

Systematic errors in weighted two-point correlation functions: an application to interaction-induced star formation

Aday R. Robaina^{1★} and Eric F. Bell²

¹*Institut de Ciències del Cosmos, Universitat de Barcelona, Martí y Franqués 1, E-08028 Barcelona, Spain*

²*Department of Astronomy, University of Michigan, 550 Church Street, Ann Arbor, MI 48109, USA*

Accepted 2011 December 2. Received 2011 November 7; in original form 2009 December 10

ABSTRACT

Weighted correlation functions are an increasingly important tool for understanding how galaxy properties depend on their separation from each other. We use a mock galaxy sample drawn from the Millennium Simulation, assigning weights using a simple prescription to illustrate and explore how well a weighted correlation function recovers the true separation dependence of the input weights. We find that the use of a weighted correlation function results in a dilution of the magnitude of any separation dependence of properties and a smearing out of that dependence in radius, compared to the input behaviour. We present a quantitative discussion of the dilution in the magnitude of separation dependence in properties in the special case of a constant enhancement at $r < r_c$. In this particular case where there was a star formation rate (SFR) enhancement at small radius $r < r_c = 35$ kpc, the matching of one member of an enhanced pair with an unenhanced galaxy in the same group gives an artificial enhancement out to large radius ~ 200 kpc. We compare this with the observations of SFR enhancement from the Sloan Digital Sky Survey, finding very similar behaviour – a significant enhancement at radii < 40 kpc and a weak enhancement out to more than 150 kpc. While we explore a particular case in this paper, it is easy to see that the phenomenon is general, and precision analysis of weighted correlation functions will need to account carefully for this effect using simulated mock catalogues.

Key words: galaxies: general – galaxies: statistics.

1 INTRODUCTION

Correlation functions (two-point and higher order ones) have proved to be powerful statistical tools in order to address the study of the galaxy clustering (e.g. Peebles & Groth 1976; Groth & Peebles 1977; Peebles 1980; Davis & Peebles 1983) and are still widely used in both the local (Connolly et al. 2002; Eisenstein et al. 2005; Masjedi et al. 2006) and high-redshift Universe (Giavalisco et al. 1998; Blain et al. 2004). Studies of the two-point correlation function have matured to the point that one can study how galaxies populate dark matter haloes in detail (e.g. Zehavi et al. 2004), the typical halo masses of galaxy populations as a function of redshift (e.g. Lyman breaks; Giavalisco et al. 1998), the relative clustering of different populations [e.g. the tendency of active galactic nucleus (AGN) to cluster like the massive galaxy population as a whole; Li et al. 2006], and use of clustering measures on the smallest scale to constrain the merger history of galaxies (e.g. Patton et al. 2002; Bell et al. 2006; Robaina et al. 2010).

Furthermore, the correlation function method allows us not only to study the clustering of the galaxies themselves, but also how some of their properties are clustered. Weighted correlation functions (Boerner, Mo & Zhou 1989) or in a general sense, marked statistics (Beisbart & Kerscher 2000; Faltenbacher et al. 2002; Gottlöber et al. 2002; Skibba et al. 2006; Robaina et al. 2009) have been widely used in the last 10 years in order to study how observables depend on the separation between galaxies. In particular, weighted correlation functions are frequently used to study the dependence of star formation rate (SFR) on separation between galaxies, in great part to explore the influence of galaxy interactions on enhancing a galaxy pair's SFR (e.g. Li et al. 2008; Robaina et al. 2009).

The goal of this paper is to explore the application of weighted correlation functions to study the variation of observables (e.g. SFR, colour, AGN accretion rate, morphology) as a function of radius. We briefly introduce weighted two-point correlation functions in Section 2. We then construct a toy model with which we study the behaviour of the inferred weighted quantities relative to the input behaviour (Section 3). This toy model primarily illustrates some general features of how weighted correlation functions recover input behaviour, and we stress that the framework discussed in this paper applies generally to any application of weighted correlation

★E-mail: arobaina@gmail.com

function analysis, while noting that we choose to present a case that is most directly analogous to the study of SFR enhancement in close pairs of galaxies. We show the results of this analysis in Section 4. In Section 5, we briefly compare with observational results of star formation (SF) enhancement derived using the Sloan Digital Sky Survey (SDSS; Li et al. 2008). In Section 6, we present our conclusions. When necessary, we have assumed $H_0 = 70 \text{ km s}^{-1}$, $\Omega_{\text{m}0} = 0.3$ and $\Omega_{\Lambda 0} = 0.7$.

2 BACKGROUND

In this work, we explore the possible artefacts that the use of a marked correlation function could introduce when studying the clustering of galaxy properties. A full explanation of the methodology followed in this work has been already presented in Robaina et al. (2009), and is similar to the methodology adopted by Skibba et al. (2006) and Li et al. (2008); we summarize here the basics of the method but we refer the reader to those papers for a deeper explanation.

The two-point correlation function $\xi(r)$ is the *excess* probability of finding a galaxy at a given distance r from another galaxy:

$$dP = n[1 + \xi(r)]dV, \quad (1)$$

where dP is the probability of finding a galaxy in volume element dV at a distance r from a galaxy, and n is the galaxy number density. A simple estimator of the unweighted correlation function is $\xi(r) \simeq \text{DD}/\text{RR} - 1$, where DD is the histogram of separations between galaxies and RR is the histogram of separations between galaxies in a randomly distributed catalogue. In a similar way, one can estimate the weighted correlation function as $W(r) \simeq \text{PP}/\text{PP}_R - 1$, where PP is the weighted histogram of real galaxies and PP_R the weighted histogram of separations from the catalogues with randomized coordinates.

We choose to use an additive weighting scheme (the weight of the pair is the sum of the weights of individual galaxies) for concreteness (e.g. Robaina et al. 2009), while noting that a multiplicative weighting would yield a qualitatively similar result. Then, we can define the ‘mark’ $E(r)$ as the excess clustering of the weighted correlation function compared to the unweighted correlation function:

$$E(r) = \frac{1 + W(r)}{1 + \xi(r)}. \quad (2)$$

3 AN IDEALIZED EXPERIMENT

We use De Lucia et al. (2006) catalogue at $z = 0$ derived from the Millennium Simulation (Springel et al. 2005) in order to study how the enhancement in a physical quantity caused by a galaxy–galaxy interaction (e.g. a SF enhancement) would be recovered by weighted two-point correlation function techniques. We manually assign a weight (we refer to it as the mark) to every galaxy in the sample, giving a mark = 1 to galaxies which are *not* closer than $r_c = 35 \text{ kpc}$ to any other galaxy and mark = ϵ (with $\epsilon > 1$) to those galaxies which are in close, 3D pairs with separation $r < r_c$ kpc. For concreteness, we consider simulated galaxies with stellar masses $M_* > 2.5 \times 10^{10} M_\odot$, noting that the conclusions reached in this paper are generally applicable, in a qualitative sense.

We now examine how the marks of galaxy pairs relate to the actual behaviour of the enhancement as a function of separation from their nearest neighbour. The mark is estimated by dividing the weighted correlation function by its unweighted counterpart, and recall that the correlation function relates every galaxy to every

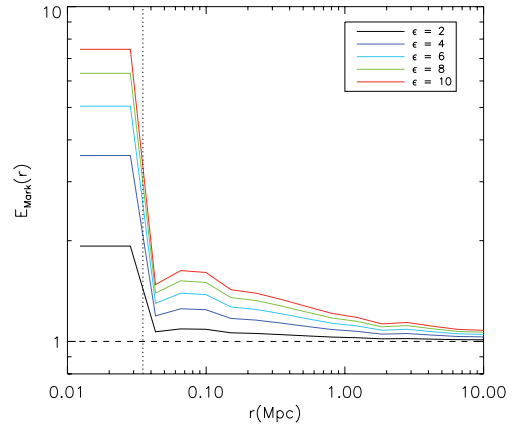


Figure 1. Apparent enhancement as a function of distance and ‘real’ enhancement ϵ (which acts only at $r < r_c$ kpc) computed for a sample with a minimum stellar mass of $2.5 \times 10^{10} M_\odot$ and different values of ϵ . The vertical dotted line shows the separation r_c (35 kpc in this case). A tail of artificial enhancement extending to large separations is produced as an artefact of the weighted correlation function method. The enhancement recovered in the close pairs $r < r_c$ kpc is reduced with respect to ϵ and the level of reduction is a strong function of ϵ (see the text and Fig. 2 for more details). The apparent minimum at $r \simeq 0.04 \text{ Mpc}$ is an artefact due to the relation between the finite bin size and the chosen scale r_c .

other galaxy in the sample.¹ The weight is additive, and since every galaxy with a companion closer than r_c has weight ϵ , the mark of a close pair is 2ϵ . Yet, the galaxies in this close pair will be matched also to every other galaxy in the sample. Therefore, when a galaxy in the same group or cluster at a distance $r > r_c$ from the enhanced pair is matched with a galaxy in the pair, the mark of that pair will be $\epsilon + 1$ (1 being the default weight of non-enhanced galaxies). We see that a pair with $r > r_c$ will show an enhancement when, in reality, there is no physical interaction-induced enhancement at that radius. As that third galaxy will be matched with *both* galaxies in the neighbour close pair, two pairs with mark = $\epsilon + 1$ will be contributed. Furthermore, imagine now that there is another real close pair of galaxies placed at several Mpc from the first close pair, in which both galaxies will also have mark = ϵ . From matching all those four galaxies, the final product will be six galaxy pairs displaying mark = 2ϵ . This will clearly affect both the normalization of the mark and the recovered value for the enhancement, producing a tail of false enhancement in the regions where more companions would be found (representing dense regions of the Universe) and decreasing the enhancement found at $r < r_c$.

4 RESULTS

We show this effect in Fig. 1. Clearly, a relatively weak tail of enhancement is recovered out to large separations. The amplitude of this tail has a radial dependence, as close pairs of galaxies tend to be found in dense regions of the Universe (Barton et al. 2007). As the magnitude of this tail depends on the distribution of neighbours as a function of the separation it will be more relevant for galaxy samples in which the clustering is stronger (e.g. massive galaxies, or non-star-forming galaxies).

Also visible in Fig. 1 is the dilution of the recovered enhancement compared with the actual enhancement ϵ for pairs with $r < r_c$,

¹ Even in the case in which some criteria for pair-matching are imposed, such as line-of-sight constraints, mass ratio, etc., one particular galaxy will be matched with many secondaries at very different separations.

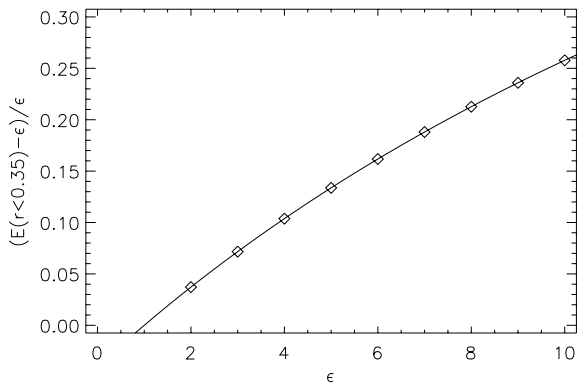


Figure 2. Relative error in $E(r < r_c)$, the enhancement recovered by the marked correlation function in close pairs as a function of the ‘real’ enhancement ϵ in those pairs. For this example we have used, as in Fig. 1, a lower mass cut of $2.5 \times 10^{10} M_\odot$. Diamonds: recovered values from the method. Solid line: expected value using the proper normalization shown in equation (4). The error when the intrinsic enhancement is small is modest; when $\epsilon < 4$ then the discrepancy between $E(r < r_c)$ and ϵ is < 10 per cent.

$E(r < r_c)$. The value of $E(r < r_c)$ is lower than the ‘real’ enhancement ϵ by a factor which increases with ϵ . This effect is better seen in Fig. 2 where we show the relative discrepancy between $E(r < r_c)$ and ϵ as a function of ϵ . In this idealized case, this discrepancy can be exactly recovered by accounting carefully for the different pairs formed by galaxies in the sample. The relationship between $E(r < r_c)$ and ϵ is

$$E(r < r_c) = \frac{\epsilon N_{p,\text{tot}}}{W_{\text{cp,max}} N_{\text{cp,max}} + W_{\text{mp}} N_{\text{mp}} + W_{\text{fp}} N_{\text{fp}}}, \quad (3)$$

where $N_{p,\text{tot}}$ is the total number of pairs which can be formed from the galaxy sample,² $N_{\text{cp,max}}$ is the total number of pairs which can be formed with galaxies belonging to close pairs,³ $W_{\text{cp,max}}$ is the weight associated with those pairs, N_{pm} is the number of pairs in which only one galaxy belongs to a close pair, W_{mp} is the weight associated with them, and W_{fp} and $N_{\text{p, far}}$ are, respectively, the weight and the number of pairs in which none of the galaxies belongs to a close pair.

In our particular case of an additive weight, this expression reduces to

$$E(r < r_c) = \frac{2\epsilon}{(f^2 + f)(\epsilon - 1) + 2}, \quad (4)$$

where f is the fraction of galaxies in close pairs. The degree of clustering of the sample is reflected in the value of f , so this expression is valid under different clustering conditions. For the purposes of this work, we calculate f *directly* from the mock catalogue, but real galaxy surveys lack of accurate 3D information. It is common to calculate f from the inferred real-space correlation function by integrating equation (1) out to r_c (Bell et al. 2006; Masjedi et al. 2006). In their analysis, in the limit of small r_c , and if the correlation function is parametrized as a power law $\xi(r) = (r/r_0)^{-\gamma}$, then

$$P(r < r_c) = f = \int_0^{r_c} n[1 + \xi(r)]dV, \quad (5)$$

$$f \simeq \frac{4\pi n}{3 - \gamma} r_0^\gamma r_c^{3-\gamma}. \quad (6)$$

² When performing an autocorrelation, the total number of unique pairs would be $N(N - 1)/2$, N being the number of galaxies in the sample.

³ This is *not* the same as the number of close pairs, as we already explained.

It is worth noting that in the above example we have studied the simple case in which the enhancement is present only in close galaxy pairs, with the enhancement represented by a step function. When applying weighted correlation functions to more complex problems, like those involving clustering of the mass or colour, the function describing the behaviour of the weight on separation would be much more complex. In that case, an expression for the behaviour of the weight as a function of separation will have to be derived on a case-by-case basis and matched with the data. Yet, even in that more complex case, the underlying problem is very similar: the magnitude of any radial dependence in properties will be diluted and smeared out in radius by the use of weighted two-point correlation function methods.

5 AN EXAMPLE APPLICATION TO OBSERVATIONS

In order to test the relevance of this analysis to the real Universe, we compare our predictions with a well-established phenomenon: the enhancement of the SFR in galaxy interactions. This observable has two obvious advantages. First, there are a number of works in which this enhancement has been studied (Barton, Geller & Kenyon 2000; Lambas et al. 2003; Li et al. 2008; Robaina et al. 2009). Secondly, the SFR is expected to be enhanced only at scales at which galaxy–galaxy interactions are relevant; beyond that scale star formation is not only not expected to be enhanced, but should be depressed because of the well-known SFR–density anticorrelation (e.g. Balogh et al. 2002). From the above-mentioned works we choose to compare with Li et al. (2008) for three reasons: (a) they use marked statistics, (b) their large sample allowed an accurate estimate of enhancement to be made and (c) SDSS clustering has been shown to be similar to the one present in the De Lucia et al. (2006) mock catalogue from the Millennium Simulation in the local Universe (Springel et al. 2005).

Real galaxy surveys, even spectroscopic surveys, have no access to the real-space separation of galaxies. Li et al. (2008) used a projected correlation function $w(r_p)$ to circumvent this difficulty, where the projected correlation function is related to the 3D correlation function via

$$w(r_p) = \int_{-\infty}^{\infty} \xi \left(\left[r_p^2 + \pi^2 \right]^{1/2} \right) d\pi, \quad (7)$$

where π is the coordinate along the line of sight, and r_p is the projected separation transverse to the line of sight. We use for this exercise galaxies more massive than $3 \times 10^{10} M_\odot$ in order to match the selection criteria in Li et al. (2008). Moreover, they did not use an additive weight but used the Specific Star Formation Rate (SSFR) of the primary galaxy as the weight of the pair. We also use such a scheme here to perform our weighted analysis in the simulation. Li et al. (2008) calculated the cross-correlation between a subsample of galaxies which form stars (primaries) and all the galaxies in the sample (secondaries). We choose to mimic this selection, to good approximation, by equating star-forming and non-star-forming galaxies to observations of the halo mass dependence of blue and red galaxies, respectively. We determine the dependence of red galaxy fraction on group halo mass for galaxies above our stellar mass cut in the observational group catalogue of Yang et al. (2007). We then label galaxies in the Millennium Simulation as star-forming and non-star-forming following the observed trend between halo mass and red fraction. In this way, we accurately reproduce the overall trend in non-star-forming fraction with halo mass. With this halo mass dependent prescription for star-forming fraction, the number

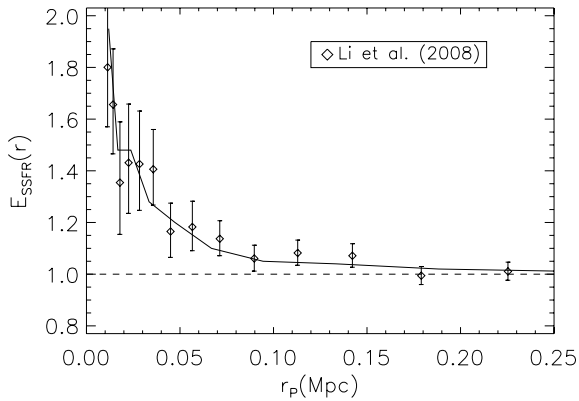


Figure 3. Specific SFR enhancement in (massive) galaxy pairs as function of the projected separation as measured by Li et al. (2008; diamonds) and our prediction including the tail of artificial enhancement out to several hundred kiloparsecs (solid line). In this example a value of $\epsilon = 2$ has been used for galaxies in pairs with $r_p < 15$ kpc and $\epsilon = 1.5$ for galaxies in pairs with $15 < r_p < 40$ kpc. The galaxy sample has been selected to be consistent with the massive sample in Li et al. (2008).

densities and clustering strength of the star-forming and non-star-forming populations will be as close to realistic as is currently feasible using the Millennium Simulation (Li & White 2010). We then assign an enhancement to all star-forming galaxies found physically in very close pairs, but in order to mimic the pair selection in Li et al. (2008), who cross-correlate a sample of spectroscopically defined star-forming galaxies with a photometric catalogue of all galaxies above the stellar mass limit, we run the correlation function selecting galaxy pairs with ‘line-of-sight’ separations of less than 100 Mpc. Our results are not sensitive to this choice of maximum separation; correlations between galaxies on scales larger than 100 Mpc are extremely weak, in comparison to the strong clustering on < 1 Mpc scales. Motivated by the star formation enhancement observed in galaxy samples selected in a similar manner at different redshifts (Li et al. 2008; Robaina et al. 2009) we choose to model the data with $\epsilon = 2$ for star-forming galaxies in pairs with separations $r_p < 15$ kpc and $\epsilon = 1.5$ for those in pairs with $15 < r_p < 40$ kpc. We also neglect any environmental suppression of star formation at separations $r > r_c$ (Barton et al. 2000; Balogh et al. 2002). These are clearly oversimplifications, as the real dependence of enhancement (and suppression at large radii) on separation will be considerably more complex. Yet, this simple model suffices to illustrate the recovered enhancement signature expected from a model in which SF is enhanced only at small radii.

Notwithstanding these limitations, we compare the results of our simple model with the data in Fig. 3. Strikingly, we find that the tail of enhanced SF out to ~ 200 kpc seen in the data may, in great part, be a reflection of the use of marked correlation function statistics to explore the radial dependence of SF enhancement in galaxies. This has direct relevance in the interpretation of the results from Li et al. (2008). If one argued that the enhancement at ~ 100 kpc (or much of it) was real, one would need to fulfil two criteria to produce such an effect. First, assuming that the triggering event is the first pass, one would need an enhancement lifetime of at least 300 Myr (longer than the internal dynamical time) for typical orbital velocities of 300 km s^{-1} or less. Secondly, a significant fraction of the secondaries would need to have near-radial orbits in order to produce such an enhancement. If, as we suggest instead, the enhanced SF at ~ 100 kpc is an artefact of the use of the two-point correlation function, then one would argue that enhancement

happens only for close pairs and shorter interaction-induced SF time-scales and a greater diversity of orbits would be permitted. While developing a model that realistically reproduces the data is beyond the scope of this paper, one can clearly see that this effect needs to be accounted for in order to robustly interpret the behaviour of marked correlation functions.

6 CONCLUSIONS

Weighted correlation functions are an increasingly important tool for understanding how galaxy properties depend on their separation from each other. We use a mock galaxy sample drawn from the Millennium Simulation, assigning weights using a simple prescription to illustrate and explore how well a weighted correlation function recovers the true radial dependence of the input weights. We find that the use of a weighted correlation function results in a dilution of the magnitude of any radial dependence of properties and a smearing out of that radial dependence in radius, compared to the input behaviour. We present a quantitative discussion of the dilution in the magnitude of radial dependence in properties in the special case of a constant enhancement ϵ for pairs separated by $r < r_c$. In this particular case the matching of one member of an enhanced pair with an unenhanced galaxy in the same group gives an artificial enhancement of $\sim 0.1\epsilon$ out to large radii $\lesssim 5r_c$, and matching of one member of an enhanced pair with a member of another very distant enhanced pair pulls down the value of the recovered enhancement, with the discrepancy between the input and recovered enhancement being a function of the fraction of galaxies in close pairs and the value of the input enhancement. This systematic error is < 10 percent for enhancements $\epsilon < 4$, but precision measurements should account for this effect. We compare these results with observations of SFR enhancement from the SDSS Li et al. (2008), finding very similar behaviour – a significant enhancement at radii < 40 kpc and a weak enhancement out to more than 150 kpc, lending credibility to the notion that weak enhancement in SFR seen out to large radii is an artefact of the use of weighted correlation function statistics. While we explored a particular case in this paper, it is easy to see that the phenomenon is general.

Given this difference between input weights and those recovered by the weighted two-point correlation function, one might ask if one should not use a different method to explore radial trends in observables. We would argue that most different methods boil down to weighted two-point correlation functions implicitly anyway, and that one is stuck at least at the qualitative level with the differences between input and recovered weights that we have discussed above. For example, partnering projected pairs into different ‘pairs’ (i.e. not matching every galaxy with every other galaxy) suffers from two drawbacks: this is still a projected analysis, and many projected close pairs will be separated by significant distances along the line of sight; and secondly, one may choose the wrong galaxy to partner with, a particularly acute issue for triplets or groups of galaxies. One can see that such a method will suffer from a similar suppression of enhancement from the inclusion of non-pairs in the pair sample; of course, radial smearing is not possible in such a case, as there is only one radial bin. We conclude that those wishing to quantitatively analyse weighted correlation functions (or related observables) will need to account carefully for this effect using an analysis of simulated mock catalogues.

ACKNOWLEDGMENTS

We thank Cheng Li for sharing with us the electronic version of the results of their work and the referee for useful comments which

improved this manuscript. ARR acknowledges support FP7-IDEAS-Phys.LSS 240117.

REFERENCES

- Balogh M. L., Couch W. J., Smail I., Bower R. G., Glazebrook K., 2002, *MNRAS*, 335, 10
- Barton E. J., Geller M. J., Kenyon S. J., 2000, *ApJ*, 530, 660
- Barton E. J., Arnold J. A., Zentner A. R., Bullock J. S., Wechsler R. H., 2007, *ApJ*, 671, 1538
- Beisbart C., Kerscher M., 2000, *ApJ*, 545, 6
- Bell E. F., Phleps S., Somerville R. S., Wolf C., Borch A., Meisenheimer K., 2006, *ApJ*, 652, 270
- Blain A. W., Chapman S. C., Smail I., Ivison R., 2004, *ApJ*, 611, 725
- Boerner G., Mo H., Zhou Y., 1989, *A&A*, 221, 191
- Connolly A. J. et al., 2002, *ApJ*, 579, 42
- Davis M., Peebles P. J. E., 1983, *ApJ*, 267, 465
- De Lucia G., Springel V., White S. D. M., Croton D., Kauffmann G., 2006, *MNRAS*, 366, 499
- Eisenstein D. J. et al., 2005, *ApJ*, 633, 560
- Faltenbacher A., Gottlöber S., Kerscher M., Muumliller V., 2002, *A&A*, 395, 1
- Giavalisco M., Steidel C. C., Adelberger K. L., Dickinson M. E., Pettini M., Kellogg M., 1998, *ApJ*, 503, 543
- Gottlöber S., Kerscher M., Kravtsov A. V., Faltenbacher A., Klypin A., Muumliller V., 2002, *A&A*, 387, 778
- Groth E. J., Peebles P. J. E., 1977, *ApJ*, 217, 385
- Lambas D. G., Tissera P. B., Alonso M. S., Coldwell G., 2003, *MNRAS*, 346, 1189
- Li C., White S. D. M., 2010, *MNRAS*, 407, 515
- Li C., Kauffmann G., Wang L., White S. D. M., Heckman T. M., Jing Y. P., 2006, *MNRAS*, 373, 457
- Li C., Kauffmann G., Heckman T. M., Jing Y. P., White S. D. M., 2008, *MNRAS*, 385, 1903
- Masjedi M. et al., 2006, *ApJ*, 644, 54
- Patton D. R. et al., 2002, *ApJ*, 565, 208
- Peebles P. J. E., 1980, Research supported by the National Science Foundation. *The Large-Scale Structure of the Universe*. Princeton Univ. Press, Princeton, NJ
- Peebles P. J. E., Groth E. J., 1976, *A&A*, 53, 131
- Robaina A. R. et al., 2009, *ApJ*, 704, 324
- Robaina A. R., Bell E. F., van der Wel A., Somerville R. S., Skelton R. E., McIntosh D. H., Meisenheimer K., Wolf C., 2010, *ApJ*, 719, 844
- Skibba R., Sheth R. K., Connolly A. J., Scranton R., 2006, *MNRAS*, 369, 68
- Springel V. et al., 2005, *Nat*, 435, 629
- Yang X., Mo H. J., van den Bosch F. C., Pasquali A., Li C., Barden M., 2007, *ApJ*, 671, 153
- Zehavi I. et al., 2004, *ApJ*, 608, 16

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