plotted in different colors.

For each sample, we exclude one cell at a time in our measurement and record the resulting covariance matrices as our jackknife covariance. Note that for $w_{\rm p}(r_{\rm p})$, we exclude a pair of objects if either or both are in the excluded cell, whereas for $P(N_{\rm CIC})$, we keep all the galaxies as companions, and only exclude cylinder centers that fall in the excluded cell from the probability distribution.

4. Measurement Results

The $w_{\rm p}(r_{\rm p})$ values we measure from data are shown in Section IV.B.4, where each luminosity threshold is plotted individually in the first five panels, and compared alongside each other in the last panel. The error bars show the jackknife error of each $r_{\rm p}$ bin. For the $M_r < -20.0$ and $M_r < 19.5$ thresholds, the samples with the alternative shallower redshift ranges are plotted in gray. Number densities are also shown in text in the corresponding panels, with jackknife errors included in parentheses. $n_{\rm gal}$ values for the alternative samples are marked with asterisks. The $P(N_{\rm CIC})$ statistics are similarly shown in Fig. IV.B.4. The values of the statistics are listed in Table IV.2, Table IV.3, Table IV.4, and Table IV.5.

Brighter samples have lower number densities, but higher $w_{\rm p}$ values in most $r_{\rm p}$ bins. Excluding the Sloan Great Wall results in lower number densities and weaker large-scale clustering, as is expected. The $P(N_{\rm CIC})$ values in all the bins sum up to unity, as $P(N_{\rm CIC})$ is a probability distribution. Brighter galaxies are rarer, and tend to have fewer companions, resulting in higher probabilities of smaller $N_{\rm CIC}$, and lower probabilities of larger $N_{\rm CIC}$. The comparison between the different redshift limit samples for the $M_r < -20.0$ and $M_r < -19.5$ thresholds reflects the non-trivial impact of the Sloan Great Wall on the counts-in-cylinders statistic. For both thresholds, excluding the Great Wall increases the probability of galaxies having only a few companions, and reduces that of having more companions in general. However, the shallower $M_r < -20.0$ sample has higher $P(N_{\rm CIC})$ values at high $N_{\rm CIC}$, though the effect is not statistically significant. The samples that have smaller volumes have larger jackknife errors for all of the observables.

C. Mock Building for Algorithm Validation

In order to fit SDSS data using a simulation, we need measurements of the observable statistics from the simulation to reflect the behavior of real data. However, light cone mocks that mimic observed data are computationally expensive, and therefore infeasible for sampling from the high dimensional model parameter space, whereas measurements from the original cubic volume of the simulation can be obtained much more rapidly. Consistency between statistic measurements from the cube and the light cone mocks that have the same underlying physics would validate the use of the cube for fitting data. Before describing the consistency check in Section IV.D, here we detail the construction of our light cone mocks that incorporate the relevant geometrical and observational effects present in SDSS data.

1. Halo Populating in Simulation Cube

We populate the halo catalog at $z = 0.1$ from SMDPL using the stellar mass–(sub)halo mass relation model described in Ref. [11] to get a galaxy catalog in the cubic simulation volume (hereafter the "**cube mock**"). This model is implemented in the `halotools` software package

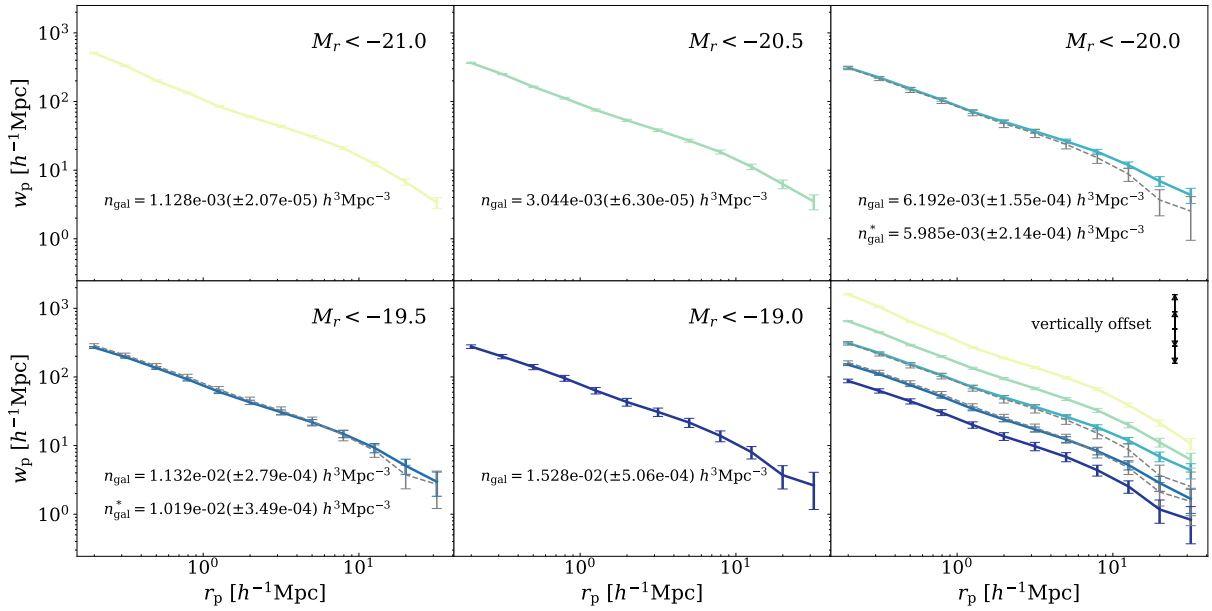


Figure IV.3: Measurement of the galaxy number density $n_{\rm gal}$ and the projected two-point function $w_{\rm p}$ from SDSS data. Each of the first five panels shows the measurement for one luminosity sample with jackknife error bars that represent the cosmic variance. For the -20.0 and -19.5 thresholds, results for the alternative samples without the Sloan Great Wall are shown in the respective panels as values with asterisks and gray lines. The bottom right panel shows the $w_{\rm p}(r_{\rm p})$ of the five thresholds together, with vertical offsets of 0.25 dex each, starting from the -20.0 sample, for visual clarity.

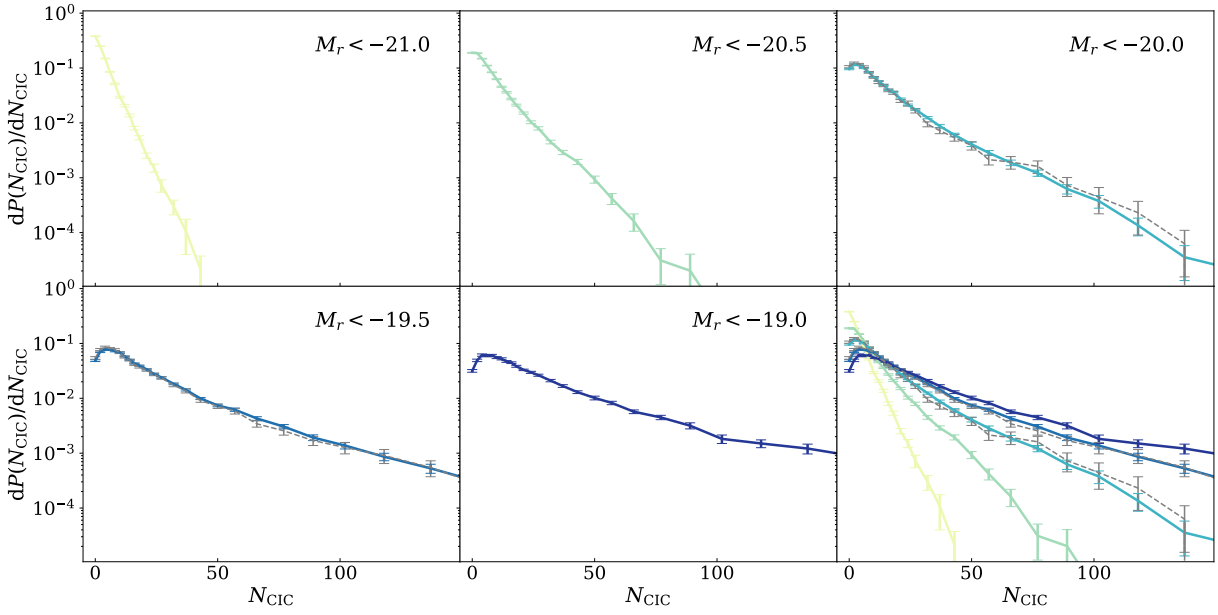


Figure IV.4: Measurement of the counts-in-cylinders statistic $P(N_{\rm CIC})$ from SDSS data. Similar to Section IV.B.4, each of the first five panels shows the measurement for one luminosity sample with jackknife error bars that represent the cosmic variance. The statistic is represented as the probability distribution of $N_{\rm CIC}$, normalized by the bin widths. For the -20.0 and -19.5 thresholds, results for the alternative samples without the Sloan Great Wall are shown in the respective panels as gray lines. The bottom right panel shows the $P(N_{\rm CIC})$ of the five thresholds together.

	$M_r < -21.0$	$M_r < -20.5$	$M_r < -20.0$	$M_r < -19.5$
$n_{\rm gal}$	1.128(0.021)	3.044(0.063)	6.192(0.155)	11.32(0.279)
0.200	508.36(13.00)	365.33(7.38)	311.92(8.25)	268.89(8.67)
0.316	332.75(8.30)	251.26(4.73)	224.05(6.71)	195.09(7.68)
0.501	201.96(4.74)	163.95(4.17)	154.16(5.82)	134.95(6.13)
0.794	134.25(3.51)	111.89(3.23)	106.07(4.84)	92.18(4.97)
1.259	85.17(2.32)	75.32(2.58)	70.86(3.78)	60.96(3.72)
1.995	59.77(1.76)	53.23(2.04)	50.78(3.06)	42.72(3.13)
3.162	43.54(1.51)	38.23(1.62)	36.87(2.46)	30.64(2.47)
5.012	30.86(1.13)	26.89(1.35)	26.36(2.04)	21.87(2.07)
7.943	20.96(1.00)	18.39(1.21)	18.47(1.66)	14.80(1.79)
12.59	12.35(0.81)	11.21(1.03)	11.84(1.32)	9.27(1.42)
19.95	6.75(0.70)	6.26(0.91)	6.92(1.11)	5.07(1.26)
31.62	3.37(0.63)	3.50(0.86)	4.35(1.09)	2.97(1.15)

Table IV.2: Measured values of the observable statistics from SDSS data. $n_{\rm gal}$ and $w_{\rm p}(r_{\rm p})$ measurements are listed. The leftmost column shows the $r_{\rm p}$ bin center in units of h^{-1} Mpc for $w_{\rm p}$. The number density values $n_{\rm gal}$ are listed in units of $10^{-3}h^3{\rm Mpc}^{-3}$, and the values listed for $w_{\rm p}(r_{\rm p})$ are in units of h^{-1} Mpc. Jackknife errors are shown in brackets. This table shows results for four samples, and is continued in Table IV.3.

	$M_r < -19.0$	$M_r < -20.0^*$	$M_r < -19.5^*$
$n_{\rm gal}$	15.28(0.506)	5.985(0.214)	10.19(0.349)
0.200	276.92(16.64)	308.88(17.98)	288.03(17.14)
0.316	198.23(14.26)	216.09(14.45)	207.49(14.93)
0.501	139.73(12.11)	148.19(12.91)	144.09(12.32)
0.794	95.93(9.47)	103.46(9.92)	99.67(9.62)
1.259	63.29(6.79)	68.98(6.96)	65.68(6.67)
1.995	43.15(5.57)	47.56(5.88)	45.20(5.55)
3.162	30.89(4.32)	34.55(4.74)	32.29(4.31)
5.012	21.65(3.29)	23.78(3.38)	22.65(3.34)
7.943	13.86(2.50)	15.16(2.61)	14.27(2.54)
12.59	8.03(1.65)	8.77(1.91)	8.48(1.74)
19.95	3.72(1.38)	3.67(1.51)	3.75(1.41)
31.62	2.63(1.45)	2.53(1.58)	2.73(1.52)

Table IV.3: Measured values of the observable statistics from SDSS data (Part 2). Table IV.2 continued, for the other three samples.

	$M_r < -21.0$	$M_r < -20.5$	$M_r < -20.0$	$M_r < -19.5$
-0.5, 0.5	0.379(4.82e-3)	0.189(3.87e-3)	0.096(2.82e-3)	0.049(1.82e-3)
0.5, 1.5	0.249(2.49e-3)	0.187(2.80e-3)	0.117(2.91e-3)	0.071(2.48e-3)
1.5, 2.5	0.149(2.07e-3)	0.148(1.87e-3)	0.113(2.43e-3)	0.078(2.47e-3)
2.5, 3.5	0.085(1.76e-3)	0.111(1.49e-3)	0.099(1.96e-3)	0.077(2.22e-3)
3.5, 4.5	0.052(1.62e-3)	0.083(1.27e-3)	0.084(1.59e-3)	0.072(2.04e-3)
4.5, 5.5	0.029(1.21e-3)	0.063(1.19e-3)	0.069(1.14e-3)	0.068(1.75e-3)
5.5, 6.5	0.021(1.08e-3)	0.046(1.05e-3)	0.058(1.05e-3)	0.058(1.31e-3)
6.5, 7.5	0.014(8.48e-4)	0.036(1.09e-3)	0.049(1.06e-3)	0.050(1.03e-3)
7.5, 8.5	0.008(5.99e-4)	0.027(8.79e-4)	0.040(9.48e-4)	0.043(1.05e-3)
8.5, 9.5	0.005(4.64e-4)	0.021(8.50e-4)	0.036(9.49e-4)	0.039(8.89e-4)
9.5, 11.5	0.005(5.55e-4)	0.030(1.38e-3)	0.055(1.60e-3)	0.067(1.70e-3)
11.5, 12.5	0.002(2.56e-4)	0.010(6.21e-4)	0.021(7.80e-4)	0.028(8.75e-4)
12.5, 14.5	0.001(3.72e-4)	0.016(1.10e-3)	0.035(1.36e-3)	0.048(1.53e-3)
14.5, 17.5	0.001(2.71e-4)	0.013(1.05e-3)	0.037(1.69e-3)	0.055(2.01e-3)
17.5, 19.5	0.000(1.35e-4)	0.006(5.12e-4)	0.018(1.07e-3)	0.028(1.28e-3)
19.5, 23.5	0.000(6.37e-5)	0.008(8.23e-4)	0.024(1.83e-3)	0.040(1.74e-3)
23.5, 26.5	0	0.003(4.13e-4)	0.012(1.13e-3)	0.022(1.25e-3)
26.5, 30.5	0	0.002(3.80e-4)	0.011(1.13e-3)	0.025(1.69e-3)
30.5, 35.5	0	0.001(2.83e-4)	0.009(1.09e-3)	0.021(1.63e-3)
35.5, 41.5	0	0.000(1.19e-4)	0.007(9.51e-4)	0.018(1.66e-3)
41.5, 47.5	0	0.000(1.21e-4)	0.004(7.29e-4)	0.012(1.38e-3)
47.5, 54.5	0	0.000(1.87e-5)	0.003(6.83e-4)	0.010(1.30e-3)
54.5, 63.5	0	0	0.001(4.34e-4)	0.008(1.21e-3)
63.5, 73.5	0	0	0.000(2.22e-4)	0.005(9.92e-4)
73.5, 84.5	0	0	0.000(2.32e-4)	0.003(8.14e-4)
84.5, 97.5	0	0	0.000(1.43e-4)	0.002(7.83e-4)
97.5, 112.5	0	0	0	0.001(4.39e-4)
112.5, 129.5	0	0	0	0.000(7.32e-5)
129.5, 149.5	0	0	0	0

Table IV.4: Measured values of the observable statistics from SDSS data. $P(N_{\rm CIC})$ measurements are listed. The leftmost column shows the edges of each bin in $N_{\rm CIC}$ for $P(N_{\rm CIC})$. The values listed for $P(N_{\rm CIC})$ are the probability in each bin, which would sum up to unity. Jackknife errors are shown in brackets. This table shows results for four samples, and is continued in Table IV.5.

	$M_r < -19.0$	$M_r < -20.0^*$	$M_r < -19.5^*$
-0.5, 0.5	0.032(1.79e-3)	0.106(4.95e-3)	0.054(2.97e-3)
0.5, 1.5	0.049(2.53e-3)	0.123(5.43e-3)	0.077(3.82e-3)
1.5, 2.5	0.061(2.98e-3)	0.117(4.86e-3)	0.086(3.51e-3)
2.5, 3.5	0.060(2.87e-3)	0.106(3.92e-3)	0.081(3.22e-3)
3.5, 4.5	0.061(2.50e-3)	0.084(2.93e-3)	0.077(3.25e-3)
4.5, 5.5	0.055(2.24e-3)	0.068(2.50e-3)	0.068(2.47e-3)
5.5, 6.5	0.052(2.30e-3)	0.055(2.87e-3)	0.056(2.01e-3)
6.5, 7.5	0.047(1.76e-3)	0.048(2.42e-3)	0.050(1.88e-3)
7.5, 8.5	0.045(1.60e-3)	0.039(2.25e-3)	0.041(1.61e-3)
8.5, 9.5	0.038(1.54e-3)	0.036(2.36e-3)	0.037(1.59e-3)
9.5, 11.5	0.066(1.98e-3)	0.050(2.98e-3)	0.065(2.93e-3)
11.5, 12.5	0.030(1.29e-3)	0.023(2.38e-3)	0.027(1.41e-3)
12.5, 14.5	0.053(2.28e-3)	0.033(2.71e-3)	0.047(2.61e-3)
14.5, 17.5	0.063(3.12e-3)	0.028(2.97e-3)	0.052(3.21e-3)
17.5, 19.5	0.034(1.93e-3)	0.014(1.75e-3)	0.027(2.26e-3)
19.5, 23.5	0.052(3.00e-3)	0.022(2.99e-3)	0.036(2.44e-3)
23.5, 26.5	0.030(2.39e-3)	0.012(2.00e-3)	0.022(1.90e-3)
26.5, 30.5	0.033(2.35e-3)	0.009(1.72e-3)	0.023(2.80e-3)
30.5, 35.5	0.028(2.19e-3)	0.010(2.49e-3)	0.017(2.24e-3)
35.5, 41.5	0.027(2.45e-3)	0.010(2.49e-3)	0.015(2.55e-3)
41.5, 47.5	0.019(2.42e-3)	0.004(1.66e-3)	0.010(1.92e-3)
47.5, 54.5	0.013(2.24e-3)	0.003(1.55e-3)	0.009(2.09e-3)
54.5, 63.5	0.014(2.20e-3)	0.002(1.26e-3)	0.008(2.33e-3)
63.5, 73.5	0.012(2.46e-3)	0.001(4.72e-4)	0.005(1.77e-3)
73.5, 84.5	0.009(2.32e-3)	0	0.003(1.36e-3)
84.5, 97.5	0.006(1.91e-3)	0	0.003(1.57e-3)
97.5, 112.5	0.003(1.24e-3)	0	0.001(8.22e-4)
112.5, 129.5	0.003(1.41e-3)	0	0.000(1.98e-4)
129.5, 149.5	0.003(1.59e-3)	0	0

Table IV.5: Measured values of the observable statistics from SDSS data (Part 4). Table IV.4 continued, for the other three samples.

[71]. The model provides the position, velocity, and stellar mass information of the galaxies. To avoid adopting a specific conversion between stellar mass and magnitude, we assume a constant mass-to-light ratio for all galaxies, such that the ranking of galaxy stellar masses from large to small is equivalent to the ranking of absolute magnitudes from bright to faint.

Since we are only interested in studying the consistency of galaxy statistics between differently constructed mocks, the specific choice of any reasonable galaxy–halo connection model in this step should not affect our results.

2. Light Cone Building

We build light cone mocks (hereafter "**cone mocks**") from the cube mock described in Section IV.C.1, with a routine that is similar to Ref. [94]. Note that we do not populate the halo catalog again, but directly use the cubic galaxy catalog, such that the cube mock and cone mocks have identical underlying cosmologies and galaxy–halo connections. The steps are as follows:

- Choose a random position in the cubic volume to place our virtual observer;
- Periodically repeat the cube mock out to the desired depth;
- Calculate and record the true redshifts $z_{\rm true}$ of the galaxies, accounting for both distance and velocity information;
- Apply redshift measurement uncertainty according to the model described in Appendix A of Ref. [67], and record the resulting redshifts with error, $z_{\rm true,err}$;
- Choose a random direction of observation, convert the galaxy positions into angular coordinates $(\mathrm{ra}, \mathrm{dec})$, and apply the 7461 deg^2-SDSS footprint;
- Again assuming a constant mass-to-light ratio for all galaxies, rank the r-band fluxes of galaxies, which are anti-proportional to the distance squared;
- Keep galaxies with $0.01 < z_{\rm true,err} < 0.18$ and make a flux limit selection based on the flux ranking and the total number of galaxies in the same redshift range in SDSS;
- Assign fiber collision status to galaxies, using the method described in Ref. [64];
- Perform the nearest-neighbor correction for the fiber collision effect, as was done in real data, and record the resulting $z_{\rm obs}$.

We repeat this process with different random seeds to generate 100 cone mocks.

3. Sample Selection

For validation purposes, we define samples in the mock catalogs by galaxy number density, that correspond to the five data samples in the first part of Table IV.1.

For the cube mock, we multiply the number densities listed in Table IV.2 and Table IV.3 by the simulation volume to get the total number of galaxies $N_{\rm cube}$ in each sample. We then select the $N_{\rm cube}$ brightest galaxies from the cube mock, according to the magnitude ranking, to get the luminosity-threshold samples.

For the cone mocks, we apply the same redshift range limits on $z_{\rm obs}$ as those used for selecting data samples[3], and apply the same magnitude ranking thresholds as in the cube. In other words, only galaxies bright enough to be included in the cube samples can be included in the corresponding cone samples.

D. Algorithm Validation

In this section, we compare measurements of galaxy spatial statistics from cube and cone mocks. We obtain measurements on the light cone mocks with the exact same algorithm used for SDSS data, which we described in Section IV.B.2 and Section IV.B.3. We demonstrate that our treatment of observational effects present in light cones ensures that cube mocks yield unbiased estimates of the statistics measured in cone mocks, and hence observational data.

1. Cube Mock Algorithm

a. Statistics

In cube mocks, we apply periodic conditions and adopt the plane-parallel approximation. We place the line of sight along the three axes in turn and average the measurements for each observable.

By construction, the number density $n_{\rm gal}$ of each mock sample is simply that of the corresponding data sample. For cube measurements of $w_{\rm p}(r_{\rm p})$, we use the natural estimator for $\xi(r_{\rm p}, \pi)$,

[3]To avoid exceeding the flux limit, we adopt a conservative redshift cut for the $M_r < -19.5$ at $cz_{\rm max} = 23450\,{\rm km\,s^{-1}}$ in the mocks instead of the $25450\,{\rm km\,s^{-1}}$ limit for data, which does not affect our results qualitatively.

$$\hat{\xi}_{\rm N} = \frac{DD}{RR} - 1, \tag{22}$$

where DD is the normalized galaxy–galaxy pair count, and RR is the normalized random–random pair count. Given the simplicity of the geometry, we use analytic randoms instead of actually drawing random points to reduce the computational cost. Galaxy pairs are counted in $r_{\rm p}$ bins and $\xi(r_{\rm p}, \pi)$ is integrated along the chosen axis out to $\pi_{\rm max}$.

To measure $P(N_{\rm CIC})$, we center a cylinder on every galaxy in the sample, and count the number of companion galaxies that fall in the cylinder, excluding the cylinder center itself. The histogram of the counts is calculated with our specified bins.

b. Jackknife Covariance

To test the consistency between cube and cone measurements, we need to understand the uncertainty of both. Because the cube and cone mocks have the same galaxy population, the only component of the covariance is the jackknife covariance, which provides an estimate of the uncertainty due to finite volume. We randomly select 10 cone mocks for which to measure jackknife covariances, using the exact same procedure with which we measure the data covariance in Section IV.B.3, and take the average of their covariance matrices. The division of the cube into jackknife cells is trivial. We divide the simulation volume into 100 cuboids of $40 \times 40 \times 400 (h^{-1}\,{\rm Mpc})^3$, where the long axis is the same length as the simulation, and lies along the line of sight. We repeat the process for each of the three projections and take the average jackknife covariance. The total covariance matrix that we use for the consistency test is the sum of the cube jackknife covariance and the cone jackknife covariance.

2. Comparison Between Cube and Cone Mocks

We compare the measurements of $n_{\rm gal}$ (Section IV.D.2), $w_{\rm p}(r_{\rm p})$ (Fig. IV.D.2), and $P(N_{\rm CIC})$ (Fig. IV.D.2) between the cube mock and the mean of the cone mocks generated from it. We consider the measurements consistent between the cube and the cones if the deviation is within the total jackknife error. In the figures, we show the cube measurements with error bars, individual cone measurements, and their mean. We find that for all the luminosity samples we consider, and for

all three of our statistics, the measurements are consistent within error. This confirms that with our algorithm, measurements from the simulation cube can be used as unbiased estimates of statistics measured from SDSS-like datasets. In particular, we note that the nearest neighbor correction is sufficient to account for fiber collision, for the statistics that we consider. We will therefore proceed to make measurements on the SDSS data using the light cone algorithm in Section IV.B, and fit the data using the simulation cube in Section IV.E.

E. MCMC Fit

In this section, we fit our measurements of the SDSS data using the halo occupation distribution (HOD) and decorated halo occupation distribution (dHOD) models, which we described in Section III.B.2.

1. Covariance Matrix

In fitting data measurement with mock measurements in the simulation, we need to account for both the uncertainty in the observation and the uncertainty in the mock estimation. Therefore, the covariance matrix that we use for the fitting is the sum of the data component and the theory component. In Fig. IV.8, we show the normalized covariance matrix for the $M_r < -20.0$ sample as an example. The two panels on the left are the jackknife covariances from the SDSS data and the SMDPL mock separately, and the rightmost panel is the sum of both. In plotting the matrices, they are normalized by the diagonal elements, such that all the diagonal elements are 1 by construction, and the off-diagonal elements range between -1 and 1, which are color coded in the figure. The data covariance is noisier due to its smaller volume than the simulation covariance. In both components, $n_{\rm gal}$ is positively correlated with $w_{\rm p}$ at all scales except the smallest, and also positively correlated with higher counts of companions in cylinders. The is consistent with our expectation that the clustering in denser regions are stronger. The correlations between $w_{\rm p}(r_{\rm p})$ values across different scales are positive within the range we investigate, and their correlations with $P(N_{\rm CIC})$ approximately follow those of $n_{\rm gal}$. On the other hand, $P(N_{\rm CIC})$ values at the higher

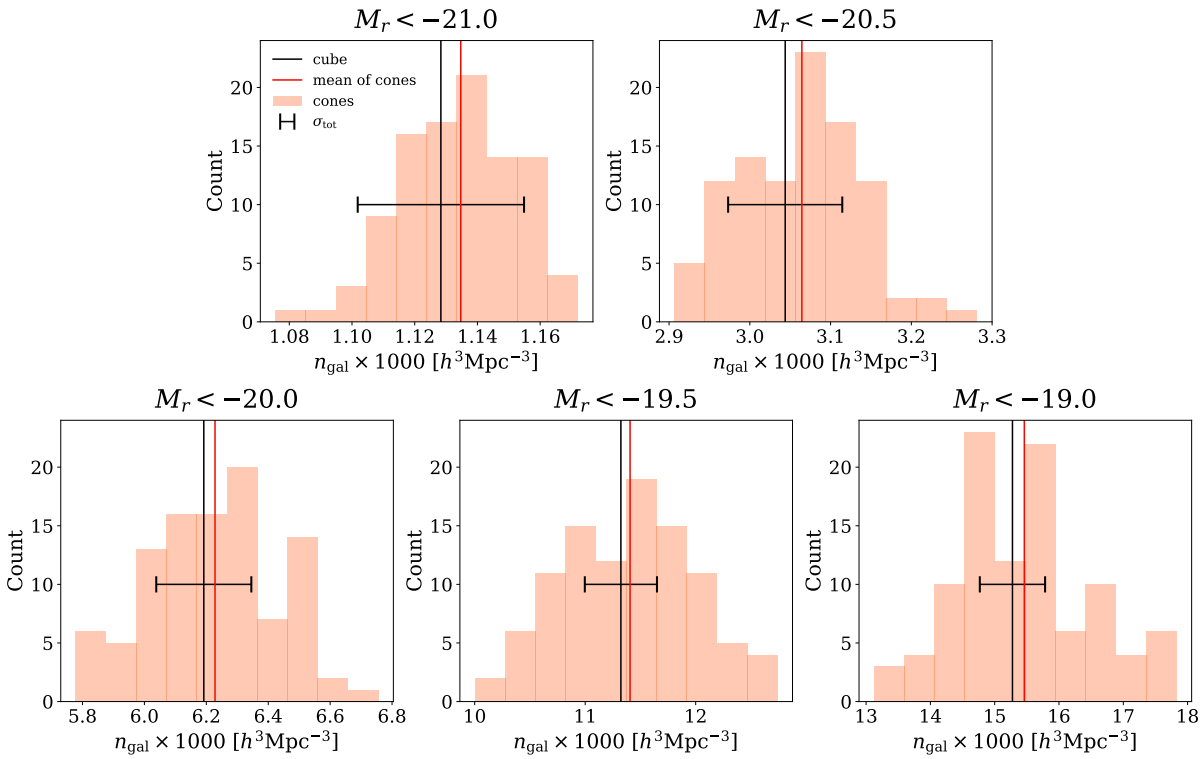


Figure IV.5: Comparison between the cube mock and the cone mocks of the galaxy number density $n_{\rm gal}$. Each panel compares a different luminosity sample. The underlying cube value is shown as a vertical black line. The horizontal black error bars are the total jackknife error from the cube and the cones. The values measured from the different light cone mocks are plotted as a histogram, and the mean is marked by the vertical red line. This figure shows that the measurement of the galaxy number density is consistent between our cube and cone mocks for all five samples.

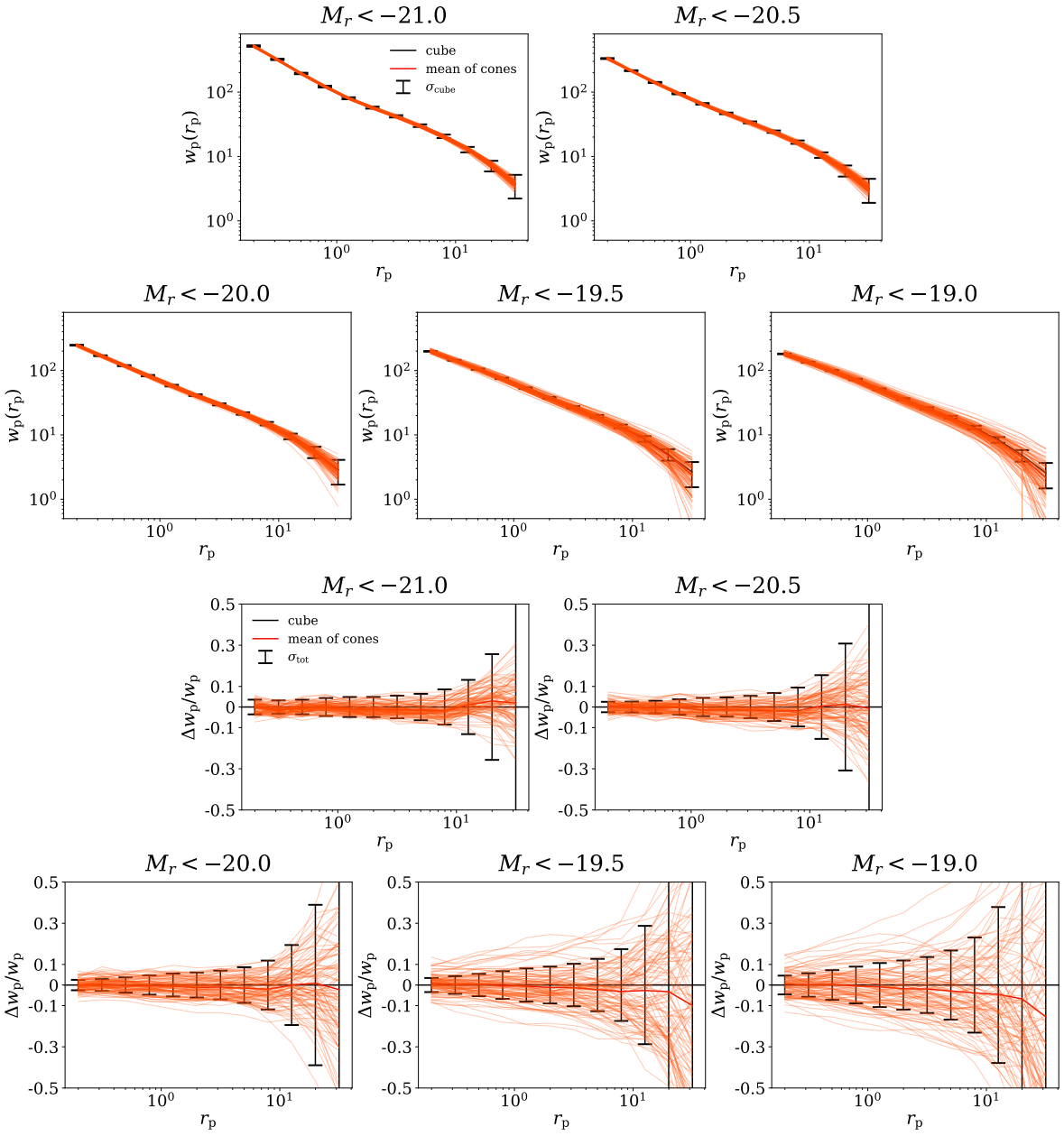


Figure IV.6: Comparison between the cube mock and the cone mocks of the projected two-point function $w_{\rm p}$. The top five panels show a comparison of the measurements of $w_{\rm p}(r_{\rm p})$, and the bottom five panels show the fractional deviation between the cube and the cones. The luminosity sample corresponding to each panel is labeled at the top.

Figure IV.6: *(cont.)* In the top panels, the cube $w_{\rm p}$ is plotted as solid black lines, and error bars show the jackknife error of the cube. Individual cones are plotted as thin lines in the background, and the mean of the cones is shown by the solid red line. In the bottom panels, the y-axis is the fractional deviation of the cones from the cube measurements. The horizontal black line marks zero deviation, and the error bars show the total jackknife error from the cube and the cones. Measurements for all the samples are consistent within error.

and lower ends are anti-correlated by construction, as the probabilities sum up to unity.

2. Constraints on Model

We use the Markov Chain Monte Carlo (MCMC) method to infer the parameter constraints from the measurements. For each SDSS galaxy sample, we consider four fitting cases: (1) the standard HOD inferred from $n_{\rm gal} + w_{\rm p}(r_{\rm p})$; (2) the standard HOD inferred from $n_{\rm gal} + w_{\rm p}(r_{\rm p}) + P(N_{\rm CIC})$; (3) the decorated HOD inferred from $n_{\rm gal} + w_{\rm p}(r_{\rm p})$; and (4) the decorated HOD inferred from $n_{\rm gal} + w_{\rm p}(r_{\rm p}) + P(N_{\rm CIC})$.

For each case, we use the full covariance matrix of the corresponding observables, including the cross covariances between different statistics, described in the previous subsection. We assume likelihood $\mathcal{L} \propto e^{-\chi^2/2}$, with

$$\chi^2 = \sum_{i,j} \Delta f_i [\mathbf{C}^{-1}]_{ij} \Delta f_j, \tag{23}$$

where f_i and f_j are the ith and jth element of the joint statistic vector $\mathbf{f}$, and $\mathbf{C}$ is the full covariance. Because the three types of statistics, $n_{\rm gal}$, $w_{\rm p}(r_{\rm p})$, and $P(N_{\rm CIC})$, have a wide range of values, we perform pseudo-inversion of the covariance, which we described in Section III.B.5. We assume uniform priors within certain intervals on the parameters, listed in Table IV.6.

With the above likelihood and prior, we use the `emcee` [56] package, which is an implementation of MCMC to sample from the posterior distribution in the parameter space. Due to the long computational time required for MCMC chains to converge, we only present results for the $M_r < -20.0$ sample in this dissertation.

In Fig. IV.9 and Fig. IV.10, we show the corner plots for the four fitting cases of the $M_r < -20.0$

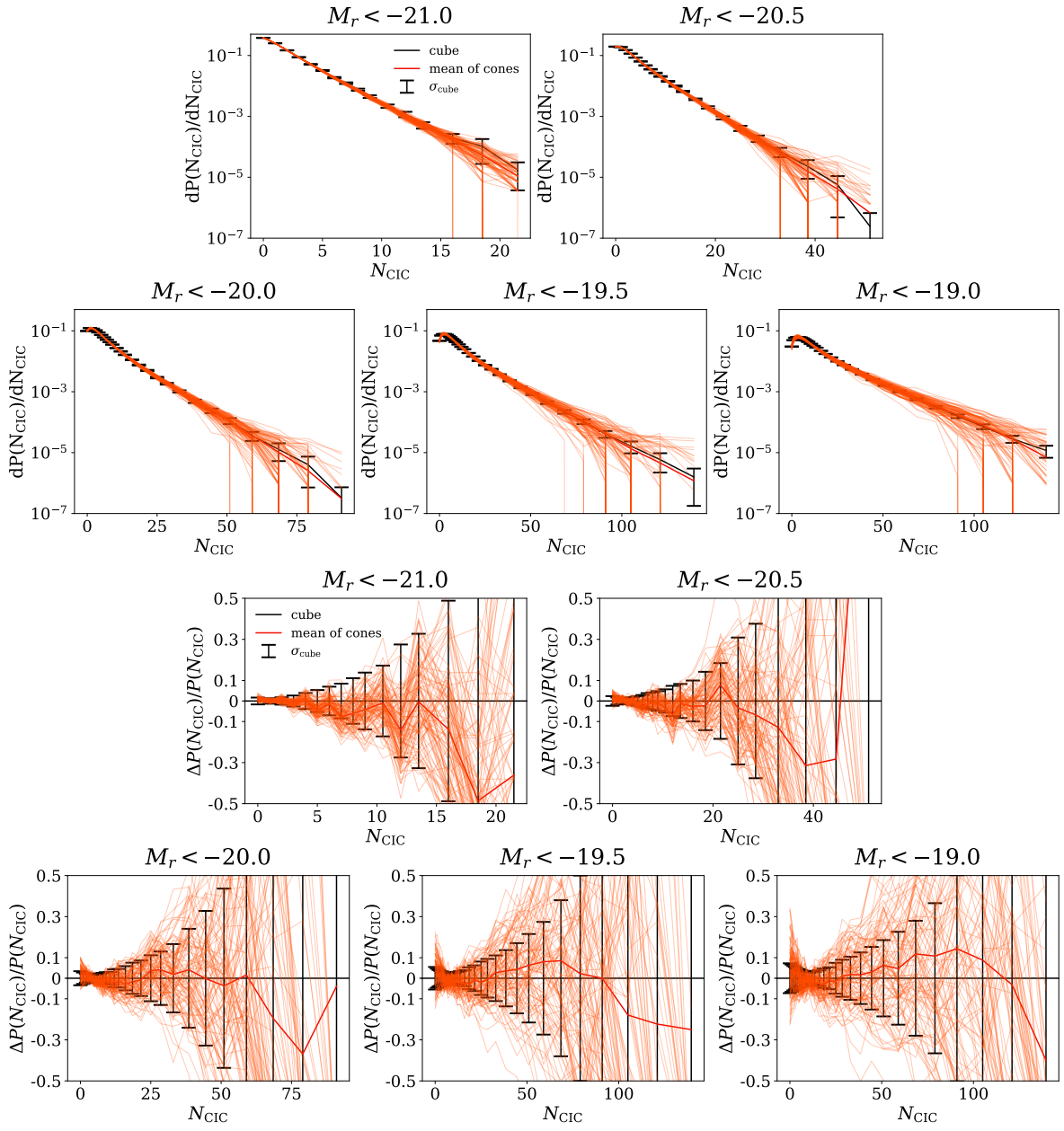


Figure IV.7: Comparison between the cube mock and the cone mocks of the counts-in-cylinders statistic $P(N_{\rm CIC})$. Same as Fig. IV.D.2, but for counts-in-cylinders instead of $w_{\rm p}$. In the top panels the probability of counts in each bin is normalized by the bin width. Note that the range of the x-axis is different for each sample, though y-axis ranges are the same for all. Measurements are again consistent for all the samples.

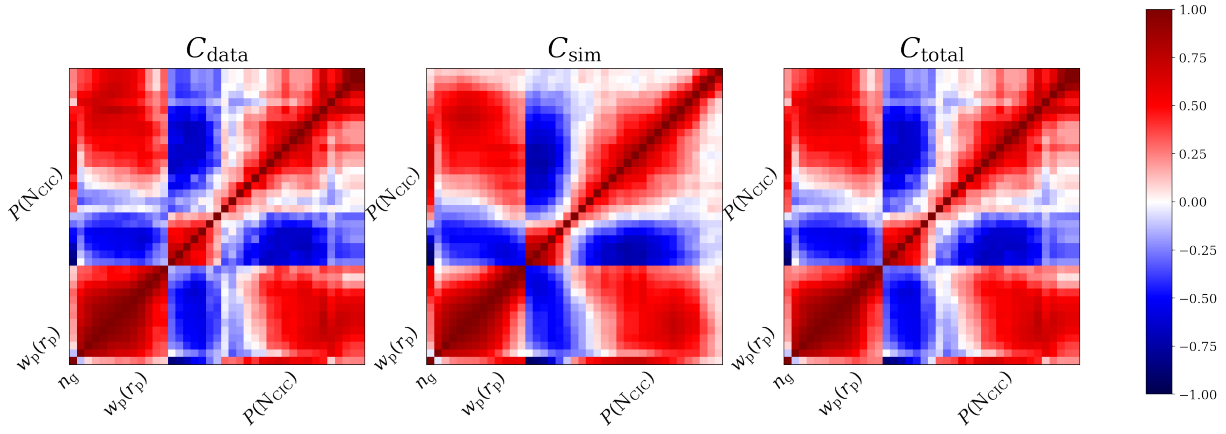


Figure IV.8: Normalized covariance matrices (i.e., correlation matrices) for the $M_r < -20.0$ sample. The left panel is the jackknife covariance of the SDSS data sample, the middle panel that of the SMDPL simulation, and the right panel shows the sum of the two. The diagonal elements of the correlation matrices are 1 by construction.

Parameter	Prior interval
α	[0, 1.5]
$\log M_1$	[11.5, 15.0]
$\sigma_{\log M}$	[0.02, 1.5]
$\log M_0$	[9.0, 14.0]
$\log M_{\rm min}$	[11.0, 14.0]
$A_{\rm cen}$	[-1, 1]
$A_{\rm sat}$	[-1, 1]

Table IV.6: Prior intervals adopted for MCMC fits. The $A_{\rm cen}$ and $A_{\rm sat}$ priors only apply to the decorated HOD fits.

sample listed above. The corner plots show contours of parameter constraints in two-dimensional projections, as well as one-dimensional distributions of each parameter. Comparing between the constraints from $n_{\rm gal} + w_{\rm p}(r_{\rm p})$ and $n_{\rm gal} + w_{\rm p}(r_{\rm p}) + P(N_{\rm CIC})$, we see that the latter is indeed much more constraining. For example, our preference of positive $A_{\rm cen}$ values is much stronger with $P(N_{\rm CIC})$ included, improving from a marginal 2σ detection of non-zero central galaxy assembly bias to over 5σ.

We quantify the consistency between the two sets of parameters using the method described in Section 2.2 of Ref. [90]. We find a marginal (0.88σ) tension between $n_{\rm gal} + w_{\rm p}(r_{\rm p})$ and $n_{\rm gal} + w_{\rm p}(r_{\rm p}) + P(N_{\rm CIC})$ in the standard HOD case, and no tension (0.07σ) in the decorated HOD case. We also compare the goodness of fit between the two models, through the Akaike Information Criterion (AIC) [3]

$$\mathrm{AIC} = 2k - 2\ln\mathcal{L}_{\rm max}, \tag{24}$$

and the Bayesian Information Criterion (BIC) [161]

$$\mathrm{BIC} = k\ln n - 2\ln\mathcal{L}_{\rm max}, \tag{25}$$

where k is the number of parameters in the model, n is the number of data points involved in the fit, and $\mathcal{L}_{\rm max}$ the maximum likelihood, in our case $-2\ln\mathcal{L}_{\rm max} = \chi^2_{\rm min}$. The model with smaller AIC or BIC values is considered superior. The information criteria, along with the posterior 1σ intervals and minimum χ^2 values normalized by the number of degrees of freedom, are listed in Table IV.7. We find that $n_{\rm gal} + w_{\rm p}(r_{\rm p})$ does not prefer one model over the other, whereas $n_{\rm gal} + w_{\rm p}(r_{\rm p}) + P(N_{\rm CIC})$ shows a strong preference of the decorated HOD model, with ΔAIC = 14.8 and ΔBIC = 10.5.

F. Discussion and Conclusions

In this chapter, we have confirmed that the counts-in-cylinders statistic complements the canonical observational measurement of the galaxy two-point correlation function, in constraining the galaxy–halo connection. Using the observable set that includes counts-in-cylinders, we have found that for the $M_r < -20.0$ galaxy sample that we study, the standard halo occupation distribution

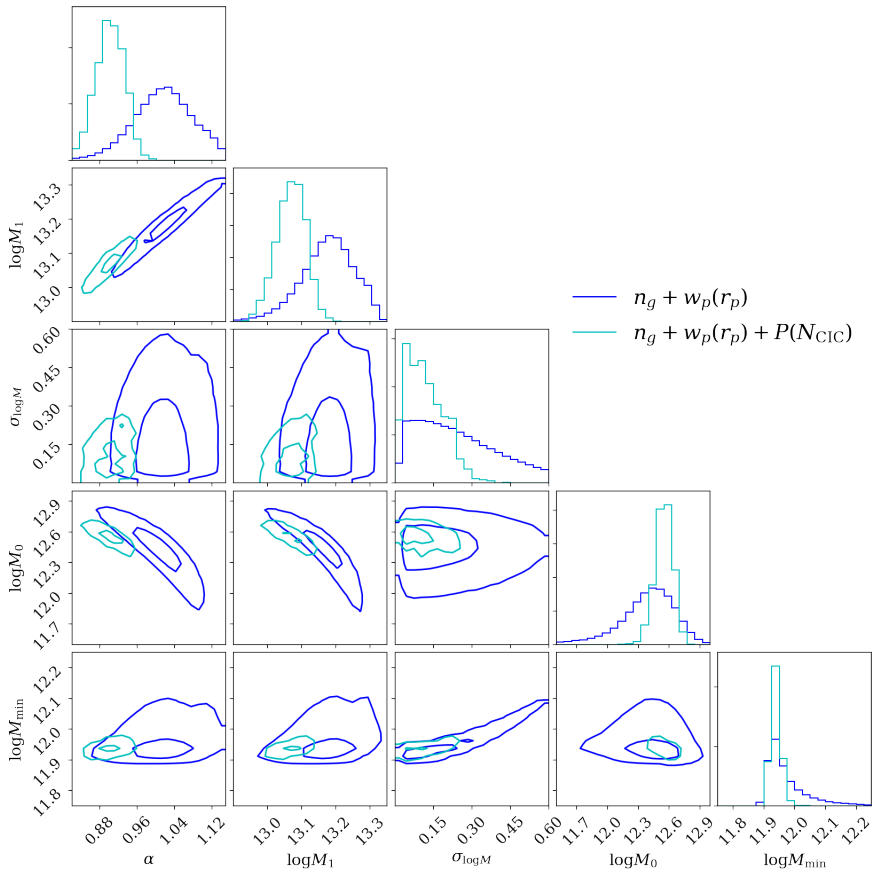


Figure IV.9: Constraints on the HOD parameters from data. The contours show the 1σ and 2σ constraints on the parameter space in two-dimensional planes, and the one-dimensional histograms are marginalized for each individual parameter. The blue contours show the constraints from $n_{\rm gal}$ and $w_{\rm p}(r_{\rm p})$, and the cyan contours show the constraints from $n_{\rm gal}$, $w_{\rm p}(r_{\rm p})$, and $P(N_{\rm CIC})$ combined.

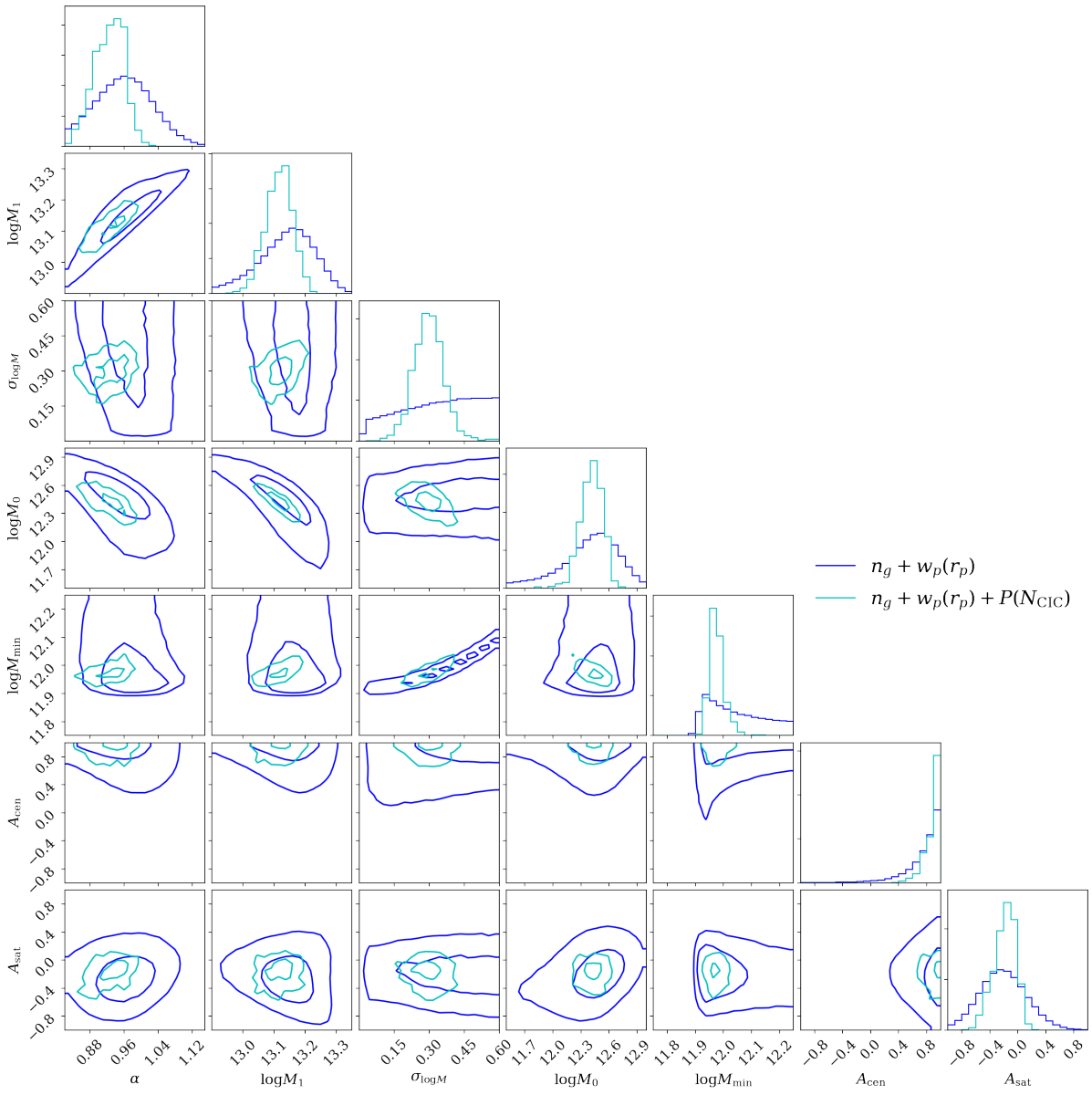


Figure IV.10: Constraints on the decorated HOD parameters from data. Same as Fig. IV.9, but for the decorated HOD model.

	Standard HOD, $n_{\rm gal} + w_{\rm p}(r_{\rm p})$	Decorated HOD, $n_{\rm gal} + w_{\rm p}(r_{\rm p})$
α	$1.018^{+0.063}_{-0.064}$	$0.955^{+0.070}_{-0.080}$
$\log M_1$	$13.180^{+0.073}_{-0.081}$	$13.144^{+0.082}_{-0.110}$
$\sigma_{\log M}$	$0.238^{+0.243}_{-0.151}$	$0.690^{+0.450}_{-0.397}$
$\log M_0$	$12.393^{+0.234}_{-0.398}$	$12.382^{+0.288}_{-0.681}$
$\log M_{\rm min}$	$11.953^{+0.091}_{0.032}$	$12.167^{+0.402}_{-0.204}$
$A_{\rm cen}$	/	$0.826^{+0.130}_{-0.301}$
$A_{\rm sat}$	/	$-0.211^{+0.340}_{-0.320}$
χ^2	11.91	7.08
χ^2/dof	1.49	1.18
AIC	21.91	21.08
BIC	24.74	25.04
	Standard HOD, $n_{\rm gal} + w_{\rm p}(r_{\rm p}) + P(N_{\rm CIC})$	Decorated HOD, $n_{\rm gal} + w_{\rm p}(r_{\rm p}) + P(N_{\rm CIC})$
α	$0.902^{+0.030}_{-0.032}$	$0.927^{+0.033}_{-0.038}$
$\log M_1$	$13.071^{+0.041}_{-0.046}$	$13.118^{+0.043}_{-0.049}$
$\sigma_{\log M}$	$0.109^{+0.093}_{-0.066}$	$0.298^{+0.061}_{-0.062}$
$\log M_0$	$12.552^{+0.085}_{-0.088}$	$12.420^{+0.112}_{-0.123}$
$\log M_{\rm min}$	$11.937^{+0.018}_{-0.014}$	$11.974^{+0.027}_{-0.020}$
$A_{\rm cen}$	/	$0.917^{+0.064}_{-0.145}$
$A_{\rm sat}$	/	$-0.167^{+0.140}_{-0.185}$
χ^2	52.49	33.69
χ^2/dof	1.54	1.05
AIC	62.49	47.69
BIC	70.81	59.34

Table IV.7: Derived model parameters and goodness of fit measures for the $M_r < -20.0$ galaxy sample.

(HOD) is not as good a description of the data as the decorated HOD, which the original combination of galaxy number density and two-point correlation function fails to distinguish. We make statistically significant detection of galaxy assembly bias in this sample.

The implication of our results is twofold. Firstly, we have demonstrated the constraining power of galaxy count statistics. Count statistics incorporate the higher-order information of the field, and are easy to measure from observational data. We advocate the use of these statistics in combination with the commonly used two-point measurements, which has the potential not only to improve constraints on the galaxy–halo connection, but also on cosmological models and other aspects of galaxy physics. Secondly, we have detected galaxy assembly bias, which shows that the first-order assumption that galaxy occupation only depends on halo mass is inaccurate. This effect needs to be properly understood and modeled. Traditional models that adopt the mass-only ansatz will not suffice in the new era of precision cosmology, and developing new models that account for galaxy assembly bias is a timely task, which is attracting increasing attention (see, for example, [219, 37]). We aim to explore other applications of galaxy count statistics, and further look into the physical origin of galaxy assembly bias in future work.

V. Conclusions and Outlook

Knowledge of the connection between observable galaxies and the underlying large-scale structure facilitates the comparison between predictions from cosmological models and observations of the real Universe. This comparison is pivotal in the collective effort to answer some of the most prominent questions in astronomy today — What is dark energy? How does it impact the evolution of the Universe? How did galaxies form and evolve to shape today's visible Universe? Is our Galaxy unique in any way? While these all depend upon the reliability of our theoretical framework, the uncertainty in the galaxy–halo connection has become a pressing concern in the era of precision cosmology.

In this thesis, we have addressed the galaxy–halo connection, from theoretical, computational, and observational perspectives. In Chapter II, we investigate how the halo concentration parameter emerges from the halo mass assembly history, and distinguish between contributions from different modes of halo growth. Central to the methodology of this work is the examination of dynamical processes in dynamical timescales. This methodology can be adapted to any other halo property of interest, and a study of the co-evolution of multiple halo properties can provide deep insight into how halos are shaped by physical events. Our conclusion that merger events are of crucial importance in halo structure evolution, together with the fact that mergers strongly depend on the environment density, suggests that they may be a major factor that causes halos to have secondary biases — dependences of halo clustering on internal properties other than mass: halos with different clustering strengths have different internal properties because they have different merger rates. Galaxy mergers often accompany halo mergers, and similar mechanisms may well lead to a dependence of galaxy abundance and properties on the environment. If we include galaxy properties, which are available in hydrodynamical simulations, in addition to halo properties in the study of co-evolution, we may be able to uncover the origin of any possible galaxy assembly bias. This is also one of the subjects I plan to pursue in the near future.

Chapters III and IV approach the topic from a different perspective. While a number of theoretical and numerical works suggest it is likely that both halo clustering and galaxy clustering are correlated with secondary halo properties, the effect has thus far eluded detection in observation.

There are two possible avenues for obtaining more conclusive observational evidence: either to use larger datasets and reduce error, or to exploit existing data using more efficient statistical probes. The former is very promising, especially given the great effort that is going into new generations of surveys, which will provide enormous amounts of data, and boost signal-to-noise ratios. However, this does not solve the problem of degeneracies in the model parameter space, for they are intrinsic to the statistics, and determined by the manner of their dependence on the model. The alternative approach, which is to develop new statistics, can be more helpful in this respect. Because of their different dependences on the model, complementary statistic sets break degeneracies and yield tighter constraints on the parameters. Our work in these two chapters makes use of this idea, and reintroduces the use of count statistics to complement the canonical probes. Using these statistics, we make definitive detection of galaxy assembly bias, which necessitates explorations of its physical origin and better models that incorporate the effect.

These projects lead to advanced understanding of the physical processes that drive the formation and evolution of galaxies and halos on nonlinear scales, and more accurate models of galaxy clustering in the Universe that ensure correct interpretation of data for cosmological analyses.

Appendix A Concentration–Mass Relation

In Fig. A.1, we show the concentration–mass relation of the random sample described in Section II.B.2.b, and the dependence of the relation on half-mass scales. The random sample is divided into 5 quintiles based on the mass normalized marks of the half-mass scales. In the figure, we show the mean concentration relation along with the standard deviation in the relation for each subsample. The standard error of the mean is naturally much smaller.

From the figure we observe that, in general, the concentration–mass relation depends monotonically on the half-mass scale, again in consistence with previous findings that earlier forming haloes tend to be more concentrated. The scatter in the relation is larger for later forming haloes except in a few cases, which can be explained by the fact that they are more likely to have had recent mergers. This figure complements our findings in Fig. II.6.

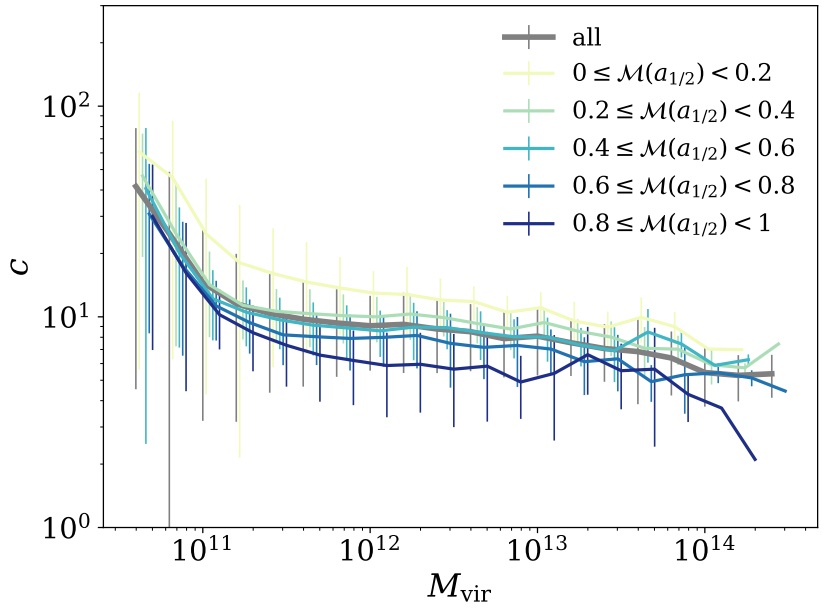


Figure A.1: Concentration–mass relation for halo samples. The solid lines show the mean relations, while the error bars show the standard deviation of the relation within each sample. The thicker gray line shows the concentration–mass relation for all the haloes in the random sample described in Section II.B.2.b, while the other lines show subsamples with different mark values of $a_{1/2}$, as are labeled in the legend.

Appendix B Derivative Fitting

In this appendix, we describe our approach to obtain the derivative matrix $\partial\mathbf{f}/\partial\mathbf{p}$.

We begin from the fiducial parameter set specified in Table III.1, and add perturbations to each dHOD parameter (except M_0, which we do not vary in our analysis) in turn, keeping all other parameters fixed, to examine the dependence of observables on each individual parameter (i.e., the set of **f**–**p** relations). The process is non-trivial because we construct our observables from mock catalogs based on the population of a simulation of finite size with galaxies drawn from the (d)HOD. The fact that we construct our observables using mock catalogs has the advantage of accuracy compared to analytic approximation methods, but it also has the unfortunate consequence that individual observables can exhibit non-negligible stochasticity from one mock catalog realization to another. In order to mitigate the impact of this variability, rather than taking $\Delta\mathbf{f}/\Delta\mathbf{p}$ at a single perturbed point, we consider a series of perturbed models in the neighborhood of the fiducial value, along each dimension of the parameter space. We then fit our set of **f**–**p** relations for a slope, as an estimate of the partial derivative in the neighborhood of the fiducial point. We will elaborate on these procedures in the remainder of this appendix.

As we have discussed in Section III.B.2.b, in the `Halotools` implementation of the (d)HOD model, the galaxy occupation of each halo is randomly drawn from the probability distribution function determined by the properties of the halo. The mean occupation varies with the (d)HOD parameters, whereas any particular realization of the mock galaxy catalog is also dependent on the sequence of random numbers used in this process. To generate the mock catalogs that are suitable for Fisher analysis, we use a fixed random seed for each random variable in the (d)HOD model. The values of random seeds are set independently of the (d)HOD parameter values. In this fashion, the number of galaxies in each halo will always have the same random deviate each time the halo is populated from the underlying (d)HOD model, regardless of the parameter values. Had we not implemented this, the **f**–**p** relations we measure would be severely impacted by the significantly greater stochasticity introduced by the random process. It would then be required to construct a very large number of mock catalogs to marginalize over this stochasticity. By employing a common seed, the differences between these catalogs become primarily attributable to parameter differences

and not statistical fluctuations from catalog to catalog that could be induced by finite sampling of the (d)HOD models. This exercise minimizes the variation of the observables attributable to stochasticity; however, **f** is generally not a smooth function of **p**, due to the intricate nature of the dependence of the observables on the galaxy distribution.

Determining the **f**–**p** relations from a single random number seed is not sufficient because the **f**–**p** relation has a small dependence upon the random number seed. Consequently, the values of the resulting derivatives will vary slightly with different random seeds. For this reason, we repeat this entire process for a large number (or order ~ 100) of different random number seeds and take the trimmed mean of **f**, averaging only the central 68% values. In this manner, we construct smooth **f**–**p** relations that do not dependent on the choice of the random seed. In the example shown in Fig. B.1, we assess the relation between the projected two-point correlation function $w_{\rm p}$ at $r_{\rm p} = 1.74h^{-1}$ Mpc and the central galaxy assembly bias parameter $A_{\rm cen}$. Thin colored lines correspond to different random number seeds, and for clarity we only plot results for a subset of all random seeds used. The thicker black line shows the trimmed mean from the central 68% of the fixed-seed **f**–**p** data points, which serves as our estimate of the observable–parameter relation. This procedure provides us with a set of **f**–**p** relations, one for each observable–parameter pair, that we fit as described below.

We use the `R` [149] package `locfit` [107] to fit local linear derivatives to the **f**–**p** relation that we have obtained. We choose the degree of local polynomials to be two, which captures the shape of the curve without excessive overfitting. The `locfit` package provides 5 commonly used weighting kernels. We have confirmed that different kernels yield similar results. We present results obtained with the default `tricube` kernel. Because some of the **f**–**p** relations are strongly non-linear, the smoothing scale for the local fit needs to be chosen with care. We choose the smoothing scale of fitting following the principle that the range of the parameter considered for the local derivative fit should be comparable with the posterior 1σ constraint for each observable combination respectively. Qualitatively, this is motivated by the fact that the 1σ constraints defines what it means to be in the "neighborhood" of the fiducial point. To this criterion, we add two additional restrictions:

I. For each **f**-**p** relation, we obtain the optimal smoothing parameter from generalized cross validation (GCV), using the `loess.as` function in the `R` package `fANCOVA` [200]. We use this

smoothing scale as a lower limit, in order to avoid fitting numerical noise.

II. We apply another lower limit of smoothing scale for each parameter in each threshold, based on the physical interpretation of **f**-**p** relations, as some of the unphysical effects from the mock are not recognized by GCV. We have tested that our conclusions are not sensitive to this choice within a reasonable range.

In the example of Fig. B.1, we fit a local linear derivative to the relation at the fiducial parameter set, marked by the vertical gray dashed line, i.e., $[\partial w_{\rm p}(r_{\rm p} = 1.74h^{-1}\,{\rm Mpc})/\partial A_{\rm cen}]|_{A_{\rm cen}=0}$. The smoothing scale is shown as a gray band that is symmetric around the fiducial parameter. The local linear fit is shown by the solid red line. The fitted slopes are the derivatives we use in our forecasts.

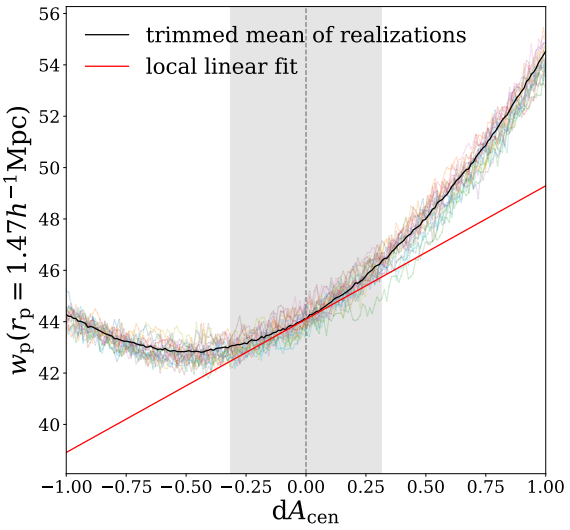


Figure B.1: An example of our function–parameter relations and its fitted derivative. In this figure, we show an example of our **f**–**p** relations and its fitted derivative. In this example, we study the observable $w_{\rm p}(r_{\rm p} = 1.74h^{-1}\,{\rm Mpc})$ against the perturbation in $A_{\rm cen}$. Each thin colored line is the dependence of the observable on the parameter, obtained with a different random number seed. The solid black line shows the trimmed mean of these fixed-seed **f**-**p** data points. The fiducial parameter is marked by the vertical gray dashed line (in this case at $A_{\rm cen} = 0$, so that $dA_{\rm cen} = A_{\rm cen}$). The smoothing scale adopted is shown as a vertical gray band, and the solid red line is the local linear fit.)

Appendix C Forecast Constraints on HOD Parameters

In this appendix, we present a comprehensive list of our Fisher forecast results for constraints on dHOD parameters.

In Fig. C.1 and Fig. C.2, we show bar plots of constraints on the decorated HOD parameters besides $A_{\rm cen}$ and $A_{\rm sat}$ (shown in Figure III.4), excluding $\log M_0$ which is poorly constrained. These include α, $\log M_1$, $\log M_{\rm min}$, and $\sigma_{\log M}$, which are also the original parameters of the standard HOD.

We also list the posterior constraint values in Tables C, C, C, C, and C, for the 6 parameters that we allow to vary and for all 5 of the luminosity samples we study. The constraints we include are from all the individual observables, all the possible combinations of two observables, and the combination of all 6 observables.

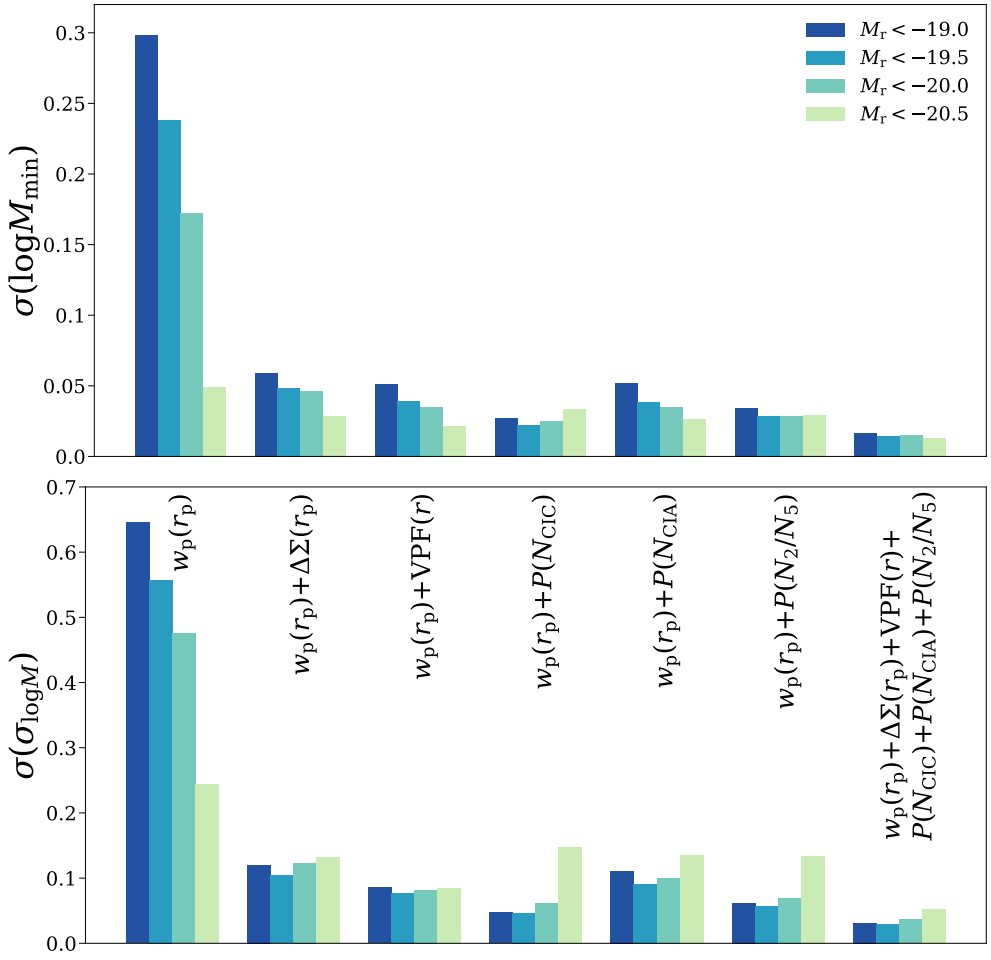


Figure C.1: The same as Figure III.4, but showing the constraints for the parameters α and $\log M_1$.

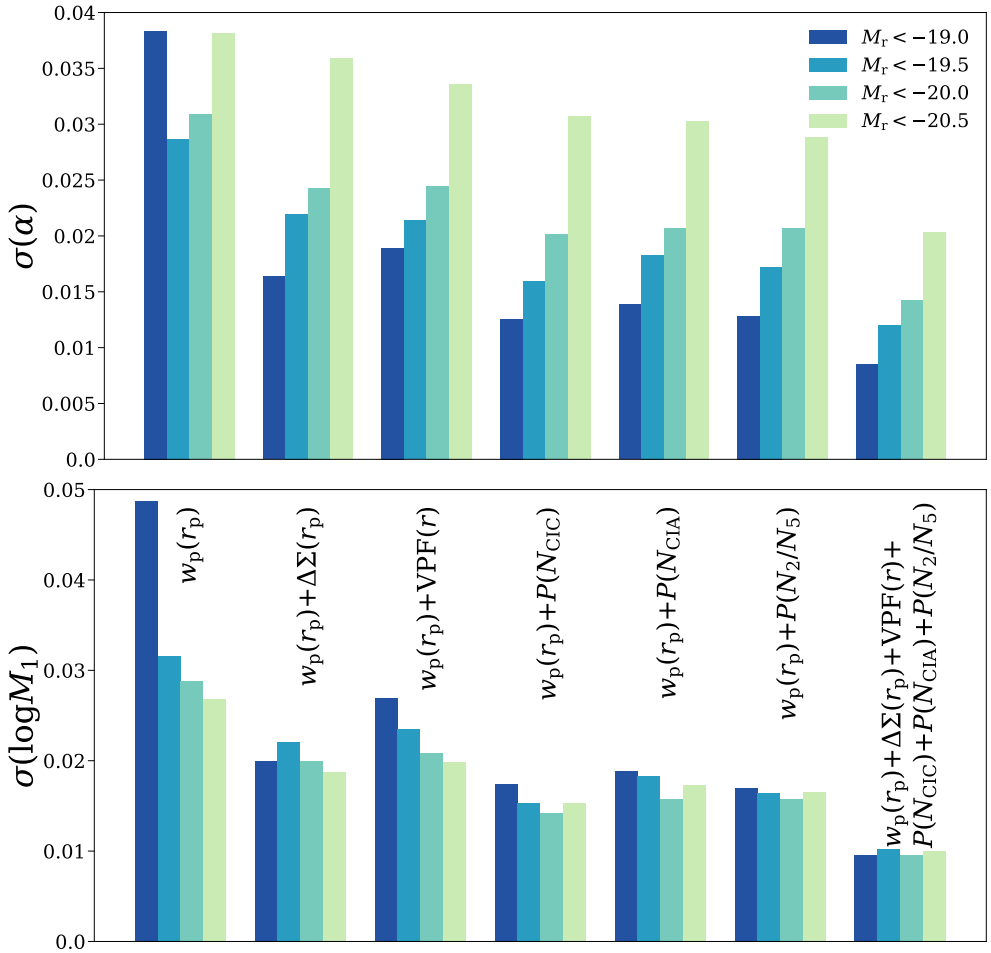


Figure C.2: The same as Figure III.4, but showing the constraints for the parameters $\log M_{\rm min}$ and $\sigma_{\log M}$.

Observable combination	$\sigma(\log M_{\rm min})$	$\sigma(\sigma_{\log M})$	$\sigma(\alpha)$	$\sigma(\log M_1)$	$\sigma(A_{\rm cen})$	$\sigma(A_{\rm sat})$
$w_{\rm p}(r_{\rm p})$	0.30	0.645	0.038	0.05	0.71	0.34
$\Delta\Sigma(r_{\rm p})$	0.10	0.157	0.066	0.09	0.79	0.32
VPF(r)	0.22	0.129	0.689	0.37	0.63	3.99
$P(N_{\rm CIC})$	0.18	0.179	0.091	0.18	1.34	0.20
$P(N_{\rm CIA})$	0.06	0.151	0.049	0.07	0.46	0.36
$P(N_2/N_5)$	0.05	0.073	0.040	0.06	0.57	0.21
$w_{\rm p}(r_{\rm p}) + \Delta\Sigma(r_{\rm p})$	0.06	0.121	0.016	0.02	0.41	0.20
$w_{\rm p}(r_{\rm p}) +$ VPF(r)	0.05	0.086	0.019	0.03	0.31	0.25
$w_{\rm p}(r_{\rm p}) + P(N_{\rm CIC})$	0.03	0.048	0.013	0.02	0.15	0.15
$w_{\rm p}(r_{\rm p}) + P(N_{\rm CIA})$	0.05	0.111	0.014	0.02	0.16	0.19
$w_{\rm p}(r_{\rm p}) + P(N_2/N_5)$	0.03	0.062	0.013	0.02	0.27	0.15
$\Delta\Sigma(r_{\rm p}) +$ VPF(r)	0.06	0.084	0.036	0.04	0.31	0.19
$\Delta\Sigma(r_{\rm p}) + P(N_{\rm CIC})$	0.04	0.054	0.022	0.03	0.19	0.12
$\Delta\Sigma(r_{\rm p}) + P(N_{\rm CIA})$	0.05	0.082	0.030	0.03	0.17	0.19
$\Delta\Sigma(r_{\rm p}) + P(N_2/N_5)$	0.04	0.064	0.022	0.03	0.29	0.14
VPF(r) + $P(N_{\rm CIC})$	0.05	0.057	0.030	0.06	0.43	0.19
VPF(r) + $P(N_{\rm CIA})$	0.05	0.061	0.046	0.04	0.23	0.35
VPF(r) + $P(N_2/N_5)$	0.03	0.050	0.030	0.04	0.38	0.20
$P(N_{\rm CIC}) + P(N_{\rm CIA})$	0.04	0.055	0.017	0.03	0.24	0.14
$P(N_{\rm CIC}) + P(N_2/N_5)$	0.04	0.054	0.019	0.03	0.23	0.13
$P(N_{\rm CIA}) + P(N_2/N_5)$	0.03	0.056	0.023	0.03	0.23	0.17
All six	0.02	0.031	0.009	0.01	0.10	0.07

Table C.1: Halo occupation distribution (HOD) parameter constraints for the $M_r < -19.0$ sample.

Observable combination	$\sigma(\log M_{\rm min})$	$\sigma(\sigma_{\log M})$	$\sigma(\alpha)$	$\sigma(\log M_1)$	$\sigma(A_{\rm cen})$	$\sigma(A_{\rm sat})$
$w_{\rm p}(r_{\rm p})$	0.24	0.558	0.029	0.03	0.63	0.35
$\Delta\Sigma(r_{\rm p})$	0.07	0.138	0.080	0.08	0.76	0.33
VPF(r)	0.13	0.179	0.657	0.30	0.62	3.10
$P(N_{\rm CIC})$	0.09	0.119	0.088	0.14	1.06	0.16
$P(N_{\rm CIA})$	0.05	0.141	0.075	0.11	0.72	0.37
$P(N_2/N_5)$	0.04	0.076	0.052	0.05	0.58	0.28
$w_{\rm p}(r_{\rm p}) + \Delta\Sigma(r_{\rm p})$	0.05	0.104	0.022	0.02	0.40	0.23
$w_{\rm p}(r_{\rm p}) +$ VPF(r)	0.04	0.077	0.021	0.02	0.32	0.27
$w_{\rm p}(r_{\rm p}) + P(N_{\rm CIC})$	0.02	0.046	0.016	0.02	0.17	0.12
$w_{\rm p}(r_{\rm p}) + P(N_{\rm CIA})$	0.04	0.091	0.018	0.02	0.18	0.22
$w_{\rm p}(r_{\rm p}) + P(N_2/N_5)$	0.03	0.057	0.017	0.02	0.31	0.17
$\Delta\Sigma(r_{\rm p}) +$ VPF(r)	0.05	0.079	0.047	0.04	0.30	0.21
$\Delta\Sigma(r_{\rm p}) + P(N_{\rm CIC})$	0.03	0.050	0.026	0.03	0.20	0.12
$\Delta\Sigma(r_{\rm p}) + P(N_{\rm CIA})$	0.04	0.071	0.038	0.03	0.21	0.21
$\Delta\Sigma(r_{\rm p}) + P(N_2/N_5)$	0.03	0.060	0.028	0.03	0.31	0.16
VPF(r) + $P(N_{\rm CIC})$	0.04	0.052	0.039	0.06	0.44	0.14
VPF(r) + $P(N_{\rm CIA})$	0.04	0.058	0.062	0.04	0.26	0.35
VPF(r) + $P(N_2/N_5)$	0.03	0.052	0.038	0.04	0.34	0.27
$P(N_{\rm CIC}) + P(N_{\rm CIA})$	0.03	0.052	0.023	0.03	0.27	0.12
$P(N_{\rm CIC}) + P(N_2/N_5)$	0.03	0.052	0.025	0.03	0.26	0.13
$P(N_{\rm CIA}) + P(N_2/N_5)$	0.03	0.054	0.032	0.03	0.26	0.22
All six	0.01	0.030	0.012	0.01	0.11	0.07

Table C.2: Halo occupation distribution (HOD) parameter constraints for the $M_r < -19.5$ sample.

Observable combination	$\sigma(\log M_{\rm min})$	$\sigma(\sigma_{\log M})$	$\sigma(\alpha)$	$\sigma(\log M_1)$	$\sigma(A_{\rm cen})$	$\sigma(A_{\rm sat})$
$w_{\rm p}(r_{\rm p})$	0.17	0.475	0.031	0.03	0.78	0.45
$\Delta\Sigma(r_{\rm p})$	0.06	0.149	0.081	0.07	0.86	0.43
VPF(r)	0.10	0.155	0.783	0.21	0.56	2.73
$P(N_{\rm CIC})$	0.09	0.157	0.081	0.11	1.32	0.16
$P(N_{\rm CIA})$	0.04	0.137	0.088	0.11	0.91	0.33
$P(N_2/N_5)$	0.05	0.112	0.064	0.05	0.80	0.32
$w_{\rm p}(r_{\rm p}) + \Delta\Sigma(r_{\rm p})$	0.05	0.123	0.024	0.02	0.53	0.29
$w_{\rm p}(r_{\rm p})$ + VPF(r)	0.03	0.082	0.024	0.02	0.39	0.32
$w_{\rm p}(r_{\rm p}) + P(N_{\rm CIC})$	0.03	0.062	0.020	0.01	0.25	0.12
$w_{\rm p}(r_{\rm p}) + P(N_{\rm CIA})$	0.04	0.100	0.021	0.02	0.26	0.22
$w_{\rm p}(r_{\rm p}) + P(N_2/N_5)$	0.03	0.069	0.021	0.02	0.41	0.19
$\Delta\Sigma(r_{\rm p})$ + VPF(r)	0.04	0.083	0.054	0.04	0.39	0.25
$\Delta\Sigma(r_{\rm p}) + P(N_{\rm CIC})$	0.03	0.065	0.031	0.02	0.30	0.12
$\Delta\Sigma(r_{\rm p}) + P(N_{\rm CIA})$	0.03	0.076	0.042	0.03	0.30	0.22
$\Delta\Sigma(r_{\rm p}) + P(N_2/N_5)$	0.03	0.070	0.033	0.02	0.43	0.19
VPF(r) + $P(N_{\rm CIC})$	0.03	0.059	0.033	0.04	0.44	0.13
VPF(r) + $P(N_{\rm CIA})$	0.03	0.069	0.071	0.04	0.32	0.32
VPF(r) + $P(N_2/N_5)$	0.03	0.069	0.040	0.03	0.38	0.30
$P(N_{\rm CIC}) + P(N_{\rm CIA})$	0.03	0.067	0.027	0.03	0.38	0.11
$P(N_{\rm CIC}) + P(N_2/N_5)$	0.03	0.074	0.030	0.03	0.36	0.12
$P(N_{\rm CIA}) + P(N_2/N_5)$	0.03	0.073	0.036	0.03	0.34	0.24
All six	0.01	0.037	0.014	0.01	0.15	0.08

Table C.3: Halo occupation distribution (HOD) parameter constraints for the $M_r < -20.0$ sample.

Observable combination	$\sigma(\log M_{\rm min})$	$\sigma(\sigma_{\log M})$	$\sigma(\alpha)$	$\sigma(\log M_1)$	$\sigma(A_{\rm cen})$	$\sigma(A_{\rm sat})$
$w_{\rm p}(r_{\rm p})$	0.05	0.244	0.038	0.03	1.07	0.60
$\Delta\Sigma(r_{\rm p})$	0.05	0.197	0.098	0.06	1.05	0.52
VPF(r)	0.07	0.117	1.023	0.11	0.71	2.57
$P(N_{\rm CIC})$	0.07	0.282	0.070	0.06	1.53	0.25
$P(N_{\rm CIA})$	0.04	0.209	0.126	0.12	1.88	0.40
$P(N_2/N_5)$	0.04	0.186	0.060	0.03	1.42	0.41
$w_{\rm p}(r_{\rm p}) + \Delta\Sigma(r_{\rm p})$	0.03	0.132	0.036	0.02	0.70	0.34
$w_{\rm p}(r_{\rm p}) +$ VPF(r)	0.02	0.084	0.034	0.02	0.53	0.45
$w_{\rm p}(r_{\rm p}) + P(N_{\rm CIC})$	0.03	0.148	0.031	0.02	0.60	0.19
$w_{\rm p}(r_{\rm p}) + P(N_{\rm CIA})$	0.03	0.136	0.030	0.02	0.60	0.30
$w_{\rm p}(r_{\rm p}) + P(N_2/N_5)$	0.03	0.134	0.029	0.02	0.79	0.25
$\Delta\Sigma(r_{\rm p}) +$ VPF(r)	0.03	0.083	0.065	0.03	0.51	0.32
$\Delta\Sigma(r_{\rm p}) + P(N_{\rm CIC})$	0.04	0.140	0.039	0.02	0.57	0.20
$\Delta\Sigma(r_{\rm p}) + P(N_{\rm CIA})$	0.02	0.106	0.054	0.03	0.53	0.27
$\Delta\Sigma(r_{\rm p}) + P(N_2/N_5)$	0.03	0.118	0.039	0.02	0.67	0.24
VPF(r) + $P(N_{\rm CIC})$	0.02	0.084	0.036	0.03	0.60	0.22
VPF(r) + $P(N_{\rm CIA})$	0.02	0.083	0.073	0.03	0.53	0.37
VPF(r) + $P(N_2/N_5)$	0.02	0.088	0.042	0.03	0.58	0.38
$P(N_{\rm CIC}) + P(N_{\rm CIA})$	0.03	0.130	0.034	0.03	0.78	0.18
$P(N_{\rm CIC}) + P(N_2/N_5)$	0.03	0.142	0.037	0.02	0.87	0.20
$P(N_{\rm CIA}) + P(N_2/N_5)$	0.03	0.125	0.042	0.03	0.77	0.27
All six	0.01	0.052	0.020	0.01	0.29	0.11

Table C.4: Halo occupation distribution (HOD) parameter constraints for the $M_r < -20.5$ sample.

Observable combination	$\sigma(\log M_{\rm min})$	$\sigma(\sigma_{\log M})$	$\sigma(\alpha)$	$\sigma(\log M_1)$	$\sigma(A_{\rm cen})$	$\sigma(A_{\rm sat})$
$w_{\rm p}(r_{\rm p})$	0.19	0.308	0.097	0.03	0.84	0.90
$\Delta\Sigma(r_{\rm p})$	0.08	0.101	0.228	0.07	0.67	0.81
VPF(r)	0.12	0.149	1.558	0.10	0.53	4.10
$P(N_{\rm CIC})$	0.19	0.273	0.148	0.11	2.46	0.36
$P(N_{\rm CIA})$	0.06	0.070	0.277	0.15	1.48	0.88
$P(N_2/N_5)$	0.05	0.081	0.091	0.03	0.80	0.62
$w_{\rm p}(r_{\rm p}) + \Delta\Sigma(r_{\rm p})$	0.05	0.074	0.081	0.02	0.38	0.57
$w_{\rm p}(r_{\rm p}) + {\rm VPF}(r)$	0.05	0.073	0.084	0.03	0.41	0.70
$w_{\rm p}(r_{\rm p}) + P(N_{\rm CIC})$	0.06	0.083	0.065	0.02	0.56	0.26
$w_{\rm p}(r_{\rm p}) + P(N_{\rm CIA})$	0.04	0.061	0.076	0.02	0.40	0.50
$w_{\rm p}(r_{\rm p}) + P(N_2/N_5)$	0.04	0.068	0.063	0.02	0.49	0.37
$\Delta\Sigma(r_{\rm p}) + {\rm VPF}(r)$	0.05	0.062	0.159	0.04	0.36	0.56
$\Delta\Sigma(r_{\rm p}) + P(N_{\rm CIC})$	0.05	0.064	0.073	0.02	0.41	0.30
$\Delta\Sigma(r_{\rm p}) + P(N_{\rm CIA})$	0.03	0.044	0.119	0.03	0.30	0.47
$\Delta\Sigma(r_{\rm p}) + P(N_2/N_5)$	0.04	0.055	0.077	0.02	0.37	0.41
${\rm VPF}(r) + P(N_{\rm CIC})$	0.06	0.082	0.073	0.03	0.51	0.30
${\rm VPF}(r) + P(N_{\rm CIA})$	0.04	0.059	0.205	0.04	0.38	0.78
${\rm VPF}(r) + P(N_2/N_5)$	0.05	0.067	0.083	0.03	0.40	0.57
$P(N_{\rm CIC}) + P(N_{\rm CIA})$	0.04	0.064	0.062	0.03	0.61	0.25
$P(N_{\rm CIC}) + P(N_2/N_5)$	0.05	0.072	0.065	0.03	0.68	0.27
$P(N_{\rm CIA}) + P(N_2/N_5)$	0.04	0.060	0.077	0.03	0.52	0.42
All six	0.02	0.030	0.042	0.01	0.21	0.18

Table C.5: Halo occupation distribution (HOD) parameter constraints for the $M_r < -21.0$ sample.

Bibliography

[1] K. N. Abazajian, J. K. Adelman-McCarthy, M. A. Agüeros, S. S. Allam, C. Allende Prieto, D. An, K. S. J. Anderson, S. F. Anderson, J. Annis, N. A. Bahcall, and et al. The Seventh Data Release of the Sloan Digital Sky Survey. *Astrophysical Journal, Supplement*, 182:543–558, June 2009.

[2] T. M. C. Abbott, F. B. Abdalla, S. Avila, M. Banerji, E. Baxter, K. Bechtol, M. R. Becker, E. Bertin, J. Blazek, S. L. Bridle, D. Brooks, D. Brout, D. L. Burke, A. Campos, A. Carnero Rosell, M. Carrasco Kind, J. Carretero, F. J. Castander, R. Cawthon, C. Chang, A. Chen, M. Crocce, C. E. Cunha, L. N. da Costa, C. Davis, J. De Vicente, J. DeRose, S. Desai, E. Di Valentino, H. T. Diehl, J. P. Dietrich, S. Dodelson, P. Doel, A. Drlica-Wagner, T. F. Eifler, J. Elvin-Poole, A. E. Evrard, E. Fernand ez, A. Ferté, B. Flaugher, P. Fosalba, J. Frieman, J. García-Bellido, E. Gaztanaga, D. W. Gerdes, T. Giannantonio, D. Gruen, R. A. Gruendl, J. Gschwend, G. Gutierrez, W. G. Hartley, D. L. Hollowood, K. Honscheid, B. Hoyle, D. Huterer, B. Jain, T. Jeltema, M. W. G. Johnson, M. D. Johnson, A. G. Kim, E. Krause, K. Kuehn, N. Kuropatkin, O. Lahav, S. Lee, P. Lemos, C. D. Leonard, T. S. Li, A. R. Liddle, M. Lima, H. Lin, M. A. G. Maia, J. L. Marshall, P. Martini, F. Menanteau, C. J. Miller, R. Miquel, V. Miranda, J. J. Mohr, J. Muir, R. C. Nichol, B. Nord, R. L. C. Ogando, A. A. Plazas, M. Raveri, R. P. Rollins, A. K. Romer, A. Roodman, R. Rosenfeld, S. Samuroff, E. Sanchez, V. Scarpine, R. Schindler, M. Schubnell, D. Scolnic, L. F. Secco, S. Serrano, I. Sevilla-Noarbe, M. Smith, M. Soares-Santos, F. Sobreira, E. Suchyta, M. E. C. Swanson, G. Tarle, D. Thomas, M. A. Troxel, V. Vikram, A. R. Walker, N. Weaverdyck, R. H. Wechsler, J. Weller, B. Yanny, Y. Zhang, J. Zuntz, and DES Collaboration. Dark Energy Survey year 1 results: Constraints on extended cosmological models from galaxy clustering and weak lensing. *Physical Review D*, 99(12):123505, Jun 2019.

[3] H Akaike. Information theory as an extension of the maximum likelihood principle. á in: Petrov, bn and csaki, f. In *Second International Symposium on Information Theory. Akademiai Kiado, Budapest, pp. 276Á281*, 1973.

[4] Jean-Michel Alimi, Alain Blanchard, and Richard Schaeffer. Nonlinear Clustering in the CfA Redshift Survey. *Astrophysical Journal*, 349:L5, Jan 1990.

[5] Lauren Anderson, Eric Aubourg, Stephen Bailey, Dmitry Bizyaev, Michael Blanton, Adam S. Bolton, J. Brinkmann, Joel R. Brownstein, Angela Burden, Antonio J. Cuesta, Luiz A. N. da Costa, Kyle S. Dawson, Roland de Putter, Daniel J. Eisenstein, James E. Gunn, Hong Guo, Jean-Christophe Hamilton, Paul Harding, Shirley Ho, Klaus Honscheid, Eyal Kazin, David Kirkby, Jean-Paul Kneib, Antoine Labatie, Craig Loomis, Robert H. Lupton, Elena Malanushenko, Viktor Malanushenko, Rachel Mandelbaum, Marc Manera,

Claudia Maraston, Cameron K. McBride, Kushal T. Mehta, Olga Mena, Francesco Montesano, Demetri Muna, Robert C. Nichol, Sebastián E. Nuza, Matthew D. Olmstead, Daniel Oravetz, Nikhil Padmanabhan, Nathalie Palanque-Delabrouille, Kaike Pan, John Parejko, Isabelle Pâris, Will J. Percival, Patrick Petitjean, Francisco Prada, Beth Reid, Natalie A. Roe, Ashley J. Ross, Nicholas P. Ross, Lado Samushia, Ariel G. Sánchez, David J. Schlegel, Donald P. Schneider, Claudia G. Scóccola, Hee-Jong Seo, Erin S. Sheldon, Audrey Simmons, Ramin A. Skibba, Michael A. Strauss, Molly E. C. Swanson, Daniel Thomas, Jeremy L. Tinker, Rita Tojeiro, Mariana Vargas Magaña, Licia Verde, Christian Wagner, David A. Wake, Benjamin A. Weaver, David H. Weinberg, Martin White, Xiaoying Xu, Christophe Yèche, Idit Zehavi, and Gong-Bo Zhao. The clustering of galaxies in the SDSS-III Baryon Oscillation Spectroscopic Survey: baryon acoustic oscillations in the Data Release 9 spectroscopic galaxy sample. *Monthly Notices of the RAS*, 427:3435–3467, Dec 2012.

[6] M. Celeste Artale, Idit Zehavi, Sergio Contreras, and Peder Norberg. The impact of assembly bias on the halo occupation in hydrodynamical simulations. *Monthly Notices of the RAS*, 480(3):3978–3992, Nov 2018.

[7] Roger Balian and Richard Schaeffer. Galaxies: Fractal Dimensions, Counts in Cells, and Correlations. *Astrophysical Journal*, 335:L43, December 1988.

[8] J. M. Bardeen, J. R. Bond, N. Kaiser, and A. S. Szalay. The statistics of peaks of Gaussian random fields. *Astrophysical Journal*, 304:15–61, May 1986.

[9] Elizabeth J. Barton, Jacob A. Arnold, Andrew R. Zentner, James S. Bullock, and Risa H. Wechsler. Isolating Triggered Star Formation. *Astrophysical Journal*, 671:1538–1549, December 2007.

[10] C. M. Baugh, E. Gaztanaga, and G. Efstathiou. A comparison of the evolution of density fields in perturbation theory and numerical simulations - II. Counts-in-cells analysis. *Monthly Notices of the RAS*, 274:1049–1070, June 1995.

[11] P. S. Behroozi, C. Conroy, and R. H. Wechsler. A Comprehensive Analysis of Uncertainties Affecting the Stellar Mass-Halo Mass Relation for $0 < z < 4$. *Astrophysical Journal*, 717:379–403, July 2010.

[12] P. S. Behroozi, R. H. Wechsler, and C. Conroy. The Average Star Formation Histories of Galaxies in Dark Matter Halos from z = 0-8. *Astrophysical Journal*, 770:57, June 2013.

[13] Peter S. Behroozi, Risa H. Wechsler, and Hao-Yi Wu. The ROCKSTAR Phase-space Temporal Halo Finder and the Velocity Offsets of Cluster Cores. *Astrophysical Journal*, 762(2):109, Jan 2013.

[14] Peter S. Behroozi, Risa H. Wechsler, Hao-Yi Wu, Michael T. Busha, Anatoly A. Klypin, and Joel R. Primack. Gravitationally Consistent Halo Catalogs and Merger Trees for Precision Cosmology. *Astrophysical Journal*, 763(1):18, Jan 2013.

[15] Andrew J. Benson, Aaron Ludlow, and Shaun Cole. Halo concentrations from extended Press-Schechter merger histories. *Monthly Notices of the RAS*, 485(4):5010–5020, June 2019.

[16] A. A. Berlind and D. H. Weinberg. The Halo Occupation Distribution: Toward an Empirical Determination of the Relation between Galaxies and Mass. *Astrophysical Journal*, 575:587–616, August 2002.

[17] Heather D. Berrier, Elizabeth J. Barton, Joel C. Berrier, James S. Bullock, Andrew R. Zentner, and Risa H. Wechsler. Counts-in-cylinders in the Sloan Digital Sky Survey with Comparisons to N-body Simulations. *Astrophysical Journal*, 726:1, January 2011.

[18] Marc Betoule, J. Guy, R. Kessler, J. Mosher, P. Astier, R. Biswas, P. El Hage, D. Hardin, J. Marriner, R. Pain, and N. Regnault. Improved cosmological constraints from a joint analysis of the SNLS and SDSS surveys. In *American Astronomical Society Meeting Abstracts #223*, volume 223, page 427.04, January 2014.

[19] Suman Bhattacharya, Salman Habib, Katrin Heitmann, and Alexey Vikhlinin. Dark Matter Halo Profiles of Massive Clusters: Theory versus Observations. *Astrophysical Journal*, 766(1):32, Mar 2013.

[20] M. R. Blanton and S. Roweis. K-Corrections and Filter Transformations in the Ultraviolet, Optical, and Near-Infrared. *Astronomical Journal*, 133:734–754, February 2007.

[21] M. R. Blanton, D. J. Schlegel, M. A. Strauss, J. Brinkmann, D. Finkbeiner, M. Fukugita, J. E. Gunn, D. W. Hogg, Ž. Ivezić, G. R. Knapp, R. H. Lupton, J. A. Munn, D. P. Schneider, M. Tegmark, and I. Zehavi. New York University Value-Added Galaxy Catalog: A Galaxy Catalog Based on New Public Surveys. *Astronomical Journal*, 129:2562–2578, June 2005.

[22] Michael R. Blanton, Daniel Eisenstein, David W. Hogg, and Idit Zehavi. The Scale Dependence of Relative Galaxy Bias: Encouragement for the "Halo Model" Description. *Astrophysical Journal*, 645:977–985, July 2006.

[23] G. R. Blumenthal, S. M. Faber, J. R. Primack, and M. J. Rees. Formation of galaxies and large-scale structure with cold dark matter. *Nature*, 311:517–525, Oct 1984.

[24] J. R. Bond, S. Cole, G. Efstathiou, and N. Kaiser. Excursion Set Mass Functions for Hierarchical Gaussian Fluctuations. *Astrophysical Journal*, 379:440, Oct 1991.

[25] J. R. Bond, A. H. Jaffe, and L. Knox. Estimating the power spectrum of the cosmic microwave background. *Physical Review D*, 57:2117–2137, February 1998.

[26] Sownak Bose, Daniel J. Eisenstein, Lars Hernquist, Annalisa Pillepich, Dylan Nelson, Federico Marinacci, Volker Springel, and Mark Vogelsberger. Revealing the galaxy-halo connection in IllustrisTNG. *arXiv e-prints*, page arXiv:1905.08799, May 2019.

[27] Greg L. Bryan and Michael L. Norman. Statistical Properties of X-Ray Clusters: Analytic and Numerical Comparisons. *Astrophysical Journal*, 495(1):80–99, Mar 1998.

[28] J. S. Bullock, T. S. Kolatt, Y. Sigad, R. S. Somerville, A. V. Kravtsov, A. A. Klypin, J. R. Primack, and A. Dekel. Profiles of dark haloes: evolution, scatter and environment. *Monthly Notices of the RAS*, 321(3):559–575, Mar 2001.

[29] M. Cacciato, F. C. van den Bosch, S. More, R. Li, H. J. Mo, and X. Yang. Galaxy clustering and galaxy-galaxy lensing: a promising union to constrain cosmological parameters. *Monthly Notices of the RAS*, 394:929–946, April 2009.

[30] M. Cacciato, F. C. van den Bosch, S. More, H. Mo, and X. Yang. Cosmological constraints from a combination of galaxy clustering and lensing - III. Application to SDSS data. *Monthly Notices of the RAS*, 430:767–786, April 2013.

[31] Jonás Chaves-Montero, Raul E. Angulo, Joop Schaye, Matthieu Schaller, Robert A. Crain, Michelle Furlong, and Tom Theuns. Subhalo abundance matching and assembly bias in the EAGLE simulation. *Monthly Notices of the RAS*, 460(3):3100–3118, Aug 2016.

[32] J. Chen. Color dependence in the spatial distribution of satellite galaxies. *Astronomy and Astrophysics*, 484:347–354, June 2008.

[33] Hillary L. Child, Salman Habib, Katrin Heitmann, Nicholas Frontiere, Hal Finkel, Adrian Pope, and Vitali Morozov. Halo Profiles and the Concentration-Mass Relation for a ΛCDM Universe. *Astrophysical Journal*, 859(1):55, May 2018.

[34] S. Colombi, F. R. Bouchet, and R. Schaeffer. A Count Probability Cookbook: Spurious Effects and the Scaling Model. *The Astrophysical Journal Supplement Series*, 96:401, Feb 1995.

[35] Julia M. Comerford and Priyamvada Natarajan. The observed concentration-mass relation for galaxy clusters. *Monthly Notices of the RAS*, 379(1):190–200, July 2007.

[36] C. Conroy, R. H. Wechsler, and A. V. Kravtsov. Modeling Luminosity-dependent Galaxy Clustering through Cosmic Time. *Astrophysical Journal*, 647:201–214, August 2006.

[37] S. Contreras, R. E. Angulo, and M. Zennaro. A flexible modelling of galaxy assembly bias. *Monthly Notices of the RAS*, 504(4):5205–5220, July 2021.

[38] A. Cooray. Halo model at its best: constraints on conditional luminosity functions from measured galaxy statistics. *Monthly Notices of the RAS*, 365:842–866, January 2006.

[39] A. Cooray and R. Sheth. Halo models of large scale structure. *Physics Reports*, 372:1–129, December 2002.

[40] Camila A. Correa, J. Stuart B. Wyithe, Joop Schaye, and Alan R. Duffy. The accretion history of dark matter haloes - I. The physical origin of the universal function. *Monthly Notices of the RAS*, 450(2):1514–1520, June 2015.

[41] Camila A. Correa, J. Stuart B. Wyithe, Joop Schaye, and Alan R. Duffy. The accretion history of dark matter haloes - II. The connections with the mass power spectrum and the density profile. *Monthly Notices of the RAS*, 450(2):1521–1537, June 2015.

[42] Camila A. Correa, J. Stuart B. Wyithe, Joop Schaye, and Alan R. Duffy. The accretion history of dark matter haloes - III. A physical model for the concentration-mass relation. *Monthly Notices of the RAS*, 452(2):1217–1232, September 2015.

[43] J. Coupon, S. Arnouts, L. van Waerbeke, T. Moutard, O. Ilbert, E. van Uitert, T. Erben, B. Garilli, L. Guzzo, C. Heymans, H. Hildebrandt, H. Hoekstra, M. Kilbinger, T. Kitching, Y. Mellier, L. Miller, M. Scodeggio, C. Bonnett, E. Branchini, I. Davidzon, G. De Lucia, A. Fritz, L. Fu, P. Hudelot, M. J. Hudson, K. Kuijken, A. Leauthaud, O. Le Fèvre, H. J. McCracken, L. Moscardini, B. T. P. Rowe, T. Schrabback, E. Semboloni, and M. Velander. The galaxy-halo connection from a joint lensing, clustering and abundance analysis in the CFHTLenS/VIPERS field. *Monthly Notices of the RAS*, 449:1352–1379, May 2015.

[44] William I. Cowley, Karina I. Caputi, Smaran Deshmukh, Matthew L. N. Ashby, Giovanni G. Fazio, Olivier Le Fèvre, Johan P. U. Fynbo, Olivier Ilbert, Henry J. McCracken, Bo Milvang-Jensen, and Rachel S. Somerville. The Galaxy-Halo Connection for $1.5 \lesssim z \lesssim 5$ as Revealed by the Spitzer Matching Survey of the UltraVISTA Ultra-deep Stripes. *Astrophysical Journal*, 853:69, Jan 2018.

[45] D. J. Croton, L. Gao, and S. D. M. White. Halo assembly bias and its effects on galaxy clustering. *Monthly Notices of the RAS*, 374:1303–1309, February 2007.

[46] Joseph DeRose, Risa H. Wechsler, Jeremy L. Tinker, Matthew R. Becker, Yao-Yuan Mao, Thomas McClintock, Sean McLaughlin, Eduardo Rozo, and Zhongxu Zhai. The AEMULUS Project. I. Numerical Simulations for Precision Cosmology. *Astrophysical Journal*, 875(1):69, April 2019.

[47] Vincent Desjacques, Donghui Jeong, and Fabian Schmidt. Large-scale galaxy bias. *Physics Reports*, 733:1–193, Feb 2018.

[48] B. Diemer and A. V. Kravtsov. Dependence of the Outer Density Profiles of Halos on Their Mass Accretion Rate. *Astrophysical Journal*, 789:1, July 2014.

[49] B. Diemer, S. More, and A. V. Kravtsov. The Pseudo-evolution of Halo Mass. *ApJ*, 766:25, March 2013.

[50] Benedikt Diemer. COLOSSUS: A Python Toolkit for Cosmology, Large-scale Structure, and Dark Matter Halos. *Astrophysical Journal, Supplement*, 239(2):35, Dec 2018.

[51] Benedikt Diemer and Andrey V. Kravtsov. A Universal Model for Halo Concentrations. *Astrophysical Journal*, 799(1):108, Jan 2015.

[52] S. Dodelson. *Modern cosmology*. Elsevier, 2003.

[53] A. Dressler. Galaxy morphology in rich clusters - Implications for the formation and evolution of galaxies. *Astrophysical Journal*, 236:351–365, March 1980.

[54] Alan R. Duffy, Joop Schaye, Scott T. Kay, and Claudio Dalla Vecchia. Dark matter halo concentrations in the Wilkinson Microwave Anisotropy Probe year 5 cosmology. *Monthly Notices of the RAS*, 390(1):L64–L68, October 2008.

[55] J. Einasto. Influence of the atmospheric and instrumental dispersion on the brightness distribution in a galaxy. *Trudy Inst. Astrofiz. Alma-Ata 51, 87*, 1965.

[56] D. Foreman-Mackey, D. W. Hogg, D. Lang, and J. Goodman. emcee: The mcmc hammer. *PASP*, 125:306–312, 2013.

[57] O. Friedrich, D. Gruen, J. DeRose, D. Kirk, E. Krause, T. McClintock, E. S. Rykoff, S. Seitz, R. H. Wechsler, G. M. Bernstein, J. Blazek, C. Chang, S. Hilbert, B. Jain, A. Kovacs, O. Lahav, F. B. Abdalla, S. Allam, J. Annis, K. Bechtol, A. Benoit-Lévy, E. Bertin, D. Brooks, A. Carnero Rosell, M. Carrasco Kind, J. Carretero, C. E. Cunha, C. B. D'Andrea, L. N. da Costa, C. Davis, S. Desai, H. T. Diehl, J. P. Dietrich, A. Drlica-Wagner, T. F. Eifler, P. Fosalba, J. Frieman, J. García-Bellido, E. Gaztanaga, D. W. Gerdes, T. Giannantonio, R. A. Gruendl, J. Gschwend, G. Gutierrez, K. Honscheid, D. J. James, M. Jarvis, K. Kuehn, N. Kuropatkin, M. Lima, M. March, J. L. Marshall, P. Melchior, F. Menanteau, R. Miquel, J. J. Mohr, B. Nord, A. A. Plazas, E. Sanchez, V. Scarpine, R. Schindler, M. Schubnell, I. Sevilla-Noarbe, E. Sheldon, M. Smith, M. Soares-Santos, F. Sobreira, E. Suchyta, M. E. C. Swanson, G. Tarle, D. Thomas, M. A. Troxel, V. Vikram, J. Weller, and DES Collaboration. Density split statistics: Joint model of counts and lensing in cells. *Physical Review D*, 98:023508, Jul 2018.

[58] J. N. Fry and P. J. E. Peebles. Statistical analysis of catalogs of extragalactic objects. IX. The four- point galaxy correlation function. *Astrophysical Journal*, 221:19–33, April 1978.

[59] L. Gao, J. F. Navarro, S. Cole, C. S. Frenk, S. D. M. White, V. Springel, A. Jenkins, and A. F. Neto. The redshift dependence of the structure of massive Λ cold dark matter haloes. *Monthly Notices of the RAS*, 387:536–544, June 2008.

[60] L. Gao and S. D. M. White. Assembly bias in the clustering of dark matter haloes. *Monthly Notices of the RAS*, 377:L5–L9, April 2007.

[61] Liang Gao, Volker Springel, and Simon D. M. White. The age dependence of halo clustering. *Monthly Notices of the RAS*, 363(1):L66–L70, Oct 2005.

[62] J. Goodman and J Weare. Ensemble samplers with affine invariance. *Comm. App. Math. and Comp. Sci*, 5(1), 2010.

[63] D. Gruen, O. Friedrich, E. Krause, J. DeRose, R. Cawthon, C. Davis, J. Elvin-Poole, E. S. Rykoff, R. H. Wechsler, A. Alarcon, G. M. Bernstein, J. Blazek, C. Chang, J. Clampitt, M. Crocce, J. De Vicente, M. Gatti, M. S. S. Gill, W. G. Hartley, S. Hilbert, B. Hoyle, B. Jain, M. Jarvis, O. Lahav, N. MacCrann, T. McClintock, J. Prat, R. P. Rollins, A. J. Ross, E. Rozo, S. Samuroff, C. Sánchez, E. Sheldon, M. A. Troxel, J. Zuntz, T. M. C. Abbott, F. B. Abdalla, S. Allam, J. Annis, K. Bechtol, A. Benoit-Lévy, E. Bertin, S. L. Bridle, D. Brooks, E. Buckley-Geer, A. Carnero Rosell, M. Carrasco Kind, J. Carretero, C. E. Cunha, C. B. D'Andrea, L. N. da Costa, S. Desai, H. T. Diehl, J. P. Dietrich, P. Doel, A. Drlica-Wagner, E. Fernandez, B. Flaugher, P. Fosalba, J. Frieman, J. García-Bellido, E. Gaztanaga, T. Giannantonio, R. A. Gruendl, J. Gschwend, G. Gutierrez, K. Honscheid, D. J. James, T. Jeltema, K. Kuehn, N. Kuropatkin, M. Lima, M. March, J. L. Marshall, P. Martini, P. Melchior, F. Menanteau, R. Miquel, J. J. Mohr, A. A. Plazas, A. Roodman, E. Sanchez, V. Scarpine,

M. Schubnell, I. Sevilla-Noarbe, M. Smith, R. C. Smith, M. Soares-Santos, F. Sobreira, M. E. C. Swanson, G. Tarle, D. Thomas, V. Vikram, A. R. Walker, J. Weller, Y. Zhang, and DES Collaboration. Density split statistics: Cosmological constraints from counts and lensing in cells in DES Y1 and SDSS data. *Physical Review D*, 98:023507, Jul 2018.

[64] H. Guo, I. Zehavi, and Z. Zheng. A New Method to Correct for Fiber Collisions in Galaxy Two-point Statistics. *Astrophysical Journal*, 756:127, September 2012.

[65] H. Guo, Z. Zheng, I. Zehavi, H. Xu, D. J. Eisenstein, D. H. Weinberg, N. A. Bahcall, A. A. Berlind, J. Comparat, C. K. McBride, A. J. Ross, D. P. Schneider, R. A. Skibba, M. E. C. Swanson, J. L. Tinker, R. Tojeiro, and D. A. Wake. The clustering of galaxies in the SDSS-III Baryon Oscillation Spectroscopic Survey: modeling of the luminosity and colour dependence in the Data Release 10. *ArXiv:1401.3009*, January 2014.

[66] Hong Guo, Zheng Zheng, Idit Zehavi, Peter S. Behroozi, Chia-Hsun Chuang, Johan Comparat, Ginevra Favole, Stefan Gottloeber, Anatoly Klypin, and Francisco Prada. Redshift-space clustering of SDSS galaxies - luminosity dependence, halo occupation distribution, and velocity bias. *Monthly Notices of the RAS*, 453(4):4368–4383, Nov 2015.

[67] Hong Guo, Zheng Zheng, Idit Zehavi, Peter S. Behroozi, Chia-Hsun Chuang, Johan Comparat, Ginevra Favole, Stefan Gottloeber, Anatoly Klypin, Francisco Prada, David H. Weinberg, and Gustavo Yepes. Redshift-space clustering of SDSS galaxies - luminosity dependence, halo occupation distribution, and velocity bias. *Monthly Notices of the RAS*, 453(4):4368–4383, November 2015.

[68] Q. Guo, S. Cole, V. Eke, and C. Frenk. Satellite Galaxy Number Density Profiles in the Sloan Digital Sky Survey. *ArXiv:1201.1296*, January 2012.

[69] Alan H. Guth. Inflationary universe: A possible solution to the horizon and flatness problems. *Phys. Rev. D*, 23:347–356, Jan 1981.

[70] Ed Hawkins, Steve Maddox, Shaun Cole, Ofer Lahav, Darren S. Madgwick, Peder Norberg, John A. Peacock, Ivan K. Baldry, Carlton M. Baugh, Joss Bland-Hawthorn, Terry Bridges, Russell Cannon, Matthew Colless, Chris Collins, Warrick Couch, Gavin Dalton, Roberto De Propris, Simon P. Driver, George Efstathiou, Richard S. Ellis, Carlos S. Frenk, Karl Glazebrook, Carole Jackson, Bryn Jones, Ian Lewis, Stuart Lumsden, Will Percival, Bruce A. Peterson, Will Sutherland, and Keith Taylor. The 2dF Galaxy Redshift Survey: correlation functions, peculiar velocities and the matter density of the Universe. *Monthly Notices of the RAS*, 346:78–96, Nov 2003.

[71] A. Hearin, D. Campbell, E. Tollerud, P. Behroozi, B. Diemer, N. J. Goldbaum, E. Jennings, A. Leauthaud, Y.-Y. Mao, S. More, J. Parejko, M. Sinha, B. Sipocz, and A. Zentner. High-Precision Forward Modeling of Large-Scale Structure: An open-source approach with Halotools. *ArXiv:1606.04106*, June 2016.

[72] A. P. Hearin and D. F. Watson. The Dark Side of Galaxy Color. *ArXiv:1304.5557*, April 2013.

[73] A. P. Hearin, A. R. Zentner, J. A. Newman, and A. A. Berlind. Mind the Gap: Tightening the Mass-Richness Relation with Magnitude Gaps. *ArXiv:1207.1074*, July 2012.

[74] A. P. Hearin, A. R. Zentner, F. C. van den Bosch, D. Campbell, and E. Tollerud. Introducing Decorated HODs: modeling assembly bias in the galaxy-halo connection. *Monthly Notices of the RAS*, May 2016.

[75] Andrew P. Hearin, Andrew R. Zentner, Andreas A. Berlind, and Jeffrey A. Newman. SHAM beyond clustering: new tests of galaxy-halo abundance matching with galaxy groups. *Monthly Notices of the RAS*, 433(1):659–680, Jul 2013.

[76] Katrin Heitmann, Hal Finkel, Adrian Pope, Vitali Morozov, Nicholas Frontiere, Salman Habib, Esteban Rangel, Thomas Uram, Danila Korytov, Hillary Child, Samuel Flender, Joe Insley, and Silvio Rizzi. The Outer Rim Simulation: A Path to Many-core Supercomputers. *Astrophysical Journal, Supplement*, 245(1):16, November 2019.

[77] David W. Hogg, Michael R. Blanton, Jarle Brinchmann, Daniel J. Eisenstein, David J. Schlegel, James E. Gunn, Timothy A. McKay, Hans-Walter Rix, Neta A. Bahcall, J. Brinkmann, and Avery Meiksin. The Dependence on Environment of the Color-Magnitude Relation of Galaxies. *Astrophysical Journal*, 601:L29–L32, January 2004.

[78] E. Hubble. A Relation between Distance and Radial Velocity among Extra-Galactic Nebulae. *Proceedings of the National Academy of Science*, 15:168–173, March 1929.

[79] F. Jiang and F. C. van den Bosch. Statistics of dark matter substructure - III. Halo-to-halo variance. *Monthly Notices of the RAS*, 472:657–674, November 2017.

[80] Fangzhou Jiang and Frank C. van den Bosch. Statistics of dark matter substructure - I. Model and universal fitting functions. *Monthly Notices of the RAS*, 458(3):2848–2869, May 2016.

[81] Turner Johnson, Andrew J. Benson, and Daniel Grin. A Random Walk Model for Dark Matter Halo Concentrations. *arXiv e-prints*, page arXiv:2006.15231, June 2020.

[82] Howard L. Jones. Investigating the properties of a sample mean by employing random subsample means. *Journal of the American Statistical Association*, 51(273):54–83, 1956.

[83] N. Kaiser. On the spatial correlations of Abell clusters. *Astrophysical Journal, Letters*, 284:L9–L12, September 1984.

[84] Guinevere Kauffmann, Simon D. M. White, Timothy M. Heckman, Brice Ménard, Jarle Brinchmann, Stéphane Charlot, Christy Tremonti, and Jon Brinkmann. The environmental dependence of the relations between stellar mass, structure, star formation and nuclear activity in galaxies. *Monthly Notices of the RAS*, 353:713–731, September 2004.

[85] Rita Seung Jung Kim and Michael A. Strauss. Measuring High-Order Moments of the Galaxy Distribution from Counts in Cells: The Edgeworth Approximation. *Astrophysical Journal*, 493:39–51, Jan 1998.

[86] A. Klypin, S. Gottlöber, A. V. Kravtsov, and A. M. Khokhlov. Galaxies in N-Body Simulations: Overcoming the Overmerging Problem. *Astrophysical Journal*, 516:530–551, May 1999.

[87] Anatoly Klypin, Gustavo Yepes, Stefan Gottlöber, Francisco Prada, and Steffen Heß. MultiDark simulations: the story of dark matter halo concentrations and density profiles. *Monthly Notices of the RAS*, 457(4):4340–4359, Apr 2016.

[88] Anatoly A. Klypin, Sebastian Trujillo-Gomez, and Joel Primack. Dark Matter Halos in the Standard Cosmological Model: Results from the Bolshoi Simulation. *Astrophysical Journal*, 740(2):102, Oct 2011.

[89] Anatoly A. Klypin, Sebastian Trujillo-Gomez, and Joel Primack. Dark Matter Halos in the Standard Cosmological Model: Results from the Bolshoi Simulation. *Astrophysical Journal*, 740(2):102, October 2011.

[90] Fabian Köhlinger, Benjamin Joachimi, Marika Asgari, Massimo Viola, Shahab Joudaki, and Tilman Tröster. A Bayesian quantification of consistency in correlated data sets. *Monthly Notices of the RAS*, 484(3):3126–3153, April 2019.

[91] E. Komatsu, K. M. Smith, J. Dunkley, C. L. Bennett, B. Gold, G. Hinshaw, N. Jarosik, D. Larson, M. R. Nolta, L. Page, D. N. Spergel, M. Halpern, R. S. Hill, A. Kogut, M. Limon,

S. S. Meyer, N. Odegard, G. S. Tucker, J. L. Weiland, E. Wollack, and E. L. Wright. Seven-year Wilkinson Microwave Anisotropy Probe (WMAP) Observations: Cosmological Interpretation. *Astrophysical Journal, Supplement*, 192(2):18, Feb 2011.

[92] A. V. Kravtsov, A. A. Berlind, R. H. Wechsler, A. A. Klypin, S. Gottlöber, B. Allgood, and J. R. Primack. The Dark Side of the Halo Occupation Distribution. *Astrophysical Journal*, 609:35–49, July 2004.

[93] S. D. Landy and A. S. Szalay. Bias and variance of angular correlation functions. *Astrophysical Journal*, 412:64–71, July 1993.

[94] J. U. Lange, F. C. van den Bosch, A. R. Zentner, K. Wang, and A. S. Villarreal. Updated Results on the Galaxy-Halo Connection from Satellite Kinematics in SDSS. *arXiv e-prints*, November 2018.

[95] Johannes U. Lange, Frank C. van den Bosch, Andrew R. Zentner, Kuan Wang, and Antonio S. Villarreal. Maturing Satellite Kinematics into a Competitive Probe of the Galaxy-Halo Connection. *ArXiv e-prints*, page arXiv:1810.10511, October 2018.

[96] Johannes U. Lange, Frank C. van den Bosch, Andrew R. Zentner, Kuan Wang, and Antonio S. Villarreal. Updated Results on the Galaxy-Halo Connection from Satellite Kinematics in SDSS. *arXiv e-prints*, page arXiv:1811.03596, Nov 2018.

[97] A. Leauthaud et al. New Constraints on the Evolution of the Stellar-to-dark Matter Connection: A Combined Analysis of Galaxy-Galaxy Lensing, Clustering, and Stellar Mass Functions from z = 0.2 to z =1. *Astrophysical Journal*, 744:159, January 2012.

[98] Christoph T. Lee, Joel R. Primack, Peter Behroozi, Aldo Rodríguez-Puebla, Doug Hellinger, and Avishai Dekel. Properties of dark matter haloes as a function of local environment density. *Monthly Notices of the RAS*, 466(4):3834–3858, April 2017.

[99] Christoph T. Lee, Joel R. Primack, Peter Behroozi, Aldo Rodríguez-Puebla, Doug Hellinger, and Avishai Dekel. Tidal stripping and post-merger relaxation of dark matter haloes: causes and consequences of mass-loss. *Monthly Notices of the RAS*, 481(3):4038–4057, December 2018.

[100] Benjamin V. Lehmann, Yao-Yuan Mao, Matthew R. Becker, Samuel W. Skillman, and Risa H. Wechsler. The Concentration Dependence of the Galaxy-Halo Connection: Modeling Assembly Bias with Abundance Matching. *Astrophysical Journal*, 834:37, Jan 2017.

[101] Michael Levi, Chris Bebek, Timothy Beers, Robert Blum, Robert Cahn, Daniel Eisenstein, Brenna Flaugher, Klaus Honscheid, Richard Kron, Ofer Lahav, Patrick McDonald, Natalie Roe, David Schlegel, and representing the DESI collaboration. The DESI Experiment, a whitepaper for Snowmass 2013. *arXiv e-prints*, page arXiv:1308.0847, Aug 2013.

[102] Pengfei Li, Federico Lelli, Stacy S. McGaugh, Nathaniel Starkman, and James M. Schombert. A constant characteristic volume density of dark matter haloes from SPARC rotation curve fits. *Monthly Notices of the RAS*, 482(4):5106–5124, February 2019.

[103] Yin Li, Yueying Ni, Rupert A. C. Croft, Tiziana Di Matteo, Simeon Bird, and Yu Feng. AI-assisted superresolution cosmological simulations. *Proceedings of the National Academy of Science*, 118(19):2022038118, May 2021.

[104] Yun Li, H. J. Mo, and L. Gao. On halo formation times and assembly bias. *Monthly Notices of the RAS*, 389(3):1419–1426, Sep 2008.

[105] Yun Li, H. J. Mo, Frank C. van den Bosch, and W. P. Lin. On the assembly history of dark matter haloes. *Monthly Notices of the RAS*, 379(2):689–701, Aug 2007.

[106] Yen-Ting Lin, Joseph J. Mohr, and S. Adam Stanford. K-Band Properties of Galaxy Clusters and Groups: Luminosity Function, Radial Distribution, and Halo Occupation Number. *Astrophysical Journal*, 610:745–761, August 2004.

[107] Catherine Loader. *locfit: Local Regression, Likelihood and Density Estimation.*, 2013. R package version 1.5-9.1.

[108] LSST Science Collaboration, Paul A. Abell, Julius Allison, Scott F. Anderson, John R. Andrew, J. Roger P. Angel, Lee Armus, David Arnett, S. J. Asztalos, Tim S. Axelrod, Stephen Bailey, D. R. Ballantyne, Justin R. Bankert, Wayne A. Barkhouse, Jeffrey D. Barr, L. Felipe Barrientos, Aaron J. Barth, James G. Bartlett, Andrew C. Becker, Jacek Becla, Timothy C. Beers, Joseph P. Bernstein, Rahul Biswas, Michael R. Blanton, Joshua S. Bloom, John J. Bochanski, Pat Boeshaar, Kirk D. Borne, Marusa Bradac, W. N. Brandt, Carrie R. Bridge, Michael E. Brown, Robert J. Brunner, James S. Bullock, Adam J. Burgasser, James H. Burge, David L. Burke, Phillip A. Cargile, Srinivasan Chand rasekharan, George Chartas, Steven R. Chesley, You-Hua Chu, David Cinabro, Mark W. Claire, Charles F. Claver, Douglas Clowe, A. J. Connolly, Kem H. Cook, Jeff Cooke, Asantha Cooray, Kevin R. Covey, Christopher S. Culliton, Roelof de Jong, Willem H. de Vries, Victor P. Debattista, Francisco Delgado, Ian P. Dell'Antonio, Saurav Dhital, Rosanne Di Stefano, Mark Dickinson, Benjamin Dilday, S. G. Djorgovski, Gregory Dobler, Ciro Donalek, Gregory Dubois-Felsmann, Josef Durech, Ardis Eliasdottir, Michael Eracleous, Laurent Eyer, Emilio E. Falco, Xiaohui Fan, Christopher D. Fassnacht, Harry C. Ferguson, Yanga R. Fernandez, Brian D. Fields,

Douglas Finkbeiner, Eduardo E. Figueroa, Derek B. Fox, Harold Francke, James S. Frank, Josh Frieman, Sebastien Fromenteau, Muhammad Furqan, Gaspar Galaz, A. Gal-Yam, Peter Garnavich, Eric Gawiser, John Geary, Perry Gee, Robert R. Gibson, Kirk Gilmore, Emily A. Grace, Richard F. Green, William J. Gressler, Carl J. Grillmair, Salman Habib, J. S. Haggerty, Mario Hamuy, Alan W. Harris, Suzanne L. Hawley, Alan F. Heavens, Leslie Hebb, Todd J. Henry, Edward Hileman, Eric J. Hilton, Keri Hoadley, J. B. Holberg, Matt J. Holman, Steve B. Howell, Leopoldo Infante, Zeljko Ivezic, Suzanne H. Jacoby, Bhuvnesh Jain, R, Jedicke, M. James Jee, J. Garrett Jernigan, Saurabh W. Jha, Kathryn V. Johnston, R. Lynne Jones, Mario Juric, Mikko Kaasalainen, Styliani, Kafka, Steven M. Kahn, Nathan A. Kaib, Jason Kalirai, Jeff Kantor, Mansi M. Kasliwal, Charles R. Keeton, Richard Kessler, Zoran Knezevic, Adam Kowalski, Victor L. Krabbendam, K. Simon Krughoff, Shrinivas Kulkarni, Stephen Kuhlman, Mark Lacy, Sebastien Lepine, Ming Liang, Amy Lien, Paulina Lira, Knox S. Long, Suzanne Lorenz, Jennifer M. Lotz, R. H. Lupton, Julie Lutz, Lucas M. Macri, Ashish A. Mahabal, Rachel Mandelbaum, Phil Marshall, Morgan May, Peregrine M. McGehee, Brian T. Meadows, Alan Meert, Andrea Milani, Christopher J. Miller, Michelle Miller, David Mills, Dante Minniti, David Monet, Anjum S. Mukadam, Ehud Nakar, Douglas R. Neill, Jeffrey A. Newman, Sergei Nikolaev, Martin Nordby, Paul O'Connor, Masamune Oguri, John Oliver, Scot S. Olivier, Julia K. Olsen, Knut Olsen, Edward W. Olszewski, Hakeem Oluseyi, Nelson D. Padilla, Alex Parker, Joshua Pepper, John R. Peterson, Catherine Petry, Philip A. Pinto, James L. Pizagno, Bogdan Popescu, Andrej Prsa, Veljko Radcka, M. Jordan Raddick, Andrew Rasmussen, Arne Rau, Jeonghee Rho, James E. Rhoads, Gordon T. Richards, Stephen T. Ridgway, Brant E. Robertson, Rok Roskar, Abhijit Saha, Ata Sarajedini, Evan Scannapieco, Terry Schalk, Rafe Schindler, Samuel Schmidt, Sarah Schmidt, Donald P. Schneider, German Schumacher, Ryan Scranton, Jacques Sebag, Lynn G. Seppala, Ohad Shemmer, Joshua D. Simon, M. Sivertz, Howard A. Smith, J. Allyn Smith, Nathan Smith, Anna H. Spitz, Adam Stanford, Keivan G. Stassun, Jay Strader, Michael A. Strauss, Christopher W. Stubbs, Donald W. Sweeney, Alex Szalay, Paula Szkody, Masahiro Takada, Paul Thorman, David E. Trilling, Virginia Trimble, Anthony Tyson, Richard Van Berg, Daniel Vand en Berk, Jake VanderPlas, Licia Verde, Bojan Vrsnak, Lucianne M. Walkowicz, Benjamin D. Wand elt, Sheng Wang, Yun Wang, Michael Warner, Risa H. Wechsler, Andrew A. West, Oliver Wiecha, Benjamin F. Williams, Beth Willman, David Wittman, Sidney C. Wolff, W. Michael Wood-Vasey, Przemek Wozniak, Patrick Young, Andrew Zentner, and Hu Zhan. LSST Science Book, Version 2.0. *arXiv e-prints*, page arXiv:0912.0201, Dec 2009.

[109] A. D. Ludlow, J. F. Navarro, M. Boylan-Kolchin, P. E. Bett, R. E. Angulo, M. Li, S. D. M. White, C. Frenk, and V. Springel. The mass profile and accretion history of cold dark matter haloes. *Monthly Notices of the RAS*, 432:1103–1113, June 2013.

[110] Aaron D. Ludlow, Sownak Bose, Raúl E. Angulo, Lan Wang, Wojciech A. Hellwing, Julio F. Navarro, Shaun Cole, and Carlos S. Frenk. The mass-concentration-redshift relation of cold and warm dark matter haloes. *Monthly Notices of the RAS*, 460(2):1214–1232, Aug 2016.

[111] Aaron D. Ludlow, Julio F. Navarro, Raúl E. Angulo, Michael Boylan-Kolchin, Volker Springel, Carlos Frenk, and Simon D. M. White. The mass-concentration-redshift relation of cold dark matter haloes. *Monthly Notices of the RAS*, 441(1):378–388, Jun 2014.

[112] Aaron D. Ludlow, Julio F. Navarro, Ming Li, Raul E. Angulo, Michael Boylan-Kolchin, and Philip E. Bett. The dynamical state and mass-concentration relation of galaxy clusters. *Monthly Notices of the RAS*, 427(2):1322–1328, December 2012.

[113] Andrea V. Macciò, Aaron A. Dutton, and Frank C. van den Bosch. Concentration, spin and shape of dark matter haloes as a function of the cosmological model: WMAP1, WMAP3 and WMAP5 results. *Monthly Notices of the RAS*, 391(4):1940–1954, Dec 2008.

[114] R. Mandelbaum, U. Seljak, R. J. Cool, M. Blanton, C. M. Hirata, and J. Brinkmann. Density profiles of galaxy groups and clusters from SDSS galaxy-galaxy weak lensing. *Monthly Notices of the RAS*, 372:758–776, October 2006.

[115] R. Mandelbaum, U. Seljak, G. Kauffmann, C. M. Hirata, and J. Brinkmann. Galaxy halo masses and satellite fractions from galaxy-galaxy lensing in the Sloan Digital Sky Survey: stellar mass, luminosity, morphology and environment dependencies. *Monthly Notices of the RAS*, 368:715–731, May 2006.

[116] Rachel Mandelbaum, Uroš Seljak, and Christopher M. Hirata. A halo mass—concentration relation from weak lensing. *Journal of Cosmology and Astroparticle Physics*, 2008(8):006, August 2008.

[117] Yao-Yuan Mao, Marc Williamson, and Risa H. Wechsler. The Dependence of Subhalo Abundance on Halo Concentration. *Astrophysical Journal*, 810:21, September 2015.

[118] Yao-Yuan Mao, Andrew R. Zentner, and Risa H. Wechsler. Beyond assembly bias: exploring secondary halo biases for cluster-size haloes. *Monthly Notices of the RAS*, 474(4):5143–5157, Mar 2018.

[119] James McBride, Onsi Fakhouri, and Chung-Pei Ma. Mass accretion rates and histories of dark matter haloes. *Monthly Notices of the RAS*, 398(4):1858–1868, Oct 2009.

[120] Kevin Spencer McCarthy, Zheng Zheng, and Hong Guo. The Effects of Galaxy Assembly Bias on the Inference of Growth Rate from Redshift-Space Distortions. *arXiv e-prints*, page arXiv:1810.05183, Oct 2018.

[121] Joseph E. McEwen and David H. Weinberg. The effects of assembly bias on the inference of matter clustering from galaxy-galaxy lensing and galaxy clustering. *Monthly Notices of the RAS*, 477:4348–4361, Jul 2018.

[122] J. Merten, M. Meneghetti, M. Postman, K. Umetsu, A. Zitrin, E. Medezinski, M. Nonino, A. Koekemoer, P. Melchior, D. Gruen, L. A. Moustakas, M. Bartelmann, O. Host, M. Donahue, D. Coe, A. Molino, S. Jouvel, A. Monna, S. Seitz, N. Czakon, D. Lemze, J. Sayers, I. Balestra, P. Rosati, N. Benítez, A. Biviano, R. Bouwens, L. Bradley, T. Broadhurst, M. Carrasco, H. Ford, C. Grillo, L. Infante, D. Kelson, O. Lahav, R. Massey, J. Moustakas, E. Rasia, J. Rhodes, J. Vega, and W. Zheng. CLASH: The Concentration-Mass Relation of Galaxy Clusters. *Astrophysical Journal*, 806(1):4, June 2015.

[123] H. J. Mo and S. D. M. White. An analytic model for the spatial clustering of dark matter haloes. *Monthly Notices of the RAS*, 282:347–361, September 1996.

[124] Houjun Mo, Frank C. van den Bosch, and Simon White. *Galaxy Formation and Evolution*. Cambridge University Press, 2010.

[125] B. Moore, F. Governato, T. Quinn, J. Stadel, and G. Lake. Resolving the Structure of Cold Dark Matter Halos. *Astrophysical Journal, Letters*, 499(1):L5–L8, May 1998.

[126] S. More, B. Diemer, and A. V. Kravtsov. The Splashback Radius as a Physical Halo Boundary and the Growth of Halo Mass. *Astrophysical Journal*, 810:36, September 2015.

[127] Surhud More, Frank C. van den Bosch, Marcello Cacciato, H. J. Mo, Xiaohu Yang, and Ran Li. Satellite kinematics - II. The halo mass-luminosity relation of central galaxies in SDSS. *Monthly Notices of the RAS*, 392:801–816, January 2009.

[128] B. P. Moster, R. S. Somerville, C. Maulbetsch, F. C. van den Bosch, A. V. Macciò, T. Naab, and L. Oser. Constraints on the Relationship between Stellar Mass and Halo Mass at Low and High Redshift. *Astrophysical Journal*, 710:903–923, February 2010.

[129] J. F. Navarro, A. Ludlow, V. Springel, J. Wang, M. Vogelsberger, S. D. M. White, A. Jenkins, C. S. Frenk, and A. Helmi. The diversity and similarity of simulated cold dark matter haloes. *Monthly Notices of the RAS*, 402:21–34, February 2010.

[130] Julio F. Navarro, Carlos S. Frenk, and Simon D. M. White. The assembly of galaxies in a hierarchically clustering universe. *Monthly Notices of the RAS*, 275(1):56–66, Jul 1995.

[131] Julio F. Navarro, Carlos S. Frenk, and Simon D. M. White. The Structure of Cold Dark Matter Halos. *Astrophysical Journal*, 462:563, May 1996.

[132] Julio F. Navarro, Carlos S. Frenk, and Simon D. M. White. A Universal Density Profile from Hierarchical Clustering. *Astrophysical Journal*, 490(2):493–508, Dec 1997.

[133] Angelo F. Neto, Liang Gao, Philip Bett, Shaun Cole, Julio F. Navarro, Carlos S. Frenk, Simon D. M. White, Volker Springel, and Adrian Jenkins. The statistics of Λ CDM halo concentrations. *Monthly Notices of the RAS*, 381(4):1450–1462, Nov 2007.

[134] J. A. Newman, M. C. Cooper, M. Davis, S. M. Faber, A. L. Coil, P. Guhathakurta, D. C. Koo, A. C. Phillips, C. Conroy, A. A. Dutton, D. P. Finkbeiner, B. F. Gerke, D. J. Rosario, B. J. Weiner, C. N. A. Willmer, R. Yan, J. J. Harker, S. A. Kassin, N. P. Konidaris, K. Lai, D. S. Madgwick, K. G. Noeske, G. D. Wirth, A. J. Connolly, N. Kaiser, E. N. Kirby, B. C. Lemaux, L. Lin, J. M. Lotz, G. A. Luppino, C. Marinoni, D. J. Matthews, A. Metevier, and R. P. Schiavon. The DEEP2 Galaxy Redshift Survey: Design, Observations, Data Reduction, and Redshifts. *ArXiv e-prints*, March 2012.

[135] Yueying Ni, Yin Li, Patrick Lachance, Rupert A. C. Croft, Tiziana Di Matteo, Simeon Bird, and Yu Feng. AI-assisted super-resolution cosmological simulations II: Halo substructures, velocities and higher order statistics. *arXiv e-prints*, page arXiv:2105.01016, May 2021.

[136] Jr. Oemler, Augustus. The Systematic Properties of Clusters of Galaxies. Photometry of 15 Clusters. *Astrophysical Journal*, 194:1–20, Nov 1974.

[137] Masamune Oguri and Yen-Ting Lin. Inferring Host Dark Matter Halo Masses of Individual Galaxies from Neighboring Galaxy Counts. *Astrophysical Journal*, 801:94, Mar 2015.

[138] Ana Laura O'Mill, Fernanda Duplancic, Diego García Lambas, Carlos Valotto, and Laerte Sodré. Galaxy triplets in Sloan Digital Sky Survey Data Release 7 - I. Catalogue. *Monthly Notices of the RAS*, 421(3):1897–1907, Apr 2012.

[139] P. J. E. Peebles. *The large-scale structure of the universe*. Princeton university press, 1980.

[140] R. Penrose. A generalized inverse for matrices. *Mathematical Proceedings of the Cambridge Philosophical Society*, 51(3):406–413, 1955.

[141] Planck Collaboration, P. A. R. Ade, N. Aghanim, C. Armitage-Caplan, M. Arnaud, M. Ashdown, F. Atrio-Barand ela, J. Aumont, C. Baccigalupi, A. J. Banday, and et al. Planck 2013 results. XVI. Cosmological parameters. *Astronomy and Astrophysics*, 571:A16, Nov 2014.

[142] Planck Collaboration, N. Aghanim, Y. Akrami, M. Ashdown, J. Aumont, C. Baccigalupi, M. Ballardini, A. J. Banday, R. B. Barreiro, N. Bartolo, S. Basak, R. Battye, K. Benabed, J. P. Bernard, M. Bersanelli, P. Bielewicz, J. J. Bock, J. R. Bond, J. Borrill, F. R. Bouchet, F. Boulanger, M. Bucher, C. Burigana, R. C. Butler, E. Calabrese, J. F. Cardoso, J. Carron, A. Challinor, H. C. Chiang, J. Chluba, L. P. L. Colombo, C. Combet, D. Contreras, B. P. Crill, F. Cuttaia, P. de Bernardis, G. de Zotti, J. Delabrouille, J. M. Delouis, E. Di Valentino, J. M. Diego, O. Doré, M. Douspis, A. Ducout, X. Dupac, S. Dusini, G. Efstathiou, F. Elsner, T. A. Enßlin, H. K. Eriksen, Y. Fantaye, M. Farhang, J. Fergusson, R. Fernandez-Cobos, F. Finelli, F. Forastieri, M. Frailis, A. A. Fraisse, E. Franceschi, A. Frolov, S. Galeotta, S. Galli, K. Ganga, R. T. Génova-Santos, M. Gerbino, T. Ghosh, J. González-Nuevo, K. M. Górski, S. Gratton, A. Gruppuso, J. E. Gudmundsson, J. Hamann, W. Handley, F. K. Hansen, D. Herranz, S. R. Hildebrandt, E. Hivon, Z. Huang, A. H. Jaffe, W. C. Jones, A. Karakci, E. Keihänen, R. Keskitalo, K. Kiiveri, J. Kim, T. S. Kisner, L. Knox, N. Krachmalnicoff, M. Kunz, H. Kurki-Suonio, G. Lagache, J. M. Lamarre, A. Lasenby, M. Lattanzi, C. R. Lawrence, M. Le Jeune, P. Lemos, J. Lesgourgues, F. Levrier, A. Lewis, M. Liguori, P. B. Lilje, M. Lilley, V. Lindholm, M. López-Caniego, P. M. Lubin, Y. Z. Ma, J. F. Macías-Pérez, G. Maggio, D. Maino, N. Mandolesi, A. Mangilli, A. Marcos-Caballero, M. Maris, P. G. Martin, M. Martinelli, E. Martínez-González, S. Matarrese, N. Mauri, J. D. McEwen, P. R. Meinhold, A. Melchiorri, A. Mennella, M. Migliaccio, M. Millea, S. Mitra, M. A. Miville-Deschênes, D. Molinari, L. Montier, G. Morgante, A. Moss, P. Natoli, H. U. Nørgaard-Nielsen, L. Pagano, D. Paoletti, B. Partridge, G. Patanchon, H. V. Peiris, F. Perrotta, V. Pettorino, F. Piacentini, L. Polastri, G. Polenta, J. L. Puget, J. P. Rachen, M. Reinecke, M. Remazeilles, A. Renzi, G. Rocha, C. Rosset, G. Roudier, J. A. Rubiño-Martín, B. Ruiz-Granados, L. Salvati, M. Sandri, M. Savelainen, D. Scott, E. P. S. Shellard, C. Sirignano, G. Sirri, L. D. Spencer, R. Sunyaev, A. S. Suur-Uski, J. A. Tauber, D. Tavagnacco, M. Tenti, L. Toffolatti, M. Tomasi, T. Trombetti, L. Valenziano, J. Valiviita, B. Van Tent, L. Vibert, P. Vielva, F. Villa, N. Vittorio, B. D. Wandelt, I. K. Wehus, M. White, S. D. M. White, A. Zacchei, and A. Zonca. Planck 2018 results. VI. Cosmological parameters. *Astronomy and Astrophysics*, 641:A6, September 2020.

[143] Planck Collaboration, Y. Akrami, F. Arroja, M. Ashdown, J. Aumont, C. Baccigalupi, M. Ballardini, A. J. Banday, R. B. Barreiro, N. Bartolo, S. Basak, R. Battye, K. Benabed, J. P. Bernard, M. Bersanelli, P. Bielewicz, J. J. Bock, J. R. Bond, J. Borrill, F. R. Bouchet, F. Boulanger, M. Bucher, C. Burigana, R. C. Butler, E. Calabrese, J. F. Cardoso, J. Carron, B. Casaponsa, A. Challinor, H. C. Chiang, L. P. L. Colombo, C. Combet, D. Contreras, B. P. Crill, F. Cuttaia, P. de Bernardis, G. de Zotti, J. Delabrouille, J. M. Delouis, F. X. Désert, E. Di Valentino, C. Dickinson, J. M. Diego, S. Donzelli, O. Doré, M. Douspis, A. Ducout, X. Dupac, G. Efstathiou, F. Elsner, T. A. Enßlin, H. K. Eriksen, E. Falgarone, Y. Fantaye, J. Fergusson, R. Fernandez-Cobos, F. Finelli, F. Forastieri, M. Frailis, E. Franceschi, A. Frolov, S. Galeotta, S. Galli, K. Ganga, R. T. Génova-Santos, M. Gerbino, T. Ghosh, J. González-Nuevo, K. M. Górski, S. Gratton, A. Gruppuso, J. E. Gudmundsson, J. Hamann, W. Hand ley, F. K. Hansen, G. Helou, D. Herranz, E. Hivon, Z. Huang, A. H. Jaffe, W. C. Jones, A. Karakci, E. Keihänen, R. Keskitalo, K. Kiiveri, J. Kim, T. S. Kisner, L. Knox, N. Krachmalnicoff, M. Kunz, H. Kurki-Suonio, G. Lagache, J. M. Lamarre,

M. Langer, A. Lasenby, M. Lattanzi, C. R. Lawrence, M. Le Jeune, J. P. Leahy, J. Lesgourgues, F. Levrier, A. Lewis, M. Liguori, P. B. Lilje, M. Lilley, V. Lindholm, M. López-Caniego, P. M. Lubin, Y. Z. Ma, J. F. Macías-Pérez, G. Maggio, D. Maino, N. Mand olesi, A. Mangilli, A. Marcos-Caballero, M. Maris, P. G. Martin, E. Martínez-González, S. Matarrese, N. Mauri, J. D. McEwen, P. D. Meerburg, P. R. Meinhold, A. Melchiorri, A. Mennella, M. Migliaccio, M. Millea, S. Mitra, M. A. Miville-Deschênes, D. Molinari, A. Moneti, L. Montier, G. Morgante, A. Moss, S. Mottet, M. Münchmeyer, P. Natoli, H. U. Nørgaard-Nielsen, C. A. Oxborrow, L. Pagano, D. Paoletti, B. Partridge, G. Patanchon, T. J. Pearson, M. Peel, H. V. Peiris, F. Perrotta, V. Pettorino, F. Piacentini, L. Polastri, G. Polenta, J. L. Puget, J. P. Rachen, M. Reinecke, M. Remazeilles, A. Renzi, G. Rocha, C. Rosset, G. Roudier, J. A. Rubiño-Martín, B. Ruiz-Granados, L. Salvati, M. Sandri, M. Savelainen, D. Scott, E. P. S. Shellard, M. Shiraishi, C. Sirignano, G. Sirri, L. D. Spencer, R. Sunyaev, A. S. Suur-Uski, J. A. Tauber, D. Tavagnacco, M. Tenti, L. Terenzi, L. Toffolatti, M. Tomasi, T. Trombetti, J. Valiviita, B. Van Tent, L. Vibert, P. Vielva, F. Villa, N. Vittorio, B. D. Wandelt, I. K. Wehus, M. White, S. D. M. White, A. Zacchei, and A. Zonca. Planck 2018 results. I. Overview and the cosmological legacy of Planck. *arXiv e-prints*, page arXiv:1807.06205, Jul 2018.

[144] M. Postman and M. J. Geller. The morphology-density relation - The group connection. *Astrophysical Journal*, 281:95–99, Jun 1984.

[145] Douglas Potter, Joachim Stadel, and Romain Teyssier. PKDGRAV3: beyond trillion particle cosmological simulations for the next era of galaxy surveys. *Computational Astrophysics and Cosmology*, 4(1):2, May 2017.

[146] Francisco Prada, Anatoly A. Klypin, Antonio J. Cuesta, Juan E. Betancort-Rijo, and Joel Primack. Halo concentrations in the standard Λ cold dark matter cosmology. *Monthly Notices of the RAS*, 423(4):3018–3030, Jul 2012.

[147] William H. Press and Paul Schechter. Formation of Galaxies and Clusters of Galaxies by Self-Similar Gravitational Condensation. *Astrophysical Journal*, 187:425–438, Feb 1974.

[148] M. H. Quenouille. Notes on bias in estimation. *Biometrika*, 43(3/4):353–360, 1956.

[149] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2016.

[150] Rachel M. Reddick, Risa H. Wechsler, Jeremy L. Tinker, and Peter S. Behroozi. The Connection between Galaxies and Dark Matter Structures in the Local Universe. *Astrophysical Journal*, 771:30, Jul 2013.

[151] B. A. Reid, H.-J. Seo, A. Leauthaud, J. L. Tinker, and M. White. A 2.5 per cent measurement of the growth rate from small-scale redshift space clustering of SDSS-III CMASS galaxies. *Monthly Notices of the RAS*, 444:476–502, October 2014.

[152] Beth A. Reid and David N. Spergel. Constraining the Luminous Red Galaxy Halo Occupation Distribution Using Counts-In-Cylinders. *Astrophysical Journal*, 698:143–154, Jun 2009.

[153] Martin P. Rey, Andrew Pontzen, and Amélie Saintonge. Sensitivity of dark matter haloes to their accretion histories. *Monthly Notices of the RAS*, 485(2):1906–1915, May 2019.

[154] Aldo Rodríguez-Puebla, Peter Behroozi, Joel Primack, Anatoly Klypin, Christoph Lee, and Doug Hellinger. Halo and subhalo demographics with Planck cosmological parameters: Bolshoi-Planck and MultiDark-Planck simulations. *Monthly Notices of the RAS*, 462(1):893–916, Oct 2016.

[155] Emilio Romano-Díaz, Enrico Garaldi, Mikolaj Borzyszkowski, and Cristiano Porciani. ZOMG - II. Does the halo assembly history influence central galaxies and gas accretion? *Monthly Notices of the RAS*, 469(2):1809–1823, Aug 2017.

[156] Nina Roth, Andrew Pontzen, and Hiranya V. Peiris. Genetically modified haloes: towards controlled experiments in ΛCDM galaxy formation. *Monthly Notices of the RAS*, 455(1):974–986, January 2016.

[157] Douglas H. Rudd, Andrew R. Zentner, and Andrey V. Kravtsov. Effects of Baryons and Dissipation on the Matter Power Spectrum. *Astrophysical Journal*, 672:19–32, Jan 2008.

[158] Shun Saito, Alexie Leauthaud, Andrew P. Hearin, Kevin Bundy, Andrew R. Zentner, Peter S. Behroozi, Beth A. Reid, Manodeep Sinha, Jean Coupon, Jeremy L. Tinker, Martin White, and Donald P. Schneider. Connecting massive galaxies to dark matter haloes in BOSS - I. Is galaxy colour a stochastic process in high-mass haloes? *Monthly Notices of the RAS*, 460:1457–1475, Aug 2016.

[159] Eduard Salvador-Solé, José María Solanes, and Alberto Manrique. Merger versus Accretion and the Structure of Dark Matter Halos. *Astrophysical Journal*, 499(2):542–547, May 1998.

[160] Joop Schaye, Robert A. Crain, Richard G. Bower, Michelle Furlong, Matthieu Schaller, Tom Theuns, Claudio Dalla Vecchia, Carlos S. Frenk, I. G. McCarthy, John C. Helly, Adrian Jenkins, Y. M. Rosas-Guevara, Simon D. M. White, Maarten Baes, C. M. Booth, Peter Camps, Julio F. Navarro, Yan Qu, Alireza Rahmati, Till Sawala, Peter A. Thomas, and

James Trayford. The EAGLE project: simulating the evolution and assembly of galaxies and their environments. *Monthly Notices of the RAS*, 446(1):521–554, January 2015.

[161] G. Schwarz. Estimating the Dimension of a Model. *Ann. Statist.*, 6(2):461–464, 03 1978.

[162] U. Seljak. Analytic model for galaxy and dark matter clustering. *Monthly Notices of the RAS*, 318:203–213, October 2000.

[163] U. Seljak, A. Makarov, R. Mandelbaum, C. M. Hirata, N. Padmanabhan, P. McDonald, M. R. Blanton, M. Tegmark, N. A. Bahcall, and J. Brinkmann. SDSS galaxy bias from halo mass-bias relation and its cosmological implications. *PRD*, 71(4):043511–+, February 2005.

[164] Uros Seljak, Anze Slosar, and Patrick McDonald. Cosmological parameters from combining the Lyman-α forest with CMB, galaxy clustering and SN constraints. *Journal of Cosmology and Astro-Particle Physics*, 2006:014, Oct 2006.

[165] R. K. Sheth, A. Diaferio, L. Hui, and R. Scoccimarro. On the streaming motions of haloes and galaxies. *Monthly Notices of the RAS*, 326:463–472, September 2001.

[166] R. K. Sheth, H. J. Mo, and G. Tormen. Ellipsoidal collapse and an improved model for the number and spatial distribution of dark matter haloes. *Monthly Notices of the RAS*, 323:1–12, May 2001.

[167] Manodeep Sinha. Corrfunc: Corrfunc-1.1.0. *http://dx.doi.org/10.5281/zenodo.55161*, June 2016.

[168] Manodeep Sinha, Andreas A. Berlind, Cameron K. McBride, Roman Scoccimarro, Jennifer A. Piscionere, and Benjamin D. Wibking. Towards accurate modelling of galaxy clustering on small scales: testing the standard ΛCDM + halo model. *Monthly Notices of the RAS*, 478:1042–1064, Jul 2018.

[169] Samuel W. Skillman, Michael S. Warren, Matthew J. Turk, Risa H. Wechsler, Daniel E. Holz, and P. M. Sutter. Dark Sky Simulations: Early Data Release. *arXiv e-prints*, page arXiv:1407.2600, Jul 2014.

[170] D. Spergel, N. Gehrels, C. Baltay, D. Bennett, J. Breckinridge, M. Donahue, A. Dressler, B. S. Gaudi, T. Greene, O. Guyon, C. Hirata, J. Kalirai, N. J. Kasdin, B. Macintosh, W. Moos, S. Perlmutter, M. Postman, B. Rauscher, J. Rhodes, Y. Wang, D. Weinberg, D. Benford, M. Hudson, W. S. Jeong, Y. Mellier, W. Traub, T. Yamada, P. Capak, J. Colbert,

D. Masters, M. Penny, D. Savransky, D. Stern, N. Zimmerman, R. Barry, L. Bartusek, K. Carpenter, E. Cheng, D. Content, F. Dekens, R. Demers, K. Grady, C. Jackson, G. Kuan, J. Kruk, M. Melton, B. Nemati, B. Parvin, I. Poberezhskiy, C. Peddie, J. Ruffa, J. K. Wallace, A. Whipple, E. Wollack, and F. Zhao. Wide-Field InfrarRed Survey Telescope-Astrophysics Focused Telescope Assets WFIRST-AFTA 2015 Report. *arXiv e-prints*, page arXiv:1503.03757, Mar 2015.

[171] V. Springel, S. D. M. White, A. Jenkins, C. S. Frenk, N. Yoshida, L. Gao, J. Navarro, R. Thacker, D. Croton, J. Helly, J. A. Peacock, S. Cole, P. Thomas, H. Couchman, A. Evrard, J. Colberg, and F. Pearce. Simulations of the formation, evolution and clustering of galaxies and quasars. *Nature*, 435:629–636, June 2005.

[172] T. Sunayama, A. P. Hearin, N. Padmanabhan, and A. Leauthaud. The scale-dependence of halo assembly bias. *Monthly Notices of the RAS*, 458:1510–1516, May 2016.

[173] Istvan Szapudi, Avery Meiksin, and Robert C. Nichol. Higher Order Statistics from the Edinburgh/Durham Southern Galaxy Catalogue Survey. I. Counts in Cells. *Astrophysical Journal*, 473:15, December 1996.

[174] T. Tal, D. A. Wake, P. G. van Dokkum, F. C. van den Bosch, D. P. Schneider, J. Brinkmann, and B. A. Weaver. Mass Growth and Mergers: Direct Observations of the Luminosity Function of LRG Satellite Galaxies Out to z = 0.7 from SDSS and BOSS Images. *Astrophysical Journal*, 746:138, February 2012.

[175] Argyro Tasitsiomi, Andrey V. Kravtsov, Stefan Gottlöber, and Anatoly A. Klypin. Density Profiles of ΛCDM Clusters. *Astrophysical Journal*, 607(1):125–139, May 2004.

[176] The Dark Energy Survey Collaboration. The Dark Energy Survey. *arXiv e-prints*, pages astro–ph/0510346, Oct 2005.

[177] J. Tinker, A. V. Kravtsov, A. Klypin, K. Abazajian, M. Warren, G. Yepes, S. Gottlöber, and D. E. Holz. Toward a Halo Mass Function for Precision Cosmology: The Limits of Universality. *Astrophysical Journal*, 688:709–728, December 2008.

[178] J. L. Tinker, C. Conroy, P. Norberg, S. G. Patiri, D. H. Weinberg, and M. S. Warren. Void Statistics in Large Galaxy Redshift Surveys: Does Halo Occupation of Field Galaxies Depend on Environment? *Astrophysical Journal*, 686:53–71, October 2008.

[179] J. L. Tinker, A. Leauthaud, K. Bundy, M. R. George, P. Behroozi, R. Massey, J. Rhodes, and R. Wechsler. Evolution of the Stellar-to-Dark Matter Relation: Separating Star-Forming and Passive Galaxies from z=1 to 0. *ArXiv:1308.2974*, August 2013.

[180] J. L. Tinker, B. E. Robertson, A. V. Kravtsov, A. Klypin, M. S. Warren, G. Yepes, and S. Gottlöber. The Large-scale Bias of Dark Matter Halos: Numerical Calibration and Model Tests. *Astrophysical Journal*, 724:878–886, December 2010.

[181] S. Trujillo-Gomez, A. Klypin, J. Primack, and A. J. Romanowsky. Galaxies in ΛCDM with Halo Abundance Matching: Luminosity-Velocity Relation, Baryonic Mass-Velocity Relation, Velocity Function, and Clustering. *Astrophysical Journal*, 742:16, November 2011.

[182] John W. Tukey. Abstracts of papers. *The Annals of Mathematical Statistics*, 29(2):614–623, 1958.

[183] Mohammadjavad Vakili and ChangHoon Hahn. How Are Galaxies Assigned to Halos? Searching for Assembly Bias in the SDSS Galaxy Clustering. *Astrophysical Journal*, 872(1):115, Feb 2019.

[184] A. Vale and J. P. Ostriker. Linking halo mass to galaxy luminosity. *Monthly Notices of the RAS*, 353:189–200, September 2004.

[185] F. C. van den Bosch, D. Aquino, X. Yang, H. J. Mo, A. Pasquali, D. H. McIntosh, S. M. Weinmann, and X. Kang. The importance of satellite quenching for the build-up of the red sequence of present-day galaxies. *Monthly Notices of the RAS*, 387:79–91, June 2008.

[186] F. C. van den Bosch, S. More, M. Cacciato, H. Mo, and X. Yang. Cosmological constraints from a combination of galaxy clustering and lensing - I. Theoretical framework. *Monthly Notices of the RAS*, 430:725–746, April 2013.

[187] F. C. van den Bosch, X. Yang, and H. J. Mo. Linking early- and late-type galaxies to their dark matter haloes. *Monthly Notices of the RAS*, 340:771–792, April 2003.

[188] F. C. van den Bosch, X. Yang, H. J. Mo, S. M. Weinmann, A. V. Macciò, S. More, M. Cacciato, R. Skibba, and X. Kang. Towards a concordant model of halo occupation statistics. *Monthly Notices of the RAS*, 376:841–860, April 2007.

[189] Frank C. van den Bosch. The universal mass accretion history of cold dark matter haloes. *Monthly Notices of the RAS*, 331(1):98–110, Mar 2002.

[190] Frank C. van den Bosch. Dissecting the evolution of dark matter subhaloes in the Bolshoi simulation. *Monthly Notices of the RAS*, 468(1):885–909, June 2017.

[191] Frank C. van den Bosch, Fangzhou Jiang, Andrew Hearin, Duncan Campbell, Douglas Watson, and Nikhil Padmanabhan. Coming of age in the dark sector: how dark matter haloes grow their gravitational potential wells. *Monthly Notices of the RAS*, 445(2):1713–1730, December 2014.

[192] Frank C. van den Bosch, Xiaohu Yang, H. J. Mo, and Peder Norberg. The abundance and radial distribution of satellite galaxies. *Monthly Notices of the RAS*, 356:1233–1248, February 2005.

[193] R. P. van der Marel, J. Magorrian, R. G. Carlberg, H. K. C. Yee, and E. Ellingson. The Velocity and Mass Distribution of Clusters of Galaxies from the CNOC1 Cluster Redshift Survey. *Astronomical Journal*, 119:2038–2052, May 2000.

[194] E. van Uitert, M. Cacciato, H. Hoekstra, M. Brouwer, C. Sifón, M. Viola, I. Baldry, J. Bland-Hawthorn, S. Brough, M. J. I. Brown, A. Choi, S. P. Driver, T. Erben, C. Heymans, H. Hildebrandt, B. Joachimi, K. Kuijken, J. Liske, J. Loveday, J. McFarland, L. Miller, R. Nakajima, J. Peacock, M. Radovich, A. S. G. Robotham, P. Schneider, G. Sikkema, E. N. Taylor, and G. Verdoes Kleijn. The stellar-to-halo mass relation of GAMA galaxies from 100 deg^2 of KiDS weak lensing data. *Monthly Notices of the RAS*, 459:3251–3270, July 2016.

[195] Antonio S. Villarreal, Andrew R. Zentner, Yao-Yuan Mao, Chris W. Purcell, Frank C. van den Bosch, Benedikt Diemer, Johannes U. Lange, Kuan Wang, and Duncan Campbell. The immitigable nature of assembly bias: the impact of halo definition on assembly bias. *Monthly Notices of the RAS*, 472(1):1088–1105, Nov 2017.

[196] Mark Vogelsberger, Shy Genel, Volker Springel, Paul Torrey, Debora Sijacki, Dandan Xu, Greg Snyder, Dylan Nelson, and Lars Hernquist. Introducing the Illustris Project: simulating the coevolution of dark and visible matter in the Universe. *Monthly Notices of the RAS*, 444(2):1518–1547, October 2014.

[197] Mark Vogelsberger, Federico Marinacci, Paul Torrey, and Ewald Puchwein. Cosmological simulations of galaxy formation. *Nature Reviews Physics*, 2(1):42–66, January 2020.

[198] J. V. Wall and C. R. Jenkins. *Practical Statistics for Astronomers*. Cambridge Observing Handbooks for Research Astronomers. Cambridge University Press, 2003.

[199] J. Wang, J. F. Navarro, C. S. Frenk, S. D. M. White, V. Springel, A. Jenkins, A. Helmi, A. Ludlow, and M. Vogelsberger. Assembly history and structure of galactic cold dark matter haloes. *Monthly Notices of the RAS*, 413(2):1373–1382, May 2011.

[200] Xiao-Feng Wang. *fANCOVA: Nonparametric Analysis of Covariance*, 2010. R package version 0.5-1.

[201] Michael S. Warren. 2HOT: An Improved Parallel Hashed Oct-Tree N-Body Algorithm for Cosmological Simulation. *arXiv e-prints*, page arXiv:1310.4502, Oct 2013.

[202] D. F. Watson, A. A. Berlind, C. K. McBride, D. W. Hogg, and T. Jiang. The Extreme Small Scales: Do Satellite Galaxies Trace Dark Matter? *Astrophysical Journal*, 749:83, April 2012.

[203] R. H. Wechsler and J. L. Tinker. The Connection Between Galaxies and Their Dark Matter Halos. *Annual Review of Astron and Astrophys*, 56:435–487, September 2018.

[204] Risa H. Wechsler, James S. Bullock, Joel R. Primack, Andrey V. Kravtsov, and Avishai Dekel. Concentrations of Dark Halos from Their Assembly Histories. *Astrophysical Journal*, 568(1):52–70, Mar 2002.

[205] Risa H. Wechsler, Andrew R. Zentner, James S. Bullock, Andrey V. Kravtsov, and Brandon Allgood. The Dependence of Halo Clustering on Halo Formation History, Concentration, and Occupation. *Astrophysical Journal*, 652(1):71–84, Nov 2006.

[206] S. M. Weinmann, F. C. van den Bosch, X. Yang, and H. J. Mo. Properties of galaxy groups in the Sloan Digital Sky Survey - I. The dependence of colour, star formation and morphology on halo mass. *Monthly Notices of the RAS*, 366:2–28, February 2006.

[207] A. Weyant, C. Schafer, and W. M. Wood-Vasey. Likelihood-free Cosmological Inference with Type Ia Supernovae: Approximate Bayesian Computation for a Complete Treatment of Uncertainty. *Astrophysical Journal*, 764:116, February 2013.

[208] S. D. M. White and M. J. Rees. Core condensation in heavy halos: a two-stage theory for galaxy formation and clustering. *Monthly Notices of the RAS*, 183:341–358, May 1978.

[209] Benjamin D. Wibking, Andrés N. Salcedo, David H. Weinberg, Lehman H. Garrison, Douglas Ferrer, Jeremy Tinker, Daniel Eisenstein, Marc Metchnik, and Philip Pinto. Emulating galaxy clustering and galaxy-galaxy lensing into the deeply non-linear regime: methodology, information, and forecasts. *Monthly Notices of the RAS*, 484:989–1006, Mar 2019.

[210] H.-Y. Wu, E. Rozo, and R. H. Wechsler. The Effects of Halo Assembly Bias on Self-Calibration in Galaxy Cluster Surveys. *Astrophysical Journal*, 688:729–741, December 2008.

[211] Hao-Yi Wu, Oliver Hahn, Risa H. Wechsler, Yao-Yuan Mao, and Peter S. Behroozi. Rhapsody. I. Structural Properties and Formation History from a Statistical Sample of Resimulated Cluster-size Halos. *Astrophysical Journal*, 763(2):70, Feb 2013.

[212] Haojie Xu, Zheng Zheng, Hong Guo, Ying Zu, Idit Zehavi, and David H. Weinberg. The conditional colour-magnitude distribution - I. A comprehensive model of the colour-magnitude-halo mass distribution of present-day galaxies. *Monthly Notices of the RAS*, 481(4):5470–5500, Dec 2018.

[213] Xiaoju Xu and Zheng Zheng. Dependence of halo bias and kinematics on assembly variables. *Monthly Notices of the RAS*, 479(2):1579–1594, Sep 2018.

[214] Xiaoju Xu and Zheng Zheng. Galaxy assembly bias of central galaxies in the Illustris simulation. *arXiv e-prints*, page arXiv:1812.11210, Dec 2018.

[215] X. Yang, H. J. Mo, Y. P. Jing, F. C. van den Bosch, and Y. Chu. Populating dark matter haloes with galaxies: comparing the 2dFGRS with mock galaxy redshift surveys. *Monthly Notices of the RAS*, 350:1153–1173, June 2004.

[216] X. Yang, H. J. Mo, and F. C. van den Bosch. Constraining galaxy formation and cosmology with the conditional luminosity function of galaxies. *Monthly Notices of the RAS*, 339:1057–1080, March 2003.

[217] S. Yuan, D. J. Eisenstein, and L. H. Garrison. Using galaxy pairs to investigate the three-point correlation function in the squeezed limit. *Monthly Notices of the RAS*, 472:577–590, November 2017.

[218] S. Yuan, D. J. Eisenstein, and L. H. Garrison. Exploring the squeezed three-point galaxy correlation function with generalized halo occupation distribution models. *Monthly Notices of the RAS*, 478:2019–2033, August 2018.

[219] Sihan Yuan, Boryana Hadzhiyska, Sownak Bose, Daniel J. Eisenstein, and Hong Guo. Evidence for galaxy assembly bias in BOSS CMASS redshift-space galaxy correlation function. *Monthly Notices of the RAS*, 502(3):3582–3598, April 2021.

[220] I. Zehavi et al. The Luminosity and Color Dependence of the Galaxy Correlation Function. *Astrophysical Journal*, 630:1–27, September 2005.

[221] I. Zehavi et al. Galaxy Clustering in the Completed SDSS Redshift Survey: The Dependence on Color and Luminosity. *Astrophysical Journal*, 736:59, July 2011.

[222] A. R. Zentner. The Excursion Set Theory of Halo Mass Functions, Halo Clustering, and Halo Growth. *International Journal of Modern Physics D*, 16:763–815, 2007.

[223] A. R. Zentner, A. Hearin, F. C. van den Bosch, J. U. Lange, and A. Villarreal. Constraints on Assembly Bias from Galaxy Clustering. *MNRAS Submitted*, June 2016.

[224] A. R. Zentner, A. P. Hearin, and F. C. van den Bosch. Galaxy assembly bias: a significant source of systematic error in the galaxy-halo relationship. *Monthly Notices of the RAS*, 443:3044–3067, October 2014.

[225] Andrew R. Zentner, Andreas A. Berlind, James S. Bullock, Andrey V. Kravtsov, and Risa H. Wechsler. The Physics of Galaxy Clustering. I. A Model for Subhalo Populations. *Astrophysical Journal*, 624:505–525, May 2005.

[226] Zhongxu Zhai, Jeremy L. Tinker, Matthew R. Becker, Joseph DeRose, Yao-Yuan Mao, Thomas McClintock, Sean McLaughlin, Eduardo Rozo, and Risa H. Wechsler. The Aemulus Project III: Emulation of the Galaxy Correlation Function. *arXiv e-prints*, page arXiv:1804.05867, Apr 2018.

[227] D. H. Zhao, Y. P. Jing, H. J. Mo, and G. Börner. Accurate Universal Models for the Mass Accretion Histories and Concentrations of Dark Matter Halos. *Astrophysical Journal*, 707(1):354–369, Dec 2009.

[228] D. H. Zhao, H. J. Mo, Y. P. Jing, and G. Börner. The growth and structure of dark matter haloes. *Monthly Notices of the RAS*, 339(1):12–24, Feb 2003.

[229] Z. Zheng, A. A. Berlind, D. H. Weinberg, A. J. Benson, C. M. Baugh, S. Cole, R. Davé, C. S. Frenk, N. Katz, and C. G. Lacey. Theoretical Models of the Halo Occupation Distribution: Separating Central and Satellite Galaxies. *Astrophysical Journal*, 633:791–809, November 2005.

[230] Z. Zheng, A. L. Coil, and I. Zehavi. Galaxy Evolution from Halo Occupation Distribution Modeling of DEEP2 and SDSS Galaxy Clustering. *Astrophysical Journal*, 667:760–779, October 2007.

[231] Z. Zheng, I. Zehavi, D. J. Eisenstein, D. H. Weinberg, and Y. P. Jing. Halo Occupation Distribution Modeling of Clustering of Luminous Red Galaxies. *Astrophysical Journal*, 707:554–572, December 2009.

[232] Rongpu Zhou, Jeffrey A. Newman, Yao-Yuan Mao, Aaron Meisner, John Moustakas, Adam D. Myers, Abhishek Prakash, Andrew R. Zentner, David Brooks, Yutong Duan, Martin Landriau, Michael E. Levi, Francisco Prada, and Gregory Tarle. The clustering of DESI-like luminous red galaxies using photometric redshifts. *Monthly Notices of the RAS*, 501(3):3309–3331, March 2021.

[233] Fritz Zwicky. Die rotverschiebung von extragalaktischen nebeln. *Helvetica physica acta*, 6:110–127, 1933.