CGL

Computational Geometric Learning

D3.3: Heterogeneous and Missing Data

Frederic Cazals

Ebrahim Ehsanfar

Dan Halperin

Rien van de Weijgaert

CGL Technical Report No.: CGL-TR-26

Part of deliverable:	WP-III/D3
Site:	INRIA, RUG, TAU, TUDO
Month:	24

Project co-funded by the European Commission within FP7 (2010–2013)
under contract nr. IST-255827

1 Introduction and Overview

We give a survey of the main approaches that are commonly used for dealing with missing and heterogeneous data in different geometric contexts. Missing data occurs when no value exists for some of the variables in an observation, e.g. when input techniques are not accurate enough or input devices have limited strength and can provide only partial data for a system. This can have a significant effect on the results of a given input data. The system then has to make up for the missing data in order to decide and plan the actions. Also in many cases, data might be from different sources, largely unknown or unlimited, and in many varying formats which form a heterogeneous input for the system. In this survey, we focus on overcoming missing and heterogeneous data in the following topics:

- Robotics
- Structural Biology
- Reconstructions of the cosmic density distribution

In some cases, missing data can be ignored, e.g. in [1] where the probability that a value is missing is independent of the unobserved information such as the value itself. When data is non-ignorably missing, it means that the probability that a value is missing depends on unobserved information, the model for generating the input data has to take into account the missing data mechanism. Dealing with imperfect data is a growing challenge and new ideas and approaches that can handle difficulties in this regards are encouraged, e.g. NASA and other American agencies launched Big Data Challenge[1] which is a public competition to develop new ways to deal with heterogeneous parts of the US government's vast stores of data.

Section 2 discusses the methods for dealing with imperfect data in robotics. Heterogeneous and missing data commonly occur in robot planings, even for highly developed systems. The knowledge of robots is usually modeled in form of a probability distribution and missing and heterogeneous data is studied in form of *uncertainty* on robot knowledge. We then introduce different tools to plan the robotic systems under uncertainty. Section 2 also describes overcoming missing and heterogeneous data in geometric approaches and provides some examples in this regard.

Section 3 consists of the missing data problem statement in structural biology. We describe the state of the art in handling heterogeneous and missing data in biophysical experimental techniques in reconstructing macro-molecular models; include X-ray crystallography, Nuclear Magnetic Resonance (NMR) and Cryo-Electron Microscopy (cryoEM). We also analyze difficulties inherent to in-silico modeling and illustrate the experimental approaches used to cope with them.

In Section 4, we investigate how serious deficiencies, distortions and missing data affect reconstructions of the cosmic density distribution which are beset by numerous observational artifacts with the measurement of various geometric and topological parameters. In Section 4, intention is to develop instruments for correcting the deficiencies such that we will be able to infer the correct parameters of the real underlying cosmology and cosmic mass distribution.

Section 5 includes a brief summary and conclusion.

2 Robotics

Robots are machines that perceive their environment and then plan and execute their actions based on this perception. Perception is carried out via sensors and it is

[1] http://www.nasa.gov/home/hqnews/2012/oct/HQ_12-346_Bio_Data_Challenge.txt

intertwined with action. When the robotic task is non-trivial, the robot may need to acquire a large variety of heterogeneous data. For example, a robotic sailing boat [2] needs to learn about wind speed, wind direction, wave height, near-by vessels or other obstacles, and more. It then has to fuse these different types of data in order to plan its next steps. More often than not, sensors are limited in strength or accuracy, providing the robotic system with only partial data. The system then has to make up for the missing data in order to plan its action.

Sizable portions of the robot algorithms literature assume that the robot has perfect knowledge of its environment. This approach could be justified in several ways: (i) the robot is acting in a carefully engineered environment, like a robot arm in an assembly line, or (ii) the robot system is equipped with high-end sensors providing accurate and sufficient information for the desired task. Even under such favorable conditions, assuming perfect knowledge of the environment may be risky: something unexpected can happen even on a factory assembly line, and sensors may fail (possibly only few of many).

The issue of imperfect or incomplete data is more strongly pronounced with modern robot systems that operate in unpredictable environments in the air, sea or on solid ground, quite often acting in the presence of (moving) humans. The data fusion and processing problems are accentuated by imprecision of the robot actions (for example, the robot has to rotate by 30 degrees but cannot carry out this rotation to perfection) and by the fact that the models of the physical world are only approximate. All these factors together led roboticists to seek a sound way to cope with missing, incomplete or uncertain data.

A holistic approach to handling incomplete data in robotics substitutes the traditional description of a robot state at a fixed time as a single configuration by a *probability distribution* (PD, for short) of configurations. At one extreme, when the robot has very good knowledge of its state, the PD will have one sharp peak. At the other extreme, when a lot of data is missing, the PD will be flat. A set of tools has been developed to update the PD as the robot acquires more information or loses information when the system dynamically changes. The tools can be collectively termed Bayes filters which aim to estimate the robot configuration in the form of a PD. These filters include Kalman filters, hidden Markov models, dynamic Bayesian networks to name just a few. Based on this estimation, the action can now be derived by utility optimization. In their book "Probabilistic Robotics" [3], Thrun, Burgard and Fox give a systematic and comprehensive coverage of this approach.

The basic approaches to plan under uncertainty, like Partially Observable Markov Decision Processes (POMPD), are computationally expensive and may be too costly to directly apply to complex settings. New and improved variants of planning under uncertainty constitute and active area of current robotics research (e.g., [4], [5], [6], [7]).

The robotics methods that plan with only partial observation (missing data) do not render the algorithmic approaches for the fully-observable setting useless. More tractable methods for planning under uncertainty are hybrid in the sense that they combine the Bayesian filters approach with information obtained from standard algorithms.

We also remark that alternative approaches exist to overcoming missing data in the robotics context in a purely geometric fashion (see, e.g., [8],[9],[10]).

Missing data has a curious incarnation in research on motion planning of robots with many degrees of freedom in the presence of known static obstacles. The entire geometric setting is known, but in raw form—in the physical workspace. However, we do not know how to plan motion in a complete fashion in the workspace. We therefore resort to the high-dimensional configuration space. An effective method to translate the full geometric knowledge from the workspace to high-dimensional C-spaces is by sampling. Obviously we cannot sample exhaustively to get a complete

knowledge of the C-space, and we are faced with difficult questions where exactly to search for pertinent missing data. This is a largely unresolved problem that has only heuristic solutions thus far (see, e.g., [11, Chapter 7], [12]).

Heterogeneous data are commonplace in robotics, as implied in the opening paragraph of this section, which refers to robotic sailing boats. It is based on our own experience from the world robotic sailing championship [2]. We now provide another example, where the issue of heterogeneous data is meticulously recorded, discussed and archived. In the 2007 DARPA urban challenge, eleven autonomous vehicles had to drive along a 90 kilometer track in mock-urban environment. The MIT team recorded the varied sensor information they collected in order to plan their vehicle's route. The data include input from: twelve planar laser range scanners, five cameras, a high precision navigation system, and a high density laser range scanner [13]. They describe in detail how the data were used for the task at hand [14]. This scenario is typical of modern autonomous robotic systems. In general, fusion of heterogeneous data into a coherent and useful structure remains a challenging problem.

3 Structural Biology

While proteins and nucleic acids are the fundamental components of an organism, Biology itself is based on the interactions they make with each other. Structural biology precisely aims at unraveling the relationship between the structure and the function of these macro-molecules and complexes, an endeavor which is especially challenging due to missing and heterogeneous data.

In the following, we discuss these difficulties in two directions: first, we focus on experimental techniques aiming at reconstructing macro-molecular models; second, we analyze difficulties inherent to in-silico modeling.

3.1 Reconstructing Models from Bio-physical Experiments

Structural information obtained for macro-molecular systems has proven essential in interpreting physical, biochemical, and functional data. An elite club of experimental techniques dominated by X-ray crystallography and nuclear magnetic resonance (NMR), and to a lesser extent by cryo-electron microscopy (cryoEM) are being used to this end. The structural information obtained is stored in a public repository, the Protein Databank (PDB, http://www.rcsb.org/pdb) [15], which as of October 2012 contains about 75,000 entries, with hundreds added each month.

X-ray crystallography. In X ray crystallography, a crystal made of (complexes of) macro-molecules is bombarded by X rays, yielding a diffraction spectrum. The challenge consists in solving an inverse problem, the outcome being the x, y, z coordinates of the atoms accounting for the diffraction spectrum. The atoms whose coordinates are reported are those constituting the so-called asymmetric unit of the crystal, namely the smallest object which under a group action generates the periodic crystal. X-ray crystallography has proved to be particularly well adapted to biological structure determination, yielding the atomic coordinates of a wide variety of sizes of structures. Although structures of virus particles having a high degree of symmetry have been solved, at 30 nanometers the ribosome is currently the largest asymmetric structure solved by X-ray crystallography. Multi-domain proteins, oligomers and complexes can be much larger than this and must be investigated using different techniques – see below.

[2] `http://acg.cs.tau.ac.il/courses/workshop/spring-2011/projects/roboat/world-robotic-sailing-competition-2011/training-and-competition`

Heterogeneous and missing data are commonly faced in crystallography. For a crystal structure, heterogeneity typically corresponds to a molecular segment (the portion of a polypeptide chain) which does not exist in a single conformation, but instead as a mixture of well identified alternative conformations. Practically, for each atom involved in such a region, *alternate* locations are reported in the PDB file. If the region is even more flexible, the signal accumulated in the diffraction spectrum is not significant enough, preventing the reconstruction. Thus, highly flexible regions result in missing data in PDB files. A typical example is that of flexible loops, namely sections up to 30 amino-acids in length. The reconstruction of such loops borrows upon modeling, and raises challenges since conformations compliant with the two fixed extremities need to be generated [16, 17], and the quality and diversity of these candidates must be assessed [18].

Nuclear Magnetic Resonance. As opposed to X ray crystallography, NMR is targeting small molecules in solution, thus providing conditions which are closer to the in-vivo ones. In contrast to crystallography, an NMR experiment results in an ensemble of atomic resolution structures, i.e. a family of conformations fluctuating about one stable conformation. This ensemble is especially interesting to assess the flexibility of the molecule, which may vary significantly along the backbone. NMR reconstruction is based on the analysis of spectra. One peak in such a spectrum encodes the spatial proximity between the two nuclei and can be translated into an interval of possible distances between them. The assignment of peaks to atoms thus yields a set of distance intervals, also called *restraints*, from which the following distance geometry problem must be solved: given (selected) distance intervals between selected pairs of atoms of a molecule, infer the position of these atoms [19]. Missing data correspond to pairs of atoms located nearby in 3D space, but which do not yield any restraint. Various algorithms have been developed to solve the distance geometry problem, a classical strategy being based on the Bayesian approach [20].

Cryo-Electron Microscopy (cryoEM). In electron microscopy, a frozen sample consisting of large protein assembly or even a whole cell is bombarded by electrons, from which structural information can be recovered [21]. The output of a cryoEM experiment is not a list of atoms, but a 3D density map encoding the density of matter in a cubical volume containing the model. Typically, low (less than 10 Å, domains visible) to medium (around 5Å, secondary structure elements visible) resolutions are achieved.

CryoEM actually refers to two different techniques. In single particle analysis, isolated samples are bombarded, yielding 2D images corresponding to different viewpoints. Combining these 2D images into a 3D image yields a model. But the process is challenging precisely due to the heterogeneity of the samples: the conformations of the molecules imaged may vary, or even worse, the sample preparation process (freezing) may have damaged the molecules. In cryoEM tomography, a given sample is instead bombarded at incremental degrees of rotation, from which a 3D model can also be reconstructed. In both cases, the result is a 3D density map, where each voxel encodes the density of matter. This density is in general very noisy due to the low electron doses used to avoid damaging biological specimens. Choosing a density level for contouring a surface (called the envelope) that encloses the model is non trivial, as the intensity is generally high for globular domains of the proteins but low for unstructured regions such as linkers connecting these domains. As discussed above for crystallography, linkers typically yield missing regions in cryoEM maps. In favorable cases, atomic resolution models can be built from cryoEM maps by fitting existing and/or modeled structural elements into the maps.

On databases of biological complexes. Having discussed difficulties inherent to bio-physical experiments, one comment is in order about biological databases, i.e. databases containing structural information for protein complexes of biological interest.

As it should be clear from the description of crystallography, NMR and cryoEM experiments, the experiment best suited to acquire high-resolution information on complexes is crystallography. However, given the asymmetric unit of a crystal, one central question consists of deciding whether a contact between two polypeptide chains is a crystallization artifact or a real biological contact — e.g. that between an enzyme and its substrate or between an immunoglobulin and its antigen. Although tools have been developed to qualify contacts within the asymmetric unit [22], the question remains open in general. A consequence is that databases of biological complexes require manual curation, which explains their modest size. For example, the protein docking benchmark `http://zlab.umassmed.edu/zdock/benchmark.shtml` contains 176 test cases. Similarly, the binding affinity benchmark (`http://bmm.cancerresearchuk.org/~bmmadmin/Affinity/`) contains 144 complexes whose binding affinities have been measured experimentally. This is a clear limitation of *homology modeling*, a strategy which consists of inferring information on the partners of a molecular A from the partners of a molecule B which is homologous i.e. similar to A.

3.2 In-Silico Modeling

Modeling by data integration. We mentioned above that large macro-molecular assemblies are not amenable to crystallographic studies. One such complex is the nuclear pore complex (NPC), a cylindrical channel-like assembly located on the nuclear membrane of eukaryotic cells [23]. The NPC has a diameter circa 100 nanometers, and involves about 450 proteins of 30 different species, those located at the periphery of the lumen being highly flexible — these proteins actually regulate the nucleo-cytoplasmic transport. Tens of such molecular machines are known, and of the order of hundreds probably exist in eukaryotic cells. The challenges faced to reconstruct them are different from those inherent to small (typically binary) complexes. First, the exact composition is terms of proteins present in subject to variation along time — e.g. the NPC evolves in composition along the cell cycle. Second, some of the proteins involved are inherently flexible.

To cope with these difficulties, a strategy known as *reconstruction by data integration* has recently been proposed [24]. In a nutshell, this strategy is reminiscent of NMR and consists of mixing experimental data from a variety of sources, so as to find out the model(s) best complying with the data. This paradigm actually requires three ingredients:

- A geometric model of the system studied. For coarse-grain models, a collection of balls representing protein domains is typically used. The geometric parameters describing these balls (center plus radius, whence $4n$ parameters for n balls) define the configuration space of the model.

- Various experimental data (see below), shedding complementary light on the system. These data are turned into so-called *restraints*, and their sum defines a scoring function. This function aims at highlighting which regions of the configuration space of the model best comply with the data.

- An optimization strategy aiming at finding the most significant local minima of the scoring function.

These ingredients can then be used in a process which iteratively mixes computation

and data generation. This process has been used for the reconstruction of plausible models of the NPC [23].

To illustrate the heterogeneity of the data used, and also the missing pieces of information, we now briefly comment on the experimental techniques used in this context:

- **Tandem Affinity Purification (TAP)**. TAP experiments give access to all the protein types found in all complexes containing a prescribed protein type [25], and can thus be used to constrain the spatial proximity of the protein instances within the model. However, one piece of information is critically missing: if instances of two protein types B and C are known to interact with an instance of type A, one does not know whether there exists a single complex involving instances of A, B, C or two complexes involving instances of A, B and A, C respectively.

- **Overlay assays.** In contrast with TAP data, overlay assays aim at detecting pairwise protein-protein contacts, allowing to directly constrain protein contacts in the model. A protein P_b called *bait* is first purified and immobilized on nitrocellulose. Then, a fusion protein P_p called *probe* is also purified and overlaid with P_b. After a period of incubation and a washing step for eliminating unbound probes, the detection of overlaid proteins is carried out [26], yielding a signal $S_{p,b}$, which is specific from the protein complex probe-bait.

- **Ultracentrifugation**. Ultracentrifugation allows determining the shape of globular (domains of) proteins, which is useful to assign rough shapes to proteins. The experiment consists in rotating a protein sample in the centrifuge. By measuring the sedimentation velocity of the protein sample, one can determine an abstract value called the sedimentation coefficient S, which is constant among the marker proteins having the same density. S is directly related to the molecular mass, the volume and the shape of the involved proteins [27].

- **Immuno-electron microscopy.** In immuno-EM, one wishes to locate specific proteins within an assembly [28], this positional information being used to favor the location of proteins in the model. To this end, the protein of interest is attached to a specific antibody or a big tag, which is itself labeled with gold particles. Tracking these gold particles under a microscope yields positional information (up to the microscope resolution) for the proteins of interest.

- **Cryo-electron microscopy (cryoEM).** See our description above.

Accommodating such heterogeneous data is a challenge for two main reasons: first, optimizing the scoring function across the configuration space of the model is generally a complex non convex problem — directly influenced by the relative weights assigned to the various restraints used; second, assessing the coherence between the reconstructed models and the input data is non trivial [29, 30].

Sampling conformational spaces. The regulation of the activity of proteins is tightly coupled to their dynamic properties — a conformational change typically yields the active conformation of the protein from unstructured ones. Understanding such dynamics in detail is a key problem, but since no bio-physical experiment provides the required time resolution, numerical simulations are resorted to. The dynamics that occur in protein folding, conformational changes, and association principally involve changes in non-covalent interactions, which can be treated using conventional classical mechanics, and notably the integration of Newton's equations

of motion, such as is used for all-atom molecular dynamics (MD) simulations. In a MD simulation, a time-step on the order of the fastest nuclear vibrations, with a period on the order of 1 femtosecond, is used. Significant time and computational resources are called for in simulating microsecond (10^9 steps) processes, but dedicated hardware has allowed *in silico* folding of small proteins on the sub-msec scale to be approached [31]. However, the conformational space of larger systems cannot be sampled exhaustively, yielding partial views and thus biased thermodynamic quantities. To remedy this problem, accelerated MD simulations are being developed, based in particular on the idea of biasing the potential from which forces are computed [32], so as to avoid re-visiting regions already sampled.

4 Cosmological Observations and Geometric Biases

The main focus of our research is the use - and development - of instruments from computational geometry and computational topology for the analysis and characterization of the spatial distribution of galaxies and matter in the Universe. On scales of a few up to a hundred Megaparsec, galaxies have assembled in a weblike pattern of sheetlike walls, prominent filaments and clusters, which surround large near-empty voids.

One of the main aspects of the study of the large scale distribution of galaxies is its use as a means to reconstruct the (continuous) cosmic mass distribution. Galaxies form a minor constituent of the total cosmic matter in the Universe. Most of the mass s in the form of the as yet unidentified nonbaryonic dark matter component. Dark matter cannot be observed, but is noticeable via its gravitational influence.

Ideally, we would like to reconstruct the spatial distribution of dark matter throughout a sampled volume. For our cosmological project we are faced by the challenge to extract geometric and topological information from data which are obtained by astronomical observing campaigns. A major and crucial assumption for many of the involved reconstruction techniques is that the galaxy distribution is an unbiased and representative - be it sparse - sample of the underlying dark matter field. In nearly all astronomical situations, however, this is not the case and even unfeasible. A range of observational selection effects play a role in the produced databases. They are beset by a large range of errors, deficiencies, distortions and missing data. We will address the situation in which we know the systematics of these effects. On the basis of this, we will be able to investigate the way in which they affect the obtained geometric and topological parameters.

To appreciate the background of the errors and biases that affect our capability to extract information in the observational reality of astronomy, one may also consider its unique context. The principal reason is that astronomy is a data starved science, entirely dependent on the photons that nature, i.e. the objects populating our universe, are sending to earth. As one cannot influence the information that we measure, a range of errors and biases are affecting our capability to extract information. As in all physical sciences, astronomical observations involve simple statistical measurement errors. Also instrumental effects are a major source of error. In addition, there are also substantial error contributions as a result of atmospheric circumstances and errors induced by the reduction of the raw observed data to analyzable information. However, even far more serious are systematic errors and biases induced by - often inescapable - effects and choices. Because astronomy cannot influence the information that reaches us from space, it is impossible to obtain data that represent a complete and fair sample of space and/or time. Moreover, nearly all astronomical objects are extremely faint, which makes it a demanding

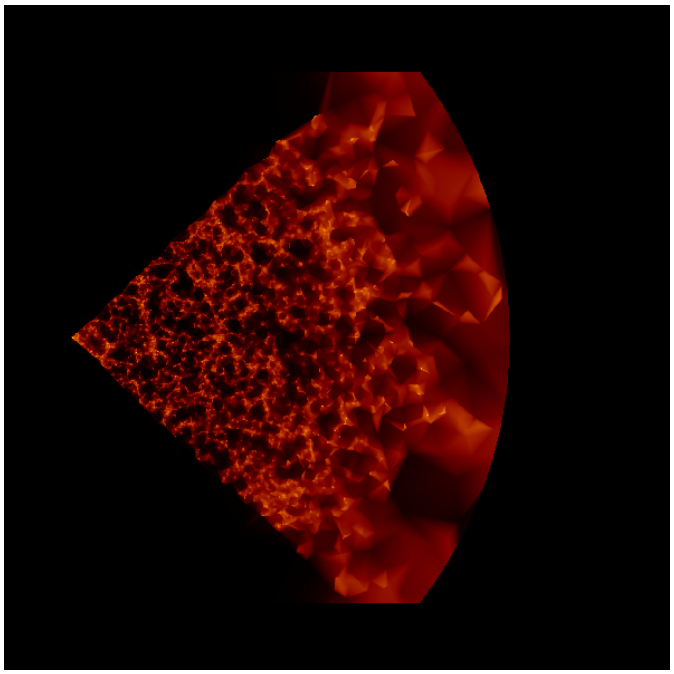

Figure 1: Illustration of typical cosmological survey bias. DTFE reconstructed cosmological density field. We, ie. the observers, are at cone, and distance increases along radial lines. Notice the decreased spatial resolution as a function of distance.

task to gather the necessary photons to be able to extract information. This limits the number of objects that can be observed in a given time interval.

Astronomy is also limited in its ability to study the time evolution of systems and objects. Timescales of cosmic processes are usually in the order of millions to even billions of years. Astronomical observations are not more than snapshots, and to get an idea of the evolutionary processes astronomers rely on physical models, computer modelling and observations of a large number of similar objects residing in different stage of development.

In cosmological circumstances, a first serious artefact is introduced by the choice to observe only galaxies down to a certain brightness threshold and discard fainter galaxies. This translates into a sampling density which is a function of distance. At a close distance, it is therefore possible to see galaxies that are intrinsically faint, while at a high distance only the brightest galaxies can be seen. This produces a strong gradient in sampling density, with sufficiently resolved patterns at nearby distance but an increasingly diluted structure at larger distances. A second important factor is that limited resources usually force astronomers to survey only a limited region of the sky. Atmospheric and instrumental problems during long observing campaigns also tend to lead to a partial coverage and regions which have been missed.

Studies based on the full three-dimensional distribution of matter and galaxies, including our own, are affected by an additional major systematic effect. While position of a galaxy on the sky is known to high precision, it is almost impossible to determine the distances to galaxies with sufficient accuracy. Even for galaxies with a distance of a few tens of Megaparsec, we are usually not capable of determining the distance more accurate than 10-20%. Instead, cosmologists use the expansion of the Universe itself to get a reasonable estimate of the distance. According to the Hubble expansion of the Universe, the velocity with which galaxies move away from us is proportional to their distance. The velocity can be quite accurately determined from the redshift of the galaxy's electromagnetic spectrum. Hence, redshift is used as a reasonably good approximation to distance. However, galaxies not only move because of the expansion of the Universe, they also move with respect

to the Universe itself. These motions are strongly correlated with the nature of the structures in which they are embedded as they are the result of the cosmic structure formation process itself. It leads to socalled redshift distortions and means that the estimated redshift distance of galaxies is beset by systematic distortions which are often impossible to disentangle from the underlying structure. This leads to a distortion of the patterns mapped by cosmological galaxy surveys, the maps on which we apply our geometric and topological tools. In all, we may conclude that astronomical datasets usually comprise a very sparse, incomplete and distorted representation of the underlying physical reality.

In the previous years, we have been succesful in designing a range of toolboxes for analyzing and reconstructing the intricate complex pattern of the Cosmic Web. The central procedure is the Delaunay Tessellation Field Estimator (DTFE) [33, 34, 35], which translates the discrete galaxy distribution into a piecewise linear density field that retains the intricate and complex structure visible in the galaxy distribution. Tests of DTFE have revealed that it reproduces accurately the anisotropic and multiscale mass distribution seen in the galaxy distribution and in computer models of cosmic structure formation. In addition, with the Watershed Void Transform (WVF) we have implemented the watershed transform towards identification of voids [36, 37]. The Nexus algorithm processes the DTFE density field with a morphological scalespace analysis towards the identification of filaments and walls [38, 39, 40]. The Spineweb procedure [39] is a generalization of the WVF algorithm towards mapping the spine of the Cosmic Web. Finally, over the past years we have been developing routines for homology analysis of the cosmic mass distributions, i.e. the determination of Betti number curves (as function of density threshold) and persistence diagrams. At the moment, we are in the process of relating these to the evolving patterns in the cosmic mass distribution.

4.1 Project Definition

The ultimate intention of our project is to apply our geometric and topological tools on the observational reality in an attempt to use the observed Cosmic Web towards discriminating between cosmological models and the determination of cosmic parameters. In other words, given the succesfull application of these geometric and topological tools on idealized computer simulations of cosmic structure formation, we need to turn to the crucial question how they can be used to analyze the observational reality. In this context, we may identify a few key questions and aspects:

- It is of major importance to assess in how far the field reconstructions and inferred parameters are affected by the deficiencies, distortions and non-uniformities in the data.
- Given insight into the way in which the various deficiencies affect the geometric and topological measurements, we wish to develop correction algorithms which may (partially) correct the induced artefacts.
- Find whether there are robust geometric and topological measurements that are less affected by the various astronomical biases and deficiencies.
- A major share of our work will involve the comparison of data with theoretical models, often in the form of idealized computer simulations of cosmic structure formation that are post-processed such that they are affected by the same observational deficiencies as the data. The resulting parameter measurement will not necessarily reflect the underlying reality, even though a comparison may lead to a useful discrimination between models.

- The best option would be to pre-process the observed data to a (partially) reconstructed mass field. The subsequent application of our geometric and topological tools would then produce parameters close to the one of the actual cosmic density field. Such reconstructions of the cosmic density field based on the available data have been developed. They are the realm of constrained field reconstructions (see e.g [41, 42, 43, 44] and Kriging reconstructions (see e.g. [45]).

5 Summary and Conclusion

Imperfect data can reduce the performance of approaches based on observations, particularly if missing a value depends on unobserved information. In this survey, we described dealing with heterogeneous and missing data mechanism in three different disciplines including robotic sciences, structural biology and cosmology. We have seen also significant approaches that are commonly used for dealing with missing and heterogeneous data provided in a geometric fashion of mentioned fields.

Acknowledgement. The project CG Learning acknowledges the financial support of the Future and Emerging Technologies (FET) programme within the Seventh Framework Programme for Research of the European Commission, under FET-Open grant number: 255827

References

[1] Donald B. Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.

[2] Alexander Scläfer and Ole Blaureock, editors. *Proceedings of the 4th International Robotic Sailing Conference*. Springer, 2011.

[3] S. Thrun, W. Burgard, and D. Fox. *Probabilistic Robotics*. MIT Press, Cambridge, MA, 2005.

[4] Haoyu Bai, David Hsu, Mykel J. Kochenderfer, and Wee Sun Lee. Unmanned aircraft collision avoidance using continuous-state pomdps. In *Robotics: Science and Systems*, 2011.

[5] Adam Bry and Nicholas Roy. Rapidly-exploring random belief trees for motion planning under uncertainty. In *ICRA*, pages 723–730, 2011.

[6] Robert Platt, Leslie Pack Kaelbling, Tomás Lozano-Pérez, and Russ Tedrake. Non-gaussian belief space planning: Correctness and complexity. In *ICRA*, pages 4711–4717, 2012.

[7] Jur van den Berg, Sachin Patil, and Ron Alterovitz. Motion planning under uncertainty using iterative local optimization in belief space. *I. J. Robotic Res.*, 31(11):1263–1278, 2012.

[8] Mark de Berg, Leonidas J. Guibas, Dan Halperin, Mark H. Overmars, Otfried Schwarzkopf, Micha Sharir, and Monique Teillaud. Reaching a goal with directional uncertainty. *Theor. Comput. Sci.*, 140(2):301–317, 1995.

[9] Michael Erdmann. Using backprojections for fine motion planning with uncertainty. *I. J. Robotic Res.*, 5(1):19–45, 1986.

[10] Anthony Lazanas and Jean-Claude Latombe. Landmark-based robot navigation. *Algorithmica*, 13(5):472–501, 1995.

[11] H. Choset, W. Burgard, S. Hutchinson, G. Kantor, L. E. Kavraki, K. Lynch, and S. Thrun. *Principles of Robot Motion: Theory, Algorithms, and Implementation*. MIT Press, June 2005.

[12] Oren Salzman, Michael Hemmer, and Dan Halperin. On the power of manifold samples in exploring configuration spaces and the dimensionality of narrow passages. *CoRR*, abs/1202.5249, 2012.

[13] Albert S. Huang, Matthew E. Antone, Edwin Olson, Luke Fletcher, David Moore, Seth J. Teller, and John J. Leonard. A high-rate, heterogeneous data set from the darpa urban challenge. *I. J. Robotic Res.*, 29(13):1595–1601, 2010.

[14] Albert Huang, David Moore, Matthew E. Antone, Edwin Olson, and Seth J. Teller. Multi-sensor lane finding in urban road networks. In *Robotics: Science and Systems*, 2008.

[15] P.W. Rose, B. Beran, C. Bi, W.F. Bluhm, D. Dimitropoulos, D.S. Goodsell, A. Prlic, M. Quesada, G.B. Quinn, J.D. Westbrook, J. Young, B. Yukich, C. Zardecki, H.M. Berman, and P.E. Bourne. The rcsb protein data bank: redesigned web site and web services. *Nucleic Acids Res*, 39(Database issue):D392–D401, Jan 2011.

[16] D.J. Mandell, E.A. Coutsias, and T. Kortemme. Sub-angstrom accuracy in protein loop reconstruction by robotics-inspired conformational sampling. *Nature Methods*, 6:551 – 552, 2009.

[17] A. Dhanik, P. Yao, N. Marz, R. Propper, C. Kou, G. Liu, H. van den Bedem, and J.C. Latombe. Efficient algorithms to explore conformation spaces of flexible protein loops. In *7th Workshop on Algorithms in Bioinformatics (WABI 2007)*, pages 265–276, September 2007.

[18] S. Loriot, S. Sachdeva, K. Bastard, C. Prevost, and F. Cazals. On the characterization and selection of diverse conformational ensembles. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 8(2):487–498, 2011.

[19] B.R. Donald and J. Martin. Automated nmr assignment and protein structure determination using sparse dipolar coupling constraints. *Progress in nuclear magnetic resonance spectroscopy*, 55(2):101, 2009.

[20] W. Rieping, M. Habeck, and M. Nilges. Inferential structure determination. *Science*, 309:303–6, 2005.

[21] J. Frank. *Three-dimensional electron microscopy of macromolecular assemblies: visualization of biological molecules in their native state*. Oxford University Press, USA, 2006.

[22] J. Bernauer, R. P. Bahadur, F. Rodier, J. Janin, and A. Poupon. DiMoVo: a Voronoi tessellation-based method for discriminating crystallographic and biological protein-protein interactions. *Bioinformatics*, 24(5):652–8, 2008.

[23] F. Alber, S. Dokudovskaya, L. M. Veenhoff, W. Zhang, J. Kipper, D. Devos, A. Suprapto, O. Karni-Schmidt, R. Williams, B.T. Chait, M.P. Rout, and A. Sali. Determining the Architectures of Macromolecular Assemblies. *Nature*, 450(7170):683–694, Nov 2007.

[24] F. Alber, F. Frster, D. Korkin, M. Topf, and A. Sali. Integrating Diverse Data for Structure Determination of Macromolecular Assemblies. *Ann. Rev. Biochem.*, 77:11.1–11.35, 2008.

[25] O. Puig, F. Caspary, G. Rigaut, B. Rutz, E. Bouveret, E. Bragado-Nilsson, M. Wilm, and B. Séraphin. The tandem affinity purification method: A general procedure of protein complex purification. *Methods*, 24:218–229, 2001.

[26] R.A. Hall. Studying protein-protein interactions via blot overlay or far western blot. *Methods in molecular biology*, 261:167–174, 2004.

[27] H.P. Erickson. Size and shape of protein molecules at the nanometer level determined by sedimentation, gel filtration, and electron microscopy. *Biol Proced Online*, 11:32–51, 2009.

[28] H. Schwartz and H. Hohenberg. *Immuno-electron Microscopy*. Wiley Online Library, 2001.

[29] T. Dreyfus, V. Doye, and F. Cazals. Assessing the reconstruction of macromolecular assemblies with toleranced models. *Proteins: structure, function, and bioinformatics*, 80(9):2125–2136, 2012.

[30] T. Dreyfus, V. Doye, and F. Cazals. Probing a continuum of macro-molecular assembly models with graph templates of sub-complexes. 2013. Submitted.

[31] D.E. Shaw, P. Maragakis, K. Lindorff-Larsen, S. Piana, R.O. Dror, M.P. Eastwood, J.A. Bank, J.M. Jumper, J.K. Salmon, Y. Shan, and W. Wriggers. Atomic-level characterization of the structural dynamics of proteins. *Science*, 330(6002):341–346, Oct 2010.

[32] A. Laio and M. Parrinello. Escaping free-energy minima. *Proceedings of the National Academy of Sciences*, 99(20):12562–12566, 2002.

[33] W.E. Schaap and R. van de Weygaert. Continuous fields and discrete samples: reconstruction through delaunay tessellations. 2000.

[34] I. Szapudi. Introduction to higher order spatial statistics in cosmology. 665:457–492, 2009.

[35] Marius C. Cautun and Rien van de Weygaert. The DTFE public software - The Delaunay Tessellation Field Estimator code. 2011.

[36] Erwin Platen, Rien Van De Weygaert, and Bernard J. T. Jones. A cosmic watershed: the wvf void detection technique. *Monthly Notices of the Royal Astronomical Society*, 380(2):551–570, 2007.

[37] E. G. Patrick Bos, Rien van de Weygaert, Klaus Dolag, and Valeria Pettorino. The darkness that shaped the void: dark energy and cosmic voids. *Monthly Notices of the Royal Astronomical Society*, 426(1):440–461, 2012.

[38] M. A. Aragón-Calvo, B. J. T. Jones, R. van de Weygaert, and J. M. van der Hulst. The multiscale morphology filter: identifying and extracting spatial patterns in the galaxy distribution. *A&A*, 474(1):315–338, 2007.

[39] Miguel A. Aragn-Calvo, Rien van de Weygaert, and Bernard J. T. Jones. Multiscale phenomenology of the cosmic web. *Monthly Notices of the Royal Astronomical Society*, 408(4):2163–2187, 2010.

[40] Marius Cautun, Rien van de Weygaert, and Bernard J.T. Jones. NEXUS: Tracing the Cosmic Web Connection. 2012.

[41] E. Bertschinger. Path integral methods for primordial density perturbations - Sampling of constrained Gaussian random fields. , 323:L103–L106, December 1987.

[42] Y. Hoffman and E. Ribak. Constrained realizations of Gaussian fields - A simple algorithm. , 380:L5–L8, October 1991.

[43] R. van de Weygaert and E. Bertschinger. Peak and gravity constraints in Gaussian primordial density fields: An application of the Hoffman-Ribak method. , 281:84, July 1996.

[44] Francisco-Shu Kitaura. The Initial Conditions of the Universe from Constrained Simulations. 2012.

[45] Erwin Platen, Rien van de Weygaert, Bernard J. T. Jones, Gert Vegter, and Miguel A. Aragn Calvo. Structural analysis of the sdss cosmic web i. nonlinear density field reconstructions. *Monthly Notices of the Royal Astronomical Society*, 416(4):2494–2526, 2011.