

Measuring the distribution of galaxies between haloes

Andrew J. Benson[★]

California Institute of Technology, MC 105-24, Pasadena, CA 91125, USA

Accepted 2001 February 23. Received 2001 January 30; in original form 2000 December 8

ABSTRACT

We develop a method to measure the probability, $P(N; M)$, of finding N galaxies in a dark matter halo of mass M from the theoretically determined clustering properties of dark matter haloes and the observationally measured clustering properties of galaxies. Knowledge of this function and the distribution of the dark matter completely specifies all clustering properties of galaxies on scales larger than the size of dark matter haloes. Furthermore, $P(N; M)$ provides strong constraints on models of galaxy formation, since it depends upon the merger history of dark matter haloes and the galaxy–galaxy merger rate within haloes. We show that measurements from a combination of the Two Micron All Sky Survey and Sloan Digital Sky Survey or Two-degree Field Galaxy Redshift Survey data sets will allow $P(N; M)$ averaged over haloes occupied by bright galaxies to be accurately measured for $N = 0–2$.

Key words: galaxies: haloes – dark matter – large-scale structure of Universe.

1 INTRODUCTION

Recent work on the clustering properties of galaxies has focused on the connection between galaxies and dark matter haloes, using theoretical models of galaxy formation (Kauffmann, Nusser & Steinmetz 1997; Diaferio et al. 1999; Kauffmann et al. 1999a,b; Benson et al. 2000a,b; Seljak 2000; Somerville et al. 2001) or observational data (Peacock & Smith 2000) to determine the number of galaxies that reside within haloes of given mass. Models of this type have been successful in explaining the near-power-law nature of the galaxy–galaxy correlation function (Kauffmann et al. 1999a; Benson et al. 2000a), and the strong clustering of Lyman-break galaxies at $z \approx 3$ (Baugh et al. 1998; Governato et al. 1998; Wechsler et al. 2000).

In this context, Benson et al. (2000a) calculated the quantity $P(N; M)$, the probability of finding N galaxies brighter than a specified luminosity, L_0 , in a dark matter halo of mass M , from the galaxy formation model of Cole et al. (2000). This quantity is particularly powerful since, if the population of galaxies residing in a halo is determined by the halo mass alone, then $P(N; M)$ fully determines all clustering properties of galaxies on scales larger than the size of dark matter haloes once a model for the distribution of dark matter haloes is chosen (on smaller scales the spatial distribution of galaxies within individual haloes becomes important). As such, $P(N; M)$ may be thought of as a complete description of the galaxy–dark matter bias including any non-linearity and stochasticity. Furthermore, if $P(N; M)$ can be

measured from the observed clustering pattern of galaxies it provides a direct and powerful constraint for models of galaxy formation since it is sensitive to the merger history of dark matter haloes, and to the rate of galaxy–galaxy mergers within haloes.

In this paper we describe how $P(N; M)$ may be measured directly from a volume-limited galaxy redshift survey by using a counts-in-cells analysis to determine the probability of finding N galaxies in a cell, $S(N)$. The remainder of this paper is laid out as follows. In Section 2 we describe our method and give the formulae relating $P(N; M)$ and $S(N)$ for all N . In Section 3 we investigate how well $P(N; M)$ can be measured from a combination of the Two Micron All Sky Survey (Skrutskie et al. 1995; 2MASS) and Two-degree Field Galaxy Redshift Survey (Dalton 2000; 2dFGRS) or Sloan Digital Sky Survey (Blanton et al. 2000; SDSS) data sets using the mock galaxy catalogues of Benson et al. (2000a), and finally in Section 4 we present our conclusions.

2 METHOD

We will assume that the galaxy population of a dark matter halo is determined only by the mass of that halo. While it is the distribution of halo masses that varies most significantly as a function of environment (Lemson & Kauffmann 1999), other quantities are also known to correlate with environment, for example the concentration of the halo (Bullock et al. 2001). In practice the properties of galaxies may depend upon such variables thereby altering the clustering properties of the galaxies. (If the properties of the galaxy population depend on some variable that is uncorrelated with environment it will have no effect on the clustering properties of galaxies.) In principle, other variables could be included in our analysis by defining a function $P(N; M, x_1,$

[★]E-mail: abenson@astro.caltech.edu

..., x_n), where x_1, \dots, x_n represents other variables upon which the properties of galaxies may depend. However, current data sets are insufficient to allow meaningful measurements of such a function to be made and so we will restrict ourselves to considering $P(N; M)$ only at present.

As we will derive clustering statistics of dark matter haloes from N -body simulations (this being the most accurate way to determine the statistics we are interested in) we must also assume that $P(0; M) = 1$ for all $M < M_0$; i.e. haloes below mass M_0 never contain any galaxies brighter than L_0 , since we can only measure the clustering properties of haloes above a certain mass (set by the resolution of the simulation). This is a reasonable assumption – if haloes of arbitrarily low mass could host bright galaxies then, since there are an infinite number of haloes per unit volume (at least according to the Press–Schechter theory) there would be an infinite number of galaxies per unit volume. Having made this assumption we can ignore haloes of mass less than M_0 as they make no contribution to the galaxy population that we are considering.

Ideally, M_0 should be as large as possible to permit the best estimates of the dark matter halo clustering statistics to be made (a small value would require a high-resolution simulation with necessarily small volume, resulting in poor statistics for the more massive haloes). How may we determine M_0 for a given galaxy population? One approach would be to make use of dynamical mass estimates (e.g. Vogt et al. 1997). However, these are not available for all types of galaxy. An alternative method is to use galaxy samples selected at near-infrared wavelengths from which we can infer a stellar mass from the sample magnitude limit (Kauffmann & Charlot 1998). Then

$$M_0 \geq \frac{\Omega_0}{\Omega_b} M_*, \quad (1)$$

where M_* is the stellar mass. This lower limit on M_0 corresponds to the case where the entire gaseous mass of a halo is turned into stars. The halo must have at least this mass to make the observed galaxy. The conversion from K -band light to stellar mass is uncertain by a factor of approximately two (Brinchmann & Ellis 2000), so M_0 should realistically be taken to be two to three times lower than the value inferred from equation (1).

While any clustering statistic can be written in terms of $P(N; M)$ and the clustering properties of dark matter haloes [for example, the two-point correlation function expressed in terms of $P(N; M)$ is given in Appendix A], a particularly simple relation can be found for $S(N)$, the probability of finding N galaxies brighter than L_0 in a cell of given size and shape. While these statistics can in principle reveal $P(N; M)$ for any N and for a range of M , in practice measurement is severely limited by unavoidable noise in the data as will be shown in Section 3. Nevertheless, useful constraints can still be obtained from this analysis. In the remainder of this section we develop the relations necessary to determine $P(N; M)$ for all N and M , but will only make use of the simplest forms of these relations in Section 3.

The probability of finding N galaxies in a cell of given size and geometry can be expressed in terms of the probability of finding a certain combination of haloes in that cell and the probabilities of finding different numbers of galaxies in each of those haloes. In order to measure $P(N; M)$ it is necessary to divide haloes into a number of mass ranges, or bins. We will then refer to the mean value of $P(N; M)$ averaged over all haloes in mass bin i as $P_i(N)$, such that $P_i(N)$ is the probability of finding N galaxies in a halo

selected at random from mass bin i , i.e.

$$P_i(N) = \frac{\int_{M_i}^{M_{i+1}} P(N; M) \frac{dn}{dM}(M) dM}{\int_{M_i}^{M_{i+1}} \frac{dn}{dM}(M) dM}, \quad (2)$$

where M_i is the lower bound of the i th mass bin and dn/dM is the mass function of dark matter haloes.

Let $S(N)$ be the probability of finding N galaxies in a cell (of given size and geometry). For a particular choice of cosmology and dark matter let $Q(N_1, N_2, \dots, N_n)$ be the probability of finding the centres of N_1 haloes in mass bin 1, N_2 in bin 2 etc. in a cell, where we have used a total of n mass bins. (We take the centre of mass to define the halo centre.) Note that in general $Q(N_1, N_2, \dots, N_n) \neq Q(N_1)Q(N_2)\dots Q(N_n)$ since the distribution of haloes is typically correlated. Note that $S(N)$ is an observationally measurable quantity, and $Q(N_1, N_2, \dots, N_n)$ can be obtained from a structure formation model. As we show below, these two quantities are related, and that relation depends upon $P(N; M)$. Measurement of $S(N)$ therefore allows us to measure $P(N; M)$.

We can write $S(N)$ as the sum over all possible combinations of N_1, N_2, \dots, N_n of $Q(N_1, N_2, \dots, N_n)$ multiplied by the probability of finding i_1 galaxies in the first halo, i_2 in the second etc. summed over all combinations of i_1, i_2, \dots that satisfy the constraint $\sum_j i_j = N$ (i.e. only those combinations that produce the correct number of galaxies in the cell contribute to the total probability).

For example, $S(0)$ is given by

$$S(0) = \sum_{N_1=0}^{\infty} \sum_{N_2=0}^{\infty} \dots \sum_{N_n=0}^{\infty} Q(N_1, N_2, \dots, N_n) \prod_{j=1}^n P_j^{N_j}(0), \quad (3)$$

while $S(1)$ and $S(2)$ are given by

$$S(1) = \sum_{N_1=0}^{\infty} \sum_{N_2=0}^{\infty} \dots \sum_{N_n=0}^{\infty} Q(N_1, N_2, \dots, N_n) \times \sum_{i_1=1}^n C(N_1^{(0)}, N_1^{(1)}, N_2^{(0)}, N_2^{(1)}) \frac{P_{i_1}(1)}{P_{i_1}(0)} \prod_{j=1}^n P_j^{N_j}(0) \quad (4)$$

and

$$S(2) = \sum_{N_1=0}^{\infty} \sum_{N_2=0}^{\infty} \dots \sum_{N_n=0}^{\infty} Q(N_1, N_2, \dots, N_n) \times \left[\sum_{i_1=1}^n C(N_1^{(0)}, \dots, N_1^{(2)}, N_2^{(0)}, \dots, N_2^{(2)}) \frac{P_{i_1}(2)}{P_{i_1}(0)} + \sum_{i_1=1}^n \sum_{i_2=1}^n C(N_1^{(0)}, N_1^{(1)}, N_2^{(0)}, N_2^{(1)}) \frac{P_{i_1}(1) P_{i_2}(1)}{P_{i_1}(0) P_{i_2}(0)} \right] \prod_{j=1}^n P_j^{N_j}(0), \quad (5)$$

where $N_i^{(j)}$ is the number of times a halo in mass bin i is populated by j galaxies and $C(N_i^{(0)}, N_i^{(1)}, \dots)$ is the number of distinct permutations of each term that contribute to the probability. The weighting factor $C(N_i^{(0)}, N_i^{(1)}, \dots)$ is the number of ways to populate the available haloes with the galaxies divided by the number of times such terms appear in the

summation. In general,

$$\begin{aligned}
 C(N_{i_1}^{(0)}, \dots, N_{i_1}^{(k)}, \dots, N_{i_n}^{(0)}, \dots, N_{i_n}^{(k)}) \\
 &= \prod_{l=1}^n \frac{N_{i_l}!}{\prod_{j=0}^k N_{i_l}^{(j)}!} \bigg/ \prod_{m=1}^k \frac{[\sum_{l=1}^n N_{i_l}^{(m)}]!}{\prod_{l=1}^n N_{i_l}^{(m)}!} \\
 &= \prod_{l=1}^n \frac{N_{i_l}!}{N_{i_l}^{(0)}! [\sum_{m=1}^k N_{i_l}^{(m)}]!}, \quad (6)
 \end{aligned}$$

where $N_i = \sum_{j=0}^{\infty} N_i^{(j)}$. For example,

$$C(N_{i_1}^{(0)}, N_{i_1}^{(1)}) = \frac{[N_{i_1}^{(0)} + N_{i_1}^{(1)}]!}{N_{i_1}^{(0)}! N_{i_1}^{(1)}!}. \quad (7)$$

As expected, $S(N)$ depends only upon those $P_i(j)$ for which $j \leq N$. Therefore, we may begin by finding the $P_i(0)$ s using the expression for $S(0)$, then proceed to find the $P_i(1)$ s using the expression for $S(1)$ and the previously calculated $P_i(0)$ and so on. Each expression therefore involves n unknowns [for $S(N)$ these are the $P_i(N)$], and so we must have a measure of $S(N)$ for at least n different cell sizes to solve the equations. While the above equations cannot be solved analytically for the $P_i(N)$, solutions can be found relatively simply using the method of Powell (Press et al. 1992) to minimize the quantity $\chi^2 = \sum_i \{[S_i^{(\text{obs})}(N) - S_i^{(\text{model})}(N)]/\Delta S_i^{\text{obs}}(N)\}^2$ for example, where the sum is taken over all cell sizes considered.

In general, the expression for $S(N)$ will be of the form

$$\begin{aligned}
 S(N) &= \sum_{N_1=0}^{\infty} \sum_{N_2=0}^{\infty} \dots \sum_{N_n=0}^{\infty} Q(N_1, N_2, \dots, N_n) \\
 &\times \left[\sum_{i_1=1}^n C(N_{i_1}^{(0)}, \dots, N_{i_1}^{(N)}) \frac{P_{i_1}(N)}{P_{i_1}(0)} \right. \\
 &+ \sum_{i_1=1}^n \sum_{i_2=1}^n [C(N_{i_1}^{(0)}, \dots, N_{i_1}^{(N-1)}, N_{i_2}^{(0)}, \dots, N_{i_2}^{(N-1)}) \\
 &\times \frac{P_{i_1}(N-1) P_{i_2}(1)}{P_{i_1}(0) P_{i_2}(0)} + C(N_{i_1}^{(0)}, \dots, N_{i_1}^{(N-2)}, N_{i_2}^{(0)}, \dots, N_{i_2}^{(N-2)}) \\
 &\times \frac{P_{i_1}(N-2) P_{i_2}(2)}{P_{i_1}(0) P_{i_2}(0)} + 3 \text{ halo terms} + 4 \text{ halo terms} \\
 &\left. + \dots + N \text{ halo terms} \right] \prod_{j=1}^n P_j^{N_j}(0), \quad (8)
 \end{aligned}$$

where the expression ‘ N halo terms’ refers to all terms corresponding to galaxies shared between N different haloes (i.e. the first two sums in the above expression are therefore ‘1 halo terms’ and ‘2 halo terms’).

At this point it is instructive to consider briefly the assumptions made in obtaining the above relations. First, we have assumed that all galaxies lie at the centre of the halo they occupy. Then, a halo being in a cell guarantees that any galaxies it contains are also in the cell. In reality galaxies are likely to be spread throughout the halo with some unknown spatial distribution, and so some galaxies may lie outside of the cell even though their halo centre is inside (and conversely some galaxies may lie inside even though their halo centre is outside). While our analysis could be extended to account for such ‘edge effects’ this would require us to assume a distribution for galaxies within individual haloes. We prefer to concentrate on scales where these effects are negligible. In

Section 3 we demonstrate that edge effects are an insignificant source of error.

Secondly, we assume that the galaxy occupancy of all haloes in a mass bin is well described by a single set of $P_i(N)$. Providing $P(N; M)$ varies little across the mass bin, this is a reasonable assumption. However, as we will see in Section 3, noisy data may limit us to considering a single mass bin, extending from M_0 to infinity, for which the above assumption is unlikely to hold true. While we implicitly assume that the $P_i(N)$ are independent of the number of haloes found in a cell, the Press–Schechter (Press & Schechter 1974) formalism tells us that high-density regions of the Universe will contain preferentially higher mass haloes than low-density regions. Consequently cells that contain many haloes will preferentially contain high-mass haloes, while in cells containing few haloes the haloes are likely to be of low mass. If, for example, $P(0; M)$ is a decreasing function of M then cells with few haloes (which are typically the most abundant) will contain zero galaxies more often than our model assumes. The resulting increase in $S(0)$ can be seen in the synthetic data sets used in Section 3. While this has a non-negligible effect on $S(0)$, particularly for large cell sizes, the value of $P(0)$ recovered is quite insensitive to this since most of the signal comes from small cell sizes.

3 APPLICATION TO SYNTHETIC DATA SETS

Perhaps the most suitable data set to which this technique can be applied will be a combination of the 2MASS survey with a large-redshift survey (e.g. the 2dFGRS or the SDSS). The 2MASS survey provides near-infrared photometry which allows M_0 to be estimated, but must be complemented by a redshift survey in order to provide a 3D map of the galaxy distribution.¹ A volume-limited 2MASS sample of galaxies brighter than $M_K - 5 \log h = -23.5$ would have a volume of order $3 \times 10^6 h^{-3} \text{ Mpc}^3$ in the 2dFGRS survey area (or around four times this volume in the SDSS survey area). As this is very similar to the volume of the GIF Λ CDM N -body simulation used by Benson et al. (2000a), we will use their synthetic galaxy catalogues to estimate how well $P(N; M)$ could be recovered from such a data set. We do not attempt here to reproduce the full details of the survey geometry or selection function, but merely consider a synthetic data set with comparable volume and number density of galaxies in order to estimate the accuracy with which $P(N; M)$ may be recovered from such a survey.

We consider only galaxies brighter than $M_K - 5 \log h = -23.5$ to ensure that we need only consider haloes that are well resolved by the GIF simulation. These galaxies live in haloes with masses greater than $10^{12} h^{-1} M_{\odot}$ in this model (the particle mass in the GIF Λ CDM simulation is $1.4 \times 10^{10} h^{-1} M_{\odot}$). Inferring M_0 from the K -band magnitude of the galaxies we find $M_0 = 8 \times 10^{11} h^{-1} M_{\odot}$. We therefore conservatively set $M_0 = 4 \times 10^{11} h^{-1} M_{\odot}$. We consider only one bin of halo mass, i.e. all haloes more massive than $4 \times 10^{11} h^{-1} M_{\odot}$. While this technique can in principle be applied to several halo mass bins, we find that in practice this is very difficult. Typically the values of $P_i(N)$ for the more massive bins are poorly constrained since there are very few haloes in the mass range, or else the solutions of equation (8) for different cell sizes are degenerate in the $P_i(N)$ values and so only allow certain combinations of $P_i(N)$ values to be accurately measured. Very large data sets, with correspondingly small errors,

¹ While $P(N; M)$ could be measured from a 2D data set, the 3D information will provide a much stronger constraint.

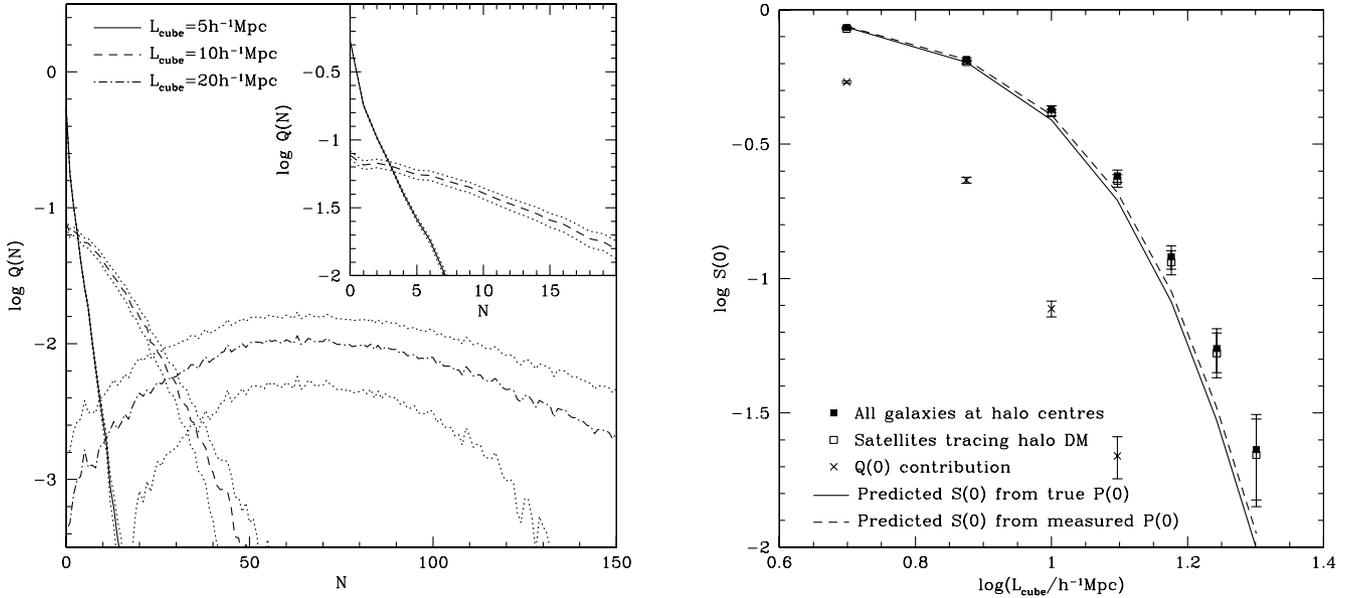


Figure 1. Left-hand panel: the probability of finding N haloes more massive than $4 \times 10^{11} h^{-1} M_{\odot}$ in cubes of sides 5 (solid line), 10 (dashed line) and $20 h^{-1} \text{Mpc}$ (dot-dashed line) in redshift space. Dotted lines indicate errors on these quantities assuming Poisson statistics. The inset shows an expanded view of the low- N region. Right-hand panel: the probability of finding zero galaxies brighter than $M_K - 5 \log h = -23.5$ in cubic cells of side L_{cube} in redshift space in the simulations of Benson et al. (2000a). Solid squares show the result when all galaxies are placed at the centres of dark matter haloes, while open squares indicate the result when satellite galaxies are made to trace the dark matter in their halo. Errors are calculated assuming Poisson statistics. The solid line shows $S(0)$ calculated from the measured $Q(N)$ (as shown in the left-hand panel) and the value of $P(0)$ measured directly from the models of Benson et al. (2000a). Crosses with errorbars show the contribution of $Q(0)$ to $S(0)$. The $Q(0)$ contribution is around 65 per cent for $L_{\text{cube}} = 5 h^{-1} \text{Mpc}$, and falls for larger values of L_{cube} .

may allow a measurement of $P_i(N)$ in more than one mass bin, although this will probably require a treatment of edge effects which must eventually become the dominant source of error.

The distribution of halo masses within this bin has a mean, defined as

$$\bar{M} = \frac{\int_{M_0}^{\infty} M \frac{dn}{dM} dM}{\int_{M_0}^{\infty} \frac{dn}{dM} dM},$$

of $\bar{M} = 3.4 \times 10^{12} h^{-1} M_{\odot}$, and a dispersion, defined as

$$\sigma^2 = \frac{\int_{M_0}^{\infty} (M - \bar{M})^2 \frac{dn}{dM} dM}{\int_{M_0}^{\infty} \frac{dn}{dM} dM},$$

of $\sigma = 4.5$. This particular bin may therefore be expected to probe the contents of galactic-sized haloes (as opposed to groups or clusters). However, it should be noted that for a bin spanning such a wide range of masses, computing a galaxy number weighted mean halo mass,

$$\bar{M}_g = \frac{\int_{M_0}^{\infty} M \bar{N}(M) \frac{dn}{dM} dM}{\int_{M_0}^{\infty} \frac{dn}{dM} dM},$$

where $\bar{N}(M)$ is the mean number of galaxies per halo of mass M , can give a very different result. In the case of the synthetic data sets used here we find $\bar{M}_g = 3.8 \times 10^{13} h^{-1} M_{\odot}$.

We measure $S(N)$ in cubic cells of side $L_{\text{cube}} = 5.0, 7.5, 10.0, 12.5, 15.0, 17.5$ and $20.0 h^{-1} \text{Mpc}$, and measure $Q(N)$ for the same

cell sizes. Both $S(N)$ and $Q(N)$ are calculated for galaxy/halo positions in redshift space. For smaller cubes, edge effects begin to become a significant source of error, while for larger cubes the GIF simulation contains very few independent volumes.

The left-hand panel of Fig. 1 shows $Q(N)$ for $L_{\text{cube}} = 5, 10$ and $20 h^{-1} \text{Mpc}$, while the right-hand panel shows $S(0)$ (squares) and $Q(0)$ (crosses) as functions of L_{cube} . Errors are estimated assuming Poisson statistics and that there are $(L_{\text{GIF}}/L_{\text{cube}})^3$ independent volumes in the simulation, where $L_{\text{GIF}} = 141.3 h^{-1} \text{Mpc}$ is the size of the GIF ΛCDM simulation volume. This is known to underestimate the true errors (e.g. Kim & Strauss 1998), but is sufficient for our present purposes. Note that placing all galaxies at the halo centre (solid squares), or placing one galaxy at the centre and making satellite galaxies trace the dark matter of their halo (open squares) has little effect on the measured $S(0)$, i.e. edge effects are unimportant for this sample. For the smallest cells we consider that $Q(0)$ accounts for around 65 per cent of the value of $S(0)$, and makes a smaller contribution for the larger cells. Also shown is the value of $S(0)$ predicted by equation (8) with the recovered value of $P_1(0)$ (dashed line) and the true value of $P_1(0)$ (solid line). For the larger cell sizes neither of these gives a good fit to the mock data points. This is due to the failure of our assumption that $P_1(N; M)$ is roughly constant throughout the mass bin (as discussed in Section 2). However, as we discuss below this does not drastically alter the recovered values of $P_1(N)$.

We determine $P_1(N)$ from the measured $S(N)$ and $Q(N)$ by solving equation (8) for $P_1(N)$ by minimizing χ^2 (as described in Section 2). Fig. 2 shows the true $P_1(N)$ as measured directly from the full model and from the GIF synthetic galaxy catalogue (solid squares and solid triangles respectively), with errorbars computed assuming Poisson statistics, and the $P_1(N)$ recovered from the synthetic galaxy catalogue via the $S(N)$ values with all galaxies at their halo centre (open triangles) and with satellite galaxies tracing

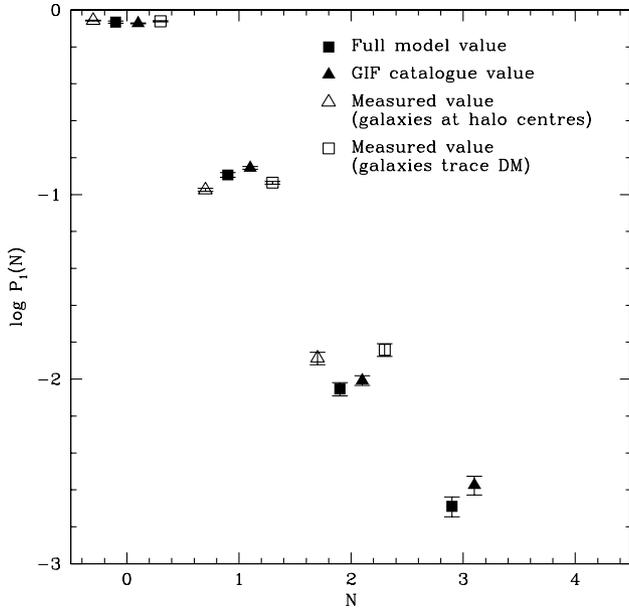


Figure 2. $P_1(N)$ for haloes more massive than $4 \times 10^{11} h^{-1} \text{Mpc}$ and galaxies brighter than $M_K - 5 \log h = -23.5$. Solid squares show $P_1(N)$ taken directly from the full model of Benson et al. (2000a), while solid triangles show those taken directly from the GIF synthetic galaxy catalogue of Benson et al. (2000a). Open triangles show the $P_1(N)$ recovered from the GIF synthetic galaxy catalogue via determinations of $S(N)$ when all galaxies are placed at the centre of their halo (error bars are computed assuming Poisson statistics), while open squares show the result when satellite galaxies are made to trace the dark matter content of their halo (errorbars computed from $\Delta\chi^2$). The points are offset slightly in N for clarity.

the dark matter of their halo (open squares), with errorbars estimated from $\Delta\chi^2$. The first three $P_1(N)$ are recovered with reasonable accuracy from the synthetic galaxy catalogues. [For $N = 0, 1$ and 2 the recovered $P_1(N)$ differ from the true values by 3, 17 and 46 per cent, respectively, although we caution that these values are from a single realization of the synthetic galaxy catalogue and so may not be representative.]

A weakness of this approach is that the equation for $P_i(N)$ depends upon all $P_i(N')$ where $N' < N$. Hence, any errors in the determination of $P_i(0)$ affect the estimate of $P_i(1)$ etc. In the case of the synthetic galaxy catalogues used here we can recover $P_1(N)$ accurately for $N = 0, 1, 2$. When we consider $P_1(3)$, however, we find that the contribution to $S(3)$ from terms involving only $P_1(0)$, $P_1(1)$ and $P_1(2)$ already exceeds the measured value. Thus the solution to the equation requires that $P_1(3)$ be negative, which is of course impossible. Thus with a data set of this size only the first few $P_1(N)$ can be measured.

4 DISCUSSION

We have described how the distribution of galaxies amongst haloes, as described by the function $P(N; M)$ (the probability of finding N galaxies brighter than a specified luminosity L_0 in a halo of mass M), can be measured directly from a galaxy redshift survey once a model for the spatial distribution of dark matter haloes is assumed. Specifically we derive relations between the observationally measurable quantity $S(N)$ (the probability of finding N galaxies in a cell) and the theoretically determinable quantity $Q(N_1, N_2, \dots, N_n)$ (the probability of finding different numbers of dark matter haloes in a cell). These relations depend upon $P(N; M)$,

thereby allowing $P(N; M)$ to be determined from observational determinations of $S(N)$ and a model of structure formation.

The distribution function $P(N; M)$ provides a complete description of galaxy bias (at least on scales larger than the size of haloes) in terms of physically meaningful quantities, and will also be sensitive to the merging history of dark matter haloes and the rate of galaxy–galaxy mergers within dark matter haloes. We have presented the technique in its simplest form. We defer a more detailed study of errors (including edge effects) and the limitations imposed by the simplifying assumptions made to a future paper.

Our approach assumes a model for the underlying distribution of dark matter haloes, and the results obtained will therefore be dependent on that model. Measurements of key cosmological parameters, perhaps from measurements of the cosmic microwave background (CMB) (Jungman et al. 1996; Bind, Efstathiou & Tegmark 1997), and the dark matter power spectrum, from weak lensing (e.g. Tyson, Wittman & Angel 2001) or Lyman α forest studies (Croft et al. 1998), in the near future should allow the halo distribution to be fully determined.

Using the mock galaxy catalogues produced by Benson et al. (2000a) we have shown that $P(N; M)$ averaged over all haloes more massive than $4 \times 10^{11} h^{-1} M_\odot$ is measurable for the first few values of N from a combination of the 2MASS data set with a redshift survey such as the SDSS or 2dFGRS. To measure $P(N; M)$ for higher N or as a function of M would require larger data sets and a detailed consideration of edge effects. Measurement of this quantity from forthcoming galaxy redshift surveys will therefore provide strong constraints for models of galaxy formation and clustering, and reveal a great deal about the connection between galaxies and dark matter.

ACKNOWLEDGMENTS

We would like to thank Marc Kamionkowski for a careful reading of this work, Carlton Baugh, Shaun Cole, Carlos Frenk and Cedric Lacey for making available results from their galaxy formation model, Simon White for useful discussions, and the Virgo Consortium for making available the GIF simulations used in this work. We would also like to thank the anonymous referee for helpful suggestions.

REFERENCES

- Baugh C. M., Cole S., Frenk C. S., Lacey C. G., 1998, *ApJ*, 498, 504
- Benson A. J., Cole S., Frenk C. S., Baugh C. M., Lacey C. G., 2000a, *MNRAS*, 311, 793
- Benson A. J., Baugh C. M., Cole S., Frenk C. S., Lacey C. G., 2000b, *MNRAS*, 316, 107
- Blanton M. et al., 2000, *BAAS*, 196, 53.12
- Bond J. R., Efstathiou G., Tegmark M., 1997, *MNRAS*, 291, L33
- Brinchmann J., Ellis R. S., 2000, *ApJ*, 536, 77
- Bullock J. S., Kolatt T. S., Sigad Y., Somerville R. S., Kravtsov A. V., Klypin A. A., Primack J. R., Dekel A., 2001, *MNRAS*, 321, 559
- Cole S. M., Lacey C. G., Baugh C. M., Frenk C. S., 2000, *MNRAS*, 319, 168
- Croft R. A. C., Weinberg D. H., Katz N., Hernquist L., 1998, *ApJ*, 495, 44
- Dalton G. B. et al., 2000, *BAAS*, 196, 56.05
- Diaferio A., Kauffmann G., Colberg J. M., White S. D. M., 1999, *MNRAS*, 307, 537
- Governato F., Baugh C. M., Frenk C. S., Cole S., Lacey C. G., Quinn T., Stadel J., 1998, *Nat*, 392, 359
- Jungman G., Kamionkowski M., Kosowsky A., Spergel D. N., 1996, *Phys. Rev. D*, 54, 1332

- Kauffmann G., Nusser A., Steinmetz M., 1997, MNRAS, 286, 795
 Kauffmann G., Charlot S., 1998, MNRAS, 297, L23
 Kauffmann G., Colberg J. M., Diaferio A., White S. D. M., 1999a, MNRAS, 303, 188
 Kauffmann G., Colberg J. M., Diaferio A., White S. D. M., 1999b, MNRAS, 307, 529
 Kim R. S. J., Strauss M. A., 1998, ApJ, 493, 39
 Lemson G., Kauffmann G., 1999, MNRAS, 302, 111
 Peacock J. A., Smith R. E., 2000, MNRAS, 318, 1144
 Press W. H., Schechter P., 1974, ApJ, 187, 425
 Press W. H., Flannery B. P., Teukolsky S. A., Vetterling W. H., 1992, Numerical Recipes: The Art of Scientific Computing, 2nd edn. Cambridge Univ. Press, Cambridge
 Seljak U., 2000, MNRAS, 318, 203
 Skrutskie M. F. et al., 1995, BAAS, 187, 75.07
 Somerville R. S., Lemson G., Sigad Y., Dekel A., Kauffmann G., White S. D. M., 2001, MNRAS, 320, 289
 Tyson J. A., Wittman D., Angel J. R. P., 2001, in Cline D. B., ed., Sources and Detection of Dark Matter and Dark Energy in the Universe. Springer, Heidelberg (astro-ph/0005381)
 Vogt N. P. et al., 1997, ApJ, 479, 121
 Wechsler R. H., Somerville R. S., Bullock J. S., Kolatt T. S., Primack J. R., Blumenthal G. R., Dekel A., 2000, ApJ, submitted (astro-ph/0011261)

APPENDIX A: THE TWO-POINT CORRELATION FUNCTION OF GALAXIES

The two-point correlation function is a familiar clustering statistic easily expressed in terms of $P(N; M)$. Suppose there is a halo of mass M_1 to $M_1 + dM_1$ in a small volume element dV_1 . Let dQ_{12} be the probability of finding a halo of mass M_2 to $M_2 + dM_2$ in a small volume element dV_2 a distance r away from the first halo. We can write

$$dQ_{12}(r) = [1 + \xi_{12}(r)] \frac{dn}{dM}(M_1) \frac{dn}{dM}(M_2) dM_1 dM_2 dV_1 dV_2, \quad (\text{A1})$$

where $\xi_{12}(r)$ is the cross-correlation function of these haloes. A single halo pair may contribute many galaxy pairs. On average the above halo pair will contribute

$$dN_{12}(r) = \bar{N}(M_1)\bar{N}(M_2) dQ_{12}(r) \quad (\text{A2})$$

galaxy pairs, where $\bar{N}(M) = \sum_{i=0}^{\infty} iP(i; M)$ is the mean number of galaxies in a halo of mass M . For a random distribution of galaxies we would expect

$$dN_{12}^{(r)}(r) = \bar{N}(M_1)\bar{N}(M_2) \frac{dn}{dM}(M_1) \frac{dn}{dM}(M_2) dM_1 dM_2 dV_1 dV_2. \quad (\text{A3})$$

Integrating equations (A2) and (A3) over all halo masses we find the total number of galaxy pairs in the clustered and random cases

to be

$$dN_{\text{gg}}(r) = \int_{M_0}^{\infty} \int_{M_0}^{\infty} \bar{N}(M_1)\bar{N}(M_2)[1 + \xi_{12}(r)] \frac{dn}{dM}(M_1) \times \frac{dn}{dM}(M_2) dM_1 dM_2 dV_1 dV_2, \quad (\text{A4})$$

$$dN_{\text{gg}}^{(r)}(r) = \int_{M_0}^{\infty} \int_{M_0}^{\infty} \bar{N}(M_1)\bar{N}(M_2) \frac{dn}{dM}(M_1) \times \frac{dn}{dM}(M_2) dM_1 dM_2 dV_1 dV_2 = n_{\text{gal}}^2 dV_1 dV_2, \quad (\text{A5})$$

where n_{gal} is the mean number density of the galaxies. The galaxy–galaxy correlation function is defined to be

$$\xi_{\text{gg}}(r) = \frac{dN_{\text{gg}}(r)}{dN_{\text{gg}}^{(r)}(r)} - 1 = \int_{M_0}^{\infty} \int_{M_0}^{\infty} \xi_{12}(r) \frac{\bar{N}(M_1)\bar{N}(M_2)}{n_{\text{gal}}^2} \frac{dn}{dM}(M_1) \times \frac{dn}{dM}(M_2) dM_1 dM_2. \quad (\text{A6})$$

Equation (A6) is valid only on scales larger than the size of dark matter haloes since $P(N; M)$ does not specify the distribution of galaxies within an individual halo. It should be noted, however, that $P(N; M)$ does specify the mean number of galaxy pairs within a halo of mass M , $\bar{N}_{\text{p}}(M)$, since

$$\bar{N}_{\text{p}}(M) = \sum_{N=0}^{\infty} N(N-1)P(N; M). \quad (\text{A7})$$

Given $N_{\text{p}}(M)$ and an assumption about the shape of the galaxy density profile within a halo, the correlation function (and other statistics) can be predicted over all scales (see e.g. Seljak 2000). Determining \bar{N} and \bar{N}_{p} from a measured $P_i(N)$ is somewhat difficult due to the weights applied to each $P_i(N)$ [i.e. N and $N(N-1)$ for \bar{N} and \bar{N}_{p} , respectively]. These make the contributions from large values of N significant, while present data sets may only be able to recover $P_i(N)$ for relatively low N . For example, using the recovered values of $P_1(N)$ from the synthetic data sets of Section 3, \bar{N} and \bar{N}_{p} are underestimated by 25 and 70 per cent, respectively. This inability to measure $P_i(N)$ for large N is therefore the largest source of error in measuring \bar{N} and \bar{N}_{p} at present.

This paper has been typeset from a $\text{\TeX}/\text{\LaTeX}$ file prepared by the author.