Introduction

> "Make your choice, adventurous Stranger,
> Strike the bell and bide the danger,
> Or wonder, till it drives you mad,
> What would have followed if you had."
> — Clive Staples Lewis (1955), *The Chronicles of Narnia, The Magician's Nephew*

Large-scale structure surveys during the age of precision cosmology

Understanding the structure of the Universe at the largest scales is one of the main goals of cosmology. The existence of such a structure has been suggested by early observational projects aimed at mapping the distribution of galaxies, which resulted in a number of discoveries of individual elements – filamentary bridges between superclusters, and large voids – on scales of tens of megaparsecs (Gregory & Thompson, 1978; Gregory, Thompson & Tifft, 1981; Kirshner *et al.*, 1981; Zel'dovich, Einasto & Shandarin, 1982). In 1986, the results of the Center for Astrophysics redshift survey marked a milestone, with the discovery of bubble-like structures separated by sheets on which galaxies tend to lie (de Lapparent, Geller & Huchra, 1986). These results renewed interest for large-scale structure cartography, leading to new galaxy catalogs up to a depth of $\sim$ 400 Mpc (Geller & Huchra, 1989; Shectman *et al.*, 1996; Vettolani *et al.*, 1997; Schuecker & Ott, 1991). In spite of their incompleteness, these maps conclusively confirmed the existence of a large-scale organization of galaxies into a hierarchical structure, the *cosmic web*. At the turn of the century, massive surveys, aimed at obtaining the spectra of hundreds of thousands of galaxies (e.g. the 2dFGRS, Colless *et al.*, 2003; the SDSS, Strauss *et al.*, 2002; WiggleZ, Drinkwater *et al.*, 2010 or the 2MASS redshift survey, Huchra *et al.*, 2012), mapped large volumes of the nearby Universe. They allowed to largely increase the completeness of observations and to obtain large enough samples for statistical analyses. Other observational programs (e.g. DEEP2, Davis *et al.*, 2003, 2007; VVDS, Le Fèvre *et al.*, 2005, 2013; zCOSMOS, Lilly *et al.*, 2007; GAMA, Driver *et al.*, 2009; VIPERS, Guzzo *et al.*, 2014) focused on targeting galaxies in a smaller area on the sky, but at higher redshift.

In the coming decade, ongoing or planned cosmological programs will measure the distribution of galaxies at an unprecedented level. These include wide photometric surveys (DES, Dark Energy Survey Collaboration, 2005; HSC, Miyazaki *et al.*, 2012; J-PAS, Benítez *et al.*, 2015; LSST, LSST Science Collaboration, 2009, 2012), deep spectroscopic surveys (eBOSS; HETDEX, Hill *et al.*, 2008; the Subaru Prime Focus Spectrograph, Takada *et al.*, 2014; DESI, Schlegel *et al.*, 2011; Abdalla *et al.*, 2012; Levi *et al.*, 2013), and the Euclid (**?**Amendola *et al.*, 2013) and WFIRST (Green *et al.*, 2012; Spergel *et al.*, 2013) satellites.

How do we compare this avalanche of data to cosmological models? The standard picture of LSS formation, developed over the last three decades, relies of the gravitational self-evolution of a set of initial density fluctuations, giving rise to the complex structures observed in galaxy surveys. Extracting the wealth of information that surveys contain thus requires a quantitative understanding of both the generation of the initial seed perturbations and of the dynamics of gravitational instability.

Early Universe physics and generation of the initial conditions of the Universe

Inflation and the Hot Big Bang scenario provide an observationally well-supported physical model for the initial conditions. The inflationary paradigm (see e.g. Baumann, 2011, for a review) is generally favored over other theories for the origin of seed perturbations, since it also provides explanations for some shortcomings of the standard Hot Big Bang picture, e.g. the statistical homogeneity and isotropy of the Universe, and the horizon problem (Guth, 1981; Linde, 1982; Albrecht & Steinhardt, 1982). According to this model, during the inflationary era, the equation of state of the Universe is governed by a potential-dominated quantum scalar field with negative pressure, the so-called *inflaton field*. This quantum field drives an exponential growth of the cosmic scale factor. What is remarkable with inflation is that the accelerated expansion in the very early Universe can magnify the vacuum quantum fluctuations of the inflaton into macroscopic cosmological perturbations.

This model naturally provides us with a statistically homogeneous and isotropic density field with small, very nearly Gaussian-distributed, and nearly scale-invariant density perturbations (Guth & Pi, 1982; Hawking, 1982; Starobinsky, 1982; Bardeen, Steinhardt & Turner, 1983).

Phenomena such as primordial nucleosynthesis (Alpher, Bethe & Gamow, 1948), decoupling and recombination, free-streaming of neutrinos, acoustic oscillations of the photon-baryon plasma, and transition from radiation to matter domination, come next. They are predicted by the Hot Big Bang model, which remains a cornerstone of our understanding of the past and present Universe (see e.g. Kolb & Turner, 1990; Peacock, 1999). They change the post-inflationary density field into what we call the "initial conditions" for gravitational evolution. Then, during the matter and dark-energy dominated eras, self-gravity and the expansion of the Universe modify these initial conditions into an evolved density field, at first through linear transfer and then through non-linear structure formation.

Due to their quantum origin, the process of generating seed perturbations is stochastic (see e.g. Baumann, 2011, section 2.3). Therefore, a probability distribution function is the most fundamental characterization of the large-scale structure of the Universe. As a consequence, it is now standard to describe in a probabilistic way the generation of the initial density fluctuations by the above-mentioned early Universe processes.

Large-scale structure evolution and galaxy formation

According to the current picture of cosmic structure formation, all presently observed structures have their origins in primordial seed fluctuations. Zel'dovich & Novikov (1983) recognized the central role played by gravitational instability. Peebles (1982a, 1984) realized that baryonic models of structure formation are insufficient to explain observed galaxies morphology and distribution, and consequently proposed the introduction of cold dark matter. The ensuing controversy between the "top-down" (in which large structures form first, then fragment; as is the case when hot dark matter, such as neutrinos, dominates) and "bottom-up" (in which small structures such as galaxies form first, then aggregate; as is the case when cold dark matter dominates) structure formation scenarios was subsequently settled in favor of the latter (Bond, Szalay & Turner, 1982; Melott *et al.*, 1983; Blumenthal *et al.*, 1984). Therefore, it is currently believed that structure formation is mostly governed by the gravitational aggregation of a dark matter fluid. As proposed by Rees & Ostriker (1977); Silk (1977); White & Rees (1978), luminous objects such as galaxies form via condensation and cooling of baryonic matter in gravitational potential wells shaped by the dark matter structure.

Physical processes and information content

The detailed appearance of the presently observed galaxy distribution contains a record of its origin and formation history. Large-scale structure formation therefore encodes information on a wide range of processes involving very different physics, ranging from quantum field theory and general relativity, to the dynamics of collisionless dark matter and the hydrodynamics and radiation transfer processes involved in galaxy formation. The next generation of galaxy surveys is therefore expected to provide insights into many fundamental physics questions: What is the Universe made of? What is the microphysics of dark matter? How does dark energy behave? What is the mass of neutrinos? Is general relativity complete or does it require modifications? What were the conditions in the early Universe?

All LSS observations are informative in some ways about these questions, but due to an incomplete understanding of the dark matter-galaxy connection (the "bias problem": see in particular the "peak-background split model", Bardeen *et al.*, 1986; Cole & Kaiser, 1989, and the "halo model", Seljak, 2000; Peacock & Smith, 2000; Cooray & Sheth, 2002) and observational effects (the Alcock-Paczynski effect, Alcock & Paczynski, 1979; redshift-space distortions, Kaiser, 1987; Peacock *et al.*, 2001; Hawkins *et al.*, 2003; Guzzo *et al.*, 2008; non-trivial selection functions; see e.g. Percival, 2014, for a review), the message is encoded and sometimes hard to extricate. Hence, crucial to the aim of answering the above questions is identifying where is the information content and developing efficient tools to extract it.

The usual strategy is to look at the shape and length scales imprinted in the galaxy power spectrum, such as the baryon acoustic oscillation scale (Percival *et al.*, 2001; Cole *et al.*, 2005; Eisenstein *et al.*, 2005; Percival *et al.*, 2010). However, at small scales and at late times, non-linear dynamics shifts the information content away from the two-point function to the higher-order correlators. One of the main goals of this thesis is to access the untapped information in late-time, non-linear modes. The number of modes accessible for cosmological

analyses grows like k^3, where k is the largest usable wavenumber. In the case of BAOs, a technique known as "reconstruction" has been designed to correct for the effects of non-linearities, and has been shown to improve distance measurements (Eisenstein *et al.*, 2007; Padmanabhan *et al.*, 2012). Hence, our strategy is twofold: pushing down the smallest scale that can be both modeled and resolved; and inferring the initial conditions that give rise to the observed LSS. The reward for undertaking this project is a potentially vast gain of information for the determination of model parameters.

The scientific method and the process of data assimilation

Generally, contact between theory and observations cannot be established directly. Historically, scientific progress relied on experimental assessment (the "first paradigm" of science) and theoretical modeling (the "second paradigm"). In the last few decades, with growing complexity of the real-world processes to be described, testable predictions of theories often had to be obtained through numerical simulations of phenomena. Additionally, even elaborate experiments do not allow for direct comparison to the results of simulations, due to the fact that there exists no ideal observation in reality, as they are subject to a variety of uncertainties. One has to model the response functions of devices and treat their outputs, a step we refer to as signal processing (for example, the representation of a real-world signal and the application of fast Fourier transforms require its computer representation to be discrete both in configuration and in frequency space). Numerical simulations and data processing constitute the "third paradigm" of science. Their outputs are compared to judge and evaluate current models. These results can be used to perform inferences, i.e. update our knowledge on theoretical parameters, test hypotheses and compare competing models. They can also be used to optimize the design of new experiments. These last two steps close loops that go from theory to data, and from data to theory, as illustrated in figure 1. Scientific progress in any of the physical sciences crucially depends on these steps.

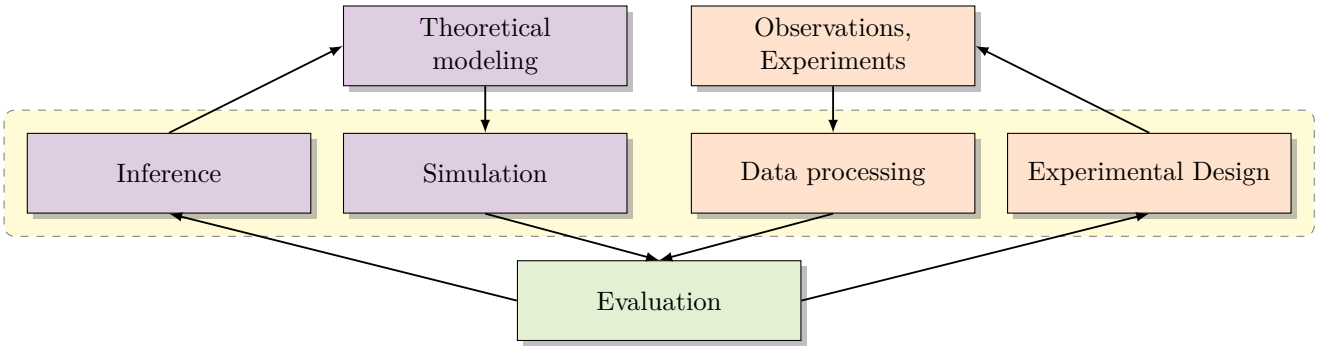

Figure 1: Diagram illustrating steps in the scientific method. Progress in physical sciences depends on each of these steps: experimental assessment (the first paradigm), theoretical modeling (the second paradigm), computational studies (simulation of phenomena and data processing – the third paradigm). The outputs of simulations and data processing are compared to judge and evaluate current models. These results are used to infer theoretical parameters and to design new experiments. The yellow rectangle shows the emergence of a fourth research paradigm: data-intensive scientific discovery, where extremely large data sets captured by instruments and generated by complex simulations are used.

Several authors are now describing the emergence of a so-called "fourth paradigm" of research: data-intensive scientific discovery (Microsoft Research, 2009; Szalay, 2014). Scientific insights are wrested from extremely vast data sets. This transition from hypothesis-driven to data-driven research is made possible by new technologies for gathering, processing, manipulating, analyzing, mining, and displaying data. For example, exascale computers, expected around 2018, will be of the order of processing power of the human brain at the neural level.

This thesis falls within the context of this emerging fourth paradigm. Its field is cosmostatistics, the discipline that uses stochastic quantities as seeds of structure to make the connection between cosmological models and observations. This area is at the interface between theory and observational data (see Leclercq, Pisani & Wandelt, 2014):

- It consists of predicting cosmological observables from stochastic quantities as seeds of structure in the Universe (*from theory to data*). Theoretical hypotheses are used to model, predict and anticipate results.

- It uses the departures from homogeneity and isotropy, observed in astronomical surveys, to distinguish between cosmological models (*from data to theory*). Data sets are used to infer parameters of the theoretical models and to compare their relative suitability.

More specifically, this work focuses on the process of *data assimilation* for the large-scale structure, i.e. the process by which observations are incorporated into a numerical model of a real system. Data assimilation is a field of statistics, widely understood and used outside of astrophysics (e.g. in meteorology, geophysics, oceanography and climate sciences). The general mechanism consists of iteratively correcting the state of a prediction based on a physical model, using successive observations. In this work, we borrow ideas from these sciences and apply them to large-scale structure data analysis. For all quantities of interest, we do not only provide a best-guess estimate, but fully account for all credible regions by a detailed quantification of the probability density.

Goal and structure of this thesis

The ambition of this work is to describe progress towards enriching the standard for the analysis of galaxy surveys. The central ingredient is the recently proposed BORG (Bayesian Origin Reconstruction from Galaxies, Jasche & Wandelt, 2013a; Jasche, Leclercq & Wandelt, 2015) algorithm. BORG is a Bayesian data assimilation code, which provides a fully probabilistic, physical model of the non-linearly evolved density field as probed by LSS surveys.

The goal of this thesis is to demonstrate that Bayesian large-scale structure inference with the BORG algorithm has moved beyond the proof-of-concept stage. In particular, it describes the first application to real cosmological data from the Sloan Digital Sky Survey, and shows how these results can be used for cosmic web classification and analysis.

This thesis is organized as follows. Part I focuses on the analytical and numerical description of the morphology and growth of the LSS. Chapter 1 is an introduction to the theory of structure formation, as relevant for this thesis. As Lagrangian perturbation theory is a key ingredient of the BORG algorithm, the reliability of its numerical predictions is investigated in chapter 2.

Part II introduces Bayesian large-scale structure inference. In chapter 3, we present the framework of Bayesian probability theory. Chapter 4 reviews the latest version of the BORG algorithm and its implementation. Chapter 5 presents the application of BORG to the Sloan Digital Sky Survey data. These results quantify the distribution of initial conditions as well as the possible formation history of the observed structures.

As LSS surveys contain a wealth of information that cannot be trivially extracted due to the non-linear dynamical evolution of the density field, part III discusses methods designed to improve upon standard techniques by including non-Gaussian and non-linear data models for the description of late-time structure formation. Chapter 6 presents a computationally fast and flexible model of mildly non-linear density fields via the technique of remapping LPT. In chapter 7, we introduce the concept of non-linear filtering, designed to improve LSS samples at non-linear scales.

Finally, part IV exploits the BORG SDSS analysis for different cosmographic projects aiming at characterizing and analyzing the cosmic web. Chapter 8 demonstrates that the inference of voids in the dark matter distribution is possible, and that, in addition, our method yields a drastic reduction of statistical uncertainty in void catalogs. In chapter 9, we describe a probabilistic classification of the dynamic cosmic web into four distinct components: voids, sheets, filaments, and clusters. Subsequently, chapter 10 introduces a new decision criterion for labeling different regions, on the basis of posterior probabilities and the strength of data constraints.

In the last chapter, we summarize our results and give our conclusions. Prospective applications and possible directions for future investigations are also mentioned.

The appendices provide complementary material: a mathematical exposition of Gaussian random fields (appendix A), a review of the particle-mesh technique for dark matter simulations (appendix B), and a description of the cosmic web analysis algorithms used in this thesis (appendix C).