Detection of cosmic filaments using the Candy model

Radu S. Stoica*

Departament de Matemàtiques, Universitat Jaume I,

Campus Riu Sec, E-12071 Castelló, Spain

Vicent J. Martínez†

Observatori Astronòmic de la Universitat de València,

Apartat de correus 22085, 46075 València, Spain

Jorge Mateu‡

Departament de Matematiques, Universitat Jaume I,

Campus Riu Sec, E-12071 Castelló, Spain

Enn Saar§

Tartu Observatoorium, Tõravere, 61602 Estonia

(Dated: May 19, 2004)

Abstract

We propose to apply a marked 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 simulated catalogues, describe the procedure, and characterize the results. We find the distribution of the length of the filaments, and suggest how to use this approach to obtain other statistical characteristics of filamentary networks.

PACS numbers: 98.65.Dx, 02.70.Rr, 42.30.Sy

*Electronic address: stoica@guest.uji.es

†Electronic address: martinez@uv.es

‡Electronic address: mateu@mat.uji.es

§Electronic address: saar@aai.ee

I. 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 [1]. As an example, we present here two maps. The first map comes from a well-known recent galaxy survey, the Las Campanas Redshift Survey (LCRS, [23]). This survey measured the redshifts (recession velocities) of galaxies in six slices of width of 80° and of thickness of 1.5°; a typical number of galaxies in a slice is about 4000, and the depth of the survey is about $575\,h^{-1}$Mpc[31] (corresponding to a redshift $z = 0.2$ for the standard cosmological model). A map of one of the slices (the -42° slice) is shown in Fig. 1, in the upper panel.

The other map we show is from the most recent, ongoing survey, the Sloan Digital Sky Survey (SDSS, [29]). This survey will measure fainter galaxies than the LCRS, will reach deeper in space and will finally cover a full adjoint π sterradians of the sky. The data that have been released by now consist also of separate slices; we chose a contiguous slice of 2.5° thick and 60° wide (Data Release 1, the slice with the mean 'survey coordinate' $\eta \approx -23^\circ$). The depth of the SDSS main galaxy sample is about $840\,h^{-1}$Mpc (the redshift $z = 0.3$). The map of the selected slice, containing 10886 galaxies, is shown in Fig. 1, in the lower panel.

The dominant feature of these maps, 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 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.

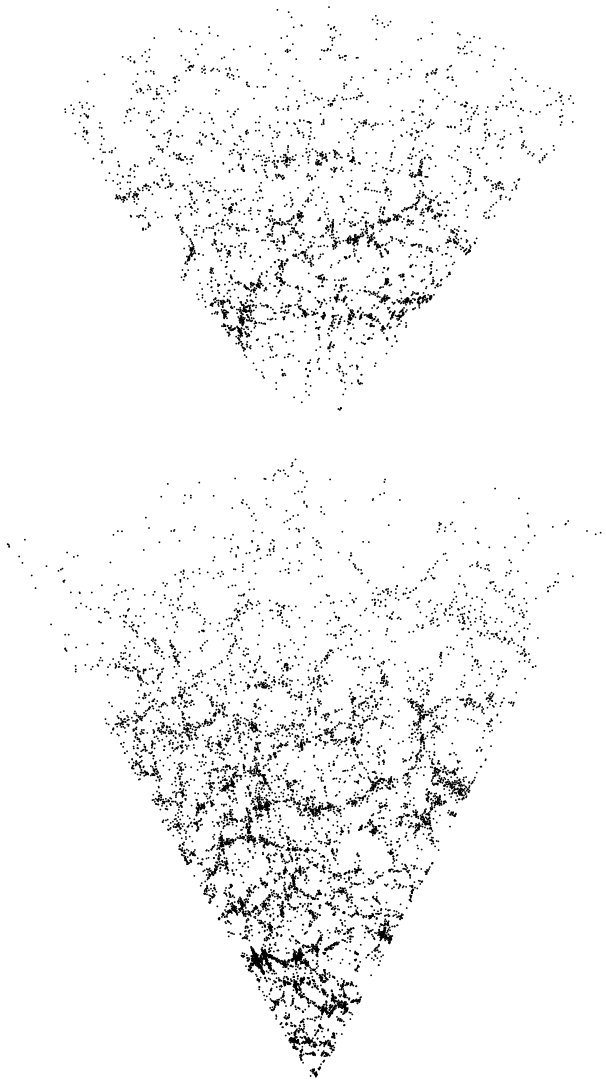

FIG. 1: Galaxy maps for two recent surveys, the LCRS, top panel, and the SDSS, bottom panel. The observers (we) are situated at the bottom of the figures. Both slices are thin (the thickness of the LCRS slice is 1.5° and that of the SDSS slice is 2.6°). The scale is the same in both panels; the depth of the SDSS slice in the right panel is $900\,h^{-1}$ Mpc. The filamentary network of galaxies is clearly seen; the disappearance of structure with depth (towards the top of the figure) is caused by luminosity selection.

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 function 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 [1]). The first application of the MST formalism to describe the filamentary networks of galaxy maps was that of [2]; many later studies have used it.

The minimal spanning tree is unique for a given point set, which is good, and it connects all the 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 (we shall show an example of a minimal spanning tree in sec. V). The filamentary network seen by 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 [19]. 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.

We propose 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 [14, 26, 27]. This method is based on the Candy model, a marked point process, where segments serve as marks. As this is the first time such a method is used for the galaxy distribution, we describe it in detail below. We test it also on 2-D simulated galaxy maps, justifying our data choice. The task differs considerably from road network detection, as the noise is larger, and we have no continuous roads, but sparsely populated ridges instead.

The present approach allows us to find the length distribution for the filaments; we give examples of this distribution for different data samples. In this paper, we choose the Candy process parameters by trial and error following a reversible jump process. As the method is automated, it can also be used to estimate those parameters by using maximum likelihood methods; these will serve then as new statistics for filament networks.

II. MARKED POINT PROCESSES

Let $(K, \mathcal{B}, \nu)$ be a measure space, where K is a compact subset of $\mathbb{R}^2$ of strictly positive Lebesgue measure $0 < \nu(K) < \infty$ and $\mathcal{B}$ the associated Borel $\sigma-$algebra of subsets of K. For $n \in \mathbb{N}$ let K_n be the set of all unordered configurations $\mathbf{k} = \{k_1, k_2, \ldots, k_n\}$ that consists of n not necessarily distinct points $k_i \in K$. Let us consider the configuration space $\Omega = \cup_{n=0}^{\infty} K_n$ equipped with the $\sigma-$algebra $\mathcal{F}$ generated by the mappings $\{k_1, k_2, \ldots, k_n\} \rightarrow \sum_{i=1}^{n} \mathbf{1}\{k_i \in B\}$ counting the number of points in Borel sets $B \in \mathcal{B}$. A point process on K is a measurable map from a probability space into $(\Omega, \mathcal{F})$.

The reference measure is given by the unit rate Poisson process that distributes the points in K according to a Poisson process with intensity ν.

Different characteristics or marks may be attached to the points. Under these circumstances, we consider a point process on $K \times M$ as the random sequence $\mathbf{x} = \{(k_1, m_1), \ldots, (k_n, m_n)\}$ where $n \in \mathbb{N}_0$, $k_i \in K$ and $m_i \in M$ for all $i = 1, \ldots, n$. The characteristics space M is equipped with its corresponding Borel $\sigma-$algebra and the probability measure ν_M. A marked point process X with locations in K and marks in M is a point process on $K \times M$ such that the distribution of locations only is a point process on K.

In this case, the reference measure is the unit rate Poisson process on $K \times M$, with the locations distributed according to a Poisson process with intensity ν and i.i.d marks according to ν_M. When the marks represent parameters of an object, such a process is sometimes called an object point process.

The reference measure exhibits no interaction between points or objects. Indeed, we can construct a much more complicated marked point process by specifying a probability density with respect to the reference measure :

$$p(\mathbf{x}) = \alpha \exp[-U(\mathbf{x})], \tag{1}$$

with α the normalizing constant and $U(\mathbf{x})$ the interaction energy of the system. The energy function is written as the sum

$$U(\mathbf{x}) = \sum_{j=1}^{q} \sum_{\{x_{i1}, \ldots, x_{ij}\} \in \mathbf{x}} \omega^{(j)}(x_{i1}, \ldots, x_{ij})$$

where $\omega^{(j)} : (K \times M)^j \rightarrow \mathbb{R}$ for $j = 1, \ldots, q$ are the interaction potentials. The marked point processes with the probability density of the form given by (1) are known in physics

under the name of Gibbs point processes. If there exists a positive real $C > 0$ such that $U(\mathbf{x}) - U(\mathbf{x} \cup \{(k,m)\}) \leq \log C$ for all $(k,m) \in K \times M$ the process is said to be locally stable.

This relation implies the Ruelle's stability condition [22], which ensures the integrability of the given probability density function. Furthermore, local stability is essential in establishing convergence proofs for the Monte Carlo dynamics simulating such a model [9].

For our problem, $\mathbf{y}$, the data to be analysed, consists of points (galaxies) spread in a finite window K. We want to extract the filamentary structure of this data. It is natural to consider the filaments $\mathbf{x}$ we want to detect as a set of random segments being the realization of a marked point process.

The probability density of such a marked point process is given by

$$p_{\mathbf{y}}(\mathbf{x}) \propto \exp[-(U_{\mathbf{y}}(\mathbf{x}) + U_r(\mathbf{x}))] \tag{2}$$

with the terms $U_{\mathbf{y}}(\mathbf{x})$ and $U_r(\mathbf{x})$ being the data energy and the interaction energy, respectively. The first term is related to the location of the filaments among galaxies, whereas the second is related to the geometrical properties of the filaments, playing the role of a regularization term.

The configuration of segments composing the filamentary network is estimated by the minimum of the total energy of the sysytem

$$\hat{\mathbf{x}} = \arg\min_{\mathbf{x}} \{U_{\mathbf{y}}(\mathbf{x}) + U_r(\mathbf{x})\}. \tag{3}$$

In the following we will present the two components of the energy function, considerations about the simulation of such models using the MCMC dynamics will be given and a simulated annealing algorithm will be presented. Finally, we will apply the model to describe two-dimensional filamentary networks of galaxies.

III. A PROBABILISTIC MODEL FOR THE FILAMENTARY STRUCTURE OF GALAXY MAPS

A. The interaction energy : Candy model

The filaments we want to extract are composed of non-overlapping connected segments. Locally, the curvature of one filament does not vary too much. In the data we dispose we

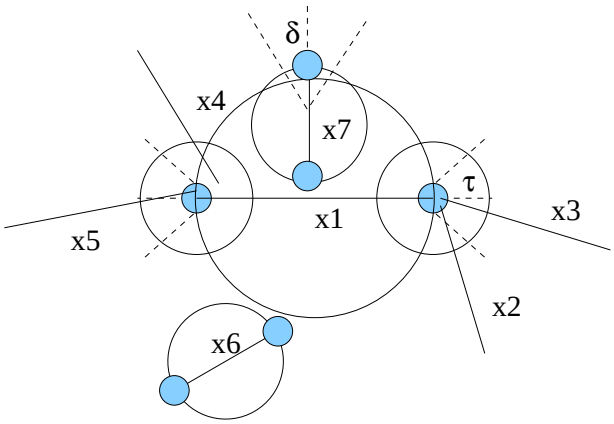

FIG. 2: Connection and alignment interaction between segments.

can notice just a few short filaments, which can be represented by isolated segments.

Under these considerations a natural choice for the interaction energy becomes the Candy model, a marked point process simulating random networks of segments. Here, a segment is seen as a random object $\zeta = (k, (\theta, w, l))$ that is characterized by its center location $k \in K$ and its geometrical parameters $(\theta, w, l) \in [0, \pi] \times [w_{\min}, w_{\max}] \times [l_{\min}, l_{\max}] = M$, representing its orientation, width and length respectively. The Candy model exhibits three types of interactions between segments: connectivity, alignment and rejection.

Historically speaking, the model was introduced for the first time as a prior distribution for thin network extraction in remotely sensed images [25–27]. Properties of the model such as local stability and Markovianity, convergence proofs of an adapted Metropolis-Hastings dynamics for simulating the model, as well as parameter estimation, were further investigated in [17]. Different versions of the model were analysed and compared for the special case of road network detection [14].

A segment has a connection region formed by the union of the two circles centered at its extremities and of a radius r_c. Two segments $\eta = (k_\eta, (\theta_\eta, w_\eta, l_\eta))$ and $\zeta = (k_\zeta, (\theta_\zeta, w_\zeta, l_\zeta))$ are connected $\eta \sim_c \zeta$ if only one extremity of a segment is in the connection region of the other segment and if $\|\theta_\eta - \theta_\zeta\| \leq \tau$. With respect to this definition, a segment is doubly connected if both of its extremities are connected, singly connected if only one of its extremities is connected and free if none of its extremities is connected. The Candy model favors doubly connected segments whereas free and singly connected segments are penalized.

In Fig. 2 we show an example of a configuration of segments. The free segments are x_2, x_4, x_6 and x_7, this because the segment x_2 does not fullfil the orientation requirements

for the connection and the others do not respect the connection condition. The segments x_3 and x_5 are singly connected, whereas the segment x_1 is doubly connected.

Similarly, the attraction region of a segment η is the union of both circles centered at each extremity with a radius $r_o = l_\eta/4$. Two segments η and ζ exhibit alignment interaction $\eta \sim_o \zeta$ if $d(k_\eta, k_\zeta) > \frac{1}{2}\max\{l_\eta, l_\zeta\}$, if only one extremity of a segment is in the attraction region of the other segment, and if $\min\{\|\theta_\eta - \theta_\zeta\|, \pi - \|\theta_\eta - \theta_\zeta\|\} > \tau$, with τ a threshold value. The Candy model penalizes the segments having alignment interaction.

In the configuration shown in Fig. 2 $x_3 \not\sim_o x_1$ and $x_5 \not\sim_o x_1$, while the segments $x_2 \sim_o x_1$ and $x_4 \sim_o x_1$ because these pairs of segments exhibit high differences between their orientations.

Connectivity is a stronger interaction than alignment. Still, as we look for the filaments fitting the data in a random way, this last interaction gives us the possibility not to eliminate from the current configuration the segments with low data energy, which are almost connected.

Every segment η is provided with a rejection region given by a circle centered in k_η and of a radius $r_r = l_\eta/2$. Two segments η and ζ exhibit rejection interaction if $d(k_\eta, k_\zeta) < \frac{l_\eta + l_\zeta}{2}$ and if $\|\|\theta_\eta - \theta_\zeta\| - \pi/2\| > \delta$, where δ is a threshold value. The Candy model forbids configurations containing rejecting segments, avoiding configurations containing overlapping segments.

If $d(k_\eta, k_\zeta) \leq \frac{1}{2}\max\{l_\eta, l_\zeta\}$ and if $\|\|\theta_\eta - \theta_\zeta\| - \pi/2\| \leq \delta$, then the segments may cross or form a "T" junction. The configurations with crossing segments $\eta \sim_x \zeta$ are forbidden by the Candy model, whereas the "T" junctions are allowed.

Clearly, in Fig. 2 the segments x_1 and x_6 do not reject each other since they are far enough, while the segments x_1 and x_7 do not cross, forming a "T" junction.

For any configuration of segments $\mathbf{x} = \{x_1, \ldots, x_n\}$ with $i = 1, \ldots, n$, we are able now to write for the probability density of the Candy model

$$\begin{aligned} p_r(\mathbf{x}) \;\propto\; & \left\{\prod_{i=1}^{n(\mathbf{x})} \exp\left[\tfrac{l_i - l_{\max}}{l_{\max}} + \tfrac{w_i - w_{\max}}{w_{\max}}\right]\right\} \times \\ & \gamma_d^{n_d(\mathbf{x})} \gamma_f^{n_f(\mathbf{x})} \gamma_s^{n_s(\mathbf{x})} \gamma_o^{n_o(\mathbf{x})} \times \\ & \textstyle\prod_{i<j} \mathbf{1}\{x_i \not\sim_r x_j\} \mathbf{1}\{x_i \not\sim_x x_j\} \end{aligned} \tag{4}$$

where $\gamma_d, \gamma_f, \gamma_s > 0$ and $\gamma_o \in (0, 1)$ are the model parameters, $n_d(\mathbf{x}), n_f(\mathbf{x}), n_s(\mathbf{x})$ are the numbers of doubly, free and singly connected segments, and $n_o(\mathbf{x})$ is the number of pairs of segments which are not well aligned. In order to avoid too much small segments in the

configuration, the model favors segments covering a big area. Clearly the interaction energy is obtained taking $U_r(\mathbf{x}) = -\log p_r(\mathbf{x})$.

With respect to the classical definition of the Candy model in [17], the model described by (4) contains differences in the definition of interactions between segments. We kept the same name for our model, as we believe that the modifications required to apply it to cosmological data do not change the basic premises of the classical Candy model. Concerning connectivity, the present modifications were introduced in order to eliminate some "undesired" configurations, as a segment being connected with itself or a segment being connected at one extremity with both extremities of another segment. Furthermore, the new modifications allow us to build more appropriate tailored moves for the Metropolis-Hastings dynamics simulating the model. The rejection region was extended, as the filaments we observe may be rather wide, hence we want to avoid overlapping of segments when the data is good enough. This is also the reason why the width penalty was introduced. Nevertheless, it is easy to prove that under these modifications, together with the crossing interaction the Candy model is still locally stable and (Ripley-Kelly) Markov [21].

B. The data energy

The data energy checks whether a segment belongs to the network or not [25–27]. A segment x is considered a part of the filament network, if its geometrical shape $\tilde{x}$ covers as many galaxies as possible. Still, we want to avoid the case where segments are found in a cloud of points rather than in a filament. To do this, we consider the shadow segments x_r and and x_l – the segments situated to the right and to the left of the segment x, as in Fig. 3. The above-mentioned case is avoided if the number of galaxies covered by $\tilde{x_r}$ and $\tilde{x_l}$ is small. Therefore, let us define the quantity $v_{\mathbf{y}}(x)$ given by

$$v_{\mathbf{y}}(x) = 2n(\mathbf{y} \cap \tilde{x}) - n(\mathbf{y} \cap \tilde{x_r}) - n(\mathbf{y} \cap \tilde{x_l}),$$

where $n(\mathbf{y} \cap \tilde{x})$ is the number of galaxies covered by the geometrical shape of the segment x. Now, if the following three conditions: $v_{\mathbf{y}}(x) \geq 3$, $n(\mathbf{y} \cap \tilde{x}) > n(\mathbf{y} \cap \tilde{x_r})$, and $n(\mathbf{y} \cap \tilde{x}) > n(\mathbf{y} \cap \tilde{x_l})$ are simultaneously fulfilled, the data energy contribution of a segment is $V_{\mathbf{y}}(\{x\}) = -v_{\mathbf{y}}(x)$. If only one of the three conditions is not verified then $V_{\mathbf{y}}(\{x\}) = V_{\max}$, with $V_{\max} > 0$ a positive fixed value.

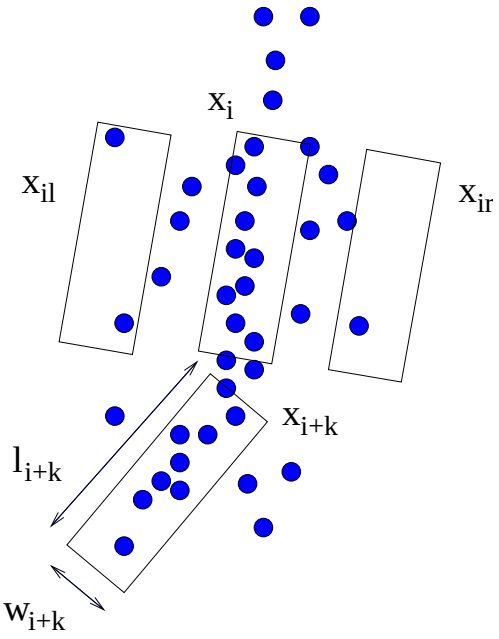

FIG. 3: Locating segments in a pattern of points.

The total data energy is defined as the sum of the data energy contributions of every segment in the configuration

$$U_{\mathbf{y}}(\mathbf{x}) = \sum_{x \in \mathbf{x}} V_{\mathbf{y}}(\{x\}) \tag{5}$$

C. Simulation dynamics and optimization

The equations (4) and (5) provide us with all the ingredients needed to construct the Gibbs point process given by (2). The estimate of the network (3) is obtained by means of a simulated annealing algorithm.

This algorithm iteratively samples the law $[p_{\mathbf{y}}(\mathbf{x})]^{\frac{1}{T}}$ while slowly decreasing the temperature T. At high positive values of temperature the space state is explored. When the temperature goes down to zero, $T \to 0$, the configurations of minimal energy are reached. A polynomial decreasing scheme $T_{k+1} = cT_k$ with $c \in [0.99, 1.00]$ may be used for cooling.

To sample from a probability density of a point process several Monte Carlo methods are available, such as the spatial birth-and-death processes, the Metropolis-Hastings and reversible jumps dynamics, or the much more recent exact simulation techniques as coupling from the past or clan of ancestors [8–10, 13, 16, 18, 20]. The Candy point process exhibits rather complicate interactions, hence the use of the spatial birth-and-death process or the

cited exact simulation techniques are useful in practice only for a limited range of the model parameters. Therefore, for our present model we opted for a sampling algorithm based on the Metropolis-Hastings dynamics. Details concerning the implementation of samplers for the Candy model based on Metropolis-Hastings or reversible jumps processes can be found in [17, 25–27].

IV. DATA

The Candy process and its applications have been developed for 2-D maps. So the natural way to introduce them in cosmology is to consider 2-D cases, also. It will allow us to compare the results, and will make it easier to understand the problems arising. Our final goal is to apply the Candy process to describe 3-D networks of filaments, as the large-scale structure maps fill the space. The 3-D network consists of complete filaments, as do the 2-D road geographical road maps, so the filaments in the test data should also be complete.

The observational galaxy maps showing filaments (see Fig. 1) have mainly the geometry of a thin slice, as those shown in the figure. Although such data have been used to study the large-scale filamentary structure, the slices do not provide proper data for that. The thickness of these slices is much smaller than the typical size of a filament, and although the maps give a visual impression of filaments, the filaments we see are pieces of real filaments, obtained by planar cuts through the real 3-D structure.

Another possibility is to use thicker slices, which can be selected, e.g., from the only large-scale contiguous data for the moment, the 2dF survey [6]. But this choice carries its own difficulties – thick slices give us the 2-D projection of the 3-D network, smearing essential details.

Simulations of the formation and evolution of large-scale structure can also provide us with galaxy maps. As a demonstration that we understand the basic features of the process, these maps show filamentary structure. How and why an initial Gaussian random density field develops filaments under self-gravitation, is an interesting matter, well explained by [5].

The usual simulations give us 3-D worlds, but it easy to also simulate the evolution of structure in a 2-D world. This has been done before, to obtain better numerical resolution (see, e.g. [3]); we used 2-D simulations to get complete cosmological networks of model galaxies.

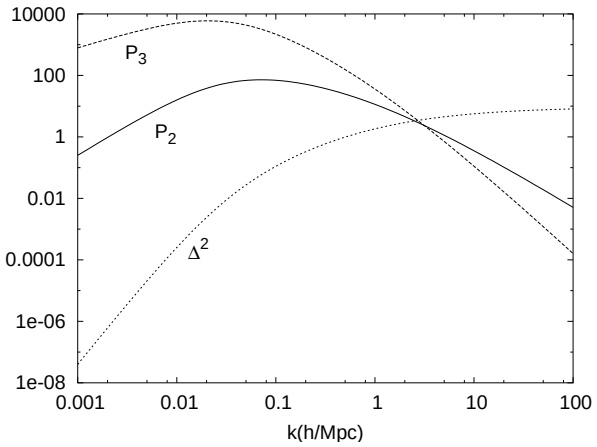

FIG. 4: The spectral density used for the 2-D simulation (P_2), the corresponding spectral density for the 3-D case (P_3), and the spectral energy per unit logarithmic wavenumber interval Δ^2, versus the wavenumber k.

The present-day large-scale structure is determined, first, by the expansion history of the cosmological model, and, secondly, by the initial density and velocity fields at the start of the simulation. We chose the standard 'concordance' cosmological model [28] to describe the expansion. As the initial fields are assumed to be Gaussian random fields, they are described by their power spectra (the spectral density of the density perturbations $P(k)$, where k is the module of the wave-vector; see, e.g. [1]). We chose a simple expression for the spectral density that describes reasonably well the CDM (Cold Dark Matter) model [12] and modified it to get the same spectral energy contribution to the variance per unit logarithmic wavenumber interval, $\Delta^2(k)$, in our 2-D world, as in the real 3-D world. In a 3-D world this quantity is defined as

$$\Delta_3^2(k) = \frac{1}{2\pi^2} P_3(k) k^3,$$

and in a 2-D world as

$$\Delta_2^2(k) = \frac{1}{2\pi} P_2(k) k^2;$$

the equality of the above quantities gives

$$P_2(k) = \frac{k}{\pi} P_3(k)$$

(the lower indices show the dimensionality of the space). This is the spectral density we used, with $P_3(k)$ taken from [12]. Both the spectral densities and the spectral energy used

are shown in Fig. 4. As usual, the wavenumber is given in units of h/Mpc, the spectral densities are in units of Mpc^3/h^3 ($P_3(k)$), Mpc^2/h^2 ($P_2(k)$), and $\Delta^2(k)$ is dimensionless.

We selected the scales and spectrum amplitudes to get a picture similar to that we see in 3-D models (the size of the patch we modelled was $128h^{-1}$Mpc, and we used a 256^2 grid with the same number of cold dark matter particles). These numbers are not really important, as this is a mock model, anyway. Then we ran a 2-D dynamical N-body simulation, developing the initial perturbations into large-scale structures – the present-day density and velocity fields.

TABLE I: Parameters of the data sets: a is the cosmological expansion factor, n is the number of galaxies, α is the void density threshold and β is the biasing amplitude.

Case	a	n	α	β
A	1.0	4127	0.5	0.20
B	0.6	4249	0.5	0.18
C	0.2	8879	1.0	0.49

These density fields describe the dark matter content of the universe. Populating model universes with galaxies is a complex problem, but for our purposes simple recipes are sufficient. We used two well-known properties of the large-scale galaxy distribution. First, galaxies avoid large low-density regions, known as voids; we modeled this by selecting a density threshold α (all our densities are given in the units of the mean density). In regions with density lower than this threshold no galaxies were placed. Secondly, galaxy density is biased in respect to the dark matter density. We found that the model galaxy distribution resembled best the observational maps for a nonlinear biasing law:

$$\rho_{\mathrm{gal}} = \beta\sqrt{\rho_{\mathrm{CDM}}}, \qquad \rho_{\mathrm{CDM}} \geq \alpha. \tag{6}$$

We chose the amplitudes α and β to produce approximately the same number of galaxies as observed in cosmological slices of similar size.

Finally, we generated a realization of a Cox point process, using the galaxy density given by (6) as the driving probability. In order to see how well the Candy model works in different situations, we chose three different time moments from the simulation, with a different filamentary structure. As the earliest of them (our case C) has a very rich set of

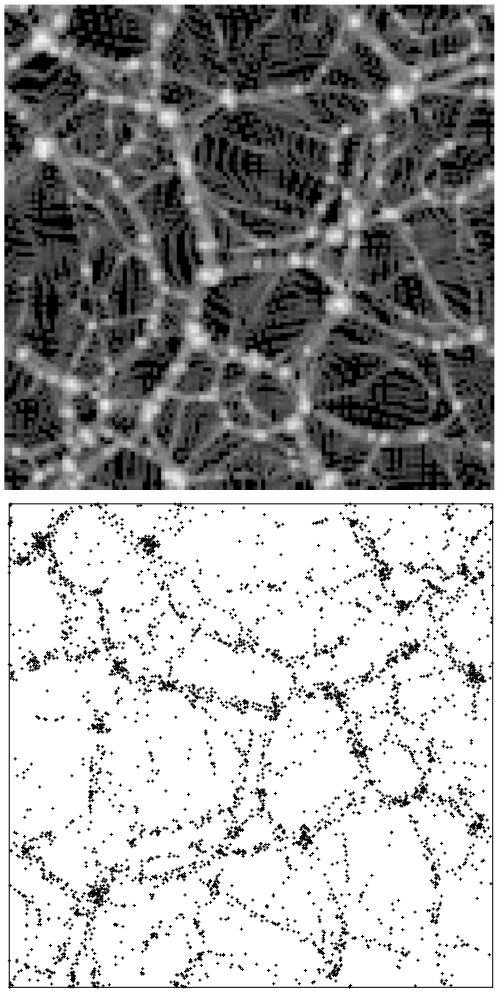

FIG. 5: Dark matter density (upper panel, logarithmic scale) and galaxy distribution (lower panel) for the data set B.

filaments, we generated about twice as many galaxies for that data set as for other sets. As usual in cosmology, we characterize the time moments by the value of the expansion factor a. This factor equals unity at the present epoch and the earlier the epoch, the smaller is the expansion factor (our universe expands). The parameters for our three data sets are given in Table I, and the dark matter density and galaxy distributions are compared for the set B in Fig. 5.

All the data sets are shown in Fig. 6.

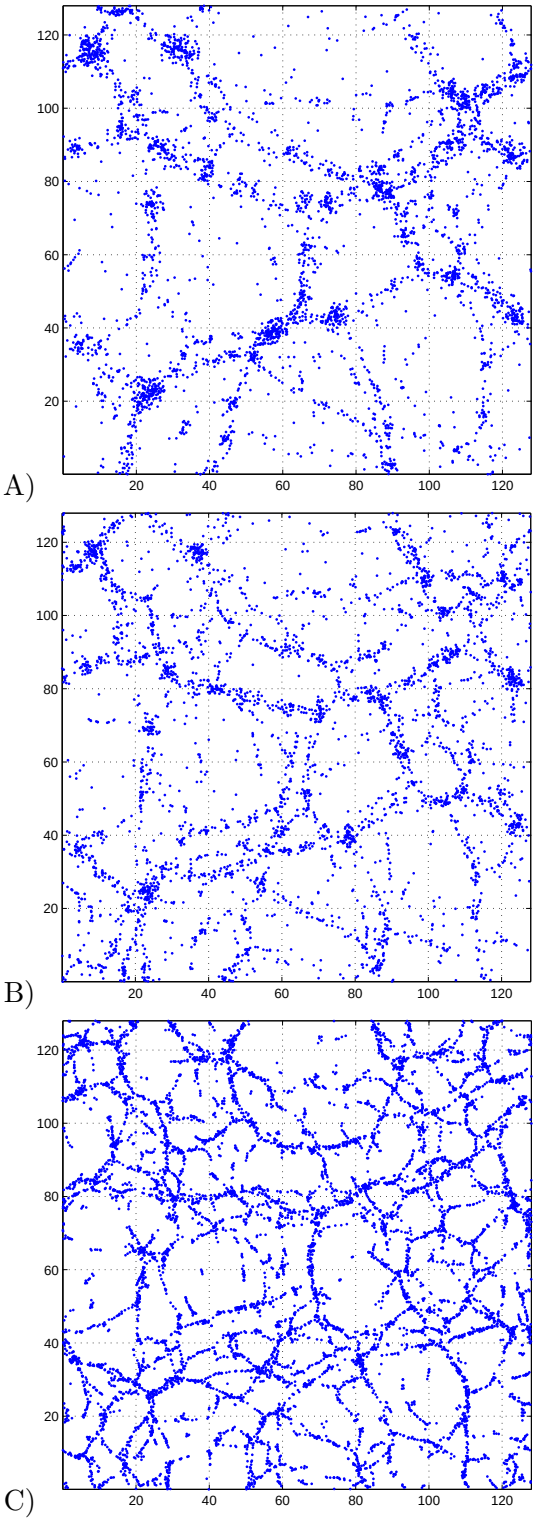

FIG. 6: Three sets of data

A. Experimental results

A simulated annealing algorithm was implemented based on Metropolis-Hastings dynamics. The parameter for the cooling scheme was taken as $c = 0.9995$ and the initial temperature was set to 10. The algorithm was run for 10^7 iterations, whereas the temperature was lowered every 10^3 steps.

The Candy model has a large number of parameters, and these should be chosen rather carefully in order to get a good representation of the filaments in data. The segment parameters (segment lengths and widths) have to be chosen to let the model filaments follow those in data. Thus, for the first two data sets, the segment parameters were $l_{\min} = 3, l_{\max} = 5, w_{\min} = 1, w_{\max} = 2$; for the third data set, smaller segments were considered: $l_{\min} = 2, l_{\max} = 3, w_{\min} = 0.95, w_{\max} = 1.05$ (all distances are given in h^{-1} Mpc). The interaction regions were defined by choosing the radius of the connecting region $r_c = 0.5$ and the rejection parameter that forbids segments to cross, $\delta = 0.1$ radians. The orientation parameter, which limits the local curvature of filaments, was chosen as $\tau = 0.5$ radians for the first two data sets and as $\tau = 0.75$ radians for the data set C.

We experimented with a large number of interaction parameters. Here we show the results for the three sets, which give almost equally good results. The interaction parameters for these sets are given in Table II. The optimization method was run for each data set. High potentials were given to undesired configurations as single and free segments, badly aligned pairs of segments with respect to the parameter τ, and badly placed segments with respect to the data term.

TABLE II: The interaction parameters.

Parameters	Sets		
	1	2	3
$-\log\gamma_d$	-10	-10	-10
$-\log\gamma_f$	8	7	7
$-\log\gamma_s$	2	2	1
$-\log\gamma_o$	3	3	3
$V_{\max}$	25	25	25

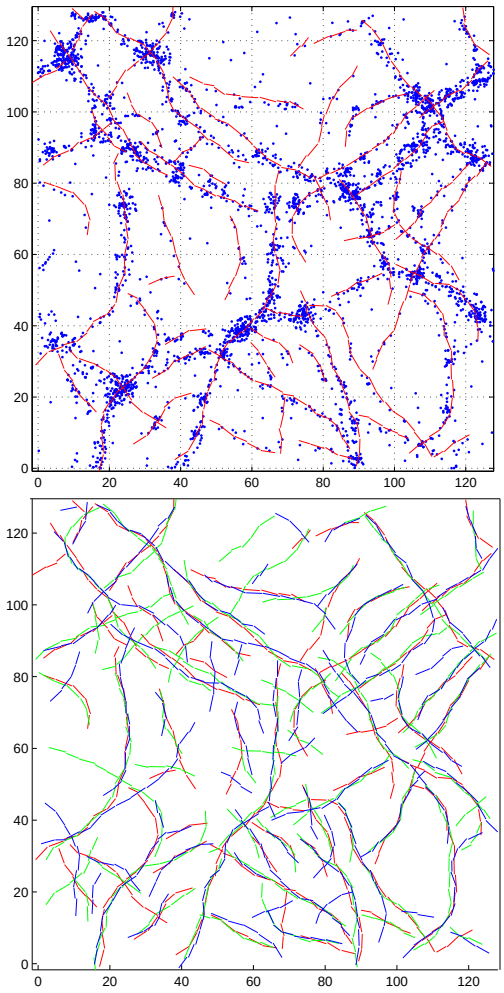

FIG. 7: Results obtained for the data set A: upper panel — the "best network" extraction superposed on the data, lower panel — the three networks superposed. The network for the first parameter set is shown by black lines, for the set 2 — by grey lines, and for the set 3 — by light grey lines.

In the Figs. 7,8 and 9 we compare for each set the data and the best filament network, and compare all three networks. We note that we do not use the periodicity of the data — although numerical simulations are mostly periodic, the real galaxy distribution is not. Thus it has no sense to complicate the numerical procedure.

The best set of parameters for the data set A was set 1. Examining Fig. 7 we see that the procedure works well. All obvious filaments, which one would draw by eye, are found, and the placement of "secondary" filaments in more sparsely populated regions is also good. Note also that galaxy concentrations ("clusters") do not destroy the filamentary pattern;

filaments usually branch in these regions.

The difference between the sets is slight, all parameter sets represent the network fairly well. All strong filaments coincide, the difference is in the small and weak filaments, built on a few points only. This is well seen in the lower panel, where all three networks are superposed. The parameter set 2, e.g., generates spurious filaments in a sparsely populated upper top of the data region, and the set 3 produces several very short isolated filaments. On the other hand, it also provides a perfect branching point at $x = 90$, $y = 30$, that sets 1 and 2 do not find.

Figure 8 illustrates the filament networks found for the data set B. This data set has a richer and more uniform selection of filaments than the set A. As these sets have approximately the same number of galaxies, individual filaments in the set B are more sparse and harder to identify. Nevertheless, the method works well, especially for the parameter set 1 – the best set; this network is shown in the upper panel of Fig. 8. There are only a few questionable short filaments, e.g., around $x = 70$, $y = 120$ and $x = 10$, $y = 50$. The parameter set 2 generates considerably more short isolated filaments, which do not represent the data well, and the filaments for the parameter set 3 tend to deviate in wrong directions.

The data set C has the richest set of filaments. Those are shorter and not as pronounced as the filaments in the two first data sets — this is the way the large scale structure develops in the universe. The early structure that the set C describes evolves by concentrating into larger and larger clusters and filaments; small-scale structure gets weaker and disappears gradually. In order to apply the Candy model, we had to generate about twice more galaxies for this set than for the other two. As shown in Fig. 9, our procedure delineates the filamentary network satisfactorily here, too, although, probably, the segments should have been smaller yet. As seen in the upper panel for the best parameter set (set 1 in this case, too), segments sometime jump from an obvious filament to another (e.g., at $x = 47$, $y = 20$); there is also a tendency to form short filaments for a collection of a few points, as at $x = 7$, $y = 60$ and $x = 75$, $y = 107$.

The parameter set 2 is in this case about as good as the set 1; it tends to miss a few obvious filaments, however (e.g., at $x = 124$, $y = 50$), and has difficulties in resolving interaction regions (knots in the network), see the region at $x = 70$, $y = 110$. This region has been equally difficult to model for all three parameter sets. And, finally, the parameter set 3 gives the worst filament placement between the three.

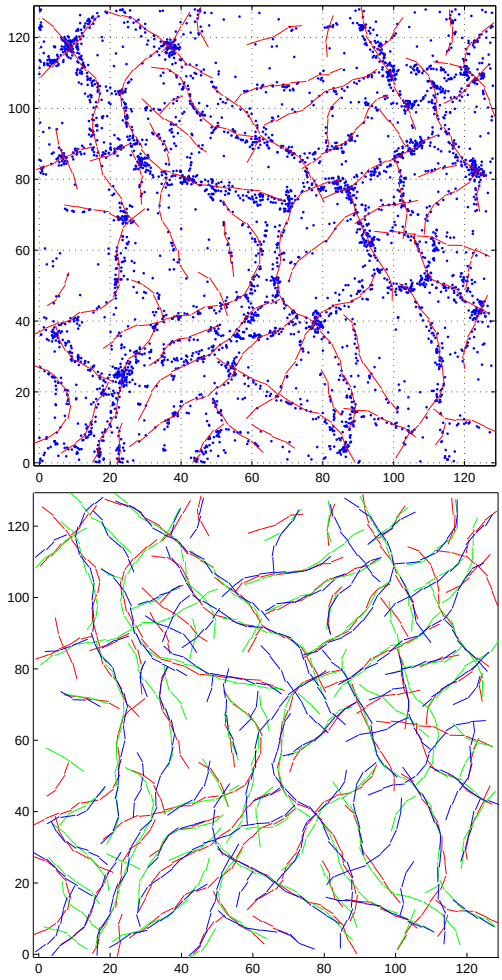

FIG. 8: Results obtained for the data set B: upper panel — the "best network" extraction superposed on the data, lower panel — the three networks superposed. The network for the first parameter set is shown by black lines, for the set 2 — by grey lines, and for the set 3 — by light grey lines.

B. Length distribution

As the Candy model is able to reconstruct the filamentary network, given a point process (galaxy map), the collection of its parameters can be considered as a description of the network. When determined from the data by a likelihood procedure, they can serve as statistics of the network. But there are simple statistics we can already study; the simplest one is the probability distribution of the lengths of individual filaments (sets of connected segments). A similar problem, that of the length of the largest filament, has been addressed recently, using a pixel-based method to define filaments and finding the pixel size, where the

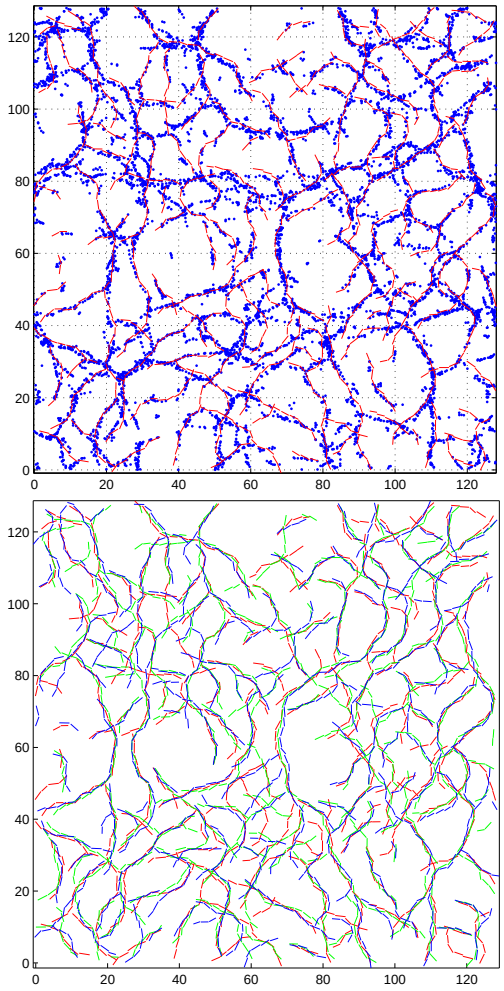

FIG. 9: Results obtained for the data set C: upper panel — the "best network" extraction superposed on the data, lower panel — the three networks superposed. The network for the first parameter set is shown by black lines, for the set 2 — by grey lines, and for the set 3 — by light grey lines.

filaments are yet statistically significant [4]. As filaments delineate voids, the distribution of the length of filaments is also connected with the distribution of void sizes. This subject has a long history; see, e.g. [1] and an interesting recent theoretical paper [24].

Comparison of the length histograms in Fig. 10 reveals a series of peaks, several distinct characteristic lengths in the filamentary network[32]. These peaks are especially prominent for the data sets A and C, and smeared out for the set B. Also, the locations of the peaks do not depend much on the specific parameter set, the peaks more or less coincide. And although the sample sizes are a bit too small for the first two data sets to draw firm

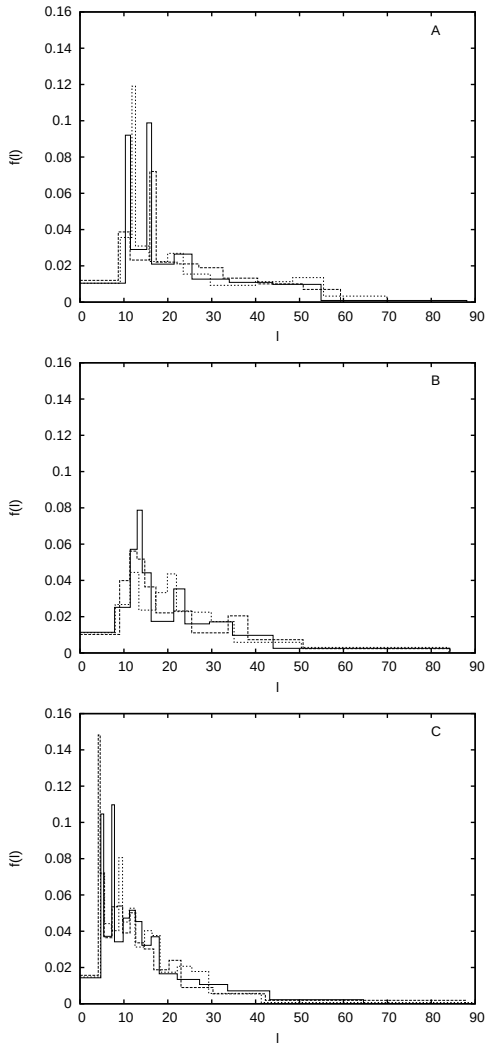

FIG. 10: Filament length distribution histograms for the three data sets (marked in the panels). Solid lines indicates the parameter set 1, dashed lines – set 2, and dotted lines – set 3.

conclusions, inspection of the integral probability distributions shows that the peaks are real.

Another feature of the distributions is their long wings – the lengths of filaments reach about 90, almost the full size of the box. Inspection of Figs. 7,8 and 9 shows that long filaments are those that pass through the branching points and are really collections of several filaments. So, we have to find in future a recipe that would locate the branching points and break the filaments; otherwise we shall loose connection with the void distribution. For the histograms in Fig. 10 this means that there would be additional contributions to the 20-40 length range, which are presently missing.

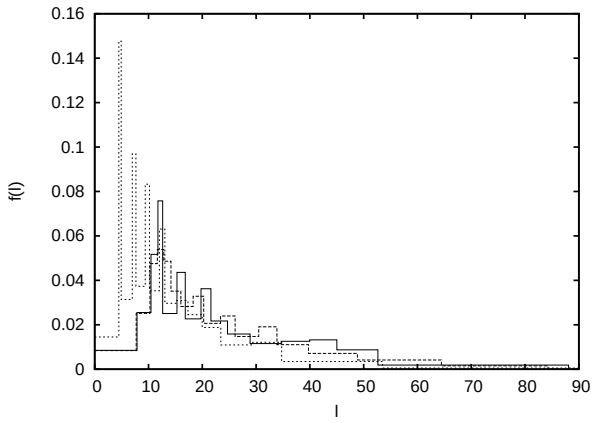

FIG. 11: Combined filament length distribution histograms for the three data sets. Solid line indicates the data set A, dashed line – set B, and dotted line – set C.

As the locations of the peaks almost do not depend on the parameter set used, we combined the length data for the three parameter sets together. These distributions for the three data sets are compared in Fig. 11. Thus these peaks are significant, revealing discrete scales in the data. We also see that the overall length distribution is shifted to the shorter sizes for the set C, compared with the set A.

V. OTHER APPROACHES

There exist only a few methods to describe the observed filamentary networks of galaxies. The best known approach is that of minimal spanning trees (MST) [2]. The minimal spanning tree connects all data points, and its length distribution function describes mainly the nearest-neighbour distance distributions, not the large-scale network we see. The MST has to connect also all points in clusters, while the Candy model can be tuned to ignore them (as we have seen, clusters become usually branching points of the filament network in the Candy model). The differences between the MST and the Candy model can be well seen in Fig. 12. Nevertheless, the MST has been extensively used to describe the filamentary network; as the Candy model looks much better, it would probably lead to a better description of the cosmic filaments.

Using a technique based on multiscale geometric analysis Arias-Castro et al. [30] have shown how filaments can be detected, when they are embedded in a uniformly distributed background of points. This algorithm is specially focused for finding hidden filamentary

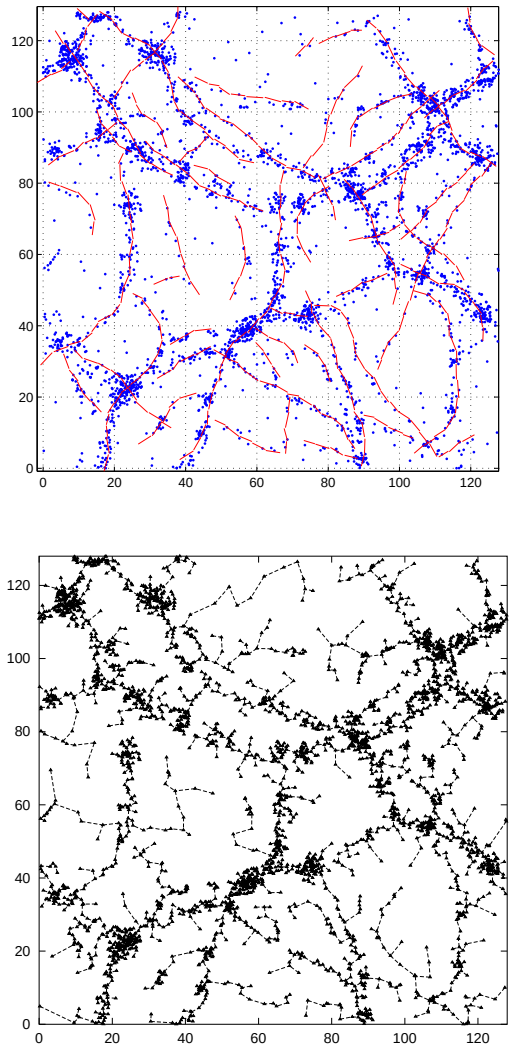

FIG. 12: A Candy model (upper panel) and the minimal spanning tree (lower panel) for the same set of data (set A)).

patterns in images or nearly Poissonian point processes. Our approach is different and is better suited for finding many filaments in clustered point processes.

Another, more recent approach to describe filaments [4] proceeds by binning the map (calculating a density field), and using Minkowski functionals of the isodensity contours to estimate filamentarity of the objects. While this approach will classify all objects, it has two free parameters, the smoothing length (size of the density bin), and the isodensity level. True, in some respect our approach is similar to that, as the segments of the Candy model have a finite width (we are also estimating a density field). But our density estimator is anisotropic and adaptive, in principle, and we trace only filamentary structures.

A third approach that is also based on a density field, is to determine the saddle points

and to build a network of field lines (directed along the gradient of the field), connecting saddle points with local maxima – the skeleton [19]. This approach could reconstruct well the cellular network of filaments (so far it has been applied to study the pixelized cosmic microwave background data [7]), but it will also depend on the density estimation procedure. And, as the authors say, this approach is computationally complex.

VI. CONCLUSION AND PERSPECTIVES

The parameter values for our method were chosen after several trials and errors. Under these circumstances, parameter estimation using Monte Carlo maximum likelihood methods may be considered [9, 17]. These parameters could then be considered as statistics describing the filamentary network. These will certainly be much better suited for this task than the moments of the density distribution in real or in Fourier space, which are commonly used in cosmology.

The data term is a very simple test. Much more sophisticated techniques as testing the alignment of the points covered by a segment, or statistical tests such as the complete randomness test need investigation.

To our knowledge there is no proof for the existence of an optimal cooling scheme when using Metropolis-Hastings dynamics for simulating point processes in a simulated annealing algorithm. There is such a proof for the spatial birth-and-death process, but in practice the authors sample the model using a fixed cold temperature [15]. The choice we opted for, a slow polynomial decreasing scheme, does not guarantee that the global optimum is reached.

But overall, as we have seen, the results are good, the Candy model can be tuned to trace well the filamentary network. And it can be naturally generalized to describe the real 3-D filamentary networks of galaxy maps; see [11]. As we already said above, it can also be considered as a tool to provide statistics of filamentary networks. These are the future directions of our work.

VII. ACKNOWLEDGEMENTS

Enn Saar and Radu Stoica want to thank the hospitality of the Observatori Astronòmic de la Universitat de València where part of this work was done. This work has been supported

by Valencia University through a visiting professorship for Enn Saar, by the Spanish MCyT project AYA2003-08739-C02-01 (including FEDER), by the Generalitat Valenciana project GRUPOS03/170, and by the Estonian Science Foundation grant 4695. We thank Rien van de Weygaert for his code to calculate the MST. The work of Jorge Mateu and Radu Stoica was carried out, respectively, under the grants BFM2001-3286 of the Spanish MCyT and SB2001-0130 from MECD.

---

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

[2] J.D. Barrow, D.H. Sonoda, and S.P. Bhavsar. Minimal spanning tree, filaments and galaxy clustering. *Mon. Not. R. Astr. Soc.*, 216:17–35, 1985.

[3] J.F. Beacom, K.G. Dominik, A.L. Melott, S.P. Perkins, and S.F. Shandarin. Gravitational clustering in the expanding universe - Controlled high-resolution studies in two dimensions. *Astrophysical Journal*, 372, 351–363, 1991.

[4] S. Bharadwaj, S.P. Bhavsar, and J.V. Sheth. The size of the longest filaments in the universe. *astro-ph/0311342*, 2003.

[5] J.R. Bond, L. Kofman, and D. Pogosyan. How Filaments are Woven into the Cosmic Web. *Nature*, 380, 603-606, 1996.

[6] M. Colless et al. (The 2dFGRS Team). The 2dF Galaxy Redshift Survey: Final data release. *astro-ph/0306581*, 2003.

[7] H.K. Eriksen, D.I. Novikov, and P.B. Lilje. Testing for non-Gaussianity in the WMAP data: Minkowski functionals and the length of the skeleton. *astro-ph/0401276*, 2004.

[8] C.J. Geyer and J. Møller. Simulation procedures and likelihood inference for spatial point processes. *Scandinavian Journal of Statistics*, 21:359–373, 1994.

[9] C. J. Geyer. 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, 1999.

[10] P.J. Green. Reversible jump MCMC computation and Bayesian model determination. *Biometrika*, 82:711-732, 1995.

[11] P. Gregori, J. Mateu, and R.S. Stoica. A marked point process for modelling three-dimensional

patterns. *Spatial point process modelling and its applications* (eds. A. Baddeley, P. Gregori, J. Mateu, R.S. Stoica, D.Stoyan), ISBN : 84-8021-475-9, Publicacions de la Universitat Jaume I, 91–99, Castellon, 2004.

[12] A. Jenkins, C.S. Frenk, F.R. Pearce, P.A. Thomas, J.M. Colberg, S.D.M. White, H.M.P. Couchman, J.A. Peacock, G. Efstathiou, and A.H. Nelson. Evolution of structure in cold dark matter universes. *Astrophysical Journal*, 499, 20–40, 1998.

[13] W.S. Kendall and J. Møller. Perfect simulation using dominating processes on ordered spaces, with application to locally stable point processes. *Advances in Applied Probability (SGSA)*, 32:844–865, 2000.

[14] C. Lacoste, X. Descombes and J. Zerubia. A comparative ctudy of point processes for line network extraction in remote sensing. *Research report No. 4516*, INRIA Sophia-Antipolis, 2002.

[15] M. N. M. van Lieshout. Stochastic annealing for nearest-neighbour point processes with application to object recognition. *Advances in Applied Probability*, 26 : 281 –300, 1994.

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

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

[18] M. N. M. van Lieshout and R. S. Stoica. Perfect simulation for marked point processes. *CWI Research Report* PNA-0306, 2003.

[19] D. Novikov, S. Colombi, and O. Doré. Skeleton as a probe of the cosmic web: the 2D case. *astro-ph/0307003*, 2003.

[20] C.J. Preston. Spatial birth-and-death processes. *Bulletin of the International Statistical Institute*, 46:371–391, 1977.

[21] B.D. Ripley and F.P. Kelly. Markov point processes. *Journal of the London Mathematical Society*, 15:188–192, 1977.

[22] D. Ruelle. *Statistical mechanics*. Wiley, New York, 1969.

[23] S.A. Shectman, S.D. Landy, A. Oemler, D.L. Tucker, H. Lin, R.P. Kirchner, and P.L. Schecter. The Las Campanas Redshift Survey. *Astrophysical Journal*, 470:172–188, 1996.

[24] R.K. Sheth and R. van de Weygaert. A hierachy of voids: Much ado about nothing. *astro-ph/0311260*, 2003.

[25] R.S. Stoica. *Processus ponctuels pour l'extraction des réseaux linéiques dans les images satellitaires et aériennes*. PhD Thesis (in French), Nice Sophia-Antipolis University, 2001.

[26] 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 (eds.), *Spatial statistics through applications*, WIT Press, Southampton, UK, 2002.

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

[28] M. Tegmark, M. Zaldarriaga, and A.J.S. Hamilton. Towards a refined cosmic concordance model: Joint 11-parameter constraints from CMB and large-scale structure. *Physical Review D*, 63, 043007.

[29] D.G. York *et al* (The SDSS Collaboration). The Sloan Digital Sky Survey: Technical summary. *Astronomical Journal*, 120:1579–1587, 2000.

[30] E. Arias-Castro, D. Donoho, and X. Huo. Adaptive multiscale detection of filamentary structures embedded in a background of uniform random points. Stanford Technical Report, 2003.

[31] 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 $h \approx 0.71$.

[32] The histograms shown are equal-probability histograms, with equal areas in every bin. Although not common, it is a good, non-parametrical way to represent probability densities.