Part IV

Cosmic web analysis

Chapter 8

Dark matter voids in the SDSS galaxy survey

Contents

> "Lost and forgotten be, darker than the darkness,
> Where gates stand for ever shut, till the world is mended."
> — John Ronald Reuel Tolkien (1954), *The Fellowship of the Ring*

Abstract

What do we know about voids in the dark matter distribution given the Sloan Digital Sky Survey and assuming the ΛCDM model? In chapter 5, application of the Bayesian inference algorithm BORG to the SDSS Data Release 7 main galaxy sample has generated detailed Eulerian and Lagrangian representations of the large-scale structure as well as the possibility to accurately quantify corresponding uncertainties. Building upon these results, we present constrained catalogs of voids in the Sloan volume, aiming at a physical representation of dark matter underdensities and at the alleviation of the problems due to sparsity and biasing on galaxy void catalogs. To do so, we generate data-constrained reconstructions of the presently observed large-scale structure using a fully non-linear gravitational model. We then find and analyze void candidates using the VIDE toolkit. Our methodology therefore predicts the properties of voids based on fusing prior information from simulations and data constraints. For usual void statistics (number function, ellipticity distribution and radial density profile), all the results obtained are in agreement with dark matter simulations. Our dark matter void candidates probe a deeper void hierarchy than voids directly based on the observed galaxies alone. The use of our catalogs therefore opens the way to high-precision void cosmology at the level of the dark matter field. We have made the void catalogs used in this work available at http://www.cosmicvoids.net.

This chapter is adapted from its corresponding publication, Leclercq *et al.* (2015).

8.1 Introduction

Observations of the cosmic large-scale structure have revealed that galaxies tend to lie in thin wall-like structures surrounding large underdense regions known as voids, which constitute most of the volume of the Universe. Although the discovery of cosmic voids dates back to some of the first galaxy redshift surveys

(Gregory & Thompson, 1978; Kirshner *et al.*, 1981; de Lapparent, Geller & Huchra, 1986) and their significance was assessed in some early studies (Martel & Wasserman, 1990; van de Weygaert & van Kampen, 1993; Goldberg & Vogeley, 2004), the systematic analysis of void properties has only been considered seriously as a source of cosmological information in the last decade (e.g. Sheth & van de Weygaert, 2004; Colberg *et al.*, 2005; Viel, Colberg & Kim, 2008; Betancort-Rijo *et al.*, 2009; Lavaux & Wandelt, 2010; Biswas, Alizadeh & Wandelt, 2010; van de Weygaert & Platen, 2011; Lavaux & Wandelt, 2012, and references therein). Like overdense tracers of the density field such as clusters, voids can be studied by statistical methods in order to learn about their distribution and properties compared to theoretical predictions.

Generally, direct sensitivity of void statistics to cosmology is only guaranteed for the underdense regions of the overall matter density field, which includes a large fraction of dark matter. These are the physical voids in the LSS, for which theoretical modeling is established. However, absent direct measurements of dark matter underdensities, current void catalogs are defined using the locations of galaxies in large redshift surveys (Pan *et al.*, 2012; Sutter *et al.*, 2012b, 2014d; Nadathur & Hotchkiss, 2014). Since galaxies trace the underlying mass distribution only sparsely, void catalogs are subject to uncertainty and noise. Additionally, numerical simulations show that there exists a population of particles in cosmic voids. This is an indication of physical biasing in galaxy formation: there is primordial dark and baryonic matter in voids, but due to the low density, little galaxy formation takes place there. Additionally, due to complex baryonic physics effects during their formation and evolution, galaxies are biased tracers of the underlying density field, which gives rise to qualitatively different void properties.

The sensitivity of void properties to the sampling density and biasing of the tracers has only been recently analyzed in depth on simulations, by using synthetic models to mimic realistic surveys. Little & Weinberg (1994); Benson *et al.* (2003); Tinker & Conroy (2009); Sutter *et al.* (2014c) found that the statistical properties of voids in galaxy surveys are not the same as those in dark matter distributions. At lower tracer density, small voids disappear and the remaining voids are larger and more spherical. Their density profiles get slightly steeper, with a considerable increase of their compensation scale, which potentially may serve as a static ruler to probe the expansion history of the Universe (Hamaus *et al.*, 2014b). Hamaus, Sutter & Wandelt (2014) recently proposed a universal formula for the density profiles of voids, describing in particular dark matter voids in simulations (see also Colberg *et al.*, 2005; Paz *et al.*, 2013; Ricciardelli, Quilis & Varela, 2014; Nadathur *et al.*, 2015). The connection between galaxy voids and dark matter voids on a one-by-one basis is difficult due to the complex internal hierarchical structure of voids (Dubinski *et al.*, 1993; van de Weygaert & van Kampen, 1993; Sahni, Sathyaprakah & Shandarin, 1994; Sheth & van de Weygaert, 2004; Aragon-Calvo & Szalay, 2013; Sutter *et al.*, 2014d,b). However, the nature of this relationship determines the link between a survey, with its particular tracer density, and the portion of the cosmic web that it represents. Understanding this connection is of particular importance in light of recent results that probe the LSS via its effect on photons geodesics. These results include Melchior *et al.* (2014); Clampitt & Jain (2015), which probe the dark matter distribution via weak gravitational lensing; Ilić, Langer & Douspis (2013); Planck Collaboration (2014a) for the detection of the integrated Sachs-Wolfe effect in the cosmic microwave background, sensitive to the properties of dark energy. As a response to this demand, Sutter *et al.* (2014b) found that voids in galaxy surveys always correspond to underdensities in the dark matter, but that their centers may be offset and their size can differ, in particular in sparsely sampled surveys where void edges suffer fragmentation.

While previous authors offer broad prescriptions to assess the effects of sparsity and biasing of the tracers on voids, the connection between galaxy voids of a particular survey and dark matter underdensities remains complex. In particular, disentangling these effects from cosmological signals in presence of the uncertainty inherent to any cosmological observation (selection effects, survey mask, noise, cosmic variance) remains an open question. In this work, we propose a method designed to circumvent the issues due to the conjugate and intricate effects of sparsity and biasing on galaxy void catalogs. In doing so, we will show that voids in the dark matter distribution can be constrained by the *ab initio* analysis of surveys of tracers, such as galaxies. We will demonstrate the feasibility of our method and obtain catalogs of dark matter voids candidates in the Sloan Digital Sky Survey Data Release 7.

Our method is based on the identification of voids in the dark matter distribution inferred from large-scale structure surveys. The constitution of such maps from galaxy positions, also known as "reconstruction", is a field in which Bayesian methods have led to enormous progress over the last few years. Initial approaches typically relied on approximations such as a multivariate Gaussian or log-normal distribution for density fields, with a prescription for the power spectrum to account for the correct two-point statistics (Lahav *et al.*, 1994; Zaroubi, 2002; Erdoğdu *et al.*, 2004; Kitaura & Enßlin, 2008; Kitaura *et al.*, 2009; Kitaura, Jasche & Metcalf,

2010; Jasche & Kitaura, 2010; Jasche *et al.*, 2010b,a). However, due to their potentially complex shapes, proper identification of structures such as voids requires reconstructions correct not only at the level of the power spectrum, but also higher-order correlators. Inferences of this kind from observational data have only been made possible very recently by the introduction of physical models of structure formation in the likelihood. This naturally moves the problem to the inference of the initial conditions from which the large-scale structure originates (Jasche & Wandelt, 2013a; Kitaura, 2013; Wang *et al.*, 2013).

This work exploits the recent application of the BORG (Bayesian Origin Reconstruction from Galaxies, Jasche & Wandelt (2013a), see chapter 4) algorithm to the Sloan Digital Sky Survey galaxies (Jasche, Leclercq & Wandelt, 2015, see chapter 5), and on the subsequent generation of constrained non-linear realizations of the present large-scale distribution of dark matter. BORG is a full-scale Bayesian framework, permitting the four-dimensional physical inference of density fields in the linear and mildly non-linear regime, evolving gravitationally from the initial conditions to the presently observed large-scale structure. By exploring a highly non-linear and non-Gaussian LSS posterior distribution via efficient Markov Chain Monte Carlo methods, it also provides naturally and fully self-consistently accurate uncertainty quantification for all derived quantities. A straightforward use of reconstructed initial conditions is to resimulate the considered volume (Lavaux, 2010; Kitaura, 2013; Heß, Kitaura & Gottlöber, 2013). In the same spirit, building upon the inference of the initial conditions by BORG, one can generate a set of data-constrained realizations of the present large-scale structure via full N-body dynamics. As we will show, we make use of initial conditions reconstructed by BORG without any further post-processing, which demonstrates the high quality of inference results.

Due to the limited number of phase-space foldings, the influence of non-linearity in cosmic voids is expected to be milder as compared to galaxies and dark matter halos (Neyrinck, 2012; Neyrinck & Yang, 2013; Leclercq *et al.*, 2013, see also Abel, Hahn & Kaehler, 2012; Falck, Neyrinck & Szalay, 2012; Shandarin, Habib & Heitmann, 2012). For this reason, voids are more closely related to the initial conditions of the Universe, which makes them the ideal laboratories for physical application of Bayesian inference with BORG. In this work, we apply the void finder algorithm VIDE (Sutter *et al.*, 2015b), based on ZOBOV (Neyrinck, 2008), to data-constrained, non-linear reconstructions of the LSS. Each of them is a full physical realization of densely-sampled particles tracing the dark matter density field. In this fashion, we construct catalogs of dark matter voids in the SDSS volume robust to sparsity and biasing of galaxies. As we will show, this procedure drastically reduces statistical uncertainty in void catalogs. Additionally, the use of data-constrained reconstructions allows us to extrapolate the void identification in existing data (e.g. at very small or at the largest scales, at high redshift or near the survey boundary).

As described in chapters 4 and 5, (see Jasche & Wandelt, 2013a; Jasche, Leclercq & Wandelt, 2015), the BORG inference framework possesses a high degree of control on observational systematic and statistical uncertainties such as noise, survey geometry and selection effects. Uncertainty quantification is provided via efficient sampling of the corresponding LSS posterior distribution. The resultant set of initial and final density field realizations yields a numerical representation of the full posterior distribution, capturing all data constraints and observational uncertainties. Building upon these results, in this work, we will extend our Bayesian reasoning to void catalogs. Specifically, we apply full non-linear N-body dynamics to a set of data-constrained initial conditions to arrive at a set of non-linear dark matter density fields at the present epoch. As a result, we obtain a probabilistic description of non-linear density fields constrained by SDSS observations. Applying the VIDE void finder to this set of reconstructions yields N data-constrained realizations of the catalog, representing the posterior probability distribution for dark matter voids given observations. In this fashion, we have fully Bayesian access to uncertainty quantification via the variation between different realizations. In particular, we are now able to devise improved estimators for any void statistics by the use of Blackwell-Rao estimators. To assess the robustness of this technique for cosmological application, we focus on three key void observables: number functions, ellipticity distributions and radial density profiles. These are especially sensitive probes of non-standard cosmologies (Bos *et al.*, 2012) and are well understood in both data and simulations (e.g Sutter *et al.*, 2014d).

As a general matter, we stress that these data-constrained realizations of dark matter void catalogs were obtained assuming a ΛCDM prior. Using our products for model testing therefore requires care: in the absence of data constraints, one will simply be dealing with realizations of the ΛCDM prior. Consequently, any departure from unconstrained ΛCDM predictions are driven by the data. Conversely, for model tests where the data are not strongly informative, agreement with ΛCDM is the default answer.

This chapter is organized as follows. In section 8.2, we describe our methodology: Bayesian inference with the BORG algorithm, non-linear filtering of the results, void identification technique and Blackwell-Rao estimators

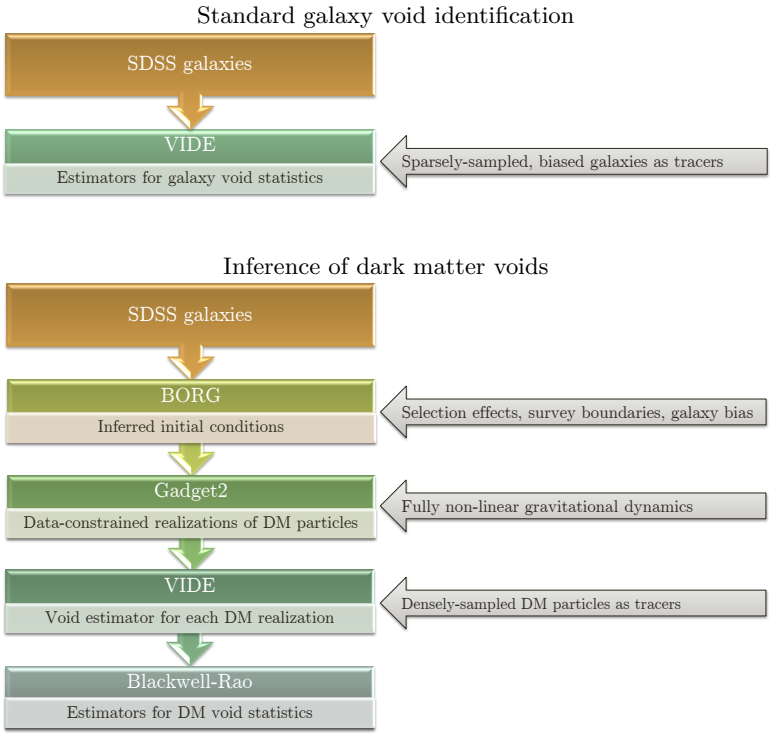

Figure 8.1: Schematic representation of our methodology for the inference of dark matter voids (lower panel) in comparison to the standard approach for the identification of galaxy voids (upper panel).

for void statistics. In section 8.3, we examine the properties of the dark matter voids in our catalogs. Finally, in section 8.4 we summarize our results, discuss perspectives for existing and upcoming galaxy surveys and offer concluding comments.

8.2 Methodology

In this section, we describe our methodology step by step:

1. inference of the initial conditions with BORG (section 8.2.1),
2. generation of data-constrained realizations of the SDSS volume (section 8.2.2),
3. void finding and processing (section 8.2.3),
4. combination of different void catalogs with Blackwell-Rao estimators (section 8.2.4).

In section 8.2.5, we describe the void catalogs used as references for comparison with our results. These are galaxy void catalogs directly based on SDSS galaxies without use of our methodology, and catalogs of voids in dark matter simulations.

A schematic representation of our procedure is represented in figure 8.1, in comparison to the standard approach of finding voids using galaxies as tracers.

8.2.1 Bayesian large-scale structure inference with the BORG algorithm

This works builds upon previous results, obtained by the application of BORG (Bayesian Origin Reconstruction from Galaxies, Jasche & Wandelt, 2013a) to SDSS main galaxy data (Jasche, Leclercq & Wandelt, 2015). In the rest of this section, we summarize its most stringent features; the reader is referred to chapters 4 and 5 for all details.

The BORG algorithm is a fully probabilistic inference machinery aiming at the analysis of linear and mildly-non-linear density and velocity fields in galaxy observations. It incorporates a physical model of cosmological

structure formation, which translates the traditional task of reconstructing the non-linear three-dimensional density field into the task of inferring corresponding initial conditions from present cosmological observations. This approach yields a highly non-trivial Bayesian inference, requiring to explore very high-dimensional and non-linear spaces of possible solutions to the initial conditions problem from incomplete observations. Typically, these parameter spaces comprise on the order of 10^6 to 10^7 parameters, corresponding to the elements of the discretized observational domain.

Specifically, the BORG algorithm explores a posterior distribution consisting of a Gaussian prior, describing the statistical behavior of the initial density field at a cosmic scale factor of $a = 10^{-3}$, linked via second-order Lagrangian perturbation theory to a Poissonian model of galaxy formation at the present epoch (for details see Jasche & Wandelt, 2013a and Jasche, Leclercq & Wandelt, 2015). As pointed out by previous authors (see e.g. Moutarde *et al.*, 1991; Buchert, Melott & Weiß, 1994; Bouchet *et al.*, 1995; Scoccimarro, 2000; Bernardeau *et al.*, 2002; Scoccimarro & Sheth, 2002, and chapter 2), 2LPT describes the one-, two- and three-point statistics correctly and represents higher-order statistics very well. Consequently, the BORG algorithm naturally accounts for features of the cosmic web, such as filaments, that are typically associated to high-order statistics induced by non-linear gravitational structure formation processes.

Besides physical structure formation, the posterior distribution also accounts for survey geometry, selection effects and noise, inherent to any cosmological observation (see section 4.2). Corresponding full Bayesian uncertainty quantification is provided by exploring this highly non-Gaussian and non-linear posterior distribution via an efficient Hamiltonian Markov Chain Monte Carlo sampling algorithm (see Jasche & Wandelt, 2013a, and sections 3.4.3, 4.3.2, for details). In order to account for luminosity dependent galaxy bias (Jasche & Wandelt, 2013b) and to make use of automatic noise calibration, we further use modifications introduced to the original BORG algorithm by Jasche, Leclercq & Wandelt (2015) (see section 4.3.1).

In this work, we make use of the 12,000 samples of the posterior distribution generated by Jasche, Leclercq & Wandelt (2015), described in chapter 5, which constitute highly-detailed and accurate reconstructions of the initial and present-day density fields constrained by SDSS observations.

8.2.2 Generation of data-constrained reconstructions

Starting form 11 statistically independent initial conditions realizations from the BORG SDSS analysis, we generated a set of fully non-linear, constrained reconstructions of the LSS. This step is achieved via optimal filtering of BORG results with the GADGET-2 (Springel, Yoshida & White, 2001; Springel, 2005) cosmological code. For details on the non-linear filtering procedure, see chapter 7, in particular section 7.2 for the description of the set of realizations used in this chapter.

8.2.3 Void finding and processing

8.2.3.1 Void finding

We identify and post-process voids with the VIDE (Void IDentification and Examination) toolkit[1] (Sutter *et al.*, 2015b, also described in section C.1 of appendix C), which uses a highly modified version of ZOBOV (Neyrinck, 2008; Lavaux & Wandelt, 2012; Sutter *et al.*, 2012b) to create a Voronoi tessellation of the tracer particle population and the watershed transform to group Voronoi cells into zones and voids (Platen, van de Weygaert & Jones, 2007). The watershed transform identifies catchment basins as the cores of voids, and ridgelines, which separate the flow of water, as the boundaries of voids. It naturally builds a nested hierarchy of voids (Lavaux & Wandelt, 2012; Bos *et al.*, 2012). For the purposes of this work, we examine all voids regardless of their position in the hierarchy. The pipeline imposes a density-based threshold within the void finding operation: voids only include as additional members Voronoi zones if the minimum ridge density between that zone and the void is less than 0.2 times the mean particle density (Platen, van de Weygaert & Jones, 2007; see Blumenthal *et al.*, 1992; Sheth & van de Weygaert, 2004 for the role of the corresponding $\delta = -0.8$ underdensity). If a void consists of only a single zone (as they often do in sparse populations) then this restriction does not apply.

VIDE provides several useful definitions used in this work, such as the effective radius,

$$R_{\mathrm{v}} \equiv \left(\frac{3}{4\pi}V\right)^{1/3}, \tag{8.1}$$

[1] http://www.cosmicvoids.net

where V is the total volume of the Voronoi cells that contribute to the void. We use this radius definition to ignore voids with R_v below the mean particle spacing $\bar{n}^{-1/3}$ of the tracer population, as these are increasingly affected by Poisson fluctuations. VIDE also reports the volume-weighted center, or macrocenter, as

$$\mathbf{x}_\mathrm{v} \equiv \frac{1}{\sum_i V_i} \sum_i \mathbf{x}_i V_i, \tag{8.2}$$

where $\mathbf{x}_i$ and V_i are the positions and Voronoi volumes of each tracer particle i, respectively.

In each tracer population, the VIDE pipeline provides void estimators ; in particular, the three statistics we will focus on in section 8.3: number count, ellipticity distribution and radial density profile.

In figure 8.2, we show slices through different data-constrained realizations. The density of dark matter particles identified by VIDE as being part of a void is represented in gray scale. Note that, since ZOBOV essentially performs a division of space in different void regions with vanishingly-thin ridges, almost all particles initially present in the dark matter field are conserved. For clarity of the visualization, the quantity represented is $\ln(2+\delta)$ where δ is the density contrast of particles in voids. The SDSS galaxies used for the BORG analysis are overplotted as red dots. The core of dark matter voids (using a density threshold $\delta < -0.3$) is shown in color. As can be observed, dark matter voids also correspond to underdensities in the field traced by galaxies, which is in agreement with the results obtained by Sutter *et al.* (2014b) in simulations.

8.2.3.2 Selection of voids

The VIDE pipeline identifies all dark matter voids in the non-linear data-constrained realizations described in section 8.2.2. These live in boxes of 750 Mpc/h side length with periodic boundary conditions. In order to select physically meaningful dark matter void candidates, we have to select a subsample of voids which intersect the volume of the box actually constrained by SDSS galaxies.

As described in chapter 5, unobserved and observed regions in the inferred final density fields do not appear visually distinct, a consequence of the excellent performance of the 2LPT model implemented in BORG as a physical description of structure formation. In addition, due to the non-local transport of observational information between initial and final conditions, the region influenced by data extends beyond the survey boundaries and the large-scale structure appears continuous there. The fact that data constraints can radiate out of the survey volume has been known since the first constrained reconstructions of the mass distribution (Bertschinger, 1987; Hoffman & Ribak, 1991; van de Weygaert & Bertschinger, 1996), where a power spectrum prior was assumed to sample constrained Gaussian random fields. Here, as detailed in chapters 4 and 5, constraints are propagated by the structure formation model assumed in the inference process (2LPT), which accounts not only for two-point statistics, but for the full hierarchy of correlators, in its regime of validity. Therefore, dark matter voids candidates intersecting the survey boundaries can be considered as physical if a significant fraction of their volume is influenced by the data.

The survey response operator R is a voxel-wise function representing simultaneously the survey geometry (observed and unobserved regions) and the selection effects in galaxy catalogs. Here, we kept for R the average over the six luminosity bins used in the BORG SDSS run (for details see chapter 5). For the purpose of this work, we keep all void candidates whose center is in a region where R is strictly positive. This region represents 7.9×10^7 cubic Mpc/h, around 18.7% of the full box. In each of the 11 realizations used in this work, we kept around 166,000 data-constrained voids out of 886,000 voids in the entire box.

In figure 8.2, the survey response operator is shown in color from purple (totally unobserved region) to blue (region fully constrained by the data). One can see the correct propagation of information operated by BORG, as voids appear continuous at the survey boundaries.

8.2.4 Blackwell-Rao estimators for dark matter void realizations

A particular advantage of our Bayesian methodology is the ability to provide accurate uncertainty quantification for derived dark matter void properties. In particular, the Markovian samples described in chapter 5 permit us to employ a Blackwell-Rao estimator to describe the posterior distribution for inferred dark matter voids. Specifically, we are interested in deriving the posterior distribution $\mathcal{P}(x|d)$ of a dark matter void property x given observations d. Using the realizations of the initial conditions δ^i and the dark matter void realizations V generated by the approach described in sections 8.2.2 and 8.2.3, we obtain

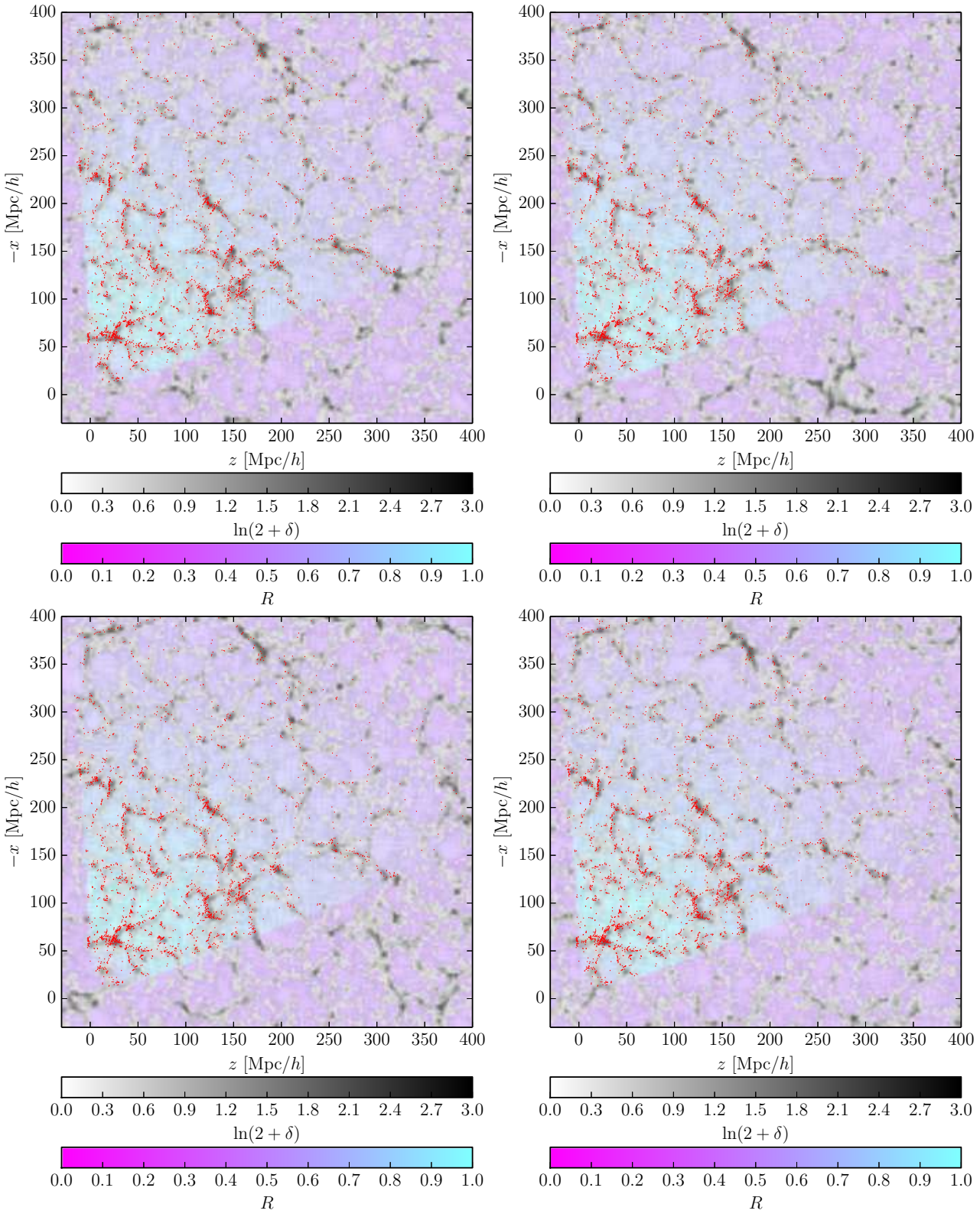

Figure 8.2: Slices through different data-constrained realizations used to build samples of the dark matter void catalog. The SDSS galaxies used for the inference with BORG are represented as red dots. The density of dark matter particles identified by VIDE as being part of a void is shown in gray scale. In color, we show the particles that live in the core of dark matter voids (in a density environment smaller than -0.3 times the average density). The survey response operator R shows how well the results are constrained by the data (see text for details). In the observed region, the data are strongly informative about the cosmic web in general and voids in particular; the reconstructions are not prior-dominated.

$$\begin{aligned}
\mathcal{P}(x|d) &= \int \mathcal{P}(x|V)\,\mathcal{P}(V,\delta^{\mathrm{i}}|d)\,\mathrm{d}V\,\mathrm{d}\delta^{\mathrm{i}} \\
&= \int \mathcal{P}(x|V)\,\mathcal{P}(V|\delta^{\mathrm{i}},d)\,\mathcal{P}(\delta^{\mathrm{i}}|d)\,\mathrm{d}V\,\mathrm{d}\delta^{\mathrm{i}} \\
&= \int \mathcal{P}(x|V)\,\delta_{\mathrm{D}}(V-\tilde{V}(\delta^{\mathrm{i}}))\,\mathcal{P}(\delta^{\mathrm{i}}|d)\,\mathrm{d}V\,\mathrm{d}\delta^{\mathrm{i}} \\
&= \int \mathcal{P}(x|\tilde{V}(\delta^{\mathrm{i}}))\,\mathcal{P}(\delta^{\mathrm{i}}|d)\,\mathrm{d}\delta^{\mathrm{i}} \\
&\approx \frac{1}{N}\sum_k \mathcal{P}\left(x|\tilde{V}(\delta^{\mathrm{i}}_k)\right) \\
&= \frac{1}{N}\sum_k \mathcal{P}\left(x|V_k\right)\,, \qquad (8.3)
\end{aligned}$$

where we assumed the dark matter void templates V to be conditionally independent of the data d given the initial conditions δ^{i}, and to derive uniquely from the initial density field via the procedure described in sections 8.2.2 and 8.2.3, yielding $\mathcal{P}(V|\delta^{\mathrm{i}},d) = \mathcal{P}(V|\delta^{\mathrm{i}}) = \delta_{\mathrm{D}}(V-\tilde{V}(\delta^{\mathrm{i}}))$. We also exploited the fact that we have a sampled representation of the initial conditions posterior distribution $\mathcal{P}(\delta^{\mathrm{i}}|d) \approx 1/N\sum_k \delta_{\mathrm{D}}(\delta^{\mathrm{i}}-\delta^{\mathrm{i}}_k)$, where k labels one of the N samples. The last line of equation (8.3) represents the Blackwell-Rao estimator for void property x to be inferred from our dark matter void catalogs V_k, providing thorough Bayesian means to quantify uncertainties. It consists of a mixture distribution over different realizations of dark matter void templates.

The VIDE pipeline provides estimated means and variances for derived quantities x, allowing us to model the distributions $\mathcal{P}(x|V_k)$ as Gaussians with mean x_k and variance σ_k^2, for respective dark matter void templates. The final expression for the posterior distribution of x given the data is therefore

$$\mathcal{P}(x|d) \approx \frac{1}{N}\sum_k \frac{1}{\sqrt{2\pi\sigma_k^2}}\,\exp\left(-\frac{1}{2}\frac{(x-x_k)^2}{\sigma_k^2}\right). \qquad (8.4)$$

Even though we have access to non-Gaussian uncertainty quantification via the posterior distribution given in equation (8.4), for the presentation in this chapter we will be content with estimating means and variances. The mean for x given d is

$$\langle x|d\rangle \approx \frac{1}{N}\sum_k x_k, \qquad (8.5)$$

and the variance is

$$\langle (x-\langle x\rangle)^2\,|d\rangle \approx \frac{1}{N}\sum_k (x_k^2+\sigma_k^2) - \langle x|d\rangle^2. \qquad (8.6)$$

As described in section 8.2.3.2, we select voids in the data-constrained regions of reconstructions of the dark matter density field. Since these regions are the same in different reconstructions, the different void catalogs describe the same region of the actual Universe. For this reason, while estimating uncertainties, it is not possible to simply use all the voids in our catalogs as if they were independent.[2] However, using an increasing number of reconstructions, we shall still see a decrease of statistical uncertainty. Indeed, from (8.5) and (8.6) it follows that

$$\langle (x-\langle x\rangle)^2\,|d\rangle \leq \frac{1}{N}\sum_k \sigma_k^2, \qquad (8.7)$$

which means that the combination of different realizations will generally yield an improved estimator for any original statistics.

Note that this procedure is completely general and applies to any estimator provided by the VIDE pipeline.

8.2.5 Void catalogs for comparison of our results

In section 8.3, we will compare our results for dark matter voids to state-of-the-art results for galaxy voids. To do so, we use the catalogs of Sutter *et al.* (2012b) based on the SDSS DR7 galaxies, publicly available at

[2] We generally recommend special care for proper statistical treatment while working with the data-constrained realizations of our dark matter void catalog, especially if one wants to use frequentist estimators of void properties.

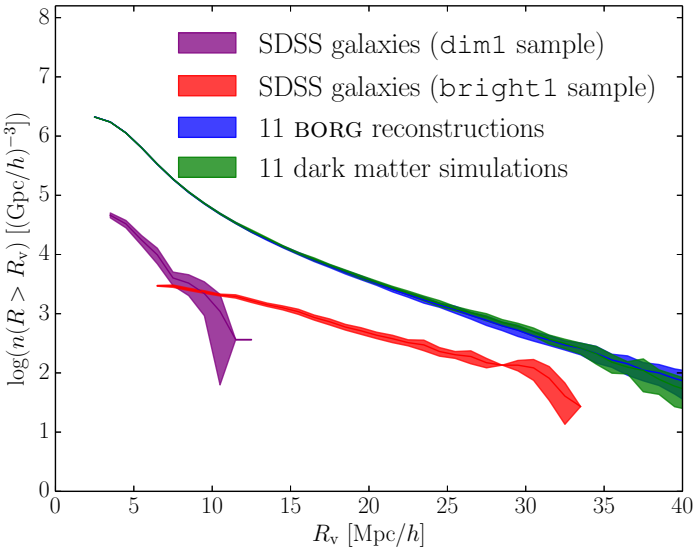

Figure 8.3: Cumulative void number functions. The results from 11 BORG reconstructions (blue) are compared to a dark matter N-body simulation (green) and to the galaxy voids directly found in two volume-limited sub-samples of the SDSS DR7 (`dim1`, purple and `bright1`, red). The solid lines are the measured or predicted number functions and the shaded regions are the 2-σ Poisson uncertainties. Fewer voids are found in observations than in dark matter simulations, due to the sparsity and bias of tracers, as well as observational uncertainty coming from the survey geometry and selection effects. Number functions from BORG reconstructions agree with simulations at all scales.

http://www.cosmicvoids.net. In particular, we compare to the voids found in the `bright1` and `dim1` volume-limited galaxy catalogs, for which the mean galaxy separations are 8 and 3 Mpc/h, respectively (for details, see Sutter *et al.*, 2012b).

Assessment of our results for dark matter voids in SDSS data also require systematic comparison to dark matter voids found in cosmological simulations. We ran 11 such unconstrained simulations with the same setup as described in section 8.2.2 for the generation of data-constrained realizations. We started from Gaussian random fields with an Eisenstein & Hu (1998, 1999) power spectrum using the fiducial cosmological parameters of the BORG analysis ($\Omega_\mathrm{m} = 0.272$, $\Omega_\Lambda = 0.728$, $\Omega_\mathrm{b} = 0.045$, $h = 0.702$, $\sigma_8 = 0.807$, $n_\mathrm{s} = 0.961$, see chapter 5). These initial density fields, defined in a 750 Mpc/h cubic box of 256^3 voxels, are occupied by a Lagrangian lattice of 512^3 dark matter particles. These are evolved to $z = 69$ with 2LPT and from $z = 69$ to $z = 0$ with GADGET-2. As for constrained realizations, in our simulations we selected the voids located inside the observed SDSS volume (see section 8.2.3.2) and combined properties using Blackwell-Rao estimators (see section 8.2.4).

8.3 Properties of dark matter voids

In this section, we describe the statistical properties of the dark matter voids found in the data-constrained parts of our reconstructions of the SDSS volume. We focus on three key statistical summaries abundantly described in the literature: number count, ellipticity distribution and radial density profiles.

8.3.1 Number function

The number function of voids provides a simple, easily accessible, and surprisingly sensitive cosmological probe. For example, the number function has been shown to respond to coupled dark matter-dark energy (Li & Zhao, 2009; Sutter *et al.*, 2015a), modified gravity (Li, Zhao & Koyama, 2012; Clampitt, Cai & Li, 2013), and variations in fundamental cosmological parameters (Pisani *et al.*, 2015). While most studies of the number function take place in N-body simulations, there has also been significant theoretical and analytical work, beginning with the excursion set formulation of Sheth & van de Weygaert (2004) and continuing through further enhancements to account for the complex nature of void shapes (Jennings, Li & Hu, 2013). As previous authors (Müller *et al.*, 2000; Sutter *et al.*, 2012b, 2014d; Nadathur & Hotchkiss, 2014; Nadathur *et al.*, 2015)

have noted, there tend to be fewer voids in observations than in numerical simulations, especially for small voids. This is due to the conjugate effects of sparsity and biasing of tracers, which can modify the number function in complex ways (Furlanetto & Piran, 2006; Sutter *et al.*, 2014c,d), as well as survey geometries and selection effects, which can non-trivially diminish the void population. However, recently Sutter *et al.* (2014d) showed a correspondence between observed and theoretical number functions once these factors are taken into account.

Figure 8.3 shows the cumulative void number function in BORG reconstructions (blue) compared to dark matter simulations using the same setup (green) and to galaxy voids in the SDSS DR7 (red and purple). The confidence regions are 2-σ Poisson uncertainties and the blue and green lines use Blackwell-Rao estimators to combine the results in 11 realizations.

We can immediately note the excellent agreement between simulations and dark matter voids candidates in the SDSS as found by our methodology. The two void populations are almost indistinguishable at all scales, which demonstrates that the data-constrained number function predicted by our methodology is exactly that of dark matter voids in numerical simulations. In particular, this proves that our framework correctly permits to circumvent the effects of sparsity and biasing of SDSS galaxies on void number count. Indeed, dark matter voids in our reconstructions are densely-sampled with the same number density as in simulations, $\bar{n} = 0.318\ (\mathrm{Mpc}/h)^{-3}$ (512^3 particles in $(750\ \mathrm{Mpc}/h)^3$) compared to $\bar{n} \approx 10^{-3}\ (\mathrm{Mpc}/h)^{-3}$ for SDSS galaxies (Sutter *et al.*, 2012b). Furthermore, any incorrect treatment of galaxy bias by the BORG algorithm would result in a residual bias in our reconstructions that would yield an erroneous void number function as compared to simulations (Sutter *et al.*, 2014c). The absence of any such feature confirms that galaxy bias is correctly accounted for in our analysis and further validates the framework described in chapter 5 (Jasche, Leclercq & Wandelt, 2015).

Additionally, due to the high density of tracer particles, we find at least around one order of magnitude more voids at all scales than the voids directly traced by the SDSS galaxies, which sample the underlying mass distribution only sparsely. This results in a drastic reduction of statistical uncertainty in void catalogs, as we demonstrate in sections 8.3.2 and 8.3.3.

8.3.2 Ellipticity distribution

The shape distribution of voids is complementary to overdense probes of the dark matter density field such as galaxy clusters. Indeed, as matter collapses to form galaxies, voids expand and can do so aspherically. While Icke (1984) argued that voids are expected to become more spherical as they expand, Platen, van de Weygaert & Jones (2008) found that the shape distribution of voids remains complex at late times and showed that the aspherical expansion of voids is strongly linked to the external tidal influence.[3] Therefore, the shapes of empty regions generally change during cosmic evolution and retain information on their formation history. In particular, the void shape distribution potentially serves as a powerful tracer of the equation of state of dark energy (Lee & Park, 2006; Park & Lee, 2007; Biswas, Alizadeh & Wandelt, 2010; Lavaux & Wandelt, 2012; Bos *et al.*, 2012). In addition, the mean stretch of voids along the line of sight may be used for an application of the Alcock-Paczynski test (Alcock & Paczynski, 1979; Ryden, 1995; Lavaux & Wandelt, 2012; Sutter *et al.*, 2012a, 2014a; Hamaus *et al.*, 2014a).

For these applications, it is of crucial importance for the void catalog to be unaffected by systematics due to baryonic physics. Furthermore, as pointed out by Bos *et al.* (2012), in sparse populations such as galaxies it is very difficult to statistically separate ΛCDM from alternative cosmologies using void shapes. As we now show, our framework allows to access void shapes at the level of the dark matter distribution, deeper than with the galaxies, and to reduce the statistical uncertainty due to their sparsity. Note that all the phase information and spatial organization of the LSS is unaffected by our prior assumptions, which generally affect the density amplitudes via the cosmological power spectrum. The geometry of voids discussed here is therefore strongly constrained by the observations.

We simplify the discussion by focusing on the ellipticity, computed by the VIDE toolkit using the eigenvalues of the inertia tensor (for details, see section C.1.3.2 and Sutter *et al.*, 2015b). Figure 8.4 shows the mean ellipticity and the standard error on the mean (i.e. $\sigma/\sqrt{N_\mathrm{v}}$, where σ is the standard deviation and N_v is the number of voids) as a function of void effective radius. The red line represents the galaxy voids directly found in the SDSS data, the blue line the dark matter voids of our data-constrained catalogs, and the green line the voids

[3] Tidal effects are taken into account in our analysis since BORG models gravitational evolution up to second order in Lagrangian perturbation theory.

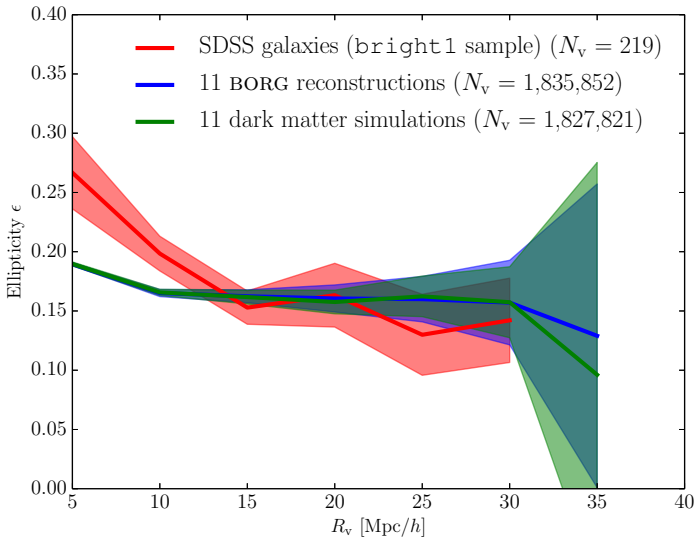

Figure 8.4: Distribution of ellipticities ε versus effective radii of voids. The solid line shows the mean, and the shaded region is the 2-σ confidence region estimated from the standard error on the mean in each radial bin. Small galaxy voids are found more elliptical than dark matter voids because of important Poisson fluctuations below the mean galaxy separation (8 Mpc/h). Ellipticities of dark matter voids in BORG reconstructions and simulations agree at all scales, and the statistical uncertainty in their determination is drastically reduced in comparison to galaxy void catalogs.

found in dark matter simulations prepared with the same setup. The blue and green lines use Blackwell-Rao estimators to combine the results of 11 realizations. For the interpretation of the ellipticity of small galaxy voids, it is useful to recall that the mean galaxy separation in the `bright1` sample is 8 Mpc/h, meaning that Poisson fluctuations will be of importance for voids whose effective radius is below this scale.

The comparison between dark matter voids of BORG reconstructions and of simulations shows that the predicted ellipticities fully agree with the expectations at all scales. This further demonstrates that our candidates qualify as dark matter voids as defined by numerical simulations, in particular alleviating the galaxy bias problem. Furthermore, as already noted, our inference framework produces many more voids than sparse galaxy catalogs, especially at small scales. This results in a radical reduction of statistical uncertainty in the ellipticity prediction for small dark matter voids as compared to galaxy voids, as can be observed in figure 8.4.

8.3.3 Radial density profiles

The radial density profile of voids, reconstructed in real space using techniques such as those described in Pisani *et al.* (2014), can be used to test general relativity and constrain dynamical dark energy models (Shoji & Lee, 2012; Spolyar, Sahlén & Silk, 2013). More generally, it shows a self-similar structure (Colberg *et al.*, 2005; Ricciardelli, Quilis & Varela, 2014; Hamaus, Sutter & Wandelt, 2014; Nadathur *et al.*, 2015), and characterizes the LSS in a fundamental way (van de Weygaert & van Kampen, 1993). All results presented in this section assume that dark matter particles in BORG reconstructions and in simulations live in physical space. The BORG algorithm automatically mitigates redshift-space distortions by treating anisotropic features in the data as noise (Jasche, Leclercq & Wandelt, 2015). Furthermore, as pointed out by Padilla, Ceccarelli & Lambas (2005), redshift-space distortions have very mild effects on void density profiles. We therefore expect our results to be robust under the transformation from real to redshift space.

Using VIDE, we construct the one-dimensional radial density profiles of stacked voids for various void sizes. Note that we do not apply any rescaling to the void sizes as we stack. Figure 8.5 shows two such profiles, for voids of effective radius in the range 6-8 Mpc/h (left panel) and 20-25 Mpc/h (right panel). The solid lines show the mean and the shaded regions are the 2-σ confidence regions estimated from the standard error on the mean, using Blackwell-Rao estimators for BORG reconstructions and dark matter simulations. At the level of statistical error in our results, our reconstructions show radial density profiles in agreement with simulations at all radii and for all void sizes. Note that, if small voids essentially reflect the prior information used for the BORG analysis and N-body filtering, bigger voids are strongly constrained by the data. The profile shapes agree

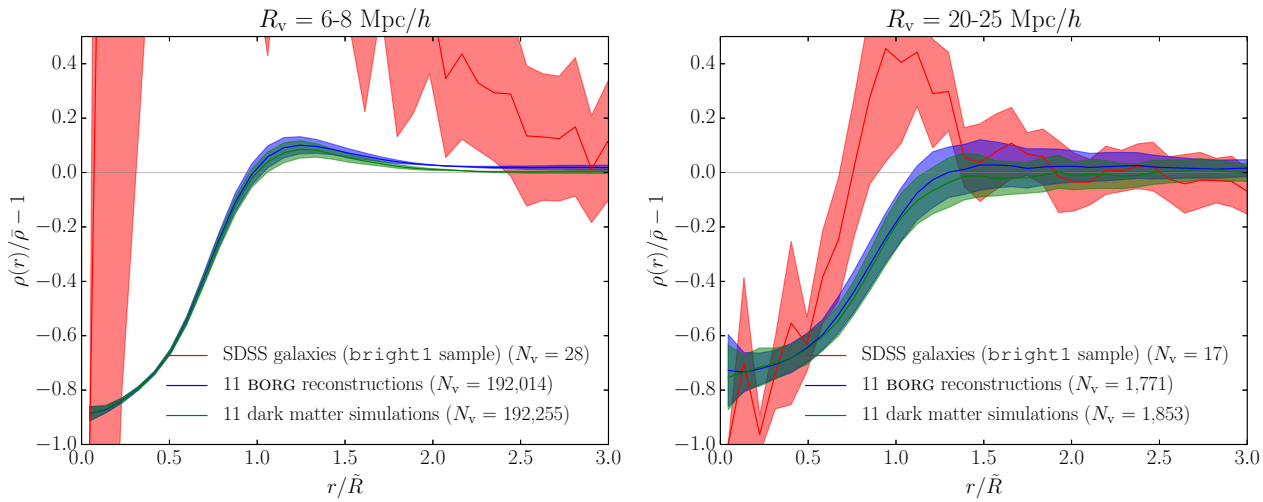

Figure 8.5: One-dimensional radial density profiles of stacked voids, for voids of effective radius in the range 6-8 Mpc/h (left) and 20-25 Mpc/h (right). $\tilde{R}$ corresponds to the median void size in the stack. The solid line shows the mean, and the shaded region is the 2-σ confidence region estimated from the standard error on the mean in each radial bin. Galaxy void profiles are strongly noise-dominated, contrary to dark matter voids. The heights of compensation ridges are different because dark matter voids are identified in a higher density of tracers, which induces a deeper void hierarchy.

nicely with the results of Sutter *et al.* (2014c); Hamaus, Sutter & Wandelt (2014) from dark matter simulations: higher ridges and lower central densities in smaller voids. Specifically, our reconstructions exhibit the same behaviour as simulations, with a transition scale between small overcompensated to large undercompensated voids (Ceccarelli *et al.*, 2013; Paz *et al.*, 2013; Cai *et al.*, 2014; Hamaus *et al.*, 2014b).

In contrast, galaxy void profiles at the same scales are strongly noise-dominated. This is due to the sparsity and biasing of galaxies, which are alleviated with the present approach. In particular, our methodology performs a meaningful compromise between data and prior information, which predicts corrected shapes and smaller variance for the profiles of dark matter voids as compared to galaxy voids. Note that at the same physical scales (e.g. 20 Mpc/h), galaxy voids and dark matter voids have different ridge heights. This is because a deeper void hierarchy emerges in higher tracer sampling densities, affecting the compensation of voids at a given size (Sutter *et al.*, 2014c).

In addition to the location of all dark matter particles, our inference framework also provides their individual velocity vectors, which are predicted from gravitational clustering. While the direct measurement of individual galaxy velocities is very difficult in most observations, our reconstruction technique readily allows to infer the velocity profile of voids. This allows to make a connection between a static (based on the density profiles) and a dynamic (based on the velocity profiles) characterization of voids. In particular, as mentioned before, our results agree with the existence of a transition scale between two regimes: undercompensated, inflowing voids and overcompensated, outflowing voids, respectively known as void-in-cloud and void-in-void in the terminology originally introduced by Sheth & van de Weygaert (2004).

8.4 Summary and conclusions

This chapter is an example of the rich variety of scientific results that have been produced by the recent application (Jasche, Leclercq & Wandelt, 2015) of the Bayesian inference framework BORG (Jasche & Wandelt, 2013a) to the Sloan Digital Sky Survey main sample galaxies. We proposed a method designed to find dark matter void candidates in the Sloan volume. In doing so, we proved that physical voids in the dark matter distribution can be correctly identified by the *ab initio* analysis of galaxy surveys.

Our method relies on the physical inference of the initial conditions for the entire LSS (Jasche & Wandelt, 2013a; Jasche, Leclercq & Wandelt, 2015). Starting from these, we generated realizations of the LSS using a fully non-linear cosmological code. In this fashion, as described in section 8.2.2, we obtained a set of data-constrained reconstructions of the present-day dark matter distribution. The use of fully non-linear dynamics as a filter allowed us to extrapolate the predictions of BORG to the unconstrained non-linear regimes and to obtain an

accurate description of structures. As described in section 8.2.3, we identified the voids in these reconstructions using the void finder of the VIDE pipeline (Sutter *et al.*, 2015b) and applied an additional selection criterion to limit the final catalogs of dark matter voids candidates to regions covered by observations. To check that these candidates qualify for physical voids, we analyzed our catalogs in terms of a set of statistical diagnostics. We focused on three key void statistics, well understood both in data and in simulations, provided by the VIDE toolkit: number function, ellipticity distribution and radial density profile. As mentioned in section 8.2.5, for comparison of our results, we used the void catalog of Sutter *et al.* (2012b), directly based on SDSS main sample galaxies, and unconstrained dark matter simulations produced with the same setup as our reconstructions.

For quantifying the uncertainty, we adopted the same Bayesian philosophy as in the LSS inference framework: several void catalogs are produced, based on different samples of the BORG posterior probability distribution function. Each of them represents a realization of the actual dark matter voids in the Sloan volume, and the variation between these catalogs quantifies the remaining uncertainties of various sources (in particular, survey geometry and selection effects, see chapter 5 for a complete discussion). In order to produce a statistically meaningful combination of our different dark matter void catalogs, in section 8.2.4, we introduced Blackwell-Rao estimators. We showed that the combination of different realizations generally yields an improved estimator for any original void statistic.

For all usual void statistics (number function in section 8.3.1, ellipticity distribution in section 8.3.2 and radial density profiles in section 8.3.3), we found remarkably good agreement between predictions for dark matter voids in our reconstructions and expectations from numerical simulations. This validates our inference framework and qualifies the candidates to physically reasonable dark matter voids, probing a level deeper in the mass distribution hierarchy than galaxies. Further, since sparsity and biasing of tracers modify these statistics (Sutter *et al.*, 2014c), it means that these effects have been correctly accounted for in our analysis. Indeed, in chapter 5 we showed that BORG accurately accounts for luminosity-dependent galaxy bias and performs automatic calibration of the noise level within a fully Bayesian approach. Building on the detailed representation of initial density fields, our reconstructions possess a high density of tracers, $\bar{n} = 0.318\ (\mathrm{Mpc}/h)^{-3}$, contrary to galaxies, which sample the underlying mass distribution only sparsely ($\bar{n} \approx 10^{-3}\ (\mathrm{Mpc}/h)^{-3}$).

Another important aspect of our methodology is that the use of full-scale physical density fields instead of a scarce population of galaxies allows to adjust the density of tracers to reduce shot noise at the desired level. In our analysis, we found at least one order of magnitude more voids at all scales. This yields a radical reduction of statistical uncertainty in noise-dominated void catalogs, as we have shown for ellipticity distributions and density profiles.

In summary, our methodology permits to alleviate the issues due to the conjugate and intricate effects of sparsity and biasing on galaxy void catalogs, to drastically reduce statistical uncertainty in void statistics, and yields new catalogs of dark matter voids for a variety of cosmological applications. For example, these enhanced data sets can be used for cross-correlation with other cosmological probes such as the cosmic microwave background, to study the integrated Sachs-Wolfe effect, or gravitational lensing shear maps. Along with the ensemble mean density field and corresponding standard deviations inferred by BORG, published as supplementary material with Jasche, Leclercq & Wandelt (2015), we believe that the catalogs of our dark matter voids candidates in the Sloan volume can be of interest to the scientific community. For this reason, all the void catalogs used to produce the results described in this chapter have been made publicly available at http://www.cosmicvoids.net, along with the paper corresponding to this chapter (Leclercq *et al.*, 2015).

Our Bayesian methodology, based on inference with BORG and subsequent non-linear filtering of the results, assumes some prior information, namely the standard ΛCDM cosmological framework and initially Gaussian density fluctuations. We want to emphasize that any analysis using our constrained catalogs will be biased toward the confirmation of these assumptions. Therefore, this method will be only applicable if the data contain sufficient support for the presence of non-standard cosmology to overrule the preference for ΛCDM and Gaussianity in our prior. However, any significant departure from standard cosmology means that the prior has been overridden by the likelihood and that such deviations really are supported by the data.

While the recommendations of Sutter *et al.* (2014c) for quantifying and disentangling the effects of sparsity and biasing depend on specific survey details, our inference framework is extremely general. It allows to translate void statistics from current and future galaxy surveys to theory-like, high-resolution dark matter predictions. In this fashion, it is straightforward to decide if any particular void statistic can be directly informative about cosmology. These results indicate a new promising path towards effective and precise void cosmology at the level of the dark matter field.

Chapter 9

Bayesian analysis of the dynamic cosmic web in the SDSS galaxy survey

Contents

> "I just wonder how things were put together."
> — Claude Shannon

Abstract

Recent application of the Bayesian algorithm BORG to the Sloan Digital Sky Survey main sample galaxies resulted in the physical inference of the formation history of the observed large-scale structure from its origin to the present epoch. In this work, we use these inferences as inputs for a detailed probabilistic cosmic web-type analysis. To do so, we generate a large set of data-constrained realizations of the large-scale structure using a fast, fully non-linear gravitational model. We then perform a dynamic classification of the cosmic web into four distinct components (voids, sheets, filaments, and clusters) on the basis of the tidal field. Our inference framework automatically and self-consistently propagates typical observational uncertainties to web-type classification. As a result, this study produces accurate cosmographic classification of large-scale structure elements in the SDSS volume. By also providing the history of these structure maps, the approach allows an analysis of the origin and growth of the early traces of the cosmic web present in the initial density field and of the evolution of global quantities such as the volume and mass filling fractions of different structures. For the problem of web-type classification, the results described in this chapter constitute the first connection between theory and observations at non-linear scales including a physical model of structure formation and the demonstrated capability of uncertainty quantification. A connection between cosmology and information theory using real data also naturally emerges from our probabilistic approach. Our results constitute quantitative chrono-cosmography of the complex web-like patterns underlying the observed galaxy distribution.

This chapter is adapted from its corresponding publication, Leclercq, Jasche & Wandelt (2015c).

9.1 Introduction

The large-scale distribution of matter in the Universe is known to form intricate, complex patterns traced by galaxies. The existence of this large-scale structure, also known as the *cosmic web* (Bond, Kofman & Pogosyan, 1996), has been suggested by early observational projects aiming at mapping the Universe (Gregory & Thompson, 1978; Kirshner *et al.*, 1981; de Lapparent, Geller & Huchra, 1986; Geller & Huchra, 1989; Shectman *et al.*, 1996), and has been extensively analyzed since then by massive surveys such as the 2dFGRS (Colless *et al.*, 2003), the SDSS (e.g. Gott *et al.*, 2005) or the 2MASS redshift survey (Huchra *et al.*, 2012). The cosmic web is usually segmented into different elements: voids, sheets, filaments, and clusters. At late times, low-density regions (voids) occupy most of the volume of the Universe. They are surrounded by walls (or sheets) from which departs a network of denser filaments. At the intersection of filaments lie the densest clumps of matter (clusters). Dynamically, matter tends to flow out of the voids to their compensation walls, transits through filaments and finally accretes in the densest halos.

Describing the cosmic web morphology is an involved task due to the intrinsic complexity of individual structures, but also to their connectivity and the hierarchical nature of their global organization. First approaches (e.g. Barrow, Bhavsar & Sonoda, 1985; Gott, Dickinson & Melott, 1986; Babul & Starkman, 1992; Mecke, Buchert & Wagner, 1994; Sahni, Sathyaprakash & Shandarin, 1998) often characterized the LSS with a set of global and statistical diagnostics, without providing a way to locally identify cosmic web elements. In the last decade, a variety of methods has been developed for segmenting the LSS into its components and applied to numerical simulations and observations. Among them, some focus on investigating one component at a time, in particular filaments (e.g. the Candy model – Stoica *et al.*, 2005; Stoica, Martínez & Saar, 2007, 2010, the skeleton analysis – Novikov, Colombi & Doré, 2006; Sousbie *et al.*, 2008, and DisPerSE – Sousbie, 2011; Sousbie, Pichon & Kawahara, 2011) or voids (e.g. Plionis & Basilakos, 2002; Colberg *et al.*, 2005; Shandarin *et al.*, 2006; Platen, van de Weygaert & Jones, 2007; Neyrinck, 2008; Sutter *et al.*, 2015b; Elyiv *et al.*, 2015, see also Colberg *et al.*, 2008 for a void finder comparison project). Unfortunately, this approach does not allow an analysis of the connections between cosmic web components, identified in the same framework. Another important class of web classifiers dissects clusters, filaments, walls, and voids at the same time. In particular, several recent studies deserve special attention due to their methodological richness. The "T-web" and "V-web" (Hahn *et al.*, 2007a; Forero-Romero *et al.*, 2009; Hoffman *et al.*, 2012) characterize the cosmic web based on the tidal and velocity shear fields. DIVA (Lavaux & Wandelt, 2010) rather uses the shear of the Lagrangian displacement field. ORIGAMI (Falck, Neyrinck & Szalay, 2012) identifies single and multi-stream regions in the full six-dimensional phase-space information (Abel, Hahn & Kaehler, 2012; Neyrinck, 2012; Shandarin, Habib & Heitmann, 2012). The Multiscale Morphology Filter (Aragón-Calvo *et al.*, 2007) and later refinements NEXUS/NEXUS+ (Cautun, van de Weygaert & Jones, 2013) follow a multiscale approach which probes the hierarchical nature of the cosmic web.

In the standard theoretical picture, the cosmic web arises from the anisotropic nature of gravitational collapse, which drives the formation of structure in the Universe from primordial fluctuations (Peebles, 1980). The capital importance of the large-scale tidal field in the formation and evolution of the cosmic web was first pointed out in the seminal work of Zel'dovich (1970). In the Zel'dovich approximation, the late-time morphology of structures is linked to the eigenvalues of the tidal tensor in the initial conditions. Gravitational collapse amplifies any anisotropy present in the primordial density field to give rise to highly asymmetrical structures. This picture explains the segmented nature of the LSS, but not its connectivity. The cosmic web theory of Bond, Kofman & Pogosyan (1996) asserted the deep connection between the tidal field around rare density peaks in the initial fluctuations and the final web pattern, in particular the filamentary cluster-cluster bridges. More generally, the shaping of the cosmic web through gravitational clustering is essentially a deterministic process described by Einstein's equations and the main source of stochasticity in the problem enters in the generation of initial conditions, which are known, from inflationary theory, to resemble a Gaussian random field to very high accuracy (Guth & Pi, 1982; Hawking, 1982; Bardeen, Steinhardt & Turner, 1983). For these reasons, considerable effort has been devoted to a theoretical understanding of the LSS in terms of perturbation theory in the Eulerian and Lagrangian frames (for a review, see Bernardeau *et al.*, 2002). While this approach offers important analytical insights, it only permits to describe structure formation in the linear and mildly non-linear regimes and it is usually limited to the first few correlation functions of the density field. The complete description of the connection between primordial fluctuations and the late-time LSS, including a full phase-space treatment and the entire hierarchy of correlators, has to rely on a numerical treatment through N-body simulations. The characterization of cosmic web environments in the non-linear regime and

the description of their time evolution has only been treated recently, following the application of web classifiers to state-of-the-art simulations. In particular, Hahn *et al.* (2007a); Aragón-Calvo, van de Weygaert & Jones (2010) presented a local description of structure types in high-resolution cosmological simulations. Hahn *et al.* (2007b); Bond, Strauss & Cen (2010); Cautun *et al.* (2014) analyzed the time evolution of the cosmic web in terms of the mass and volume content of web-type components, their density distribution, and a set of new analysis tools especially designed for particular elements.

To the best of our knowledge, neither the classification of cosmic environments at non-linear scales in physical realizations of the LSS nor the investigation of their genesis and growth, using real data and with demonstrated capability of uncertainty quantification, have been treated in the existing literature. In this work, we propose the first probabilistic web-type analysis conducted with observational data in the deeply non-linear regime of LSS formation. We build accurate maps of dynamic cosmic web components with a resolution of around 3 Mpc/h, constrained by observations. In addition, our approach leads to the first quantitative inference of the formation history of these environments and allows the construction of maps of the embryonic traces in the initial perturbations of the late-time morphological features of the cosmic web.

Cosmographic descriptions of the LSS in terms of three-dimensional maps, and in particular a dynamic structure type cartography carry potential for a rich variety of applications. Such maps characterize the anisotropic nature of gravitational structure formation, the clustering behavior of galaxies as a function of their tidal environment and permit to describe the traces of the cosmic web already imprinted in the initial conditions. So far, most investigations focused on understanding the physical properties of dark halos and galaxies in relation to the LSS. Hahn *et al.* (2007a,b, 2009); Hahn, Angulo & Abel (2015); Aragón-Calvo, van de Weygaert & Jones (2010) found a systematic dependence of halo properties such as morphological type, color, luminosity and spin parameter on their cosmic environment (local density, velocity and tidal field). In addition, a correlation between halo shapes and spins and the orientations of nearby filaments and sheets, predicted in simulations (Altay, Colberg & Croft, 2006; Hahn *et al.*, 2007a,b, 2009; Paz, Stasyszyn & Padilla, 2008; Zhang *et al.*, 2009; Codis *et al.*, 2012; Libeskind *et al.*, 2013; Welker *et al.*, 2014; Aragon-Calvo & Yang, 2014; Laigle *et al.*, 2015), has been confirmed by observational galaxy data (Paz, Stasyszyn & Padilla, 2008; Jones, van de Weygaert & Aragón-Calvo, 2010; Tempel, Stoica & Saar, 2013; Zhang *et al.*, 2013). Cartographic descriptions of the cosmic web also permit to study the environmental dependence of galaxy properties (see e.g. Lee & Lee, 2008; Lee & Li, 2008; Park, Kim & Park, 2010; Yan, Fan & White, 2012; Kovač *et al.*, 2014) and to make the connection between the sophisticated predictions for galaxy properties in hydrodynamic simulations (e.g. Vogelsberger *et al.*, 2014; Dubois *et al.*, 2014; Codis *et al.*, 2015) and observations. Another wide range of applications of structure type reconstructions is to probe the effect of the inhomogeneous large-scale structure on photon properties and geodesics. For example, it is possible to interpret the weak gravitational lensing effects of voids (Melchior *et al.*, 2014; Clampitt & Jain, 2015). Dynamic information can also be used to produce prediction templates for secondary effects expected in the cosmic microwave background such as the kinetic Sunyaev-Zel'dovich effect (Li *et al.*, 2014), the integrated Sachs-Wolfe and Rees-Sciama effects (e.g. Cai *et al.*, 2010; Ilić, Langer & Douspis, 2013; Planck Collaboration, 2014a). Lastly, as the cosmic web morphology arises from gravitational instability, it can be used to test general relativity (Falck, Koyama & Zhao, 2015).

Building such refined cosmographic descriptions of the Universe requires high-dimensional, non-linear inferences. In chapter 5 (Jasche, Leclercq & Wandelt, 2015), we presented a chrono-cosmography project, aiming at reconstructing simultaneously the density distribution, the velocity field and the formation history of the LSS from galaxies. To do so, we used an advanced Bayesian inference algorithm to assimilate the Sloan Digital Sky Survey DR7 data into the forecasts of a physical model of structure formation (second order Lagrangian perturbation theory). Besides inferring the four-dimensional history of the matter distribution, these results permit us an analysis of the genesis and growth of the complex web-like patterns that have been observed in our Universe. Therefore, this work constitutes a new chrono-cosmography project, aiming at the analysis of the evolving cosmic web.

Our investigations rely on the inference of the initial conditions in the SDSS volume (see chapter 5). Starting from these, we generate a large set of constrained realizations of the Universe using the COLA method (Tassev, Zaldarriaga & Eisenstein, 2013, see also section 7.3.1). This physical model allows us to perform the first description of the cosmic web in the non-linear regime, using real data, and to follow the time evolution of its constituting elements. Throughout this chapter, we adopt the Hahn *et al.* (2007a) dynamic "T-web" classifier, which segments the LSS into voids, sheets, filaments, and clusters. This choice is motivated by the close relation between the equations that dictate the dynamics of the growth of structures in the Zel'dovich formalism and the Lagrangian description of the LSS which naturally emerges with BORG. As this procedure

relies on the estimation of the eigenvalues of the tidal tensor in Fourier space, it constitutes a non-linear and non-local estimator of structure types, requiring adequate means to propagate observational uncertainties to finally inferred products (web-type maps and all derived quantities), in order not to misinterpret results. The BORG algorithm naturally addresses this problem by providing a set of density realizations constrained by the data. The variation between these samples constitute a thorough quantification of uncertainty coming from all observational effects (in particular the incompleteness of the data because of the survey mask and the radial selection functions, as well as luminosity-dependent galaxy biases, see chapter 5 for details), not only with a point estimation but with a detailed treatment of the likelihood. Hence, for all problems addressed in this work, we get a fully probabilistic answer in terms of a prior and a posterior distribution. Building upon the robustness of our uncertainty quantification procedure, we are able to make the first observationally-supported link between cosmology and information theory (see Neyrinck, 2015, for theoretical considerations related to this question) by looking at the entropy and Kullback-Leibler divergence of probability distribution functions.

This chapter is organized as follows. In section 9.2, we describe our methodology: Bayesian large-scale structure inference with the BORG algorithm, non-linear filtering of samples with COLA and web-type classification using the T-web procedure. In sections 9.3 and 9.4, we describe the cosmic web at present and primordial times, respectively. In section 9.5, we follow the time evolution of web-types as structures form in the Universe. Finally, we summarize our results and offer concluding comments in section 9.6.

9.2 Methods

In this section, we describe our methodology step by step:

1. inference of the initial conditions with BORG (section 9.2.1),
2. generation of data-constrained realizations of the SDSS volume via non-linear filtering of BORG samples with COLA (section 9.2.2),
3. classification of the cosmic web in voids, sheets, filaments, and clusters, using the T-web algorithm (section 9.2.3).

9.2.1 Bayesian large-scale structure inference with BORG

This work builds upon results previously obtained by application of the BORG (Bayesian Origin Reconstruction from Galaxies, Jasche & Wandelt, 2013a) algorithm to the Sloan Digital Sky Survey data release 7 (Jasche, Leclercq & Wandelt, 2015). BORG is a full-scale Bayesian inference code which permits to simultaneously analyze morphology and formation history of the cosmic web (see chapters 4 and 5 for a complete description).

As discussed in Jasche & Wandelt (2013a), accurate and detailed cosmographic inferences from observations require modeling the mildly non-linear and non-linear regime of the presently observed matter distribution. The exact statistical behavior of the LSS in terms of a full probability distribution function for non-linearly evolved density fields is not known. For this reason, the first full-scale reconstructions relied on phenomenological approximations, such as multivariate Gaussian or log-normal distributions, incorporating a cosmological power spectrum to accurately represent correct two-point statistics of density fields (see e.g. Lahav *et al.*, 1994; Zaroubi, 2002; Erdoğdu *et al.*, 2004; Kitaura & Enßlin, 2008; Kitaura *et al.*, 2009; Kitaura, Jasche & Metcalf, 2010; Jasche & Kitaura, 2010; Jasche *et al.*, 2010b,a). However, these prescriptions only model the one and two-point statistics of the matter distribution. Additional statistical complexity of the evolved density field arises from the fact that gravitational structure formation introduces mode coupling and phase correlations. This manifests itself not only in a sheer amplitude difference of density and velocity fields at different redshifts, but also in a modification of their statistical behavior by the generation of higher-order correlation functions. An accurate modeling of these high-order correlators is of crucial importance for a precise description of the connectivity and hierarchical nature of the cosmic web, which is the aim of this chapter.

While the statistical nature of the late-time density field is poorly understood, the initial conditions from which it formed are known to obey Gaussian statistics to very great accuracy (Planck Collaboration, 2015). Therefore, it is reasonable to account for the increasing statistical complexity of the evolving matter distribution by a dynamical model of structure formation linking initial and final conditions. This naturally turns the problem of LSS analysis to the task of inferring the initial conditions from present cosmological observations (Jasche & Wandelt, 2013a; Kitaura, 2013; Wang *et al.*, 2013). This approach yields a very high-dimensional

and non-linear inference problem. Typically, the parameter space to explore comprises on the order of 10^6 to 10^7 elements, corresponding to the voxels of the map to be inferred. For reasons linked to computational cost, the BORG algorithm employs 2LPT as an approximation for the actual gravitational dynamics linking initial three-dimensional Gaussian density fields to present, non-Gaussian density fields. As known from perturbation theory (see e.g. Bernardeau *et al.*, 2002), in the linear and mildly non-linear regime, 2LPT correctly describes the one-, two- and three-point statistics of the matter distribution and also approximates very well higher-order correlators. It accounts in particular for tidal effects in its regime of validity. Consequently, the BORG algorithm correctly transports the observational information corresponding to complex web-like features from the final density field to the corresponding initial conditions. Note that such an explicit Bayesian forward-modeling approach is always more powerful than constraining (part of) the sequence of correlation functions, as it accounts for the entire dark matter dynamics (in particular for the infinite hierarchy of correlators), in its regime of validity. This is of particular importance, since the hierarchy of correlation functions has been shown to be an insufficient description of density fields in the non-linear regime (Carron, 2012; Carron & Neyrinck, 2012).

As discussed in chapter 5 (Jasche, Leclercq & Wandelt, 2015), our analysis comprehensively accounts for observational effects such as selection functions, survey geometry, luminosity-dependent galaxy biases and noise. Corresponding uncertainty quantification is provided by sampling from the high-dimensional posterior distribution via an efficient implementation of the Hamiltonian Markov Chain Monte Carlo method (see chapter 4 and Jasche & Wandelt, 2013a, for details). In particular, luminosity-dependent galaxy biases are explicitly part of the BORG likelihood and the bias amplitudes are inferred self-consistently during the run. Though not explicitly modeled, redshift-space distortions are automatically mitigated: due to the prior preference for homogeneity and isotropy, such anisotropic features are treated as noise in the data.

In the following, we make use of the 12,000 samples of the posterior distribution for primordial density fields, obtained in chapter 5. These reconstructions, constrained by SDSS observations, act as initial conditions for the generation of constrained large-scale structure realizations. It is important to note that we directly make use of BORG outputs without any further post-processing, which demonstrates the remarkable quality of our inference results.

9.2.2 Non-linear filtering of samples with COLA

In section 2.3 (Leclercq *et al.*, 2013, section 2.A), we performed a study of differences in the representation of structure types in density fields predicted by LPT and N-body simulations. To do so, we used the same web-type classification procedure as in this work (see sections 9.2.3 and C.2). In spite of the visual similarity of LPT and N-body density fields at large and intermediate scales (above a few Mpc/h), we found crucial differences in the representation of structures. Specifically, LPT predicts fuzzier halos than full gravity, and incorrectly assigns the surroundings of voids as part of them. This manifests itself in an overprediction of the volume occupied by clusters and voids at the detriment of sheets and filaments. The substructure of voids is also known to be incorrectly represented in 2LPT (Sahni & Shandarin, 1996; Neyrinck, 2013; Leclercq *et al.*, 2013).

For these reasons, in this chapter we cannot directly make use of the final BORG density samples, which are a prediction of the 2LPT model. Instead, we rely on the inferred initial conditions, which contain the data constraints (as described in chapter 5) and on a non-linear filtering step (see chapter 7) similar to the one described in chapter 8 (Leclercq *et al.*, 2015). Due to the large number of samples to be processed for this work, we do not use a fully non-linear simulation code as in chapter 8, but the COLA method (Tassev, Zaldarriaga & Eisenstein, 2013, see also section 7.3.1).

The generation of the set of non-linear BORG-COLA samples, used in this chapter, is described in section 7.3.2.

9.2.3 Classification of the cosmic web

The BORG filtered reconstructions permit a variety of scientific analyses of the large scale structure in the observed Universe. In this work, we focus specifically on the possibility to characterize the cosmic web by distinct structure types. Generally, any of the methods cited in the introduction (section 9.1) can be employed for analysis of our density samples, however for the purpose of this chapter, we follow the "T-web" classification procedure as proposed by Hahn *et al.* (2007a), described in section C.2.

The basic idea of this dynamical classification approach is that the eigenvalues $\mu_1 \leq \mu_2 \leq \mu_3$ of the tidal tensor $\mathcal{T}_{ij} \equiv \mathrm{H}(\tilde{\Phi})_{ij}$ (Hessian of the rescaled gravitational potential) characterize the geometrical properties of each point in space. With these definitions, the three eigenvalues of the tidal tensor form a decomposition of the density contrast field, in the sense that the trace of $\mathcal{T}$ is $\mu_1 + \mu_2 + \mu_3 = \delta$. Each spatial point can then be classified as a specific web type by considering the signs of μ_1, μ_2, μ_3, according to the rules given in table C.1.

Several extensions of this classification procedure exist, that permit different classification up to sub-megaparsec scales (see section C.2.4). In this work, we will probe scales down to ~ 3 Mpc/h (the voxel size in our reconstructions). Therefore, we will be content with the original classification procedure as proposed by Hahn *et al.* (2007a).

It is important to note that the tidal tensor and the rescaled gravitational potential are both physical quantities, and hence their calculation requires the availability of a full physical density field in contrast to a smoothed mean reconstruction of the density field. As described in chapter 5, density samples obtained by the BORG algorithm provide such required full physical density fields. The tidal tensor can therefore easily be calculated in each density sample from the Fourier space representations of equations (C.6) and (C.7) (see section C.2.5 and Hahn *et al.*, 2007a; Forero-Romero *et al.*, 2009, for details on the technical implementation).

The web classifier provides four voxel-wise scalar fields that characterize the large scale structure. In a specific realization, the answer is unique, meaning that these fields obey the following conditions at each voxel position $\vec{x}_k$:

$$\mathrm{T}_i(\vec{x}_k) \in \{0,1\} \text{ for } i \in [\![0,3]\!] \quad \text{and} \quad \sum_{i=0}^{3} \mathrm{T}_i(\vec{x}_k) = 1 \tag{9.1}$$

where T_0 = void, T_1 = sheet, T_2 = filament, T_3 = cluster. In this work, we follow the Bayesian approach of Jasche, Leclercq & Wandelt (2015) and quantify the degree of belief in structure type classification. Specifically, our web classification is given in terms of four voxel-wise scalar fields that obey the following conditions at each voxel position $\vec{x}_k$:

$$\mathcal{T}_i(\vec{x}_k) \in [0,1] \text{ for } i \in [\![0,3]\!] \quad \text{and} \quad \sum_{i=0}^{3} \mathcal{T}_i(\vec{x}_k) = 1. \tag{9.2}$$

Here, $\mathcal{T}_i(\vec{x}_k) \equiv \langle \mathrm{T}_i(\vec{x}_k) \rangle_{\mathcal{P}(\mathrm{T}_i(\vec{x}_k)|d)} = \mathcal{P}(\mathrm{T}_i(\vec{x}_k)|d)$ are the posterior probabilities indicating the possibility to encounter specific structure types at a given position in the observed volume, conditional on the data. These are estimated by applying the web classification to all density samples and counting the relative frequencies at each individual spatial coordinate within the set of samples (see section 5 in Jasche *et al.*, 2010b). With this definition, the cosmic web-type posterior mean is given by

$$\langle \mathcal{P}(\mathrm{T}_i(\vec{x}_k)|d) \rangle = \frac{1}{N} \sum_{n=1}^{N} \sum_{j=0}^{3} \delta_{\mathrm{K}}^{\mathrm{T}_i(\vec{x}_k)\mathrm{T}_j^n(\vec{x}_k)}, \tag{9.3}$$

where n labels one of the N samples, $\mathrm{T}_j^n(\vec{x}_k)$ is the result of the web classifier on the n-th sample (i.e. a unit four-vector at each voxel position $\vec{x}_k$ containing zeros except for one component, which indicates the structure type), and δ_{K}^{ab} is a Kronecker symbol.

9.3 The late-time large-scale structure

In this section, we discuss the results of our analysis of the final density field, at $a = 1$. For reasons of computational time with COLA filtering (see section 9.2.2), we kept around 10% of the original set of samples obtained in chapter 5. In order to mitigate as much as possible the effects of correlation among samples, we maximally separated the samples kept for the present analysis, keeping one out of ten consecutive samples of the original Markov Chain. Hence, for all results discussed in this section, we used a total of 1,097 samples inferred by BORG and filtered with COLA.

9.3.1 Tidal environment

As a natural byproduct, the application of the T-web classifier to density samples yields samples of the pdfs for the three eigenvalues of the tidal field tensor. These pdfs account for the assumed physical model of structure formation and the data constraints, and quantify uncertainty coming in particular from selection

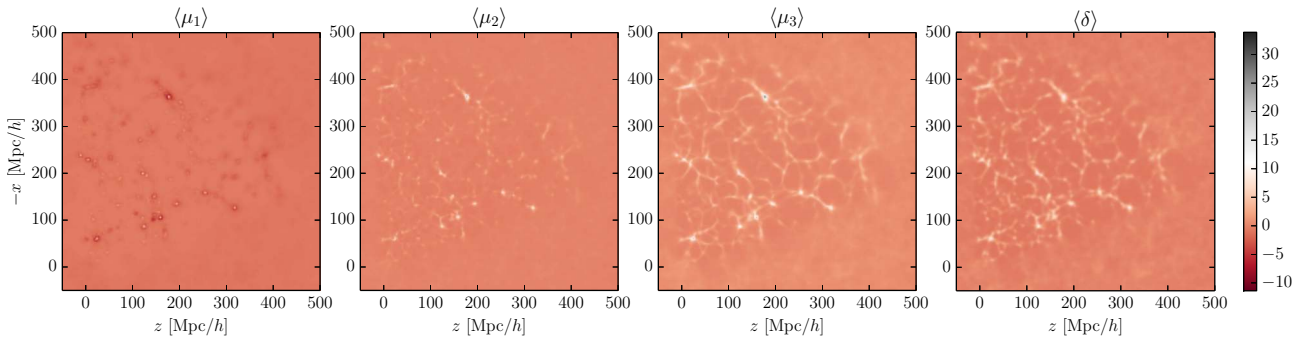

Figure 9.1: Slices through the three-dimensional ensemble posterior mean for the eigenvalues $\mu_1 \leq \mu_2 \leq \mu_3$ of the tidal field tensor in the final conditions, estimated from 1,097 samples. The rightmost panel shows the corresponding slice through the posterior mean for the final density contrast $\delta = \mu_1 + \mu_2 + \mu_3$, obtained in section 5.3.1.

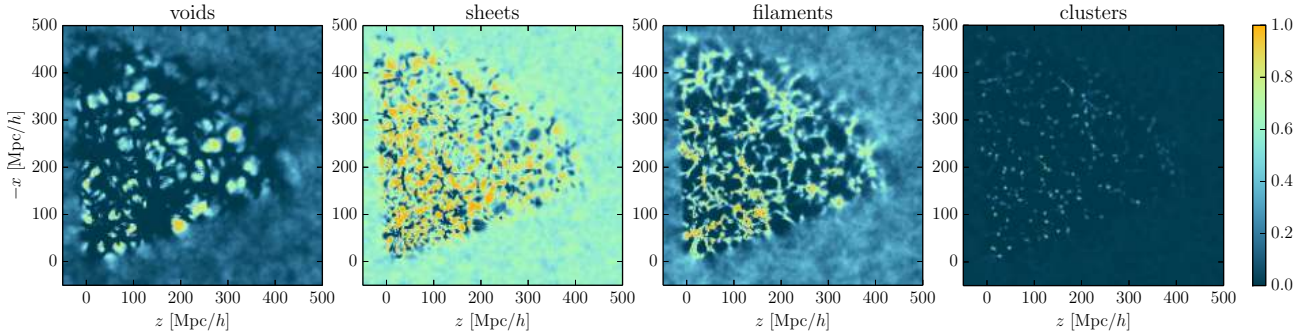

Figure 9.2: Slices through the posterior mean for different structure types (from left to right: void, sheet, filament, and cluster) in the late-time large-scale structure in the Sloan volume ($a = 1$). These four three-dimensional voxel-wise pdfs sum up to one on a voxel basis.

effects, surveys geometries and galaxy biases. In a similar fashion as described in section 5.3, the ensemble of samples permits us to provide any desired statistical summary such as mean and variance.

In figure 9.1, we show slices through the ensemble mean fields μ_1, μ_2 and μ_3. For visual comparison, the rightmost panel of figure 9.1 shows the corresponding slice through the posterior mean of the final density contrast, $\delta = \mu_1 + \mu_2 + \mu_3$, obtained in section 5.3.1. Different morphologies can be observed in the data-constrained parts of these slices: μ_1, μ_2 and μ_3 respectively trace well the clusters, filaments and sheets, as we now argue. The μ_1 field is rather homogeneous, apart for small spots where all eigenvalues are largely positive, i.e. undergoing dramatic gravitational collapse along three axes. These correspond to the dynamic clusters. Note that there exists a form of "tidal compensation": these clusters are surrounded by regions where μ_1 is smaller than its cosmic mean. More patterns can be observed in the μ_2 field: it also exhibits filaments (appearing as dots when piercing the slice). Finally, the μ_3 field is highly-structured, as it also traces sheets (which appear filamentary when sliced). Dynamic voids can also be easily distinguished in this field, wherever μ_3 is negative.

9.3.2 Probabilistic web-type cartography

Building upon previous results and using the procedure described in section 9.2.3, we obtain probabilistic maps of structures. More precisely, we obtain four probability distributions at each spatial position, $\mathcal{P}(\mathrm{T}_i(\vec{x_k})|d)$, indicating the possibility to encounter a specific structure type (cluster, filament, sheet, void) at that position. As noted in section 9.2.3, these pdfs take their values in the range $[0, 1]$ and sum up to one on a voxel-basis. Figure 9.2 shows slices through their means (see equation (9.3)). The plot shows the anticipated behavior, with a high degree of structure and values close to certainty (i.e. zero or one) in regions covered by data, while the unobserved regions approach a uniform value corresponding to the prior. At this point, it is worth noting that the T-web classifier has a prior preference for some structure types. Using unconstrained large-scale structure

Structure type	$\mu_{\mathcal{P}(\mathrm{T}_i)}$	$\sigma_{\mathcal{P}(\mathrm{T}_i)}$
Late-time large-scale structure ($a = 1$)		
Void	0.14261	6.1681×10^{-4}
Sheet	0.59561	6.3275×10^{-4}
Filament	0.24980	5.5637×10^{-4}
Cluster	0.01198	5.8793×10^{-5}

Table 9.1: Prior probabilities assigned by the T-web classifier to the different structures types, in the late-time large-scale structure ($a = 1$).

realizations produced with the same setup,[1] we measured that these prior probabilities, $\mathcal{P}(\mathrm{T}_i)$, can be well described by Gaussians whose mean and standard deviation are given in table 9.1.

In addition to their ensemble mean, the set of samples permits to propagate uncertainty quantification to web-type classification. In particular, it allows us to locally assess the strength of data constraints. In information theory, a convenient way to characterize the uncertainty content of a random source $\mathcal{S}$ is the Shannon entropy (Shannon, 1948), defined by

$$H\left[\mathcal{S}\right] \equiv -\sum_i p_i \log_2(p_i), \tag{9.4}$$

where the p_i are the probabilities of possible events. This definition yields expected properties and accounts for the intuition that the more likely an event is, the less information it provides when it occurs (i.e. the more it contributes to the source entropy). We follow this prescription and write the voxel-wise entropy of the web-type posterior, $\mathcal{P}(\mathrm{T}(\vec{x}_k)|d)$, as

$$H\left[\mathcal{P}(\mathrm{T}(\vec{x}_k)|d)\right] \equiv -\sum_{i=0}^{3} \mathcal{P}(\mathrm{T}_i(\vec{x}_k)|d) \log_2(\mathcal{P}(\mathrm{T}_i(\vec{x}_k)|d)). \tag{9.5}$$

It is a number in the range $[0, 2]$ and its natural unit is the shannon (Sh). $H = 0$ Sh in the case of perfect certainty, i.e. when the data constraints entirely determine the underlying structure type: $\mathcal{P}(\mathrm{T}_{i_0}(\vec{x}_k)|d)$ is 1 for one i_0 and 0 for $i \neq i_0$. H reaches its maximum value of 2 Sh when all $\mathcal{P}(\mathrm{T}_i(\vec{x}_k)|d)$ are equal to $1/4$. This is the case of maximal randomness: all the events being equally likely, no information is gained when one occurs.

A slice through the voxel-wise entropy of the web-type posterior is shown in the left panel of figure 9.3. Generally, the entropy map reflects the information content of the posterior pdf, which comes from augmenting the information content of the prior pdf with the data constraints, in the Bayesian way.

The entropy takes low values and shows a high degree of structure in the regions where data constraints exist, and even reaches zero in some spots where the data are perfectly informative. Comparing with figures 9.1 and 9.2, one can note that this structure is highly non-trivial and does not follow any of the previously described patterns. This is due to the facts that in a Poisson process, the signal (here the density, inferred in section 5.3.1) is correlated with the uncertainty and that structure types classification further is a non-linear function of the density field. In the unobserved regions, the entropy fluctuates around a constant value of about 1.4 Sh, which characterizes the information content of the prior. This value is consistent with the expectation, which can be computed using equation (9.5) (unconditional on the data) and the numbers given in table 9.1.

The information-theoretic quantity that measures the information gain (in shannons) due to the data is the relative entropy or Kullback-Leibler divergence (Kullback & Leibler, 1951) of the posterior from the prior,

$$\begin{aligned} D_{\mathrm{KL}}\left[\mathcal{P}(\mathrm{T}(\vec{x}_k)|d) \| \mathcal{P}(\mathrm{T})\right] &\equiv \sum_{i=0}^{3} \mathcal{P}(\mathrm{T}_i(\vec{x}_k)|d) \log_2\left(\frac{\mathcal{P}(\mathrm{T}_i(\vec{x}_k)|d)}{\mathcal{P}(\mathrm{T}_i)}\right) \\ &= -H\left[\mathcal{P}(\mathrm{T}(\vec{x}_k)|d)\right] - \sum_{i=0}^{3} \mathcal{P}(\mathrm{T}_i(\vec{x}_k)|d) \log_2(\mathcal{P}(\mathrm{T}_i)). \end{aligned} \tag{9.6}$$

[1] By this, we specifically mean realizations obtained from initial randomly-generated Gaussian density fields with an Eisenstein & Hu (1998, 1999) power spectrum using the fiducial cosmological parameters of the BORG analysis ($\Omega_\mathrm{m} = 0.272$, $\Omega_\mu = 0.728$, $\Omega_\mathrm{b} = 0.045$, $h = 0.702$, $\sigma_8 = 0.807$, $n_\mathrm{s} = 0.961$, see equation (5.1)). The density field is defined on a 750 Mpc/h cubic grid of 256^3-voxels and populated by 512^3 dark matter particles, which are evolved to $z = 69$ with 2LPT and from $z = 69$ to $z = 0$ with COLA, using 30 timesteps logarithmically-spaced in the scale factor. The particles are binned on a 256^3-voxel grid with the CiC scheme to get the final density field.

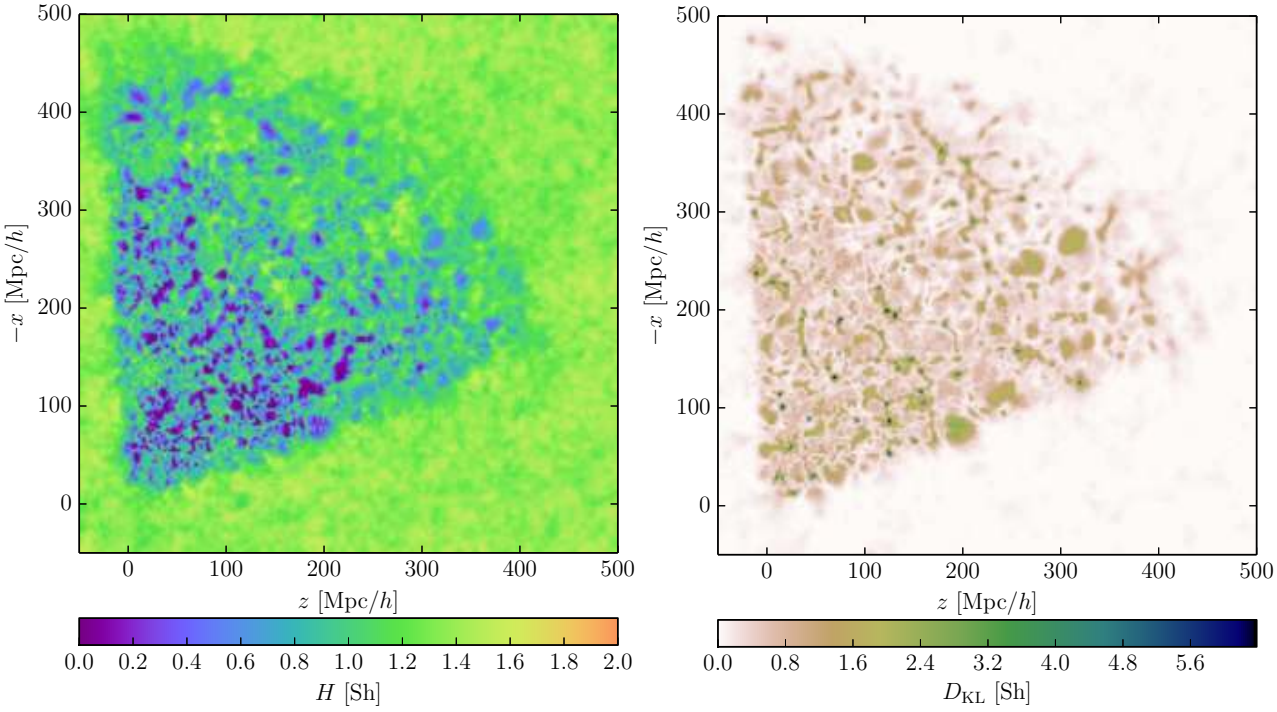

Figure 9.3: Slices through the entropy of the structure types posterior (left panel) and the Kullback-Leibler divergence of the posterior from the prior (right panel), in the final conditions. The entropy H, defined by equation (9.5), quantifies the information content of the posterior pdf represented in figure 9.2, which results from fusing the information content of the prior and the data constraints. The Kullback-Leibler divergence D_{KL}, defined by equation (9.6), represents the information gained in moving from the prior to the posterior. It quantifies the information that has been learned on structure types by looking at SDSS galaxies.

It is a non-symmetric measure of the difference between the two probability distributions.

A slice through the voxel-wise Kullback-Leibler divergence of the web-type posterior from the prior is shown in the right panel of figure 9.3. As expected, the information gain is zero out of the survey boundaries. In the observed regions, SDSS galaxies are informative on underlying structure types at the level of at least ~ 1 Sh. This number can go to ~ 3 Sh in the interior of deep voids and up to ~ 6 Sh in the densest clusters. This map permits to visualize the regions where additional data would be needed to improve structure type classification, e.g. in some high-redshift regions where uncertainty remains due to selection effects.

9.3.3 Volume and mass filling fractions

A characterization of large scale environments commonly found in literature involves evaluating global quantities such as the volume and mass content of these structures. In a particular realization, the volume filling fraction (VFF) for structure type T_i is the number of voxels of type T_i divided by the total number of voxels in the considered volume,

$$\mathrm{VFF}(\mathrm{T}_i) \equiv \frac{\sum_{\vec{x}_k} \sum_{j=0}^{3} \delta_{\mathrm{K}}^{\mathrm{T}_i(\vec{x}_k)\mathrm{T}_j^n(\vec{x}_k)}}{N_{\mathrm{vox}}}. \tag{9.7}$$

The mass filling fraction (MFF) can be obtained in a similar manner by weighting the same sum by the local density $\rho(\vec{x}_k) = \bar{\rho}\,(1 + \delta(\vec{x}_k))$,

$$\mathrm{MFF}(\mathrm{T}_i) \equiv \frac{\sum_{\vec{x}_k} \sum_{j=0}^{3} (1 + \delta(\vec{x}_k)) \delta_{\mathrm{K}}^{\mathrm{T}_i(\vec{x}_k)\mathrm{T}_j^n(\vec{x}_k)}}{\sum_{\vec{x}_k} (1 + \delta(\vec{x}_k))}. \tag{9.8}$$

To ensure that results are not prior-dominated, we measured the VFFs and MFFs in the data-constrained parts of our realizations. More precisely, we limited ourselves to the voxels where the survey response operator (representing simultaneously the survey geometry and the selection effects, see sections 4.2 and 5.1) is strictly positive. This amounts to $N_{\mathrm{vox}} = 3{,}148{,}504$ out of $256^3 = 16{,}777{,}216$ voxels, around 18.7% of the full box (see also section 8.2.3.2 and figure 8.2). In equations (9.7) and (9.8), $\vec{x}_k$ labels one of these voxels.

By measuring the VFF and MFF of different structure types in each constrained realization of our ensemble, we obtained the posterior pdfs, $\mathcal{P}(\mathrm{VFF}(\mathrm{T}_i)|d)$ and $\mathcal{P}(\mathrm{MFF}(\mathrm{T}_i)|d)$, conditional on the data. Similarly, we computed the prior pdfs, $\mathcal{P}(\mathrm{VFF}(\mathrm{T}_i))$ and $\mathcal{P}(\mathrm{MFF}(\mathrm{T}_i))$, using unconstrained realizations produced with the same setup. We found that all these pdfs can be well described by Gaussians, the mean and variance of which are given in tables 9.2 and 9.3.

Previous studies on this topic (e.g. Doroshkevich, 1970b; Shen *et al.*, 2006; Hahn *et al.*, 2007a; Forero-Romero *et al.*, 2009; Jasche *et al.*, 2010b; Aragón-Calvo, van de Weygaert & Jones, 2010; Shandarin, Habib & Heitmann, 2012; Cautun *et al.*, 2014) have found a wide range of values for the VFF and MFF of structures (see e.g. table 3 in Cautun *et al.*, 2014). For example, existing studies found that clusters occupy at most a few percent of the volume of the Universe but contribute significantly to the mass content, with a MFF ranging from $\sim 10\%$ (Hahn *et al.*, 2007a; Cautun *et al.*, 2014) to $\sim 40\%$ (Shandarin, Habib & Heitmann, 2012). The void volume fraction can vary from $\sim 10\%$ (Hahn *et al.*, 2007a) to $\sim 80\%$ (Aragón-Calvo, van de Weygaert & Jones, 2010; Shandarin, Habib & Heitmann, 2012; Cautun *et al.*, 2014); in the Forero-Romero *et al.* (2009) formalism (see section C.2.4), it is a very sensitive function of the threshold μ_{th} (figure 9 in Jasche *et al.*, 2010b). These large disparities in the literature arise because different algorithms use various information and criteria for classifying the cosmic web. For this reason, we believe that it is only relevant to make relative statements for the same setup, i.e. to compare our results to the corresponding prior quantities, as done in tables 9.2 and 9.3. In this purpose, the large number of samples used allowed a precise characterization of the pdfs so that all digits quoted in the tables are significant. Note that all our analyses are repeatable for different setups, which allows in principle a comparison with any previous work.

As expected for a Bayesian update of the degree of belief, the posterior quantities generally have smaller variance and a mean value displaced from the prior mean. For the MFF, the posterior means are always within two standard deviations of the corresponding prior means. The analysis shows that in the SDSS, a larger mass fraction is occupied by clusters, sheets, and voids, at the detriment of filaments, in comparison to the prior expectation. The data also favor a smaller filling of the Sloan volume by filaments and sheets and larger filling by voids and clusters. For the cluster VFF, the posterior mean, $\mu_{\mathrm{VFF}(\mathrm{T}_3)|d} = 0.01499$ is at about 15 standard deviations ($\sigma_{\mathrm{VFF}(\mathrm{T}_3)} = 1.9194 \times 10^{-4}$) of the prior mean, $\mu_{\mathrm{VFF}(\mathrm{T}_3)} = 0.01198$. Given other results on the VFF and MFF, we believe that the data truly favor a higher volume content in clusters as compared to the

Structure type	μ_{VFF}	σ_{VFF}	μ_{VFF}	σ_{VFF}
	Late-time large-scale structure ($a = 1$)			
	Posterior		Prior	
Void	0.14897	1.8256×10^{-3}	0.14254	6.2930×10^{-3}
Sheet	0.58914	1.3021×10^{-3}	0.59562	2.2375×10^{-3}
Filament	0.24689	1.1295×10^{-3}	0.24986	4.4440×10^{-3}
Cluster	0.01499	8.7274×10^{-5}	0.01198	1.9194×10^{-4}

Table 9.2: Mean and standard deviation of the prior and posterior pdfs for the volume filling fraction of different structure types in the late-time large-scale structure ($a = 1$).

Structure type	μ_{MFF}	σ_{MFF}	μ_{MFF}	σ_{MFF}
	Late-time large-scale structure ($a = 1$)			
	Posterior		Prior	
Void	0.04050	8.3531×10^{-4}	0.03876	2.3352×10^{-3}
Sheet	0.35605	1.2723×10^{-3}	0.35286	3.6854×10^{-3}
Filament	0.47356	1.5661×10^{-3}	0.48170	4.2215×10^{-3}
Cluster	0.12990	6.4966×10^{-4}	0.12666	1.8284×10^{-3}

Table 9.3: Mean and standard deviation of the prior and posterior pdfs for the mass filling fraction of different structure types in the late-time large-scale structure ($a = 1$).

structure formation model used as prior. However, this surprising result should be treated with care; part of the discrepancy is likely due to the original BORG analysis, which optimizes the initial conditions for evolution with 2LPT (instead of the non-linear evolution with COLA used for the present work). LPT predicts fuzzier halos than N-body dynamics, which results in the incorrect prediction of a high cluster VFF (see section 2.3; Leclercq *et al.*, 2013).

9.4 The primordial large-scale structure

In this section, we discuss the results of our analysis of the initial density field, at $a = 10^{-3}$. Since the analysis of the primordial large-scale structure does not involve an additional filtering step, we have been able to keep a larger number of samples of the posterior pdf for initial conditions, obtained in chapter 5. Hence, for all results described in this section, we used a total of $4,473$ samples.

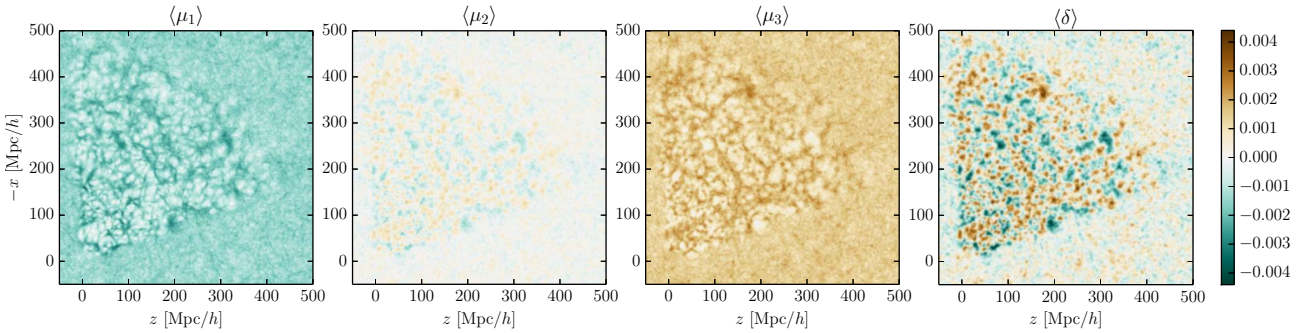

Figure 9.4: Slices through the three-dimensional ensemble posterior mean for the eigenvalues $\mu_1 \leq \mu_2 \leq \mu_3$ of the tidal field tensor in the initial conditions, estimated from $4,473$ samples. The rightmost panel shows the corresponding slice through the posterior mean for the initial density contrast $\delta = \mu_1 + \mu_2 + \mu_3$, obtained in section 5.3.1.

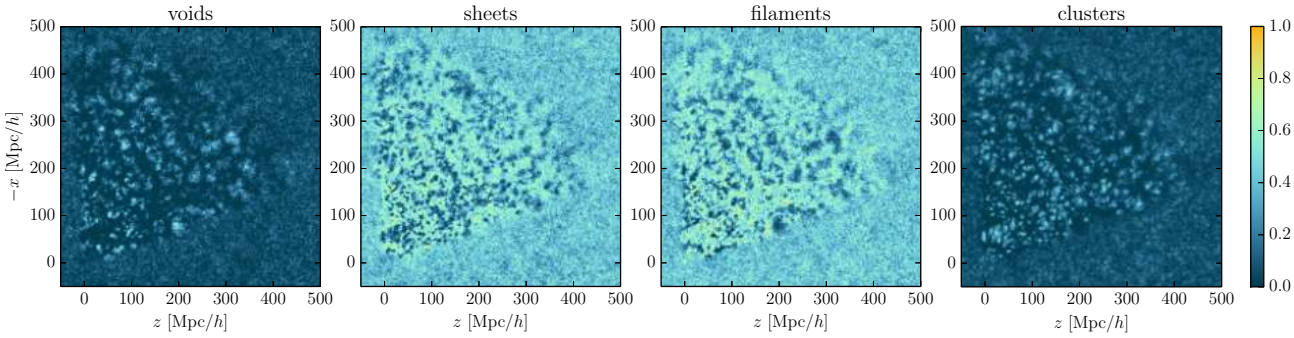

Figure 9.5: Slices through the posterior mean for different structure types (from left to right: void, sheet, filament, and cluster) in the primordial large-scale structure in the Sloan volume ($a = 10^{-3}$). These four three-dimensional voxel-wise pdfs sum up to one on a voxel basis.

9.4.1 Tidal environment

In a similar fashion as in section 9.3.1, the application of the T-web classifier to initial density samples yields the posterior pdf for the three eigenvalues, μ_1, μ_2 and μ_3, of the initial tidal field tensor. Figure 9.4 shows slices through their means. For visual comparison, the rightmost panel shows the corresponding slice through the posterior mean of the initial density contrast, $\delta = \mu_1 + \mu_2 + \mu_3$, obtained in section 5.3.1.

In a Gaussian random field, μ_1 is generally negative, μ_3 is generally positive and μ_2 close to zero (see the unobserved parts of the slices in figure 9.4). In addition, μ_2 closely resembles the total density contrast δ up to a global scaling. In the constrained regions, the eigenvalues of the initial tidal tensor follow this behavior. The structure observed in their maps is visually consistent with the decomposition of Gaussian density fluctuations as shown by the right panel.

9.4.2 Probabilistic web-type cartography

Looking at the sign of the eigenvalues of the initial tidal tensor and following the procedure described in section 9.2.3, we obtain a probabilistic cartography of the primordial large-scale structure. As before, we obtain four voxel-wise pdfs $\mathcal{P}(\mathrm{T}_i(\vec{x_k})|d)$, taking their values in the range $[0, 1]$ and summing up to one. Figure 9.5 shows slices through their means, defined by equation (9.3). As in the final conditions, the maps exhibit structure in the data-constrained regions and approach uniform values in the unobserved parts, corresponding to the respective priors. Using unconstrained realizations of Gaussian random fields produced with the same setup,[2] we measured these prior probabilities. Their means and standard deviations are given in table 9.4.

At this point, it is worth mentioning that there exists an additional symmetry for Gaussian random fields. Since the definition of the tidal tensor is linear in the density contrast (see equations (C.6) and (C.7)) and since positive and negative density contrasts are equally likely, a positive and negative value for a given μ_i have the same probabilities. Because of this sign symmetry, the pdfs for voids and clusters (0 or 3 positive/negative eigenvalues) and the pdfs for sheets and filaments (1 or 2 positive/negative eigenvalues) are equal. This can be checked both in table 9.4 and in the unconstrained regions of the maps in figure 9.5. In the constrained regions, a qualitative complementarity between pdfs for voids and clusters and for sheets and filaments can be observed. This can be explained by the following. As $\sum_i \mathcal{P}(\mathrm{T}_i(\vec{x_k})|d) = 1$ and assuming that $\mathcal{P}(\mathrm{T}_i(\vec{x_k})|d) \approx \mathcal{P}(\mathrm{T}_{3-i}(\vec{x_k})|d)$ for unlikely events, consistently with the previous remark, we get $\mathcal{P}(\mathrm{T}_0(\vec{x_k})|d) \approx 1 - \mathcal{P}(\mathrm{T}_3(\vec{x_k})|d)$ wherever $\mathcal{P}(\mathrm{T}_1(\vec{x_k})|d) \approx \mathcal{P}(\mathrm{T}_2(\vec{x_k})|d)$ is sufficiently small; and $\mathcal{P}(\mathrm{T}_1(\vec{x_k})|d) \approx 1 - \mathcal{P}(\mathrm{T}_2(\vec{x_k})|d)$ wherever $\mathcal{P}(\mathrm{T}_0(\vec{x_k})|d) \approx \mathcal{P}(\mathrm{T}_3(\vec{x_k})|d)$ is sufficiently small. These results are therefore consistent with expectations based on Gaussianity for the primordial large-scale structure in the Sloan volume.

In a similar fashion as in section 9.3.2, the ensemble of samples permits us to propagate uncertainties to structure type classification and to characterize the strength of data constraints. In the left panel of figure 9.6, we show a slice through the voxel-wise entropy of the web-type posterior pdf in the initial conditions, defined by equation (9.5). This function quantifies the information content of the posterior, which comes from both the prior and the data constraints. As in the final conditions, the entropy takes lower values inside the survey

[2] We used the initial conditions of our set of unconstrained simulations (see footnote 1).

Structure type	$\mu_{\mathcal{P}(\mathrm{T}_i)}$	$\sigma_{\mathcal{P}(\mathrm{T}_i)}$
Primordial large-scale structure ($a = 10^{-3}$)		
Void	0.07979	5.4875×10^{-5}
Sheet	0.42022	1.0240×10^{-4}
Filament	0.42022	1.0412×10^{-4}
Cluster	0.07978	5.6337×10^{-5}

Table 9.4: Prior probabilities assigned by the T-web classifier to the different structures types, in the primordial large-scale structure ($a = 10^{-3}$).

Structure type	μ_{VFF}	σ_{VFF}	μ_{VFF}	σ_{VFF}
	Primordial large-scale structure ($a = 10^{-3}$)			
	Posterior		Prior	
Void	0.07994	4.0221×10^{-4}	0.07977	1.0200×10^{-3}
Sheet	0.41994	6.1770×10^{-4}	0.42019	1.7885×10^{-3}
Filament	0.42048	6.3589×10^{-4}	0.42024	1.7820×10^{-3}
Cluster	0.07964	3.8043×10^{-4}	0.07980	1.0260×10^{-3}

Table 9.5: Mean and standard deviation of the prior and posterior pdfs for the volume filling fraction of different structure types in the primordial large-scale structure ($a = 10^{-3}$).

region. In the unobserved parts, the entropy fluctuates around 1.6 Sh, value which characterizes the information content of the prior. Using equation (9.5) (unconditional on the data) and the numbers given in table 9.4, one can check that this number is consistent with the expectation. In the right panel of figure 9.6, we show a map of the Kullback-Leibler divergence of the posterior from the prior, which represents the information gain due to the data.

9.4.3 Volume and mass filling fractions

We computed the volume and mass filling fractions (defined by equations (9.7) and (9.8)) of different structure types in the primordial large-scale structure in the Sloan volume. As for the final conditions, we kept only the regions where the survey response operator is strictly positive. Consequently, we obtained the posterior pdfs $\mathcal{P}(\mathrm{VFF}(\mathrm{T}_i)|d)$ and $\mathcal{P}(\mathrm{MFF}(\mathrm{T}_i)|d)$. Using a set of unconstrained Gaussian random fields, we also measured $\mathcal{P}(\mathrm{VFF}(\mathrm{T}_i))$ and $\mathcal{P}(\mathrm{MFF}(\mathrm{T}_i))$ and found that all these pdfs are well described by Gaussians, the means and standard deviations of which are given in table 9.5 and 9.6.

All posterior quantities obtained are within two standard deviations of the corresponding prior means, and show smaller variance, as expected. Hence, all results obtained are consistent with Gaussian initial conditions.

9.5 Evolution of the cosmic web

In addition to the inference of initial and final density fields, BORG allows to simultaneously analyze the formation history and morphology of the observed large-scale structure, a subject that we refer to as chrono-

Structure type	μ_{MFF}	σ_{MFF}	μ_{MFF}	σ_{MFF}
	Primordial large-scale structure ($a = 10^{-3}$)			
	Posterior		Prior	
Void	0.07958	4.0122×10^{-4}	0.07941	1.0163×10^{-3}
Sheet	0.41933	6.1907×10^{-4}	0.41957	1.7912×10^{-3}
Filament	0.42110	6.3543×10^{-4}	0.42087	1.7785×10^{-3}
Cluster	0.07999	3.8206×10^{-4}	0.08015	1.0293×10^{-3}

Table 9.6: Mean and standard deviation of the prior and posterior pdfs for the mass filling fraction of different structure types in the primordial large-scale structure ($a = 10^{-3}$).

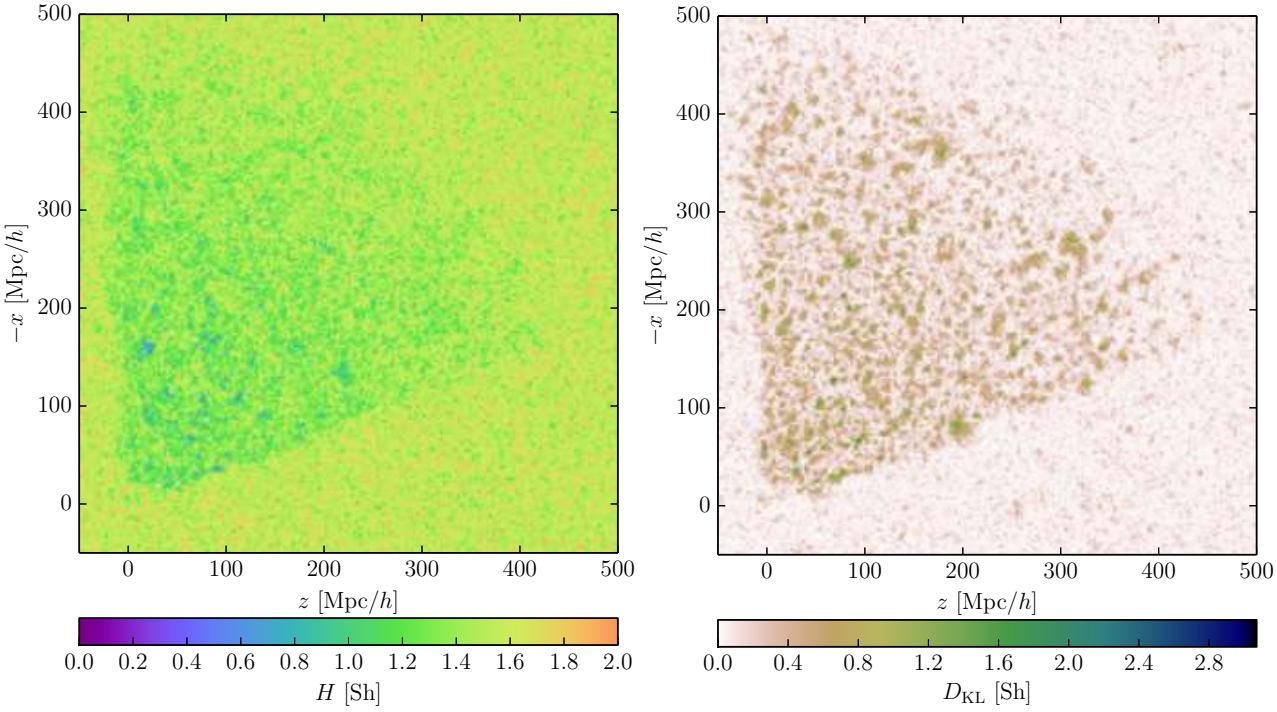

Figure 9.6: Slices through the entropy of the structure types posterior (left panel) and the Kullback-Leibler divergence of the posterior from the prior (right panel), in the initial conditions. The entropy H, defined by equation (9.5), quantifies the information content of the posterior pdf represented in figure 9.5, which results from fusing the information content of the prior and the data constraints. The Kullback-Leibler divergence $D_{\rm KL}$, defined by equation (9.6), represents the information gained in moving from the prior to the posterior. It quantifies the information that has been learned on structure types by looking at SDSS galaxies.

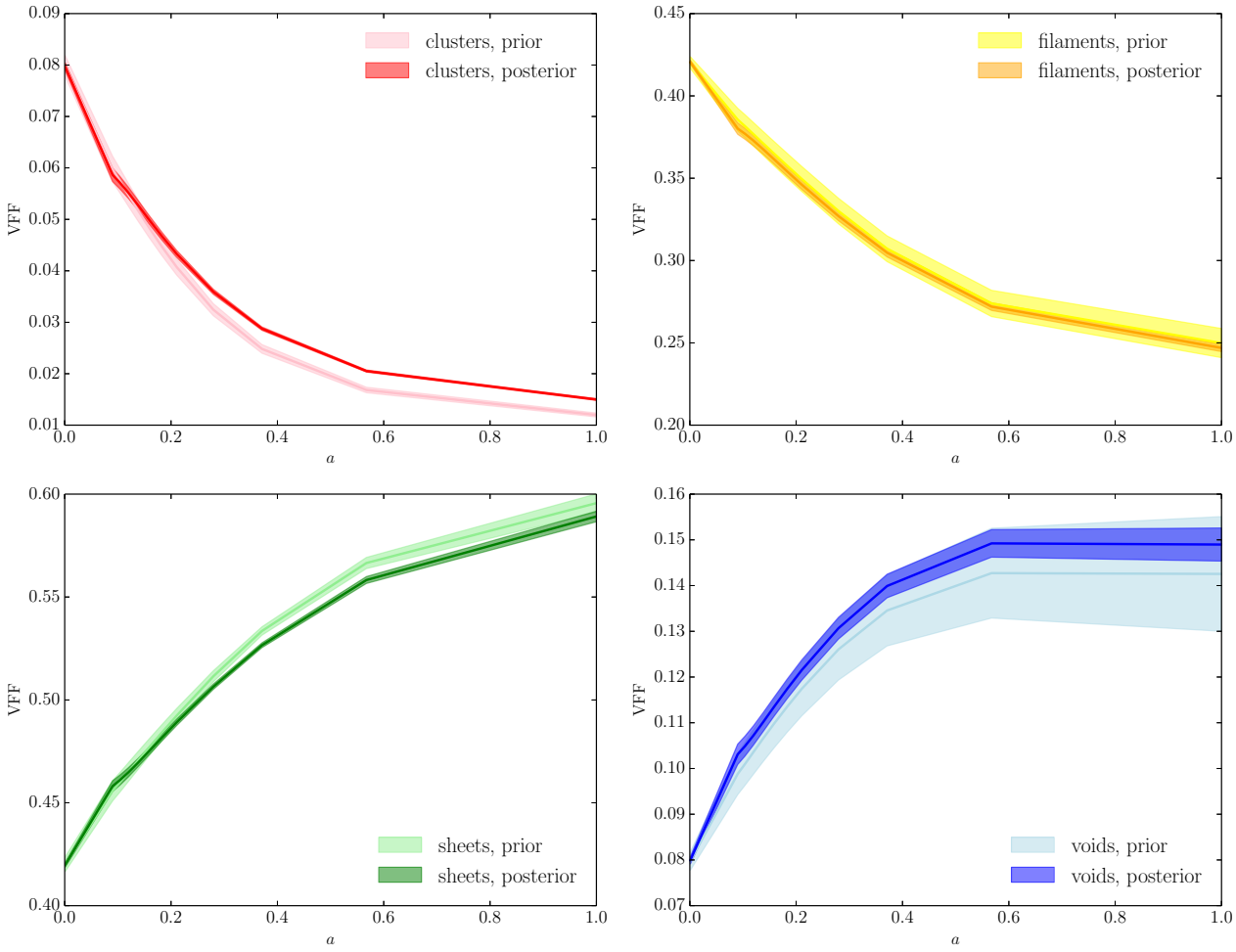

Figure 9.7: Time evolution of the volume filling fractions of different structure types (from left to right and top to bottom: clusters, filaments, sheets, voids). The solid lines show the pdf means and the shaded regions are the 2-σ credible intervals. Light colors are used for the priors and dark colors for the posteriors.

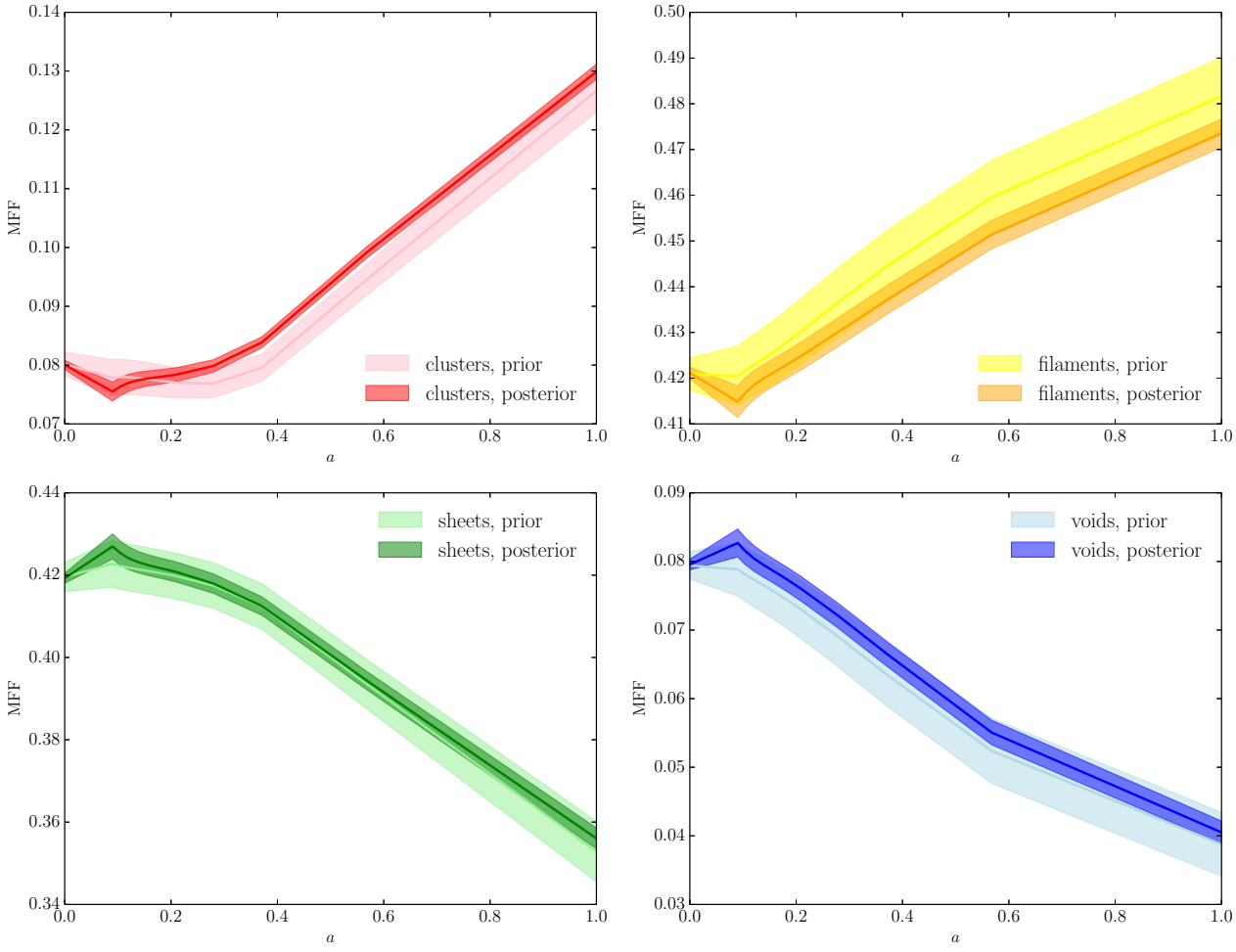

Figure 9.8: Same as figure 9.7 but for the mass filling fractions.

cosmography. In this section, we discuss the evolution of the cosmic web from its origin ($a = 10^{-3}$, analyzed in section 9.4) to the present epoch ($a = 1$, analyzed in section 9.3). To do so, we use 11 snapshots saved during the COLA filtering of our results (see section 9.2.2). These are linearly separated in redshift from $z = 10$ to $z = 0$. We perform this analysis in the 1,097 samples filtered with COLA considered in section 9.3. For each of these samples and for each redshift, we follow the procedure described in sections 9.2.2 and 9.2.3 to compute the density field and to classify the structure types.

9.5.1 Evolution of the probabilistic maps

We followed the time evolution of the probabilistic web-type maps from the primordial (figure 9.5) to the late-time large-scale structure (figure 9.2). In unconstrained regions, these maps show the evolution of the prior preference for specific structure types (see tables 9.1 and 9.4), in particular the breaking of the initial symmetry between voids and clusters and between sheets and filaments, discussed in section 9.4.2.

In data-constrained regions, the time evolution of web-type maps permits to visually check the expansion history of individual regions where the posterior probability of one specific structure is high. In particular, it is easy to see that, as expected from their dynamical definition, voids expand and clusters shrink in comoving coordinates, from $a = 10^{-3}$ to $a = 1$ (the reader is invited to compare the leftmost and rightmost panels of figures 9.2 and 9.5). Similarly, regions corresponding with high probability to sheets and filaments expand along two and one axis, respectively, and shrink along the others. This phenomenon is more difficult to see in slices, however, as the slicing plane intersects randomly the eigendirections of the tidal tensor.

The time evolution of maps of the web-type posterior entropy (absolute and relative to the prior) also exhibit some interesting features. There, it is possible to simultaneously check the increase of the information content of the prior (from $H \approx 1.6$ Sh to $H \approx 1.4$ Sh) and the displacement of observational information operated by the physical model. As the large-scale structure forms in the Sloan volume, data constraints are propagated and the complex structure of the final entropy map (figure 9.3), discussed in section 9.3.2, takes shape.

9.5.2 Volume filling fraction

Our ensemble of snapshots allows us to check the time evolution of global characterizations of the large-scale structure such as the volume and mass filling fractions of different structures. As in sections 9.3.3 and 9.4.3, we computed these quantities using only the volume where the survey response operator is non-zero. In figure 9.7, we plot these VFF as a function of the scale factor. There, the solid lines correspond to the pdf means and the shaded regions to the 2-σ credible intervals, with light colors for the priors and dark colors for the posteriors.

The time variation of the VFF in figure 9.7 is consistent with the expected dynamical behavior of structures. As voids and sheets expand along three and two axes, respectively, their volume fraction increases. Here, the posterior probabilities are mild updates of this prediction. Conversely, as clusters and filaments shrink along three and two axes, respectively, their volume fraction decreases. An explanation for the substantial displacement of the posterior from the prior, observed for clusters, can be found in section 9.3.3.

As already noted, the VFF is a very sensitive function of the precise definition of structures, grid size, density assignment scheme, smoothing scale, etc. For this reason, even for prior probabilities, our results can be in qualitative disagreement with previous authors (e.g. figure 23 in Cautun *et al.*, 2014), due to their very different definitions of structures. Therefore, we only found relevant to compare our posterior results with the prior predictions based on unconstrained realizations. The same remark applies to the MFF in the following section.

9.5.3 Mass filling fraction

In figure 9.8, we show the time evolution of the mass filling fractions using the same plotting conventions. Results are consistent with an interpretation based on large scale flows of matter. According to this picture, voids always loose mass while clusters always become more massive. The behavior of sheets and filaments can in principle be more complex, since these regions have both inflows and outflows of matter depending on the detail of their expansion profiles. In our setup, we found that the number of axes along which there is expansion dominates in the determination of the balance of inflow versus outflow, for global quantities such as the MFF. Therefore, filaments always gain mass and sheets always loose mass. Summing up our prior predictions, as they expand along at least two axes, matter flows out of voids and sheets and streams towards filaments and clusters.

The posterior probabilities slightly update this picture. Observations support smaller outflowing of matter from voids. For structures globally gaining matter, the priors are displaced towards less massive filaments and more massive clusters. All posterior predictions fall within the ~ 2-σ credible interval from corresponding prior means.

9.6 Summary and Conclusion

Along with chapter 8 (Leclercq *et al.*, 2015), this work exploits the high quality of inference results produced by the application of the Bayesian code BORG (chapter 4; Jasche & Wandelt, 2013a) to the Sloan Digital Sky Survey main galaxy sample (chapter 5; Jasche, Leclercq & Wandelt, 2015). We presented a Bayesian cosmic web analysis of the nearby Universe probed by the northern cap of the SDSS and its surrounding. In doing so, we produced the first probabilistic, four-dimensional maps of dynamic structure types in real observations.

As described in section 9.2.1, our method relies on the physical inference of the initial density field in the LSS (Jasche & Wandelt, 2013a; Jasche, Leclercq & Wandelt, 2015). Starting from these, we generated a large set of data-constrained realizations using the fast COLA method (section 9.2.2). The use of 2LPT as a physical model in the inference process and of the fully non-linear gravitational dynamics, provided by COLA, as a filter allowed us to describe structures at the required statistical accuracy, by very well representing the full hierarchy of correlation functions. Even though initial conditions were inferred with the approximate 2LPT model, we checked that the clustering statistics of constrained non-linear model evaluations agree with theoretical expectations up to scales considered in this work. As described in section 9.2.3, we used the dynamic web-type classification algorithm proposed by Hahn *et al.* (2007a) to dissect the cosmic web into voids, sheets, filaments, and clusters.

In sections 9.3 and 9.4, we presented the resulting maps of structures in the final and initial conditions, respectively, and studied the distribution of global quantities such as volume fraction and mass filling fractions. In section 9.5, we further analyzed the time evolution of our results, in a rigorous chrono-cosmographic framework.

For all results presented in this chapter, we demonstrated a thorough capability of uncertainty quantification. Specifically, for all inferred maps and derived quantities, we got a probabilistic answer in terms of a prior and a posterior distribution. The variation between samples of the posterior distribution quantifies the remaining uncertainties of various origins (in particular noise, selection effects, survey geometry and galaxy bias, see chapters 4 and 5 for a detailed discussion). Building upon our accurate probabilistic treatment, we looked at the entropy of the structure type posterior and at the relative entropy between posterior and prior. In doing so, we quantified the information gain due to SDSS galaxy data with respect to the underlying dynamic cosmic web and analyzed how this information is propagated during cosmic history. This study constitutes the first link between cosmology and information theory using real data.

In summary, our methodology yields an accurate cosmographic description of web types in the non-linear regime of structure formation, permits to analyze their time evolution and allows a precise uncertainty quantification in a full-scale Bayesian framework. These inference results can be used for a rich variety of applications, ranging from studying galaxies inside their environment to cross-correlating with other cosmological probes. They count among the first steps towards accurate chrono-cosmography, the subject of simultaneously analyzing the morphology and formation history of the inhomogeneous Universe.

Note added: As we were finalizing the paper corresponding to this chapter (Leclercq, Jasche & Wandelt, 2015c) for submission, the works by Zhao *et al.* (2015) and Shi, Wang & Mo (2015) appeared where the relationship between halos and the cosmic web environment defined by the tidal tensor is being studied.

Chapter 10

Cosmic-web type classification using decision theory

Contents

> "If no mistake have you made, yet losing you are... a different game you should play."
> Master Yoda, in recollections of Mace Windu,
> — Matthew Stover (2003), *Star Wars: Shatterpoint*

Abstract

We propose a decision criterion for segmenting the cosmic web into different structure types (voids, sheets, filaments, and clusters) on the basis of their respective probabilities and the strength of data constraints. Our approach is inspired by an analysis of games of chance where the gambler only plays if a positive expected net gain can be achieved based on some degree of privileged information. The result is a general solution for classification problems in the face of uncertainty, including the option of not committing to a class for a candidate object. As an illustration, we produce high-resolution maps of web-type constituents in the nearby Universe as probed by the Sloan Digital Sky Survey main galaxy sample. Other possible applications include the selection and labeling of objects in catalogs derived from astronomical survey data.

This chapter is adapted from its corresponding publication, Leclercq, Jasche & Wandelt (2015a).

10.1 Introduction

Building accurate maps of the cosmic web from galaxy surveys is one of the most challenging tasks in modern cosmology. Rapid progress in this field took place in the last few years with the introduction of inference techniques based on Bayesian probability theory (Kitaura *et al.*, 2009; Jasche *et al.*, 2010b; Nuza *et al.*, 2014; Jasche, Leclercq & Wandelt, 2015). This facilitates the connection between the properties of the cosmic web, thoroughly analyzed in simulations (e.g. Hahn *et al.*, 2007a; Aragón-Calvo, van de Weygaert & Jones, 2010; Cautun *et al.*, 2014), and observations (see chapter 3 and Leclercq, Pisani & Wandelt, 2014, for a review on the interface between theory and data in cosmology).

In chapter 9 (Leclercq, Jasche & Wandelt, 2015c), we conducted a fully probabilistic analysis of structure types in the cosmic web as probed by the Sloan Digital Sky Survey main galaxy sample. This study capitalized on the large-scale structure inference performed by Jasche, Leclercq & Wandelt (2015, chapter 5) using the BORG (Bayesian Origin Reconstruction from Galaxies, Jasche & Wandelt, 2013a, chapter 4) algorithm. As the full gravitational model of structure formation COLA (COmoving Lagrangian Acceleration, Tassev, Zaldarriaga & Eisenstein, 2013; see also section 7.3.1) was used, our approach resulted in the first probabilistic and time-dependent classification of cosmic environments at non-linear scales in physical realizations of the large-scale

structure conducted with real data. Using the Hahn *et al.* (2007a) definition (appendix C.2, see also its extensions, Forero-Romero *et al.*, 2009; Hoffman *et al.*, 2012), we obtained three-dimensional, time-dependent maps of the posterior probability for each voxel to belong to a void, sheet, filament or cluster.

These posterior probabilities represent all the available structure type information in the observational data assuming the framework of ΛCDM cosmology. Since the large-scale structure cannot be uniquely determined from observations, uncertainty remains about how to assign each voxel to a particular structure type. The question we address in this chapter is how to proceed from the posterior probabilities to a particular choice of assigning a structure type to each voxel. Decision theory (see, for example, Berger, 1985) offers a way forward, since it addresses the general problem of how to choose between different actions under uncertainty. A key ingredient beyond the posterior is the utility function that assigns a quantitative profit to different actions for all possible outcomes of the uncertain quantity. The optimal decision is that which maximizes the expected utility.

After setting up the problem using our example and briefly recalling the relevant notions of Bayesian decision theory, we will discuss different utility functions and explore the results based on a particular choice.

10.2 Method

The decision problem for structure-type classification can be stated as follows. We have four different web-types that constitute the "space of input features:" $\{\mathrm{T}_0 = \text{void}, \mathrm{T}_1 = \text{sheet}, \mathrm{T}_2 = \text{filament}, \mathrm{T}_3 = \text{cluster}\}$. We want to either choose one of them, or remain undecided if the data constraints are not sufficient. Therefore our "space of actions" consists of five different elements: $\{a_0 =$ "decide void," $a_1 =$ "decide sheet," $a_2 =$ "decide filament," $a_3 =$ "decide cluster," and $a_{-1} =$ "do not decide."$\}$ The goal is to write down a decision rule prescribing which action to take based on the posterior information.

Bayesian decision theory states that the action a_j that should be taken is that which maximizes the expected utility function (conditional on the data d), given in this example by

$$U(a_j(\vec{x}_k)|d) = \sum_{i=0}^{3} G(a_j|\mathrm{T}_i)\, \mathcal{P}(\mathrm{T}_i(\vec{x}_k)|d), \tag{10.1}$$

where $\vec{x}_k$ labels one voxel of the considered domain, $\mathcal{P}(\mathrm{T}_i(\vec{x}_k)|d)$ are the posterior probabilities of the different structure types given the data, and $G(a_j|\mathrm{T}_i)$ are the gain functions that state the profitability of each action, given the "true" underlying structure. Formally, G is a mapping from the space of input features to the space of actions. For our particular problem, it can be thought of as a 5×4 matrix $\mathbf{G}$ such that $\mathbf{G}_{ij} \equiv G(a_j|\mathrm{T}_i)$, in which case eq. (10.1) can be rewritten as a linear algebra equation, $\mathbf{U} = \mathbf{G}.\mathbf{P}$ where the 5-vector $\mathbf{U}$ and the 4-vector $\mathbf{P}$ contain the elements $\mathbf{U}_j \equiv U(a_j(\vec{x}_k)|d)$ and $\mathbf{P}_i \equiv \mathcal{P}(\mathrm{T}_i(\vec{x}_k)|d)$, respectively.

Let us consider the choice of gain functions. Several choices are possible. For example, the 0/1-gain functions reward a correct decision with 1 for each voxel, while an incorrect decision yields 0. This leads to choosing the structure type with the highest posterior probability. While this seems like a reasonable choice, we need to consider that a decision is made in each voxel, whereas we are interested in identifying structures as objects that are made of many voxels. For instance, since clusters are far smaller than voids, the *a priori* probability for a voxel to belong to a cluster is much smaller than for the same voxel to belong to a void. To treat different structures on an equal footing, it makes sense to reward the correct choice of structure type T_i by an amount inversely proportional to the average volume V_i of one such structure. In the following, we use the prior probability as a proxy for the volume fractions,

$$\mathcal{P}(\mathrm{T}_i) \approx \frac{V_i}{V_0 + V_1 + V_2 + V_3}. \tag{10.2}$$

We further introduce an overall cost for choosing a structure with respect to remaining undecided, leading to the following specification of the utility,

$$G(a_j|\mathrm{T}_i) = \begin{cases} \dfrac{1}{\mathcal{P}(\mathrm{T}_i)} - \alpha & \text{if } j \in [\![0,3]\!] \text{ and } i = j, \\ -\alpha & \text{if } j \in [\![0,3]\!] \text{ and } i \neq j, \\ 0 & \text{if } j = -1. \end{cases} \tag{10.3}$$

This choice limits 20 free functions to only one free parameter, α. With this set of gain functions, making (or not) a decision between structure types can be thought of as choosing to play or not to play a gambling game costing α. Not playing the game, i.e. remaining undecided ($j = -1$), is always free ($G(a_{-1}|\mathrm{T}_i) = 0$ for all i). If the gambler decides to play the game, i.e. to make a decision ($j \in [\![0, 3]\!]$), they pay α but may win a reward, $\frac{1}{\mathcal{P}(\mathrm{T}_i)}$, by betting on the correct underlying structure ($i = j$).

In the absence of data, the posterior probabilities in equation (10.1) are the prior probabilities $\mathcal{P}(\mathrm{T}_i)$, which are independent of the position $\vec{x}_k$, and the utility functions are, for $j \in [\![0, 3]\!]$,

$$\begin{aligned} U(a_j) &= \sum_{i=0}^{3} G(a_j|\mathrm{T}_i)\,\mathcal{P}(\mathrm{T}_i) \\ &= \left(\frac{1}{\mathcal{P}(\mathrm{T}_j)} - \alpha\right)\mathcal{P}(\mathrm{T}_j) - \sum_{\substack{i=0 \\ i\neq j}}^{3} \alpha\,\mathcal{P}(\mathrm{T}_i) \\ &= 1 - \alpha\left(\mathcal{P}(\mathrm{T}_j) + \sum_{\substack{i=0 \\ i\neq j}}^{3} \mathcal{P}(\mathrm{T}_i)\right) \\ &= 1 - \alpha, \qquad (10.4) \\ \text{and} \quad U(a_{-1}) &= 0. \qquad (10.5) \end{aligned}$$

Equations (10.4) and (10.5) mean that, in the absence of data, this reduces to the roulette game utility function, where, if correctly guessed, *a priori* unlikely outcomes receive a higher reward, inversely proportional to the fraction of the probability space they occupy. Betting on outcomes according to the prior probability while paying $\alpha = 1$ leads to a *fair game* with zero expected net gain. The gambler will always choose to play if the cost per game is $\alpha \leq 1$ and will never play if $\alpha > 1$.

The posterior probabilities update the prior information in light of the data, providing an advantage to the gambler through privileged information about the outcome. In the presence of informative data, betting on outcomes based on the posterior probabilities will therefore ensure a positive expected net gain and the gambler will choose to play even if $\alpha > 1$. Increasing the parameter α therefore represents a growing *aversion for risk* and limits the probability of losing. Indeed, for high α, the gambler will only play in cases where the posterior probabilities give sufficient confidence that the game will be won, i.e. that the decision will be correct.

10.3 Maps of structure types in the SDSS

We applied the above decision rule to the web-type posterior probabilities presented in chapter 9 (Leclercq, Jasche & Wandelt, 2015c), for different values of $\alpha \geq 1$ as defined by equation (10.3). In doing so, we produced various maps of the volume of interest, consisting of the northern Galactic cap of the SDSS main galaxy sample and its surroundings. Slices through these three-dimensional maps are shown in figure 10.1 for the late-time large-scale structure (at $a = 1$) and in figure 10.2 for the primordial large-scale structure (at $a = 10^{-3}$).

When the game is fair (namely when $\alpha = 1$), it is always played, i.e. a decision between one of the four structure types is always made. This results in the *speculative map* of structure types (top left panel of figures 10.1 and 10.2). There, a decision is made even in regions that are not constrained by the data (at high redshift or outside of the survey boundaries), based on prior betting odds.

By increasing the value of $\alpha > 1$, we demand higher confidence in making the correct decision. This yields increasingly *conservative maps* of the Sloan volume (see figures 10.1 and 10.2). In particular, at high values of α, the algorithm makes decisions in the regions where data constraints are strong (see figures 9.3 and 9.6), but often stays undecided in the unobserved regions. It can be observed that even at very high values, $\alpha \gtrsim 3$, a decision for one structure is made in some unconstrained voxels (typically in favor of the structure for which the reward is the highest: clusters in the final conditions, and clusters or voids in the initial conditions). This effect is caused by the limited number of samples used in our analysis. Indeed, because of the finite length of the Markov Chain, the sampled representation of the posterior has not yet fully converged to the true posterior. For this reason, the numerical representation of the posterior can be artificially displaced too much from the prior, which results in an incorrect web-type decision. This effect could be mitigated by obtaining more samples in the original BORG analysis (for an increased computational cost); or can be avoided by further increasing α,

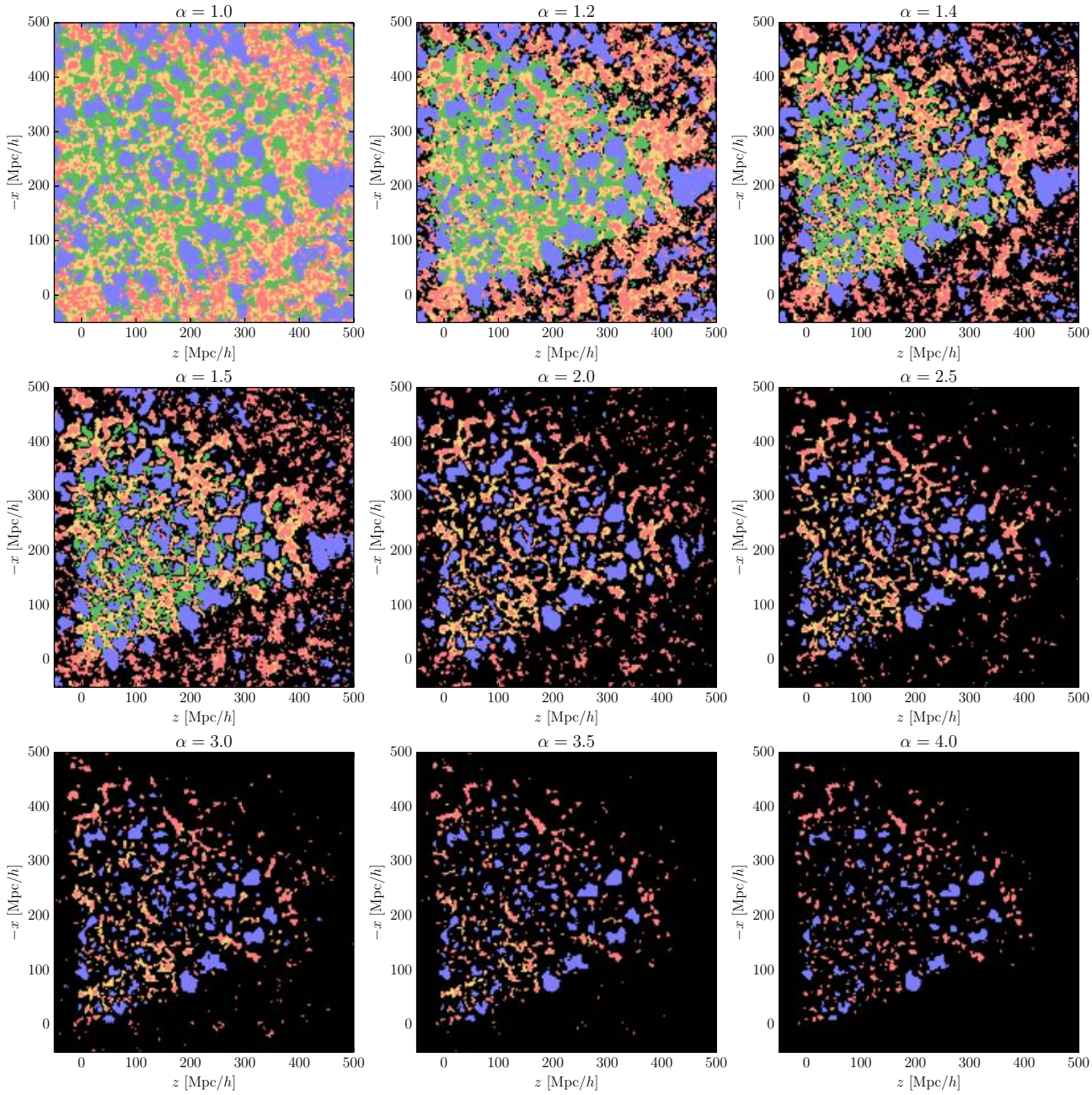

Figure 10.1: Slices through maps of structure types in the late-time large-scale structure, at $a = 1$. The color coding is blue for voids, green for sheets, yellow for filaments, and red for clusters. Black corresponds to regions where data constraints are insufficient to make a decision. The parameter α, defined by equation (10.3), quantifies to the risk aversion in the map: $\alpha = 1.0$ corresponds to the most speculative map of the large-scale structure, and maps with $\alpha \geq 1$ are increasingly conservative. These maps are based on the posterior probabilities inferred in chapter 9 and on the Bayesian decision rule subject of the present chapter.

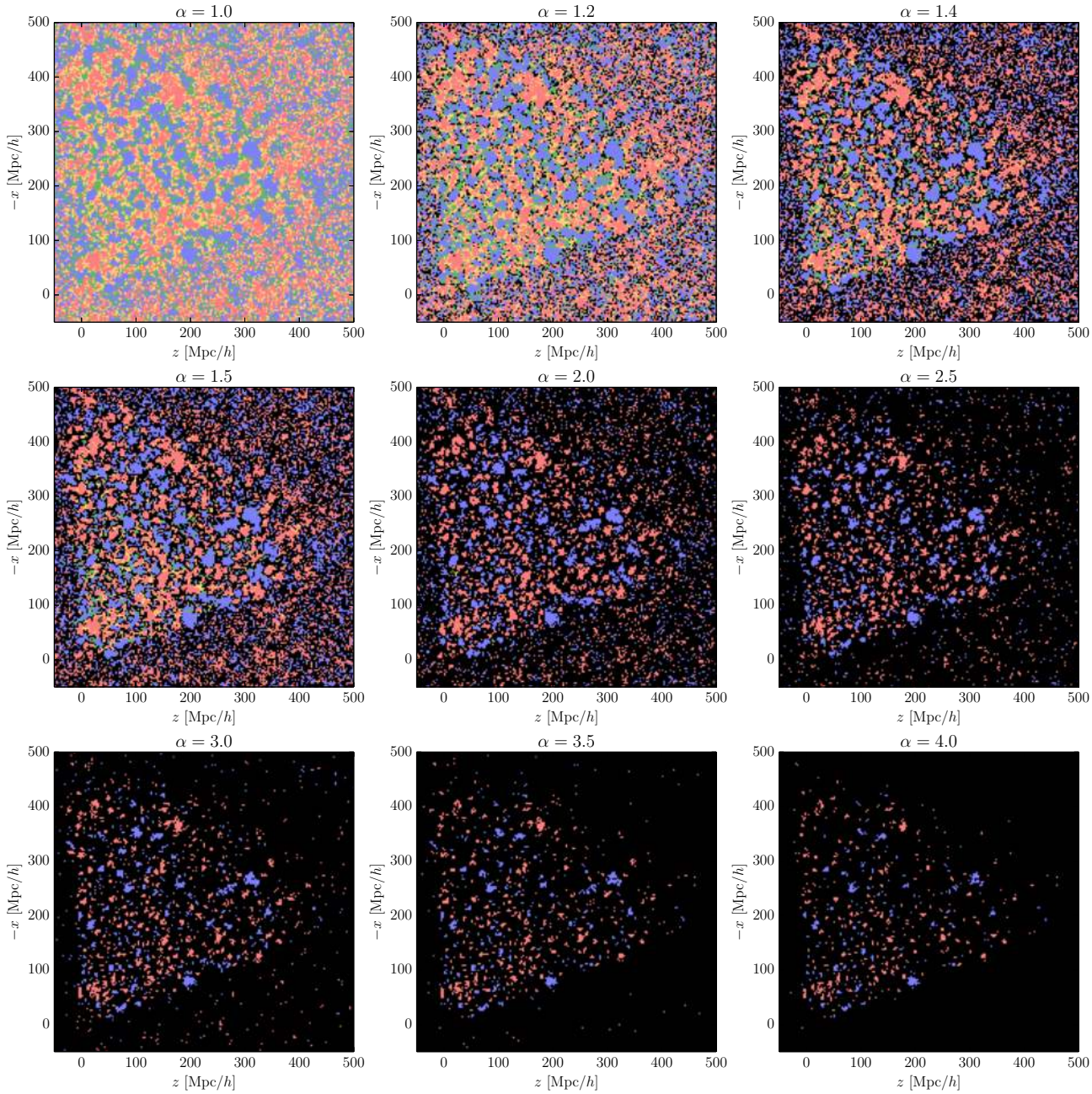

Figure 10.2: Same as figure 10.1 for the primordial large-scale structure, at $a = 10^{-3}$.

at the expense of also degrading the map in the observed regions. We found the value of $\alpha = 4$ (bottom right panel of figures 10.1 and 10.2) to be the best compromise between reducing the number of unobserved voxels in which a decision is made to a tiny fraction and keeping information in the volume covered by the data.

As expected, structures for which the prior probabilities are the highest disappear first from the map when one increases α: betting on these structures being poorly rewarded, this choice is avoided in case of high risk aversion. In the final conditions (figure 10.1), we found that sheets completely disappear for $\alpha \approx 1.68$ and filaments for $\alpha \approx 4.01$. In the initial conditions (figure 10.2), the critical value is around $\alpha \approx 2.36$ for both sheets and filaments. In the most conservative maps displayed in figures 10.1 and 10.2 ($\alpha = 4.0$), the SDSS data provide extremely high evidence for the voids and clusters shown. In constrained parts, extended regions belonging to a given structure type may not have the expected shape. This is true in particular for filamentary regions. Several factors can explain this: first, slicing through filaments make them appear as dots; second, with the dynamic T-web definition, filament regions often extend out into sheets and voids, and their static skeleton geometry is not the most prominent at the voxel scale (3 Mpc/h in this work).

As detailed in chapters 4 and 5, data constraints are propagated by the structure formation model assumed in the inference process (second-order Lagrangian perturbation theory) and therefore radiate out of the SDSS boundaries. For this reason, for moderate values of α, web-type classification can be extended beyond the survey boundaries to regions influenced by data. This can be observed in figures 10.1 and 10.2, where one can see, for instance, that the shape of voids that intersect the mask is correctly recovered. Similarly, the classification of high-redshift structures confirms that the treatment of selection effects by BORG is correctly propagated to web-type analysis.

We finally comment on the required computational resources for the complete chain for running BORG, computing the web-type posterior, and making a decision. Inference with BORG is the most expensive part: on average, one sample is generated in 1500 seconds on 16 cores (chapter 5; Jasche, Leclercq & Wandelt, 2015). Then, in each sample, tidal shear analysis (chapter 9; Leclercq, Jasche & Wandelt, 2015c) is a matter of a few seconds. Once the web-type posterior is known, making a decision, which is the subject of the present chapter, is almost instantaneous. Therefore, once the density field has been inferred, which is useful for a much larger variety of applications, our method is substantially cheaper than several state-of-the-art techniques for cosmic web analysis (e.g. the method of Tempel, Stoica & Saar, 2013; Tempel *et al.*, 2014, for detecting filaments).

10.4 Conclusions

In this chapter, we proposed a rule for optimal decision making in the context of cosmic web classification. We described the problem set-up in Bayesian decision theory and proposed a set of gain functions that permit an interpretation of the problem in the context of game theory. This framework enables the dissection of the cosmic web into different elements (voids, sheets, filaments, and clusters) given their prior and posterior probabilities and naturally accounts for the strength of data constraints.

As an illustration, we produced three-dimensional templates of structure types with various risk aversion, describing a volume covered by the SDSS main galaxy sample and its surrounding. These maps constitute an efficient statistical summary of the inference results presented in chapter 9 (Leclercq, Jasche & Wandelt, 2015c) for cross-use with other astrophysical and cosmological data sets.

Beyond this specific application, our approach is more generally relevant to the solution of classification problems in the face of uncertainty. For example, the construction of catalogs from astronomical surveys is directly analogous to the problem we describe here: it simultaneously involves a decision about whether or not to include a candidate object and which class label (e.g. star or galaxy) to assign to it.