PUB. IRMA, LILLE 2009
Vol. 69, N^{o} IX

Filaments in observed and mock galaxy catalogues

R. S. Stoica*
Université Lille 1
Laboratoire Paul Painlevé
59655 Villeneuve d'Ascq Cedex France

V. J. Martínez†
Observatori Astronòmic, Universitat de València
Apartat de correus 22085
46075 València, Spain

E. Saar‡
Tartu Observatoorium, Tõravere
61602 Estonia

Résumé

The main feature of the spatial large-scale galaxy distribution is an intricate network of galaxy filaments. The present paper compares the filaments in the real data and in the numerical models, to see if our best models reproduce statistically the filamentary network of galaxies.

Résumé

Le réseau filamentaire est la caractéristiques principale de la distribution spatiale à large échelle des galaxies. Ce papier compare la structure filamentaire dans des données réelles et synthétiques afin de vérifier si les meilleurs modèles sont capables de reproduire des réseaux ayant les mêmes caractéristiques statistiques que les observations.

Keywords: marked point processes, statistical inference, cosmology observations, large-scale structure of the Universe.

1 Introduction

The large-scale structure of the Universe traced by the three-dimensional distribution of galaxies shows intriguing patterns: filamentary structures connecting

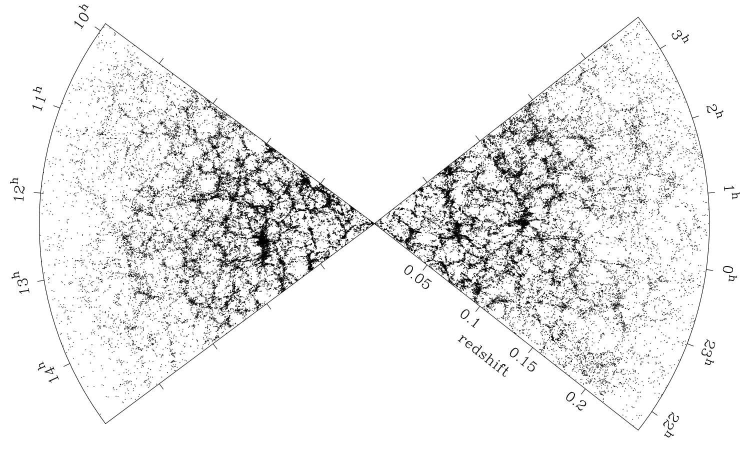

Figure 1: Galaxy map for two 2dFGRS slices of thickness of 2.6°. The filamentary network of galaxies is clearly seen.

in huge clusters surround nearly empty regions, the so-called voids. As an example, we show here a map from the 2dF Galaxy Redshift Survey (2dFGRS, [8]). In order to better show the filamentary network, Fig. 1 shows the positions of galaxies in two 2.6° thick slices from two spatial regions that the 2dFGRS covered. The distances are given in redshifts z; approximately, the physical distance $D \approx 3000\,h^{-1}\,z\,\mathrm{Mpc}$[1].

Filaments visually dominate the galaxy maps. Real three-dimensional filaments have been extracted from the galaxy distribution as a result of special observational projects ([22]), or by searching for filaments in the 2dFGRS catalogue ([23]). These filaments have been searched for between galaxy clusters, determining the density distribution and deciding if it is filamentary, individually for every filament. Filaments are also suspected to hide half of the warm gas in the Universe; an example of a discovery of such gas is the paper by [38].

However, there are still no standard methods to describe the observed filamentary structure, but much work is being done in this direction. The usual second-order summary statistics as the two-point correlation function or the power spectrum do not provide morphological information. Minkowski functionals, minimal spanning tree (MST), percolation and shapefinders have been introduced for this purpose (for a review see [18]).

The minimal spanning tree was introduced in cosmology by [4]. It is a unique graph that connects all points of the process without closed loops, but it describes mainly the local nearest-neighbour distribution and does not give us the global and large-scale properties of the filamentary network. A recent development of these ideas is presented by [6]. He applies a minimal spanning tree on a grid, and works close to the percolation regime – this allows to study the global structure of the galaxy distribution. We note that using a grid introduces

[1] Here and below h is the dimensionless Hubble constant, $H = h \cdot 100\mathrm{km/sec/Mpc}$.

a smoothed density, and this is typical for other recent approaches, too.

In order to describe the filamentary structure of continuous density fields, a *skeleton* method has been proposed and developed by [9, 21]. The skeleton is determined by segments parallel to the gradient of the field, connecting saddle points to local maxima. Calculating the skeleton involves interpolation and smoothing the point distribution, which introduces an extra parameter, the band-width of the kernel function used to estimate the density field from the point distribution, typically a Gaussian function. This is typical for most of the density-based approaches. The skeleton method was first applied for two-dimensional maps, approach for studying the cosmic microwave sky background [9], The method was generalized to 3-D maps [26] and was applied to the Sloan Digital Sky Survey by [27], providing, by means of the length of the skeleton, a good discriminant tool for the analysis of the filamentary structures. The formalism has recently been developed further and applied to study the evolution of filamentary structure in simulations [25].

Another approach is that of [1]. They use DFTE (Delaunay Triangulation Field Estimator) to reconstruct the density field for the galaxy distribution, and apply the Multiscale Morphology Filter (MMF) to identify various structures, as clusters, walls, filaments and voids. As a further development, this group has used the watershed algorithm to describe the global properties of the density field [2].

A new direction is to use the second-order properties (the Hessian matrix) of the density field [5] or the deformation tensor [10]. As shown in these papers, this also allows to trace and classify different features of the fields.

Our approach does not introduce the density estimation step; we consider the galaxy distribution as a marked point process. In an early paper [32], we proposed to use an automated method to trace filaments for realisations of point processes, that has been shown to work well for detection of road networks in remote sensing situations [16, 28, 29]. This method is based on the Candy model, a marked point process where segments serve as marks. The Candy model can be applied to 2-D filaments, and we tested it on simulated galaxy distributions. The filaments we found delineated well the filaments detected by eye.

Based on our previous experience with the Candy process, we generalised the approach for three dimensions. As the interactions between the structure elements are more complex in three dimensions, we had to define a more complex model, the Bisous model ([31]). This model gives a general framework for the construction of complex patterns made of simple interacting objects. In our case, it can be seen as a generalisation of the Candy model. We applied the Bisous model to trace and describe the filaments in the 2dFGRS ([33]) and demonstrated that it works well.

In the paper cited above we chose the observational samples from the main magnitude-limited 2dFGRS catalogue, selecting the spatial regions to have approximately constant spatial densities. Strict application of the Bisous process demands, however, a truly constant spatial density (intensity). In this paper, we will apply the Bisous process to compare the observational data with mock catalogues, specially built to represent the 2dFGRS survey. In order to obtain strict statistical test results, we use here volume-limited subsamples of the 2dFGRS, and of the mock catalogues. We trace the filamentary network in all our catalogues and compare its properties.

2 Mathematical tools

In this section we describe the main tools we use to study the large-scale filaments. The key idea is to see this filamentary structure as a realisation of a marked point process. Under this hypothesis, the cosmic web can be considered as a random configuration of segments or thin cylinders that interact, forming a network of filaments. Hence, the morphological and quantitative characteristics of these complex geometrical objects can be obtained by following a straightforward procedure: constructing a model, sampling the probability density describing the model, and, finally, applying the methods of statistical inference.

We have given a more rigorous description of these methods in our previous paper ([33]). As this journal is not frequently read by astronomers, we recapitulate here the methods in a more understandable language.

2.1 Marked point processes

A popular model for the spatial distribution of galaxies is a point process on K (a compact subset of $\mathbf{R}^3$, the cosmologists's sample volume), a random configuration of points $\mathbf{k} = \{k_1, \ldots, k_n\}$, lying in K. Let $\nu(K)$ be the volume of K.

We may associate characteristics or marks to the points. For instance, to each point in a configuration $\mathbf{k}$, shape parameters describing simple geometrical objects may be attached. Let $(M, \mathcal{M}, \nu_M)$ be the probability measure space defining these marks. A marked or object point process on $K \times M$ is random configuration $\mathbf{y} = \{(k_1, m_1), (k_2, m_2), \ldots, (k_n, m_n)\}$, with $y_i = (k_i, m_i) \in K \times M$ for all $i = 1, \ldots, n$ such that the locations process is a point process on K. For our purposes, the point process is considered finite and simple, *i.e.* the number of points in a configuration is finite and $k_i \neq k_j$, for any i, j such that $1 \leq i, j \leq n$.

In case of the simplest marked point process, the objects do not interact. The Poisson object point process is the most appropriate choice for such a situation. This process chooses a number of objects according to a Poisson law of the intensity parameter $\nu(K)$, gives to each object a random independent location uniformly in K and a random shape or mark chosen independently according to ν_M. The Poisson object point process has the great adavantage that it can be described by analytical formulae. Still, it is too simple, whenever the interactions of objects are to be taken into account.

The solution to the latter problem is to specify a probability density $p(\mathbf{y})$ that takes into account interactions between the objects. This probability density is specified with respect to the reference measure given by the Poisson object point process. There is a lot of freedom in building such densities, provided that they are integrable with respect to the reference measure and locally stable. This second condition requires that there exists $\Lambda > 0$ such that $p(\mathbf{y} \cup \{(k, m)\})/p(\mathbf{y}) \leq \Lambda$ for any $(k, m) \in K \times M$. Local stability implies integrability. It is also an important condition, guaranteeing that the simulation algorithms for sampling such models have good convergence properties.

For further reading and a comprehensive mathematical presentation of object point processes, we recommend the monographs by [35, 19, 34, 14]

2.2 Bisous model

In this section, we shall describe the probability density of the Bisous model for the network of cosmic filaments. The Bisous model is a marked point process that was designed for generating and analysing random spatial patterns [31, 33].

Random spatial patterns are complex geometrical structures composed of rather simple objects that interact. We can describe our problem as follows: in a region K of a finite volume, we observe a finite number of galaxies $\mathbf{d} = \{d_1, d_2, \ldots, d_r\}$. The positions of these galaxies form a complicated filamentary network. Modeling it by a network of thin cylinders that can get connected and aligned in a certain way, a marked point process – the Bisous model – can be built in order to describe it.

A random cylinder is an object characterized by its center k and its mark giving the shape parameters. The shape parameters of a cylinder are the radius r, the height h and the orientation vector ω. We consider the radius and height parameters as fixed, whereas the orientation vector parameters $\omega = \phi(\eta, \tau)$ are uniformly distributed on $M = [0, 2\pi) \times [0, 1]$ such that

$$\omega = (\sqrt{1-\tau^2} cos(\eta), \sqrt{1-\tau^2} sin(\eta), \tau). \tag{1}$$

For our purposes, throughout this paper the shape of a cylinder is denoted by $s(y) = s(k, r, h, \omega)$, a compact subset of $\mathbf{R}^3$ of a finite volume $\nu(s(y))$. The shape of a random cylinder configuration $\mathbf{y}$ is defined by the random set $Z(\mathbf{y}) = \cup_{y \in \mathbf{y}} s(y)$.

A cylinder (k, ω) has $q = 2$ extremity rigid points. We centre around each of these points a sphere of radius r_a. These two spheres form an attraction region that plays an important role in defining connectivity and alignment rules for cylinders. We illustrate the basic cylinder in Fig. 2, where it is centred at the coordinate origin and its symmetry axis is parallel to Ox. The coordinates of the extremity points are

$$e_u = ((-1)^{u+1}(\frac{h}{2} + r_a), 0, 0), \quad u \in \{1, 2\} \tag{2}$$

and the orientation vector is $\omega = (1, 0, 0)$.

The probability density for a marked point process based on random cylinders can be written using the Gibbs modeling framework:

$$p(\mathbf{y}|\theta) = \frac{\exp\left[-U(\mathbf{y}|\theta)\right]}{\alpha} \tag{3}$$

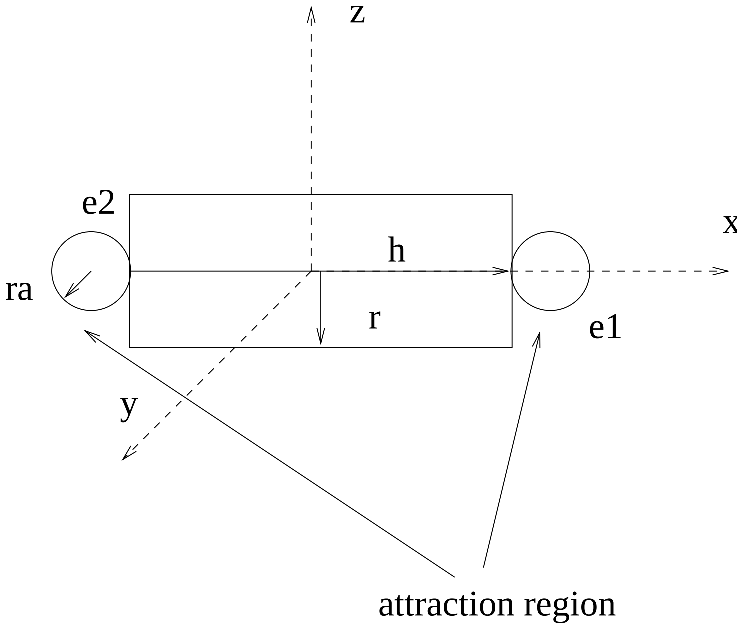

Figure 2: A thin cylinder that generates the filamentary network.

where α is the normalizing constant, θ is the vector of model parameters and $U(\mathbf{y}|\theta)$ is the energy function of the system.

Modeling the filamentary network induced by the galaxy positions needs two assumptions. The first assumption is that locally, galaxies may be grouped together inside a rather small thin cylinder. The second assumption is that such small cylinders may combine to prolongate a filament, if neighbouring cylinders are aligned in similar directions.

Following these two ideas the energy function given by (3) can be specified as:

$$U(\mathbf{y}|\theta) = U_{\mathbf{d}}(\mathbf{y}|\theta) + U_i(\mathbf{y}|\theta) \tag{4}$$

where $U_{\mathbf{d}}(\mathbf{y}|\theta)$ is the data energy and $U_i(\mathbf{y}|\theta)$ is the interaction energy, associated to the first and second assumptions above, respectively. In fact, it is perfectly reasonable to think that the data energy places the cylinders in the galaxy field, while the interaction energy plays a regularization role, by encouraging the cylinders to form filamentary patterns.

2.3 Data energy

The data energy of a configuration of cylinders $\mathbf{y}$ is defined as the sum of the energy contributions corresponding to each cylinder:

$$U_{\mathbf{d}}(\mathbf{y}|\theta) = -\sum_{y\in\mathbf{y}} v(y) \tag{5}$$

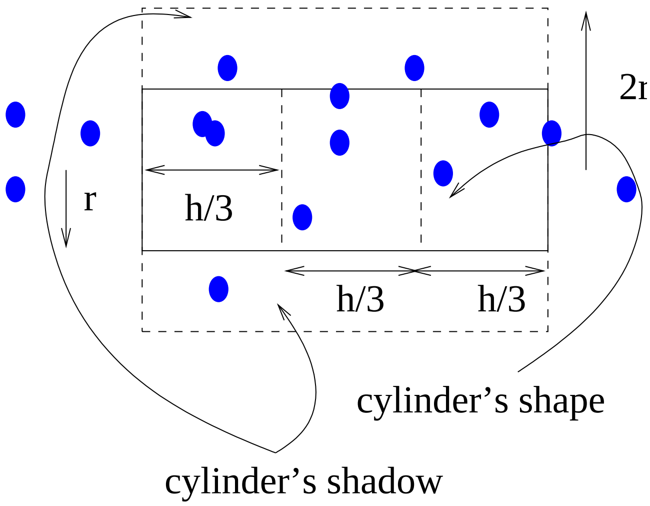

Figure 3: Two-dimensional projection of a thin cylinder with its shadow within a pattern of galaxies.

where $v(\cdot)$ is the potential function associated to a cylinder that depends on $\mathbf{d}$ and the model parametrs.

The cylinder potential is built taking into account local criteria such as the density, spread and number of galaxies. To formulate these criteria, an extra cylinder is attached to each cylinder y, with exactly the same parameters as y, except for the radius which equals $2r$. Let $\tilde{s}(y)$ be the shadow of $s(y)$ obtained by the substraction of the initial cylinder from the extra cylinder, as shown in Fig. 3. Then, each cylinder y is divided in three equal volumes along its main symmetry axis, and we denote by $s_1(y)$, $s_2(y)$ and $s_3(y)$ their corresponding shapes.

The local density condition verifies that the density of galaxies inside $s(y)$ is higher than the density of galaxies in $\tilde{s}(y)$, and it can be expressed as follows:

$$n(\mathbf{d} \cap s(y))/\nu(s(y)) > n(\mathbf{d} \cap \tilde{s}(y))/\nu(\tilde{s}(y)), \tag{6}$$

where $n(\mathbf{d} \cap s(y))$ and $n(\mathbf{d} \cap \tilde{s}(y))$ are the numbers of galaxies covered by the cylinder and its shadow, and $\nu(s(y))$ and $\nu(\tilde{s}(y))$ are the volumes of the cylinder and its shadow, respectively.

The spread condition checks that the galaxies are located everywhere along the cylinder main axis. This is formulated below:

$$\prod_{i=1}^{3} n(\mathbf{d} \cap s_i(y)) > 0, \tag{7}$$

where $n(\mathbf{d} \cap s_i(y))$ is the number of galaxies belonging to $s_i(y)$.

If both these conditions are fulfilled, then $v(y)$ is given by the difference betwen the number of galaxies contained in the cylinder and the number of galaxies contained in its shadow:

$$v(y) = n(\mathbf{d} \cap s(y)) - n(\mathbf{d} \cap \tilde{s}(y)). \tag{8}$$

Whenever any of the previous conditions is violated, a positive value $v_{\max}$ is assigned to the potential of a cylinder.

The parameter $v_{\max}$ gives some very small chances to a segment not fulfilling the required conditions to be a part of the network. This should give more complete networks and better mixing properties to the method.

We note that we have chosen cylinders as the objects here in order to trace filaments in the galaxy distribution. Such objects are tools at our disposal and any object can be chosen; as an example, [31] have built systems of flat elements (walls) and of regular polytopes (galaxy clusters), based on the Bisous process.

2.4 Interaction energy

The interaction energy takes into account the interactions between cylinders. It is the model component ensuring that the cylinders form a filamentary network, and it is given by

$$U_i(\mathbf{y}|\theta) = -n_\kappa(\mathbf{y}) \log \gamma_\kappa - \sum_{s=0}^{2} n_s(\mathbf{y}) \log \gamma_s, \tag{9}$$

where n_κ is the number of repulsive cylinder pairs and n_s is the number of cylinders connected to the network through s extremity points. The variables $\log \gamma_\kappa$ and $\log \gamma_s$ are the potentials associated to these configurations, respectively.

We define the interactions that allow the configuration of cylinders to trace the filamentary network, below. To illustrate these definitions, we show an example configuration of cylinders (in two dimensions) in Fig. 4.

Two cylinders are considered repulsive, if they are rejecting each other and if they are not orthogonal. We declare that two cylinders $y_1 = (k_1, \omega_1)$ and $y_2 = (k_2, \omega_2)$ reject each other if their centers are closer than the cylinder height, $d(k_1, k_2) < h$. Two cylinders are considered to be orthogonal if $|\omega_1 \cdot \omega_2| \leq \tau_\perp$, where $\cdot$ is the scalar product of the two orientation vectors and $\tau_\perp \in (0, 1)$ is a predefined parameter. So, we allow a certain range of mutual angles between cylinders that we consider orthogonal.

Two cylinders are connected if they attract each other, they do not reject each other and they are well aligned. Two cylinders attract each other if only one extremity point of the first cylinder is contained in the attraction region of the other cylinder. The cylinders are "magnetized" in the sense that they cannot attract each other through extremity points having the same index. Two cylinders are well aligned if $\omega_1 \cdot \omega_2 \geq 1 - \tau_\|$, where $\tau_\| \in (0, 1)$ is a predefined parameter.

Take now a look at Fig. 4. According to the previous definitions, we observe that the cylinders $c1$, c_2 and c_3 are connected. The cylinders c_1 and c_3 are connected to the network through one extremity point, while c_2 is connected

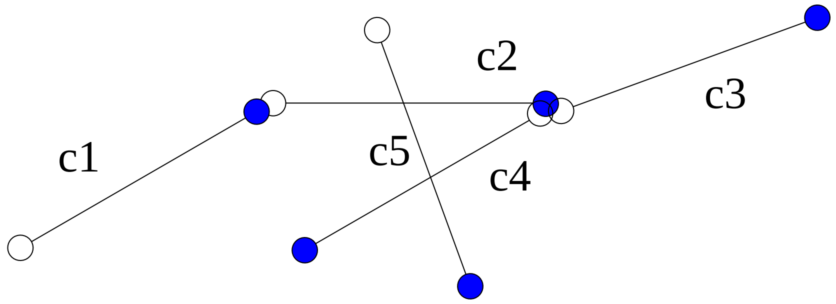

Figure 4: Two-dimensional representation of interacting cylinders.

to the network through both extremity points. The cylinders c_4 and c_5 are not connected to anything – c_4 is not well aligned with c_2, the angle between their directions is too large, and c_5 is not attracted to any other cylinder. It is important to notice that the cylinders c_3 and c_4 are not interacting – they are wrongly 'polarized', their overlapping extremity points have the same index. The cylinder c_5 is rejecting the cylinders c_2 and c_4 (the centres of these cylinders are close), but as it is rather orthogonal both to $c2$ and c_4, it is not repulsing them. The cylinders c_2 and c_4 reject each other and are not orthogonal, so they form a repulsive pair.

Alltogether, the configuration at Fig. 4 adds to the interaction energy contributions from three connected cylinders (one doubly-connected, c_2, and two single-connected, c_1 and c_3), and from one repulsive cylinder pair (c_2–c_4).

The complete model (3) that includes the definitions of the data energy and of the interaction energy given by (5) and (9) is well defined for parameters such that $v_{\max} < 0$, $\gamma_0, \gamma_1, \gamma_2 > 0$ and $\gamma_\kappa \in [0, 1]$. The definitions of the interactions and the parameter ranges chosen ensure that the complete model is locally stable and Markov in the sense of Ripley-Kelly [31]. For cosmologists – it means that we can safely use this model without expecting any dangers (numerical, convergence, etc.).

2.5 Simulation

Several Monte Carlo techniques are available to simulate marked point processes: spatial birth-and-death processes, Metropolis-Hastings algorithms, reversible jump dynamics or more recent exact simulation techniques [12, 11, 13, 15, 35, 36, 24].

In this paper, we need to sample from the joint probability density law $p(\mathbf{y}, \theta)$. This is done by using an iterative Monte Carlo algorithm. An iteration of the algorithm consists of two steps. First, a value for the parameter θ is chosen with respect to $p(\theta)$. Then, conditionally on θ, a disk pattern is sampled from $p(\mathbf{y}|\theta)$ using a Metropolis-Hastings algorithm [12, 11].

The Metropolis-Hastings algorithm for sampling the conditional law $p(\mathbf{y}|\theta)$ has a transition kernel based on three types of moves. The first move is called *birth* and it proposes to add a new cylinder to the present configuration. This new cylinder can be added uniformly in K or can be randomly connected with the rest of the network. This mechanism helps to build a connected network. The second move is called *death*, and it proposes to eliminate a randomly chosen cylinder. The role of this second move is to ensure the detailed balance of the simulated Markov chain and its convergence towards the equilibrium distribution. In order to improve the mixing properties of the sampling algorithm a third move can be added. This third move is called *change*; it chooses randomly a cylinder in the configuration and proposes to "slightly " change its parameters using simple probability distributions. For specific details concerning the implementation of this dynamics we recommend [17, 31].

Whenever the maximisation of the joint law $p(\mathbf{y}, \theta)$ is needed, the previously described sampling mechanism can be integrated into a simulated annealing algorithm. The simulated annealing algorithm is built by sampling from $p(\mathbf{y}, \theta)^{1/T}$, while T goes slowly to zero. [31] proved the convergence of such simulated annealing for simulating marked point processes, when a logarithmic cooling schedule is used. According to this result, the temperature is lowered as

$$T_n = \frac{T_0}{\log n + 1}; \tag{10}$$

we use $T_0 = 10$ for the initial temperature.

2.6 Statistical inference

One straightforward application of the simulation dynamics is tthe estimation of the filamentary structure in a field of galaxies together with the parameter estimates. These estimates are given by :

$$\begin{aligned}(\widehat{\mathbf{y}}, \widehat{\theta}) &= \arg\max_{\Omega\times\Psi} p(\mathbf{y}, \theta) = \arg\max_{\Omega\times\Psi} p(\mathbf{y}|\theta)p(\theta) \\ &= \arg\min_{\Omega\times\Psi}\{\frac{U_{\mathbf{d}}(\mathbf{y}|\theta) + U_i(\mathbf{y}|\theta)}{\alpha(\theta)} + \frac{U_p(\theta)}{\alpha_p(\theta)}\},\end{aligned} \tag{11}$$

where $\alpha(\theta)$ is the normalizing constant, $p(\theta) = \exp[-U_p(\theta)]/\alpha_p(\theta)$ is the prior law for the model parameters and Ψ is the model parameters space.

However, the solution we obtain is not unique. In practise, the shape of the prior law $p(\theta)$ may influence the solution, making the result to look more random compared with a result obtained for fixed values of parameters. Therefore, it is reasonable to wonder how precise the estimate is, that is if an element of the pattern really belongs to the pattern, or if its presence is due to random effects [30, 33])

For compact subregions $\mathcal{R} \subseteq K$ of finite volume, we can compute or give Monte Carlo approximations for average quantities such as

$$\mathrm{E}_{(\mathbf{Y},\Theta)}\left[f(\mathcal{R}, Z(\mathbf{Y}))\right], \tag{12}$$

where E denotes the expectation value over the data and model parameter space, and $f(\mathcal{R}, \cdot)$ is a real measurable function with respect to the σ-algebra associated to the configuration space of the marked point process.

If $f(\mathcal{R}, Z(\mathbf{Y})) = \mathbf{1}\{\mathcal{R} \subseteq Z(\mathbf{Y})\}$, then the expression (12) represents the probability of how often the considered model includes or visits the region $\mathcal{R}$. Furthermore, if K is partitioned into a finite collection of small disjoint cells $\{\mathcal{R}_1, \mathcal{R}_2, \ldots, \mathcal{R}_q\}$, then a visit probability map can be obtained. This map is given by the partition together with the value $P_i = \mathrm{E}\left[\mathbf{1}\{\mathcal{R}_i \subseteq Z(\mathbf{Y})\}\right]$ associated to each cell. The map is defined by the model and by the parameters of the simulation algorithm; its resolution is given by the cell partition.

The sufficient statistics of the model (9) – the interaction parameters n_κ and $n_s, s = (0, 1, 2)$ – describe the size of the filamentary network and quantify the morphological properties of the network. Therefore, they are suitable as a general characterisation of the filamentarity of a galaxy catalogue. Hence, comparing the networks of different regions and/or different catalogues is perfectly possible. Here, we use the sufficient statistics for characterising the real data and the mock catalogues.

The visit maps show the location and configuration of the filament network. Still, the detection of filaments and this verification test depend on the selected model. It is reasonable to ask if these results are obtained because the data exhibits a filamentary structure or just because of how the model parameters are selected.

The sufficient statistics can be used to build a statistical test in order to answer the previous question. For a given data catalog, samples of the model are obtained, so means of the sufficient statistics can be then computed. The same operation, using exactly the same model parameters, can be repeated whenever an artificial point field – or a synthetic data catalog – is used. If the artificial field is the realisation of a binomial point process having the same number of points as the number of galaxies in the original data set, the sufficient statistics are expected to have very low values – there is no global structure in a such binomial field. If the values of the sufficient statistics for these binomial fields were large, this would mean that the filamentary structure is due to the parameters, not to the data. Comparing the values obtained for the original data sets with Monte Carlo envelopes found for artificial point fields, we can compute Monte Carlo p-values for testing the hypothesis of the existence of the filamentary structure in the original data catalogue [30, 33]).

3 Data

We apply our algorithms to a real data catalogue and compare the results with those obtained for twenty two mock catalogues, specially generated to simulate all main features of the real data.

3.1 Observational data

At the moment there are two large galaxy redshift (spatial position) catalogs that are natural candidates for filament search. When the work reported here was carried out a few years ago, the best available redshift catalogue to study the morphology of the galaxy distribution was the 2dF Galaxy Redshift Survey [**?**, 2dFGRS,]]2dFGRS; the much larger Sloan Digital Sky Survey (SDSS) [**?**, see the description of its final status in]]SDSS7 was yet in its first releases. Also, only the 2dFGRS had at that time a collection of mock catalogues that were specially generated to mimic the observed data. So this study is based on the 2dFGRS; we shall certainly apply our algorithms to the SDSS in the future, too.

The 2dFGRS covers two separate regions in the sky, the NGP (North Galactic Cap) strip, and the SGP (South Galactic Cap) strip, with a total area of about 1500 square degrees. The nominal (extinction-corrected) magnitude limit of the 2dFGRS catalogue is $b_j = 19.45$; reliable redshifts exist for 221414 galaxies. The effective depth for the catalogue is about $z = 0.2$ or $D \approx 600\,h^{-1}\,\text{Mpc}$.

The 2dFGRS catalogue is a flux-limited catalogue and therefore the density of galaxies decreases with distance. For statistical analysis of such surveys, a weighting scheme that compensates for the missing galaxies at large distances, has to be used. However, such a weighting is suitable only for specific statistical problems, as, e.g., the calculation of correlation functions. When studying the local structure, such a weighting cannot be used; it would only amplify the shot noise.

We can eliminate weighting by using volume-limited samples. The 2dF team has generated these for scaling studies [**?**, see, e.g.,]]croton1; they kindly sent these samples to us. The volume-limited samples are selected in one-magnitude intervals; we chose as our sample that with the largest number of galaxies, for the absolute magnitude interval $\mathrm{M}_b \in [-19.0, -20.0]$. The total number of galaxies in this sample is 44713.

The borders of the two volumes covered by the sample are rather complex. As our algorithm is recent, we have not yet the estimates of the border effects, and we cannot correct for these. So we limited our analysis to the simplest volumes – bricks. As the Southern half of the galaxy sample has a convex geometry (it is limited by two conical sections of different opening angles), the bricks that is possible to cut from there have small volumes. Thus we used only the Northern data that has a geometry of a slice, and chose the brick of a maximum volume that could be cut from the slice. We will compare below the results obtained for this sample (2dF) by those obtained for a smaller sample in a previous paper (NGP250); the geometry and galaxy content of these data two sets is described in Table 1.

3.2 Mock catalogues

We compare the observed filaments with those built for mock galaxy catalogues that try to simulate the observations as closely as possible. The construction of these catalogues is described in detail by [20]; we give a short summary here.

sample	N_{gal}	depth	width	height	d
2dF	8487	133.1	254.0	31.1	5.0
NGP250	7588	88.6	169.1	20.7	3.4

Table 1: Galaxy content and geometry for the data bricks (sizes are in h^{-1}Mpc). N_{gal} is the number of galaxies in the sample, and d is the mean distance between galaxies in the sample.

The 2dF mock catalogues are based on the "Hubble Volume" simulation [7], a N-body simulation of a $3h^{-1}$ Gpc cube of 10^9 mass points. These mass points are considered as galaxy candidates, and are sampled according to a set of rules that include:

1. Biasing: the probability for a galaxy to be selected is calculated on the basis of the smoothed (with a $\sigma = 2h^{-1}$Mpc Gaussian filter) final density. This probability (biasing) is exponential [**?**, rule 2 of]]Cole98, with parameters chosen to reproduce the observed power spectrum of galaxy clustering.

2. Local structure: the observer is placed in a place similar to our local cosmological neighbourhood.

3. A survey volume is selected, following the angular and distance selection factors of the real 2dFGRS.

4. Luminosity distribution: luminosities are assigned to galaxies according to the observed (Schechter) luminosity distribution; $k + e$-corrections are added.

These "ideal"catalogues are then combined with observational errors to produce the final mock catalogues:

1. Galaxy redshifts are modified by adding random dynamical velocities.

2. Observational random errors are added to galaxy magnitudes.

3. Based on galaxy positions, survey incompleteness factors are calculated.

. These catalogues are as close to the observed catalogues as currently possible – the spatial coverage, galaxy density, clustering, luminosities, observational errors are the same. So, we expect that the filamentary structure of the mock catalogues should be close to that we observe.

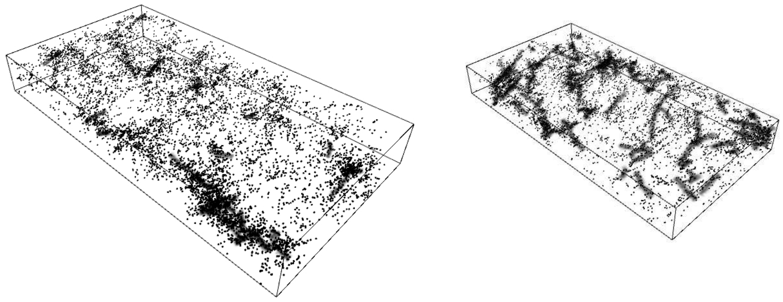

Figure 5: Filaments in the main data set 2dF (left panel) and in a smaller, but more dense data set N250 (right panel). The volumes are shown at the same scale.

4 Filaments

4.1 Experimental setup

As described above, we use the data sets drawn from the galaxy distribution in the Northern subsample of the 2dFGRS survey and from the 22 mock catalogues. For mock catalogues, we use the same absolute magnitude range and cut the same bricks that from the 2dFGRS survey.

The sample region K is the brick. In order to choose the values for the dimensions of the cylinder we use the physical dimensions of the galaxy filaments that have been observed in more detail [22, e.g.,]; we used the same values also in our previous paper [33]: $r = 0.5$, $h = 6.0$ (all sizes are in h^{-1}Mpc). The radius of the cylinder is close to the minimal one can choose, taking into account the data resolution. Its height is also close to the shortest possible, as our shadow cylinder has to have a cylindrical geometry, too (the ratio of its height to the diameter is presently 3:1). We choose the attraction radius as $r_a = 0.5$, giving the value 1.5 for the maximum distance between the connected cylinders, and for the cosines of the maximum curvature angles we choose $\tau_{\|} = \tau_{\perp} = 0.15$. This allows for a maximum of $\approx 30^\circ$ between the direction angles of connected cylinders, and considers the cylinders orthogonal, if the angle between their directions is larger than $\approx 80^\circ$.

For the data energy, we limit the parameter domain by $u_{\max} = [-25, 20]$. For the interaction energy, we choose the parameter domain as follows: $\log \gamma_0 \in [-12.5, -7.5]$, $\log \gamma_1 \in [-5, 0]$ and $\log \gamma_2 \in [0, 5]$. The hard repulsion parameter is $\gamma_k = 0$, so the configurations with repulsing cylinders are forbidden. The domain of the connection parameters was chosen such that 2-connected cylinders are generally encouraged, 1-connected cylinders are penalised and 0-connected segments are strongly penalised. This choice encourages the cylinders to group in filaments in those regions where the data energy is good enough. Still, we have no information about the relative strength of those parameters. Therefore, we have decided to use for the prior parameter density $p(\theta)$ the uniform law

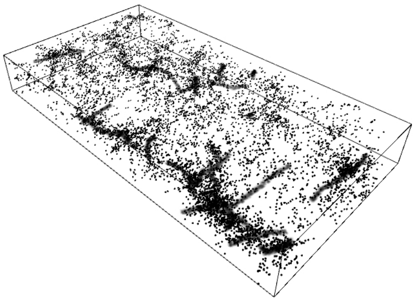

Figure 6: Filaments in the main data set 2dF, for the rescaled basic cylinder.

over the parameter domain.

4.2 Observed filaments

We ran the simulated annealing algorithm for 250000 iterations; samples were picked up every 250 steps. We ran the simulated annealing algorithm for 250000 iterations.

The cylinders obtained after running the simulated annealing outline the filamentary network. But as simulated annealing requires infinitely many iterations till convergence, and also because of the fact that an infinity of solutions are proposed (slightly changing the orientation of cylinders gives us another solution that is as good as the original one), we shall use visit maps to "average" the shape of the filaments.

Fig. 5 shows the cells that that have been visited by our model with a frequency higher than 50%, together with the galaxy field. Filamentary structure is seen, but the filaments tend to be short, and the network is not very well developed. For comparison, we show a similar map for the smaller volume (NGP250), where the galaxy density is about three times higher. We see that the effectiveness of the algorithm depends strongly on the galaxy density; too much a dilution destroys the filamentary structure.

A simple way to remedy the situation is to rescale the basic cylinder; as the density difference is three times, we rescaled the cylinder dimensions by $3^{1/3} = 1.44$. The filamentary network for this case is shown in Fig. 6. This is better developed, but not so well delineated as that for the smaller volume. Thus, if we want to effectively search for filaments, we need to know the positions of less luminous galaxies, also.

A problem that has been adressed in most of the papers about galaxy filaments is the typical filament length (or the length distribution). As our algorithm allows branching of filaments (cylinders that are approximately orthogonal), it is difficult to separate filaments. Instead of that, we find easily the total volumes of filaments, counting the cells on the visiting map. There are problems where this characteristic of the filamentary network is the most important one, as in the search for missing baryons. These are thought to be hidden as warm

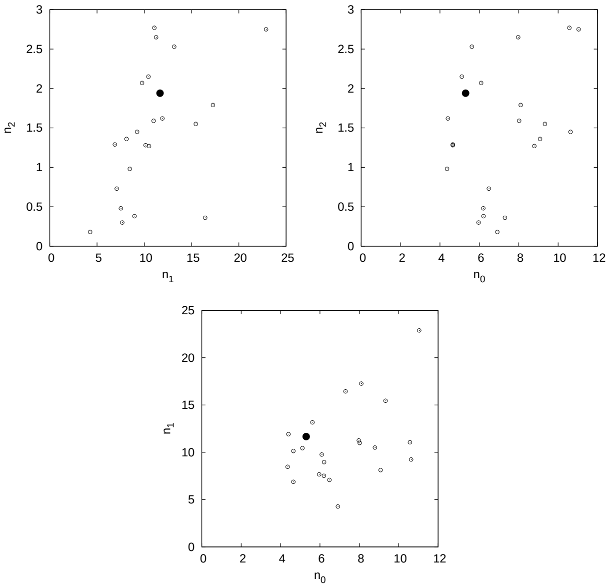

Figure 7: Filamentarity of the real data (black dot) compared with the mocks (open circles). $\bar{n}_2$ is the mean number of the 2-connected cylinders, $\bar{n}_1$ – the mean number of the 1-connected cylinders and $\bar{n}_0$ – the mean number of single cylinders.

intergalactic gas (WHIM, [37, see, e.g.,]), and knowledge of the total volume of filaments available to this gas is certainly important. As an example, for the cases considered here, the relative volumes are $1.8 \cdot 10^{-3}$ (2df, smaller cylinder), $3.3 \cdot 10^{-3}$ (2df, rescaled cylinder), and $1.6 \cdot 10^{-2}$ (NGP250).

4.3 Statistics

As we explained before, in order to compare the filamentarity of the observed data set (2dF) and the mocks, we had to run the Metropolis-Hastings algorithm at a fixed temperature $T = 1.0$ (sampling from $p(\mathbf{y}, \theta)$). The algorithm was run for 250000 iterations, and samples were picked up every 250 iterations. The means of the sufficient statistics of the model were computed using these samples. The obtained results are shown in Table 2.

As we see, the range of the interaction parameters for the 22 mocks covers the data value, although the mocks differ seriously between themselves, and the statistics are more on the lower side (the filamentary network is less developed than in reality). We illustrate the situation in Fig. 7. As an example, we compare the single-temperature visit map for the data with two extreme cases for the mocks (8 and 16) in Fig. 8.

In order to see the influence of rescaling to the sufficient statistics, we re-

Data sets	Sufficient statistics		
	$\bar{n_2}$	$\bar{n_0}$	$\bar{n_1}$
2dF	1.94	5.30	11.66
MOCK 1	2.53	5.62	13.16
MOCK 2	0.48	6.20	7.52
MOCK 3	1.29	4.65	6.88
MOCK 4	1.55	9.33	15.45
MOCK 5	1.45	10.63	9.24
MOCK 6	0.38	6.21	8.96
MOCK 7	1.36	9.08	8.12
MOCK 8	0.18	6.91	4.27
MOCK 9	2.07	6.09	9.76
MOCK 10	1.62	4.40	11.91
MOCK 11	1.28	4.65	10.14
MOCK 12	2.65	7.97	11.25
MOCK 13	0.73	6.48	7.08
MOCK 14	0.36	7.30	16.44
MOCK 15	0.98	4.36	8.47
MOCK 16	2.75	11.04	22.88
MOCK 17	0.30	5.96	7.67
MOCK 18	2.15	5.11	10.44
MOCK 19	1.59	8.02	10.99
MOCK 20	1.27	8.79	10.50
MOCK 21	2.77	10.57	11.06
MOCK 22	1.79	8.10	17.26

Table 2: The mean of the sufficient statistics for the data and the mocks: $\bar{n_2}$ is the mean number of the 2-connected cylinders, $\bar{n_1}$ is the mean number of the 1-connected cylinders and $\bar{n_0}$ is the mean number of the 0-connected cylinders.

Data sets	Sufficient statistics		
	$\bar{n_2}$	$\bar{n_0}$	$\bar{n_1}$
NGP250	11.31	32.76	56.15
2dF	7.13	6.72	33.43
MOCK 8	1.53	9.57	12.64
MOCK 16	6.67	12.48	37.81

Table 3: The mean of the sufficient statistics for the data and the mocks, for the rescaled basic cylinder. The columns are the same as in the previous table.

Binomial data sets	Sufficient statistics		
	$\max \bar{n_2}$	$\max \bar{n_0}$	$\max \bar{n_1}$
MOCK 1	0	0.02	0
MOCK 2	0	0.015	0
MOCK 3	0	0.01	0
MOCK 5	0	0.015	0
MOCK 6	0	0.03	0
MOCK 7	0	0.02	0
MOCK 8	0	0.015	0

Table 4: The maximum of mean of the sufficient statistics over binomial fields generated for some mock catalogs (the same number of points): $\max \bar{n_2}$ is the maximum mean number of the 2-connected cylinders, $\max \bar{n_1}$ is the maximum mean number of the 1-connected cylinders and $\max \bar{n_0}$ is the maximum mean number of the 0-connected cylinders.

peated the procedure with the rescaled cylinder for the data and for the mocks 8 and 16. The data are given in Table 3.

Rescaling the basic cylinder improves the network, but not so much as expected – the interaction parameters remain lower than those obtained for the NGP250 sample.

In order to see if the filamentary network we find is really hidden in the data, we re-distributed uniformly the points inside the domain K. So, the points follow a binomial distribution that depends only on the total number of points. For each (mock) data set this operation was done 100 times, hence obtaining 100 point fields. For each point field the method was launched during 50000 iterations at fixed $T = 1.0$, while samples were picked up every 250 iterations. The model parameters were the same as previously described. The mean of the sufficient statistics was then computed. The maximum values for the all 100 means for each data set are shown in Table 4.

As we see, for a random distribution the algorithm does not find any connected cylinders, both the numbers of the 1-connected and 2-connected cylinders are strictly zero. Only in a few cases the data allows to place a single cylinder. Thus, the filaments our algorithm discovers in galaxy surveys and in mock catalogs are real, they are hidden in the data, and are not the result of a lucky

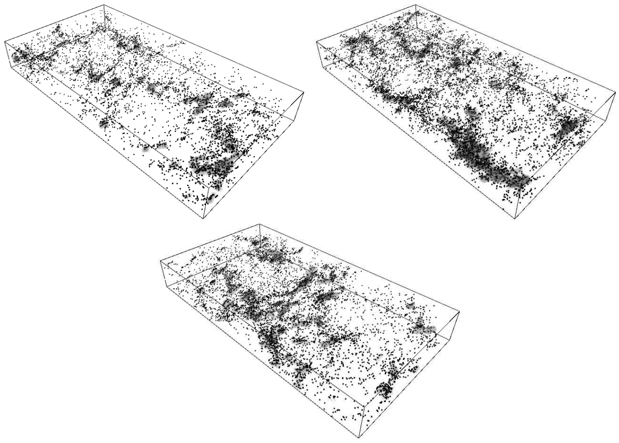

Figure 8: Visit maps for two mocks (mock8 – left panel, mock16 – right panel) and the real data (middle panel).

choice of the model parameters.

5 Conclusions and perspectives

In a previous paper [33] we developed a new approach to locate and characterize filaments hidden in three-dimensional point fields. We applied it to a galaxy catalogue (2dFGRS), found the filaments and described their properties by the sufficient statistics (interaction parameters) of our model.

As there are numerical models (mocks) that are carefully constructed to mimic all local properties of the 2dFGRS, we were interested if these models have also global properties similar to the observed data. An obvious test for that is to find and compare the filamentary networks in the data and in the mocks. We did that, using fixed shape parameters for the basic building blocks for the filaments, and fixed interaction potentials. These priors had led to good results before.

In order to strictly compare the observed catalogue and the mocks, we had to work with constant-density samples (volume-limited catalogues). This inevitably led to a smaller spatial density, and the filament networks we recovered were not so good as those found in the previous paper. Rescaling the basic cylinder helped, but not so much as expected.

There are several new results in our paper:

1. The filamentarity of the real galaxy catalogue, as described by the sufficient statistics of our model (the interaction parameters), lies within the range covered by the mocks.

2. The filamentarity of the mocks themselves differs greatly, showing that

local properties of the galaxy distribution are insufficient to model all the features of the large-scale structure.

3. Filamentarity models based on the Bisous process describe well the filament network, and can be used to construct better mock catalogues.

4. Finally, we compared our catalogues with the random (binomial) catalogues with the same number of data points and found that these do not exhibit any filamentarity at all. This proves that the filaments we find exist in the data.

There are many ways to improve on the work we have done so far. We have seen above that it is difficult to find the scale (lengths) of the filaments for our model; this problem has to be solved. Second, we have used fixed parameters for the data term (cylinder sizes); these should be found from the data. Third, the filament network seems to be hierarchical, with filaments of different widths and sizes; a good model should include this. Fourth, parameter estimation and detection validation should be also included ; the uniform law does not allow the characterization of the model parameters distribution and for the moment we cannot say that the detected filamentary pattern is correctly detected ; the only statistical statement that we can do is that this pattern is hidden in the data and we have some good ideas about where it can be found ; but we do not give any precision measure about it.

Also, it would be good if our model could be extended to describe inhomogeneous point processes – magnitude-limited catalogues that have much more galaxies and where the filaments can be traced much better. The first rescaling attempt we made in this paper could be astep in this direction, but as we saw, it is not perfect. And, as usual in astronomy – we would understand nature much better, if we had more data. The more galaxies we see at a given location, the better can we trace their large-scale structure.

The Bayesian framework and the theory of marked point process allow the mathematical formulation for filamentary pattern detection methodologies introducing the previously mentioned improvements (inhomogeneity, different size of objects, parameter estimation, etc.). The numerical implementation and the construction of these improvements in harmony with the astronomical observations and theoretical knowledge are open and challenging problems.

Acknowledgements

This work has been supported by the University of Valencia through a visiting professorship for Enn Saar, by the Spanish MCyT project AYA2003-08739-C02-01 (including FEDER), by the Estonian Ministry of Education and Science, research project SF0060067s08, and by the Estonian Science Foundation grant 8005. We thank D. Croton for the observational data (the 2dFGRS volume-limited catalogues) and the mock catalogues.

Three-dimensional visualization in this paper was conducted with the S2PLOT programming library [3]. We thank the S2PLOT team for superb work.

References

[1] 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. *Astronomy and Astrophysics*, 474:315–338, October 2007.

[2] M. A. Aragon-Calvo, E. Platen, R. van de Weygaert, and A. S. Szalay. The Spine of the Cosmic Web. *ArXiv e-prints*, September 2008.

[3] D. G. Barnes, C. J. Fluke, P. D. Bourke, and O. T. Parry. An Advanced, Three-Dimensional Plotting Library for Astronomy. *Publications of the Astronomical Society of Australia*, 23:82–93, July 2006.

[4] J. D. Barrow, D. H. Sonoda, and S. P. Bhavsar. Minimal spanning tree, filaments and galaxy clustering. *Monthly Notices of the Royal Astronomical Society*, 216:17–35, 1985.

[5] N. Bond, M. Strauss, and R. Cen. Crawling the Cosmic Network: Exploring the Morphology of Structure in the Galaxy Distribution. *ArXiv e-prints*, March 2009.

[6] J. M. Colberg. Quantifying cosmic superstructures. *Monthly Notices of the Royal Astronomical Society*, 375:337–347, February 2007.

[7] J. M. Colberg, S. D. M. White, N. Yoshida, T. J. MacFarland, A. Jenkins, C. S. Frenk, F. R. Pearce, A. E. Evrard, H. M. P. Couchman, G. Efstathiou, J. A. Peacock, P. A. Thomas, and The Virgo Consortium. Clustering of galaxy clusters in cold dark matter universes. *Monthly Notices of the Royal Astronomical Society*, 319:209–214, November 2000.

[8] M. Colless, G. Dalton, S. Maddox, W. Sutherland, P. Norberg, S. Cole, J. Bland-Hawthorn, T. Bridges, R. Cannon, C. Collins, W. Couch, N. Cross, K. Deeley, R. De Propris, S. P. Driver, G. Efstathiou, R. S. Ellis, C. S. Frenk, K. Glazebrook, C. Jackson, O. Lahav, I. Lewis, S. Lumsden, D. Madgwick, J. A. Peacock, B. A. Peterson, I. Price, M. Seaborne, and K. Taylor. The 2dF Galaxy Redshift Survey: spectra and redshifts. *Monthly Notices of the Royal Astronomical Society*, 328:1039–1063, December 2001.

[9] H. K. Eriksen, D. I. Novikov, P. B. Lilje, A. J. Banday, and K. M. Górski. Testing for Non-Gaussianity in the Wilkinson Microwave Anisotropy Probe Data: Minkowski Functionals and the Length of the Skeleton. *Astrophysical Journal*, 612:64–80, September 2004.

[10] J. E. Forero-Romero, Y. Hoffman, S. Gottloeber, A. Klypin, and G. Yepes. A Dynamical Classification of the Cosmic Web. *ArXiv e-prints*, September 2008.

[11] C. J. Geyer. Likelihood inference for spatial point processes. In O. Barndorff-Nielsen, W. S. Kendall, and M. N. M. van Lieshout, editors, *Stochastic geometry, likelihood and computation.* CRC Press/Chapman and Hall, Boca Raton, 1999.

[12] C. J. Geyer and J. Møller. Simulation procedures and likelihood inference for spatial point processes. *Scan. J. Stat.*, 21:359–373, 1994.

[13] P.J. Green. Reversible jump MCMC computation and bayesian model determination. *Biometrika*, 82:711–732, 1995.

[14] J. Illian, A. Penttinen, H. Stoyan, and D. Stoyan. *Statistical Analysis and Modelling of Spatial Point Patterns.* John Wiley and Sons Ltd., 2008.

[15] W. S. Kendall and J. Møller. Perfect simulation using dominating processes on ordered spaces, with application to locally stable point processes. *Adv. Appl. Prob.*, 32:844–865, 2000.

[16] C. Lacoste, X Descombes, and J. Zerubia. Point processes for unsupervised line network extraction in remote sensing. *IEEE Trans. Pattern Analysis and Machine Intelligence,*, 27:1568–1579, 2005.

[17] M. N. M.van Lieshout and R. S. Stoica. The Candy model revisited: properties and inference. *Statistica Neerlandica*, 57:1–30, 2003.

[18] V. J. Martínez and E. Saar. *Statistics of the Galaxy Distribution.* Chapman & Hall/CRC, Boca Raton, 2002.

[19] J. Møller and R. P. Waagepetersen. *Statistical inference for spatial point processes.* Chapman & Hall/CRC, Boca Raton, 2003.

[20] P. Norberg, S. Cole, C. M. Baugh, C. S. Frenk, I. Baldry, J. Bland-Hawthorn, T. Bridges, R. Cannon, M. Colless, C. Collins, W. Couch, N. J. G. Cross, G. Dalton, R. De Propris, S. P. Driver, G. Efstathiou, R. S. Ellis, K. Glazebrook, C. Jackson, O. Lahav, I. Lewis, S. Lumsden, S. Maddox, D. Madgwick, J. A. Peacock, B. A. Peterson, W. Sutherland, and K. Taylor. The 2dF Galaxy Redshift Survey: the b_J-band galaxy luminosity function and survey selection function. *Monthly Notices of the Royal Astronomical Society*, 336:907–931, November 2002.

[21] D. Novikov, S. Colombi, and O. Doré. Skeleton as a probe of the cosmic web: the two-dimensional case. *Monthly Notices of the Royal Astronomical Society*, 366:1201–1216, 2006.

[22] K. A. Pimbblet and M. J. Drinkwater. Intercluster Filaments of Galaxies Programme: pilot study survey and results. *Monthly Notices of the Royal Astronomical Society*, 347:137–143, January 2004.

[23] K. A. Pimbblet, M. J. Drinkwater, and M. C. Hawkrigg. Intercluster filaments of galaxies programme: abundance and distribution of filaments in the 2dFGRS catalogue. *Monthly Notices of the Royal Astronomical Society*, 354:L61, November 2004.

[24] C. J. Preston. Spatial birth-and-death processes. *Bull. Int. Stat. Inst.*, 46:371–391, 1977.

[25] T. Sousbie, S. Colombi, and C. Pichon. The fully connected N-dimensional skeleton: probing the evolution of the cosmic web. *Monthly Notices of the Royal Astronomical Society*, 393:457–477, February 2009.

[26] T. Sousbie, C. Pichon, S. Colombi, D. Novikov, and D. Pogosyan. The 3D skeleton: tracing the filamentary structure of the Universe. *Monthly Notices of the Royal Astronomical Society*, 383:1655–1670, February 2008.

[27] T. Sousbie, C. Pichon, H. Courtois, S. Colombi, and D. Novikov. The Three-dimensional Skeleton of the SDSS. *Astrophysical Journal Letters*, 672:L1–L4, January 2008.

[28] R. S. Stoica, X Descombes, M. N. M. van Lieshout, and J. Zerubia. An application of marked point processes to the extraction of linear networks from images. In J. Mateu and F. Montes, editors, *Spatial statistics through applications*. WIT Press, Southampton, UK, 2002.

[29] R. S. Stoica, X. Descombes, and J. Zerubia. A Gibbs point process for road extraction in remotely sensed images. *Int. J. Computer Vision*, 57(2):121–136, 2004.

[30] R. S. Stoica, E. Gay, and A. Kretzschmar. Cluster detection in spatial data based on monte carlo inference. *Biometrical Journal*, 49(2):1–15, 2007.

[31] R. S. Stoica, P. Gregori, and J. Mateu. Simulated annealing and object point processes : tools for analysis of spatial patterns. *Stochastic Processes and their Applications*, 115:1860–1882, 2005.

[32] R. S. Stoica, V. J. Martinez, J. Mateu, and E. Saar. Detection of cosmic filaments using the candy model. *Astronomy and Astrophysics*, 434:423–432, 2005.

[33] R. S. Stoica, V. J. Martinez, and E. Saar. A three dimensional object point process for detection of cosmic filaments. *Journal of the Royal Statistical Society : Series C (Applied Statistics)*, 55:189–205, 2007.

[34] D. Stoyan, W. S. Kendall, and J. Mecke. *Stochastic geometry and its applications*. John Wiley and Sons Ltd., 1995.

[35] M. N. M. van Lieshout. *Markov point processes and their applications*. Imperial College Press/World Scientific Publishing, London/Singapore, 2000.

[36] M. N. M. van Lieshout and R. S. Stoica. Perfect simulation for marked point processes. *Computational Statistics and Data Analysis*, 51:679–698, 2006.

[37] M. Viel, E. Branchini, R. Cen, J. P. Ostriker, S. Matarrese, P. Mazzotta, and B. Tully. Tracing the warm-hot intergalactic medium in the local Universe. *Monthly Notices of the Royal Astronomical Society*, 360:1110–1122, July 2005.

[38] N. Werner, A. Finoguenov, J. S. Kaastra, A. Simionescu, J. P. Dietrich, J. Vink, and H. Böhringer. Detection of hot gas in the filament connecting the clusters of galaxies Abell 222 and Abell 223. *Astronomy and Astrophysics*, 482:L29–L33, May 2008.