A THREE DIMENSIONAL OBJECT POINT PROCESS FOR DETECTION OF COSMIC FILAMENTS

BY

Radu S. Stoica
Vicent J. Martinez
Enn Saar

Research Report No. 17
January 2006

Unité de Biométrie
Institut National de la Recherche Agronomique
Avignon, France
http://www.avignon.inra.fr/biometrie

A three dimensional object point process for detection of cosmic filaments

R.S. Stoica [1], V.J. Martinez [2] and E. Saar [3]

[1] *INRA Unité Biomètrie, Domaine St. Paul site Agroparc, 84914 Avignon Cedex 9, France. E-mail : radu.stoica@avignon.inra.fr*

[2] *Observatori Astronòmic de la Universitat de València, Apartat de correus 22085, 46075 València, Spain. Email : martinez@uv.es*

[3] *Tartu Observatoorium, Tõravere, 61602 Estonia. Email : saar@aai.ee*

Summary. We propose to apply an object point process to automatically delineate filaments of the large-scale structure in redshift catalogues. We illustrate the feasibility of the idea on an example of the recent 2dF Galaxy Redshift Survey, describe the procedure, and characterize the results.

Keywords: Object point processes, Bisous model, filaments, cosmology, large-scale structure

1. Introduction

The large-scale structure of the Universe is studied by creating galaxy maps – positions of thousands, a few years ago, and millions, nowadays, of galaxies in space. The angular positions of galaxies are relatively easy to measure, but their distances can be estimated only by measuring their recession velocities. The latter task is difficult, especially for faint distant objects, and thus really detailed maps of galaxies have started to appear only lately. An additional caveat is that the recession velocities contain a contribution from the dynamical velocity of a galaxy, so the apparent distances of galaxies are in error. Such maps are called 'redshift space' maps, but the distance errors are not as serious as to change the overall picture of the large-scale structure.

An overview of such galaxy maps is given in Martínez and Saar (2002). As an example, we present here a map from a recently completed 2dF Galaxy Redshift Survey (2dFGRS, Colless et al. (2001)). This survey measured the redshifts (recession velocities) of galaxies in about 1500 square degrees, up to the distances of about $700\,h^{-1}$Mpc† (corresponding to a redshift $z = 0.2$ for the standard cosmological model). The redshifts were measured in two different regions of the sky; Fig. 1 shows the positions of galaxies in two 2.6° thick slices from both regions.

The dominant feature of this map, as of all other galaxy maps of the large-scale structure of the universe, is the network of filaments of different size and contrast, along with relatively empty voids between the filaments. The filamentary network contains different scales, where

†Distances between galaxies are usually measured in megaparsecs (Mpc); 1Mpc $\approx 3 \cdot 10^{24}$cm. The constant h is the dimensionless Hubble parameter; the latest determinations give for its value $h \approx 0.71$.

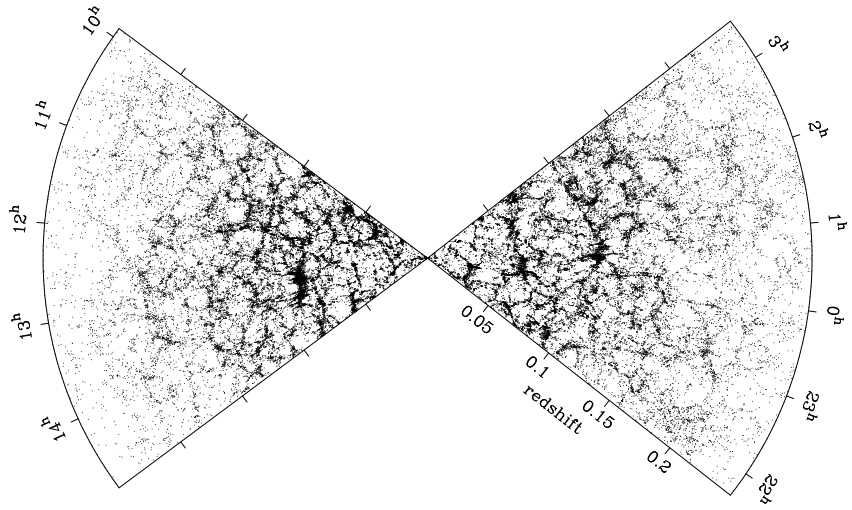

Fig. 1. Galaxy map for two 2dFGRS slices. The observers (we) are situated at the centre of the figure. Both slices are thin, with the thickness of 2.6°. The distances are given in redshifts z; approximately, the physical distance $D \approx 3000\,h^{-1}\,z$ Mpc. The numbers along the arcs show the right ascension (in hours). The filamentary network of galaxies is clearly seen; the disappearance of stucture with depth (towards the sides of the figure) is caused by luminosity selection.

smaller-scale filaments are also less prominent. The gradual disappearance of structures with increasing distance is due to the fact that the apparent luminosity of a galaxy is the fainter the more distant it is, and in more distant regions we can observe only a few of the brightest galaxies.

Although the filaments are prominent, there is no good method to describe such a filamentary structure. The usual second moment methods in real space or in the Fourier space (the two-point correlation and power spectra) do not describe well filamentary structures. The method that has been used most is the minimal spanning tree (MST, see a review in Martínez and Saar (2002). The first application of the MST formalism to describe the filamentary networks of galaxy maps was that of Barrow et al. (1985); many later studies have used it.

The minimal spanning tree is unique for a given point set, which is good, and it connects all points, which is not good. When the number of galaxies is large, the MST is rather fuzzy, and it describes mainly the local nearest-neighbour distribution. The filamentary network seen by the eye combines both local and large-scale features of the point distribution. Thus, a better notion would be that of the skeleton, proposed recently to describe continuous density fields Eriksen et al. (2004). The skeleton is formed by lines parallel to the gradient of the field, which connect the saddle points to local maxima of the field. Calculating the skeleton, however, involves smoothing the point distribution, which will introduce an extra parameter, therefore this method is not well suited for point distributions.

In a previous paper (Stoica et al. (2005)), we proposed to use an automated method to trace filaments for realizations of point processes, that has been shown to work well for detection of road networks in remote sensing situations (Lacoste et al. (2005); Stoica et al. (2002, 2004)). 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.

All the methods used to automatically detect filaments have been developed so far for two-dimensional maps. This is a natural approach for studying the cosmic microwave sky background (Eriksen et al. (2004)), which is two-dimensional, but galaxy maps are three-dimensional. The previous filament studies (e.g., Bharadwaj et al. (2004); Bharadwaj and Pandey (2004); Pandey and Bharadwaj (2005)) have also used two-dimensional galaxy maps, projections for thin slices. The main reason for that is that most of the past large-scale galaxy maps were observed for relatively thin spatial slices; also, in projection filaments seem more prominent. But, of course, both slicing of filaments and projecting the distribution onto a plane distorts the geometry and the properties of the filamentary network.

The studies of the three-dimensional filamentary network has just begun. Three-dimensional filaments have been extracted from galaxy distribution as a result of special observational projects (Pimbblet and Drinkwater (2004)), or by searching for filaments in the 2dFGRS catalogue (Pimbblet et al. (2004)). These filaments have been searched for between galaxy clusters, determining the density distribution and deciding if it is filamentary, individually for every filament. Similar studies have been carried out for N-body simulations (Colberg et al. (2005)). A review of these and previous studies of large-scale filaments is given in (Pimbblet (2005)).

No automated methods to trace filaments in the three-dimensional galaxy maps have been proposed so far. Based on our previous experience with the Candy process, we generalized 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 (Stoica et al. (2005)). We will describe the model below and shall apply it to the samples chosen from the real three-dimensional galaxy distribution, the 2dFGRS catalogue.

2. Object point processes : definitions and manipulation tools

2.1. Definitions

Let $(K, \mathcal{B}, \nu)$ be a measure space, where K is a compact subset of $\mathbb{R}^3$ of strictly positive Lebesgue measure $0 < \nu(K) < \infty$ and $\mathcal{B}$ the associated Borel $\sigma-$algebra of subsets of K. If, to points in K, shape descriptors or marks are attached, objects are formed. Let $(M, \mathcal{M}, \nu_M)$ be the probability measure space of these marks.

The considered configuration space is $\Omega = \cup_{n=0}^{\infty} \Xi_n$, with Ξ_n the set of all unordered configurations $\mathbf{y} = \{(k_1, m_1), (k_2, m_2), \ldots, (k_n, m_n)\}$ that consist of n not necessarily distinct objects $y_i = (k_i, m_i) \in K \times M$. Ξ_0 is the empty configuration. Ω is equipped with the $\sigma-$algebra $\mathcal{F}$ generated by the mappings that count the number of objects in Borel sets $A \subseteq K \times M$.

An object point process with locations of objects in K and shape characteristics in M is a measurable mapping from some probability space into $(\Omega, \mathcal{F})$.

For the sake of simplicity, throughout this paper the shape of an object is defined by a compact set $s(y_i) = s(k_i, m_i)$ that is a subset of $\mathbb{R}^3$ of finite volume $\nu(s(y_i))$. The shape of a pattern $\mathbf{y}$ is defined by the random set $Z(\mathbf{y}) = \cup_{i=1}^{n(\mathbf{y})} s(y_i)$.

The simplest object point process is the Poisson point process. This process is used as a reference measure. It generates a configuration of objects as follows : first the number of objects is chosen according to a Poisson law of intensity $\nu(K)$, then the locations of

objects are distributed uniformly in K and finally the corresponding shapes are chosen independently for each object with respect to ν_M.

The Poisson process does not take into account interactions between objects. Indeed, more realistic models can be constructed by specifying a probability density with respect to the reference measure :

$$p(\mathbf{y}|\theta) = \alpha \exp\left[-U(\mathbf{y}, \theta)\right] \tag{1}$$

with α the normalizing constant, θ the vector of model parameters and $U(\mathbf{y}, \theta)$ the energy function of the system. There is a lot of freedom for constructing the energy function, provided there exists a positive real constant $\Lambda > 0$ such that

$$U(\mathbf{y}, \theta) - U(\mathbf{y} \cup \{(k, m), \theta)\} \leq \log \Lambda \tag{2}$$

The condition (2) is known in the literature as the local stability property and it implies the integrability with respect to the reference measure, of the probability density given by (1). Furthermore, the local stability property is of major importance in establishing convergence proofs for the Monte Carlo dynamics simulating such a model Geyer (1999). The quantity $\exp[U(\mathbf{y}, \theta) - U(\mathbf{y} \cup \{(k, m), \theta)]$ is usually called the Papangelou or the conditional intensity ratio.

For further details related to the definition and the properties of object point processes we recomend the reader the following monographs (van Lieshout (2000); Møller and Waagepetersen (2003); Reiss (1993)).

2.2. Manipulation tools : sampling and inference

Since the normalizing constant α is not available analytically, Monte Carlo Markov Chain techniques are required to sample from $p(\mathbf{y}|\theta)$. Several choices are available (Geyer and Møller (1994); Geyer (1999); Green (1995); Kendall and Møller (2000); van Lieshout (2000); van Lieshout and Stoica (2003b); Preston (1977)). All these methods require local stability, to guarantee the convergence of the equilibrium distribution of the simulated Markov chain towards the probability distribution of the object point process of interest.

The spatial birth-and-death processes require the integration of the Papangelou ratio over the object parameter space $K \times M$. The computation of this integral can be avoided via a thinning operation. In practice, this procedure requires reasonable values for Λ. This is not always the case if models with complex and rather hard interactions are used (van Lieshout and Stoica (2003a); Stoica et al. (2005)). Furthermore, a simulated annealing algorithm makes all the penalizing interactions to become hard at low temperatures. Therefore, this procedure is available in practice only for a limited range of parameters. Very attractive exact simulation methods are based on spatial birth-and-death processes. Hence, they inherit the same drawback previously mentioned (Berthelsen and Møller (2002); van Lieshout and Stoica (2003b)). The Metropolis-Hastings or reversible jump methods are a good compromise solution (Geyer and Møller (1994); Geyer (1999); Green (1995)). Indeed, these last techniques do not indicate themselves when convergence is reached, in exchange they do not exhibit the mentioned drawbacks related to the methods based on spatial birth-and-death processes. Furthermore, Metropolis-Hastings dynamics allows the construction of transition kernels that "help" the model. This is a particulary efficient simulation method, when models that exhibit complex interactions between objects need to be sampled (van Lieshout and Stoica (2003a); Stoica et al. (2005)).

Sampling from the joint law $p(\mathbf{y}, \theta)$ is not trivial. The problem can be considered solved if we are able to simulate $p(\theta|\mathbf{y})$. Clearly, this can be achieved by Monte Carlo simulations. Nevertheless, implementing such a solution requires another simulated Markov chain. This extra dynamics is used to approximate ratios of the normalizing constants of $p(\theta|\mathbf{y})$ when θ explores the parameter space(Geyer and Thompson (1992); Geyer (1994, 1999)). When $\mathbf{y}$ represents the configuration of an un-marked point process, some alternative solutions are proposed in Møller et al. (2004). Still, the extension of these ideas to the case of object point processes needs further investigation.

Under these circumstances, a prior $p(\theta)$ for the model parameters may be chosen. If no a priori knowledge about θ is available, the uniform law on the parameter space is commonly asigned to $p(\theta)$. Hence, the joint law $p(\mathbf{y}, \theta)$ can be sampled using a two-step procedure. First, the θ parameter is chosen according to $p(\theta)$, then a new object configuration is obtained sampling $p(\mathbf{y}|\theta)$.

For the problem on hand, $\mathbf{d}$, the data to be analysed, consist of points (galaxies) spread in a finite volume K. We want to detect the filamentary pattern "hidden" in these data. Two hypothesis are assumed. First, this rather complex pattern is supposed to be formed of simple interacting objects. And second, the filamentary pattern is considered to be the realization of an object point process. The energy function of such a process can be written as follows :

$$U(\mathbf{y}, \theta) = U_{\mathbf{d}}(\mathbf{y}, \theta) + U_i(\mathbf{y}, \theta) \tag{3}$$

where $U_{\mathbf{d}}(\mathbf{y}, \theta)$ is the data energy and $U_i(\mathbf{y}, \theta)$ the interaction energy.

The estimator of the filamentary structure in a field of galaxies together with the parameter estimaes is given by the configuration of objects minimizing the total energy function of the system

$$(\widehat{\mathbf{y}}, \widehat{\theta}) = \arg\max_{\Omega\times\Psi} p(\mathbf{y}, \theta) = \arg\min_{\Omega\times\Psi} \{U_{\mathbf{d}}(\mathbf{y}, \theta) + U_i(\mathbf{y}, \theta) + \log p(\theta)\}$$

with Ψ the model parameters space.

The minimization of the energy function can be performed by means of a simulated annealing algorithm (van Lieshout (1994); Stoica et al. (2005)). This method iteratively samples from $p(\mathbf{y}, \theta)^{1/T}$ while slowly decreasing the temperature parameter T. When the system is frozen *i.e.* $T \to 0$, the simulated annealing samples uniformly on the sub-space of configurations minimizing the energy function (3).

In this case, the solution obtained is not unique. Hence, it is legitimate to ask if an element of the pattern realy belongs to the pattern, or if its presence is due to random effects (Stoica and Gay (2005)).

For compact regions $\mathcal{S} \subseteq \mathbb{R}^3$ of finite volume $0 \leq \nu(\mathcal{S}) < \infty$, we write the probability that an object from the pattern $\mathbf{y}$ covers $\mathcal{S}$, as follows :

$$\begin{aligned}
\mathbb{P}\left(\mathbf{1}\left\{\sum_{i=1}^{n(\mathbf{Y})} \mathbf{1}\{\mathcal{S} \subseteq s(Y_i)\}\right\}\right) &= \\
&= \int_{\Omega\times\Psi} \mathbf{1}\left\{\sum_{i=1}^{n(\mathbf{Y})} \mathbf{1}\{\mathcal{S} \subseteq s(Y_i)\}\right\} d\mathbb{P}(\mathbf{Y}, \Theta) \qquad (4) \\
&= \mathbb{E}_{(\mathbf{Y},\Theta)} \mathbf{1}\left\{\sum_{i=1}^{n(\mathbf{Y})} \mathbf{1}\{\mathcal{S} \subseteq s(Y_i)\}\right\}
\end{aligned}$$

The cover probability given by (4) can be approximated by its Monte Carlo counterpart :

$$\widehat{\mathbb{P}}\left(\mathbf{1}\left\{\sum_{i=1}^{n(\mathbf{Y})}\mathbf{1}\{\mathcal{S}\subseteq s(Y_i)\}\right\}\right)=\frac{1}{U}\sum_{u=1}^{U}\mathbf{1}\left\{\sum_{i=1}^{n(\mathbf{Y}_u)}\mathbf{1}\{\mathcal{S}\subseteq s(Y_{ui})\}\right\} \tag{5}$$

where $\mathbf{Y}_1,...,\mathbf{Y}_U$ are obtained sampling from $p(\mathbf{y},\theta)$. For the pattern $\mathbf{Y}_u$, the set of their corresponding objects is $\{Y_{ui}, i=1,\ldots,n(\mathbf{Y}(u))\}$.

The mark behaviour in the region $\mathcal{S}$ can be analysed by computing the quantity :

$$m_{\mathcal{S}}=\frac{\mathbb{E}_{(Y,\Theta)}\sum_{v=1}^{n(\mathbf{Y})}m_v\mathbf{1}\{\mathcal{S}\subseteq s(Y_v)\}}{\mathbb{E}_{(Y,\Theta)}\sum_{v=1}^{n(\mathbf{Y})}1\{\mathcal{S}\subseteq s(Y_v)\}} \tag{6}$$

with its the corresponding Monte Carlo approximation given by :

$$\widehat{m}_{\mathcal{S}}=\frac{\sum_{u=1}^{U}\sum_{v=1}^{n(\mathbf{Y}_u)}m_v\mathbf{1}\{\mathcal{S}\subseteq s(Y_{uv})\}}{\sum_{u=1}^{U}\sum_{v=1}^{n(\mathbf{Y}_u)}\mathbf{1}\{\mathcal{S}\subseteq s(Y_{uv})\}}$$

The formulas (4) and (6) can be seen as tools for studying the average behaviour in terms of location and shape of the unknown pattern exhibited by the data .

3. Bisous model applied to filamentary pattern detection

Detection of cosmic filaments using marked point processes was first performed on two-dimensional simulated data by Stoica et al. (2005). The marked point process that was used throughout the mentioned work was the Candy model. This marked point process is able to simulate and detect two-dimensional linear networks. Clearly, for real data analysis, a three-dimensional model is needed.

In this paper, we vote for the Bisous model, which is an object point process built to model and analyse general three dimensional spatial patterns (Stoica et al. (2005)). The spatial pattern is made of simple interacting objects or generating elements. The particular property of these objects is that they exhibit a finite number q of rigid extremity points $\{e_1,\ldots,e_q\}$. Around each object y an attraction region $a(y)$ is defined. This attraction region is the disjoint union $a(y)=\cup_{u=1}^{q}b(e_u,r_a)$, where $b(e_u,r_a)$ is the ball of fixed radius r_a centered in e_u. Objects attract each other through their attraction region, getting connected, hence forming the pattern. The object connected to the pattern by means of s extremity points is called $s-$connected. Among the interactions exhibited by the objects forming the pattern we enumerate connectivity, alignment and repulsion. They are to be defined later in the next section of this paper.

For a pattern $\mathbf{y}$ the probability density of the Bisous model can be written as follows :

$$p(\mathbf{y}|\theta)\propto\left[\prod_{i=1}^{n(\mathbf{y})}\beta(y_i)\right]\left[\prod_{s=0}^{q}\gamma_s^{n_s(\mathbf{y})}\right]\prod_{\kappa\in\Gamma}\gamma_\kappa^{n_\kappa(\mathbf{y})} \tag{7}$$

with $\beta: K\times M\to\mathbb{R}_+$ the intensity function and $\gamma_s>0,\gamma_\kappa\in[0,1]$ the interaction model parameters. Γ is a set of interactions between objects that are pairwise, local (distance based), symmetric and guarantee that the density (7) is well defined. For each s and κ

the sufficient statistics $n_s(\mathbf{y})$ and $n_\kappa(\mathbf{y})$ represent, respectively, the number of $s-$connected objects and the number of pairs of objects exhibiting the interaction κ in the configuration $\mathbf{y}$.

The total energy of the model $U(\mathbf{y}, \theta)$ is computed taking the negative logarithm function of (7). We naturally define its two components, the data energy

$$U_{\mathbf{d}}(\mathbf{y}, \theta) = -\sum_{i=1}^{n(\mathbf{y})} \log \beta(y_i) \tag{8}$$

and the interaction energy

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

In the following, we will present the generating element together with its corresponding interactions that define (9), and will define the data energy (8), which is related to the location of the filamentary pattern among the galaxies.

3.1. Generating element and its interactions

The generating element of the filamentary pattern "hidden" in the galaxy data is a random thin cylinder. Such a cylinder has a random center in K and has a fixed radius r and a height h, with $r \ll h$. Its random mark is given by ω, the orientation vector of the cylinder. Its corresponding parameters $\omega = \phi(\eta, \tau)$ are uniformly distributed on $M = [0, 2\pi) \times [0, 1]$ such that

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

A cylinder (k, ω) has $q = 2$ extremity points. In Figure 2 the cylinder is centered in the origin and it its symmetry axis is parallel to Oz. The coordinates of the extremity points are

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

and is the orientation vector $\omega = (0, 0, 1)$. A randomly located and oriented cylinder is obtained by the means of a translation and two rotations (Hearn and Baker (1994)).

Two cylinders $y_1 = (k_1, \omega_1)$ and $y_2 = (k_2, \omega_2)$ attract each other, and we write $y_1 \sim_a y_2$ if the set

$$\{(u, v) : 1 \leq u, v \leq 2, u \neq v, d(e_u(y_1), e_v(y_2)) \leq r_a\}$$

contains only one element.

The same cylinders reject each other, and we write $y_1 \sim_r y_2$ if $d(k_1, k_2) < h$. They are aligned, and we write $y_1 \sim_\parallel y_2$ if

$$\omega_1 \cdot \omega_2 \geq 1 - \tau_\parallel$$

where $\tau_\parallel \in (0, 1)$ is a predefined curvature parameter and $\cdot$ designes the scalar product of the two orientation vectors. If

$$|\omega_1 \cdot \omega_2| \leq \tau_\perp$$

with $\tau_\perp \in (0, 1)$, they are said to be orthogonal, and we write $y_1 \sim_\perp y_2$.

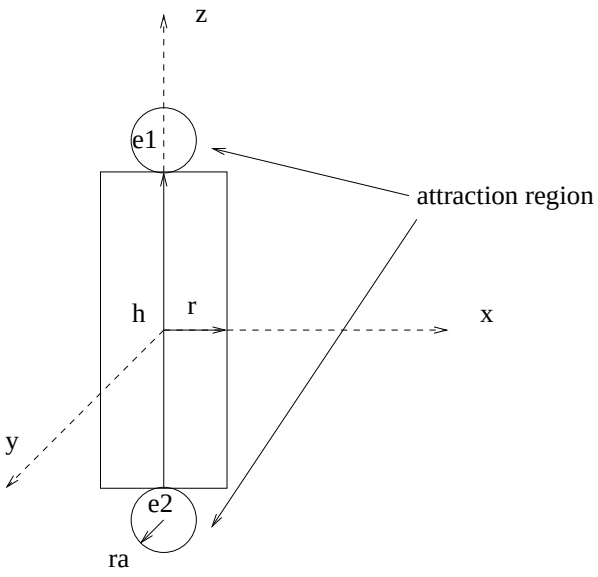

Fig. 2. Generating element for the filamentary pattern.

Two cylinders y_1 and y_2 are connected and we write $y_1 \sim_s y_2$ if the following conditions are simultaneously fulfiled :

$$y_1 \sim_a y_2$$
$$y_1 \not\sim_r y_2$$
$$y_1 \sim_{\parallel} y_2$$

If the following conditions are simultaneously verified :

$$y_1 \sim_r y_2$$
$$y_1 \not\sim_{\perp} y_2$$

the cylinders exhibit a hard repulsion, and we write $y_1 \sim_h y_2$.

The filaments forming the pattern are made of cylinders that attract each other and they are well aligned. These filaments may cross. In the same time configurations with overlapping cylinders are not desirable since they may outline clusters of galaxies instead of filaments. With respect to these considerations, all the ingredients needed for the natural construction of the interaction energy (9) are defined. The $s-$connectivity is constructed using the relation $\sim_s$. The Γ set contains here only one element given by $\sim_h$. The model parameters for the interaction energy are $\gamma_0, \gamma_1, \gamma_2 > 0$ and $\gamma_h = 0$. The definition of the interactions and the parameter ranges ensure the object point process induced by the interaction energy to be well defined, locally stable and Markov in the sense of Ripley-Kelly (Stoica et al. (2005)).

3.2. The data energy

The data energy is related to the position of the filamentary pattern in the galaxy field. To each cylinder y an extra cylinder is attached, so that it has 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 indicated in Figure 3. The cylinder y is divided along its main symmetry axis, in three equal volumes, and we denote $s_1(y)$, $s_2(y)$ and $s_3(y)$ their corresponding geometrical shapes.

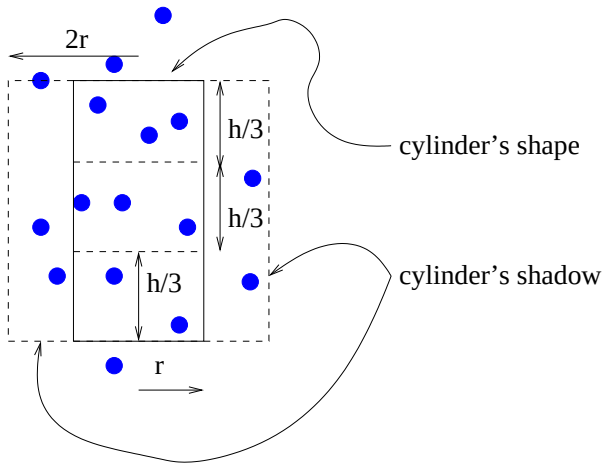

Fig. 3. A two-dimensional view of a cylinder with its shadow in a pattern of galaxies.

A cylinder may belong to the pattern if several conditions are verified. First, the density of galaxies inside $s(y)$ has to be higher than the density of galaxies in $\tilde{s}(y)$, and it can be written as follows:

$$\mathbf{1}\{\text{“density”}\} = \mathbf{1}\{n(\mathbf{d} \cap s(y))\nu(\tilde{s}(y)) > n(\mathbf{d} \cap \tilde{s}(y))\nu(s(y))\}$$

where $n(\mathbf{d} \cap s(y))$ and $n(\mathbf{d} \cap \tilde{s}(y))$ represent the number of galaxies covered by the cylinder and its shadow, respectively. Next, the galaxies covered have to be spread inside the whole volume. This is formulated below :

$$\mathbf{1}\{\text{“spread”}\} = \prod_{i=1}^{3} \mathbf{1}\{n(\mathbf{d} \cap s_i(y)) > 1\}$$

with $n(\mathbf{d} \cap s_i(y))$ the number of galaxies belonging to $s_i(y)$. Under these assumptions, $u(y)$ the energy contribution of a cylinder is defined as follows :

$$u(y) = \mathbf{1}\{\text{“density”}\}\mathbf{1}\{\text{“spread”}\}(n(\mathbf{d} \cap s(y)) - n(\mathbf{d} \cap \tilde{s}(y) + u_{\max}) - u_{\max} \qquad (10)$$

where $u_{\max}$ is a positive fixed quantity.

The data energy (8) is obtained by summing the energy contributions (10) for all the cylinders in the pattern :

$$U_{\mathbf{d}}(\mathbf{y}, \theta) = -\sum_{i=1}^{n(\mathbf{y})} u(y_i). \qquad (11)$$

It can be checked that the data energy term (11) induces a locally stable object point process.

4. Data

4.1. Observational data

The best available redshift catalog to study morphology of the galaxy distribution at present is the 2dF Galaxy Redshift Survey (2dFGRS) (Colless et al., 2001). The catalogue covers two separate regions in the sky, the NGP (North Galactic Cap) strip, and the SGP (South Galactic Cap) strip, with total area of about 1500 square degrees. The nominal (extinction-corrected) magnitude limit of the catalogue is $b_j = 19.45$; reliable redshifts were obtained for 221414 galaxies. As seen in Fig. 1, the effective depth of the catalogue is about $z = 0.2$ or $D \approx 600\,h^{-1}\,\mathrm{Mpc}$, and the total volume $V \approx 3.3 \cdot 10^7 h^{-3}\,\mathrm{Mpc}^3$.

As all galaxy redshift catalogues, the catalogue is not strictly homogeneous. First, due to different observing conditions, the magnitude limit (the depth of the catalogue) changes from observation to observation, but the changes are known and well documented. Secondly, due to the fact that the fibers that lead the light from the focal plane to the spectrograph have a finite size, the redshifts of a few per cent of selected galaxies could not be measured. This effect is quantified as redshift completeness. All the data and the programs to calculate the magnitude limits and the spectroscopic completeness are public and can be found at `http://www.mso.anu.edu.au/2dFGRS/`.

The 2dFGRS catalogue is flux-limited catalog 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. Usually, each galaxy is weighted by the inverse of the selection function (Martínez and Saar, 2002). The selection function is determined by the two main selection effects described above. 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.

At the cost of discarding many surveyed galaxies, one can alternatively use volume-limited samples. In this case, the variation in density at different locations depends only on the fluctuations of the galaxy distribution itself. We started our analysis by using the volume-limited samples prepared by the 2dF team for scaling studies (Croton et al. (2004)) and kindly sent to us by Darren Croton. Unfortunately, we found soon that the galaxy density in these samples was too low and the shot noise did not allow algorithms to find filaments; there are, typically, only of the order of ten galaxies per filament is such samples.

Thus we choose another way – we started from a full galaxy sample, and choose from that bricks in a limited distance range, where the galaxy density was approximately constant. As the SGP half of the galaxy sample has a convex geometry (it is limited by two two conical sections of different opening angles), the constant density bricks that is possible to cut from the SGP have small volumes. Thus we used only the NGP data, and chose three bricks of maximum length and height, with different depths. General characteristics of these data sets are listed in Table 1. As seen from the (comoving) galaxy density histogram in Fig. 4, the densities inside the bricks are approximately constant.

We show the arrangement of the bricks inside the 2dFGRS NGP sample limits and the spatial distribution of galaxies in Fig. 5.

4.2. Experimental results

As described above, we use three data sets, drawn from the galaxy distribution in the Northern subsample of the 2dGGRS survey. The size of the brick determines the size of

Table 1. Galaxy content and geometry for the 2dFGRS bricks (in h^{-1}Mpc). Column 7 ($D_{\rm max}$) shows the maximum distance from the observer (to the far-away brick corners) and d is the mean distance between galaxies in the sample.

sample	galaxies	near side	depth	width	height	$D_{\rm max}$	d
n150	2499	87.9	53.1	101.5	12.4	150.0	2.99
n200	4180	117.2	70.9	135.3	16.5	200.0	3.36
n250	7588	146.4	88.6	169.1	20.7	250.0	3.44

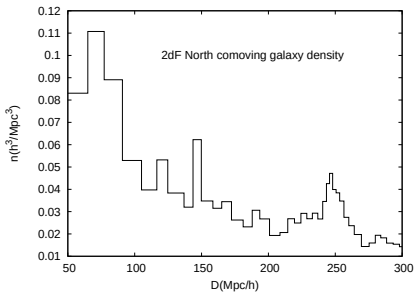

Fig. 4. Galaxy density for the full 2dFGRS NGP sample (in comoving coordinates) versus the distance D from the observer. As the nearby volume is small, single superclusters cause strong spikes in the histogram. The most prominent supercluster that creates the spike at $D \approx 250\,h^{-1}$Mpc, lies just outside of the largest data brick used in this work.

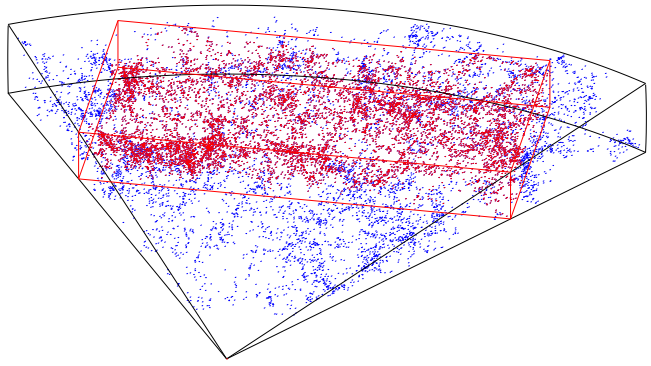

Fig. 5. The cuboidal samples analyzed in this paper drawn from the Northern Galactic Cap of the 2dFGRS.

K. As the data resolution is similar (column d in Table 1, we choose the same values for the dimensions of the cylinder for all the data sets, as follows : $r = 0.5$, $h = 6.0$. 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, 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 for the maximum distance between connected cylinders the value 1.5, and the pre-defined curvatures are $\tau_{\parallel} = \tau_{\perp} = 0.15$. This allows for a maximum of $\approx 30^{\circ}$ between the direction angles of connected cylinders, and limits the orthogonality condition to angles larger than $\approx 80^{\circ}$.

The data energy parameter is $u_{\max} = -25$. For the interaction energy, the parameter domain is chosen 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 $\gamma_h = 0$. An uniform prior on this domain was chosen for $p(\theta)$.

The joint law $p(\mathbf{y}, \theta)$ is simulated iteratively. An iteration consists of two steps. First, a parameter value is chosen according to $p(\theta)$. Then, conditionally on θ, a new configuration of cylinders is chosen with respect to $p(\mathbf{y}|\theta)$. The conditional law $p(\mathbf{y}|\theta)$ was simulated using a tailored to the model Metropolis-Hastings algorithm (van Lieshout and Stoica (2003a); Stoica et al. (2005)).

A very delicate point when building a simulated annealing procedure is the choice of an appropriate cooling schedule for the temperature. (Stoica et al. (2005)) proved that the use of a logarithmic cooling schedule guarantees the convergence of the simulated annealing algorithm when the parameters of the point process are fixed. Therefore, we have chosen the following scheme for the descent of the temperature :

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

with $T_0 = 1$.

We ran the simulated annealing algorithm for 250000 iterations. Samples were picked up every 250 iterations. The obtained cylinder configurations for the data sets $n150$, $n200$ and $n250$ are shown in Figure 6a, Figure 7a and Figure 8a, respectively.

The detected cylinders are situated in the regions where, after a visual examination, we have decided that data exhibit filamentary structure. Still, as simulated annealing requires infinitely many iterations till convergence, and also because of the fact that an infinity of solutions are proposed, we shall use coverage probabilities to "average" the shape of the filaments.

For each data set the corresponding domain K was divided into small cubic cells of size $1\,h^{-1}$Mpc. The coverage probability for each region was computed using (5). The obtained result was thresholded using three distinct values : 0.5,0.75 and 0.95. For each of these values we obtain a map of the cells that have been visited by our model with a frequency higher or equal than the given value. These visit maps for to the data sets $n150$, $n200$ and $n250$ are shown in Figure 6b-c, Figure 7b-c and Figure 8b-c respectively.

The filamentary network is clearly outlined. Here we consider a filament the geometric structure obtained recursively from a visited cell and its corresponding neighbouring cells. In this case, we have defined the neighbours of a cell, the cells that have coordinates incremented or decremented by one with respect the reference cell. Hence, a cell has 26 neighbouring cells.

Although many filaments are of simple form, there are many filaments which exhibit complex morphology, from simple branching to multibranch complexes. This is a new fact

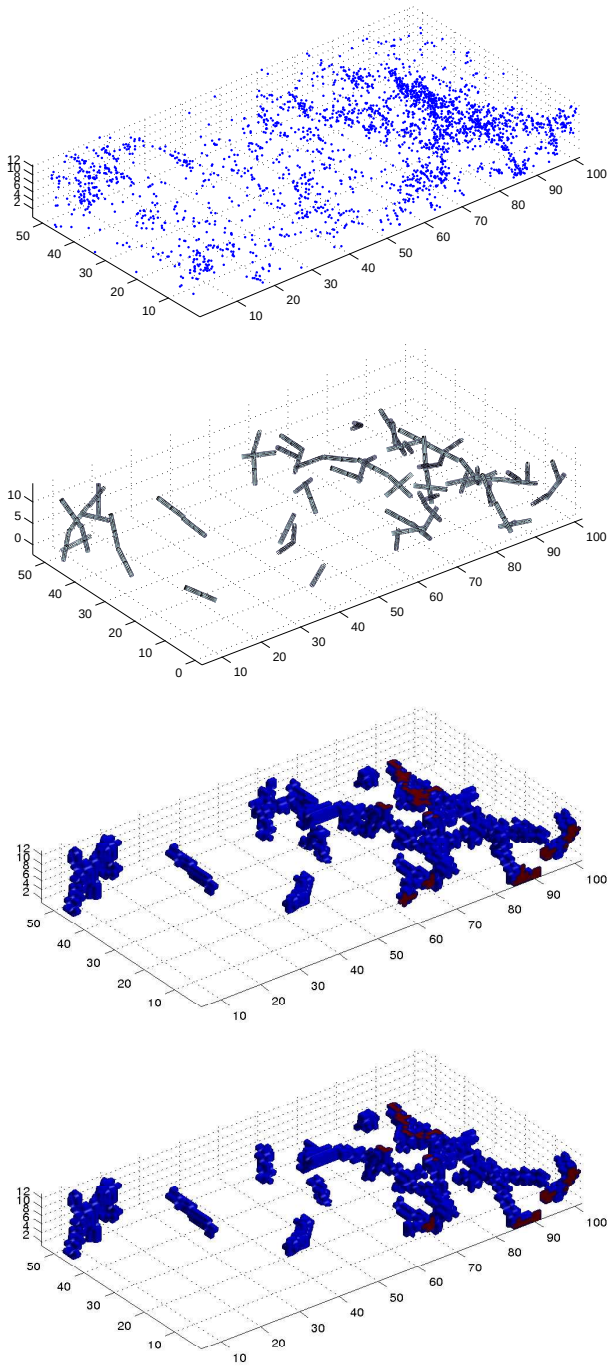

Fig. 6. Data set $n150$. a) The data. b) Cylinder configuration obtained using the simulated annealing algorithm. Cover probabilities thresholded at c) 50%, d) 95%.

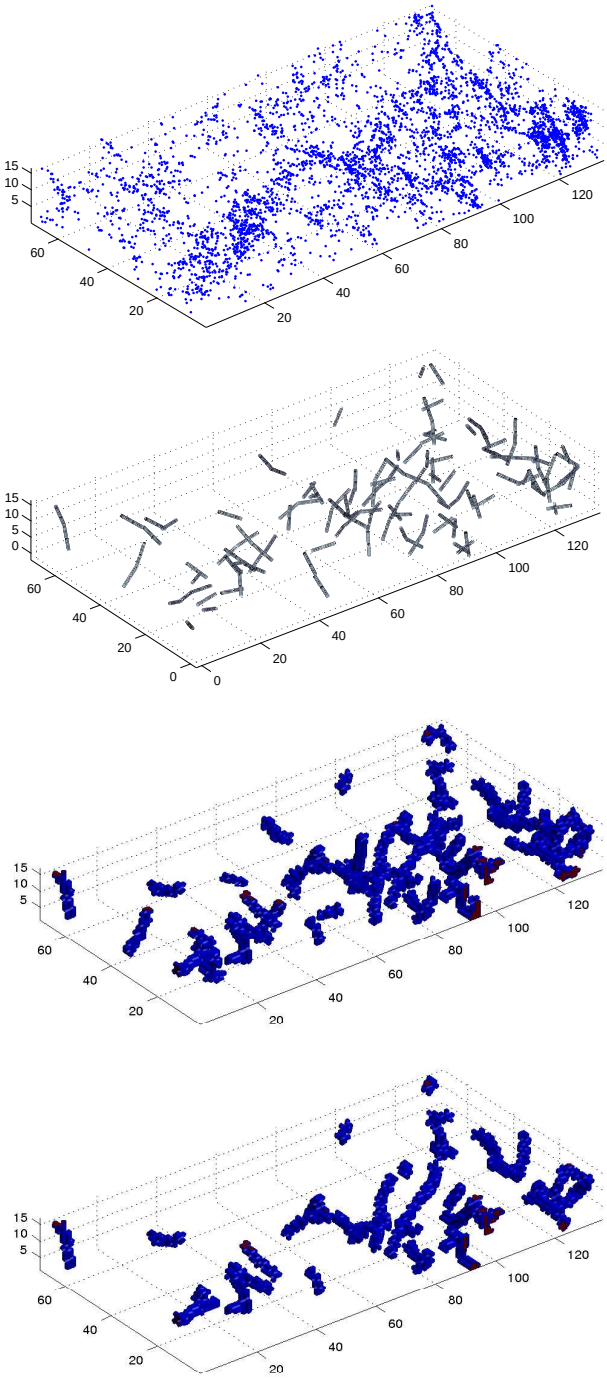

Fig. 7. Data set $n200$. a) The data, b) Cylinder configuration obtained using the simulated annealing algorithm. Cover probabilities thresholded at c) 50%, d) 95%.

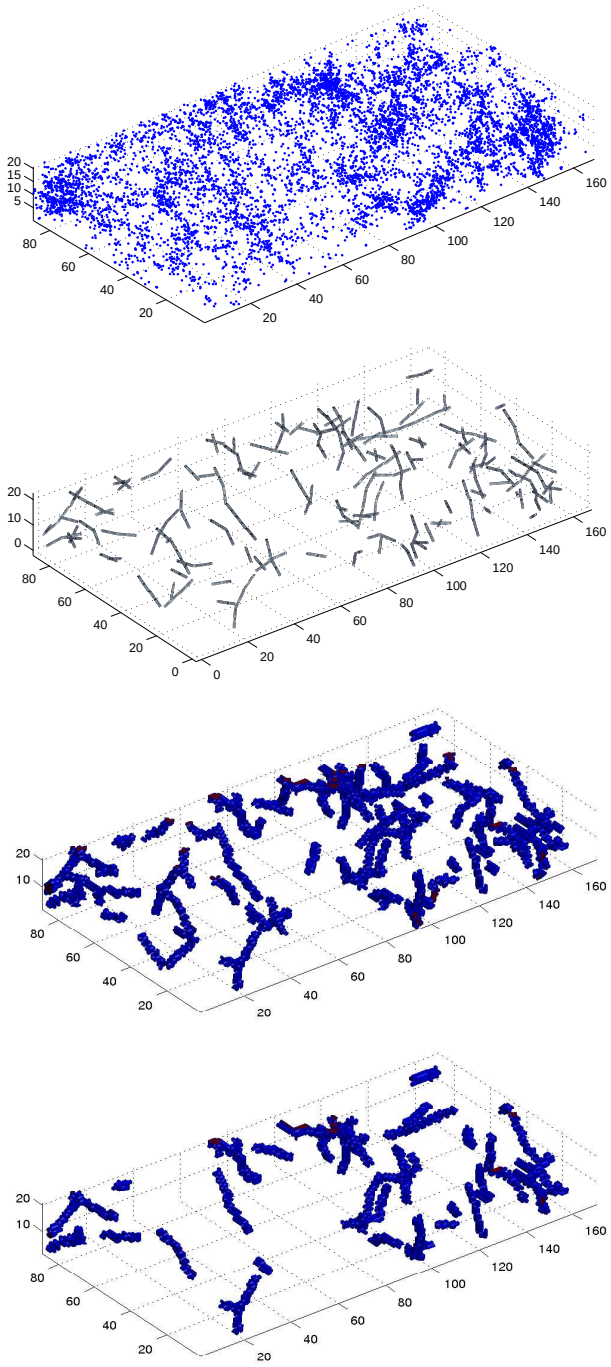

Fig. 8. Data set $n250$. a) The data, b) Cylinder configuration obtained using the simulated annealing algorithm. Cover probabilities thresholded at c) 50%, d) 95%.

that will probably force cosmologists to reconsider their usual notion of filaments as simple bridges between clusters of galaxies. We show two examples of such filaments below, one of a comparatively simple shape (Fig. 9), and another of very complex shape (Fig. 10).

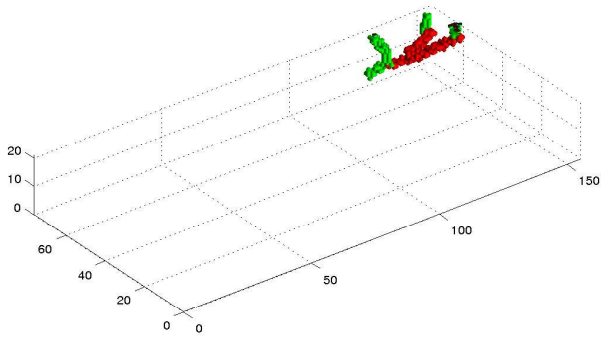

Fig. 9. A simple filament from the $n250$ sample. Lighter shading (green colour) shows the 50% cover probability threshold, darker shading (red colour) – the 95% threshold.

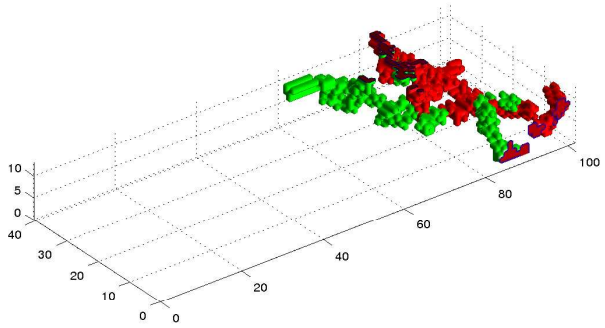

Fig. 10. A complex filament from the $n150$ sample. Lighter shading (green colour) shows the 50% cover probability threshold, darker shading (red colour) – the 95% threshold.

4.3. Statistics

The coverage probabilities are not a global test for our method, because these probabilities are computed only for small regions. The computation of the probability that a whole filament is covered by our model requires the knowledge a priori of the shape of the filament. (Stoica and Gay (2005)) used for a different problem - cluster detection in epidemiological data - a test, to check that the pattern is detected rather because of the data than of the model parameters.

Following this idea, the following experiments was carried out for each data set. First, the method was launched during 50000 iterations at fixed $T = 1.0$. 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 11.

Sufficient statistics	Data sets		
	n150	n200	n250
$\bar{n_2}$	4.13	5.83	9.88
$\bar{n_0}$	15.88	21.19	35.82
$\bar{n_1}$	21.35	35.58	46.49

Fig. 11. The mean of the sufficient statistics for the three data sets.

The second experiment consisted of re-distributing uniformly the points inside the domain K. So, the points follow a binomial distribution. For each data set this operation was done 100 times, hence obtaining 100 point fields. For each point field the method was launched in the conditions previously described. The mean of the sufficient statistics was then computed. The maximum values over all the 100 means for each data set are shwon in Table 12.

Sufficient statistics	Binomial data sets		
	n150	n200	n250
$\max \bar{n_2}$	0.015	0.05	0.015
$\max \bar{n_0}$	0.54	0.27	0.45
$\max \bar{n_1}$	0.39	0.24	0.33

Fig. 12. The maximum of mean of the sufficient statistics over binomial fields generated for each of the three data sets.

These results clearly indicate that the original data exhibit a filamentary structure. No filamentary structure is detected when the same model runs over binomial fields of points that has the same number of points as the original data. This test discriminates the two cases at a $p-$value less than 1%, and assures us that the filaments we find are due to data and not to our model.

5. Conclusion and perspectives

We have applied an object point process - Bisous model - to objectively find filaments in galaxy redshift surveys (three-dimensional galaxy maps). For that, we defined the model, fixed some of the interaction parameters and chose priors for the remaining parameters. The definition of the data term is very intuitive and rather a simple test. Much more elaborated methods testing the alignement of the points along the direction of the cylinder against the completely spatial randomness need investigation. The uniform law for the interaction parameters was prefered in order to give the same chances to a wide range of topologies of the filamentary network. Still, if concrete prior information about the topology of the filamentary network is available then this should be integrated in the model.

We have run simulated annealing sequences and select filaments on the basis of coverage probabilities for individual cells of sample volume. The coverage probabilities are to be seen

as a way of averaging the shape of the filamentary structure. They have the advantage to allow inference from statistics instead of a single realisation. Their main drawback is that the coverage probabilities are computed locally, for small regions. Still, the visualisation of the visit maps built on these probabilities brings new ideas and hypothesis about the topologies of the cosmic filaments. A global Monte Carlo statistical test was built to test the existence of the filaments in the data. To do this, we have calculated sufficient statistics for the data sets and made a comparison with the sufficient statistics obtained on binomial point fields having the same number of points as the data. Our test was indicating that the filaments we find are defined by the data, not by the chosen model.

The method used in this paper can be extended in different ways. One natural extension is to use different generating elements instead on cylinders, e.g., planar elements or clusters. (Stoica et al. (2005)). Although traces of planar structures are seen in superclusters of galaxies, these have been difficult to quantify, mainly because of their low density contrast. Another interesting application is to search for dynamically bound groups and clusters of galaxies that have a typical 'finger-of-God' signature in redshift space, extended along the line-of-sight. And there remains a question if the method could be extended to inhomogeneous point processes – this would allow us to use all the observational data, not only volume-limited subsamples.

6. 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 TO 0062465s03, and by the Estonian Science Foundation grant 6104.

References

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

Berthelsen, K. K. and J. Møller (2002). A primer on perfect simulation for spatial point processes. *Bulletin of the Brazilian Mathematical Society 33*, 351–367.

Bharadwaj, S., S. P. Bhavsar, and J. V. Sheth (2004). The size of the longest filaments in the universe. *The Astrophysical Journal 606*, 25–31.

Bharadwaj, S. and B. Pandey (2004, November). Using the Filaments in the Las Campanas Redshift Survey to Test the ΛCDM Model. *The Astrophysical Journal 615*, 1–6.

Colberg, J. M., K. S. Krughoff, and A. J. Connolly (2005, May). Intercluster filaments in a ΛCDM Universe. *Monthly Notices of the Royal Astronomical Society 359*, 272–282.

Colless, M., 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 (2001, December). The 2dF Galaxy Redshift Survey: spectra and redshifts. *Monthly Notices of the Royal Astronomical Society 328*, 1039–1063.

Croton, D. J., M. Colless, E. Gaztañaga, C. M. Baugh, P. Norberg, I. K. Baldry, J. Bland-Hawthorn, T. Bridges, R. Cannon, S. Cole, C. Collins, W. Couch, G. Dalton, R. de Propris, S. P. Driver, G. Efstathiou, R. S. Ellis, C. S. Frenk, 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 (2004, August). The 2dF Galaxy Redshift Survey: voids and hierarchical scaling models. *Monthly Notices of the Royal Astronomical Society 352*, 828–836.

Croton, D. J., E. Gaztañaga, C. M. Baugh, P. Norberg, M. Colless, I. K. Baldry, J. Bland-Hawthorn, T. Bridges, R. Cannon, S. Cole, C. Collins, W. Couch, G. Dalton, R. De Propris, S. P. Driver, G. Efstathiou, R. S. Ellis, C. S. Frenk, 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 (2004, August). The 2dF Galaxy Redshift Survey: higher-order galaxy correlation functions. *Monthly Notices of the Royal Astronomical Society 352*, 1232–1244.

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

Geyer, C. J. (1994). On the convergence of monte carlo maximum likelihood calculations. *Journal of the Royal Statistical Society, Series B 56*, 261–274.

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

Geyer, C. J. and J. Møller (1994). Simulation procedures and likelihood inference for spatial point processes. *Scan. J. Stat. 21*, 359–373.

Geyer, C. J. and E. A. Thompson (1992). Constrained monte carlo maximum likelihood for dependent data. *Journal of the Royal Statistical Society, Series B 54*, 657–699.

Green, P. (1995). Reversible jump MCMC computation and bayesian model determination. *Biometrika 82*, 711–732.

Hearn, D. and M. P. Baker (1994). *Computer graphics. Second edition*. Prentice-Hall, Englewood Cilffs, NJ.

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

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

van Lieshout, M. N. M. (1994). Stochastic annealing for nearest-neighbour point processes with application to object recognition. *Adv. Appl. Prob. 26*, 281–300.

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

van Lieshout, M. N. M. and R. S. Stoica (2003a). The Candy model revisited: properties and inference. *Stat. Neerlandica 57*, 1–30.

van Lieshout, M. N. M. and R. S. Stoica (2003b). Perfect simulation for marked point processes. Research report PNA-0306, CWI.

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

Møller, J., A. N. Pettitt, K. K. Berthelsen, and R. W. Reeves (2004). An efficient markov chain monte carlo method for distributions with intractable normalizing constants. Research report R-2004-02, Department of Mathematical Sciences, Aalborg University.

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

Pandey, B. and S. Bharadwaj (2005, March). A two-dimensional analysis of percolation and filamentarity in the Sloan Digital Sky Survey Data Release One. *Monthly Notices of the Royal Astronomical Society 357*, 1068–1076.

Pimbblet, K. A. (2005). Pulling Out Threads from the Cosmic Tapestry: Defining Filaments of Galaxies. *Publications of the Astronomical Society of Australia 22*, 136–143.

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

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

Preston, C. J. (1977). Spatial birth-and-death processes. *Bull. Int. Stat. Inst. 46*, 371–391.

Reiss, R. D. (1993). *A course on point processes*. Springer-Verlag, New York.

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

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

Stoica, R. S. and E. Gay (2005, July). Cluster detection in spatial data using marked point processes. Research report 11-2005, INRA, Avignon.

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

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