University of Groningen

The ATLAS3D project - XXIX. The new look of early-type galaxies and surrounding fields disclosed by extremely deep optical images

Duc, Pierre-Alain; Cuillandre, Jean-Charles; Karabal, Emin; Cappellari, Michele; Alatalo, Katherine; Blitz, Leo; Bournaud, Frédéric; Bureau, Martin; Crocker, Alison F.; Davies, Roger L.

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
Publisher's PDF, also known as Version of record

[Link to publication in University of Groningen/UMCG research database](#)

Submegaparsec individual photometric redshift estimation from cosmic web constraints

M. A. Aragon-Calvo,[1]* Rien van de Weygaert,[2] Bernard J. T. Jones[2]
and Bahram Mobasher[1]

[1]*Department of Physics and Astronomy, University of California, Riverside, CA 92507, USA*
[2]*Kapteyn Astronomical Institute, University of Groningen, PO Box 800, NL-9700 AV Groningen, the Netherlands*

ABSTRACT

We present a method, PhotoWeb, for estimating photometric redshifts of *individual galaxies*, and their equivalent distance, with megaparsec and even submegaparsec accuracy using the cosmic web as a constraint over photo-z estimates. PhotoWeb redshift errors of individual galaxies are of the order of $\Delta z \simeq 0.0007$, compared to errors of $\Delta z \simeq 0.02$ for current photo-z techniques. The mean redshift error is of the order of $\Delta z \simeq 5 \times 10^{-5}$–$5 \times 10^{-4}$ compared to mean errors in the range $\Delta z \simeq 0.001$–0.01 for the best available photo-z estimates in the literature. Current photo-z techniques produce redshift estimates with large errors due to the poor constraining power the galaxy's spectral energy distribution and projected clustering can provide. The cosmic web, on the other hand, provides the strongest constraints on the position of galaxies. The network of walls, filaments and voids occupy ~ 10 per cent of the volume of the Universe, yet they contain ~ 95 per cent of galaxies. The cosmic web, being a cellular system with well-defined boundaries, defines a restricted set of intermittent positions a galaxy can occupy along a given line of sight. Using the information in the density field computed from spectroscopic redshifts, we can narrow the possible locations of a given galaxy along the line of sight from a single broad probability distribution (from photo-z) to one or a few narrow peaks. Our first results improve previous photo-z errors by more than one order of magnitude allowing submegaparsec errors in some cases. Such accurate estimates for tens of millions of galaxies will allow unprecedented galaxy-Large Scale Structure (LSS) studies. In this work, we apply our technique to the Sloan Digital Sky Survey photo-z galaxy sample and discuss its performance and future improvements.

Key words: methods: data analysis – methods: numerical – methods: observational – methods: statistical – techniques: spectroscopic – large-scale structure of Universe.

1 INTRODUCTION

Spectroscopic redshifts are our main tool for measuring accurate distances to galaxies and studying the large-scale structure in the Universe. However, they are expensive to determine due to current technological limitations. As a result, only a small fraction of the galaxies observed in photometry are able to be observed in spectroscopy. Photometric redshifts (photo-z), on the other hand, offer a cheap alternative to estimate redshifts of a large number of galaxies using broad-band filters, at the expense of having large errors (typically tens of megaparsecs). Photo-z techniques perform, at their core, some sort of matching or interpolation between the

observed Spectral Energy Distribution (SED) of a galaxy and a known observed or model template (Connolly et al. 1995; Benítez 2000; Csabai et al. 2007; Mobasher et al. 2007; Cunha et al. 2009; Carliles et al. 2010), see also Hildebrandt et al. (2010) for an extensive review and comparison between different photo-z methods. A notable exception to these conventional methods is the approach of Kovač et al. (2010) in which the photo-z probability density is modified by using the local density as a constraint, reducing the photo-z errors to within the scale of the smoothing kernel used to probe the density field. Standard photo-z techniques compute redshift estimates for each galaxy independently taking into account only their SED and produce redshift estimates with typical errors of $\Delta z \sim 0.01$–0.02, equivalent to ~ 40–80 Mpc. A direct way of reducing the redshift errors is by increasing the number of observed colour bands. A state-of-the-art example is the J-PAS multiband
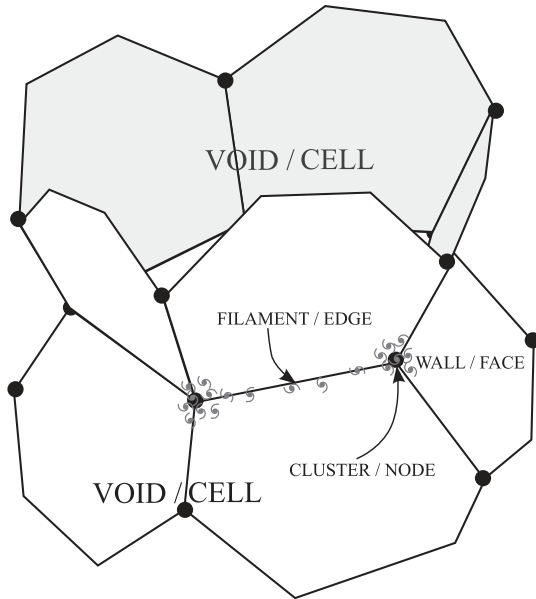
Figure 1. The distribution of galaxies as a cellular system. Voids, walls, filaments and clusters delineated by luminous galaxies correspond to cells, faces, edges and nodes of a cellular system. For illustration purposes, we draw several galaxies delineating two nodes (clusters) and their common edge (filament).

survey with more than 50 narrow bands. Such multiband surveys can be regarded as low-resolution spectroscopy but still produce errors of the order of $\Delta z = 0.003(1 + z)$. Recent improvements on photo-z estimations include clustering information from a reference sample of galaxies for which accurate redshift estimations are available, typically from spectroscopy. This is done by means of correlation functions (Schneider et al. 2006; Newman 2008; Jasche & Wandelt 2012; Ménard et al. 2013; Rahman et al. 2015) achieving *mean redshift errors* of the order of $\Delta z = 0.001$–0.01 but still providing large errors for individual galaxies. While the two-point correlation function provides a measure of clustering, it cannot fully describe the highly anisotropic distribution of galaxies. For this, high-order correlation functions are required (Fry 1984; Szapudi, Szalay & Boschan 1991), and even then still may miss critical phase information, as high-order correlation functions do not capture all clustering information in a sufficiently non-Gaussian field (Carron 2011; Carron & Neyrinck 2012). Correlation functions, while providing some indication of structure, are effectively blind to the cellular nature of the large-scale galaxy distribution (Zeldovich, Einasto & Shandarin 1982; Aragon-Calvo & Szalay 2013) which sets the strongest constraints on the position of a galaxy along the line of sight (LoS).

1.1 Constraints from the cosmic web

The distribution of galaxies forms a cellular system in which large empty voids (cells) are surrounded by a network of increasingly denser walls (faces), filaments (edges) and clusters (nodes) as illustrated in Fig. 1. The origin of the cellular nature of the Universe can be understood by the sequential anisotropic gravitational collapse of a primordial three-dimensional cloud into two-dimensional walls, one-dimensional filaments and finally dense compact clusters (Zel'dovich 1970). How the individual structural elements of the large-scale structure are arranged is (partially) described by

the *cosmic web theory* (Bond, Kofman & Pogosyan 1996) which predicts the emergence of filaments in the space in-between adjacent peaks in the density field corresponding to clusters, and in general, the emergence of a network of filaments joining clusters, groups and galaxies. Another useful picture for understanding the emergence of cellular patterns in the distribution of galaxies is a void-driven cosmic mass distribution in which voids would assume a dominant role in organizing the spatial mass distribution, as described in Icke (1984), Regos & Geller (1991), van de Weygaert (2002), Sheth & van de Weygaert (2004) and Aragon-Calvo & Szalay (2013). In this view, voids act as expanding bubbles that intersect as they fill the available space, forming the dense ridges with walls, filaments and clusters.

Voids occupy most of the volume of the Universe (∼90 per cent) while containing only a minor fraction of the observed luminous galaxies (∼5 per cent). This means that walls, filaments and clusters, accounting for ∼10 per cent of the volume in the Universe, hold 95 per cent of all the galaxies (Aragón-Calvo et al. 2010b; Cautun et al. 2014). The asymmetry between volume occupancy and galaxy number density, in addition to the cellular character of the galaxy distribution, sets strong constraints on the positions a galaxy can occupy given its surrounding matter configuration. The density field sampled along a LoS is characterized by a set of narrow, intermittent, semiperiodic peaks (Kirshner et al. 1981; Broadhurst et al. 1990; van de Weygaert 1991) separated by shallow valleys with a range of sizes delimited by the maximum void size of the order of ∼20–50 Mpc (Platen, van de Weygaert & Jones 2008). We use this information to narrow the possible locations of a given galaxy along the LoS from a broad probability distribution (from photo-z) to a few narrow peaks corresponding to structures in the cosmic web.

1.2 Bias-enhanced cosmic structures

Galaxies are *biased* tracers of the underlying matter distribution (Kaiser 1984; Mo & White 1996). While the matter density field is a volume-filling continuous field, galaxies are concentrated along the dense ridges of the density field and avoid underdense environments. The galaxy population can be considered as a stochastic sampling of a continuous locally biased matter density field. The local bias function generally produces an overconcentration (relative to the matter density field) of galaxies in dense regions (walls, filaments and clusters), and an exponential suppression within voids as their halo mass function is shifted towards low masses (Gottlöber et al. 2003; Goldberg & Vogeley 2004; Cautun et al. 2014; Neyrinck et al. 2014). This is a major reason why haloes are so rare in voids, and why locating galaxies along the cosmic web outlining voids is a good approximation. Fig. 2 shows the local biasing of dark matter haloes as a function of density and halo mass. The cosmic web features delineated by massive haloes are sharp, unlike the smoother structures delineated by less massive haloes and the even more shallow profile of the dark matter. If we consider the density field as a probability density distribution indicating the regions where it is more likely to find a galaxy, it is then clear that the biased galaxy distribution provides strong constraints on the location of a galaxy along the LoS. The constraints are more significant (the difference between peaks and valleys increases) for increasingly massive galaxies.

1.3 The data set

The analysis presented in this work is based on the Sloan Digital Sky Survey (SDSS; York et al. 2000; Abazajian et al. 2009),
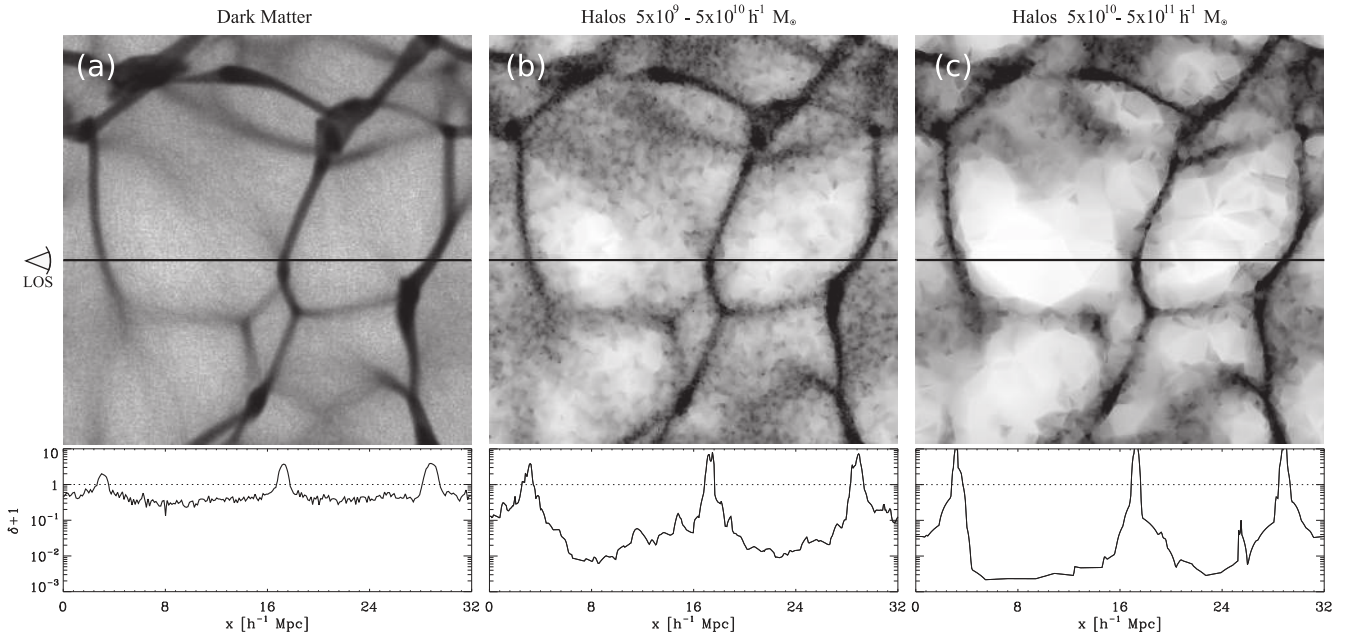
Figure 2. Halo occupation constraints in the cosmic web. Halo bias has a sharpening effect in the features of the cosmic web. (a) dark matter ensemble density field, (b) low-mass haloes ($5 \times 10^9 < M < 5 \times 10^{10}\ h^{-1}\ M_\odot$) ensemble density field, (c) high-mass haloes ($5 \times 10^{10} < M < 5 \times 10^{11}\ h^{-1}\ M_\odot$) density field. The density fields were computed by stacking 256 realizations from the *Multum In Parvo* (MIP) correlated ensemble (Aragon-Calvo 2012). The lower panels show the density profile for the LoS indicated on the left-hand side of the top panels. Halo bias produces an excess of haloes in dense regions and increasingly empty voids for increasing halo mass. Haloes more massive than $10^{11}\ h^{-1}\ M_\odot$ are practically absent from voids. Note that galaxies occupy narrow peaks along the LoS and this becomes more prominent with increasing halo mass. The ensemble density field is equivalent to observing 256 independent LoS so the lack of massive haloes is a real effect and not the product of sampling noise.

containing more than 1 million galaxies with measured spectroscopic redshift and $\sim$200 million galaxies with photometric redshift computed using a KD-tree nearest neighbour algorithm (Csabai et al. 2007). We restrict our analysis to galaxies with photo-z uncertainties below $\Delta z = 0.015$, giving a total of $\sim$6 $\times$ 10^6 galaxies with good photometric redshifts. The spectroscopic sample was used to compute density fields as described below and as a reference for error estimation for both photo-z and the method presented here (see Appendix B for details in the sample selection).

2 THE PHOTOWEB METHOD

Here we provide an outline of our redshift measurement technique, hereafter referred to as PhotoWeb. More detailed description of the method will be presented in Sections 2.1–2.4. Conceptually, PhotoWeb is based on the following two key observations.

(i) The distribution of galaxies follows a well-defined cellular structure with large empty regions surrounded by denser walls, filaments and clusters

ii) Galaxies are a biased sampling of the underlying matter density field, overpopulating the narrow high-density ridges of the cosmic web while avoiding the vast underdense voids.

These two properties of the Universe alone set strong constraints on the most probable location a galaxy can occupy. We should be able to predict the most likely position of a galaxy along the LoS from the distribution of the surrounding galaxies by combining the photo-z Probability Density Function (PDF) $P(z)_{\mathrm{photo}}$, which depends on the particular method used, with constraints derived from the cosmic web itself.

The main requirement for PhotoWeb is a reference sample of galaxies with accurate redshifts.[1] This reference sample must be dense enough to resolve the network of walls, filaments and clusters surrounding voids, since these are the structures that will be used to constrain the location of galaxies along the LoS. Strictly speaking we only need to resolve individual voids since their boundaries give the locations of walls, filaments and clusters. Finding voids is relatively easy using methods that probe the topology of the density field as described below. This means that we could, in principle, use a sample of photometric redshifts as the reference sample provided their typical redshift uncertainty is smaller than the mean void size. In this work, we will use the SDSS spectroscopic sample since the mean photo-z uncertainties are of the order of $\sim$80 Mpc which is several times the mean void size.

The cosmic structures in the reference sample are then used to constrain the position of a sample of target galaxies for which we have a redshift probability density distribution with a large uncertainty, in our case given by a photo-z method, $P(z)_{\mathrm{photo}}$. The photo-z probability distributions have in most cases a dominant peak with a dispersion of the order of $\sigma_{\mathrm{photo}} \sim 0.01$–$0.02$ (for the SDSS). On the other hand, the probability density function computed from the density field (along the LoS) $P(z)_{\mathrm{den}}$, is a complex function with multiple wide-spaced narrow peaks, each with individual dispersions of megaparsec/submegaparsec scale. The key idea in PhotoWeb is combining the broad but unique solution provided by photo-z with the narrow but degenerate constraint provided by the cosmic web in order to collapse the redshift probability distribution into, ideally, one single narrow peak.

---

[1] In practice this corresponds to spectroscopic redshifts but it can be any other method.

Our goal is to compute the probability density distribution given by the cosmic web for a given LoS. In practice we divide this function into two components, one that depends on the local density, which dominates the PDF and one that depends on the geometry of the cosmic web, which adds minor corrections to special cases as explained in the following sections. The PhotoWeb probability density distribution is then

$$P(z)_{\rm web} = P(z)_{\rm photo} \cdot P(z)_{\rm den} \cdot P(z)_{\rm geo}, \tag{1}$$

where $P(z)_{\rm den}$ is the density field sampled along the LoS and $P(z)_{\rm geo}$ is a function that follows the geometry of the cosmic web in a similar way to the density field but giving equal weight to all cosmic structures. Equation (1) assumes statistical independence between $P(z)_{\rm photo}$, $P(z)_{\rm den}$ and $P(z)_{\rm geo}$. In reality $P(z)_{\rm photo}$ is independent from $P(z)_{\rm den}$ and $P(z)_{\rm geo}$ but $P(z)_{\rm den}$ and $P(z)_{\rm geo}$ depend on each other since the cosmic web geometry is computed directly from the density. This means that errors in the density field reconstruction will be propagated to $P(z)_{\rm geo}$ and then applied again in equation (1). This highlights the importance of using robust density and Large Scale Structure (LSS) characterization methods such as the Delaunay Tessellation Field Estimator (DTFE) and Spine methods used in this work. The normalization of $P(z)_{\rm web}$ is not crucial for our purposes as explained in the following sections. In what follows we provide a detailed description of the PhotoWeb method pipeline. We describe the present state of the method, currently in active development and discuss, when appropriate, possible extensions or optimizations to be applied in future versions.

2.1 Density field constraints, $P(z)_{\rm den}$

The reconstruction of the underlying continuous density field from a discrete galaxy distribution is one of the key steps in the PhotoWeb method. The fact that we need a pre-existing density field limits the applicability of PhotoWeb to photometric surveys with overlapping spectroscopy coverage sufficiently dense to sample the structures in the cosmic web. The SDSS spectroscopic sample used in this work provides good spectroscopic coverage up to $z < 0.12$ (see Appendix C). We begin by converting spectroscopic redshifts $z_{\rm spec}$ to distances using the approximation:

$$d = z_{\rm spec}\, c/H_0, \tag{2}$$

where c is the speed of light and $H_0 = 73\,{\rm km\ s^{-1}\ Mpc^{-1}}$ is the Hubble's constant. This approximation is, for our purposes, sufficient to compute distances up to $z < 0.12$.

2.1.1 Removing redshift artefacts

When converting spectroscopic redshifts to distances one must be aware of artefacts arising from peculiar velocities characteristic galaxies in dense environments, where small-scale non-linear interactions produce large velocity dispersions. These *redshift distortions* smear compact dense groups and clusters of galaxies along the LoS, producing the so-called Fingers-of-God (FoG; Kaiser 1987). We identify and correct FoG using the approach based on a Friends of Friends group finder with an elliptical search radius introduced by Tegmark et al. (2004). Groups elongated along the LoS are identified and compressed such that their dispersion along the LoS is equal to the dispersion in the plane of the sky (see Appendix D for details). Other redshift distortions not corrected here include triple-value distortions in the turnaround region around clusters (see Kaiser 1987; Regos & Geller 1991; van Haarlem & van de Weygaert 1993) and

the *Kaiser effect* (Kaiser 1987) originating from large-scale peculiar velocities that compress low-density structures along the LoS. We do not attempt to correct the Kaiser effect as it is small and can even enhance cosmic structures in contrast to the FoG that smear groups along the LoS.

2.1.2 Density field estimation and cosmic web reconstruction

From the FoG-corrected galaxy distribution, we reconstruct the underlying density field using the DTFE technique (Schaap & van de Weygaert 2000; van de Weygaert & Schaap 2009; Platen et al. 2011). Given a spatial distribution of points (galaxies in our case), DTFE estimates the density at the position of each galaxy as the inverse volume of the adjacent Delaunay tetrahedra, i.e. its contiguous Voronoi cell. The density field is then linearly interpolated inside the tetrahedra to produce a continuous density field on a sampling grid (see Appendix A). The Delaunay tessellation follows the anisotropic features in the galaxy distribution and propagates local density estimates across the spaces between sampling points, providing a first-order *structure interpolation* scheme (Aragon-Calvo, van de Weygaert & Jones, in preparation, see Appendix A). In order to compensate for the decrease in the galaxy number density with distance we weight each individual density estimate with the galaxy selection function (see Appendix C). For each galaxy the density estimation is then given by

$$f(\boldsymbol{x_i}) = \frac{(1 + D)}{\psi(z_i)V(\mathcal{W}_i)}, \tag{3}$$

where $D = 3$ is the number of spatial dimensions in which the point distribution is embedded, $V(\mathcal{W}_i)$ is the volume of the contiguous Voronoi cell of the point i and $\psi(z_i)$ is the selection function defined as

$$\psi(z_i) = {\rm e}^{-(z_i/z_r)^\beta}, \tag{4}$$

where z_r is a characteristic redshift defining the peak in the redshift distribution and β encodes the slope of the curve (see Appendix C for details). A single density determination at the centre of a voxel tends to introduce aliasing artefacts in dense regions (where the mean intergalaxy separation can be much smaller than the distance between sampling points). Aliasing artefacts in the density field manifest as hot voxels, which can affect, among other things, the topological properties of the density field. In order to reduce aliasing artefacts, we supersample each voxel with a subgrid with three times higher resolution and set the density at each voxel as the mean of the 27 subgrid locations within the voxel. We use a fast and efficient implementation of the DTFE algorithm based on the publicly available Computational Geometry Algorithms Library (CGAL).[2] Computing the density field on a $1024 \times 512 \times 512$ grid using DTFE takes a few minutes on a modern workstation.

We then use the density field sampled along the LoS of the target galaxy directly as the probability of a galaxy residing at redshift z along the LoS, $P(z)_{\rm den}$.

2.2 LSS geometry constraints, $P(z)_{\rm geo}$

The density field reconstructed from the distribution of galaxies has a wide dynamic range, as illustrated in Fig. 2. The difference in density between the interior of a void and the centre of a cluster or filament is of several orders of magnitude. This property of the

---

[2] www.cgal.org

galaxy density field can introduce errors in the redshift estimation in cases when the LoS of a galaxy intersects a filament or wall in the vicinity of a cluster, or when the LoS intersects a cluster within a couple of standard deviations from the mean in the photo-z probability distribution of the target galaxy. In these situations $P(z)_{den}$ is dominated by the cluster, even if the cluster is relatively far from the true position of the galaxy. In order to compensate for this effect, we include an additional constraint that determines the likelihood of a galaxy being at a given position along the LoS based on the *geometry* of the cosmic web and not on its density. Since we are only interested in the geometry of the structures, we use a transformation that gives equal weight to all cosmic structures (here the term *cosmic structure* refers to the set of walls, filaments and clusters and excludes voids which are their complement). We use the distance from a target galaxy to the nearest cosmic structure to estimate the probability of the galaxy being associated with that structure. This is done by means of the *distance transform*. The distance transform is a mathematical morphology operator, usually applied to a binary image, that specifies the distance from each voxel to the nearest non-zero voxel (or boundary) according to a given metric. The distance transform of a set of points on a grid $P \subseteq G$ is computed by associating to each grid location the distance to the nearest point in P as

$$D_P(p) = \min_{q \in P} d(p, q), \tag{5}$$

where $d(p, q)$ is the Euclidean distance between p and q. If the target galaxy is located inside a cluster, the result from the density constraint alone is not affected by the geometric constraint. If however, the target galaxy is located inside a filament or wall and there is a nearby cluster then the distance transform will give more weight to the cosmic structure closest to the target galaxy instead of the cluster, providing a way to compensate for the large dynamic range of the densities in the cosmic web.

2.2.1 Cosmic web identification

In order to compute the distance transform we need a discrete representation of the density field that encodes the cosmic structures as a binary set such as

$$M = \begin{cases} 1 & \text{cluster, filament, wall} \\ 0 & \text{void.} \end{cases} \tag{6}$$

This requires a way of separating, or segmenting, voids from the other cosmic structures. We perform a full identification of voids, walls, filaments and clusters in the density field using the SPINE method (Aragón-Calvo et al. 2010b) which provides, from a continuous scalar field, a discrete label for each voxel in the density field as void, wall, filament or cluster. The SPINE method assigns morphological labels based on the topology of the density field, encoded in the *watershed transform* (Beucher & Lantuejoul 1979; Platen, van de Weygaert & Jones 2007). Fig. 3 shows the SPINE field (middle panel) computed from the density field on the top panel. The SPINE method provides a cellular segmentation of space corresponding to voids, as well as a full description of the cosmic web. In this work, we only use the spine i.e. the union of the set of walls, filaments and clusters as the basis to compute the distance transform.

From the cellular segmentation provided by SPINE, we then proceed to compute the distance transform using an efficient algorithm that takes advantage of the space-partitioning character of the watershed transform (see Appendix E), giving us an objective way to quantify association with a cosmic web structure. In practice we use
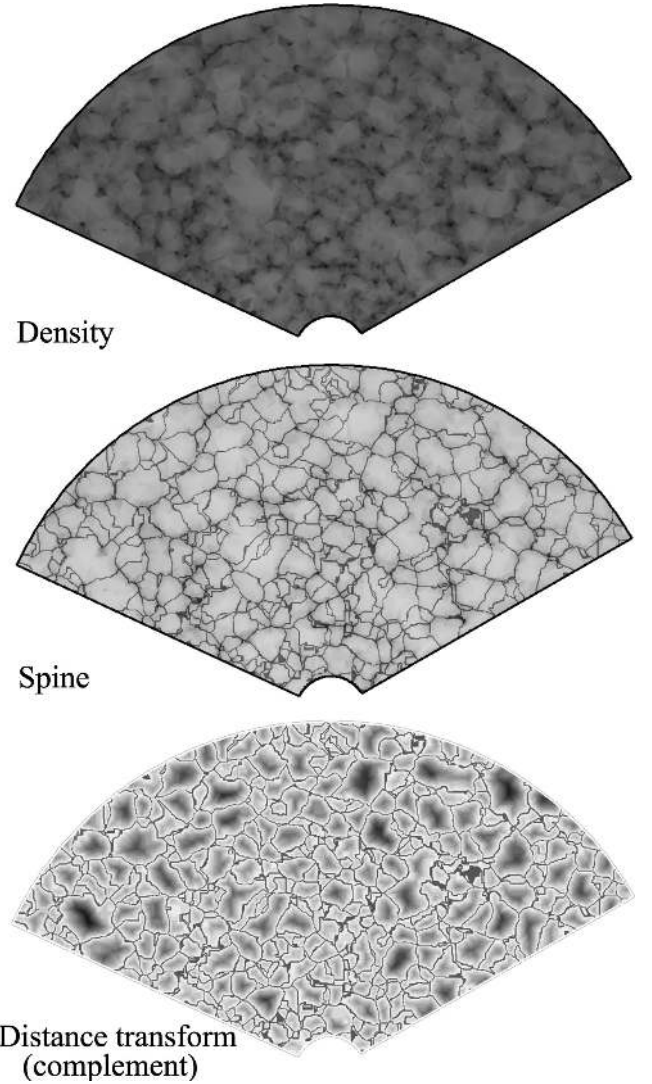

Figure 3. From top to bottom: density field, spine (walls, filaments and clusters) and distance transform on a thin slice computed from the SDSS spectroscopic sample. For clarity the density field is inverted so dark pixels indicate high-density regions. The spine in the middle panel is shown as the dark contours delineating voids. The distance transform (also inverted) has similar properties as the density field, decreasing from the spine into the centres of voids.

the complement of the distance field to produce a monotonically increasing scalar field akin to the density field but giving equal weight to clusters, filaments and walls. The distance transform along the LoS of a galaxy can then be considered a probability density distribution $P(z)_{geo}$ giving us the likelihood of finding a galaxy along the LoS on the basis of the geometry of the cosmic web (see Fig. 3).

2.3 Adding density and cosmic web constraints

The final step in the PhotoWeb pipeline is computing the product of the photometric probability density distribution and the density and cosmic web constraints as indicated in equation (1), here repeated for convenience:

$$P(z)_{web} = P(z)_{photo} \cdot P(z)_{den} \cdot P(z)_{geo}.$$

In practice, for simplicity we replace the full photo-z probability distribution of a given galaxy for a Gaussian distribution centred at

the mean redshift and with dispersion equal to the uncertainty z_{err} provided by the KD-tree photo-z method (Csabai et al. 2007, see Appendix B). This is a good approximation for cases where there is a single dominant peak in $P(z)_{\mathrm{photo}}$. However, $P(z)_{\mathrm{photo}}$ can have more complex distributions depending on the photo-z method used and the choice of filters. In general a small filter sample with broad passbands will produce a broader and less constrained $P(z)_{\mathrm{photo}}$. Also, within the same survey, galaxies at different redshifts will be sampled differently by the same set of filters which will be reflected in their $P(z)_{\mathrm{photo}}$. However, given the low redshift range of the sample used here ($z < 0.12$) and the relatively good coverage of the SDSS filters with well defined, almost independent bandpasses we can assume well-behaved photo-z estimates for the purposes of presenting the method.

Fig. 4 shows a schematic of the PhotoWeb pipeline starting from the density field reconstruction. The density field probability distribution, $P(z)_{\mathrm{den}}$, is characterized by a few narrow high peaks. Note that the dominant peak in $P(z)_{\mathrm{den}}$ is not the peak corresponding to the target galaxy $z = 0.035$. There is one significantly higher peak at $z \sim 0.02$. If the photo-z probability distribution was broader then the larger peak would likely be selected as the PhotoWeb-z. This highlights the importance of having accurate photo-z with narrow PDF. We will discuss this in detail in Section 3.3. In contrast, the features in $P(z)_{\mathrm{geo}}$ delineate cosmic structures and are also correlated to the peaks in the density field, but the peaks in $P(z)_{\mathrm{geo}}$ have all the same height, thus giving equal significance to all cosmic web structures regardless of their local density. By multiplying the broad single-peak photo-z distribution with the narrow multiple-peak density and distance transform distributions we collapse the probability distribution and obtain a unique solution given by the highest peak in the $P(z)_{\mathrm{web}}$ distribution.

2.4 Assigning PhotoWeb redshifts

In the present implementation of PhotoWeb we use a simple approach and assign the PhotoWeb -z to the redshift where the distribution $P(z)_{\mathrm{web}}$ has its maximum value. This prescription is sufficient for our purposes in this introductory paper, but it also produces artificially large galaxy biasing. Since we assign galaxies only to the peak in $P(z)_{\mathrm{web}}$ we impose a complete lack of galaxies inside voids. If a target galaxy is located inside a void (which in itself has a very low likelihood) it will be artificially shifted towards the nearest cosmic structure. In terms of the global population, galaxies in low-density environments represent a minor fraction but nevertheless an important one. One way to alleviate this would be to implement a Monte Carlo sampling of $P(z)_{\mathrm{web}}$ or some transformation of it such as $\log P(z)_{\mathrm{web}}$ which provides a more natural density variable in the non-linear regime (Coles & Jones 1991; Neyrinck, Szapudi & Szalay 2009; Carron 2011). The sampling scheme can also include a biasing prescription matched to the observed local galaxy bias.

3 RESULTS

In this section, we present our first results and perform tests in order to understand the errors and systematics in the PhotoWeb method. In order to measure redshift errors we applied PhotoWeb to galaxies in the SDSS spectroscopic sample, which also have photo-z estimates (Csabai et al. 2007, see Appendix B for details). We use spectroscopic redshifts as a reference, assuming that they provide accurate redshifts and estimate distances, which is valid for low- and medium-density environments and breaks down in dense

clusters. However, since we also correct FoG we simply assume all spectroscopic redshifts and positions to be correct.

3.1 PhotoWeb versus photo-z redshift estimation

We start by comparing how well the PhotoWeb redshifts follow the reference spectroscopic redshift. Fig. 5 shows a comparison between PhotoWeb redshifts z_{web} and the spectroscopic redshifts z_{spec} used as reference. PhotoWeb redshifts are strongly correlated with the spectroscopic redshifts. The dispersion around the slope unit line in Fig. 5 is $\Delta z \sim 0.0006$, equivalent to ~ 2 Mpc, and the FWHM is of the order of $\Delta z \sim 0.0002$ or ~ 1 Mpc. The 2D distribution is dominated by a narrow diagonal ridge and shallow broad tails. However, the distribution is non-Gaussian and the tails extend beyond what we expect for a Gaussian distribution. As a result of this, the total dispersion is higher than the one estimated from the narrow peak (this is explored in detail in Section 3.3.1). Nevertheless more than half of the galaxies fall inside the narrow ridge. For comparison we show the same plot for photo-z estimates (right-hand panel in Fig. 5). Here the correlation is not tight and the dispersion is significantly larger, of the order of $\Delta z \simeq 0.01$–0.02, corresponding to ~ 40–80 Mpc. The correlation between z_{web} and z_{spec} is very tight even at low redshifts ($z < 0.05$) where the photo-z estimates, used as starting point for PhotoWeb, have consistently higher values than their corresponding spectroscopic redshift.

3.2 PhotoWeb spatial galaxy distribution

The narrow dispersion in the PhotoWeb redshift estimations allows us to assign accurate redshifts to *individual galaxies*. Fig. 6 shows the spatial distribution of galaxies with redshifts assigned using PhotoWeb from the SDSS photo-z catalogue computed by Csabai et al. (2007). We convert redshift to distance using the approximation in equation (2). There are $\sim 2.2 \times 10^6$ galaxies with photo-z estimates in this redshift range, compared to $\sim 5 \times 10^5$ galaxies with spectroscopic redshift. For comparison we also show the galaxy positions computed using the photo-z and spectroscopic samples. There are no visible structures in the photo-z galaxy distribution. In stark contrast, the distribution of PhotoWeb-z galaxies shows a rich cosmic web. There are roughly five times more galaxies in the PhotoWeb sample compared to the spectroscopic sample. While the redshift estimation with PhotoWeb is clearly superior to photo-z there are also some visible artefacts. There is a lack of structures aligned with the LoS, reflecting the compression algorithm used to correct the FoG redshift distortions (see Jones, van de Weygaert & Aragón-Calvo 2010). This artefact could be alleviated by computing the density field from the raw spectroscopic positions, but the density field would be dominated by redshift distortions. There are also practically no galaxies inside voids. This is the result of taking the largest peak in the PhotoWeb probability distribution as the unique solution instead of performing a Monte Carlo sampling over $P(z)_{\mathrm{web}}$. In this work use direct assignment instead of a stochastic sampling for simplicity and due to our current lack of knowledge on the relation between the DTFE reconstructed galaxy density field and the un-biased matter distribution.

Even when the density and geometry constraints give multiple narrow peaks in $P(z)_{\mathrm{web}}$, this does not affect the shape of the reconstructed cosmic structures. This seems an advantage, but in reality this may introduce significant contamination for local environment studies. The intermittent nature of the peaks in the $P(z)_{\mathrm{web}}$ of an individual galaxy is manifested as a degeneracy in the PhotoWeb
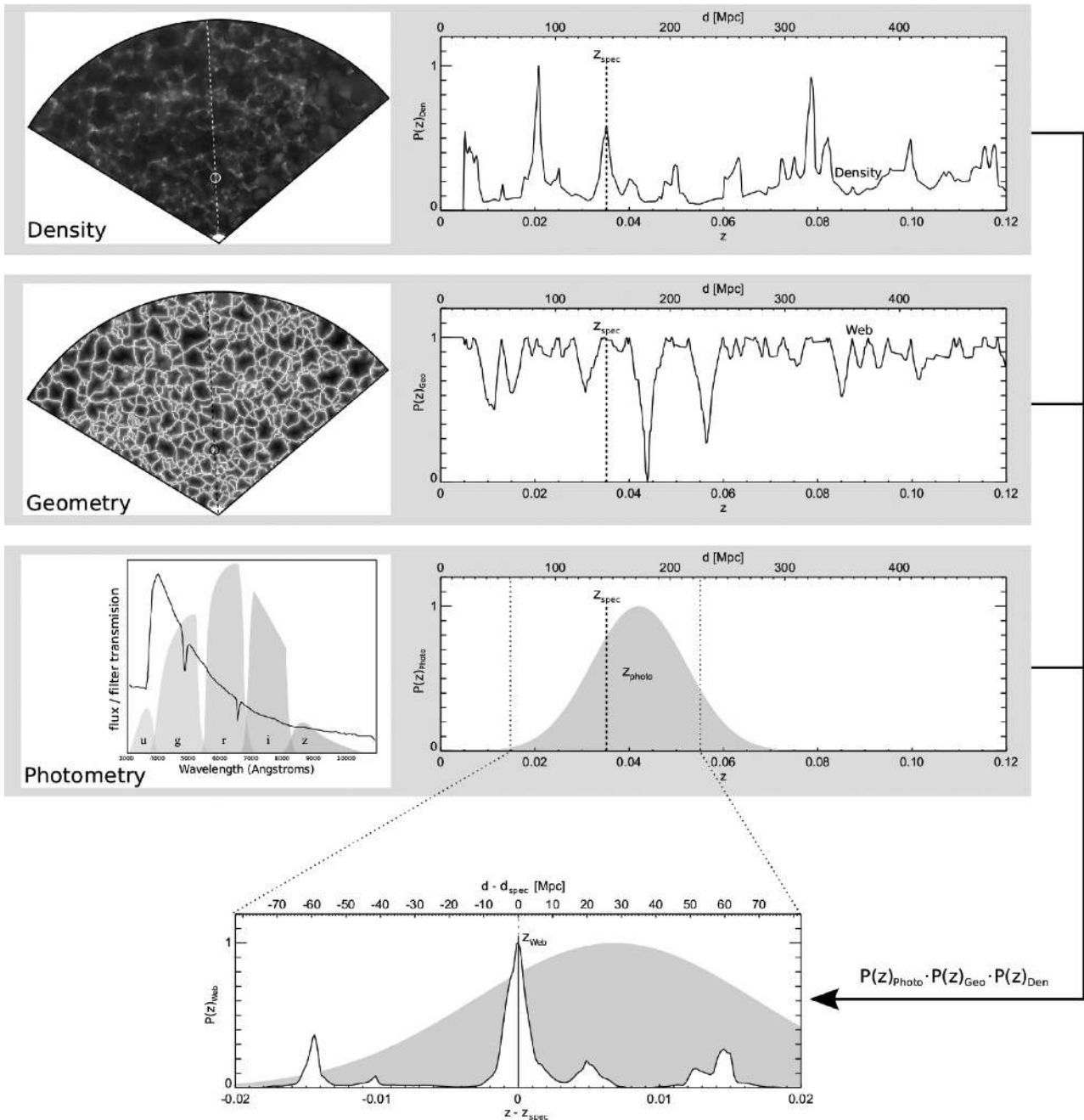
Figure 4. Computing the PhotoWeb PDF ($P(z)_\text{web}$). This diagram corresponds to a galaxy with spectroscopic redshift of $z = 0.035$. We show the three components involved in the computation of $P(z)_\text{web}$: density ($P(z)_\text{den}$ (top), geometry $P(z)_\text{geo}$ (middle) and photometry $P(z)_\text{photo}$ (bottom). The density field and the derived distance transform where computed from the spectroscopic sample. For simplicity we assume $P(z)_\text{photo}$ to be a Gaussian distribution with dispersion equal to the zErr value given by the KD-tree photo-z method (Csabai et al. 2007). The left-hand panels show the density field, distance transform and SED (the SED does not correspond to this particular galaxy and is used only as an illustration). The panels on the right-hand side show the corresponding PDF normalized to 1 at the maximum. The panel at the bottom shows the product of the three PDFs and the resulting $P(z)_\text{web}$. While the photometric redshift has a distance error of $\Delta z_\text{photo} \sim 0.02$ equivalent to ~ 30 Mpc, the assigned PhotoWeb redshift is, in this case, indistinguishable from its spectroscopic counterpart.

redshift solution and this introduces a potentially unknown contamination by placing a galaxy inside a different environment to where it really belongs. Large errors in the PhotoWeb redshift estimation originate from galaxies that are wrongly assigned to a nearby cosmic structure instead of their correct position. In cases when the photo-z $P(z)_\text{photo}$ has a large dispersion and there are dense structures far from the true galaxy's redshift, this error is particularly severe. A possible way to alleviate this is to include colour information in the density estimation and in the assignment of PhotoWeb redshifts. For instance, we expect red luminous galaxies to populate dense environments and blue star forming to avoid them. This information can be used as a prior, at least in some cases, to further constrain the PhotoWeb PDF. This is, however, beyond the scope of this work and will be explored in future work.
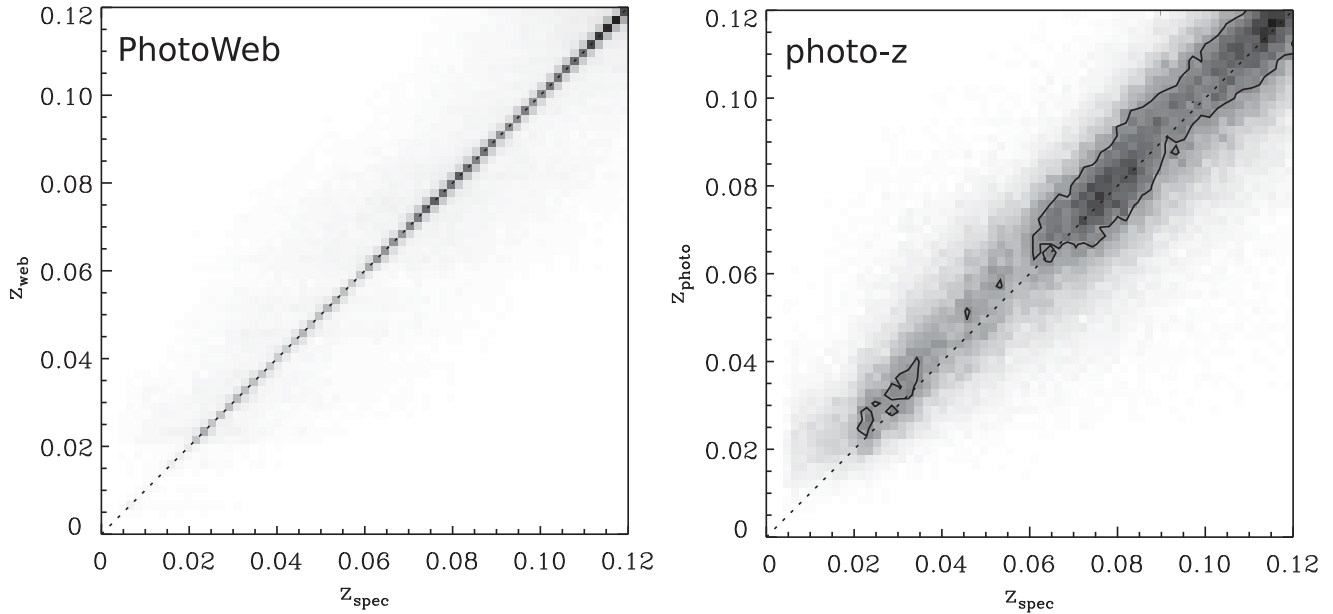
Figure 5. PhotoWeb redshift estimates versus spectroscopic redshifts. The redshift range is divided into 32 bins. Grey levels indicate the counts inside each bin. For comparison we show the photometric redshift versus spectroscopic redshift.

3.2.1 *Where is the information coming from?*

Figs 4 and 6 illustrate the fact that PhotoWeb does not add new information but it simply uses the information already present in the reference spectroscopic sample to assign the most likely locations of galaxies. PhotoWeb would not be able to reconstruct, say, a galaxy group for some (very unlikely) reason missed by the spectroscopic survey. However, the Delaunay tessellation performs a linear interpolation across structures which can be considered a first-order reconstruction. Galaxy surveys like the SDSS provide a spectroscopic sampling dense enough to trace the small features of the cosmic web. Galaxies with no spectra are expected to fill the holes in-between the sampling of galaxies with spectra.

3.3 Estimated redshift error

Fig. 7 shows the distribution of errors in the redshift and distance estimations from photo-z and PhotoWeb. While the photo-z errors have a broad Gaussian distribution, with a dispersion of $\Delta z \simeq 0.01$–0.02, equivalent to $d \simeq 40$–80 Mpc, the distribution of PhotoWeb errors has a narrow peak with dispersion $\Delta z \simeq 0.0006$, equivalent to $d \simeq 2.5$ Mpc. The cumulative probability distribution shows that almost half of the galaxies have PhotoWeb redshift errors one order of magnitude smaller than photo-z. Compare this to typical photo-z errors of $\Delta z \sim 0.02$ corresponding to a scale of ~ 80 Mpc. The cumulative distribution shows a marginal excess of PhotoWeb compared to photo-z for errors larger than ~ 70 Mpc, which is of the order of the mean dispersion in the photo-z errors. This corresponds to ~ 15 per cent of the PhotoWeb galaxies having worse redshift estimations than photo-z. On the other side of the cumulative distribution we see that as the photo-z errors decrease, their corresponding PhotoWeb errors also decrease at a higher rate. When the photo-z errors are smaller than 10 Mpc the PhotoWeb errors are one order of magnitude smaller. Fig. 8 shows that for large photo-z errors $\Delta z_{photo} > 0.01$ the PhotoWeb errors also perform poorly. Below Δz_{photo} the PhotoWeb errors are smaller, almost independently of Δz_{photo} and with a peak around $\Delta z_{web} \sim 0.0007$.

3.3.1 *The nature of the PhotoWeb errors*

The distribution of PhotoWeb errors, Δz shown in Fig. 7 has a clearly non-Gaussian shape. It consist of a high and narrow peak and tails extending well above what we expect from the exponential decline of the Gaussian function. The narrow and degenerate nature of the peaks in $P(z)_{web}$ results in either producing a very small redshift error when the galaxy is assigned to its correct cosmic structure, or a large redshift estimate when the galaxy is assigned to cosmic structures on the opposite sides of the adjacent voids to the structures where the galaxy is truly located. Note that we are assuming that $P(z)_{photo}$ is Gaussian, which although is reasonable for the galaxy sample used in our analysis, it is in general not the case (see Section 2.3). A more complex $P(z)_{photo}$ will affect the shape of the distribution of PhotoWeb errors, possibly introducing biases. We model the distribution of PhotoWeb errors into two components, one reflecting the small-scale errors from correctly assigned galaxies and another component describing the errors coming from assigning the galaxy to opposite cosmic structures:

$$F(x) = F(x)_s + F(x)_l, \tag{7}$$

where x is the redshift error Δz and $F(x)_s$, $F(x)_l$ are the contributions from small and large scales, respectively. Small-scale errors can be assumed to be Gaussian originating from uncertainties in the density estimation and non-linear processes such as peculiar velocities. Large-scale errors, on the other hand, originate from the discrete intermittent sampling of the cosmic web (see Fig. 4). An illuminating analogy is the intersection of a LoS with a Voronoi cellular distribution. The LoS will intersect the faces, edges and nodes of Voronoi cells (Icke & van de Weygaert 1991). The intersecting points will be separated by the empty spaces of Voronoi cells (van de Weygaert 1991, 1994), see also van de Weygaert (2003). The LoS of target galaxies will intersect the cosmic structures at any position producing a roughly constant error distribution which is then weighted by the semi-Gaussian probability distribution of the photo-z estimation. For simplicity we assume a Gaussian
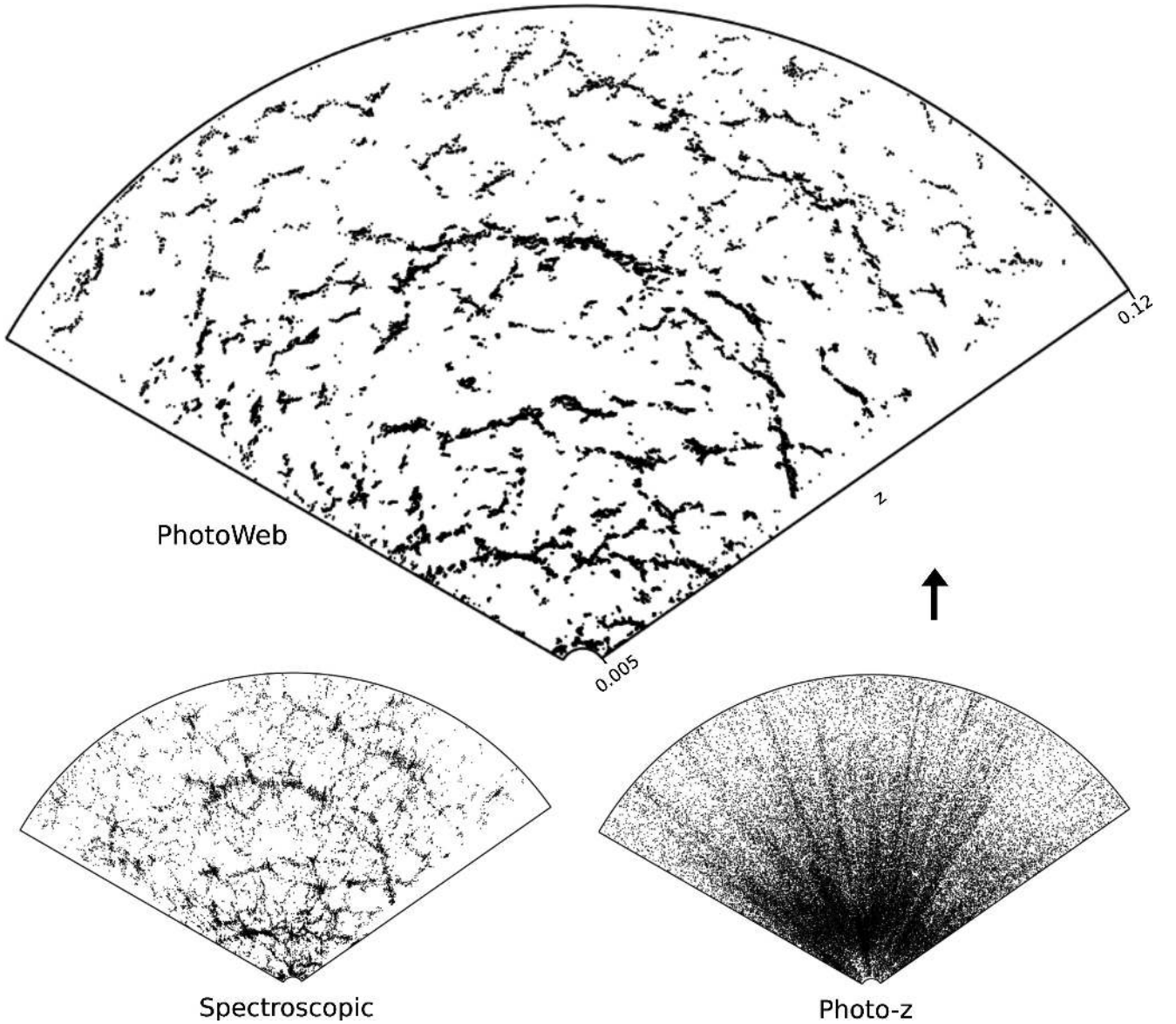
Figure 6. Galaxy distribution from PhotoWeb, photo-z and spectroscopic redshifts. We show galaxies in a 10 Mpc-thick slice across the SDSS survey. The bottom-left panel shows the galaxy positions from spectroscopic redshifts without FoG compression. The top-left panel shows the photo-z positions. The top-right panel shows the PhotoWeb positions. The distribution of galaxies from PhotoWeb follow the cosmic web delineated by the spectroscopic reference sample. Note that the distribution of galaxies from photo-z show barely any feature. The PhotoWeb positions are highly biased, there are no galaxies inside the voids and the cosmic structures have very sharp boundaries. There is also a lack of filaments aligned with the LoS as a result of the FoG compression used to derive the density field. Nevertheless the PhotoWeb positions are sticking. This diagram highlights the unique potential of PhotoWeb for small-scale environmental studies.

distribution for both small-scale and large-scale errors. The total PhotoWeb error is then modelled as

$$F(x) = Ae^{-x^2/2\sigma_s^2} + Be^{-x^2/2\sigma_l^2}. \qquad (8)$$

The best fit to the distribution in Fig. 9 (normalized to 1) gives $A = 1$, $\sigma_s = 0.0007$, $B = 0.08$ and $\sigma_l = 0.010\,9593$. The small-scale dispersion in the redshift error corresponds to a physical scale of ~ 2.5 Mpc, which is consistent with being originated from noise at the scale of cosmic structures such as walls, filaments and clusters, where most of the LoS intersect the cosmic web (Aragón-Calvo, van de Weygaert & Jones 2010a). The large-scale redshift error dispersion is closer to the value of the mean redshift error in the photo-z estimates used as input for PhotoWeb $\sigma_{\text{photo-z}} \sim 0.02$. Note that this

two-component model is not strictly correct since the large-scale component $F(x)_l$ also contributes to small-scale errors (near the mean) from correctly assigned galaxies but nevertheless, it serves as a reference to compare the contribution of small- and large-scale errors.

3.3.2 Mean redshift errors

An important quantity for cosmological studies is the mean redshift measured inside redshift bins. Even though photometric redshifts have large individual errors, their aggregate values approach the mean redshift inside a given redshift bin. Typical mean redshift errors from photo-z are of the order of 0.001–0.01 (Carliles et al.
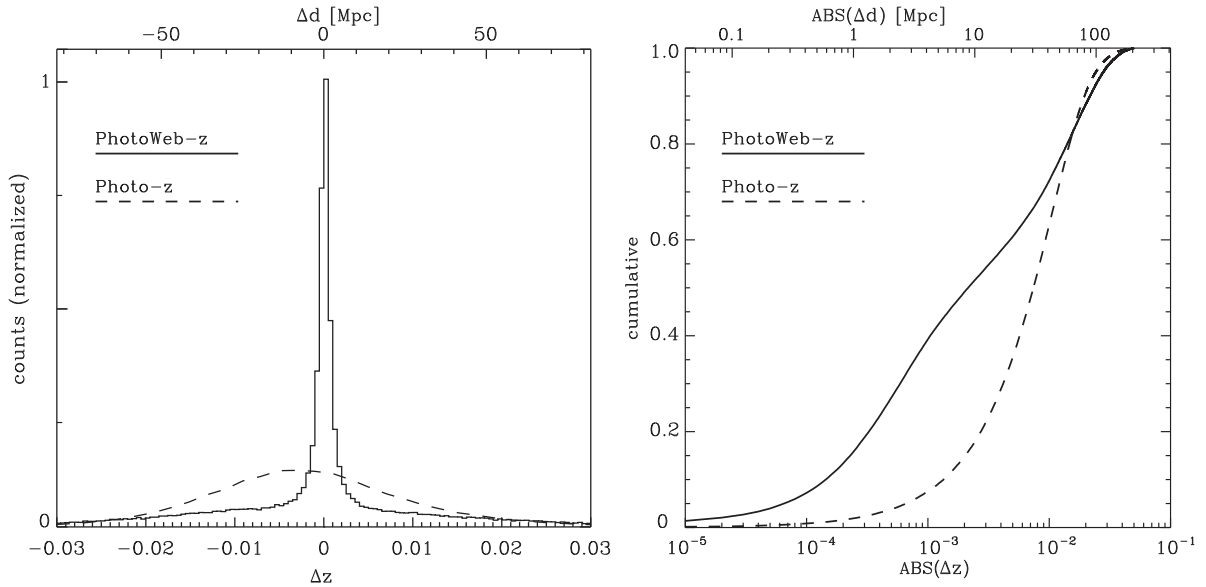
Figure 7. Redshift estimation error. Probability distribution (normalized) of the errors in PhotoWeb redshifts and photo-z redshifts. The cumulative distribution (right-hand panel) shows that almost half of the galaxies have PhotoWeb redshift estimates with errors one order of magnitude smaller than their corresponding photo-z.
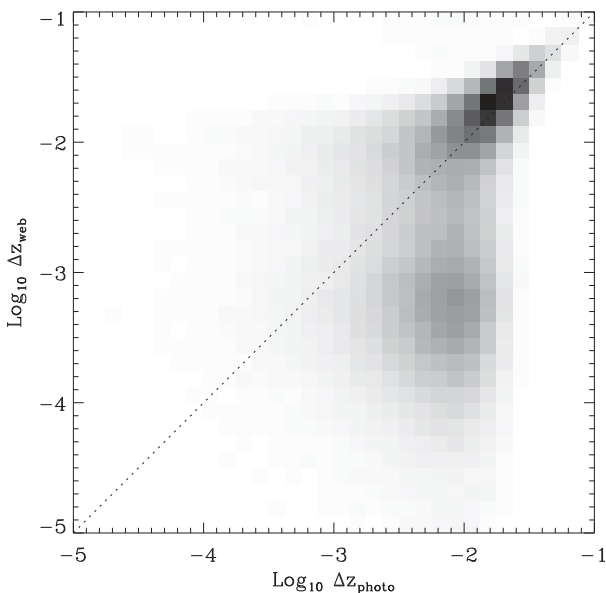

Figure 8. PhotoWeb redshift error versus photo-z redshift errors. The grey-scale indicates the counts inside each 2D bin. For large photo-z errors ($\Delta z_{web} > 0.01$) the PhotoWeb errors also perform poorly. At smaller photo-z errors ($\Delta z_{web} < 0.01$) the PhotoWeb errors are significantly smaller than photo-z.

2010; Rahman et al. 2015). However, the large dispersion in $P(z)_{photo}$ effectively acts like a smoothing (which may not be Gaussian or even isotropic as assumed here, see Section 2.3) and prevents us from resolving features in the redshift distribution at scales smaller than the typical redshift uncertainties, which for the SDSS photo-z sample is ~0.01–0.02.

Fig. 10 shows the truncated mean of the redshift errors computed in several redshift bins. PhotoWeb is sensitive to biases in the photo-z estimation that affect $P(z)_{photo}$ and may be impossible to separate from $P(z)_{den}$ and $P(z)_{geo}$ (equation 1). The SDSS photo-z catalogue used here has systematically higher estimations at low

redshift ($z < 0.05$) and as far as $z \sim 0.1$ as shown in Fig. 5. In order to correct for this effect we ignore errors larger than two standard deviations of the respective error distribution (see Section 3.3 and Fig. 7). The mean truncated PhotoWeb redshift errors are practically indistinguishable from the spectroscopic redshifts with values in the range $\Delta z \simeq 10^{-5}$–10^{-4}. These values are more than one order of magnitude smaller compared to photo-z estimates computed using only colour information (our reference sample) and even photo-z estimates reported in the literature after including galaxy correlation information $\Delta z = 0.001$–0.01 (Rahman et al. 2015). Note, however, that the mean errors were computed from the truncated distribution, this assumes that we have access to unbiased photo-z estimates which, given the recent advances in photo-z techniques is feasible.

4 CONCLUSIONS AND FUTURE WORK

The method presented here, PhotoWeb, offers an unparalleled accuracy in redshift estimation by combining the broad unique peak in the probability distribution obtained from photo-z estimations with the narrow but degenerated solutions given by the cosmic structures in the galaxy distribution. The current implementation of PhotoWeb demonstrates the constraining power of the cosmic web on galaxy positions along the LoS, achieving redshift errors of the order to $\Delta z \sim 0.0007$ and mean redshifts of the order of ~10^{-5}.

By using PhotoWeb we can increase the number of galaxies with accurate redshifts in SDSS from ~10^6 in the spectroscopic sample to ~5×10^6 (including galaxies with small photo-z uncertainty). This factor of 5 can be further increased as better photo-z estimates become available. This unprecedented galaxy sampling will allow environmental studies of galaxy formation and evolution previously unfeasible with either spectroscopic or photometric redshifts.

The major limitation of PhotoWeb (in its current implementation) is the fraction of catastrophic errors occurring when a galaxy is assigned to a different cosmic structure where it belongs. We note that this fraction is essentially equal to the one derived from photo-z. When PhotoWeb succeeds in identifying the galaxy's host structure it does so superbly. When it fails it does so in a similar
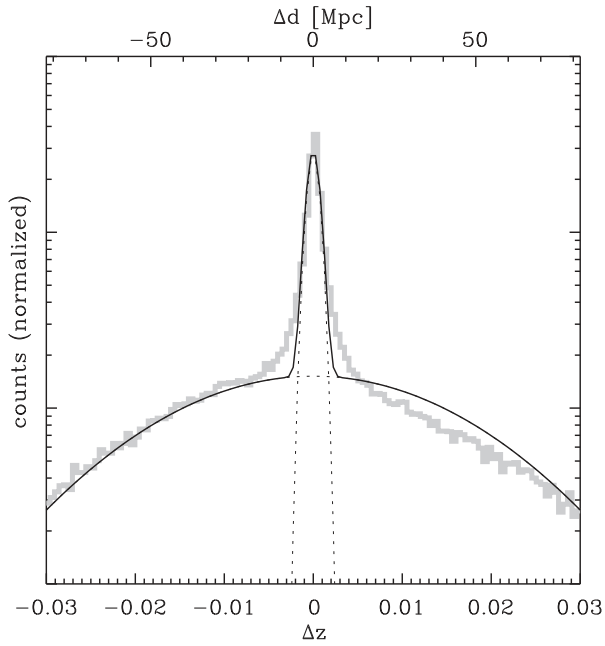
Figure 9. PhotoWeb redshift error model. The grey lines show the distribution of PhotoWeb errors. The black solid line shows the best fit from the two-component model in equation (8). The narrow peak corresponds to small-scale errors from galaxies assigned to their correct cosmic structure. The broad peak corresponds to large-scale errors produced when PhotoWeb assigns a galaxy to a structure on the opposite side of the adjacent void (or even further) to the cosmic structure where the galaxy is truly located. The large-scale dispersion is close to the mean dispersion of the photo-z errors.
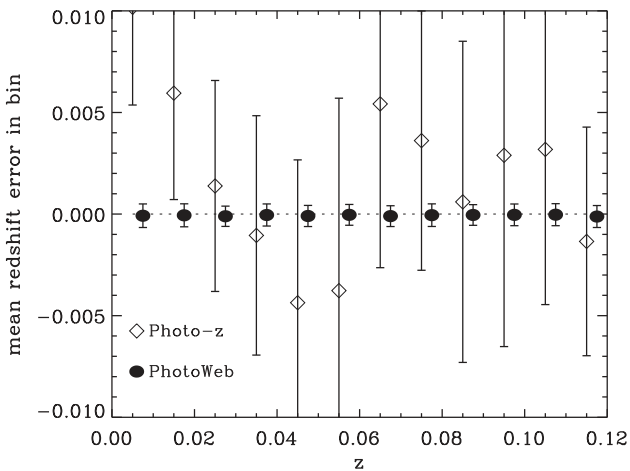
Figure 10. Mean (truncated) redshift error. The circles and diamonds mark the truncated mean redshift inside the corresponding bin and the error bars the dispersion inside the bin. The mean and dispersion where computed from their respective truncated (at 2σ) distribution. The mean redshifts from PhotoWeb are shifted to the right by 0.0025 in z for clarity.

degree as photo-z. This is in part due to the fact that when photo-z fails catastrophically there is no way PhotoWeb can identify the correct cosmic structure. The error in the peak of $P(z)_{photo}$ as well as its dispersion must be of the order of the typical void size for PhotoWeb to succeed. A major challenge is then the reduction of the fraction of galaxies erroneously assigned to nearby cosmic structures. This may seem an impossible task without the use of better

photometric redshifts but there is in fact more information available that we have not yet exploited in PhotoWeb. In particular angular correlations, galaxy colour and morphology, used as a prior, and can play a significant role in further constraining $P(z)_{photo}$. Galaxies populate the cosmic web following well-known patterns such as the colour and morphology–density relation (Dressler 1980). Our lack of knowledge on the underlying physical process does not prevent the use of these empirical relations to constrain galaxy locations along the LoS.

4.1 Future improvements

In the present implementation of PhotoWeb, we use a spectroscopic sample to derive $P(z)_{den}$ and $P(z)_{geo}$. However, for surveys with a large number of narrow-band filters, where the photo-z estimates have an uncertainty of less than a few megaparsecs it may be possible to use the photo-z sample as the reference and apply PhotoWeb iteratively. This is possible because walls, filaments and clusters are located at the boundaries of voids, and voids are relatively easy to identify as local minima in the density field.

The density field used in this work was computed in a regular cubic grid. A polar grid is an alternative option which has some desirable features like a natural degrading of spatial resolution with increasing distance, providing some compensation for the galaxy selection function and a better fit to the survey geometry. We will explore this possibility in a forthcoming paper.

The current implementation of PhotoWeb only takes advantage of the spatial coherence of the cosmic web. However, galaxies also present clustering based on their colour and morphology. The distribution of galaxies has a well-known colour segregation, red early-type galaxies tend to populate high-density regions while blue late-type avoid them and are mostly found in the intermediate- and low-density regions (Dressler 1980). We can use this information to further constrain the location of a galaxy on the LoS. For instance if the target galaxy is classified as late-type and is blue ($u - r < 2.2$) then high-density regions in the LoS density field will be assigned a low probability. The simplest way to do this would be to separate galaxies in colour bins and compute density fields for each colour bin separately but this would significantly reduce the number of available galaxies to reconstruct the density field. Instead we can weight each galaxy according to their colour and compute two separate density fields with *enhanced* red/blue galaxy populations as

$$\rho_{red} = (u - r) \, \rho, \qquad \text{if} \quad u - r > 2.2 \tag{9}$$

$$\rho_{blue} = (r - u) \, \rho, \qquad \text{if} \quad u - r < 2.2, \tag{10}$$

where ρ_{red}, ρ_{blue} are the colour-weighted density estimations and ρ is the original density estimate at the position of the galaxy computed with the DTFE.

Photometric redshift estimation methods are all based on a mapping between the observed properties of galaxies and some known model. The mapping is usually straightforward and applicable to the galaxy sample. However, as early examples of redshift estimation have shown it is possible, at least in some cases, to use extra information to further constrain the location of a galaxy, for instance by looking at signs of interaction with a companion of known redshift, or by comparing the colour/morphology of the galaxy with the galaxies along its LoS to, say, discard massive luminous elliptical galaxies from the centres of voids. These are, so far, anecdotical cases but nothing prevents us from envisioning a

general framework in which every possible piece of information is used to estimate accurate redshifts Budavári (2009). For instance, angular cross-correlations also add strong constraints to the position of a galaxy and therefore to its estimated redshift. This may help resolve some of the degeneracies that produce catastrophic errors when a galaxy is assigned to an incorrect LSS. Our long-term plan is to include all the information available to constraint the position of a galaxy along the LoS. We currently do not have the technology to include all available information in an automated way, however, new advances in image processing, machine learning and automated pattern recognition can enable this kind of analysis in the near future.

REFERENCES

Abazajian K. N. et al., 2009, ApJS, 182, 543
Aragon-Calvo M. A., 2012, preprint (arXiv:e-prints)
Aragon-Calvo M. A., Szalay A. S., 2013, MNRAS, 428, 3409
Aragón-Calvo M. A., van de Weygaert R., Jones B. J. T., 2010a, MNRAS, 408, 2163
Aragón-Calvo M. A., Platen E., van de Weygaert R., Szalay A. S., 2010b, ApJ, 723, 364
Benítez N., 2000, ApJ, 536, 571
Berlind A. A. et al., 2006, ApJS, 167, 1
Beucher S., Lantuejoul C., 1979, in International Workshop on Image Processing: Real-time Edge and Motion Detection/Estimation, CCETT/IRISA Report no. 132. Rennes, France
Bond J. R., Kofman L., Pogosyan D., 1996, Nature, 380, 603
Broadhurst T. J., Ellis R. S., Koo D. C., Szalay A. S., 1990, Nature, 343, 726
Budavári T., 2009, ApJ, 695, 747
Carliles S., Budavári T., Heinis S., Priebe C., Szalay A. S., 2010, ApJ, 712, 511
Carron J., 2011, ApJ, 738, 86
Carron J., Neyrinck M. C., 2012, ApJ, 750, 28
Cautun M., van de Weygaert R., Jones B. J. T., Frenk C. S., 2014, MNRAS, 441, 2923
Coles P., Jones B., 1991, MNRAS, 248, 1
Connolly A. J., Csabai I., Szalay A. S., Koo D. C., Kron R. G., Munn J. A., 1995, A&A, 110, 2655
Csabai I., Dobos L., Trencséni M., Herczegh G., Józsa P., Purger N., Budavári T., Szalay A. S., 2007, Astron. Nachr., 328, 852
Cunha C. E., Lima M., Oyaizu H., Frieman J., Lin H., 2009, MNRAS, 396, 2379
Dressler A., 1980, ApJ, 236, 351
Efstathiou G., Moody S. J., 2001, MNRAS, 325, 1603
Fry J. N., 1984, ApJ, 277, L5
Goldberg D. M., Vogeley M. S., 2004, ApJ, 605, 1
Gottlöber S., Łokas E. L., Klypin A., Hoffman Y., 2003, MNRAS, 344, 715
Hildebrandt H. et al., 2010, A&A, 523, A31
Huchra J. P., Geller M. J., 1982, ApJ, 257, 423
Icke V., 1984, MNRAS, 206, 1p
Icke V., van de Weygaert R., 1991, Q. J. R. Astron. Soc., 32, 85
Jasche J., Wandelt B. D., 2012, MNRAS, 425, 1042
Jones B. J. T., van de Weygaert R., Aragón-Calvo M. A., 2010, MNRAS, 408, 897
Kaiser N., 1984, ApJ, 284, L9
Kaiser N., 1987, MNRAS, 227, 1
Kirshner R. P., Oemler A., Jr, Schechter P. L., Shectman S. A., 1981, ApJ, 248, L57
Kovač K. et al., 2010, ApJ, 708, 505
Ménard B., Scranton R., Schmidt S., Morrison C., Jeong D., Budavari T., Rahman M., 2013, preprint (arXiv:1303.4722)
Mo H. J., White S. D. M., 1996, MNRAS, 282, 347
Mobasher B. et al., 2007, ApJS, 172, 117
Newman J. A., 2008, ApJ, 684, 88
Neyrinck M. C., Szapudi I., Szalay A. S., 2009, ApJ, 698, L90
Neyrinck M. C., Aragón-Calvo M. A., Jeong D., Wang X., 2014, MNRAS, 441, 646
Nguyen H., 2007, GPU Gems 3, 1st edn. Addison-Wesley, Reading, MA
Platen E., van de Weygaert R., Jones B. J. T., 2007, MNRAS, 380, 551
Platen E., van de Weygaert R., Jones B. J. T., 2008, MNRAS, 387, 128
Platen E., van de Weygaert R., Jones B. J. T., Vegter G., Calvo M. A. A., 2011, MNRAS, 416, 2494
Rahman M., Ménard B., Scranton R., Schmidt S. J., Morrison C. B., 2015, MNRAS, 447, 3500
Regos E., Geller M. J., 1991, ApJ, 377, 14
Schaap W. E., van de Weygaert R., 2000, A&A, 363, L29
Schneider M., Knox L., Zhan H., Connolly A., 2006, ApJ, 651, 14
Sheth R. K., van de Weygaert R., 2004, MNRAS, 350, 517
Steinberg E., Yalinewich A., Sari R., Duffell P., 2015, ApJS, 216, 14
Szapudi I., Szalay A. S., Boschan P., 1991, Traces of the Primordial Structure in the Universe. Max-Planck-Institut für Extraterrestrische Physik, Garching bei München, p. 39
Tegmark M. et al., 2004, ApJ, 606, 702
van de Weygaert R., 1991, MNRAS, 249, 159
van de Weygaert R., 1994, A&A, 283, 361
van de Weygaert R., 2002, in Plionis M., Cotsakis S., eds, Astrophysics and Space Science Library, Vol. 276, Modern Theoretical and Observational Cosmology. Kluwer, Dordrecht, p. 119
van de Weygaert R., 2003, in Feigelson E. D., Babu G. J., eds, The Cosmic Foam: Stochastic Geometry and Spatial Clustering Across the Universe. p. 175
van de Weygaert R., Schaap W., 2009, in Martínez V. J., Saar E., Martínez-González E., Pons-Bordería M.-J., eds, Lecture Notes in Physics, Vol. 665, Data Analysis in Cosmology. Springer-Verlag, Berlin, p. 291
van Haarlem M., van de Weygaert R., 1993, ApJ, 418, 544
York D. G. et al., 2000, A&A, 120, 1579
Zel'dovich Y. B., 1970, A&A, 5, 84
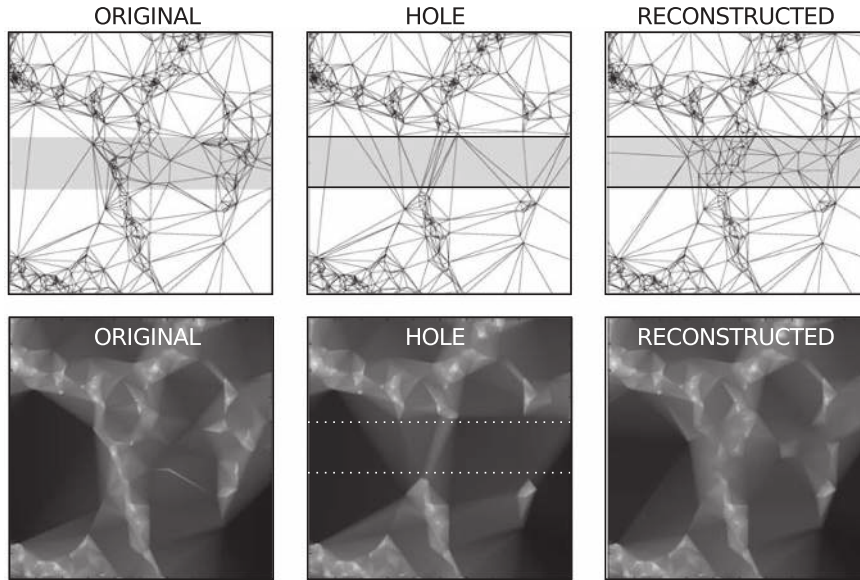Zeldovich I. B., Einasto J., Shandarin S. F., 1982, Nature, 300, 407

Figure A1. Delaunay-based structure reconstruction. The left-hand panels show the Delaunay tessellation (top) and reconstructed density field (bottom) corresponding to the biased particle distribution in an *N*-body simulation. Central panels show the corresponding Delaunay tessellation and density field after a hole has been cut from the original particle distribution. The hole is indicated by the horizontal shaded area crossing the top panels. By sampling the density field inside the hole with new particles matching the mean number density outside the hole we are able to reconstruct to a remarkable degree the original density field.

APPENDIX A: DELAUNAY-BASED STRUCTURE RECONSTRUCTION

The Delaunay tessellation not only is able to compute density estimates and interpolate their value at any point but it can also interpolate anisotropic features in the galaxy distribution. Fig. A1 shows the reconstruction of the (2D) density field from a particle distribution before and after particles have been removed to form a hole. The Delaunay tessellation interpolates the space in-between the hole joining and reconstructing the filament previously truncated by the hole. We exploit this property of the Delaunay tessellation to interpolate the space between galaxies in sparsely sampled spectroscopic surveys and reconstruct the underlying density field. It is important to note that the Delaunay tessellation interpolates the density field in a linear fashion. This limits its application to linear features. However, as long as the spectroscopic sampling used to construct the tessellation can sample the features we want to reconstruct, i.e. if its mean intergalaxy separation is smaller than the typical scale of the cosmic structures, then the linear interpolation is a good approximation.

APPENDIX B: SQL QUERY

The spectroscopic and photometric redshifts were obtained using the SDSS CASJOBS service with the following query:

```
SELECT s.ra, s.dec,s.z, p.z,p.zErr,
 p.absMagI,p.absMagR,p.absMagG,p.absMagU
FROM Photoz AS p
JOIN SpecPhoto as s
ON p.ObjID = s.ObjID
WHERE
 p.z BETWEEN 0 and 0.15
 AND p.nnIsInside = 1
 AND p.nnCount > 95
 AND p.zErr BETWEEN 0 AND 0.02
```

From this master catalogue we generated subsamples as needed. The photometric redshift catalogue from Csabai et al. (2007) was queried in the same way as above but without performing the JOIN with the spectroscopic sample. The spectroscopic galaxy catalogue used to generate density fields was produced using the following query:

```
SELECT p.ObjID, p.ra, p.dec, s.z, s.zErr,
s.zConf,
  s.cx, s.cy, s.cz,
  p.dered_u, p.dered_g, p.dered_r, ...
  p.isoPhi_u, p.isoPhi_g, p.isoPhi_r, ...
  p.isoA_u, p.isoB_u, p.isoA_g, p.isoB_g, ...
  p.petroR50_u, p.petroR90_u, p.petroR50_g, ...
  p.petroMag_u,p.petroMag_g,p.petroMag_r,...
  p.petroRad_u,p.petroRad_g,p.petroRad_r,...
  s.specclass,
  s.eClass
INTO myDB.galaxies_dr9
FROM SpecObj AS s,
  PhotoObj AS p
WHERE s.SpecObjID = p.SpecObjID
```

APPENDIX C: SELECTION FUNCTION WEIGHTING

In order to take full advantage of all the galaxies in the sample, we used a magnitude-limited catalogue. This means that the radial distribution of galaxies must be weighted in order to account for the decrease in the number density of galaxies with increasing redshift and produce an isotropic distribution. We used the formula provided by Efstathiou & Moody (2001) to model the change in the mean number of galaxies as a function of their redshift as

$$N(z)\, \mathrm{d}z = A\, z^2\, \psi(z)\mathrm{d}z, \tag{C1}$$
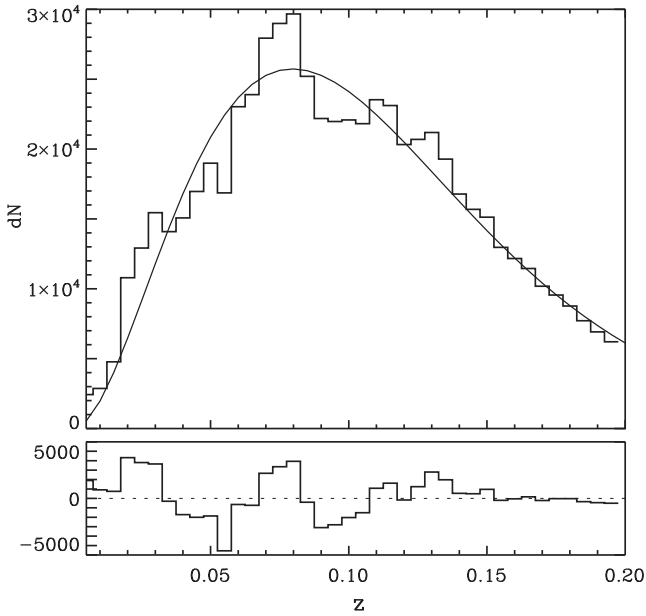
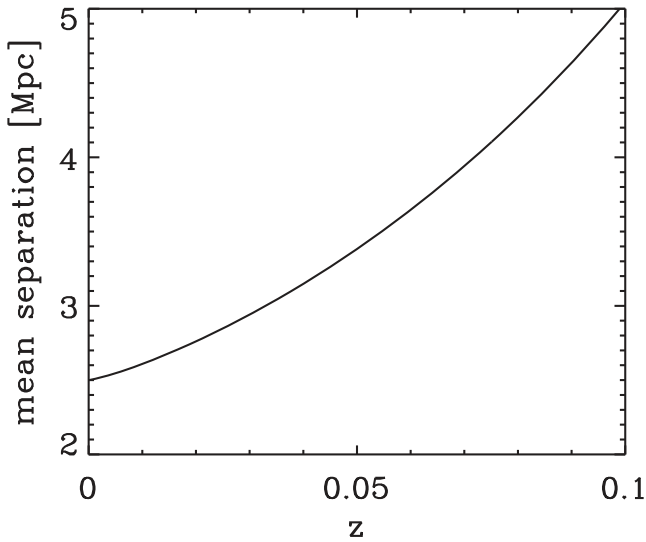**Figure C1.** Distribution of galaxies with redshift.

Figure C2. Mean intergalaxy separation as function of redshift.

where A is a normalization factor that depends on the density of galaxies, and $\psi(z)$ is the *selection function*:

$$\psi(z) = e^{-(z/z_r)^{\beta}}, \tag{C2}$$

where z_r is the characteristic redshift of the distribution and β encodes the slope of the curve. Fig. C1 shows the distribution of galaxies in the SDSS with redshift and the best fit to equation (C1).

C0.1 Mean intergalaxy separation

Fig. C2 shows the mean intergalaxy separation as a function of redshift. This plot does not account for small-scale clustering and so the values are average over the volume. We can estimate the real intergalaxy distance by considering that galaxies are contained in approximately 10 per centof the volume, the line in Fig. C2 must

be divided by $\sqrt{10} \sim 3$ giving a mean intergalaxy separation of 0.8–1.6 Mpc on the redshift range $z = 0$–0.1 .

APPENDIX D: FOG COMPRESSION

The peculiar velocities in dense non-linear regions such as clusters produce the so-called FoG and are an important source of contamination for both the density estimation and the cosmic web identification. We use a FoG compression algorithm following the approach of Tegmark et al. (2004) and identify elongated groups using a Friends of Friends group finder where galaxies are recursively linked to other galaxies within a linking volume defined by the linking parameter b:

$$b = \left(\frac{4}{3}\pi \, \bar{n}_{\text{gal}} \, r_{\text{link}}^3 \right)^{1/3}, \tag{D1}$$

where $\bar{n}_{\text{gal}}$ is the mean number density of galaxies and r_{link}^3 is the linking length. In redshift space, peculiar velocities in clusters give them an elongated shape along the LoS. Following Huchra & Geller (1982), we decompose the distance between galaxies i and j into LoS distance:

$$D_{\parallel,i,j} = (c/H_0)(z_i + z_j)\sin(\theta_{i,j}/2) \tag{D2}$$

and transversal distance:

$$D_{\perp,i,j} = (c/H_0) \mid z_i + z_j \mid, \tag{D3}$$

where z_i, z_j are the redshifts to each galaxy and $\theta_{i,j}$ is their angular separation. Two galaxies are linked if the two following conditions are satisfied:

$$D_{\parallel,i,j} \leq b_{\parallel}\bar{n}_{\text{gal}}(z)^{-1/3} \tag{D4}$$

and

$$D_{\perp,i,j} \leq b_{\perp}\bar{n}_{\text{gal}}(z)^{-1/3}, \tag{D5}$$

where $\bar{n}_{\text{gal}}(z)$ is the mean number density of galaxies as function of redshift and $b_{\parallel}$, $b_{\perp}$ are the projected linking lengths. A typical relation between LoS and transverse linking parameters is $b_{\parallel} = 8\,b_{\perp}$ (Tegmark et al. 2004; Berlind et al. 2006). The elongated groups are then *compressed* such that their dispersion along the LoS is equal to the dispersion in the plane of the sky. Fig. C3 shows the effect of the FoG compression on the galaxy distribution and on the underlying density field. One drawback of the FoG compression is that systematically compresses filaments oriented along the LoS (Jones et al. 2010).

APPENDIX E: EFFICIENT WATERSHED-BASED DISTANCE TRANSFORM COMPUTATION

There are several algorithms for computing distance transforms using the sampling grid itself as a measure of metric. However, for an Euclidean metric this operation is $O(N^2)$, which makes computing Euclidean distance transforms in large 3D grids computationally expensive and even prohibitive. However, we can take advantage of the space-partitioning property of the watershed transform and compute distance transforms inside each watershed region independently (see also Steinberg et al. 2015 for a similar application of the Voronoi tessellation for N-body computations). By doing so we reduce the number of computations from $O(N^2)$ to $O((N/M)^2)$ where M is the number of watershed regions in the volume. For cosmological volumes, such as the SDSS survey we analyse here,
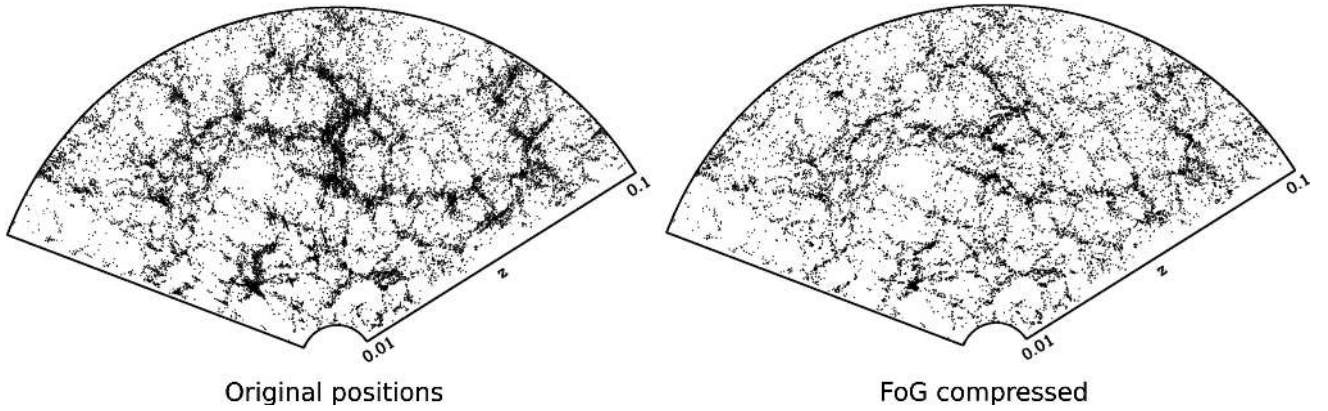
Figure C3. FoG compression. FoG removal in a slice of 6 degrees of thickness in declination. From the original galaxy distribution we extract the FoG (middle left) and compress them on the LoS direction (middle right). The final distribution without FoGs is shown in the lower slice.

the large number of watershed regions (voids) makes the brute-force computation of Euclidean distances feasible. Since each watershed region can be treated independently this optimization is trivial to parallelize. In addition to the optimization described above, the brute-force distance transform computation is done on the GPU using a CUDA code adapted from the GPU-enabled gravitational code described in Nguyen (2007). If the number of voxels inside a given

watershed region exceeds the available memory in the GPU we split the watershed region into smaller blocks and process them independently. Finally, the partial computations are combined to find the minimum distances required in equation (5).

This paper has been typeset from a TeX/LaTeX file prepared by the author.