Mon. Not. R. Astron. Soc. **000**, 000–000 (0000) Printed 7 March 2016 (MN LATEX style file v2.2)

Galaxy Zoo: improved debiasing for multiple-answer questions and the demographics of spiral arm number

Ross E. Hart[1⋆], Steven P. Bamford[1] and the Galaxy Zoo team
[1]*School of Physics & Astronomy, The University of Nottingham, University Park, Nottingham, NG7 2RD, UK*

7 March 2016

ABSTRACT
Here we study the behaviour... "climbing the question tree"

Key words: galaxies: general – galaxies: structure – galaxies: fundamental parameters – galaxies: formation

1 INTRODUCTION

Spiral galaxies are the most common type of galaxy in the local Universe (Lintott et al. 2011; Willett et al. 2013), yet formulating a single theory to account for all spiral structure still remains elusive. As most local Universe star-formation is associated with spiral galaxies, and spiral arms themselves are sites of enhanced stellar density, understanding spiral structure is key to understanding the star-formation of the local Universe. The main theories for the occurrence of spiral arm features in local galaxies initially focused on the idea of being caused by density waves in their disks (Lindblad 1963; Lin & Shu 1964), but have since been preceded by theories that consider the effects of gravity and disk dynamics (Toomre 1981; Sellwood & Carlberg 1984), with most if the work to advance the field of spiral structure theory being driven by simulations. Using observational studies to test these theories remains a challenge, as visual classifications of both the presence of spiral structure and its detailed features are required, which are difficult to obtain when considering the large samples provided by galaxy survey data.

An approach that has been used to visually classify galaxies in large survey samples has been to utilise citizen science, by allowing volunteers to morphologically classify galaxies rather than relying on classifications from a small number of experts. Galaxy Zoo 1 (GZ1, Lintott et al. (2008, 2011)) was the first project to collect visual morphologies in this way, by classifying galaxies from the Sloan Digital Sky Survey (SDSS) as either 'smooth' or 'spiral'. Using this method, each galaxy is classified by several individuals, and a likelihood or 'vote fraction' of each galaxy having a particular feature is assigned as the fraction of classifiers who claimed they saw that feature. GZ1 classifications collected in this way have been used to compare galaxy morphology with respect to colour (Masters et al. 2010b,a; Bamford et al. 2009), environment (Skibba et al. 2009; Bamford et al. 2009), and star-formation properties (Tojeiro et al. 2013; Schawinski et al. 2014; Smethurst et al. 2015).

Following from the success of GZ1, more detailed visual classifications were sought, looking for a much more comprehensive set of classifications of features that galaxies exhibit in the local Universe, including the presence of bars, and spiral arm winding and multiplicity properties. Thus, Galaxy Zoo 2 was created (GZ2,Willett et al. (2013), hereafter W13), in which volunteers were asked more questions about a subsample of SDSS galaxies than in GZ1. The main difference between GZ2 and GZ1 was that visual classifications were collected using a 'question tree' in GZ2, to gain a more exhaustive set of morphological information for each galaxy. GZ2 has already been used as a measure for detailed galaxy morphology, for example being used to compare the properties of spiral galaxies with or without bars (Masters et al. 2011, 2012; Cheung et al. 2013), look for interacting galaxies (Casteels et al. 2013), as well as looking for relationships between spiral arm structure and star-formation (Willett et al. 2015). This 'question tree' method has since been used in a similar way to measure the presence of detailed morphological features in higher redshift galaxy surveys (see Melvin et al. (2014) and Simmons et al. (2014) for examples), and other Zooniverse citizen science projects.

An issue that arises in both visual and automated methods of morphological classification is that detailed features are more difficult to observe in lower signal-to-noise images observed from a greater distance. In Galaxy Zoo, this has been termed as classification bias. It is imperative that classification bias is removed from morphological data, as it leads to sample contamination. This means that any observational differences between samples can be significantly reduced.

Classification bias manifested itself in GZ1 with galaxies at higher redshift having lower 'spiral' vote fractions, which were corrected using a numerical method (Bamford et al. 2009). The application of a question tree in GZ2 to look for more detailed features means that correcting for biases is more complicated than in GZ1. In particular, there are

⋆ E-mail: ross.hart@nottingham.ac.uk

questions with several possible answers, and debiasing one answer with respect to each of the others is therefore a less trivial process in GZ2.

The paper is organised as follows. In section 2, the sample selection and galaxy data are described. In section 3, we describe a new debiasing method that has been created to account for the classification bias in the GZ2 questions with multiple possible answers. In section 4, samples of GZ2 spiral galaxies are defined and sorted by arm multiplicity, a case where the new debiasing method is required as there are multiple responses to that question. The overall demographics of these galaxies are then compared with respect to arm multiplicity, in order for us to look for clues as to what processes influence their formation and evolution. The results are summarised in section 5.

2 DATA

2.1 Galaxy properties and sample selection

We make use of morphological information from the public data release of Galaxy Zoo 2 (GZ2; W13). The galaxies classified by GZ2 were taken from the Sloan Digital Sky Survey (SDSS) Data Release 7 (DR7; Abazajian et al. (2009)). The GZ2 sample contains essentially all well-resolved galaxies in DR7 down to a limiting absolute magnitude of $m_r \leqslant 17$, supplemented by additional sets of galaxies in Stripe 82 for which deeper, co-added imaging exists (see W13 for details). In this paper we only consider galaxies with $m_r \leqslant 17$ that were classified in normal-depth SDSS imaging and which have DR7 spectroscopic redshifts. We refer to this as our *full sample*, containing 228, 201 galaxies, to which the debiasing procedure described in § 3.3 is applied. We require redshifts in order to correct the sample for a distance-dependent bias, as described in § 3.1.

Petrosian aperture photometry in *ugriz* filters is obtained from the SDSS DR7 catalogue. Rest-frame absolute magnitudes are those computed by Bamford et al. (2009), using KCORRECT (Blanton & Roweis 2007). Galaxy stellar masses are determined from the r-band luminosity and $u-r$ colour using the calibration adopted by Baldry et al. (2006). All relevant quantities assume a flat cosmology with $\Omega_{\rm m} = 0.3$ and $H_0 = 70\,{\rm km\,s^{-1}\,Mpc^{-1}}$.

In order to study galaxy properties in a representative manner in § 4, we define a *luminosity-limited sample* with $0.03 < z < 0.085$ and $M_r \leqslant -21$, containing 62, 220 galaxies. The luminosity versus redshift distribution of our *full sample*, and the limits of our *luminosity-limited sample*, are shown in Fig. 1. These limits approximately maximize the sample size, given the $m_r \leqslant 17$ limit on the *full sample*. The lower redshift limit avoids a small number of galaxies with very large angular sizes, and hence accompanying morphological, photometric and spectroscopic complications. The upper redshift limits also corresponds to that for which we have reliable galaxy environmental density data from Baldry et al. (2006), which we will make use of in this paper.

In terms of stellar mass, the *luminosity-limited sample* is incomplete for the reddest galaxies at $\log(M/M_\odot) < 10.6$. Where necessary we therefore consider a *stellar mass-limited sample* of 41,801 galaxies, created by applying a limit of $\log(M/M_\odot) \geqslant 10.6$ to the *luminosity-limited sample*.

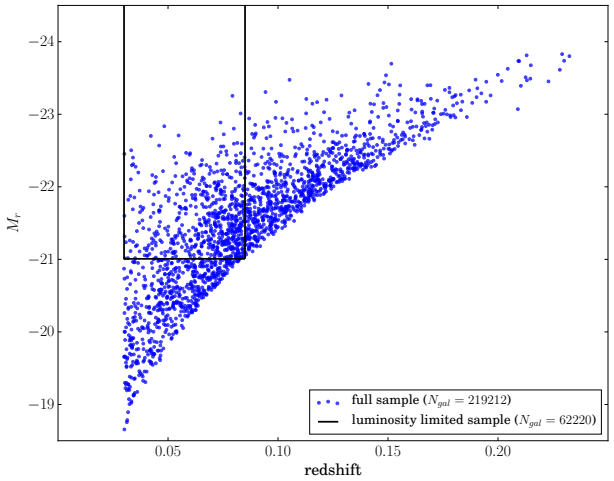

Figure 1. The r-band luminosity versus redshift distribution of our *full sample* (blue points), with the region enclosing our $0.03 < z < 0.085$, $M_r \leqslant -21$ *luminosity-limited sample* indicated by black lines. Only a random selection of 2000 points are plotted for clarity.

2.2 Stellar population models

In § 4.2.4, star-formation histories (SFHs) are compared to stellar population models. Spectral energy distributions (SEDs) are derived from Bruzual & Charlot (2003), for a range of ages and SFHs using the initial mass function from Chabrier (2003). For the models described in this paper, a single metallicity value of $Z = Z_\odot$ is used. Two dust extinction magnitudes of A_v=0 and A_v=0.4 are considered, derived in Calzetti et al. (2000). Equivalent colours for each of the star-formation and dust extinction models are calculated for each of the SDSS *ugriz* filters, with the wavelength dependent filter opacities measured in Doi et al. (2010). Full details of how the models are derived can be found in Duncan et al. (2014).

2.3 Quantifying morphology with Galaxy Zoo

In GZ2, morphological information for each galaxy was obtained by asking participants to answer a series of questions. The structure of this question tree is shown in Fig. 2. Typically, each image was viewed by $\gtrsim 40$ people (W13), but each of the questions is not answered about each galaxy. The questions further down the question tree require that another question has been answered with a particular response. For each question, the responses are each represented by the 'vote fraction', p assigned to each possible answer. For any given question, the sum of the vote fractions for all possible answers adds up to 1. Considering the 'edge-on' question (T01 in Fig. 2), a classifier would only answer that question if they had already said that they had saw features. If a galaxy was classified by 40 people, and 30 of those said they saw features, whilst the other 10 claimed it was smooth, then the corresponding vote fractions are $p_{\rm features} = 0.75$ and $p_{\rm smooth} = 0.25$. Only the 30 classifiers who saw 'features' then answered the 'spiral' question. If 15 of those said the galaxy was edge-on, and 15 said it was not, the corresponding vote fractions would be $p_{edge-on} = 0.5$ and $p_{not\,edge-on} = 0.5$.

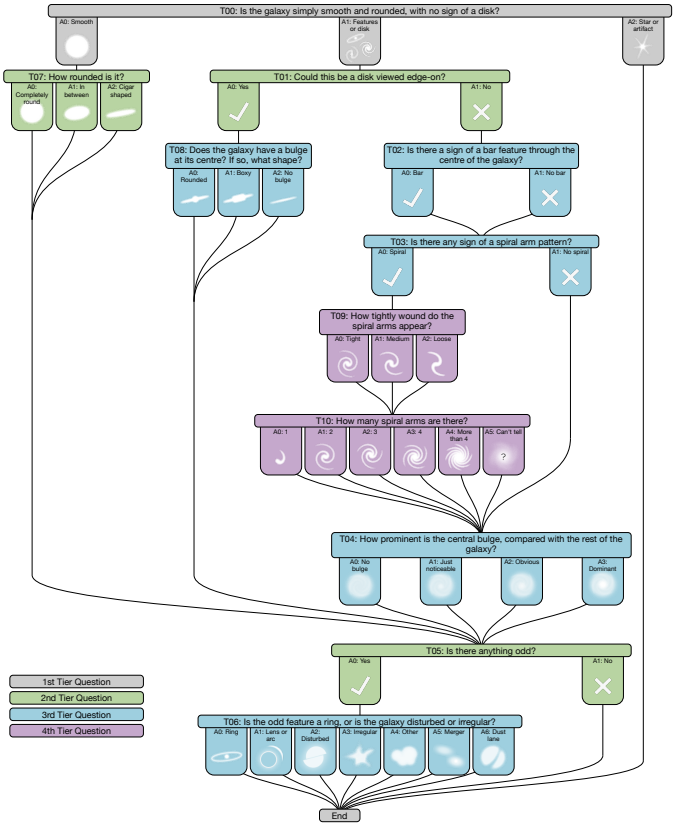

Figure 2. Diagram of the question tree that is used to classify galaxies in GZ2. The tasks are colour-coded by their depth in the question tree. As an example, the arm number question is a fourth tier question- to answer that particular question about a given galaxy, they need to have given a particular response to three previous questions (that the galaxy had features, was not edge-on and had spiral arms).

In order to reduce the influence of unreliable classifiers, W13 down-weighted individual volunteers who had poor agreement. Throughout this paper we refer to these weighted vote fractions as the 'raw' quantities. Before using these GZ2 vote fractions to study the galaxy population, we must first consider the issue of classification bias, as we shall in § 3.1.

Traditional morphologies assign each galaxy to a specific class, usually determined by one, or occasionally a few, experts. In contrast, Galaxy Zoo provides a large number of independent opinions on specific morphological features for each galaxy. This allows us to consider both the inherent 'fuzziness' and observational uncertainties of galaxy morphology, and hence control the compromise between sample contamination and completeness.

There are two principal ways in which galaxy morphologies can be quantified using Galaxy Zoo vote fractions. The first is to consider averages of the vote fractions over specific samples or bins divided by some other property. These average vote fractions can then be used to study variations in the morphological content of the galaxy population. Individual galaxies are not given specific classifications. There is no population of 'unclassified', and hence ignored, galaxies. This approach has been taken by Bamford et al. (2009), Casteels et al. (2013), Willett et al. (2015), and various other studies. With this method, the vote fractions of all galaxies can be considered together; even galaxies with a small (but non-zero) vote fraction for a given property count towards the statistics. Effectively, this approach considers the vote fractions as an estimate of the probability of a galaxy belonging to a particular class.

The second approach is to divide the galaxy sample in to different morphological categories, either by applying a threshold on the vote fractions, or choosing the class with the largest vote fraction. Such methods have been used by Land et al. (2008), Skibba et al. (2009), and many more. One advantage of this approach is that each galaxy is assigned to a definite class, with the threshold tuned to ensure a desired level of classification certainty. However, a set of 'uncertain' or 'unclassified' galaxies may remain. In some analyses these will require special attention.

These different approaches are also relevant for how questions at different levels in the tree are combined. For example, a participant is only asked if they can see spiral arms when they have already answered that they can see features in the galaxy and that the galaxy is not an edge-on disc. The vote fraction for spiral arms therefore represents the conditional probability of spiral arms *given that* features are discernible *and* that the galaxy is not edge-on. When considering whether a galaxy displays spiral arms, one should account for the answers to these previous questions in the tree. One can treat vote fractions as probabilities, multiplying them to obtain a 'probability' that a galaxy displays any features, is not edge-on and possesses spiral arms. Alternatively, one may select a set of galaxies that display features and are not edge-on and possess spiral arms, by applying some thresholds to the vote fractions for each question in turn. (See Casteels et al. (2013) for a more thorough discussion of these issues.)

The primary morphological feature we will focus on in this paper is the apparent number of spiral arms displayed by a galaxy. As we will see, some of the classes for this feature contain a relatively low fraction of the total spiral population. In addition, the vote fractions for the preferred answer are often fairly low, with votes distributed over several answers. In such cases, averaging the vote fractions over the full sample does not work particularly well, as noise from more common galaxy classes overwhelms the subtle signal from rarer classes. In this paper we therefore prefer to assign galaxies to morphological samples by applying a threshold or taking the answer with the largest vote fraction.

3 CORRECTING FOR REDSHIFT-DEPENDENT CLASSIFICATION BIAS

3.1 Biases in the Galaxy Zoo sample

Galaxies at higher redshifts appear fainter and smaller in the SDSS images, and therefore have lower signal-to-noise and resolution. Detailed features are therefore more difficult to distinguish in galaxies at higher redshift. As a result, visual galaxy classifications are biased, as fewer galaxies are classified as having the more detailed features at higher redshift, making a sample of galaxies with the these features incomplete.

It should be noted that such biases are not exclusive to Galaxy Zoo. Difficulty in detecting faint features in lower signal-to-noise galaxies is an inherent property of any visual or automated method of galaxy classification. The advantage of using Galaxy Zoo classifications is that they give a statistical method of measuring galaxy morphology. As each of the galaxies in the *full sample* has been visually classified by a number of independent observers, the apparent evolution in the presence of features can be modelled, and biases corrected accordingly.

Incompleteness and contamination are defects that arise in a sample where an inherent redshift bias affects the classifications. Incompleteness affects the 'harder to see' features: the fraction of galaxies classified as having a particular feature decreases with redshift, leaving us with poor number statistics for a sample we wish to define as having that feature. Contamination is the converse effect that appears in the 'easier to see' categories. In this case, the samples defined using the Galaxy Zoo classifications also include misclassified galaxies that should have actually been included in one of the 'harder to see' categories. Any intrinsic differences between samples that one wishes to compare may be therefore be negated.

The effect of redshift bias is shown in Fig. 3a, where the answer to the 'smooth or features' question is compared for high and low-redshift samples. The redshift range of the SDSS sample is shallow enough to argue that there should be minimal change in the overall population of galaxies (Bamford et al. 2009; Willett et al. 2013). In a *luminosity-limited sample*, the level of completeness should also be the same at all redshifts, meaning that the overall populations of the high and low redshift samples should be equivalent. However, Fig. 3a shows that the higher redshift vote fractions are dramatically skewed to lower values- generally, people are having greater difficulty in detecting the presence of features in the higher redshift images. Thus, there are fewer votes for galaxies showing 'features' and consequently more votes for galaxies being 'smooth'. If one wished to compare a sample of galaxies with 'features' against one that is 'smooth' using the raw vote fractions, the number of galaxies with 'features' would be incomplete and the 'smooth' sample would be contaminated.

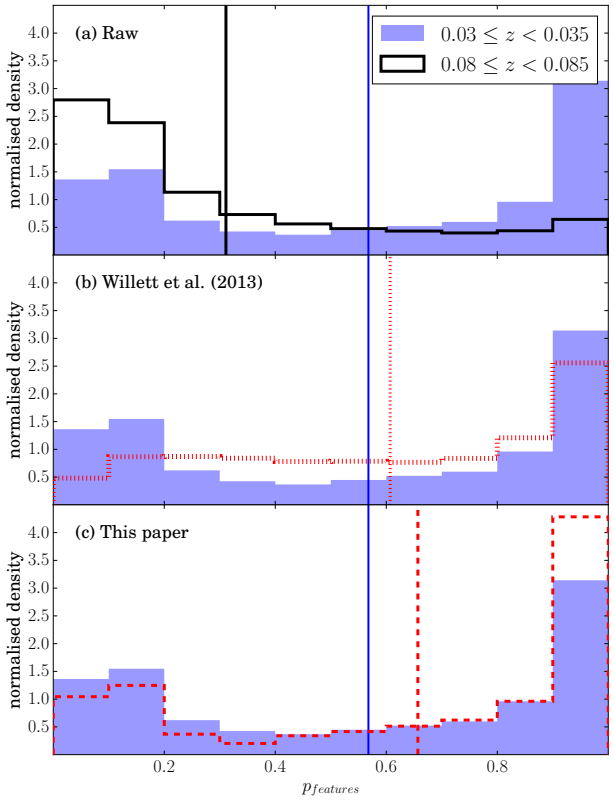

Figure 3. Histograms of vote fractions for the 'features' response to the 'smooth or features' question in GZ2. In each of the panels, the blue filled histogram shows the raw vote distribution for a low-redshift $0.03 \leqslant z < 0.035$ slice of the *luminosity limited sample*. The line histograms show the equivalent distribution for a higher-redshift $0.08 < z \leqslant 0.085$ sample. The vertical lines show the mean vote fractions.

3.2 Previous attempt to correct for redshift bias

The previous debiasing procedure applied to both GZ1 and GZ2 has focused on correcting the vote fractions of the galaxy samples by adjusting the mean vote fractions as a function of redshift. The method was first proposed in Bamford et al. (2009), and updated for GZ2 in W13. The method successfully adjusts the mean vote fractions for questions with two dominant answers, as can be seen from the vertical lines in Fig. 3b: the mean of the debiased high-redshift sample is much closer to the mean of the low-redshift sample than for raw vote distributions (Fig. 3a).

However, this technique has two limitations that make it unsuitable if we want to divide a galaxy sample in to different morphology subsets. The first issue is that adjustment of the mean vote fraction does not necessarily lead to correct adjustment of individual vote fractions. This can be seen in Fig. 3b. Although the mean vote fraction for the high-redshift sample has been correctly adjusted to approximately match the low-redshift sample, the overall distribution does not. There is an excess of debiased votes in the middle of the distribution, and fewer votes for the tails of the distribution at $p \approx 0$ and $p \approx 1$. This effect is important if we wish to divide our sample into different subsets by morphological type. As the shape of the histograms is not consistent with redshift, the fraction of galaxies with $p_{features}$ greater than a given threshold can also vary with redshift.

As described in section 2.3, GZ2 utilises multiple answered questions to obtain more detailed classifications than GZ1. In cases where the votes are split between multiple categories, the debiasing method from W13 does not always adjust the vote fractions correctly. We show this effect for the 'spiral arm number' question (T10 of Fig. 2), in Fig. 4. A sample of 'secure' spiral galaxies with $p_{\rm features} \times p_{\rm not\ edge\text{-}on} \times p_{\rm spiral} > 0.5$ is selected, (with the vote fractions corresponding to the debiased values from W13), and plot the mean vote fractions with respect to redshift for each of the arm number responses. A clear trend in $p_{\rm armnumber}$ is observed: the mean vote fractions vary systematically with redshift, even after the W13 correction has been applied. For this question, the answers with more spiral arms (3, 4, or 5+ spiral arms) are the 'harder to see' features meaning that there are fewer votes for these categories at higher redshift, which instead increase the 1 and 2 arm vote fractions. The

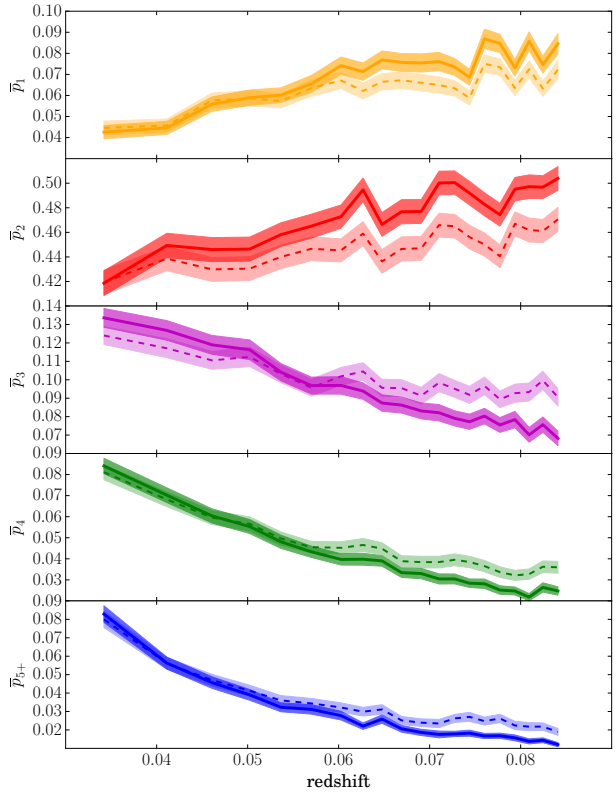

Figure 4. Mean vote fractions for each of the arm number responses to the 'arm number' question (T10 in Fig. 2. The sample consists of galaxies from the *luminosity-limited sample*, with $p_{features} \times p_{not\,edge-on} \times p_{spiral} > 0.5$ (with vote fractions taken from the W13 debiased catalogue). The solid lines show the mean arm number vote fractions obtained using the raw vote classifications, and the dashed lines indicate the same quantity obtained using the W13 debiased values. The shaded regions indicate the 1σ error on the mean.

3,4 and 5+ spiral arm samples of spiral galaxies therefore suffer from incompleteness. This is of particular importance in this case for two reasons. Firstly, as this is a 'fourth order' question, as can be seen in Fig. 2, then the sample size is limited, as three questions must have been answered 'correctly' previously for a galaxy to be classified as spiral. Secondly, the 3, 4 and 5+ arm responses have low mean vote fractions overall, of $\lesssim 0.1$. Thus, the number statistics for these categories are very low, meaning they will suffer from high levels of noise. The 1 and 2 armed spiral samples would also suffer from contamination from galaxies that should have been classified as 3, 4 or 5+ armed. Therefore, any differences between the samples may be significantly diluted, making it difficult to discern any differences in samples divided by spiral arm number.

3.3 A new method for removing redshift bias

Given the limitations described in § 3.2, we attempt to construct a new method of debiasing the GZ2 data more effectively. This is of particular importance for the question of spiral arm multiplicity (T10 of Fig. 2). For a volunteer to have answered this question, they must have given specific answers to three previous questions (that the galaxy has features, is not edge-on, and does have spiral arms). Therefore, the number of votes for such questions can be very low.

When considering a question with low number statistics, such as the spiral arm question, we prefer to use a thresholding technique, rather than using the weighted vote fractions (see § 2.3 for a descriptions of both methods). Using the arm number question as an example, the '2 spiral arms' response dominates the overall vote fractions, making up $\sim 60\%$ of the votes, as can be seen in Fig. 4. The rarer responses of 3, 4 or 5+ arms have much lower number statistics overall, with only $\sim 10\%$ of the votes. The mean values can therefore be affected by the noise in the dominant category, which will be much larger than the noise for the rarer category. We therefore prefer to divide our galaxy sample in to different sub-samples when comparing galaxies by spiral arm number.

Unlike the debiasing method in W13, our new method aims to make the vote distributions themselves as consistent with redshift rather than purely aiming for consistency in the mean vote fraction values. As each galaxy is classified by 40 or more volunteers (W13), we have enough data to model the evolution of the vote distributions as a function of redshift. Different classifiers will have different sensitivity to picking out the most detailed features. Thus, as samples at higher redshift are considered, and hence with poorer image quality, we expect the vote fraction distributions to also evolve as some classifiers become less able to see the most detailed features. We aim to account for this bias by modelling the vote fraction distributions as a function of redshift, and correcting the higher redshift vote distributions to be as similar as possible to equivalent vote distributions at low redshift.

We first define samples of galaxies for each of the questions in turn. The sample is then binned in terms of the intrinsic galaxy properties of size and luminosity, and each of these bins is divided in to redshift slices. We then attempt to model the vote distributions for each of the bins with respect to redshift, and thus match their distributions to those at low redshift. This means that if a vote fraction threshold is applied, the fraction of galaxies with a given feature remains constant: at each redshift, the sample is composed of the galaxies that are most likely to have that particular feature.

It must be noted that such a method could still be limited by number statistics at higher redshift. In the case that a feature's vote fraction drops to 0 at higher redshift, we can not 'add-in' votes- it is only possible to debias the galaxies with $p > 0$, where there is evidence for a feature being present. This remains a problem for the categories where the vote fractions are lowest, such as in the responses to the odd feature question (T06 in Fig. 2).

3.3.1 Sample selection for each question

As GZ2 morphologies are classified with a decision tree (see section 2.3), not all of the questions were answered by each of the volunteers for a given galaxy. For an individual classifier to have answered the question regarding arm number, they would also have needed to answer that the galaxy had features, was not edge-on and had spiral arms. Answering the spiral arm number question is not appropriate for all of the galaxies in the sample. If a galaxy has no spiral fea-

tures, yet a volunteer answered the spiral arm question, then such a galaxy would contribute 'noise' to the answers to that question. To avoid 'noise' introduced by incorrectly classified galaxies, clean galaxy samples are defined with $p > 0.5$. For the first question, this corresponds to all of the galaxies, as each classifier answered that particular question for each galaxy. However, when questions further down the tree are considered, this is not the case. The equiavalent $p > 0.5$ for the spiral arm question would only include the galaxies with $p_{\rm features} \times p_{\rm not\ edge\text{-}on} \times p_{\rm spiral} > 0.5$.

For each of the questions in turn, we define a sample of galaxies with which we will apply the new debiasing procedure. These samples are defined using a cut of $p > 0.5$ (corresponding to $p_{\rm features} \times p_{\rm not\ edge\text{-}on} \times p_{\rm spiral} > 0.5$ for the spiral arm question for example). A further cut of $N \geqslant 5$ (where N is the number of classifications) is also imposed to ensure that each galaxy has been classified by a significant number of people to reduce the effects of Poisson noise. In this case, the vote fractions must be the debiased vote values, to ensure each sample is as complete as possible (see § 3.1) as we look at each question. The order in which the questions are debiased is important: to define a sample for the debiasing of a particular question, all questions further up the question tree must have been debiased beforehand.

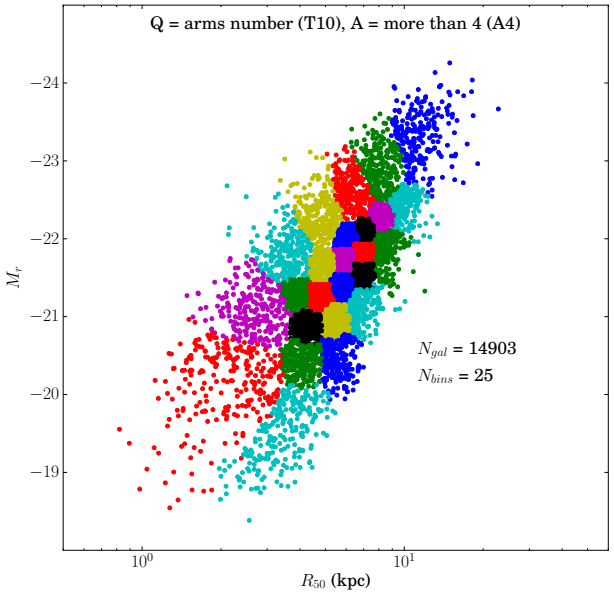

Figure 5. Distribution of the voronoi bins in terms of R_{50} and M_r for the spiral arm number question and the more than 4 spiral arms answer. Each of the voronoi bins is further divided in to redshift bins, each with $\geqslant 50$ galaxies.

3.3.2 Binning the data

It is expected that the ability to discern the presence of a particular feature will depend on intrinsic galaxy properties. For example, larger, brighter galaxies may be easier to classify over a wider redshift range. Conversely, fainter galaxies may show stronger features, as the fraction of spiral galaxies is higher **RH: cite???**. To account for these possible variations, we bin the data in terms of M_r and $\log(R_{50})$ for each answer in turn. We use the `voronoi_2d_binning` package from Cappellari & Copin (2003),to ensure the bins should an approximately equal number of galaxies. Fig. 5 shows an example of the voronoi binning for the 5+ arms response to the arm number question. When voronoi binning the data for each of the answers, only the $N_{\rm gal}$ galaxies with $p > 0$ are included, meaning that the 'signal' of galaxies is evened out over all of the voronoi bins. We aim to have ~ 30 voronoi bins for each of the questions, so the desired number of galaxies in each bin is given by $N_{\rm gal}/30$.

After voronoi binning the data in terms of their intrinsic properties of size and brightness, we further divide each bin in to redshift bins, to allow us to study how the vote distributions change with redshift. To ensure that there is a good signal, each redshift bin contains $\geqslant 50$ galaxies. This binned data is used for the debiasing methods described in the next section.

3.3.3 Modelling redshift bias

For each of the possible responses to each question, a method is applied to correct for the redshift bias in the sample, aiming to make the vote distributions for each answer consistent with redshift. The two methods that we employ to achieve this are described below.

The first method we utilise to remove redshift bias simply matches the shapes of the histograms on a 'bin-by-bin' basis.The cumulative distribution for the lowest redshift sample in a given voronoi bin is used as a reference for how the shape of the histogram would look if it were viewed at low redshift. An example of this method is shown in Fig. 6, in which the 'features or disk' answer to the 'smooth or features' question is considered. For both the low redshift bin and the high redshift bin, the vote fractions are ranked in order of low to high. Each of the galaxies in the high redshift bin is then matched to its low redshift equivalent by finding the galaxy with the closest cumulative fraction in the low redshift bin. This 'matching' technique is shown by the vertical lines of Fig. 6. In this case, a galaxy with cumulative fraction of ≈ 0.8 in the high redshift bin has $p_{\rm features} \approx 0.18$. A galaxy at the same cumulative fraction in the low-redshift bin has $p_{\rm features} \approx 0.65$, so this is the debiased value assigned to that galaxy. This is repeated for each galaxy in turn, and for each of the high redshift bins in turn. Applying a vote fraction threshold for a given response gives the same fraction of the population above that threshold in all of the redshift bins, with the galaxies most likely to have a feature making up the population of galaxies above that threshold.

The main strength of this method is that any vote distribution can be modelled in this way, irrespective of the overall shape. However, a potential weakness is that noise can be introduced due to the discretisation of the data. To limit this issue, each redshift bin has a 'good' signal of $\geqslant 50$ galaxies. this effectively 'blurs' any trends with redshift, and can actually lead to an overcorrection of vote fractions, which can be seen in Fig. 3c. Although the overall histogram shape is correctly matched, when a slice at $0.08 \leqslant z < 0.085$ is considered, we see too many galaxies with $p \approx 1$ compared to the low-redshift data. This issue is purely caused by the discretisation of the individual bins- although the trends can be modelled overall, any trends within individual bins cannot. If there is a redshift trend within a bin, then the fraction of galaxies with the more difficult to see features will preferentially reside in the lower redshift ends of the bins. This

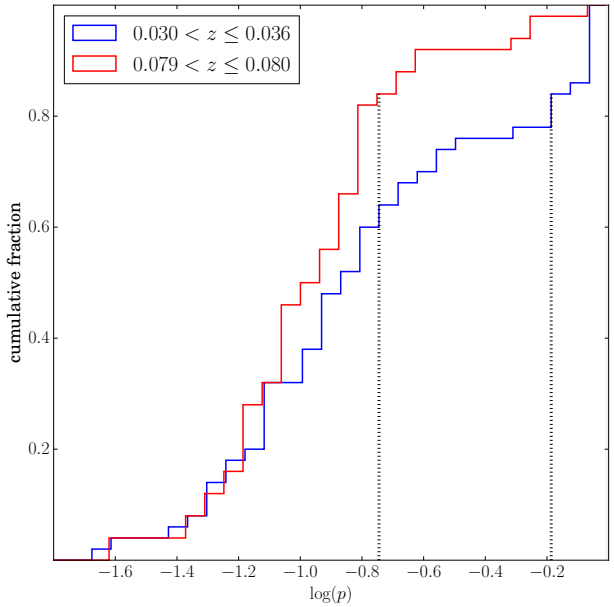

Figure 6. An example of vote distributions for an example voronoi bin for the 'features or disk' answer to the 'smooth or features' question. Each of the galaxies in the high-redshift bin (red line) is matched to its closest equivalent low-redshift galaxy (blue line) in terms of cumulative fraction. The dashed lines indicate the 'matched' values for an example galaxy with $\log(p) \approx -0.8$, and an equivalent low-redshift value of $\log(p) \approx -0.2$(corresponding to $p_{\rm raw} = 0.18$ and $p_{\rm debiased} = 0.65$).

effect leads to an overestimate of the number of galaxies with the more difficult to see features. Fig. 8a shows the debiased trends of the 'features or disk' question, which was debiased using the 'bin-by-bin' method, which shows that the method slightly overcorrects the redshift trend in the number of galaxies classified with $p_{\rm features} > 0.5$.

One potential solution would be to bin the data more finely. However, there is no 'ideal' solution to this problem, as fewer galaxies in each bin would mean that the redshift range that each bin occupies is smaller, but the noise in each of the bins is larger.

To attempt to remove the discrete nature of the correction in the 'bin-by-bin' method, an alternative approach is proposed that attempts to model vote distributions with functions. For each of the redshift bins, we plot a cumulative histogram of $\log(p)$ against cumulative fraction. An example of some of these cumulative histograms are plotted as the solid lines in Fig. 7. It can be seen that there is a clear evolution in the distributions with redshift. This effect is most prominent in the 4 and 5+ arms responses, where the distributions shift so that there are fewer galaxies with higher vote fractions. To attempt to correct for this bias, each of the cumulative histograms can be modelled with a function, and the parameters of the function can be modelled in terms of redshift(z), galaxy size (R_{50}) and intrinsic brightness (M_r). After much experimentation, a function of the following form is used to model the cumulative distributions:

$$f(p) = e^{kp^c}, \tag{1}$$

where k and c parameterise the shape of each of the curves. Best-fit k and c values are found for each of the bins, indicated by the dashed lines in Fig. 7. When fitting, the cumulative histogram is sampled evenly in $\log(p)$ to avoid the fit being weighted to the most steep parts of the curves.

After finding k and c for each of the bins, we attempt to quantify how these parameters change with respect to M_r, $\log(R_{50})$ and z. A 2σ clipping is applied to all of the k and c values to remove any fits where discrepant k or c values have been found. The data is then fitted using a continuous function of the following form:

$$\begin{aligned} A_{fit}(M_r, R_{50}, z) =& A_0 + A_M(f_M(-M_r)) + \\ & A_R(f_R(\log(R_{50}))) + A_z(f_z(z)), \end{aligned} \tag{2}$$

where A corresponds to either k or c and f_M, f_R and f_z are functions that can be either logarithmic ($\log(x)$), linear (x) or exponential (e^x). The values A_0, A_M, A_R and A_z are constants that parameterise the shape of the fit with respect to each of the terms. When fitting the data, M_r, $\log(R_{50})$ and z correspond to their respective mean values calculated using all of the galaxies in that bin. The best combination of functions is chosen by calculating A_0, A_M, A_R and A_z for each combination of f_M, f_R and f_z, and selecting the function that has the lowest square residual. We then clip any values with a $> 2\sigma$ residual to this fit and re-fit the data to find a final functional form for k and c with respect to M_r, R_{50} and z. The resulting modelled cumulative histograms for the spiral arm number question are shown by the dotted lines of Fig. 7. Limits are also applied to k and c to avoid unphysical fits at extreme values of M_R, R_{50} and z. The range of k and c is therefore set by the upper and lower limits of all of the fit k and c values within the 2σ clipping.

RH: *Do we need some kind of plot to show how k and c have been fitted here?

After finding a functional form for k and c with respect to M_r, $\log(R_{50})$ and z, each of the galaxies in the sample is debiased to find its equivalent value at low redshift. To do this for an individual galaxy, a cumulative histogram is estimated using $k_{\rm fit}(M_r, R_{50}, z)$ and $c_{\rm fit}(M_r, R_{50}, z)$, where M_r, R_{50} and z are the properties for that particular galaxy, giving the cumulative fraction for a galaxy's raw vote fraction. The equivalent cumulative histogram at $z = 0.03$ (the low redshift limit of our *luminosity-limited sample*) is also found, using $k_{\rm fit}(M_r, R_{50}, 0.03)$ and $c_{\rm fit}(M_r, R_{50}, 0.03)$. The vote fraction for the corresponding cumulative fraction is read off from the low redshift cumulative histogram in a similar way as in the 'bin-by-bin' method, this time using the fitted curves rather than the raw histograms. This is repeated for each of the galaxies in the sample to generate a set of debiased values for the *full sample* of galaxies.

As mentioned previously, function fitting avoids issues related to the discretisation of the data. However, it does introduce its own biases, as an assumption is made that the cumulative histograms can all be well-fit by a particular set of continuous functions. This may not always be the case, so we must consider which of the above methods does the best overall job of removing redshift bias. To do this, the distributions of votes for a low-redshift reference sample are compared to the distributions of higher redshift bins. Using the *luminosity-limited sample*, which is free from redshift bias across all $M_r - R_{50}$ bins, a reference sample with $0.03 \leqslant z < 0.035$ is defined. The rest of the *luminosity-limited sample* is then split in to 10 redshift slices, and the

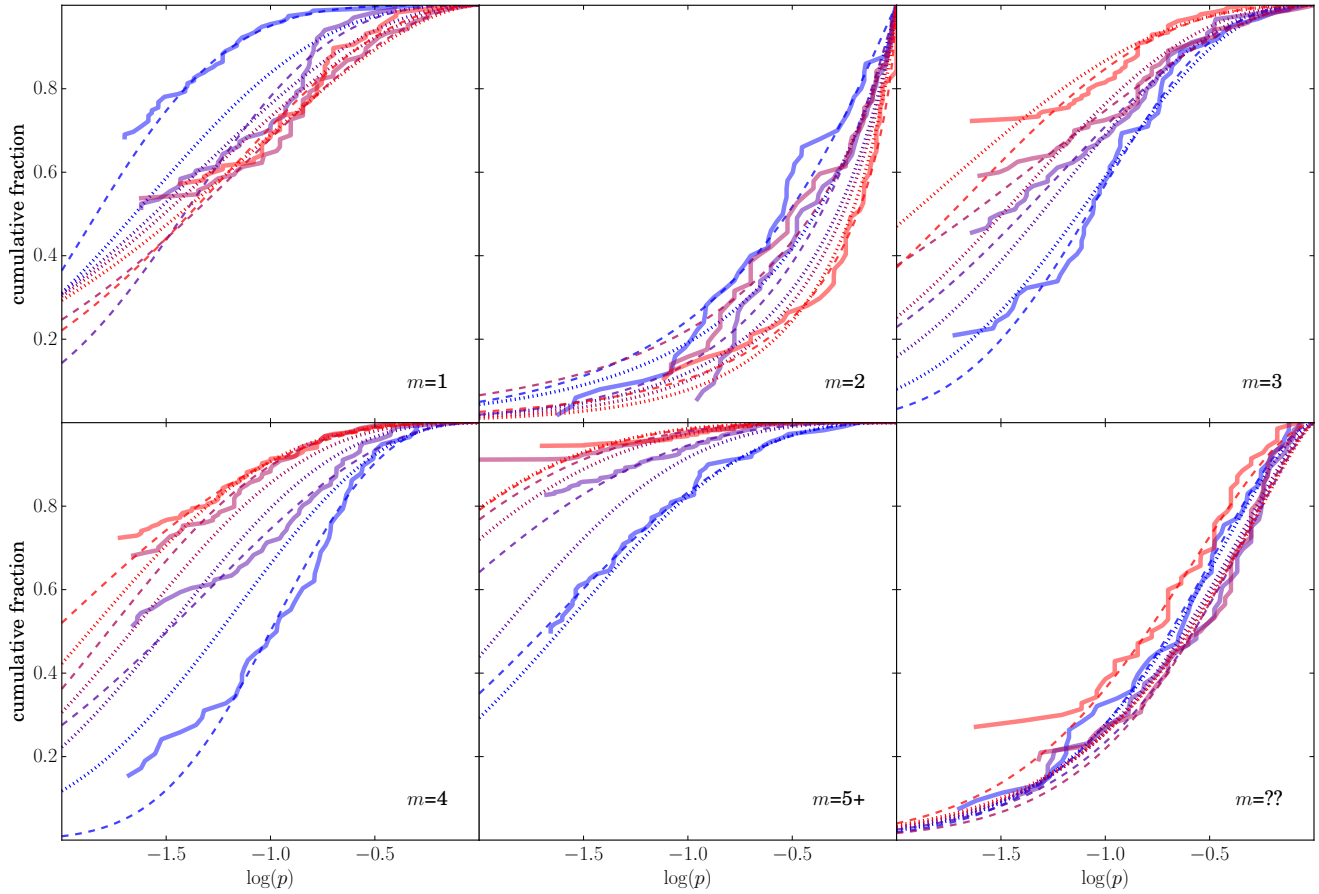

Figure 7. An example of a single voronoi bin fit for the *arm number* question. The red line indicates the highest redshift bin, and the blue line indicates the lowest redshift bin. The solid lines indicate the raw p histograms, and the dashed lines show the best fit function to each of them. The dotted lines show the corresponding approximation from the continous fit to the k and c values.

total square residual of the vote fractions from both of the debiased methods are calculated with respect to the raw vote distributions of the reference sample. The method with the lowest total square residual is taken as the preferred method, and theirs are the debiased values used as the vote fractions for that response.

3.3.4 Results from the new debiasing method

As described in § 3.3, the new method aims to keep the fraction of galaxies above a given threshold constant with redshift, rather than simply correcting the mean vote fractions with redshift, as shown in Fig. 3c. To test how successful the new debiasing method is at defining populations of galaxies above a given threshold with redshift, the fraction of galaxies with $p > 0.5$ for each of the questions is plotted in Fig. 8. It can be seen that in most cases, the new debiasing method does keep the fraction of the population with $p > 0.5$ constant with redshift, as expected. This effect is most evident when looking at the *spiral* question, in Fig. 8d. It can be seen that the original debiasing method does not adequately remove redshift bias, with fewer galaxies exhibiting spiral structure at higher redshift. However, our new method does keep this fraction approximately constant with redshift, which means the spiral sample will be more complete if we wish to use a thresholding technique to define a sample of galaxies with spiral structure. Similarly, the sample without spiral features will suffer from less contamination from incorrectly classified spiral galaxies.

Fig. 8 only shows the specific example of the threshold of $p > 0.5$. This does not give any insight into the overall vote fraction distribution, which can vary with redshift as shown in Fig. 3. Therefore, overall distributions are compared for two redshift slices in Fig. 9. It can be seen that this new method does not always 'match' the low and high redshift samples exactly, an effect that is most obvious in the 'spiral' question. Rather than getting an excess of votes towards the middle of the distribution, an excesses are more generally seen at the tails of the distributions at $p \approx 0$ and $p \approx 1$. This is because our method preferentially matches the $p \approx 1$ end of the distribution. As can be seen by the 'spiral = yes' response in Fig. 9, the top ends of the distributions are usually correctly matched; the scarcity of votes for the intermediate values of p are caused by the excess of galaxies with $p = 0$ that cannot be corrected.

4 PROPERTIES OF SPIRAL GALAXIES WITH RESPECT TO ARM NUMBER

Spiral galaxies make up as many as two-thirds of the galaxies in the local Universe (Lintott et al. 2011; Willett et al. 2013).

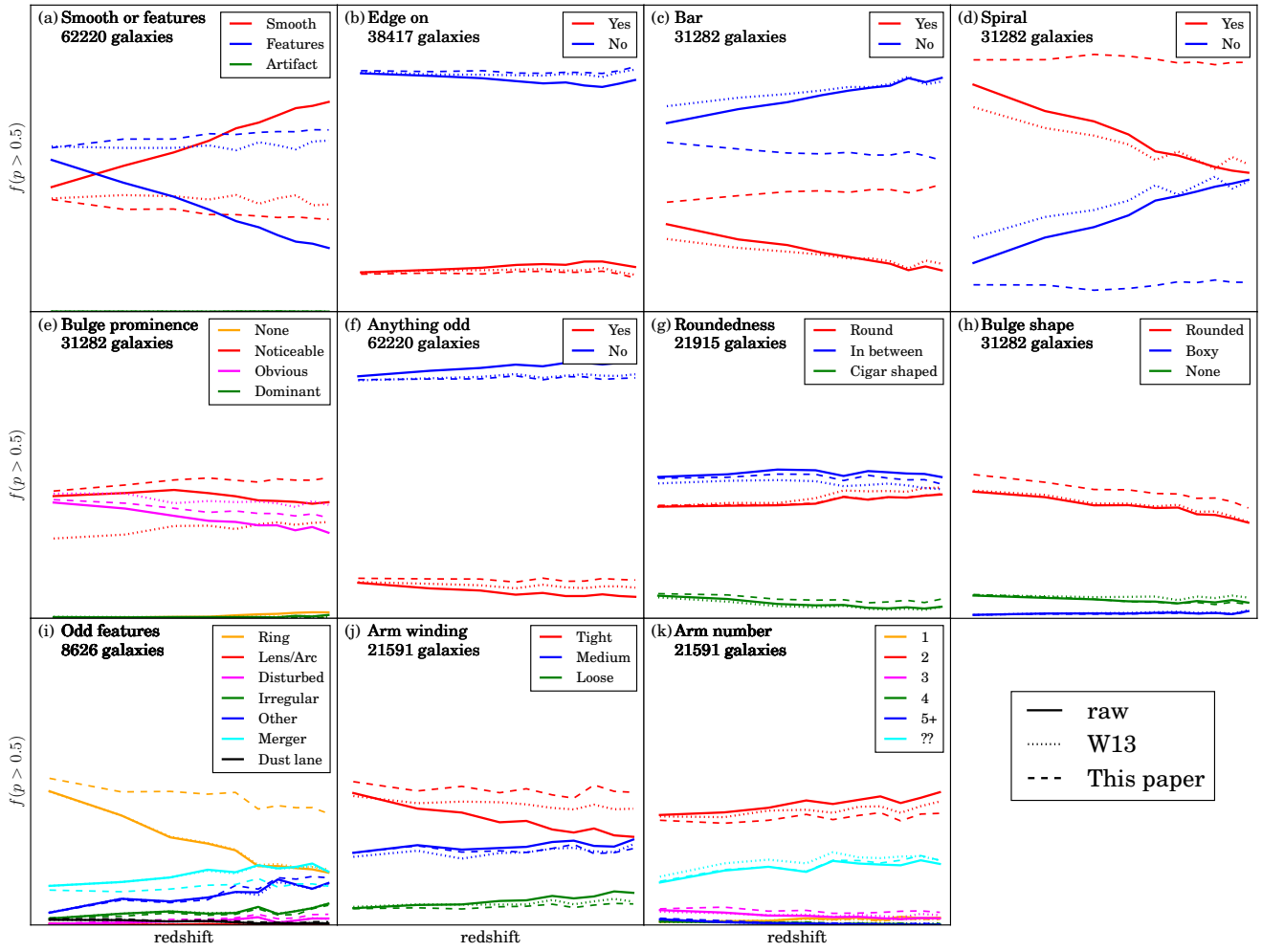

Figure 8. Number of galaxies with $p > 0.5$ for each of the questions debiased using the method described in section 3.3. The solid lines indicate the raw vote fractions and the dashed lines indicate the debiased vote fractions. The dotted lines indicate the same fractions using the W13 debiasing method. The total sample here is composed of galaxies in the *luminosity-limited sample* with $p > 0.5$ (as described in § 3.3.1).

Most of the star formation in the local Universe occurs in spirals, and in particular is concentrated in the regions of spiral arms (Grosbøl & Dottori 2012; Dobbs & Baba 2014). Understanding the physical processes responsible for spiral structure is vital in understanding how galaxies evolve, and how star-formation itself occurs and evolves.

Despite how prevalent spiral galaxies are in the local Universe, formulating a single, complete picture as to how they form and evolve is still elusive. One of the key reasons why this is the case is because spiral structure can take many varied appearances. Spiral galaxies are often classified using either a Hubble-type (Hubble 1926) or an Elmegreen-type classification scheme (Elmegreen & Elmegreen 1982, 1987). The Hubble method is based on bulge size and spiral arm pitch angle. However, as those properties are weakly correlated (Kennicutt 1981; Seigar & James 1998), the physical processes responsible for bulge growth and spiral arm pitch angle may actually be unrelated. The Elmegreen-type classifications scheme instead divides galaxies in to two types depending on the spiral arm structure itself, rather than any properties related to the galactic bulge. This scheme generally classifies galaxies as one of three types: grand design, multiple-armed or flocculent. Grand design spiral structure is associated with two symmetric spiral arms, whereas multiple-armed structure is associated with more than two spiral arms and flocculent galaxies have many, shorter, less well-defined arms. The distinct advantage to classifying spiral galaxies in this way is that contrasting physical mechanisms tare thought to play a role in the formation of these two different types of spiral structure.

Grand design spiral structure was initially thought to be due to the presence of a density wave in a galaxy's disk (Lindblad 1963; Lin & Shu 1964). In the mechanism proposed by Lin & Shu (1964), gas is 'shocked' in to star-formation in regions of high density in the disk. However, this mechanism is no longer favoured, as there is no evidence for the enhancement of star-formation in grand design spiral galaxies compared to many-armed spiral galaxies of the same stellar mass(Romanishin 1985; Elmegreen & Elmegreen 1986; Kendall, Clarke & Kennicutt 2015), or any evidence for enhancement in star-formation in the individual arms of such galaxies (Foyle et al. 2011; Choi et al. 2015). Instead, it is thought that grand design spiral structure may actually occur as a result of strong bars in galaxy disks

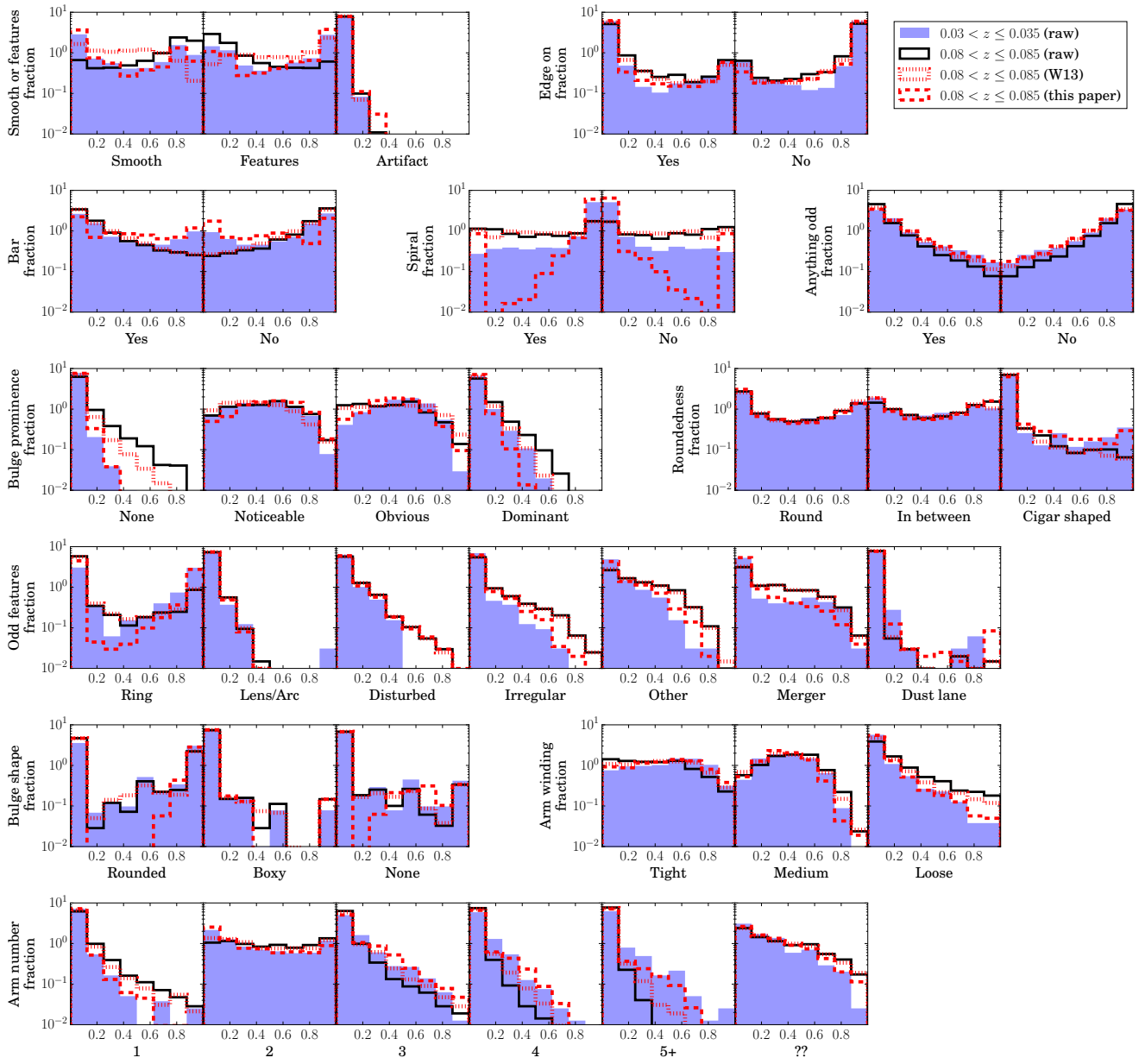

Figure 9. Vote distribution histograms for each of the answers in the GZ2 question tree. The blue filled histogram shows the distribution for a low redshift $0.03 < z \leqslant 0.035$ sample, which should have minimal redshift-dependent bias. The black solid, red dotted and red dashed histograms show the higher redshift $0.08 < z \leqslant 0.085$ distribution of the raw, W13 debiased and debiased data from this paper respectively. Both the low and higher redshift distributions are drawn from galaxies with $p > 0.5$ (as described in § 3.3.1) from the *luminosity-limited sample*.

or tidal interactions (Kormendy & Norman 1979). Early observational evidence supports the theory that hints of grand design structure can be induced via interactions, with two-armed structure being favoured over many-armed structure in high density environments (Elmegreen & Elmegreen 1982, 1987; Ann 2014), and simulations showing that galaxy-galaxy interactions can lead to the types of grand design spirals seen in the local Universe (Dobbs et al. 2010; Semczuk & Lokas 2015).

Unlike two-armed spiral structure, many-armed spiral structure arises readily in simulations without the requirement for a trigger from either a bar instability or a tidal interaction (James & Sellwood 1978; Sellwood & Carlberg 1984). Instead such structures arise readily in simulations, but require a cooling of the gas in the disk to be sustained for long periods of time (Carlberg & Freedman 1985). More recent simulations, taking the disk gravity into account, have shown that 'flocculent' structure may actually be a transient feature of spiral galaxies, with spiral arms continually being made and destroyed (Bottema 2003; Grand, Kawata & Cropper 2012; Baba et al. 2009; Baba, Saitoh & Wada 2013; D'Onghia, Vogelsberger & Hernquist 2013), rather than a long-lasting persistent structure.

Despite the recent advances in the simulations of these disk galaxies, the picture as to how all of the processes shape spiral galaxies still remains unclear. Grand design spiral

galaxies can still reside in low density environments without the presence of a bars (Elmegreen & Elmegreen 1982), meaning that they are not purely driven by these processes as described in Kormendy & Norman (1979). Additionally, the timescales of the persistence of spiral structure is still unclear, particularly as older stellar populations viewed in the infra-red show very different structure to the young stellar populations viewed at optical wavelengths (Block & Wainscoat 1991; Block et al. 1994; Thornley 1996). Most recent work on spiral structure have also mainly been focused on simulations of spiral structure. Putting observational constraints requires the visual inspection of the spiral arm structure in galaxy disks, so have been restricted to relatively small samples of order $\lesssim$ 2000 galaxies (Elmegreen & Elmegreen 1982, 1989; Ann & Lee 2013). We aim to use the GZ2 vote classifications to compare the overall demographics of spiral structure in a much larger sample of SDSS galaxies, which is complete in luminosity and stellar mass.

4.1 Spiral arms in Galaxy Zoo

In order to study how spiral properties vary, visual inspection of the number of arms in a spiral galaxy disk is required. Such classifications are provided by question T10 of the GZ2 question tree (see Fig. 2). This question has six possible responses. In this case, the responses will be referred to as m-values, and can take the value of either 1, 2, 3, 4, 5+ or can't tell.

In order to compare different spiral galaxies, a secure sample of spirals must first be defined. Galaxies with $p_{\rm features} \times p_{\rm not\ edge\text{-}on} \times p_{\rm spiral} > 0.5$ are selected. A further cut is also imposed where only galaxies with $N_{spiral} - N_{can't\,tell} \geqslant 5$ are selected. This means that the *spiral sample* is only composed of galaxies where more than 5 people classified the spiral arm number, to reduce the effects of noise caused by low numbers of classifications. The population of galaxies selected in this way from the *full sample* will hereafter be referred to as the *spiral sample*. The samples defined using these same cuts from the *luminosity-limited sample* and *stellar mass-limited sample* are referred to as the *luminosity-limited spiral sample* and *stellar mass-limited spiral sample*.

Each galaxy is then assigned a specific spiral arm number m, of either 1, 2, 3, 4 or 5+ arms, depending on which response has the highest debiased vote fraction (excluding the *can't tell* response). The *debiased* vote fractions for each of the arm number responses are hereafter referred to as p_m, where m is either 1, 2, 3, 4 or 5+. Examples of some securely classified spiral galaxies are shown in Fig. 10, where each galaxy has a dominant vote fraction of $p_m > 0.8$. The samples of galaxies assigned to each of the different m-values are referred to as the *arm number samples*.

The debiasing procedure applied to this question has shifted the vote fractions for the multiple-armed (m=3, 4, 5+) answers upwards overall, as can be seen in Fig. 12. This has the effect of making each of these samples more complete with redshift, and increasing their respective overall vote fractions. However, in the m=5+ arms case, the sample is still somewhat incomplete, as the overall fraction of galaxies that are assigned to this category decreases with redshift. The vote fractions for m=5+ fall to 0 far more quickly with redshift than any of the other categories, as can be seen from the dashed line in the bottom panel of Fig. 12, making the modelling of this redshift bias difficult. Despite this, the fraction of galaxies that make up the m=5+ category are still significantly improved compared to the sample sizes that would be defined using either the raw vote fractions or the W13 debiased vote fractions, as can be seen in from the N and f columns of Table 1.

The main result of this debiasing is that galaxies with low vote fractions for the many-armed answers are included in the many-armed categories when they were not before. As a consequence, the population of m=2 galaxies is less contaminated by galaxies that actually have 3, 4 or 5+ spiral arms. This effect is illustrated in Fig. 11, where a selection of spiral galaxies with $0.5 < p_m \leqslant 0.6$ are shown. It can be seen that the m=4 and m=5+ spiral samples at higher redshift include spiral galaxies that initially had much lower overall vote fractions. As an example, if one were to use the raw vote fractions to select 'secure' galaxy samples with $p_m > 0.5$, then the galaxy in Fig. 11(y) would be unclassified, as its highest value of p_m would only be 0.27 (which is actually for the m=4 response). Using our debiased values, it has a modal value of p_m=0.55 for the m=5+ armed response, so would be in the m=5+ sample. Even in the case of the less secure samples of Fig. 11, the galaxies classified as m=4 or m=5+ clearly have more spiral arms than those in the m=2 category. **RH: Nair classifications?**.

4.2 Comparing galaxy populations

Having defined the samples of spiral galaxies in §4.1, the demographics of the different galaxy populations separated by spiral arm number can be compared. For reference, mean stellar mass (M_*), colour ($g-r$) and local densities (Σ) are tabulated in the final three columns of table 1.

4.2.1 Comparison of sample sizes

Spiral arm multiplicity does not map exactly on to a specific Elmegreen-type for two reasons. Firstly, the arm number itself does not give any indication of the prominence of spiral arms, so cannot be used to distinguish between a galaxy with many well-defined arms and one with more flocculent spiral structure, which are usually defined differently (Elmegreen & Elmegreen 1982, 1987). The second issue is that arm structure may not necessarily be consistent at all radii (Grosbøl, Patsis & Pompei 2004) in a galaxy disk, and at all wavelengths (Block & Wainscoat 1991; Block et al. 1994; Thornley 1996), meaning that assigning a single m-value of arm number may not give a complete picture of the overall spiral arm structure. The most 'easy-to-map' categories may therefore be to compare our m=2 population with the galaxies classified as grand design, as grand design structure is usually associated with two well-defined arms across the entire disk (Elmegreen & Elmegreen 1982). In the *luminosity-limited spiral sample*, 62% of the galaxies show two-armed spiral structure. This result is consistent with optical visual classifications (Elmegreen & Elmegreen 1982) and infra-red classifications (Grosbøl, Patsis & Pompei 2004), which suggest that $\sim$ 60% of local spiral galaxies exhibit grand design spiral structure.

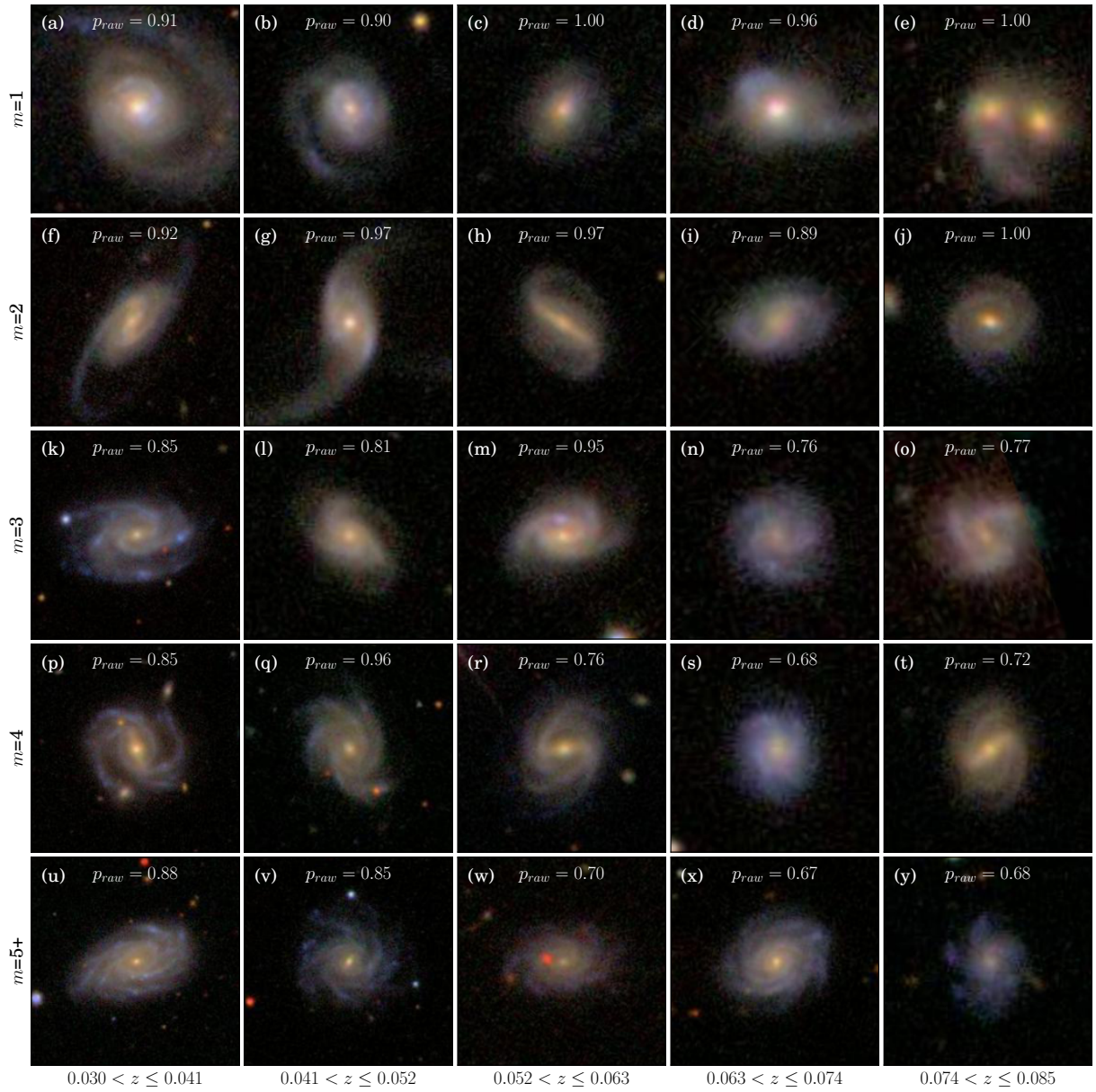

Figure 10. Galaxies classified in to each of the arm number categories (m=1,2,3,4 or 5+) for the stellar mass range $10.0 < \log M_*/M_\odot \leqslant 11.0$. All of the galaxies are taken from the *luminosity-limited spiral sample*. Each galaxy has a debiased modal vote fraction $p_m > 0.8$.

4.2.2 Stellar mass

Galaxy stellar mass is known to correlate with galaxy morphology (Bamford et al. 2009; Kelvin et al. 2014), and spiral galaxy Hubble-type (Muñoz-Mateos et al. 2015). It has also been found that stellar mass correlates with the strength of the m=2 mode in spiral galaxies, with two-armed structure more common in galaxies with greater physical size (Elmegreen & Elmegreen 1987) and stellar mass (Kendall, Clarke & Kennicutt 2015). The distributions of stellar mass for each of the *arm number samples* are shown in Fig. 13a. The overall distributions for each of the galaxy samples show that there is little evidence for a dependence of spiral arm number with respect to host galaxy stellar mass; each of the samples contains galaxies across the entire range of stellar mass from $10.0 \lesssim \log(M_*/M_\odot) \lesssim 11.5$. A slight excess of low stellar mass galaxies is found in the m=3 and m=4 samples, as well as an excess of high stellar mass spiral galaxies for the m=5+ sample. **RH: however, ...?**

The distributions of Fig. 13(a) show the distributions from the *luminosity-limited spiral sample*, so are therefore incomplete for low stellar mass galaxies (see §2.1). As we shall see in §4.2.4, higher mass galaxies are bluer, and hence more luminous for a given stellar mass. They are thus over-represented in a at low masses in a luminosity-limited sample. To look for trends in terms of stellar mass, the overall fraction of the *stellar mass-limited spiral sample* is shown in Fig. 13(b). Now, it can be seen that there do appear to be some trends between spiral arm number and host galaxy stellar mass. A significant increase in the fraction of galaxies with 5+ spiral arms is observed from the overall mean value of $0.068 pm 0.003$ to $0.14 pm 0.02$ for the highest stellar

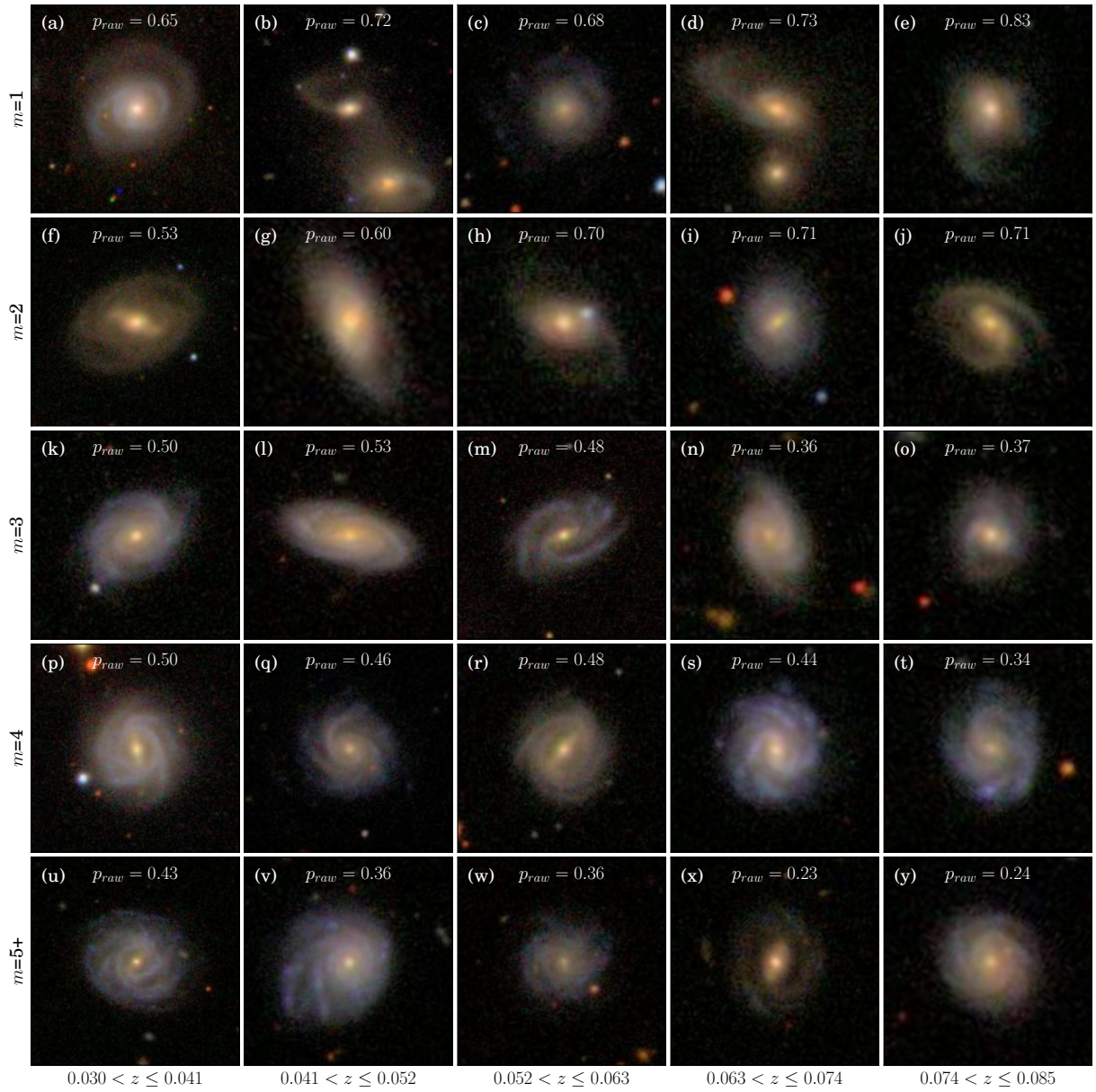

Figure 11. Galaxies classified in to each of the arm number categories (m=1, 2, 3, 4 or 5+) for the stellar mass range $10.6 < \log M_*/M_\odot \leqslant 11.0$. All of the galaxies are taken from the *luminosity-limited spiral sample*. Each of the galaxies is assigned to an arm number category by its modal p_m value. All of the modal p_m-values lie in the range $0.5 < p_m \leqslant 0.6$.

mass bin of $\log(M/M_\odot) = 11.2pm0.1$. The m=3 and m=4 samples hint at similar,but much weaker trends. Conversely, the fraction of galaxies with two spiral arms decreases from $0.643pm0.005$ for the total population to $0.55pm0.02$ in the highest stellar mass bin.

One possibility why higher mass spirals may exhibit more spiral arms is that this could purely be an effect from the visual classifications. It has already been identified that the many-armed spiral features are the most difficult to detect, so may be more easily identifiable in the largest, brightest spiral galaxies. Spiral arms are already known to have greater amplitudes (ie. be more prominent) in galaxies with larger stellar masses (Kendall, Clarke & Kennicutt 2015). It has already been demonstrated in §3.3.4 that the m=5+ galaxies are the most incomplete samples, so galaxies with greater stellar mass, that are therefore larger and brighter, may be preferentially put in this category, even after debiasing.

Another interesting scenario may be that the population of galaxies with the highest stellar mass are a population of unquenched spiral galaxies as in Ogle et al. (2016). Such galaxies still have their disks intact, so have no signatures of tidal interactions.As galaxy-galaxy interactions have been linked to both the inducement of two-armed spiral structure (Dobbs et al. 2010; Semczuk & Lokas 2015), and the depletion of gas and therefore quenching (Di Matteo et al. 2007; Li et al. 2008), then one may conclude that the disks of these galaxies have not been disturbed. A possible explanation for this is that lower mass galaxies are more susceptible to environment effects (Bamford et al. 2009),

m	N_{raw}	f_{raw}	N_{W13}	f_{W13}	$N_{debiased}$	$f_{debiased}$	$M_*(\log(M/M_\odot))$	$g-r$	$\Sigma(\mathrm{Mpc}^{-2})$
Luminosity-limited	12554	1.00	14297	1.00	17953	1.00	10.62(0.25)	0.58(0.10)	10.62(0.25)
1	563	0.04	670	0.05	922	0.05	10.63(0.27)	0.58(0.11)	10.63(0.27)
2	9044	0.72	10073	0.70	11150	0.62	10.63(0.24)	0.60(0.10)	10.63(0.24)
3	1778	0.14	2158	0.15	3555	0.20	10.59(0.26)	0.53(0.10)	10.59(0.26)
4	615	0.05	751	0.05	1162	0.06	10.60(0.26)	0.53(0.09)	10.60(0.26)
5+	554	0.04	645	0.05	1164	0.06	10.65(0.27)	0.54(0.09)	10.65(0.27)
Stellar mass-limited	6683	1.00	7226	1.00	9389	1.00	10.81(0.16)	0.63(0.08)	10.81(0.16)
1	290	0.04	331	0.05	495	0.05	10.83(0.16)	0.65(0.09)	10.83(0.16)
2	4852	0.73	5191	0.72	6039	0.64	10.81(0.15)	0.65(0.07)	10.81(0.15)
3	886	0.13	991	0.14	1655	0.18	10.82(0.16)	0.59(0.07)	10.82(0.16)
4	335	0.05	366	0.05	564	0.06	10.82(0.16)	0.58(0.07)	10.82(0.16)
5+	320	0.05	347	0.05	636	0.07	10.85(0.18)	0.58(0.07)	10.85(0.18)

Table 1. Overall properties of galaxy populations with different numbers of spiral arms. The number of galaxies with 1,2,3,4 and more than 4 arms are shown for both the *luminosity-limited* and *stellar mass-limited spiral samples*. Mean stellar masses, colours and local densities are shown for each of the populations, with 1σ standard deviations indicated in parentheses. Errors on the mean ($\sigma/\sqrt{N_{debiased}}$) are all of order < 0.01.

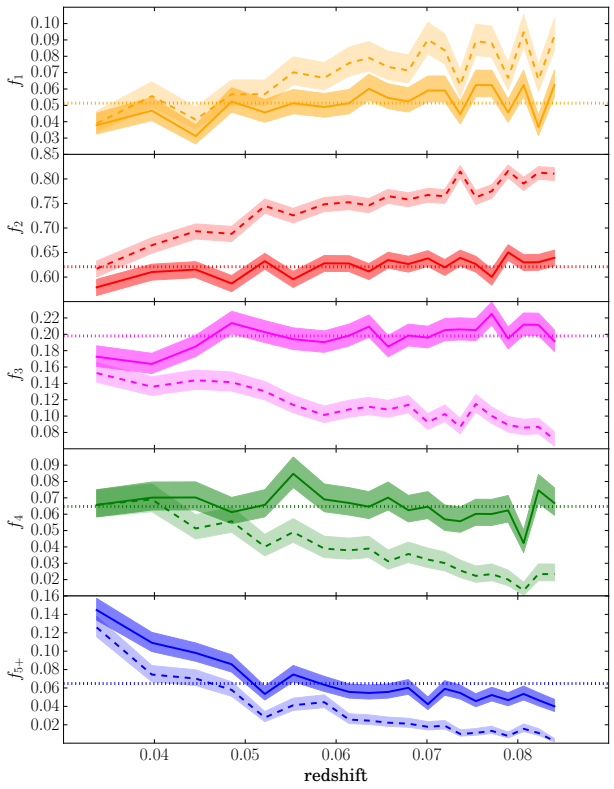

Figure 12. Fraction of galaxies in the *luminosity-limited spiral sample* classified as having 1,2,3,4, or 5+ spiral arms as a function of redshift. The solid lines indicates the fractions from the debiased values, and the dashed line indicates the same fractions using the raw vote fractions. Errors are calculated using the method described in Cameron (2011). The horizontal dotted lines show the mean fractions using the debiased values averaged over all of the bins.

so these disks are still forming stars in the transient way described in §4. As these galaxies have massive disks and high disk fractions, then theory from disk simulations predicts that higher spiral arm modes are favoured (D'Onghia 2015), explaining why the strongest trends are observed in the m=5+ spiral category, and weaker trends are seen for m=3 and m=4.

4.2.3 Local environment

It is already well established that there is a clear dependence of the type of spiral structure that galaxies exhibit with respect to their local environment. Observational evidence from comparison of visually classified galaxies has found that grand design galaxies are more prominent in high density group environments and in binary systems where a close companion galaxy is present (Elmegreen & Elmegreen 1982, 1987; Seigar, Chorney & James 2003; Elmegreen et al. 2011). These results suggest that a mechanism is responsible for the transformation of spiral structure as galaxies enter the highest density environments, with a plausible explanation being that two-armed spiral structure is the result of a recent gravitational interaction. N-body modelling of galaxies has shown that two-armed spiral structure can occur as a result of galaxy-galaxy interactions (Sundelius et al. 1987; Dobbs et al. 2010). However, the timescales of the persistence of such structures are thought to be relatively short-lived (Oh et al. 2008; Dobbs et al. 2010), meaning that an enhancement in the fraction of grand design galaxies is only observed in the highest density environments where interactions can happen on a frequent enough basis to sustain such structures (Elmegreen & Elmegreen 1986).

To compare spiral arm structure as a function of environment, a mean of Σ_4 and $\Sigma 5$ is used as an estimate of local density, as in Baldry et al. (2006); Bamford et al. (2009), denoted as Σ. $\log\Sigma$ is calculated as the mean of the density enclosed within the sphere of the 4^{th} and 5^{th} neighbour,Σ_4 and Σ_5 and is hence an adaptive scale that probes both large scales outside groups and local scales within groups.

The distributions of galaxy local densities for each of the *arm number samples* are shown in Fig. 14a. Here, the *stellar mass-limited spiral sample* is used to define the total population, as M_* and density are closely related (Baldry et al. 2006), so any biases in terms of the stellar mass distributions may have an effect on the completeness of the galaxy sample in terms of environment. The distributions show a modest dependence of spiral arm number with lo-

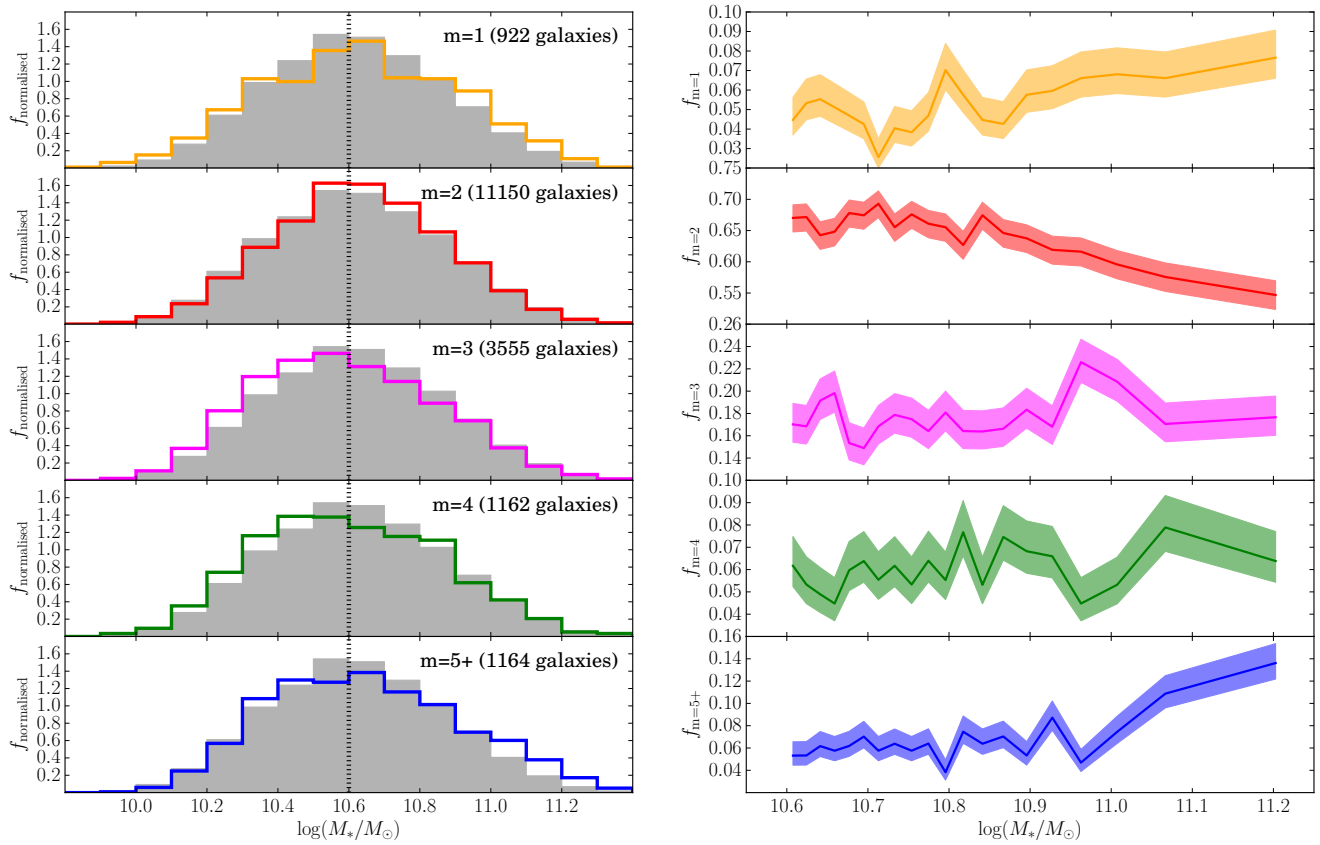

Figure 13. Left: distributions of stellar mass for the *luminosity-limited spiral sample*. The solid lines indicate the distributions for each of the *arm number samples* for each of arm numbers. The grey filled histograms show the equivalent distribution for all of the spiral galaxies for reference. The black dotted line indicates the stellar mass values above which the sample is complete in stellar mass. Right: fraction of the *stellar mass-limited spiral sample* classified as having each spiral arm number, in 20 bins of stellar mass. The shaded regions indicate the 1σ error calculated using the method described in Cameron (2011).

cal density. However, as was the case for stellar mass, each of the arm number samples spans the entire range of local density defined by Σ.

The fraction of spiral galaxies which exhibit each of the spiral arm numbers as a function of $\log\Sigma$ are shown in Fig. 14b. A clear trend is observed, with the number of two-armed spiral galaxies increasing for the highest values of local density from $0.643pm0.005$ for the overall population to $0.75pm0.02$ for the highest density bin of $\log\Sigma = 1.1pm0.2$. Conversely, all of the many-armed samples with m=3,4 or 5+ all show the opposite trends, decreasing with $\log\Sigma$ These results are in agreement with those from Elmegreen & Elmegreen (1982) and Ann (2014), in which the fraction of galaxies displaying grand design spiral structure increases in the highest density environments. As the increase seems to be most distinct in the very highest density environments, this could be indicative that two-armed spiral structure is a short-lived phase induced by galaxy-galaxy interactions, as described in Elmegreen & Elmegreen (1986). A more complete analysis of spiral structure with local environment, taking into account both interaction probabilities and overall density, would need to be considered to test this hypothesis. With our large, clean samples of galaxies with measurements of spiral arm number, we plan to take a more thorough analysis of spiral structure with environment in a future paper.

4.2.4 Galaxy colours

Colours primarily indicate stellar population ages in galaxies, although dust extinction can also have an effect. Star-formation properties have been hypothesised to correlate with spiral arm properties, where galaxies with more prominent spiral arms show enhanced star-formation (Seigar & James 1998; Kendall, Clarke & Kennicutt 2015). The presence of a density wave in a galaxy disk has been proposed as a method of inducing star-formation, but the lack of evidence for a clear enhancement of star-formation in grand design spiral galaxies (Elmegreen & Elmegreen 1986; Foyle et al. 2010; Willett et al. 2015) or a clear age gradient within spiral arms (Foyle et al. 2011; Dobbs & Baba 2014; Choi et al. 2015) suggests that this is not the case.

Galaxy colour is already known to relate to stellar mass (Baldry et al. 2006), environment (Baldry et al. 2004) and overall galaxy morphology (Bamford et al. 2009). As spiral arms are associated with recent star-formation, and also the presence of dust (Grosbøl & Dottori 2012), we expect their properties to correlate with colour. Thus, galaxy colour correlates with the presence of spiral arms, with spiral galaxies being bluer in colour than ellipticals (Bamford et al. 2009; Schawinski et al. 2014).The colour distributions are now compared to look for any trends with recent star formation history in Fig. 15(a). The colours that are plotted here are

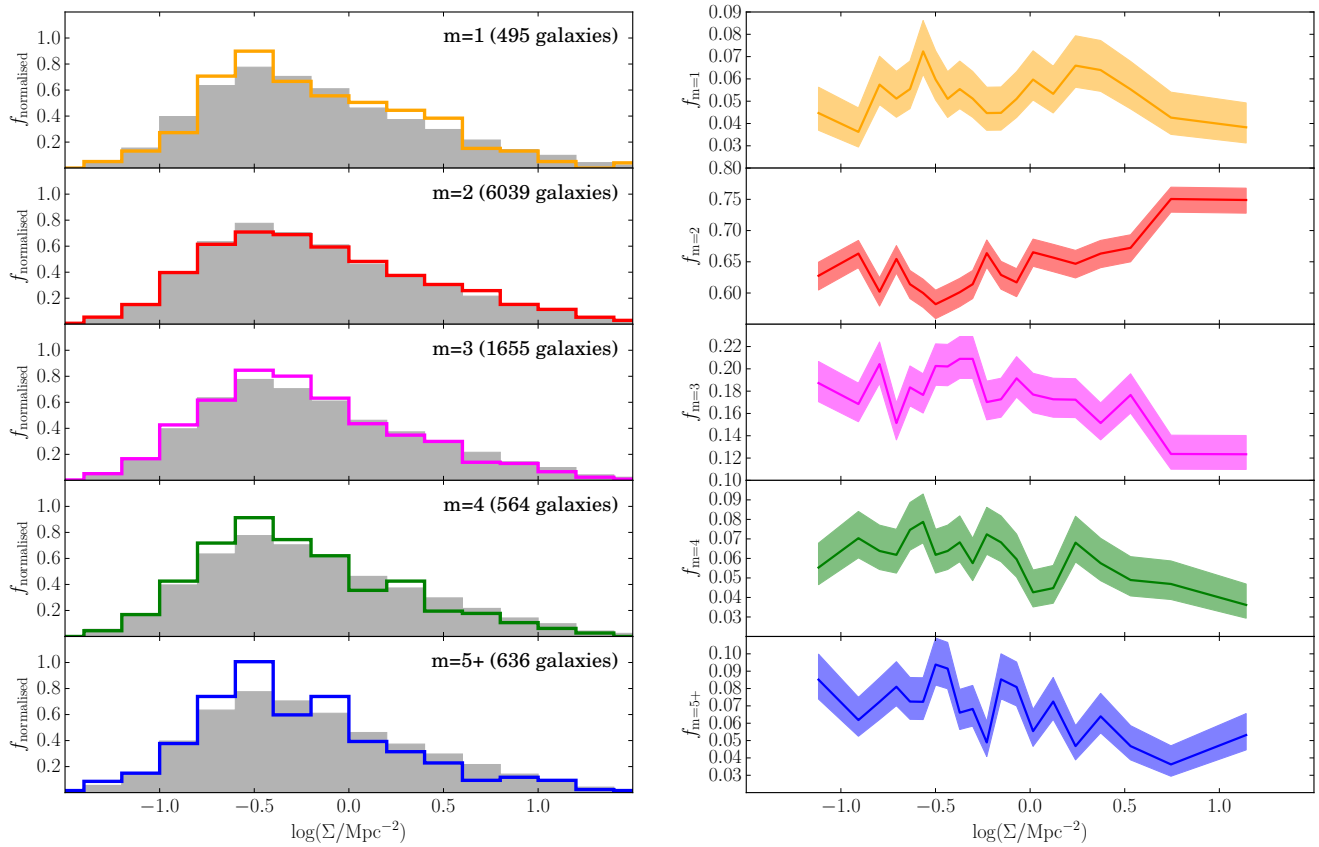

Figure 14. Left: distributions of local density (Σ) for the *stellar mass-limited spiral sample*. The solid lines indicate the distributions for each of the *arm number samples* for each of arm numbers. The grey filled histograms show the equivalent distribution for all of the spiral galaxies for reference. Right: fraction of the *stellar mass-limited spiral sample* classified as having each spiral arm number, in 20 bins of Σ. The shaded regions indicate the 1σ error calculated using the method described in Cameron (2011).

the SDSS $g-r$ optical colours, which should probe recent star-formation in galaxies. Unlike the distributions of local density and stellar mass, a strong trend is found between colour and arm multiplicity. The two-armed spiral galaxies show the reddest overall colours, with mean $g-r$ of 0.65 and a standard deviation of 0.07 in the *stellar mass-limited spiral sample*. The m=3, 4 and 5+ armed samples have corresponding colours of 0.59,0.58 and 0.58, each with standard deviation 0.07. Thus, each of the many-armed spiral samples is ≈1 standard deviation bluer than the two armed spiral galaxy population.

To further compare the overall galaxy colours, the fraction of the *stellar mass-limited spiral sample* with each of the spiral arm numbers with respect to $g-r$ is shown in Fig. 15b. Here, a clear trend is observed with the fraction of galaxies displaying two spiral arms with respect to colour. In the bluest bin ($g-r = 0.47pm0.03$), only 34±2% of galaxies have two spiral arms; in the reddest bin ($g-r = 0.79pm0.05$), $84 \pm 2\%$ have two spiral arms.

Using a single colour only gives a broad indication as to how the star-formation properties of galaxies differ. To try to gain a more detailed understanding of how the stellar populations of spiral galaxies with different numbers of spiral arms differ, the $u-r$ and $r-z$ bands are compared for each of the different arm numbers, and the results are plotted in Fig. 16. It can be seen that the most significant colour differences are observed in the $r-z$ band, which traces SFH over longer timescales, rather than than the $u-r$ band. The most significant differences are observed between the $m=2$ and $m=5+$ samples, where there is a significant offset in $r-z$ for a given $u-r$.

In order to gain more insight the m=2 and m=5+ $u-r$ vs. $r-z$ distributions are plotted are plotted in Fig. 17. The plot includes star-formation history (SFH) models for reference (Bruzual & Charlot 2003). The SFH models are for a quenching galaxy, defined with two values, t and τ, where t is the time of quenching onset and τ is the quenching timescale (a shorter τ means a faster quenching). For each of the three timescales, the dust extinction ,A_v is set to 0. The shift corresponding to a single dust attenuation of 0.4 is also indicated by the green arrow on the figure. The SFH models are described in more detail in §2.2. The plot indicates that both of the colours match SFH model colours, but that the quenching process started earlier in the m=2 population than in the m=5+ population. The m=5+ population has therefore undergone a shorter, more recent phase of star-formation. In particular, a significant population of galaxies that are red in $u-r$ and blue in $r-z$ are found, which cannot even be modelled by a very quickly declining SFR model.

Dust could also be responsible for the colour differences between the galaxy samples. The m=5+ galaxies are bluer in both the $u-r$ and $r-z$ bands, but a more significant offset is seen in the redder $r-z$ band, which could be accounted

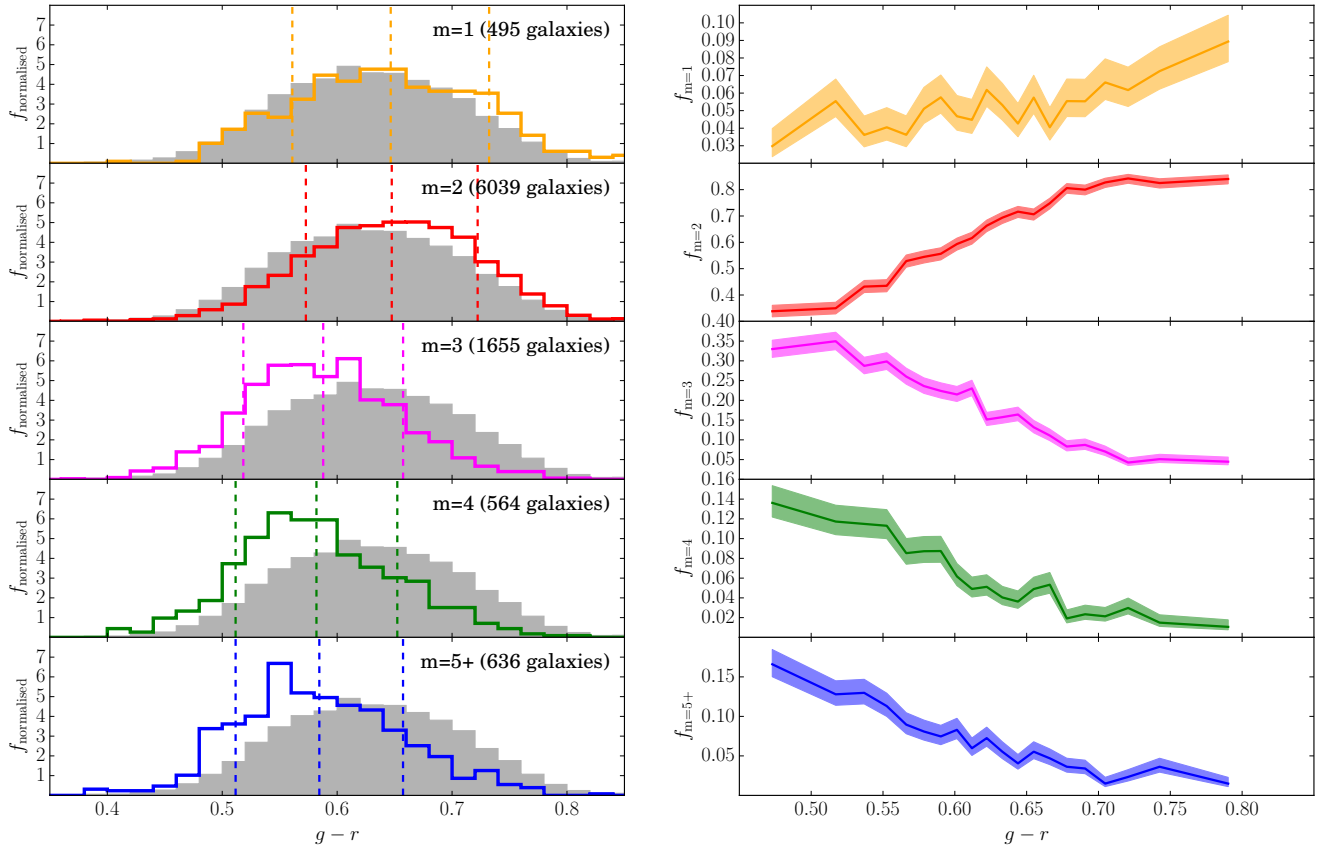

Figure 15. Left: distributions of $g-r$ colour for the *stellar mass-limited spiral sample*. The solid lines indicate the distributions for each of the *arm number samples* for each of arm numbers. The grey filled histograms show the equivalent distribution for all of the spiral galaxies for reference. Right: fraction of the *stellar mass-limited spiral sample* classified as having each spiral arm number, in 20 bins of $g-r$. The shaded regions indicate the 1σ error calculated using the method described in Cameron (2011).

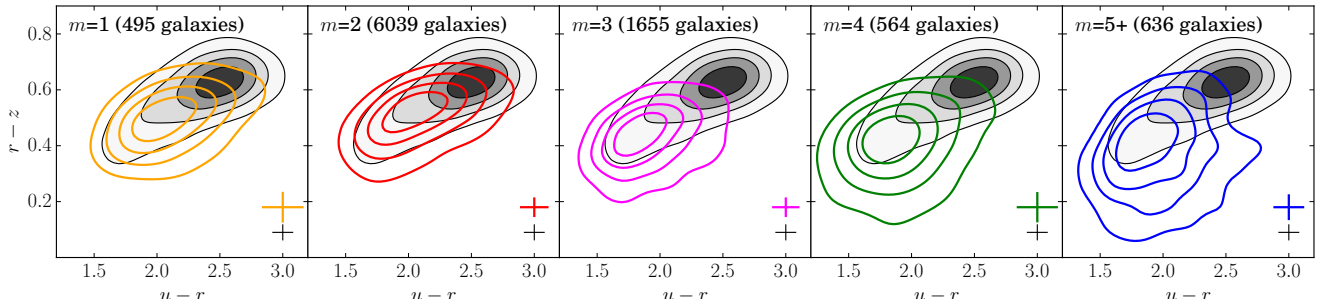

Figure 16. $u-r$ vs. $r-z$ colours for each of the *arm number samples* taken from the *stellar mass limited spiral sample*. The greyscale shaded contours show the entire *stellar mass-limited sample*, irrespective of morphology, whereas the solid lines indicate the same distribution for each *arm number sample*. The contours are plotted with a kernel density estimate, of bandwidth optimised using 5-fold cross validation, and the bandwidths are displayed in the top-left corner of each plot. The contour levels show the regions enclosing 20,40,60 and 80% of the points.

for by a greater level of dust attenuation as shown by the green arrow of Fig. 17. However, recent star-formation and dust attenuation usually correlate (Garn & Best 2010), with the most passive spiral galaxies being dust deficient (Rowlands et al. 2012). It is therefore unlikely that the two-armed spiral galaxies, which are associated with longer term star-formation (ie. longer τ), have greater dust attenuation than the m=5+ spiral galaxy population.

Recent simulations of disks in spiral galaxies have proposed that flocculent spiral structure can be sustained for long periods of time, of order ¿ 10 Gyr, (Fujii et al. 2011; D'Onghia, Vogelsberger & Hernquist 2013), with spiral arms being frequently made and broken. Our results would however suggest that flocculent spiral structure is a short-lived phase, associated with a recent star-formation event. Simulations are generally of galaxy disks in isolation, so may not account for all of the processes that affect spiral galaxies in the local Universe. In particular, the effects that environment or a central bar can have on the gas in spiral galaxies may play a key role in the inducement or transformation

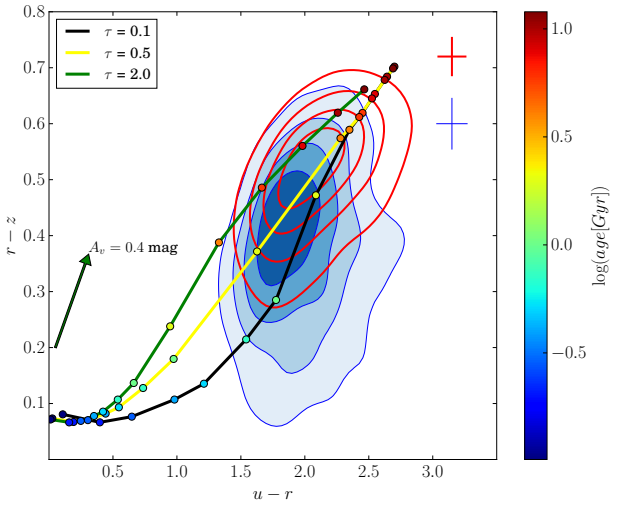

Figure 17. Contour plots for the m=2 (red solid contours)and m=5+ (blue filled contours) samples as in Fig. 16. Three evolutionary tracks for t and τ Bruzual & Charlot (2003) models are also plotted in black, yellow and green lines, and the green arrow indicates how the respective lines would shift with dust extinction.

of the spiral structure observed in local galaxies. To gain a more complete understanding of these effects, more complex SFH modelling involving multiple stellar populations and consideration of gas fractions will be considered in a future paper. **RH: Does anyone know anything about dust w.r.t spiral galaxy type?**

5 CONCLUSIONS

In this paper, the demographics of a population of local spiral galaxies have been compared with respect to spiral arm number, in order to gain an understanding of any significant differences in the physical processes that play a role in the formation and evolution of spiral structure. We make use of visual classifications of SDSS galaxies from GZ2. In order to obtain complete and clean samples of, we have developed a new method to correct redshift-dependent bias. This corrects the vote fractions to ensure sample completeness, to avoid defects arising from contamination between separate classes of galaxies. The method will also be applicable to for further studies of Galaxy Zoo data, and potentially other citizen science projects.

Using the classifications, the distributions of environment, stellar mass and colour were compared for spiral galaxies with different numbers of arms. It was found that the most massive galaxies favour many-armed spiral structure, which may be indicative that their disks have not have been sufficiently perturbed to induce two-armed spiral structure. An enhancement in the fraction of two-armed spiral galaxies was observed in the highest density environments, indicating that galaxy-galaxy interactions could play a role in the inducement of two-armed spiral structure. The most significant differences between the galaxy populations was found when comparing spiral galaxies with respect to colour and SFH. It was concluded that the many-armed spiral galaxies have had faster, more recent star-formation events than two-armed spiral galaxies, suggesting that this phase is a short-lived event in the lifetime of galaxy disks. Conversely, two-armed spiral galaxies have much longer star-formation timescales.

6 ACKNOWLEDGEMENTS

The data in this paper are the result of the efforts of the Galaxy Zoo 2 volunteers, without whom none of this work would be possible. Their efforts are individually acknowledged at authors.galaxyzoo.org.

RH acknowledges funding from the Science and Technology Funding Council.

The development of Galaxy Zoo was supported in part by the Alfred P. Sloan foundation. Galaxy Zoo was supported by the Leverhulme Trust.

Cross validation methods made use of `scikitlearn.cross_validation` (Pedregosa et al. 2011). This publication made extensive use of the `scipy` Python module (Jones et al. 2001–), including the `scipy.optimize` module for curve fitting and minimisation. This research also made use of `Astropy`, a community-developed core Python package for Astronomy (Astropy Collaboration et al. 2013).

REFERENCES

Abazajian K. N. et al., 2009, ApJS, 182, 543
Ann H. B., 2014, Journal of Korean Astronomical Society, 47, 1
Ann H. B., Lee H.-R., 2013, Journal of Korean Astronomical Society, 46, 141
Astropy Collaboration et al., 2013, AAP, 558, A33
Baba J., Asaki Y., Makino J., Miyoshi M., Saitoh T. R., Wada K., 2009, ApJ, 706, 471
Baba J., Saitoh T. R., Wada K., 2013, ApJ, 763, 46
Baldry I. K., Balogh M. L., Bower R., Glazebrook K., Nichol R. C., 2004, in American Institute of Physics Conference Series, Vol. 743, The New Cosmology: Conference on Strings and Cosmology, Allen R. E., Nanopoulos D. V., Pope C. N., eds., pp. 106–119
Baldry I. K., Balogh M. L., Bower R. G., Glazebrook K., Nichol R. C., Bamford S. P., Budavari T., 2006, MNRAS, 373, 469
Bamford S. P. et al., 2009, MNRAS, 393, 1324
Blanton M. R., Roweis S., 2007, AJ, 133, 734
Block D. L., Bertin G., Stockton A., Grosbol P., Moorwood A. F. M., Peletier R. F., 1994, A&A, 288
Block D. L., Wainscoat R. J., 1991, Nature, 353, 48
Bottema R., 2003, MNRAS, 344, 358
Bruzual G., Charlot S., 2003, MNRAS, 344, 1000
Calzetti D., Armus L., Bohlin R. C., Kinney A. L., Koornneef J., Storchi-Bergmann T., 2000, ApJ, 533, 682
Cameron E., 2011, PASA, 28, 128
Cappellari M., Copin Y., 2003, MNRAS, 342, 345
Carlberg R. G., Freedman W. L., 1985, ApJ, 298, 486
Casteels K. R. V. et al., 2013, MNRAS, 429, 1051
Chabrier G., 2003, PASP, 115, 763
Cheung E. et al., 2013, ApJ, 779, 162

Choi Y., Dalcanton J. J., Williams B. F., Weisz D. R., Skillman E. D., Fouesneau M., Dolphin A. E., 2015, ApJ, 810, 9
Di Matteo P., Combes F., Melchior A.-L., Semelin B., 2007, A&A, 468, 61
Dobbs C., Baba J., 2014, PASA, 31, e035
Dobbs C. L., Theis C., Pringle J. E., Bate M. R., 2010, MNRAS, 403, 625
Doi M. et al., 2010, AJ, 139, 1628
D'Onghia E., 2015, ApJL, 808, L8
D'Onghia E., Vogelsberger M., Hernquist L., 2013, ApJ, 766, 34
Duncan K. et al., 2014, MNRAS, 444, 2960
Elmegreen B. G., Elmegreen D. M., 1986, ApJ, 311, 554
Elmegreen B. G., Elmegreen D. M., 1989, ApJ, 342, 677
Elmegreen D. M., Elmegreen B. G., 1982, MNRAS, 201, 1021
Elmegreen D. M., Elmegreen B. G., 1987, ApJ, 314, 3
Elmegreen D. M. et al., 2011, ApJ, 737, 32
Foyle K., Rix H.-W., Dobbs C. L., Leroy A. K., Walter F., 2011, ApJ, 735, 101
Foyle K., Rix H.-W., Walter F., Leroy A. K., 2010, ApJ, 725, 534
Fujii M. S., Baba J., Saitoh T. R., Makino J., Kokubo E., Wada K., 2011, ApJ, 730, 109
Garn T., Best P. N., 2010, MNRAS, 409, 421
Grand R. J. J., Kawata D., Cropper M., 2012, MNRAS, 426, 167
Grosbøl P., Dottori H., 2012, AAP, 542, A39
Grosbøl P., Patsis P. A., Pompei E., 2004, A&A, 423, 849
Hubble E. P., 1926, ApJ, 64
James R. A., Sellwood J. A., 1978, MNRAS, 182, 331
Jones E., Oliphant T., Peterson P., et al., 2001–, SciPy: Open source scientific tools for Python. [Online; accessed 2016-03-03]
Kelvin L. S. et al., 2014, MNRAS, 444, 1647
Kendall S., Clarke C., Kennicutt R. C., 2015, MNRAS, 446, 4155
Kennicutt, Jr. R. C., 1981, AJ, 86, 1847
Kormendy J., Norman C. A., 1979, ApJ, 233, 539
Land K. et al., 2008, MNRAS, 388, 1686
Li C., Kauffmann G., Heckman T. M., Jing Y. P., White S. D. M., 2008, MNRAS, 385, 1903
Lin C. C., Shu F. H., 1964, ApJ, 140, 646
Lindblad B., 1963, Stockholms Observatoriums Annaler, 22
Lintott C. et al., 2011, MNRAS, 410, 166
Lintott C. J. et al., 2008, MNRAS, 389, 1179
Masters K. L. et al., 2010a, MNRAS, 405, 783
Masters K. L. et al., 2010b, MNRAS, 404, 792
Masters K. L. et al., 2012, MNRAS, 424, 2180
Masters K. L. et al., 2011, MNRAS, 411, 2026
Melvin T. et al., 2014, MNRAS, 438, 2882
Muñoz-Mateos J. C. et al., 2015, ApJS, 219, 3
Ogle P. M., Lanz L., Nader C., Helou G., 2016, ApJ, 817, 109
Oh S. H., Kim W.-T., Lee H. M., Kim J., 2008, ApJ, 683, 94
Pedregosa F. et al., 2011, Journal of Machine Learning Research, 12, 2825
Romanishin W., 1985, ApJ, 289, 570
Rowlands K. et al., 2012, MNRAS, 419, 2545
Schawinski K. et al., 2014, MNRAS, 440, 889
Seigar M. S., Chorney N. E., James P. A., 2003, MNRAS, 342, 1
Seigar M. S., James P. A., 1998, MNRAS, 299, 685
Sellwood J. A., Carlberg R. G., 1984, ApJ, 282, 61
Semczuk M., Lokas E. L., 2015, ArXiv e-prints
Simmons B. D. et al., 2014, MNRAS, 445, 3466
Skibba R. A. et al., 2009, MNRAS, 399, 966
Smethurst R. J. et al., 2015, MNRAS, 450, 435
Sundelius B., Thomasson M., Valtonen M. J., Byrd G. G., 1987, A&A, 174, 67
Thornley M. D., 1996, ApJL, 469, L45
Tojeiro R. et al., 2013, MNRAS, 432, 359
Toomre A., 1981, in Structure and Evolution of Normal Galaxies, Fall S. M., Lynden-Bell D., eds., pp. 111–136
Willett K. W. et al., 2013, MNRAS, 435, 2835
Willett K. W. et al., 2015, MNRAS, 449, 820