Mon. Not. R. Astron. Soc. **392,** 1080–1091 (2009) doi:10.1111/j.1365-2966.2008.14007.x

A halo model of galaxy colours and clustering in the Sloan Digital Sky Survey

Ramin A. Skibba[1⋆] and Ravi K. Sheth[2⋆]

[1]*Max-Planck-Institute for Astronomy, Königstuhl 17, D-69117 Heidelberg, Germany*

[2]*Department of Physics & Astronomy, University of Pennsylvania, 209 S. 33rd Street, Philadelphia, PA 19130, USA*

Accepted 2008 September 24. Received 2008 September 6; in original form 2008 May 5

ABSTRACT

Successful halo-model descriptions of the luminosity dependence of clustering distinguish between the central galaxy in a halo and all the others (satellites). To include colours, we provide a prescription for how the colour–magnitude relation of centrals and satellites depends on halo mass. This follows from two assumptions: (i) the bimodality of the colour distribution at a fixed luminosity is independent of halo mass and (ii) the fraction of satellite galaxies which populate the red sequence increases with luminosity. We show that these two assumptions allow one to build a model of how galaxy clustering depends on colour without any additional free parameters than those required to model the luminosity dependence of galaxy clustering. We then show that the resulting model is in good agreement with the distribution and clustering of colours in the Sloan Digital Sky Survey, both by comparing the predicted correlation functions of red and blue galaxies with measurements and by comparing the predicted colour–mark correlation function with the measured one. Mark correlation functions are powerful tools for identifying and quantifying correlations between galaxy properties and their environments: our results indicate that the correlation between halo mass and environment is the primary driver for correlations between galaxy colours and the environment; additional correlations associated with halo 'assembly bias' are relatively small. Our approach shows explicitly how to construct mock catalogues which include both luminosities *and* colours – thus providing realistic training sets for, e.g., galaxy cluster-finding algorithms. Our prescription is the first step towards incorporating the entire spectral energy distribution into the halo model approach.

Key words: methods: analytical – methods: statistical – galaxies: clusters: general – galaxies: formation – galaxies: haloes – large-scale structure of the universe.

1 INTRODUCTION

The halo model is a useful language for discussing how galaxy clustering depends on galaxy type: galaxy bias (see Cooray & Sheth 2002 for a review). To date, the halo model has been used to provide a useful framework for modelling the luminosity dependence of galaxy clustering. The main goal of this paper is to extend the halo-model description of galaxy luminosities to include colours. This is an important step towards the ultimate goal of providing a description of how the properties of a galaxy, its morphology and spectral energy distribution are correlated with those of its neighbours. The hope is that, by relating such correlations between galaxies to the properties of their parent dark matter haloes, the halo model will prove a useful guide in the study of galaxy formation.

⋆E-mail: skibba@mpia.de (RAS); shethrk@physics.upenn.edu (RKS)

The halo-model description of the luminosity dependence of clustering is usually done in three rather different ways, which have come to be known as the 'halo occupation distribution' (HOD; Jing, Mo & Börner 1998; Benson et al. 2000; Seljak 2000; Scoccimarro et al. 2001; Berlind & Weinberg 2002; Zehavi et al. 2005), the 'conditional luminosity function' (CLF; Peacock & Smith 2000; Yang, Mo & van den Bosch 2003; Cooray 2006; van den Bosch et al. 2007) and the 'subhalo abundance matching' (SHAM; Klypin et al. 1999; Kravtsov et al. 2004; Conroy, Wechsler & Kravtsov 2006; Vale & Ostriker 2006) methods. The HOD approach uses the abundance and spatial distribution of a given galaxy population (typically, just the two-point clustering statistics) to determine how the number of galaxies depends on the mass of the parent halo. This is done by studying a sequence of volume-limited galaxy catalogues, each containing galaxies more luminous than some threshold luminosity. The CLF method attempts, instead, to match the observed luminosity function by specifying how the luminosity distribution in haloes changes as a function of halo mass. One can infer the CLF from the HOD approach, and vice versa, so the question arises as to which is the more efficient description. For a given catalogue, the HOD method requires the fitting of just two free parameters, so it is relatively straightforward. The CLF method requires many more parameters to be fit simultaneously, but uses fewer volume-limited catalogues. SHAMs first identify the subhaloes within virialized haloes in simulations, and then use subhalo properties to match the subhalo abundances to the observed distribution of luminosities. Once this has been done, CLFs or HODs can be measured in the simulations.

In SPH and semi-analytic galaxy formation models, central and satellite galaxies are rather different populations (e.g. Kauffmann et al. 1999; Sheth & Diaferio 2001; Guzik & Seljak 2002; Benson et al. 2003; Sheth 2005; Zheng et al. 2005). In the HOD and CLF approaches to the halo model, the central galaxy in a halo is assumed to be very different from all the others. For example, the CLF approach must provide a description of how the central and satellite luminosity functions vary as a function of halo mass. The HOD-based analyses predict that the satellite galaxy luminosity function should be approximately independent of halo mass, and hence of group and/or cluster properties (Skibba et al. 2006). Skibba, Sheth & Martino (2007) present evidence from the Sloan Digital Sky Survey (SDSS) in support of this prediction. More recent analysis of a rather different group catalogue has confirmed this finding (Hansen et al. 2007). Skibba et al. argued that this independence can reduce the required number of free parameters in CLF-based analyses.

One of the goals of the present work is to show that the HOD-based approach also provides a rather simple way to understand how galaxy clustering depends on colour. In essence, it provides a simple algorithm for specifying how the joint CLF (i.e. the luminosity distribution in two different bands) varies with halo mass. In principle, this can be done by splitting the sample up into small bins of luminosity *and* colour, and studying how the clustering signal in each bin changes. Zehavi et al. (2005) describe a first attempt at this – for each bin in luminosity, they use two bins in colour: 'red' or 'blue'. [Croton et al. (2007a) also study the difference in clustering strengths of red and blue galaxies. They use related statistics, but do not attempt a halo-model description of their measurements.] As sample sizes increase, it will become possible to split the sample into many more colour bins. However, even for this simplest case, Zehavi et al. were led to a rather more complex parametrization of the HOD than was necessary for the luminosities – they caution that, as a result, there are more degeneracies amongst their parameter choices, and so the constraints on the HODs they obtain are considerably weaker than for luminosities alone. While such a brute force approach to determining the HOD is certainly possible, we argue below that there may be some merit to recasting the problem as one in which the physics and statistics are more closely related.

In essence, our approach exploits the fact that, to a good approximation, galaxies appear to be bimodal in their properties (e.g. Blanton et al. 2005a). In the present context, we are interested in the fact that the distribution of colours at fixed luminosity is bimodal (e.g. Baldry et al. 2004; Willmer et al. 2006). Our approach is to couple this bimodality with the centre-satellite split in the halo model.

This paper is organized as follows. Section 2 describes our approach: it shows the correlation between colour and luminosity in the SDSS sample, and then describes a model for the luminosities and colours of centrals and satellites which is designed to reproduce this bimodality. Section 3 describes how to use our model to generate mock catalogues which have the correct luminosity dependence of clustering and the observed colour–magnitude relation, as well as how to incorporate our approach into a halo-model description of the colour–mark two-point correlation function. Section 4 provides a comparison of our model predictions with measurements from the SDSS. These include the clustering signal from 'red' and 'blue' galaxies (defined as being redder or bluer than a critical luminosity-dependent colour) and the clustering signal when galaxies are weighted by colour – the colour–mark correlation function. A final section summarizes our findings.

Throughout, the rest-frame magnitudes we quote are associated with SDSS filters shifted to $z = 0.1$; the absolute magnitude of the Sun in this r-band filter is 4.76 (Blanton et al. 2003). Where necessary, we assume a flat background cosmological model in which $\Omega_0 = 0.3$, the cosmological constant is $\Lambda_0 = 1 - \Omega_0$ and $\sigma_8 = 0.9$. We write the Hubble constant as $H_0 = 100\,h$ km s^{-1} Mpc^{-1}. In addition, we always use 'log' for the 10-based logarithm and 'ln' for the natural logarithm.

2 COLOUR–MAGNITUDE BIMODALITY AND THE CENTRE-SATELLITE SPLIT

2.1 Bimodality in the SDSS

Baldry et al. (2004) report that the distribution of rest-frame $u - r$ colour at a fixed r magnitude can be well modelled as the sum of two Gaussian components. The same is true of the distribution of rest-frame $g - r$ colour (e.g. Blanton et al. 2005a); we call these the red and blue components of the distribution $p(c|L)$. The mean and rms values of these components depend on luminosity. This dependence is quite well described by simple power laws:

$$\langle g - r|M_r\rangle_{\rm red} = 0.932 - 0.032\,(M_r + 20),$$
$$\mathrm{rms}(g - r|M_r)_{\rm red} = 0.07 + 0.01\,(M_r + 20); \tag{1}$$

$$\langle g - r|M_r\rangle_{\rm blue} = 0.62 - 0.11\,(M_r + 20),$$
$$\mathrm{rms}(g - r|M_r)_{\rm blue} = 0.12 + 0.02\,(M_r + 20). \tag{2}$$

The fraction of objects in the blue component decreases with increasing luminosity:

$$f_{\rm blue}(M_r) \approx 0.46 + 0.07\,(M_r + 20), \tag{3}$$

and drops toward zero at the bright end.

Fig. 1 shows this bimodality and the two Gaussian component fits which are based on these expressions. Our model of the bimodality, which motivates an algorithm for constructing mock catalogues, and which our halo model calculation requires, uses the red and blue sequences given by equations (1) and (2). These sequences are also shown in a colour–magnitude diagram (Fig. 2) along with the colour–magnitude contours of one of the volume-limited SDSS catalogues used in Section 4.

However, it is common to make a cruder approximation to this bimodality, by simply labelling galaxies as 'red' if they are redder than

$${}^{0.1}(g - r)_{\rm cut} = 0.8 - 0.03\,({}^{0.1}M_r + 20), \tag{4}$$

and calling them 'blue' otherwise (e.g. Zehavi et al. 2005; Blanton & Berlind 2007). [The recent analysis of satellite galaxy colours by van den Bosch et al. (2008) used a stellar mass-based split, which translates into a similar colour cut as the one above, although their cut is slightly steeper with respect to r-band luminosity.] In what follows, we will only use this sharp threshold when comparing our results to previous work.

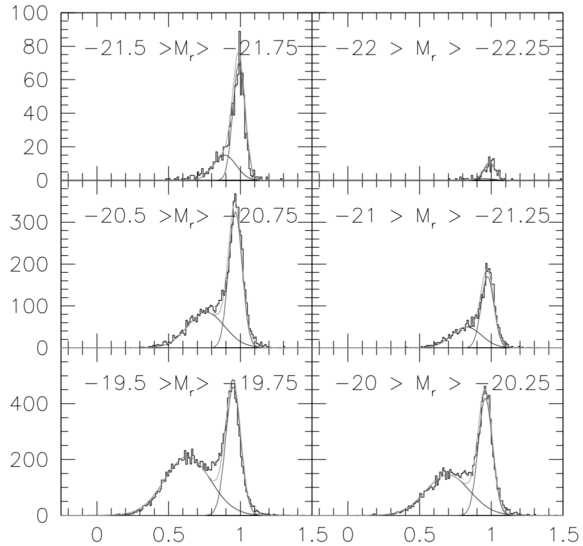

Figure 1. Bimodal distribution of $g - r$ colour in the SDSS. Smooth curves show that, at a fixed luminosity, the distribution is well modelled by the sum of two Gaussian components.

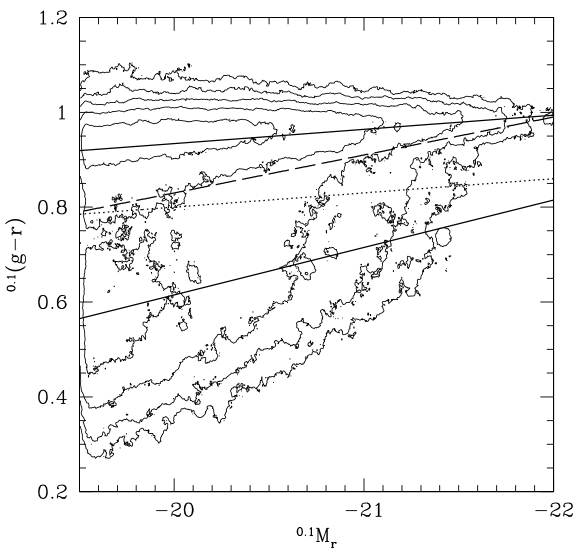

Figure 2. Colour–magnitude diagram in the $M_r < -19.5$ volume-limited SDSS catalogue. Solid lines show the mean values of the red and blue sequences (equations 1 and 2); dashed line shows the satellite sequence (equation 7) and dotted line shows equation (4) which some authors use to divide the population into red and blue.

The SDSS colours (and magnitudes) have measurement errors which contribute to the rms of the red and blue sequences, especially at faint magnitudes. However, the uncertainties in the $g - r$ galaxy colours in the SDSS are typically less than 0.02 mag, so they are unlikely to significantly affect the constraints on the model. Since the measurement errors almost certainly do not correlate with environment, they are not expected to bias the measured colour–mark correlation functions shown in Section 4; they will, however, increase the error bars on the clustering signal. We note, however, that there is an important systematic problem with the colours for which we do not correct: namely a dusty spiral will appear redder if viewed edge-on rather than face-on. In fact, a significant fraction of the objects called 'red' are not the early types which one typically associates with the 'red sequence' (Bernardi et al. 2003). Mitchell et al. (2005) estimate that this fraction is of the order of 40 per cent (also see Maller et al. 2008). Since this systematically also affects the luminosities, for which no halo-model analysis to date has yet made a correction, we have not done so here either.

2.2 Luminosities and colours of centrals and satellites

To illustrate our approach, we will begin with an extreme assumption. Suppose that (i) the bimodal colour distribution is independent of halo mass (by which we mean that the distribution of colour at a fixed luminosity is independent of halo mass; the distribution of luminosities, of course, does depend on halo mass), and (ii) satellites are drawn from the red part of the bimodal colour distribution – *no* satellites come from the blue sequence. Later in this paper, we will find it necessary to relax the second assumption, but the data do not yet require us to give up the first. We think assumption (ii) is a useful extreme which helps bring into focus the key points of the approach.

Given the constraints from the colour distribution as a function of luminosity (Section 2.1) and from luminosity-dependent clustering (Appendix A), these two assumptions allow one to model the halo mass dependence of the colours of both centrals and satellites, and in general to build a model of how galaxy clustering depends on colour, without any additional free parameters. For example, these assumptions imply that the mean satellite colour is

$$\begin{aligned}\langle c|m\rangle &\equiv \int \mathrm{d}c\,p(c|m)\,c = \int \mathrm{d}L\;p(L|m)\int \mathrm{d}c\,p(c|L,m)\,c\\ &= \int \mathrm{d}L\;p(L|m)\langle c|L,m\rangle. \qquad (5)\end{aligned}$$

Whereas the first equality is the definition, the final expression shows how one might estimate the left-hand side from the knowledge of the luminosity distribution in haloes of mass m and the mean colour at a given luminosity in such haloes.

If the distribution of satellite colours at a fixed satellite luminosity is independent of halo mass [this is not unreasonable, given that the distribution of luminosities itself is approximately independent of halo mass (see Skibba et al. 2006, 2007; Hansen et al. 2007)], then this becomes

$$\langle c|m\rangle_{\rm sat} = \int \mathrm{d}L p_{\rm sat}(L|m)\,\langle c|L\rangle_{\rm sat}. \qquad (6)$$

Thus, given m, we integrate over the distribution of satellite luminosities, weighting by $\langle c|L\rangle_{\rm sat}$.

Our simplest model (assumption ii) uses equation (1), the colour–magnitude relation along the red sequence, for $\langle c|L\rangle_{\rm sat}$. We will show later that setting

$$\langle g-r|M_r\rangle_{\rm sat} = 0.83 - 0.08\,(M_r+20) \qquad (7)$$

instead, which is bluer at faint luminosities (see Fig. 2), provides a substantially better agreement with the observations. This is best thought of as a model in which satellites are drawn from the red sequence with probability

$$p(\text{red sat}|L) = \frac{\langle c|L\rangle_{\rm sat} - \langle c|L\rangle_{\rm blue}}{\langle c|L\rangle_{\rm red} - \langle c|L\rangle_{\rm blue}}, \qquad (8)$$

and from the blue sequence with probability

$$p(\text{blue sat}|L) = 1 - p(\text{red sat}|L). \qquad (9)$$

These expressions imply that, for SDSS $g - r$ colours, p(blue sat $|L) \approx 0.4$ at $M_r = -18$, and it drops to zero at $M_r \approx -22$. Since the fraction of galaxies that are satellites has a similar dependence on luminosity (we provide explicit HOD-derived expressions for this later), this model says that although almost 60 per cent of the galaxies with $M_r = -18$ are from the blue sequence (cf. equation 3), slightly less than 20 per cent of the galaxies with $M_r = -18$ are blue satellites: only a third of the faint blue galaxies are satellites, the others are centrals. Allowing for blue-sequence satellites modifies the discussion below trivially.

It is worth reiterating that, in this model, satellite colours only depend on halo mass because satellite luminosities do. Since $p_{\rm sat}(L|m)$ depends only weakly on m (Skibba et al. 2007), we expect $\langle c|m\rangle_{\rm sat}$ to also depend only weakly on m.

In practice, we do not evaluate the integral in equation (6) as written. Rather, we use a variation of the trick we used in Skibba et al. (2006). Namely, for some function $C(L)$ of L,

$$\int_{L_{\rm min}}^{\infty}{\rm d}L\, C(L) \int_{L}^{\infty} {\rm d}L'\, p(L'|m) = \int_{L_{\rm min}}^{\infty}{\rm d}L'\, p(L'|m) \int_{L_{\rm min}}^{L'} {\rm d}L\, C(L). \qquad (10)$$

Skibba et al. studied the case where $C(L) = 1$, so the inner integral gave $L' - L_{\rm min}$. Here, we wish to set $C(L)$ to be that function of L which, when integrated over L, yields $\langle c|L'\rangle_{\rm red} - \langle c|L_{\rm min}\rangle_{\rm red}$. Thus,

$$\langle c|m\rangle_{\rm sat} = \langle c|L_{\rm min}\rangle_{\rm red} + \int_{L_{\rm min}}^{\infty}{\rm d}L\, C(L)\, P_{\rm sat}(>L|m), \qquad (11)$$

where we have defined

$$P_{\rm sat}(>L|m) \equiv \int_{L}^{\infty} {\rm d}L\, p_{\rm sat}(L|m) = \frac{N_{\rm sat}(>L|m)}{N_{\rm sat}(>L_{\rm min}|m)}. \qquad (12)$$

If colour and luminosity are in magnitudes (i.e. we work in logarithmic rather than linear variables) then the integral is simpler:

$$\langle g-r|m\rangle_{\rm sat} = \langle g-r|M_{\rm min}\rangle_{\rm sat} + C_{\rm sat,\,slope}\int_{M_{r,\rm min}}^{-\infty}{\rm d}M_r\, P_{\rm sat}(<M_r|m), \qquad (13)$$

where $P_{\rm sat}(<M_r|m) = P_{\rm sat}(>L|m)$, and $C_{\rm sat,\,slope}$ is the slope of the relation showing how the mean satellite colour changes with magnitude. That is, $C_{\rm sat,\,slope} = -0.032$ or -0.08 if satellites are drawn from the red sequence (cf. equation 1) or from equation (7).

Obtaining an expression for the typical colour associated with the central galaxies of m haloes is more complicated. Although the bimodal distribution of colour at a fixed luminosity can be thought of as arising from a mix of objects which lie along a blue or a red sequence, in what follows, it will be more useful to think in terms of the central-satellite split. In this case,

$$\langle c|L\rangle = \frac{N_{\rm cen}(L)\langle c|L\rangle_{\rm cen} + N_{\rm sat}(L)\langle c|L\rangle_{\rm sat}}{N_{\rm cen}(L) + N_{\rm sat}(L)}, \qquad (14)$$

making

$$\langle c|L\rangle_{\rm cen} = \langle c|L\rangle + \frac{N_{\rm sat}(L)}{N_{\rm cen}(L)}\,[\langle c|L\rangle - \langle c|L\rangle_{\rm sat}]. \qquad (15)$$

If, as we assumed for the satellites, the distribution of central galaxy colours at a fixed luminosity is independent of halo mass (the results of Berlind et al. 2005 support this assumption), then the mean colour as a function of halo mass is simply $\langle c|m\rangle_{\rm cen} = \langle c|L(m)\rangle_{\rm cen}$ if there is no scatter between central galaxy luminosity and halo mass (e.g. Zehavi et al. 2005). If there is scatter (e.g. Zheng, Coil & Zehavi 2007), then

$$\langle c|m\rangle_{\rm cen} = \int {\rm d}L\, P_{\rm cen}(L|m)\, \langle c|L\rangle_{\rm cen}. \qquad (16)$$

Now, by hypothesis, $\langle c|L\rangle_{\rm sat}$ is given by equation (1) (or equation 7), whereas $\langle c|L\rangle$ is simply the mean colour of all galaxies as a function of luminosity. Thus, both these quantities are observables, or are constrained by observables, for the satellites (Skibba 2008); the only unknown is $N_{\rm sat}(L)/N_{\rm cen}(L)$. Since both numbers are counted in the same volume, this is the same as the ratio of the number densities: $N_{\rm sat}(L)/N_{\rm cen}(L)$. We discuss how this ratio is determined by the luminosity-based HOD in Appendix A.

It is worth noting that the quantity in square brackets in equation (15) is negative. This means that, in general, the colours of central galaxies are *bluer* than the average for their luminosities. Although this seems counter to intuition – one is used to thinking of central galaxies as being red – it is, in fact, sensible. Essentially, the paradox is resolved when one realizes that the satellites actually inhabit more massive haloes than do centrals of the same luminosity. It may help to note that this effect is most pronounced at low L, where the mean colour is significantly bluer than the red sequence, and the number of satellites can be large. Low-luminosity galaxies that are centrals are hosted by low-mass haloes, whereas satellite galaxies of similar luminosity are more likely to reside in groups or clusters, so their parent haloes are more massive. Thus, our model has placed blue central galaxies in low-mass haloes and red satellite galaxies in massive haloes. At higher luminosities, $\langle c|L\rangle$ approaches that of the red sequence. In this limit, the term in square brackets becomes small, as does the number of satellites, so the colours of central galaxies tend to $\langle c|L\rangle$; that is, our model places luminous central galaxies on the red sequence.

3 TWO WAYS TO TEST THE MODEL OF BIMODALITY

We now describe two ways to test our model of the bimodality. The first is numerical – we provide an algorithm for constructing mock catalogues which are consistent with our model. The model can be tested by performing the same analysis on the mocks that was performed on real data. This is particularly useful for the analyses which are somewhat involved or contrived, so that an analytic description is difficult. The second is analytic – we show how our model can be implemented to provide a halo-model description of mark correlations when the mark is colour. Skibba (2008) describes the result of a third test: a direct measurement of central and satellite colours in group catalogues.

3.1 An algorithm for constructing mock catalogues with luminosities and colours

The analysis above shows that one can generate a mock galaxy catalogue in two steps: first generate luminosities, and then use them to generate colours. Note that the method used for generating luminosities is *not* important: the luminosities could have come from a HOD analysis, a CLF analysis, or may be based on a SHAM.

Our algorithm for generating luminosities comes from Skibba et al. (2006). Briefly, we specify a minimum luminosity $L_{\rm min}$ which is smaller than the minimum luminosity we wish to study. We then select the subset of haloes in the simulation which have $m > m_{\rm min}(L_{\rm min})$. Each halo is assigned a central galaxy with luminosity given by inverting the relation between halo mass and luminosity (equation A2). We specify the number of satellites the halo contains by choosing an integer from a Poisson distribution with mean $N_{\rm sat}(>L_{\rm min}|m)$. The luminosity of each satellite galaxy is specified by generating a random number u_0 distributed uniformly between 0 and 1, and finding that L for which $N_{\rm sat}(>L|m)/N_{\rm sat}(>L_{\rm min}|m) = u_0$. This ensures that the satellites have the correct luminosity distribution.

We could assign colours to each of the satellites by drawing a Gaussian random number with mean and rms given by inserting the satellite luminosity in equation (1) for the red sequence. However, as we show in Section 4, this results in a correlation between colour and environment that is too strong compared to the data. Instead, we want the satellites to have colours which are bluer than the red sequence at faint luminosities, as specified by equation (7). To implement this in our mock catalogue, we draw a uniformly distributed random number $0 \leq u_1 < 1$. The satellite is drawn from the red sequence (a Gaussian with mean and rms given by equation 1) if $u_1 \leq p(\mathrm{red\,sat}\,|L)$, where $p(\mathrm{red\,sat}\,|L)$ is given by equation (8) and from the blue sequence (Gaussian with mean and rms from equation 2) otherwise. Note that only the luminosity matters for determining the colour; the halo mass plays no additional role.

The colours for central galaxies can also be drawn from either the red or blue sequence. To determine which, we draw another uniformly distributed random number u_2. If $u_2 > f_{\mathrm{blue}}(L)/f_{\mathrm{cen}}(L)$, where L is the central object's luminosity, then the object is assigned to the red sequence, so we draw a Gaussian with luminosity-dependent mean and rms given by equation (1). Else, it is blue, and we use equation (2) instead. Equations (3) and (A8) show that this assigns all central galaxies fainter than $M_r \approx -18.5$ to the blue sequence.

Finally, we place the central galaxy at the centre of its halo, and distribute the satellites around it so that they follow a Navarro, Frenk & White (1997) profile (see Scoccimarro & Sheth 2002 for how this can be done efficiently). The resulting mock galaxy catalogue has been constructed to have the correct luminosity function as well as the correct luminosity dependence of the galaxy two-point correlation function. In addition, colours in this catalogue are assigned in accordance with the model described previously: satellite and central galaxy colours are assigned such that the galaxy population as a whole has the correct colour–luminosity distribution.

Our model makes a prediction for how the bimodality in colour differs for central and satellite galaxies. In Fig. 3, we show the colour distribution as a function of halo mass of central and satellite galaxies in a mock catalogue with $M_r < -20.5$. We normalize the central and satellite galaxy distributions by the total number of galaxies in each bin; consequently, the lower mass haloes are dominated by central galaxies, while satellites contribute most of the galaxies in massive haloes. First, note that the satellite distribution is almost the same in each panel: this is a consequence of our assumption that the distribution of satellite colours at fixed luminosity is independent of halo mass (i.e. $p_{\mathrm{sat}}(c|L, m) = p_{\mathrm{sat}}(c|L)$), and the fact that satellite luminosities are approximately independent of mass as well. On the other hand, the centrals have a more bimodal distribution in low-mass haloes, while in massive haloes most of them are on the red sequence. Secondly, it is interesting that the blue and red modes of the central galaxy bimodal colour distribution are closer together than those of the satellite colour distribution, such that the blue bump of the centrals tends to peak at the minimum in the satellite distribution.

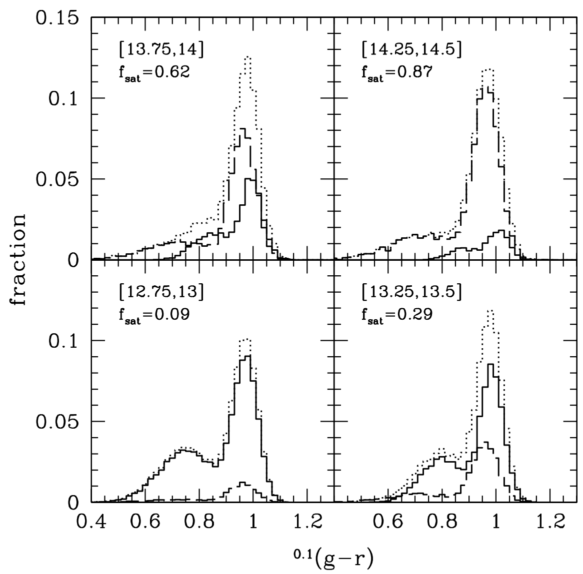

Figure 3. Bimodal distribution of $g - r$ colour for central galaxies (red histogram), satellite galaxies (blue dashed histogram) and all galaxies (centrals+satellites; black-dotted histogram) in a mock catalogue with $M_r < -20.5$. The distributions are shown for four intervals in log halo mass, indicated in square brackets in each panel.

3.2 Implicit assumptions, bells and whistles

This halo-model-based prescription for making mock catalogues uses three simplifying assumptions which are worth discussing explicitly. First, although we assume haloes are spherical and smooth, the density run of satellites around halo centres is almost certainly neither. Generating triaxial distributions is straightforward once prescriptions for how the triaxiality depends on halo mass and how it correlates with environment are available. Once these are known, they can be incorporated into the analytic halo-model description (Smith, Watts & Sheth 2006). Similarly, parametrizations of halo substructure can also be incorporated into the description (Sheth & Jain 2003). Of course, both these types of correlations can be included in the mock catalogue directly from a simulation if one simply selects the appropriate number of particles from the halo itself, rather than generating the profile shape synthetically. This is costly because now one needs the full particle distribution, rather than just the halo catalogue, to generate the mock – but note that it is not a problem of principle.

Secondly, note that the number of galaxies in a halo, the spatial distribution of galaxies within a halo and the assignment of luminosities all depend only on halo mass. None of these depends on the surrounding large-scale structure. Therefore, the mock catalogue includes only those environmental effects which arise from the environmental dependence of halo abundances. This point was made by Skibba et al. (2006); it is also true of our prescription for including colours.

Thirdly, haloes of the same mass will have had a variety of formation histories. Some will have assembled their mass and their galaxy populations more recently than others. Recent assembly means less time for dynamical friction, and, possibly, a younger stellar population. So, at a fixed halo mass, one might expect to find a correlation between the age of a halo and the galaxy population within it. In particular, the number of galaxies in a halo, their luminosities and their colours may all be correlated with the formation history. Our halo-model description (and associated mock catalogue) ignores all such correlations. To see this clearly, note that we assign luminosities and colours to the galaxies in a halo without regard for the number of galaxies in it. Had we used a SHAM to assign luminosities, then some of correlation between formation history and the galaxy population will have been included. If one is already carrying along the particle distribution from the simulation to construct the mock, then the next level of complication is to also include additional information about the merger history in the simulation, for use when making the mock.

We also assign colours to satellite galaxies without explicit consideration of the colour of the central galaxy, and we make no effort to incorporate colour gradients within a halo into our model. This is mainly because the two-point statistics we study in this paper, weighted or unweighted, are known to be not very sensitive to gradients (see Sheth et al. 2001; Scranton 2002; Sheth 2005; Skibba 2008 for more discussion and simple prescriptions for incorporating colour gradients).

These are all interesting problems for the future (and they are almost certainly not independent problems!), but the measurements described in the next section do not require these refinements.

3.3 A halo-model description of colour–mark correlations

Mark correlations are an efficient way to quantify the correlation between the properties of galaxies and their environment (Sheth, Connolly & Skibba 2005). The two-point mark correlation function is simply

$$M(r) \equiv \frac{1 + W(r)}{1 + \xi(r)}, \tag{17}$$

where $\xi(r)$ is the traditional two-point correlation function and $W(r)$ is the same sum over galaxy pairs separated by r, but now each member of the pair is weighted by the ratio of its mark to the mean mark of all the galaxies in the catalogue (e.g. Stoyan & Stoyan 1994; Beisbart & Kerscher 2000). In effect, the denominator divides out the contribution to the weighted correlation function which comes from the spatial contribution of the points, leaving only the contribution from the fluctuations of the marks.

In models where a galaxy's properties correlate with environment only because they correlate with host halo mass, but halo abundances correlate with environment, it is relatively straightforward to write down a halo model of mark correlations (Sheth 2005). Since our model of central and satellite colours is precisely of this form, we can build a halo model of colour–mark correlations. Appendix B provides a detailed description of how this is done. In principle, comparison of this prediction with measurements in the SDSS data set allows a test of our approach.

Before performing this test with data, Fig. 4 shows a comparison with measurements in the mock catalogue described in the previous section. The halo population is from the Virgo Consortium's Very Large Simulation (VLS; Yoshida, Sheth & Diaferio 2001), and the mock galaxies have $M_r < -20.5$. Luminosities and colours were assigned as described above: the top panel shows $M(r)$ as a function of real-space separation when luminosity is the mark; the bottom panel has $g - r$ as the mark. Solid curves show the halo model prediction, computed by inserting the mass dependence of the mean marks for centrals and satellites into the mark correlation formalism of Appendix B. The luminosity and colour–mark correlations are significantly above unity, which clearly shows that in denser environments we expect the luminosities of galaxies to be brighter (top panel) and the colours to be redder (bottom panel). The mark correlations also clearly show the transition from the one-halo term to the two-halo term at $r \sim \mathrm{Mpc}\, h^{-1}$, which is the virial radius of the most massive haloes at $z \sim 0$. The transition is more pronounced than in the traditional unmarked correlation function $\xi(r)$.

There is a reasonably good agreement between the halo model calculation and the mocks for both the luminosity and colour–mark correlation functions; the unmarked correlation functions $\xi(r)$ agree extremely well, so they are not shown. Both panels in the figure show a similar but small discrepancy at similar scales, approximately where the one–two-halo term transition occurs. Although statistically significant, this discrepancy is small compared to the significance with which the signal itself differs from unity: the halo model calculations are qualitatively, if not quantitatively, correct across a wide range in scales. The agreement between the model and the mocks is encouraging; it suggests that much of the environmental dependence of galaxy colour arises from the environmental dependence of host halo mass.

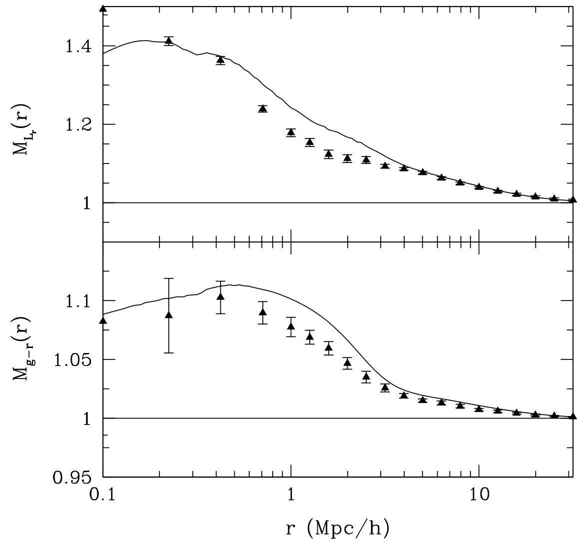

Figure 4. Luminosity (top) and $g - r$ colour (bottom) mark correlation functions in a real-space mock catalogue in which $M_r < -20.5$. Solid curves show the halo model predictions.

4 COMPARISON WITH SDSS

In this section, we compare colour mark projected correlation functions predicted by the halo model to measurements in the SDSS (York et al. 2000). We use two volume-limited large-scale structure samples built from the New York University Value-Added Galaxy Catalogue (Blanton et al. 2005b) from SDSS Data Release (DR) 4 plus, which is a subset of SDSS DR5 (Adelman-McCarthy et al. 2007). We k-correct the magnitudes to $z = 0.1$ using the KCORRECT V4_1 code of Blanton & Roweis (2007); the magnitudes are also corrected for passive evolution. Our fainter catalogue has limits $-23.5 < {}^{0.1}M_r < -19.5, 0.017 < z < 0.082$; it consists of 78 356 galaxies with mean density $\bar{n}_{\mathrm{gal}} = 0.01061\,(h^{-1}\mathrm{Mpc})^3$. Our brighter catalogue has $-23.5 < {}^{0.1}M_r < -20.5$ and $0.019 < z < 0.125$, and contains 73 468 galaxies with mean density $\bar{n}_{\mathrm{gal}} = 0.00280\,(h^{-1}\,\mathrm{Mpc})^3$. These luminosity thresholds approximately correspond to $M_r < M^* + 1$ and $M_r < M^*$, where M^* is the break in the Schechter function fit to the r-band luminosity function (Blanton et al. 2003).

For the measured correlation functions and jackknife errors, which require random catalogues and jackknife sub-catalogues, we use the hierarchical pixel scheme SDSSPix,[1] which characterizes the survey geometry, including edges and holes from missing fields and areas near bright stars. This same scheme has been used for other clustering analyses (Scranton et al. 2005; Hansen et al. 2007) and for lensing analyses (Sheldon et al. 2007).

Fig. 5 shows the distribution of $g - r$ colours in our fainter ($M_r < -19.5$) catalogue. The distributions of Petrosian and model colours

[1] http://lahmu.phyast.pitt.edu/~scranton/SDSSPix

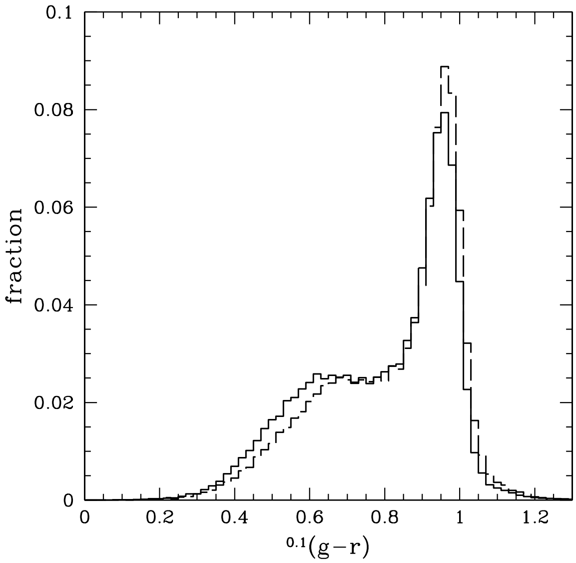

Figure 5. Distribution of Petrosian (blue histogram) and model (red dashed histogram) $g - r$ colours in the $M_r < -19.5$ volume-limited catalogue.

are similar, although the model colours are slightly redder. The mean Petrosian colour is 0.796, whereas the mean model colour is 0.825. This is not unexpected – galaxies have colour gradients, and model colours measure the colour on smaller scales. These mean values are 0.850 and 0.885 in the brighter catalogue ($M_r < -20.5$). The blue fractions of the Petrosian colours of the fainter and brighter catalogues are, respectively, 44 and 37 per cent using the fixed colour–magnitude cut (equation 4) and 47 and 43 per cent using the double-Gaussian model (equations 1–3), and they are ≈6 per cent lower for the model colours.

We now present our colour–mark correlation functions. In practice, in order to obviate redshift-space calculations in the halo model and redshift distortions in the data, we use the projected two-point correlation function

$$w_{\rm p}(r_{\rm p}) = \int {\rm d}r\, \xi(r_{\rm p}, \pi) = 2 \int_{r_{\rm p}}^{\infty} {\rm d}r\, \frac{r\, \xi(r)}{\sqrt{r^2 - {r_{\rm p}}^2}}, \qquad (18)$$

where $r = \sqrt{{r_{\rm p}}^2 + \pi^2}$, $r_{\rm p}$ and π are the galaxy separations perpendicular and parallel to the line of sight, and we integrate up to line-of-sight separations of $\pi = 40\,{\rm Mpc}\,h^{-1}$. We estimate $\xi(r_{\rm p}, \pi)$ using the Landy & Szalay (1993) estimator

$$\xi(r_{\rm p}, \pi) = \frac{DD - 2DR + RR}{RR}, \qquad (19)$$

where DD, DR and RR are the normalized counts of data–data, data–random and random–random pairs at each separation bin. We then define the marked projected correlation function

$$M_{\rm p}(r_{\rm p}) = \frac{1 + W_{\rm p}(r_{\rm p})/r_{\rm p}}{1 + w_{\rm p}(r_{\rm p})/r_{\rm p}}, \qquad (20)$$

which makes $M_{\rm p}(r_{\rm p}) \approx M(r)$ on scales larger than a few Mpc. For the SDSS measurements, we used random catalogues with 10 times as many points as in the data; the error bars show the variance of the measurements of 30 jackknife sub-catalogues.

Figs 6 and 7 compare the colour marked correlation functions for the $M_r < -19.5$ and -20.5 catalogues with our predictions. The solid and open points show the measurements for Petrosian and model colours. The colour mark signals in the bottom panels are stronger for Petrosian colours, at the 1σ level, for both luminosity

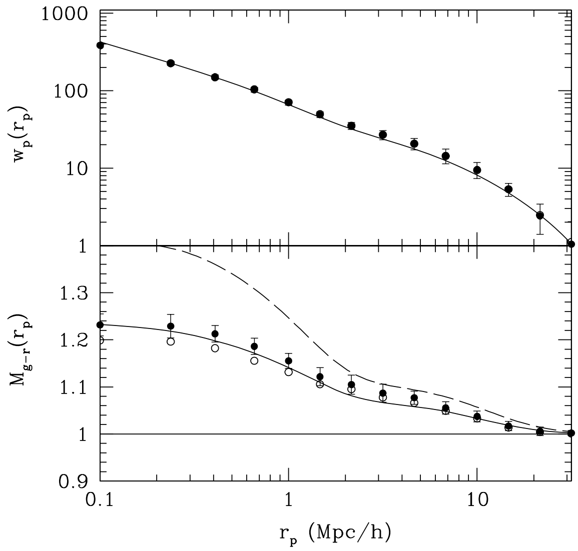

Figure 6. Projected two-point correlation function and $g - r$ colour–mark correlation function for $M_r < -19.5$. Points show SDSS measurements for Petrosian (solid points) and model colours (open points), with jackknife errors. Solid curves show the halo-model prediction when satellite galaxies can be drawn from either the red or the blue sequences (equations 7–9); dashed curve shows the prediction if satellites are drawn from the red sequence only (equation 1).

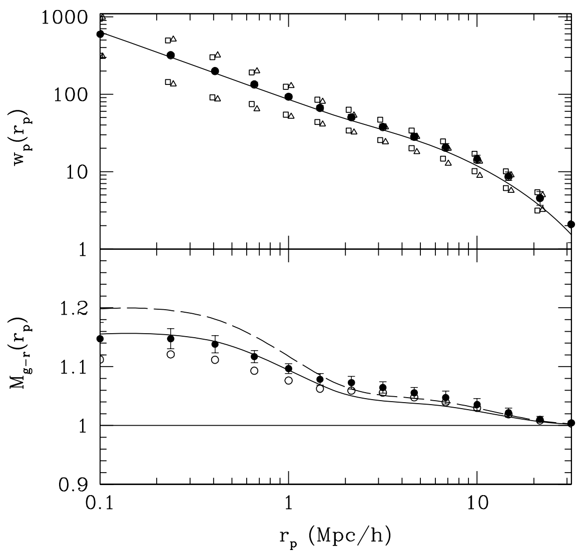

Figure 7. Projected two-point correlation function and $g - r$ colour–mark correlation function, like Fig. 6, but for $M_r < -20.5$. In the upper panel, the correlation functions for galaxies redder and bluer than the colour cut (equation 4) are also shown, for the SDSS galaxies (open squares) and mock catalogue galaxies (open triangles). For clarity, error bars are only shown for the full SDSS catalogue.

thresholds $M_r < -19.5$ and -20.5. Evidently, the environmental dependence of Petrosian colours is stronger than that of model colours. However, this is probably due to the fact that the red and blue peaks are slightly more displaced from one another for Petrosian rather than model colours.

The correlation function of galaxies split by colour is the measurement that has traditionally been used to show the environmental dependence of colour (e.g. Zehavi et al. 2005; Tinker et al. 2008). The top panel in Fig. 7 shows such measurements for galaxies redder and bluer than the colour cut given by equation (4). Open squares and triangles are for measurements in the SDSS and in a mock catalogue constructed as described in the previous section. (The SDSS galaxies were split by their Petrosian colours; the measurement is virtually the same when they are split by model colours.)

The mock catalogue is at $z = 0$, whereas the SDSS measurements (and corresponding theory curves) are at $z \sim 0.1$. Therefore, to compare the clustering of red and blue galaxies in our mock with the measurements, we measure the ratio of $w_{\rm p,red}$ to $w_{\rm p,all}$, and $w_{\rm p,blue}$ to $w_{\rm p,all}$ in our $z = 0$ mock. We then assume that this ratio would be the same at $z = 0.1$ as it is at $z = 0$; the triangles show the result of applying this ratio to $w_{\rm p,all}$ at $z = 0.1$ (i.e. the filled circles) – they represent how the clustering of red and blue galaxies differs from the full sample in our mock.

The agreement between the clustering of the mock galaxies and SDSS galaxies is very good, indicating that our model reproduces these traditional measurements of colour-dependent clustering very well.

Because mark statistics do not require binning of the data set into coarse bins in colour, or coarse bins in density, the mark correlation functions shown in the lower panels of the figures contain significantly more information about environmental correlations than more traditional measures. They allow the mark to take a continuous range in values, and they yield a clear, quantitative estimate of the correlation between the mark and the environment at a given scale. Mark statistics are also sensitive to the distribution of the marks: for example, for the fainter luminosity threshold ($M_r < -19.5$), the colour marks have a wider distribution than for the brighter threshold, so some galaxies have colours farther from the mean mark. Because these outliers also tend to be in more extreme environments, the result is a stronger mark correlation (e.g. $M_{g-r}(r_{\rm p} = 100\,h^{-1}\,{\rm kpc}) \approx 1.23$ versus 1.15). Note also that the mark correlation functions are more curved for the fainter luminosity threshold, with a more distinct transition between the one- and two-halo terms. This is because there are more satellite galaxies in the fainter sample, and the mark clustering is more sensitive to their spatial distribution within haloes.

Note that the model in which no satellites come from the blue sequence (dashed curve) produces too strong a signal: galaxies in dense environments are too red. Since most of these are satellites, this model places too many red satellites in massive haloes. The difference between the two models is greater for the faint luminosity threshold simply because there are more faint satellites, more of whose colours should be drawn from the blue sequence. The model in which satellites come from a mix of the two sequences, though they are increasingly red at large luminosities (equations 7–9), is in good agreement with the measurements on all scales where the statistic is reliably measured. This suggests that this model of the colours of central and satellite galaxies is a reasonable one. The good agreement between our model and the data also indicates that the correlation between halo mass and environment is the primary driver of the environmental dependence of galaxy colour.

5 DISCUSSION

We have developed and tested a simple model for several observed correlations between colour and environment on scales of $100\,h^{-1}\,{\rm kpc} < r_{\rm p} < 30\,h^{-1}\,{\rm Mpc}$. Our model is built upon the model of luminosity mark clustering of Skibba et al. (2006), in which the luminosity-dependent HOD was constrained by the observed luminosity-dependent correlation functions and galaxy number densities in the SDSS. The model presented here has added constraints from the bimodal distribution of the colours of SDSS galaxies as a function of luminosity. We make two assumptions: (i) that the bimodality of the colour distribution at fixed luminosity is independent of halo mass and (ii) that satellite galaxies tend to follow a particular sequence in the colour–magnitude diagram, one that approaches the red sequence with increasing luminosity (equation 7). Alternatively, this assumption can be phrased as specifying how the fraction of satellites which are drawn from the red and blue sequences depends on luminosity (equation 9).

One virtue of our model is the ease with which it allows one to include colour information into mock catalogues. Adding colours to a code which successfully reproduces luminosity-dependent clustering requires just four simple lines of code – two for centrals and two for satellites (Section 3.1). This is far more efficient than 'brute-force' approaches which are based on fitting HODs to fine bins in *L and* colour, or others which are based on using observed correlations between colour and local density. Since bimodality is also observed at $z = 1$, it would be interesting to see if our approach is similarly successful at interpreting the measurements of Coil et al. (2008) in the DEEP2 survey.

Realistic colours are necessary for providing realistic training sets for galaxy group- and cluster-finding algorithms, and a number of groups are currently developing such mock catalogues. So, we think it is worth emphasizing that our approach can be applied to *any* mock catalogue which produces the correct luminosity-dependence of clustering. Thus, although we phrased our discussion in terms of an HOD-based mock, mocks based on CLFs or SHAMs could also use our method for generating colours.

In particular, cluster-finding algorithms that exploit information about brightest cluster galaxies (BCGs), or galaxies' positions from the red sequence, or galaxies' redshift-distorted positions, or the multiplicity function or total luminosity or stellar mass of groups, could all be tested with mock catalogues constructed with the approach described in this paper. We will be happy to provide our mock catalogues to those interested, upon request.

More generally, we feel that the simplicity of our approach makes it an attractive way to begin to include the entire spectral energy distribution into the halo-model description, and hence into mock catalogues. Specifically, starting from our successful model for adding $g - r$ given L, the next step might be to add, say, $u - r$, given $g - r$ and L – again assuming that the distribution $p(u - r|g - r, L)$ is independent of halo mass. This is also attractive because we have shown that such an approach is easily described using the language of the halo model – Section 3.3 provides a halo-model description of the colour–mark correlation function. This facilitates the use of mark statistics in testing our hypothesis that the bimodal colour distribution is independent of halo mass.

Comparison of our mark correlation measurements with measurements in our mock catalogues and with our halo model calculations (Figs 6 and 7) suggests that if the bimodal colour distribution is independent of halo mass, then at least some of the non-central/satellite galaxies in a halo must be drawn from the blue sequence – this fraction of blue satellites must be larger at low luminosities. This is one of the key results of our paper.

If satellites lie on the red sequence because their star formation has been quenched by processes such as 'strangulation' (e.g. Weinmann et al. 2006), then our results suggest that quenching is still on-going at lower luminosities. Such processes are expected to modify the colours and star formation rates of satellite galaxies, but not their morphologies; we investigate this further in a subsequent paper by measuring morphology mark correlations in the SDSS Galaxy Zoo catalogue. We caution, however, that we, like all previous halo-model analyses, have ignored the fact that the inclination can affect the observed galaxy properties – luminosities and colours in the present context. Corrections for inclination-related effects are available in the literature (Giovanelli et al. 1995; Tully et al. 1998; Sheth et al. 2003), and they are not negligible. Recent work on this by Maller et al. (2008), which appeared while our work was being refereed, provides relatively straightforward corrections which may be reasonably accurate. For this reason, our work should be viewed as attempting a halo-model description of the observed colours, rather than providing a truly physical picture of the intrinsic (face-on?) colours. Of course, if the luminosities and colours had been corrected for inclination effects, we expect our analysis to also yield results which are closer to the true physical picture. But, because we have not yet included these corrections, we believe that statements about the physics of 'strangulation', especially at low luminosities, are premature.

We expect our model to be in good agreement with the findings of Zehavi et al. (2005), who analysed volume-limited SDSS samples after dividing galaxies into two bins in colour. They used a slightly redder colour cut than we did to produce the measurements shown in the top panel of Fig. 7. They found that the fraction of central galaxies which lay bluewards of this cut increased as L decreased; that there were no faint blue satellites and that, although there are blue satellites at intermediate and high L, they were about a factor of 5 less common than red satellites in haloes of the same mass. Our model is in qualitative agreement, with the mean central and satellite galaxy colours increasing with both luminosity and halo mass. Zehavi et al. inferred from their results that the majority of bright galaxies are red centrals of massive haloes, and that faint red galaxies are predominantly satellites in massive haloes. This is consistent with Swanson et al. (2008), who found that both luminous and faint red galaxies are more strongly clustered than moderately bright red galaxies. We reach a similar conclusion, although not all faint red galaxies are satellites in massive systems: some are centrals in underdense environments.

We also expect our model to be in qualitative agreement with the findings of Blanton & Berlind (2007). These authors defined blue galaxies as those lying bluewards of $g - r = 0.8 - 0.03(M_r + 20)$ (our equation 4). They then found that the colour–magnitude relation for galaxies in luminous groups tended to have $f_{\rm blue}$ decreasing with group luminosity, but that the red and blue sequences were otherwise approximately independent of group luminosity. They phrased their findings as showing that the colour–magnitude relation depends on group luminosity, presumably because they wished to draw attention to the dependence of $f_{\rm blue}$ on group luminosity. In light of the discussion above, we think this is slightly misleading. The red and blue sequences in our model are *independent* of group properties by construction. In our model, the decrease of the blue fraction in luminous groups is simply a consequence of the assumption that satellites tend to be drawn from the red rather than the blue sequence. This happens because more luminous groups will tend to have more satellites *and* redder centrals (because central galaxy luminosity increases with halo mass which is, in turn, strongly correlated with total luminosity, and luminous galaxies are red). Since our model has mainly red satellites, the red fraction is larger in more luminous groups. Skibba (2008) describes the results of a direct comparison of our model predictions with the colours of centrals and satellites in group catalogues.

In our model, *all* environmental correlations arise from the fact that massive haloes tend to reside in denser environments (Mo & White 1996; Sheth & Tormen 2002). Recent studies of the environmental dependence of halo assembly have shown that halo properties such as formation time and concentration are correlated with the environment at a fixed halo mass (Sheth & Tormen 2004; Gao, Springel & White 2005; Wechsler et al. 2006; Croton, Gao & White 2007b; Wetzel et al. 2007; Keselman & Nusser 2007; Zu et al. 2008). They have found that at a fixed mass, haloes in dense environments form at slightly earlier times than haloes in less dense environments. The success of our model suggests that such 'assembly bias' effects are not the primary drivers of the environmental dependence of galaxy colours in the real universe, thus extending previous conclusions about the insignificance of assembly bias on galaxy luminosities (Skibba et al. 2006; Abbas & Sheth 2006, 2007; Blanton & Berlind 2007; Tinker et al. 2008), at least for the relatively bright galaxies in the SDSS. Further tests, such as the analyses of luminosity and colour mark statistics of catalogues constructed from semi-analytic models with known assembly bias, would shed more light on these issues, and are the subject of a subsequent paper.

Our model does not include the galactic 'conformity' reported by Weinmann et al. (2006), in which bluer centrals are likely to be surrounded by bluer satellites, at a fixed halo mass. Including this effect is the subject of work in progress. The main quantitative predictions of our model, such as the *mean* central and satellite colours as a function of mass, and the correlations between colour and environment, are not expected to be significantly affected by this phenomenon, however. Our model also does not include colour gradients within haloes – it has long been known that satellite galaxies near halo centres tend to be redder than in the outskirts. In this case, satellite colour marks depend on both the host halo mass and on their distance from the halo centre. Halo-model analyses show that this should only matter on small scales (see discussion of fig. 4 in Sheth et al. 2001; Scranton 2002); for galaxy populations with many satellite galaxies, the one-halo term of the colour mark signal is expected to be slightly higher (Sheth 2005). Skibba (2008) incorporates this effect, and does find such an increase at small scales.

Finally, it is worth emphasizing that mark statistics are sensitive indicators of the correlations between galaxy properties and the environment, and as such are powerful tools for constraining galaxy formation models. An analysis of marked correlation with star formation rate marks in the SDSS and the Millennium Simulation is the subject of work in progress. The halo-model description of marked statistics, based on the luminosity dependence of galaxy clustering, also has many applications. In a forthcoming paper (Skibba & Sheth, in preparation), we present a model of stellar mass mark correlations and analyse them with SDSS measurements analogous to the colour–mark correlations presented here.

ACKNOWLEDGMENTS

We thank Xi Kang and Frank van den Bosch for valuable discussions, and the Aspen Center for Physics for hospitality in the Summer of 2007, where some of this work was completed. We also thank Tim McKay, the referee, for comments that helped to improve the quality of the paper. This work was supported by NASA-ATP NAG-13720 and the NSF under grant AST-0520647, and by HST-AR-10646.

We thank Jeffrey Gardner, Andrew Connolly and Cameron McBride for assistance with their Ntropy code, which was used to measure all of the correlation functions presented here. *N*tropy was funded by the NASA Advanced Information Systems Research Program grant NNG05GA60G.

Funding for the SDSS and SDSS-II has been provided by the Alfred P. Sloan Foundation, the Participating Institutions, the

National Science Foundation, the US Department of Energy, the National Aeronautics and Space Administration, the Japanese Monbukagakusho, the Max Planck Society and the Higher Education Funding Council for England. The SDSS web site is http://www.sdss.org/.

The SDSS is managed by the Astrophysical Research Consortium for the Participating Institutions. The Participating Institutions are the American Museum of Natural History, Astrophysical Institute Potsdam, University of Basel, Cambridge University, Case Western Reserve University, University of Chicago, Drexel University, Fermilab, the Institute for Advanced Study, the Japan Participation Group, Johns Hopkins University, the Joint Institute for Nuclear Astrophysics, the Kavli Institute for Particle Astrophysics and Cosmology, the Korean Scientist Group, the Chinese Academy of Sciences (LAMOST), Los Alamos National Laboratory, the Max-Planck-Institute for Astronomy (MPA), the Max-Planck-Institute for Astrophysics (MPIA), New Mexico State University, Ohio State University, University of Pittsburgh, University of Portsmouth, Princeton University, the United States Naval Observatory and the University of Washington.

REFERENCES

Abbas U., Sheth R. K., 2006, MNRAS, 372, 1749
Abbas U., Sheth R. K., 2007, MNRAS, 378, 641
Adelman-McCarthy J. K. et al., 2007, ApJS, 172, 634
Baldry I. K., Glazebrook K., Brinkmann J., Ivezić Ž., Lupton R. H., Nichol R. C., Szalay A. S., 2004, ApJ, 600, 681
Beisbart C., Kerscher M., 2000, ApJ, 545, 6
Benson A. J., Cole S., Frenk C. S., Baugh C. M., Lacey C. G., 2000, MNRAS, 311, 793
Benson A. J., Bower R. G., Frenk C. S., Lacey C. G., Baugh C. M., Cole S., 2003, ApJ, 599, 38
Berlind A. A., Weinberg D. H., 2002, ApJ, 575, 587
Berlind A. A., Blanton M. R., Hogg D. W., Weinberg D. H., Davé R., Eisenstein D. J., Katz N., 2005, ApJ, 629, 625
Bernardi M. et al., 2003, AJ, 125, 1817
Blanton M. R., Berlind A. A., 2007, ApJ, 664, 791
Blanton M. R., Roweis S., 2007, AJ, 133, 734
Blanton M. R. et al., 2003, ApJ, 592, 819
Blanton M. R., Eisenstein D., Hogg D. W., Schlegel D. J., Brinkmann J., 2005a, ApJ, 629, 143
Blanton M. R. et al., 2005b, AJ, 129, 2562
Coil A. et al., 2008, ApJ, 672, 153
Conroy C., Wechsler R. H., Kravtsov A. V., 2006, ApJ, 647, 201
Cooray A., 2006, MNRAS, 365, 842
Cooray A., Sheth R., 2002, Phys. Rep., 372, 1
Croton D. J., Norberg P., Gaztanaga E., Baugh C. M., 2007a, MNRAS, 379, 1562
Croton D. J., Gao L., White S. D. M., 2007b, MNRAS, 374, 1303
Gao L., Springer V., White S. D. M., 2005, MNRAS, 363, L66
Gardner J. P., Connolly A., M^{c}Bride C., 2007, preprint (arXiv:0709.1967)
Giovanelli R., Haynes M. P., Salzer J. J., Wegner G., da Costa L. N., Freudling W., 1995, AJ, 110, 1059
Gunn J. E. et al., 1998, AJ, 116, 3040
Gunn J. E. et al., 2006, AJ, 131, 2332
Guzik J., Seljak U., 2002, MNRAS, 335, 311
Hansen S. M., Sheldon E. S., Wechsler R. H., Koester B. P., 2007, ApJ, submitted (arXiv:0710.3780)
Jing Y., Mo H. J., Börner G., 1998, ApJ, 494, 1
Kang X., van den Bosch F. C., 2008, ApJ, 676, L101
Kauffmann G., Colberg J. M., Diaferio A., White S. D. M., 1999, MNRAS, 303, 188
Keselman J. A., Nusser A., 2007, MNRAS, 382, 1853
Klypin A., Gottlöber S., Kravtsov A. V., Khokhlov A. M., 1999, ApJ, 516, 530
Kravtsov A. V., Berlind A., Wechsler R. H., Klypin A., Gottlöber S., Allgood B., Primack J. R., 2004, ApJ, 609 35
Landy S. D., Szalay A. S., 1993, ApJ, 412, 64
Maller A. H., Berlind A. A., Blanton M. R., Hogg D. W., 2008, preprint (arXiv:0801.3286)
Mitchell J., Keeton C., Frieman J., Sheth R. K., 2005, ApJ, 622, 81
Mo H. J., White S. D. M., 1996, MNRAS, 282, 347
Navarro J. F., Frenk C. S., White S. D. M., 1997, ApJ, 490, 493
Park C., Choi Y. Y., Vogeley M. S., Gott J. R., Blanton M. R., 2007, ApJ, 658, 898
Peacock J. A., Smith R. E., 2000, MNRAS, 318, 1144
Scoccimarro R., Sheth R. K., 2002, MNRAS, 329, 629
Scoccimarro R., Sheth R. K., Hui L., Jain B., 2001, ApJ, 546, 20
Scranton R., 2002, MNRAS, 332, 697
Scranton R. et al., 2005, ApJ, 633, 589
Seljak U., 2000, MNRAS, 318, 203
Sheldon E. et al., 2007, preprint (arXiv:0709.1153)
Sheth R. K., 2005, MNRAS, 364, 796
Sheth R. K. et al., 2003, ApJ, 594, 225
Sheth R. K., Jain B., 2003, MNRAS, 345, 529
Sheth R. K., Tormen G., 1999, MNRAS, 308, 119
Sheth R. K., Tormen G., 2004, MNRAS, 350, 1385
Sheth R. K., Connolly A. J., Skibba R., 2005, preprint (astro-ph/0511773)
Sheth R. K., Diaferio A., Hui L., Scoccimarro R., 2001, MNRAS, 326, 463
Sheth R. K., Jimenez R., Panter B., Heavens A. F., 2006, ApJ, 650, L25
Skibba R. A., 2008, MNRAS, in press (doi:10.1111/j.1365-2966.2008.14152.x) (arXiv:0805.1233)
Skibba R. A., Sheth R. K., Connolly A. J., Scranton R., 2006, MNRAS, 369, 68
Skibba R. A., Sheth R. K., Martino M. C., 2007, MNRAS, 382, 1940
Smith R. E., Watts P. I. R., Sheth R. K., 2006, MNRAS, 365, 214
Stoyan D., Stoyan H., 1994, Fractals, Random Shapes, and Point Fields. Wiley, Chichester
Swanson M. E. C., Tegmark M., Blanton M., Zehavi I., 2008, MNRAS, 385, 1635
Tinker J. L., Conroy C., Norberg P., Patiri S. G., Weinberg D. H., Warren M. S., 2008, ApJ, 686, 53
Vale A., Ostriker J. P., 2006, MNRAS, 371, 1173
van den Bosch F. C. et al., 2007a, MNRAS, 376, 841
van den Bosch F. C., Aquino D., Yang X., Mo H. J., Pasquali A., McIntosh D. H., Weinmann S. M., Kang X., 2008, MNRAS, 387, 79
Wang Y., Yang X., Mo H. J., van den Bosch F. C., 2007, ApJ, 664, 608
Wechsler R. H., Zentner A. R., Bullock J. S., Kravtsov A. V., Allgood B., 2006, ApJ, 652, 71
Weinmann S. M., van den Bosch F. C., Yang X., Mo H. J., 2006, MNRAS, 366, 2
Wetzel A. R., Cohn J. D., White M., Holz D. E., Warren M. S., 2007, ApJ, 656, 139
Willmer C. N. A. et al., 2006, ApJ, 647, 853
Yang X., Mo H. J., van den Bosch F. C., 2003, MNRAS, 339, 1057
Yang X., Mo H. J., van den Bosch F. C., 2008, ApJ, 676, 248
York D. G. et al., 2000, AJ, 120, 1579
Yoshida N., Sheth R. K., Diaferio A., 2001, MNRAS, 328, 669
Zehavi I. et al., 2005, ApJ, 630, 1
Zheng Z. et al., 2005, ApJ, 633, 791
Zheng Z., Coil A. L., Zehavi I., 2007, ApJ, 667, 760
Zu Y., Zheng Z., Zhu G. T., Jing Y. P., 2008, ApJ, 686, 41

APPENDIX A: EXPLICIT EXAMPLES OF DIFFERENT HODS

The main text outlines our model; actual implementation of it depends on the form of the luminosity-based HOD. These are of two types – either the relation between halo mass and central galaxy luminosity is monotonic and deterministic, or there is some scatter.

We use the parametrization of Zehavi et al. (2005) to illustrate the former case, and that of Zheng et al. (2007) to illustrate the latter. The results described in the main text are not particularly sensitive to this choice, although the plots we are based on HODs in which there is scatter.

A1 No scatter between $L_{\rm cen}$ and halo mass

To evaluate $n_{\rm sat}/n_{\rm cen}$, suppose that the relation between halo mass and central luminosity is deterministic (i.e. there is no scatter around the $L_{\rm cen}$–m relation). Then,

$$n_{\rm cen}(L) = ({\rm d}m_L/{\rm d}L)\,({\rm d}n/{\rm d}m)_{m_L}, \tag{A1}$$

where halo model interpretations of SDSS galaxy clustering suggest that

$$\frac{m_L}{10^{12}\,h^{-1}\,{\rm M}_\odot} \approx \exp\left(\frac{L/1.12}{10^{10}\,h^{-2}\,{\rm L}_\odot}\right) - 1 \tag{A2}$$

(Zehavi et al. 2005; Skibba et al. 2006), and the halo mass function ${\rm d}n/{\rm d}m$ is described in Sheth & Tormen (1999).

The number density of satellite galaxies which have luminosity L, whatever the mass of the parent halo, is given by differentiating

$$n_{\rm sat}(>L) = \int_{m_L}^{\infty} {\rm d}m\,\frac{{\rm d}n}{{\rm d}m}\,N_{\rm sat}(>L|m) \tag{A3}$$

with respect to L. Zehavi et al. (2005) show that

$$N_{\rm sat}(>L|m) \approx \left(\frac{m}{23\,m_L}\right)^{\alpha_L}, \tag{A4}$$

where m_L is given by the expression above, and

$$\alpha_L \approx 1.16 - 0.1(M_r+20) + 0.1{\rm e}^{-1.5(M_r+21.5)^2} \tag{A5}$$

is a weakly increasing function of L. Thus,

$$n_{\rm sat}(>L) = \left[\exp\left(\frac{L/1.12}{10^{10}\,h^{-2}\,{\rm L}_\odot}\right) - 1\right]^{-\alpha_L} \times \int_{m_L}^{\infty} {\rm d}m\,\frac{{\rm d}n}{{\rm d}m}\left(\frac{m/23}{10^{12}\,h^{-1}\,{\rm M}_\odot}\right)^{\alpha_L}, \tag{A6}$$

so

$$\frac{n_{\rm sat}(L)}{n_{\rm cen}(L)} = \frac{1}{23^{\alpha_L}} + \alpha_L\,\frac{n_{\rm sat}(>L)}{({\rm d}n/{\rm d}\ln m)_{m_L}} - \alpha_L\,\frac{{\rm d}\ln\alpha_L}{{\rm d}\ln L}\,\frac{{\rm d}\ln L/{\rm d}\ln m_L}{({\rm d}n/{\rm d}\ln m)_{m_L}} \times \int_{m_L}^{\infty} {\rm d}m\,\frac{{\rm d}n}{{\rm d}m}\left(\frac{m/23}{m_L}\right)^{\alpha_L} \ln\left(\frac{m/23}{m_L}\right), \tag{A7}$$

and the fraction of objects which are centrals is

$$f_{\rm cen}(L) = \frac{n_{\rm cen}(L)}{n_{\rm cen}(L)+n_{\rm sat}(L)} = \frac{1}{1+n_{\rm sat}(L)/n_{\rm cen}(L)}. \tag{A8}$$

To see what these expressions imply, suppose that α_L were independent of L. Then,

$$\frac{n_{\rm sat}(L)}{n_{\rm cen}(L)} = \alpha\,\frac{n_{\rm sat}(>L)}{({\rm d}n/{\rm d}\ln m)_{m_L}} + \frac{1}{23^{\alpha}}, \tag{A9}$$

and

$$\langle c|m\rangle_{\rm sat} = \langle c|L_{\rm min}\rangle_{\rm red} + \int_{L_{\rm min}}^{\infty} {\rm d}L\,C(L)\left(\frac{m_{L_{\rm min}}}{m_L}\right)^{\alpha}. \tag{A10}$$

In this case, the mean satellite colour is independent of halo mass. If $\alpha = 1$ (not far off from its actual value) and $m({\rm d}n/{\rm d}m) \propto \exp(-m/m_*)/m_*$ for some fiducial value of m_* (haloes more massive than $m_* \approx 10^{13}\,h^{-1}\,{\rm M}_\odot$ are indeed exponentially rare), then $n_{\rm sat}(>L) = \exp(-m_L/m_*)/(23m_L)$ making $n_{\rm sat}(L)/n_{\rm cen}(L) = (m_*/m_L + 1)/23$. This ratio decreases as m_L increases – as L increases, the ratio of satellites to centrals decreases, and the fraction of centrals increases.

In the analyses which follow, we use the actual halo model values of these quantities rather than these approximations. A reasonable fit to the actual halo model values is given by

$$\frac{n_{\rm sat}(L)}{n_{\rm cen}(L)} \approx 0.35\left[2 - {\rm erfc}\left[0.6(M_r+20.5)\right]\right] \tag{A11}$$

This ratio tends to 0.7 at small luminosities, making the fraction of galaxies which are centrals at $L \ll 10^{10}\,h^{-2}\,{\rm L}_\odot$ about 3/5 (cf. equation A8) consistent with the satellite fraction $f_{\rm sat}(L)$ of van den Bosch et al. (2007a).

A2 Stochasticity in the $L_{\rm cen}-m$ relation

Zheng et al. (2007) allow for stochasticity in the relation between halo mass and central galaxy luminosity. They assume that

$$P(\log L_{\rm cen}|M) = \frac{1}{\sqrt{2\pi}\,\sigma_{\log L}}\exp\left[-\frac{[\log(L_{\rm cen}/\langle L_{\rm cen}|M\rangle)]^2}{2\sigma^2_{\log L}}\right], \tag{A12}$$

and then set

$$\langle N_{\rm cen}|M\rangle = \frac{1}{2}\left[1 + {\rm erf}\left(\frac{\log(M/M_{\rm min})}{\sigma_{\log M}}\right)\right] \tag{A13}$$

and

$$\langle N_{\rm sat}|M\rangle = \left(\frac{M-M_0}{M_1'}\right)^{\alpha}. \tag{A14}$$

The Poisson model for satellite counts sets

$$\langle N_{\rm sat}(N_{\rm sat}-1)|M\rangle = \langle N_{\rm sat}|M\rangle^2. \tag{A15}$$

Their table 1 shows how all of the parameters in this HOD vary with SDSS r-band luminosity. We have found that these scalings with L_r are well approximated by

$$\frac{M_{\rm min}}{10^{11.95}\,{\rm M}_\odot\,h^{-1}} \approx \exp\left(\frac{L}{10^{10.0}\,{\rm L}_\odot\,h^{-2}}\right) - 1 \tag{A16}$$

$$\sigma_{\log M} \approx \begin{cases} 0.26 \text{ if } M_r > -20.5 \\ 0.385 - 0.25\,(M_r+21), \text{ otherwise} \end{cases} \tag{A17}$$

$$M_1' \approx 17\,M_{\rm min} \tag{A18}$$

$$\frac{M_0}{10^{11.75}\,{\rm M}_\odot\,h^{-1}} \approx \left(\frac{L}{10^{9.9}\,{\rm L}_\odot\,h^{-2}}\right)^{0.6} \tag{A19}$$

$$\alpha \approx 1 - 0.07\,(M_r+18.8). \tag{A20}$$

As in Zehavi et al. (2005), the value of $M_1'/M_{\rm min}$, which determines the critical mass above which haloes typically host at least one satellite galaxy, is approximately independent of luminosity, while the $\langle N_{\rm sat}\rangle$ slope α, which characterizes the mass dependence of the efficiency of galaxy formation, increases with luminosity. The two new HOD parameters are $\sigma_{\log M}$ and M_0. They are not constrained well and their uncertainties are large (see Zheng et al. for details), but our correlation functions and colour–mark correlation functions are not very sensitive to their exact values.

For the two luminosity thresholds discussed in the main text, $M_r < -19.5$ and -20.5, the parameters above are $M_{\min} = 5.8 \times 10^{11}$ and $2.2 \times 10^{12}\, h^{-1}$ Mpc, and the effective value of $M_1/M_{\min} \approx 20$, approximately independent of luminosity, is similar to the factor of 23 in the Zehavi et al. HOD.

For our purposes, the main difference with this HOD model is the scatter in luminosity at fixed mass. We first discuss how to construct a mock catalogue that includes this scatter. We then explain how our model of the colour mark is modified.

To account for the scatter between $L_{\rm cen}$ and $M_{\rm halo}$ in the mock catalogues, we do not simply select the subset of haloes in the simulation which have $M > M_{\min}(L_{\min})$, as we do in the case of a sharp threshold (e.g. Section 3.1). Instead, we generate uniformly distributed random numbers u between 0 and 1 for each halo of mass M. Then we keep the halo if $u < \langle N_{\rm cen}\, |M\rangle$ (equation A13). As a result, only half of the haloes with $M \approx M_{\min}$ are kept, as are quite a few haloes with $M < M_{\min}$. Larger values of $\sigma_{\log M}$ increase the range of halo masses around $M_{\min}$ and increase the total number of haloes because the abundance of haloes increases with decreasing mass.

Our halo model of the colour mark is also modified by the scatter between luminosity and mass, and hence $\langle N_{\rm cen}\, |M, L_{\min}\rangle$ is no longer a step function. The central galaxy colour mark, described in Section 2, is slightly more complicated. The mean central galaxy colour as a function of luminosity $\langle c|L\rangle_{\rm cen}$ (equation 15) depends on the number density of central galaxies as a function of luminosity $n_{\rm cen}(L)$ (equation A1), which now includes an integral

$$
\begin{aligned}
n_{\rm cen}(L) &= \frac{\mathrm{d}}{\mathrm{d}L}\, n_{\rm cen}(> L) \\
&= \left(\frac{\mathrm{d}n}{\mathrm{d}M}\right)_{M_L} \left(\frac{\mathrm{d}M_L}{\mathrm{d}L}\right) \\
&\quad + \int \mathrm{d}M\, \frac{\mathrm{d}n}{\mathrm{d}M}\, \frac{\mathrm{d}}{\mathrm{d}L} \langle N_{\rm cen}|M, L_{\min}\rangle. \qquad \text{(A21)}
\end{aligned}
$$

Then the central galaxy colour mark, which is used in the colour–mark correlation functions, is also an integral (equation 16),

$$
\langle c|M\rangle_{\rm cen} = \int \mathrm{d}L\, P_{\rm cen}(L|M)\, \langle c|L\rangle_{\rm cen}. \qquad \text{(A22)}
$$

The model of the colour–mark correlation functions, described in Appendix B, is also modified. However, we reiterate that, in general, the correlation functions and colour–mark correlation functions are not sensitive to the exact amount of scatter in mass at a fixed luminosity.

APPENDIX B: A HALO MODEL OF COLOUR–MARK CORRELATIONS

We perform our halo model calculations in Fourier space. The two-point correlation function is the Fourier transform of the power spectrum

$$
\xi(r) = \int \frac{\mathrm{d}k}{k}\, \frac{k^3 P(k)}{2\pi^2}\, \frac{\sin kr}{kr}. \qquad \text{(B1)}
$$

In the halo model, $P(k)$ is written as the sum of two terms: one that arises from galaxies within the same halo and dominates on small scales (the one-halo term), and the other from galaxies in different haloes which dominate on larger scales (the two-halo term). That is,

$$
P(k) = P_{1h}(k) + P_{2h}(k), \qquad \text{(B2)}
$$

where

$$
\begin{aligned}
P_{1h}(k) = \int \mathrm{d}M\, \frac{\mathrm{d}n(M)}{\mathrm{d}M}\, \langle N_{\rm cen}|M\rangle
&\times \left[\frac{2\,\langle N_{\rm sat}|M\rangle\, u_{\rm gal}(k|M)}{\bar{n}_{\rm gal}^2} \right. \\
&\left. + \frac{\langle N_{\rm sat}(N_{\rm sat}-1)|M\rangle\, u_{\rm gal}(k|M)^2}{\bar{n}_{\rm gal}^2} \right], \qquad \text{(B3)}
\end{aligned}
$$

$$
P_{2h}(k) = \left[\int \mathrm{d}M\, \frac{\mathrm{d}n(M)}{\mathrm{d}M}\, \langle N_{\rm cen}|M\rangle \times \frac{1 + \langle N_{\rm sat}|M\rangle\, u_{\rm gal}(k|M)}{\bar{n}_{\rm gal}}\, b(M) \right]^2 P_{\rm lin}(k), \qquad \text{(B4)}
$$

where the number density of galaxies $\bar{n}_{\rm gal}$ is (*cf.*, equation A3)

$$
\bar{n}_{\rm gal} = \int dm\, \frac{\mathrm{d}n(m)}{\mathrm{d}m}\, \langle N_{\rm cen}|m\rangle \left[1 + \langle N_{\rm sat}|m\rangle\right] \qquad \text{(B5)}
$$

and $u_{\rm gal}(k|M)$ is the Fourier transform of the galaxy density profile. It is standard to assume this has the same form as the dark matter, so we use the form for u given by Scoccimarro et al. (2001). The distribution $p_{\rm sat}(N_{\rm sat})$ is expected to be well-approximated by a Poisson distribution (e.g. Kravtsov et al. 2004; Yang, Mo & van den Bosch 2008), so we set $\langle N_{\rm sat}(N_{\rm sat} - 1)|M\rangle = \langle N_{\rm sat}|M\rangle^2$. The two parts of the one-halo term in equation (B3) can be thought of as the 'centre-satellite term' and the 'satellite-satellite term.'

To describe the effect of weighting each galaxy, we use $W(k)$ to denote the Fourier transform of the weighted correlation function. Like the power spectrum, we write this as the sum of one- and two-halo terms: $W(k) = W_{1h}(k) + W_{2h}(k)$. Since central and satellite galaxies have different properties, we weight central and satellite galaxies separately by their mean mass-dependent marks: $\langle c|m\rangle_{\rm cen}$ and $\langle c|m\rangle_{\rm sat}$ (Section 2). Following Sheth (2005), we write

$$
\begin{aligned}
W_{1h}(k) = \int \mathrm{d}M\, \frac{\mathrm{d}n(M)}{\mathrm{d}M}\, \langle N_{\rm cen}|M\rangle
&\times \left[\frac{2\, c_{\rm cen}(M)\, \langle c_{\rm sat}|M, L_{\min}\rangle\, \langle N_{\rm sat}|M\rangle\, u_{\rm gal}(k|M)}{\bar{n}_{\rm gal}^2\, \bar{c}^2} \right. \\
&\left. + \frac{\langle N_{\rm sat}|M\rangle^2\, \langle c_{\rm sat}|M, L_{\min}\rangle^2\, u_{\rm gal}^2(k|M)}{\bar{n}_{\rm gal}^2\, \bar{c}^2} \right], \qquad \text{(B6)}
\end{aligned}
$$

$$
\frac{W_{2h}(k)}{P_{\rm lin}(k)} = \left[\int \mathrm{d}M\, \frac{\mathrm{d}n(M)}{\mathrm{d}M}\, \langle N_{\rm cen}|M\rangle\, b(M) \times \frac{c_{\rm cen}(M) + \langle N_{\rm sat}|M\rangle\, \langle c_{\rm sat}|M, L_{\min}\rangle\, u_{\rm gal}(k|M)}{\bar{n}_{\rm gal}\, \bar{c}} \right]^2, \qquad \text{(B7)}
$$

where we normalize by the mean colour mark

$$
\bar{c} = \int \mathrm{d}M\, \frac{\mathrm{d}n(M)}{\mathrm{d}M}\, \langle N_{\rm cen}|M\rangle \times \frac{c_{\rm cen}(M) + \langle N_{\rm sat}|M\rangle\, \langle c_{\rm sat}|M, L_{\min}\rangle}{\bar{n}_{\rm gal}}. \qquad \text{(B8)}
$$

This paper has been typeset from a TEX/LATEX file prepared by the author.