Detecting filamentary pattern in the cosmic web: a catalogue of filaments for the SDSS

E. Tempel,[1,2]★ R. S. Stoica,[3,4] V. J. Martínez,[5,6] L. J. Liivamägi,[1,7] G. Castellan[3] and E. Saar[1,8]

[1]*Tartu Observatory, Observatooriumi 1, EE-61602 Tõravere, Estonia*
[2]*National Institute of Chemical Physics and Biophysics, Rävala pst 10, EE-10143 Tallinn, Estonia*
[3]*Université Lille 1, Laboratoire Paul Painlevé, F-59655 Villeneuve d'Ascq Cedex, France*
[4]*Institut de Mécanique Céleste et Calcul d'Ephémérides, Observatoire de Paris, F-75014 Paris, France*
[5]*Observatori Astronòmic, Universitat de València, C/ Catedràtic José Beltrán 2, E-46980 Paterna, Spain*
[6]*Departament d'Astronomia i Astrofísica, Universitat de València, E-46100-Burjassot, València, Spain*
[7]*Institute of Physics, University of Tartu, EE-51010 Tartu, Estonia*
[8]*Estonian Academy of Sciences, Kohtu 6, EE-10130 Tallinn, Estonia*

ABSTRACT

The main feature of the spatial large-scale galaxy distribution is its intricate network of galaxy filaments. This network is spanned by the galaxy locations that can be interpreted as a three-dimensional point distribution. The global properties of the point process can be measured by different statistical methods, which, however, do not describe directly the structure elements. The morphology of the large-scale structure, on the other hand, is an important property of the galaxy distribution. Here, we apply an object point process with interactions (the Bisous model) to trace and extract the filamentary network in the presently largest galaxy redshift survey, the Sloan Digital Sky Survey (SDSS). We search for filaments in the galaxy distribution that have a radius of about $0.5\,h^{-1}$ Mpc. We divide the detected network into single filaments and present a public catalogue of filaments. We study the filament length distribution and show that the longest filaments reach the length of $60\,h^{-1}$ Mpc. The filaments contain 35–40 per cent of the total galaxy luminosity and they cover roughly 5–8 per cent of the total volume, in good agreement with *N*-body simulations and previous observational results.

Key words: methods: data analysis – methods: statistical – catalogues – galaxies: statistics – large-scale structure of Universe.

1 INTRODUCTION

Large galaxy redshift surveys reveal that the Universe has a salient web-like structure, called the cosmic web (Jõeveer, Einasto & Tago 1978; Bond, Kofman & Pogosyan 1996). Galaxies and matter in the Universe are arranged into a complex web-like network of dense compact clusters, elongated filaments, weak two-dimensional sheets and huge near-empty voids.

The cosmic web is one of the most intriguing and striking patterns found in nature, rendering its analysis and characterization far from trivial. The absence of objective and quantitative procedures for identifying and isolating clusters, filaments, sheets and voids in the large-scale matter distribution has been a major obstacle in investigating the structure and dynamics of the cosmic web. On the other hand, identification and quantitative description of the details of the cosmic web is important for a broad range of cosmological issues. It contains information about the structure formation physics,

and is a rich source of information on the global cosmology. The evolution, structure and dynamics of the cosmic web depend on the nature of dark matter and dark energy, and on the properties of the initial density fluctuations generated in the very early Universe. Thus, these factors must have left their imprint on the web, on its geometry and topology. Thus, probes of the large-scale structure, such as wide and deep galaxy surveys, enable us to test current physical and cosmological theories and improve our understanding of the Universe.

From an observational point of view, there is clear evidence that certain observed properties of galaxies correlate with their environment. For example, the morphology–density relation stipulates that elliptical galaxies are found preferentially in crowded environments and spiral galaxies are found in the field (Einasto et al. 1974; Dressler 1980). The same kind of correlation can be found in terms of the colours and morphology of galaxies (Blanton et al. 2005; Tempel et al. 2011), their star formation histories and ages.

Usually, in environmental studies only the local or global density is used, but various indications argue for a more intricate connection (Lee & Lee 2008). While all morphological types of galaxies

correspond to a well-defined range in density, this alone is not sufficient to differentiate between them: the connection between density and morphology is more intricate. It is also known that the spin of dark matter haloes is correlated with the underlying web elements (Navarro, Abadi & Steinmetz 2004; Aragón-Calvo et al. 2007a; Brunino et al. 2007; Zhang et al. 2009; Hahn, Teyssier & Carollo 2010; Codis et al. 2012; Libeskind et al. 2012, 2013; Aragón-Calvo 2013; Trowland, Lewis & Bland-Hawthorn 2013). Observations indicate that the rotation axes of galaxies are aligned with galaxy filaments (Trujillo, Carretero & Patiri 2006; Lee & Erdogdu 2007; Jones, van de Weygaert & Aragón-Calvo 2010; Cervantes-Sodi, Hernandez & Park 2010; Tempel & Libeskind 2013; Tempel, Stoica & Saar 2013; Zhang et al. 2013). Comparing the properties of galaxies with the structure of the cosmic web yield valuable information about the formation and evolution of galaxies.

Galaxy maps are visually dominated by filaments. Filaments are traced by galaxies and groups and often occupy the regions between massive clusters (Pimbblet, Drinkwater & Hawkrigg 2004; Murphy, Eke & Frenk 2011; Dietrich et al. 2012; Jauzac et al. 2012); however, filaments can also be located in voids (Beygu et al. 2013; Rieder et al. 2013). The prominent filamentary channels may contain up to 40 per cent of the matter in the Universe (Forero-Romero et al. 2009; Jasche et al. 2010). Also, theoretical studies (e.g. Cen & Ostriker 1999) have suggested that around half of the warm gas in the Universe, presumably accounting for the low-redshift missing baryons (Fukugita, Hogan & Peebles 1998; Viel et al. 2005), is hidden in filaments.

Translating the visual impression of the cosmic web into an algorithm that classifies the local geometry into different environments is not a trivial task, and much work is being done in this direction. Cautun, van de Weygaert & Jones (2013) gives a good overview about the various structure finding algorithms currently available. Among them are the algorithms based on the gravitational tidal tensor – the Hessian of the gravitational potential (Hahn et al. 2007; Lee & Lee 2008; Forero-Romero et al. 2009; Bond, Strauss & Cen 2010a,b; Wang et al. 2012), on the velocity field (Shandarin 2011; Hoffman et al. 2012; Wang et al. 2013), *skeleton analysis* (Novikov, Colombi & Doré 2006; Sousbie et al. 2008), *watershed segmentation* (Platen, van de Weygaert & Jones 2007; Aragón-Calvo et al. 2010b), the tessellations (Doroshkevich, Gottlöber & Madsen 1997; Aragón-Calvo et al. 2007b; Aragón-Calvo, van de Weygaert & Jones 2010a; González & Padilla 2010; Sousbie 2011; Sousbie, Pichon & Kawahara 2011; Shandarin, Habib & Heitmann 2012; Aragón-Calvo 2013), Bayesian sampling of the density field (Jasche et al. 2010), minimal spanning tree (Alpaslan et al. 2013) and multiscale probability mapping (Smith et al. 2012). All these methods are based on different assumptions and provide different results. Of course, any environment finder should be evaluated by its merits. A good algorithm should provide a quantitative classification which agrees with the visual impression and it should be based on a robust and well-defined numerical scheme.

In this work, the detection of filaments is performed using a marked point process with interactions, called Bisous model (Stoica, Gregori & Mateu 2005a). This model approximates the filamentary network by a random configuration of small segments or cylinders that interact and connect while building the network. The model was already successfully applied to observational data and to mock catalogues (Stoica, Martínez & Saar 2007b, 2010). The filaments found in these papers delineate well the filaments detected by eye and they were evaluated by Monte Carlo statistical tests. This approach has the advantage that it works directly with the original point process and does not require smoothing to create a continuous density field. Our method can be applied to relatively poorly sampled data sets, as the galaxy maps are; it can be applied both to observations and simulations. Some of the previously mentioned methods can be applied only to simulations, which makes their use limited.

Here, our marked point process methodology is adapted in order to apply it to the Sloan Digital Sky Survey (SDSS) data set. Based on the detection obtained using the Bisous model, filament spines are extracted and a filaments catalogue is built. This compiled data can be further used in order to study the properties of filaments and galaxies therein, and their relationship with galaxy clusters. Most of all the previously cited filaments detection methods are based on the calculus of some gradient, statistics or other measures characterizing locally the filament, followed by a merging, tracking or filtering procedure. The main advantage of using a marked point process methodology is that it comes freely with a natural way of integration provided by the probability theory. In this way, a simultaneous morphological and statistical characterization of the filamentary pattern is allowed. Completing this approach with the spine detection, connects this probabilistic methodology with the richness and the efficiency of the deterministic techniques already developed.

The paper is organized as follows. In Sections 2 and 3, we describe the data used and the mathematical tools. In Section 4, we define how we build the simulation and extract the spines of filaments. In Sections 5 and 6, we present and discuss our results. The description of the catalogue is given in Appendix A.

Throughout this paper, we assume the *Wilkinson Microwave Anisotropy Probe* cosmology: the Hubble constant $H_0 = 100\,h\,\mathrm{km\,s^{-1}Mpc^{-1}}$, the matter density $\Omega_\mathrm{m} = 0.27$ and the dark energy density $\Omega_\Lambda = 0.73$ (Komatsu et al. 2011).

2 SDSS DATA

Our present study is based on the SDSS data release 8 (York et al. 2000; Aihara et al. 2011). The galaxy redshifts are typically accurate to $\sim$30 km s^{-1}, making it ideal for studies of the large-scale structure. We use only the main contiguous area of the survey (the Legacy Survey) and the spectroscopic galaxy sample as compiled in Tempel, Tago & Liivamägi (2012). The lower Petrosian magnitude limit for this sample is set to $m_r = 17.77$, since for fainter galaxies, the spectroscopic sample is incomplete (Strauss et al. 2002). To exclude the Local Supercluster from the sample, the lower cosmic microwave background (CMB) corrected distance limit $z = 0.009$ was used. The upper limit was set to $z = 0.155$ (450 h^{-1} Mpc), since at larger distances the sample becomes very diluted. The sample includes 499 340 galaxies.

Due to the peculiar velocities of galaxies, which introduce Doppler effects in the redshift measurement (Jackson 1972; Davis & Peebles 1983; Kaiser 1987), the compact structures in redshift-space are elongated along the line of sight. This is the so-called *finger-of-god effect*, as first introduced by Tully & Fisher (1978). To find the filamentary structure in the SDSS data, we have to suppress first the redshift distortions for groups. For that we use the friends-of-friends (FoF) groups compiled in Tempel et al. (2012); the details of the group finding algorithm are explained in Tago et al. (2008, 2010). We spherize the groups using the rms sizes of galaxy groups in the sky and their rms radial velocities as described in Liivamägi, Tempel & Saar (2012) and Tempel et al. (2012). The method is similar to that proposed by Tegmark et al. (2004). Such a compression will lead to a better estimate of the density field and can help to find the real filamentary structure. Nevertheless,

this compression may suppress some of the line-of-sight filaments, since the FoF group finding algorithms cannot distinguish between groups and exactly line-of-sight filaments. Thus, unique recovery of the real-space structures is generally not possible. We note that in principle the redshift-space distortions can be modelled more accurately, introducing a density-dependent peculiar velocity sampling scheme (Kitaura & Enßlin 2008; Heß, Kitaura & Gottlöber 2013). However, we defer this to future work, since this will affect only a small number of filaments.

To define the filamentary structure, we use Cartesian coordinates based on the SDSS angular coordinates η and λ, allowing the most efficient placing of the galaxy sample cone inside a brick: we used the same coordinates to define the superclusters of galaxies in Liivamägi et al. (2012). The galaxy coordinates are calculated as follows:

$$x = -d_{\mathrm{gal}} \sin \lambda$$

$$y = d_{\mathrm{gal}} \cos \lambda \cos \eta$$

$$z = d_{\mathrm{gal}} \cos \lambda \sin \eta, \tag{1}$$

where d_{gal} is the finger-of-god suppressed comoving distance to a galaxy.

We refer to Tempel et al. (2012) for a more detailed description of the galaxy sample.

3 MATHEMATICAL TOOLS

In this section, we describe the main tools we use to study and extract the filamentary pattern of the galaxy distribution in the Universe. First of all, a very short and intuitive definition of marked point processes is given. For a rigorous study of this subject, we recommend as a starting point Stoyan, Kendall & Mecke (1995), van Lieshout (2000) and Møller & Waagepetersen (2004). Next, our marked point process based methodology is presented. This methodology includes: the construction of the Bisous model, a simulation algorithm and an optimization procedure. For a detailed mathematical description together with the necessary convergence proofs of the method, we recommend Stoica et al. (2005a, 2007b, 2010).

3.1 Marked point processes

Point processes are random configuration of points. If the points are labelled using a random mark, we speak about a marked point process. If the marks are the characteristics of a random geometrical object, we may say that we have an object point process. These processes were used by Martínez & Saar (2002) to study the spatial distribution of galaxies. The observed galaxies were seen as the realization of a marked point process, as follows. The centres of the galaxies were the locations in a point process, whereas the different characteristics of the galaxies (mass, luminosity, etc.) were the marks associated with the corresponding locations. The marked point processes mathematical framework allowed the authors to describe the galaxy population, to define statistical descriptors, and to derive the corresponding estimators (Martínez, Arnalte-Mur & Stoyan 2010).

The simplest marked point process is the Poisson process. In this process, the number of points is chosen according to a Poisson distribution, while the points are spread independently uniform in the location space where the marked process lives. Then, to each point a mark is attached independently identically distributed with

respect to the marks distribution. The previous process is called simple, because the independence assumption involves no interaction between objects. Such interactions can be defined by means of a probability density with respect to the reference measure given by the unit intensity marked Poisson point process (van Lieshout 2000; Møller & Waagepetersen 2004; Stoica et al. 2005a).

3.2 Bisous model

The marked point process we propose for filamentary detection is different from the ones already used in cosmology. In fact, we do not model the galaxies, but the structure outlined by the galaxy positions.

Let K be a cosmological sample of finite volume $0 < \nu(K) < \infty$, where a finite number of galaxies $\boldsymbol{d} = \{d_1, d_2, \ldots, d_n\}$ are observed. The galaxies positions are measured in Cartesian coordinates. For SDSS, equation (1) defines the coordinates. The feature we are interested in is the filamentary network outlined by the galaxies positions.

The main hypothesis of our work is that the filamentary network is made of a random configuration of connected and aligned cylinders, that is the realization of a marked point process. This marked point process is named Bisous model and it was specially designed to generate and analyse random spatial patterns (Stoica et al. 2005b, 2007b). We assume that locally, galaxies may be grouped together inside a rather small cylinder. We also assume that such small cylinders may combine to form a filament if neighbouring cylinders are aligned in similar directions. So, the elements of our marked point process are the centres of the cylinders and their corresponding geometrical shapes. Cylinders are located in the same volume where galaxies are.

A cylinder is an object characterized by its centre $\boldsymbol{k} \in K$ and shape parameters. The shape parameters of a cylinder are the radius r, the length h and the orientation vector ω. The radius is considered fixed. The length varies uniformly within the interval $[h_{\min}, h_{\max}]$ that will be specified later in this paper. The orientation vector parameters $\boldsymbol{\omega} = \phi(\eta, \tau)$ are uniformly distributed on $[0, 2\pi] \times [0, 1]$, such that

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

Hence, the mark of our process is given by $M = [h_{\min}, h_{\max}] \times [0, 2\pi] \times [0, 1]$ and its attached uniform distribution ν_M. We denote the cylinder by $s(y) = s(\boldsymbol{k}, r, h, \boldsymbol{\omega}) \subset K$.

Two extremity rigid points (end points) are attached to each cylinder $s(y)$. Around each of these points a sphere of the radius r_{a} is centred. These two spheres form an attraction region that are used to define connectivity and alignment rules for cylinders (see Section 3.4). We illustrate the basic cylinder in Fig. 1, where it is centred at the coordinate origin and its symmetry axis is parallel to the x-axis. The coordinates of the end points are

$$e_u = \left(\frac{h}{2} (-1)^{u+1}, 0, 0 \right), \qquad u \in \{1, 2\} \tag{3}$$

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

Let $\boldsymbol{y} = \{y_1 = (\boldsymbol{k}_1, m_1), y_2 = (\boldsymbol{k}_2, m_2), \ldots, y_n = (\boldsymbol{k}_n, m_n)\}$ be a configuration of cylinders, where m_i denotes the mark. The unit intensity independently marked Poisson process constructs a configuration of cylinders as follows. First, the number n of cylinders is chosen according to a Poisson law of parameter $\nu(K)$. Then the lengths and the orientation vectors are chosen independently following ν_M. Such a configuration has only very few connected and aligned cylinders. This effect is just a chance product. In order to
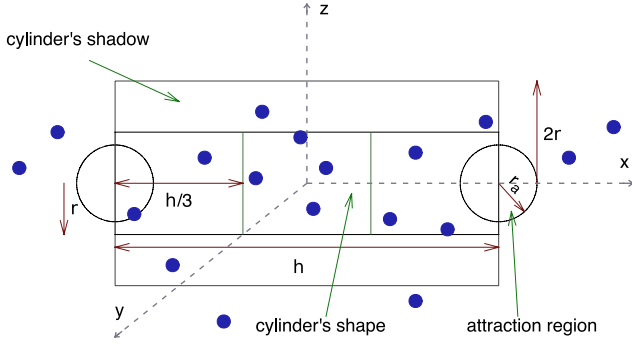
Figure 1. A two-dimensional projection of a cylinder with its shadow within a pattern of galaxies (points). The attraction regions are shown as spheres. The exact shape of the cylinder, its shadow and attraction regions depend on the model.

obtain random configurations made of connected and aligned cylinders, a model defined by a probability density is needed. Such a probability density is specified with respect to the reference Poisson process and it can be written

$$p(\mathbf{y}|\theta) = \frac{\exp\left[-U(\mathbf{y}|\theta)\right]}{Z(\theta)}, \tag{4}$$

where $Z(\theta)$ is the normalizing constant, θ is the vector of the model parameters and $U(\mathbf{y}|\theta)$ is the energy function of the system (its equivalent in physics is the total Gibbs energy of a system).

We assumed above that locally, galaxies may be grouped together inside a rather small cylinder, and such small cylinders may combine to form a filament if neighbouring cylinders are aligned in similar directions.

Following these two ideas the energy function in (4) can be specified as

$$U(\mathbf{y}|\theta) = U_{\mathrm{d}}(\mathbf{y}|\theta) + U_{\mathrm{i}}(\mathbf{y}|\theta), \tag{5}$$

where $U_{\mathrm{d}}(\mathbf{y}|\theta)$ is the data energy (see Section 3.3) and $U_{\mathrm{i}}(\mathbf{y}|\theta)$ is the interaction energy (see Section 3.4) associated with the first and second assumptions above, respectively. In fact, the data energy is related to the position of the cylinders in the galaxy field, whereas the interaction energy is related to the alignment and connection of the cylinders constructing the filamentary pattern.

Being in the possession of the model, the parameters have to be chosen. Here, the Bayesian framework is adopted and the parameters are described by a prior law $p(\theta)$ (Stoica, Gay & Kretzschmar 2007a; Stoica et al. 2007b, 2010). This allows us to write the joint estimator of the filamentary pattern and the parameters as

$$(\widehat{\mathbf{y}}, \widehat{\theta}) = \arg\max_{\Omega \times \Psi} p(\mathbf{y}, \theta) = \arg\max_{\Omega \times \Psi} p(\mathbf{y}|\theta)p(\theta)$$

$$= \arg\min_{\Omega \times \Psi}\left\{\frac{U_{\mathrm{d}}(\mathbf{y}|\theta) + U_{\mathrm{i}}(\mathbf{y}|\theta)}{Z(\theta)} + \frac{U_{\mathrm{p}}(\theta)}{Z_{\mathrm{p}}(\theta)}\right\}, \tag{6}$$

where $p(\theta) = \exp[-U_{\mathrm{p}}(\theta)]/Z_{\mathrm{p}}(\theta)$ is the prior law for the model parameters and Ψ is the model parameters space.

The Bayesian framework was preferred, since we believe, that for the problem at hand, it is much more natural to give a characterization of the parameters by a probabilistic law, instead of a fixed value. Nevertheless, even in this case, some tuning of the model based on trial and error, is needed. The solution we obtain is not unique. In practice, the shape of the prior law $p(\theta)$ may influence the solution, making the result to look more random compared with a result obtained for fixed values of parameters. Therefore, we have derived tools that are able to average the obtained solution and to

state that the obtained results are really due to the data, and not to a random effect of the presented methodology (Stoica et al. 2007a,b, 2010). Full details concerning the set-up of the method and the analysis of the results obtained are given later in this paper.

The paper continues with the presentation of the energy terms, the simulation technique and an optimization algorithm.

3.3 Data energy

The data energy term is related to the local definition of a galactic filament. This is still an important open problem. Here, we consider that locally, the galaxies positions form a filament, if they are situated inside a rather small cylinder, while fulfilling simultaneously several criteria. The first one is that the galaxies positions should be spread more or less uniformly along the main symmetry axis of the cylinder. The second one is that inside a cylinder there should be more galaxies than outside of it, that is inside the close-by neighbourhood of the cylinder. And finally, in order to avoid some clustering effect, the galaxies forming the filaments should be encouraged to get aligned as much as possible along the main symmetry axis of the cylinder.

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

$$U_{\mathrm{d}}(\mathbf{y}|\theta) = -\sum_{y \in \mathbf{y}} v(y), \tag{7}$$

where $v(y)$ is the potential function associated with the cylinder y. This potential takes into account the previously mentioned criteria and it depends on $\mathbf{d}$ (the field of galaxies) and the model parameters given by θ.

In order to give a mathematical description of these requirements, an extra cylinder is attached to each cylinder y, with exactly the same parameters as y, except for the radius which equals $2r$. Let $\tilde{s}(y)$ be the shadow of $s(y)$ obtained by the subtraction of the initial cylinder from the bigger cylinder. The cylinder and its shadow are shown in Fig. 1. Then, we divide each cylinder into three equal volumes along its main symmetry axis, and denote by $s_1(y)$, $s_2(y)$ and $s_3(y)$ their corresponding shapes.

Let us assume that locally the number of galaxies inside and around a cylinder, follows a Poisson distribution.

The first criteria requires the 'local uniform spread' of the galaxies along the main symmetry axis of the cylinder. Under the Poissonian assumption, let λ_i, $i = 1, \ldots, 3$ be the intensity parameters of the corresponding distributions for the shape regions $s_i(y)$. If the underlying Poissonian process is stationary, 'local uniform spread' requires all the λ_i to be equal. However, filaments are lumpy by nature – e.g., the filaments in Pimbblet et al. (2004) and the well-known Perseus Chain (Jõeveer & Einasto 1978). To take this into account, we relax the uniformity assumption by requiring λ_i/λ_j to be smaller than chosen threshold.

For any two regions $s_i(y)$ and $s_j(y)$ with $i \neq j$, a statistical test can be done to compare λ_i and λ_j (Przyborowski & Wilenski 1940; Kirshnamoorty & Thomson 2004). The test is

$$H_{0,i,j} : \frac{\lambda_i}{\lambda_j} \leq \rho_{\mathrm{u}} \quad \text{against} \quad H_{\mathrm{a},i,j} : \frac{\lambda_i}{\lambda_j} > \rho_{\mathrm{u}}, \tag{8}$$

for all pairs of indices $i, j \in \{1, \ldots, 3 | i \neq j\}$, where $\rho_{\mathrm{u}} \geq 1$ is a given threshold value.

Now, for a given pair (i, j), the observed number of galaxies in $s_i(y)$ and $s_j(y)$ is $X_i = m$ and $X_j = n$, respectively. Then, the p-value for this test is computed using a binomial law of parameters $n + m$

and $p(\rho_u) = \rho_u/(1 + \rho_u)$ as follows:

$$p_u(s_i(y), s_j(y)) = \boldsymbol{P}(X_i \geq m | X_i + X_j = m + n, p(\rho_u))$$

$$= \boldsymbol{P}(\mathrm{Bin}(n + m, p(\rho_u)) \geq m). \tag{9}$$

Six such tests are necessary to verify the 'local uniform spread' condition. The obtained score is

$$p_u(y) = \min\{p_u(s_i(y), s_j(y)), \ i, j \in \{1, \ldots, 3 | i \neq j\}\}. \tag{10}$$

Notice, that this global test is equivalent with verifying $1/\rho_u \leq \lambda_i/\lambda_j \leq \rho_u$ with $i < j$. Hence, it guarantees a minimum density for the galaxies inside a cylinder cell.

The second criteria demands for 'locally high density' of galaxies inside of a cylinder comparing to the density of galaxies in the close-by neighbourhood of the cylinder. Under the Poissonian assumption, the test is

$$H_0 : \frac{\lambda}{\tilde{\lambda}} \geq \rho_h \quad \text{against} \quad H_a : \frac{\lambda}{\tilde{\lambda}} < \rho_h, \tag{11}$$

with ρ_h a given threshold value. It is important to note that the volumes ratio of $s(y)$ and $\tilde{s}(y)$ plays an important role in choosing the appropriate value of ρ_h. For instance, if $\rho_h = 1/3$ this tests if the two processes have the same intensity, if $\rho_h = 1$ this tests if the intensity inside the cylinder is three times higher than inside its shadow.

If the observed number of galaxies inside $s(y)$ and $\tilde{s}(y)$ is $X = m$ and $\tilde{X} = n - m$, respectively, the p-value of the test is computed using a binomial distribution of parameters n and $p(\rho_h) = \rho_h/(1 + \rho_h)$

$$p_h(y) = \boldsymbol{P}(X \leq m | X + \tilde{X} = n, p(\rho_h))$$

$$= \boldsymbol{P}(\mathrm{Bin}(n, p(\rho_h)) \leq m). \tag{12}$$

To take into account both tests simultaneously, the following score is defined

$$p_{\mathrm{hyp}}(y) = p_u(y) \cdot p_h(y). \tag{13}$$

The previous tests are based on counts of points in some pre-defined regions. In order to take into account the spatial distribution of galaxies in a cylinder, we define the cylinder concentration as

$$\sigma^2 = \frac{1}{n - 2} \sum_{j=1}^{n} \frac{\delta_j^2}{r^2}, \tag{14}$$

with n the number of galaxies covered by the cylinder, δ_j the distance from the jth galaxy inside the cylinder to its main symmetry axis, and r is the cylinder radius. The weight $1/(n - 2)$ is chosen to eliminate some pathological cases with too few points covered by a cylinder (there must be at least three galaxies inside a cylinder). Clearly, the concentration σ^2 has a minimum when the symmetry axis of the cylinder coincides with the least mean square line passing through the cloud of points, given by the galaxy positions inside the cylinder.

It is important to note that the use of this term may induce a local Gaussian assumption. This may be considered contradictory to the Poisson hypothesis previously used. Nevertheless, the use of these two strategies is complementary: the first two requirements impose conditions on the number of galaxies inside a cylinder, while the third one imposes conditions on the spatial distribution of these galaxies. The term given by (14) is also a very good indicator of a locally high density and a better estimator for the filament axis.

The potential function $v(y)$ of the cylinder is built using the previous statistical tests and criteria. Let us assume that for the

cylinder y, the p-value p_{hyp} is computed as previously, and let $\sigma^2(y)$ be the cylinder concentration. We want $v(y)$ to be maximum for the 'best' location of the cylinder in the galaxy field. This allows the definition of the potential function as

$$v(y) = \begin{cases} -\sigma^2(y) + c_{\mathrm{hyp}} \log\left[p_{\mathrm{hyp}}(y)\right] & \text{if} \quad n \geq n_{\min} \\ -\infty & \text{if} \quad n < n_{\min} \end{cases} \tag{15}$$

with n the number of galaxies covered by the cylinder $s(y)$ and $n_{\min}$ a given threshold value. Here, the formula (14) suggests $n_{\min} \geq 3$. The parameter $c_{\mathrm{hyp}} \geq 0$ is required to make the two terms comparable: this allows us to use these two strategies safely together.

This gives for the data energy defined by equation (7):

$$U_d(\boldsymbol{y}|\theta) = -\sum_{y \in \boldsymbol{y}} \left\{ c_{\mathrm{hyp}} \log\left[p_{\mathrm{hyp}}(y)\right] - \sigma^2(y) \right\}, \tag{16}$$

and for the data term model

$$p_d(\boldsymbol{y}|\theta) \propto \exp\left[-U_d(\boldsymbol{y}|\theta)\right]$$

$$\propto \exp\left[-\sum_{y \in \boldsymbol{y}} \sigma^2(y)\right] \prod_{y \in \boldsymbol{y}} [p_{\mathrm{hyp}}(y)]^{c_{\mathrm{hyp}}}. \tag{17}$$

The data term model (17) is a super-position of inhomogeneous Poisson point processes with respect to the reference measure. Since the number of galaxies is finite and since the observed window has a limited volume, the term σ^2 is always finite. Therefore, the data term model is locally stable, hence it designs a well-defined model that has an integrable probability density.

One more point has to be retained concerning the data term. The use of p-values for constructing a potential function is different of the use of the values for 'purely' statistical tests. In this last situation, Bonferroni or Šidák corrections are required.

3.4 Interaction energy

The interaction energy is related to the relative position of the cylinders forming the network and its expression is as follows

$$U_i(\boldsymbol{y}|\theta) = -n_k(\boldsymbol{y}) \log \gamma_k - \sum_{s=0}^{2} n_s(\boldsymbol{y}) \log \gamma_s, \tag{18}$$

where $n_k(\boldsymbol{y})$ is the number of repulsive cylinder pairs and $n_s(\boldsymbol{y})$ is the number of cylinders connected to the network through s extremity points. The variables $\log \gamma_k$ and $\log \gamma_s$ are the potentials associated with these configurations, respectively.

The interaction energy (18) is defined in the same way as in Stoica et al. (2010).

Two cylinders are considered repulsive, if they are rejecting each other and if they are not orthogonal. We say that two cylinders $y_1 = (\boldsymbol{k}_1, r, h_1, \boldsymbol{\omega}_1)$ and $y_2 = (\boldsymbol{k}_2, r, h_2, \boldsymbol{\omega}_2)$ reject each other if their centres are closer than the minimum allowed distance between cylinders, $d(\boldsymbol{k}_1, \boldsymbol{k}_2) < 0.5(h_1 + h_2) - r_a$. Two cylinders are considered to be orthogonal if $|\boldsymbol{\omega}_1 \cdot \boldsymbol{\omega}_2| \leq \tau_\perp$, where $\cdot$ is the scalar product of the two orientation vectors and $\tau_\perp \in (0, 1)$ is a predefined parameter. So, a certain range of mutual angles is allowed for cylinders considered to be orthogonal.

Two cylinders are connected if they attract each other, do not reject each other and are well aligned. Two cylinders attract each other if the distance between the cylinder end points is smaller than the interaction radius r_a (see Fig. 1). Two cylinders are well aligned if $|\boldsymbol{\omega}_1 \cdot \boldsymbol{\omega}_2| \geq 1 - \tau_\parallel$, where $\tau_\parallel \in (0, 1)$ is a predefined parameter.

To illustrate these definitions, we show an example configuration of cylinders (in two dimensions) in Fig. 2. Altogether, the
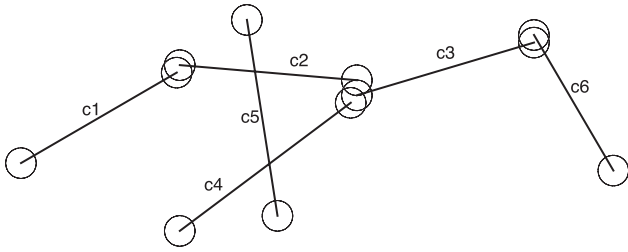
Figure 2. A two-dimensional representation of a cylinder configuration: attraction regions are shown with spheres. In this configuration, we observe that the cylinders $c1 - c2$, $c2 - c3$ and $c3 - c4$ are connected. The cylinders $c1$, $c3$ and $c4$ are connected to the network through one end point, while $c2$ is connected to the network through both end points. The cylinders $c5$ and $c6$ are not connected to anything, $c3$–$c6$ are attracting each other but they are not well aligned, and $c5$ is not attracted to any other cylinder. The cylinder $c5$ is rejecting the cylinders $c2$ and $c4$ (the centres of these cylinders are too close), but as it is rather orthogonal both to $c2$ and $c4$, it is not repulsing them. The cylinders $c2$ and $c4$ reject each other and are not orthogonal, so they form a repulsive pair.

configuration at Fig. 2 adds to the interaction energy contributions from three single-connected cylinders ($c1$, $c3$, $c4$), one doubly connected cylinder ($c2$), two free cylinders ($c5$, $c6$) and one repulsive cylinder pair ($c2$, $c4$).

The complete model (4) that includes the definition of the data energy and of the interaction energy is well defined for parameters γ_0, γ_1, $\gamma_2 > 0$, $c_{hyp} \geq 0$ and $\gamma_k \in [0, 1]$. The definition of the interactions and the parameter ranges chosen ensure that the complete model is locally stable (van Lieshout 2000; Møller & Waagepetersen 2004; Stoica et al. 2005a). This property ensures that we can safely use this model without expecting any dangers (integrability, convergence, numerical stability, etc.). The values of the interaction parameters (γ_s, γ_k) and of the data parameter c_{hyp} have to be fixed taking into account the weight of each energy component and also the underlying galaxy field. If the interaction energy parameters are too strong, then the filamentary network may appear in location where the galaxies form no filaments. If the data energy parameters are too strong, then the filamentary pattern will be well located but not really forming a filamentary network. This is a normal compromise to be found such as in solution regularization or Bayesian analysis. The Section 4.1 shows how these parameters were set.

3.5 Simulation of the model and optimization algorithm

Several Markov chains Monte Carlo (MCMC) techniques are available to simulate marked point processes: spatial birth-and-death processes, Metropolis–Hastings (MH) algorithms, reversible jump dynamics or more recent exact simulation techniques (Geyer & Møller 1994; Green 1995; Geyer 1999; Kendall & Møller 2000; van Lieshout 2000; van Lieshout & Stoica 2006).

In this paper, we need to sample from the joint probability density law $p(\mathbf{y}, \theta)$. This is done by using an iterative MCMC algorithm. An iteration of the algorithm consists of two steps. First, a value for the parameter θ is chosen with respect to $p(\theta)$. Then, conditionally on θ, a cylinder pattern is sampled from $p(\mathbf{y}|\theta)$ using an MH algorithm (Geyer & Møller 1994; Geyer 1999).

The MH algorithm we used is built using three types of moves (van Lieshout & Stoica 2003; Stoica et al. 2005a, 2007b, 2010).

(i) Birth: with a probability p_b a new cylinder ζ, sampled from the birth rate $b(\mathbf{y}, \zeta)$, is proposed to be added to the present configuration $\mathbf{y}$. The new configuration $\mathbf{y}' = \mathbf{y} \cup \{\zeta\}$ is accepted with the probability

$$\min\left\{1, \frac{p_d}{p_b} \frac{d(\mathbf{y} \cup \{\zeta\}, \zeta)}{b(\mathbf{y}, \zeta)} \frac{p(\mathbf{y} \cup \{\zeta\})}{p(\mathbf{y})}\right\}. \tag{19}$$

(ii) Death: with a probability p_d a cylinder ζ from the current configuration $\mathbf{y}$ is proposed to be eliminated according to the death proposal $d(\mathbf{y}, \zeta)$. The role of this move is to ensure the detailed balance of the simulated Markov chain and its convergence towards the equilibrium distribution. The probability of accepting the new configuration $\mathbf{y}' = \mathbf{y} \setminus \{\zeta\}$ is computed reversing the ratio (19).

(iii) Change: with a probability p_c we randomly choose a cylinder ζ_{old} in the configuration $\mathbf{y}$ and propose to slightly change its parameters using uniform proposals. For the selected element, we may change its location within the vicinity Δk of its centre and change its orientation within a small angle tolerance $\Delta \omega$ with respect its initial orientation. The new element obtained is ζ_{new}. This move improves the mixing properties of the sampling algorithm. The new configuration $\mathbf{y}' = \mathbf{y} \setminus \{\zeta_{old}\} \cup \{\zeta_{new}\}$ is accepted with the probability

$$\min\left\{1, \frac{p(\mathbf{y} \setminus \{\zeta_{old}\} \cup \{\zeta_{new}\})}{p(\mathbf{y})}\right\}. \tag{20}$$

Some practical details concerning the MH dynamics implementation are given below. For a complete description, we recommend van Lieshout & Stoica (2003) and Stoica et al. (2005a).

The uniform choices $b(\mathbf{y}, \zeta) = 1/\nu(K)$ and $d(\mathbf{y}, \zeta) = 1/n(\mathbf{y})$ are commonly adopted for their simplicity and because they guarantee the necessary convergence properties of the simulated Markov chain, such as irreducibility, Harris recurrence, and geometric ergodicity. For the probabilities p_b, p_d and p_c, all the convergence properties are preserved as long as $p_b + p_d + p_c \leq 1$. Here, $\nu(K)$ is the Lebesgue measure (volume) and $n(\mathbf{y})$ is the number of cylinders in the configuration.

Nevertheless, when the model to simulate exhibits complicated interactions, such an update mechanism built of uniform birth and death proposals may be very slow in practice. Here, the strategy proposed by van Lieshout & Stoica (2003) and Stoica et al. (2005a) is adopted. This strategy uses adapted moves that help the model. In our case, the new cylinder can be added uniformly in K (the observed volume) or can be randomly connected with the rest of the network. This mechanism helps to build a connected network and it can be implemented using a non-uniform mixture for the birth proposal

$$b(\mathbf{y}, \zeta) = \frac{p_1}{\nu(K)} + p_2 b_a(\mathbf{y}, \zeta), \tag{21}$$

with $p_1 + p_2 = 1$ (p_2 is the probability to add a connected cylinder) and $b_a(\mathbf{y}, \zeta)$ is a probability density proposing attracting and well-aligned (e.g. connected) cylinders. The expression of $b_a(\mathbf{y}, \zeta)$ is given by

$$b_a(\mathbf{y}, \zeta) = \frac{1}{n[A(\mathbf{y})]} \sum_{y \in A(\mathbf{y})} \tilde{b}(y, \zeta), \tag{22}$$

where $A(\mathbf{y})$ is the set of cylinders in the configuration $\mathbf{y}$ which have at least one end point able to create connections, and $n[A(\mathbf{y})]$ is the number of such cylinders in the configuration. Note that neglecting the edge effects, $n[A(\mathbf{y})]$ is the number of 0- and 1-connected cylinders in configuration. After choosing uniformly an

object y from the set $A(y)$, a new object $\zeta = (k_\zeta, \omega_\zeta)$ is proposed to be added using the density

$$\tilde{b}(y, \zeta) = \frac{\mathbb{1}\{k \in \tilde{a}(y)\}}{\nu[\tilde{a}(y) \cap K]} \frac{1}{\tau_\parallel}, \qquad (23)$$

where $\tilde{a}(y)$ is the region built from the union of attraction balls of y which are not containing the end of any other attracting cylinder in the configuration y, $\nu[\tilde{a}(y) \cap K]$ is the volume of those attraction balls, and $\mathbb{1}\{\cdot\}$ is the indicator function that selects the cylinders the new cylinder ζ may be connected with. Here, one end point of the proposed connected cylinder is uniformly distributed in $\tilde{a}(y)$ and the orientation ω_ζ is uniformly chosen to satisfy the well-aligned criterion $\omega_\zeta \cdot \omega_y \geq 1 - \tau_\parallel$. Clearly, the summation in equation (22) is effectively over the cylinders the new cylinder ζ can be connected with.

This birth rate leads the model to propose configurations with connected objects much more often than using the simple uniform proposal. In practice, it is also reasonable to sample only in the regions where the data potential is defined: $v(y) > -\infty$. Hence, the Lebesgue measure $\nu(K)$ in this case can be calculated as

$$\nu(K) = \int\limits_{k \in K} \mathbb{1}\left[\left(\int\limits_{m \in M} \mathbb{1}\{v[y(k, m)] > -\infty\} \, \mathrm{d}\nu_M(m) \right) > 0 \right] \mathrm{d}k. \quad (24)$$

In order to perform the maximization of $p(y, \theta)$, the previously described sampling mechanism is integrated into a simulated annealing algorithm. The simulated annealing is a global optimization method. It iteratively samples from $p_n(y, \theta) \propto [p(y|\theta)p(\theta)]^{1/T_n}$, while T_n goes slowly to zero. Stoica et al. (2005a) proved the convergence of a simulated annealing algorithm based on an MH dynamics for marked point processes, if a logarithmic cooling schedule is used. According to this result, the temperature is lowered as

$$T_n = \frac{T_0}{\log n + 1}, \qquad (25)$$

where T_0 is the initial temperature.

4 EXTRACTING AND DEFINING THE FILAMENTS

In this section, we describe how we set up the experiment and extract the filaments. Our aim is to use the result obtained using marked point processes to compile a filament catalogue. Every filament in this catalogue is represented as a spine: a set of points that define the axis of the filament.

4.1 Experimental setup

As described above, we use the data set drawn from the SDSS contiguous area. The sample region K is the observed volume in space.

In order to choose the values for the dimensions of the cylinder, we use the physical dimensions of galaxy filaments that have been observed in more detail (Pimbblet et al. 2004); we used the same values also in our previous papers (Stoica et al. 2007b, 2010): a radius $r = 0.5 \, h^{-1}$ Mpc. The same scale has been also used by Smith et al. (2012) and Tempel et al. (2013) showed that filaments of this size may influence galaxy evolution, so this seems to be the most interesting scale for galaxy filaments. Naturally, the nature of filaments is hierarchical (Shen et al. 2006; van de Weygaert & Bond 2008a,b; Aragón-Calvo & Szalay 2013) and the chosen scale of filaments can be arbitrary. In this paper, we aim to detect

filaments that have the strongest impact on galaxy evolution: for that the scale should be relatively small. Taking into account the data resolution in the SDSS, the scale $r = 0.5 \, h^{-1}$ Mpc is the minimal one we can choose. The length of a cylinder is chosen to be $h = 3.0$–$5.0 \, h^{-1}$ Mpc, which is the shortest possible (the ratio of the cylinders length to its diameter is 3:1 to 5:1). The length in this range is considered to be free to more effectively sample the low number density regions.

We choose the attraction radius $r_a = r$, which ensures that the end points of connected cylinders are not too far apart. For the cosines of the maximum curvature angles, we choose $\tau_\parallel = 0.15$ and $\tau_\perp = 0.3$. This allows for a maximum of $\approx 30°$ between the direction angles of connected cylinders and considers the cylinders to be orthogonal, if the angle between their directions is larger than $\approx 70°$.

The model parameters (r, h, r_a) influence the detection results. If they are too low, no filaments will be detected. If they are too high, the detected filaments will be too wide and/or sparse, and precision will be lost. Still, the precision can be increased and the influence of model parameters can be minimized, when sets of simulations and visit maps are used (see Section 4.2): in a certain manner, it will average the detection results. In this work, the model parameters (r, h, r_a) were chosen based on a previous knowledge and after a visual inspection of the detected filamentary pattern.

Fixing the data and interaction energy parameters require an 'initial guess' of the size of the solution. This guess does not need to be precise. The stochastic algorithmic 'machinery' will do the job, due to its mathematical theoretical properties. Nevertheless, from a practical point of view, if the range of parameters allows only very few cylinders in the configuration, then the detection may be incomplete. On the other hand, if too many cylinders are allowed, then the detection may contain a lot of false alarms. Attention should be also paid when the measure units are fixed. A transformation of the measure units induces a transformation of the model parameters so that the same probabilities are assigned to the same configurations of objects. Still, a direct relation between the change of the measure units and the model parameters is not easy to be derived, because of the non-linear character of the model. Under these circumstances, the strategy we have adopted is the following. It is generally accepted that the filaments occupy roughly 10 per cent of the observed volume (Forero-Romero et al. 2009; Aragón-Calvo et al. 2010a; Jasche et al. 2010). In the actual observed volume $\nu(K)$ of the SDSS, the observed filamentary network is made of (roughly speaking, and with respect to the chosen unit measure) about $N_{\mathrm{cyl}} = 3 \times 10^4$ cylinders. This gives a coarse cylinder density of about $N_{\mathrm{cyl}}/\nu(K)$ of cylinders per unit of volume. Hence, a very intuitive way of fixing the range of parameters is to equalize this density with the probability density given by the model of having a cylinder in an observed location. This probability density is naturally approximated by the conditional intensity of the model. Hence, we get
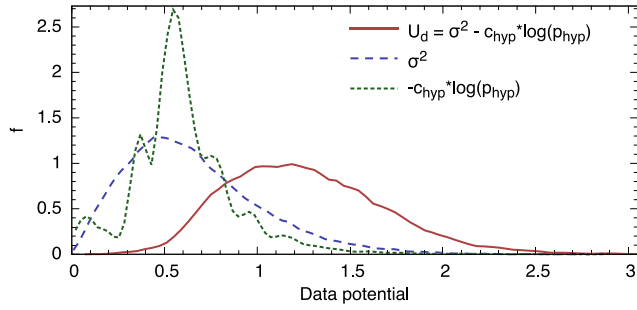
$$\frac{N_{\mathrm{cyl}}}{\nu(K)} = \lambda(\zeta; y) = \frac{p(y \cap \{\zeta\})}{p(y)} \qquad (26)$$

or

$$\log\left[\frac{N_{\mathrm{cyl}}}{\nu(K)}\right] = U(y) - U(y \cap \{\zeta\}), \qquad (27)$$

where ζ is the new cylinder to be added to the configuration y.

The definition of the data energy needs some predefined parameters. To test for the 'locally high density', we fix $\rho_h = 1.0$: the assumed number density in a filament should be at least three times larger than in its outer layers. For the 'locally uniform spread' we

Figure 3. Distribution of cylinder data potentials in a final configuration *y* (solid red line): the contribution to the data potentials from hypothesis testing (dotted green line) and from concentration (dashed blue line) are also shown. The peaks in the hypothesis testing distribution are caused by a small number of galaxies in a cylinder.

**Table 1.** All parameters used to define the model and to extract the filaments. All distances are in h^{-1} Mpc.

Param.	Description	Value
r	cylinder radius	0.5
h	cylinder length	[3.0, 5.0]
ρ_h	locally high density	1.0
ρ_u	locally uniform spread	4.0
n_min	minimum number of galaxies	3
c_hyp	hypothesis test coefficient	[0.7, 0.9]
r_a	attraction radius	0.5
$\tau_\perp$	orthogonal cylinders	0.30
$\tau_\parallel$	parallel cylinders	0.15
γ_k	repulsive cylinders	0
$\log \gamma_0$	0-connected cylinders	[−2.0, −1.0]
$\log \gamma_1$	1-connection cylinders	[−1.0, 0.0]
$\log \gamma_2$	2-connection cylinders	[0.8, 1.8]
$p_\mathrm{b}/p_\mathrm{d}/p_\mathrm{c}$	birth/death/change probabilities	0.5/0.2/0.3
p_2	connected birth probability	0.8
Δk	max shift for change move	0.2
$\Delta \omega$	min cosine for orientation change	0.95
T_0	initial temperature	5.0
δ	steps between temperature changes	100 000
N_iter	number of cycles	200 000
N_sim	number of simulations	50
$\mathcal{L}_\mathrm{lim}$	limiting visit map value for filaments	0.05
$\mathcal{D}_{\mathcal{G}\,\mathrm{lim}}$	limiting orientation strength	0.75
τ_lim	limiting angle for filaments	0.95
κ_lim	limiting curvature for filaments	1.0

set $\rho_\mathrm{u} = 4$, this allows some lumpiness along the filament and at the same time penalizes filaments that cross large clusters. The minimum number of galaxies inside a cylinder is set to be $n_\mathrm{min} = 3$. To assure balance between the concentration and hypothesis testing terms, the constants in front of the hypothesis testing term is chosen to be $c_\mathrm{hyp} \in [0.7, 0.9]$. The value is chosen uniformly within the given interval. The distribution of the cylinder data potentials after the simulation are shown in Fig. 3, together with the distributions of the hypothesis testing term and the cylinder concentration term: we see that for given data and chosen parameters, these two terms are comparable and the overall data potential is reasonable.

For interaction energy, the potentials ($\log \gamma_k$, $\log \gamma_s$) are chosen from a uniform prior density $p(\theta)$. We have opted for this choice since no information concerning the relative strength is available (Stoica et al. 2007a,b, 2010). Still, the general guidelines for fixing the prior parameters are that 2-connected cylinders are generally encouraged, while 1-connected cylinders are slightly penalized and 0-connected cylinders are strongly penalized. This choice encourages the cylinders to group in filaments in those regions where the data energy is good enough. Hence, the prior domain was set to $\log \gamma_0 \in [−2.0, −1.0]$, $\log \gamma_1 \in [−1.0, 0.0]$ and $\log \gamma_2 \in [0.8, 1.8]$. The repulsion parameter $\gamma_k = 0$, so configurations of repulsing cylinders are forbidden. The prior domain for $\log \gamma_s$ was chosen based on the distribution of data energies (see Fig. 3): the chosen domain have to be in balance (in the same range) as data potentials for cylinders.

In Table 1, the parameters of the MH algorithm and the simulated annealing algorithm are also given. For the change move, the maximum shift for the cylinder centre is Δk and the minimal cosine between the old and new direction angles is $\Delta \omega \leq |\boldsymbol{\omega}_\mathrm{old} \cdot \boldsymbol{\omega}_\mathrm{new}|$. We use a uniform prior for cylinder shift in a spherical volume with radius Δk: orientations are taken uniformly on a unit sphere.

4.2 Extracting the filamentary pattern spine

The solution provided by our model is stochastic. Therefore, some variation in the detected patterns is expected for different runs of the method. In Fig. 4 (upper-left panel), a single MCMC simulation is showed, while indicating the different types of cylinders: isolated (grey), single-connected (green) and double-connected (red). The Fig. 4 (upper-right panel) presents the superposition of 25 independent simulations. It can be seen, that the main features of the filamentary pattern are detected by most of the simulations, while differences that appear are due to the random effects of the method.

In order to have a more precise measure of these differences, a brief statistical exploratory analysis was done. To do this, the sufficient statistics of the model (the number of 0-, 1-, 2-connected cylinders) were analysed. In one simulation, the simulated annealing algorithm run during 20×10^9 steps or moves. One step consists of one iteration of the transition kernel of the MH dynamics, that is an accepted or rejected birth or death or change proposal. The temperature was lowered every 10^5 iterations (one cycle): this number of moves was considered sufficiently high in order to obtain almost un-correlated samples. Finally, we get 20×10^4 samples per simulation. In all, we have considered 50 independent simulations.

The results of the exploratory analysis are shown in Figs 5 and 6. Fig. 5 (upper panel) shows the number of cylinders in configuration as a function of cycles. We see that initially the number of 2-connected cylinders increases but it remains roughly constant after a certain time. The number of 0- and 1-connected cylinders decreases: this decrease is expected since simulated annealing penalizes these cylinders more over time. Eventually, these numbers also approach a constant value. The three lower panels in Fig. 5 show the cumulated means for the final part of the simulations: superposition of 50 simulations are shown. From this figure, we see that the number of 0-, 1-, 2-connected cylinders tend to have similar statistical values in every simulation. The cumulative standard deviations are shown in Fig. 6. The box-plots of the mean and standard deviation distributions obtained from the final realization of 50 independent simulations are shown in Fig. 7. The standard deviation is much larger than the variation in mean numbers of cylinders, showing that all the 50 simulations are statistically equivalent.

These numerical results are coherent with the detection obtained in Fig. 4. The robust part of the network is given by the 2-connected
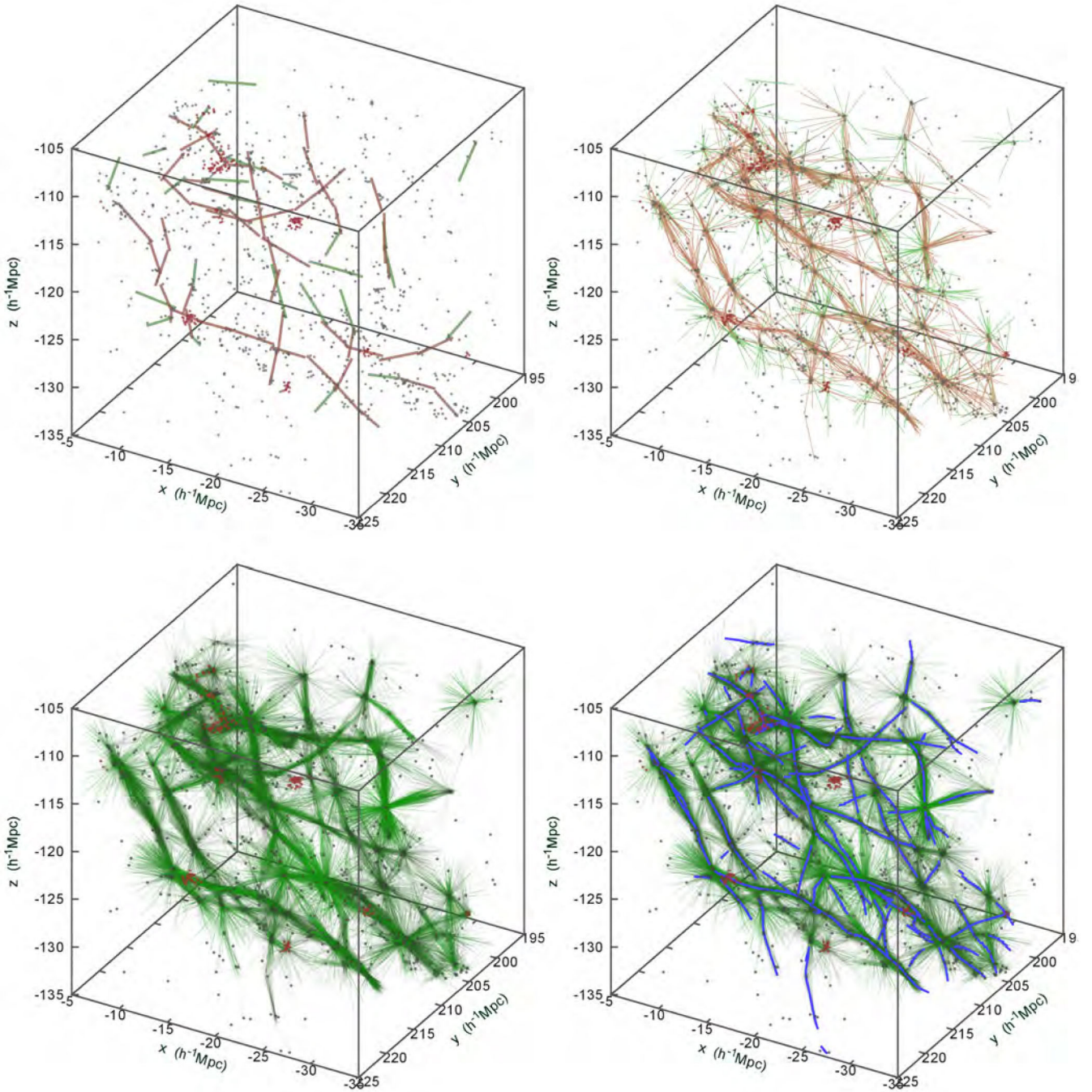
Figure 4. Detected filamentary pattern (cylinder axes) in a small sample volume within a pattern of galaxies (points). Upper-left panel: single MCMC simulation detecting the filamentary pattern; upper-right panel: the superposition of 25 independent simulations (for visual clarity, we show only half of the simulations). Cylinders are colour-coded as following: 2-connected (red), 1-connected (green) and isolated (grey). Galaxies in groups with 10 or more members are shown with red points; other galaxies are shown with grey points. *Lower* panels show the cylinders from 1000 realizations (it corresponds to the visit map) used to extract the filament spines; in lower-right panel, the extracted filament spines are also shown with blue lines. The movie, showing the MCMC in action is available at http://www.aai.ee/~elmo/sdss-filaments/.

cylinders. The part of the network made of 0- and 1-connected cylinders may be considered at a quick look as 'noisy'. Still, the question what part of this 'noisy' part is relevant for the filamentary network, is of real interest. Our manner of answering it was to leave the model parameters rather free, since we do not know exactly how the objects we are looking for look like. Averaging several simulation results and spine detection should eliminate the 'noisy' part while keeping the important short filaments.

There is another point to be outlined. Our plots show that simulated annealing does not reach convergence yet (in theory, it converges at infinity). This is due to the real computational time needed for getting the results: with modern computers, one simulation takes approximately 1000 CPU-hours in a single CPU. There was a compromise to be done here: choosing an appropriate cooling schedule and stopping the algorithm after a while, or choosing a fast cooling schedule and stating that the algorithm 'converged'. We have chosen
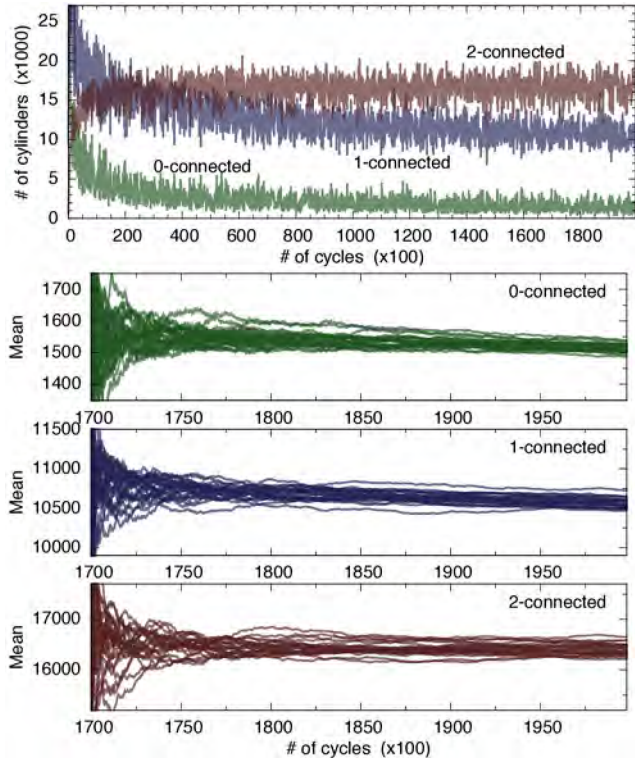
Figure 5. The time series of the numbers of cylinders (0-, 1-, 2-connected) in a configuration (upper panel). Cumulated means for 0-, 1-, 2-connected cylinders (lower three panels, respectively) computed for the final part of the simulation: superposition of 50 independent simulations.
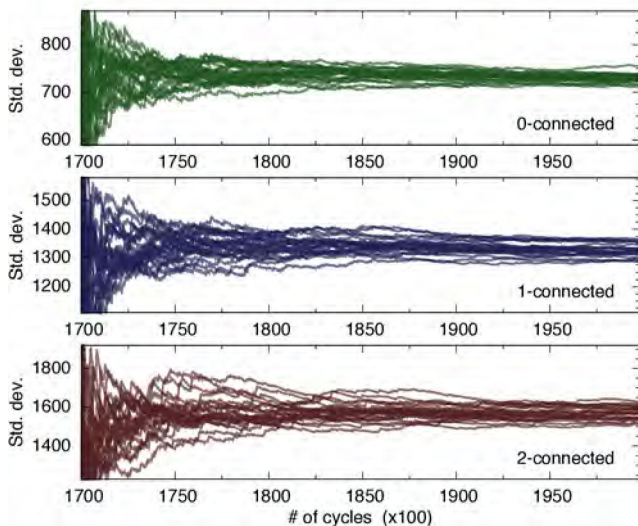
Figure 6. Cumulated standard deviations for 0-, 1-, 2-connected cylinders computed for the final part of the simulation: superposition of the 50 independent simulations.

the first approach. The convergence of the MCMC simulation methods is still an open research problem. For the reader interested to get a deeper insight of this very interesting problem we recommend as a starting point Robert & Casella (2004).

The main advantage of using such a stochastic approach consists in the ability of the model to give a simultaneous morphological and statistical characterization of the pattern. This allows the com-
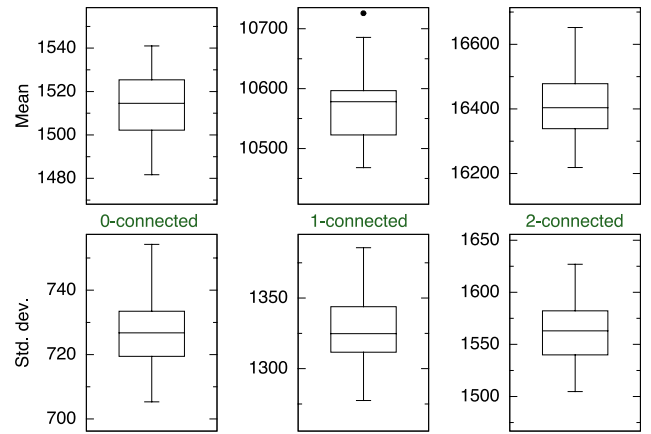
Figure 7. Box-plots of the estimated mean (upper row) and standard deviation (lower row) distributions for 0-, 1-, 2-connected cylinders. The values are computed from the final configurations of 50 independent simulations.

parison of several filamentary networks and this idea was used in a previous study to compare mock catalogues and observed data (Stoica et al. 2010). Nevertheless, it is legitimate to wish to have a smooth map of these filaments. A solution to this problem is, under the hypothesis of the model, to estimate the probability that a point belongs to the filamentary network and to look at those regions where these values are higher than a given threshold (Stoica et al. 2007a,b, 2010). Several realizations should be used to estimate this quantity and the filamentary network is smoothed. In our previous work, these quantities are called visit maps, and this name is kept in the following, due to its suggestivity. In mathematics, these quantities are known as level sets and the convergence of these type of estimators was studied in Heinrich, Stoica & Tran (2012). The visit map estimated from a number of 1000 realization is shown in Fig. 4 (lower-left panel).

Another very important aim of our work is to link the richness of our approach with the very efficient existing deterministic methods for filaments finding. Therefore, in the following a method for filamentary pattern spine detection is proposed. Fig. 4 (lower-right panel) shows the result of the introduced method. The main difference with the existing methods is that the spine detector we build uses the information provided by our stochastic approach. This information consists of different quantities that can be estimated locally using our model. These quantities estimate different probability and visit maps and also statistics related to the orientation field induced by the filamentary network.

The spine detection we propose is based on two main ideas. The first idea is that filament spines are situated at the highest density regions outlined by the filament probability maps. Next, in these regions of high probability for the filamentary network, the spines have an orientation that is aligned with the direction given by the orientation field of the filamentary network. On the contrary, in cross-sections of filaments, the filament detection probability can be high, but the orientation field is not clearly defined. The filament spine detection method is described below in detail.

First, we recall the visit map estimator $\mathcal{L}(\boldsymbol{k})$ for a given point $\boldsymbol{k} = (x, y, z)$

$$\mathcal{L}(\boldsymbol{k}) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}\{\boldsymbol{k} \in \boldsymbol{Y}_i\}, \tag{28}$$

where $Y_1, Y_2, \ldots, Y_N$ are N cylinder configurations and $\mathbb{1}\{k \in Y_i\}$ is the indicator function testing whether the point k belongs to any of the cylinders in the configuration Y_i. By this definition, the visit map is defined to be in the range $\mathcal{L}(k) \in [0, 1]$.

The density map $\mathcal{D}(k)$ of filaments is defined as a weighted visit map (level set). For a given point $k = (x, y, z)$ it is defined as

$$\mathcal{D}(k) = \frac{1}{N} \sum_{i=1}^{N} \frac{\sum_{y \in Y_i} \exp[v(y)] \mathbb{1}\{k \in y\}}{\sum_{y \in Y_i} \mathbb{1}\{k \in y\}}, \qquad (29)$$

where the first summation is over realizations and the second summation is over cylinders in that given configuration Y_i. The potential function $v(y)$ for a cylinder y is defined by equation (15). The indicator function $\mathbb{1}\{k \in y\}$ acts as a cylindrical kernel and selects the points that the cylinder y covers. Note that a point k can be covered by a several cylinders in one configuration, but effectively all configuration have equal weights. We weight the visit map to suppress weak intersecting filaments (to reduce the noise) and to encourage stronger filaments.

The orientation field $\mathcal{G}(k, \omega)$ for a point k and for an orientation $\omega = \phi(\eta, \tau)$ is defined as

$$\mathcal{G}(k, \omega) = \frac{\sum_{i=1}^{N} \sum_{y \in Y_i} \exp[v(y)] \mathbb{1}\{k \in y\} |\omega \cdot \omega_y|}{\sum_{i=1}^{N} \sum_{y \in Y_i} \exp[v(y)] \mathbb{1}\{k \in y\}}, \qquad (30)$$

where $\omega \cdot \omega_y$ denotes the scalar product between the orientation vector ω and the cylinder orientation ω_y. Using this definition, $\mathcal{G}(k, \omega) \in [0, 1]$. Using the orientation field, we also define density field for orientation strengths: the maximum of the orientation field depending on ω at a given location k is $\mathcal{D}_{\mathcal{G}}(k)$,

$$\mathcal{D}_{\mathcal{G}}(k) = \max\{\mathcal{G}(k, \omega)\}. \qquad (31)$$

This quantity is a weighted estimator of the expectation of the scalar product between the orientation ω at the location k. If the cylinder orientation ω_y is uniform on the unit sphere, then the absolute values of the scalar product is a uniform random variable between 0 and 1. Hence the value of $\mathcal{D}_{\mathcal{G}}$ under the uniform assumption should be close to 0.5. If all the cylinders are aligned with respect to ω then the value of $\mathcal{D}_{\mathcal{G}}$ should be close to 1. If we are interested in the situation that the majority of cylinders are aligned to ω, then we may test $\mathcal{D}_{\mathcal{G}} > \mathcal{D}_{\mathcal{G}\text{lim}}$ with $\mathcal{D}_{\mathcal{G}\text{lim}}$ a pre-defined threshold value rather close to 1. For our purposes, we set $\mathcal{D}_{\mathcal{G}\text{lim}} = 0.75$, since this value may suggest a half-way distance between the uniform and the completely aligned case.

The corresponding orientation of the maximum value $\mathcal{D}_{\mathcal{G}}(k)$ at location k is $\omega_{\mathcal{G}}(k)$ and it is defined as

$$\omega_{\mathcal{G}}(k) = \arg\max_{\omega}\{\mathcal{G}(k, \omega)\}. \qquad (32)$$

For computing the previous estimators, the last 32 extracted realization from a single run of the algorithm were kept: we extracted the realizations after 1000 cycles. In total we have used 50 independent runs of the algorithm. This gives in all 1600 cylinder configurations to be used for computing the previous defined quantities.

At a first look, the previous quantities can be computed locally, hence there is no need for keeping track of the cylinder configurations. In our case, we need to calculate the visit map (and orientation map with orientations) in a grid with grid-step smaller than $0.1\,h^{-1}$ Mpc to accurately determine the spine of the filaments. Due to the limitations of the computer memory, it cannot be computed globally for the entire simulation box. In order to calculate the visit

map and orientation map with sufficiently fine grid, they have to be computed locally. For that purposes, we store the cylinder configurations from every simulations and compute the visit maps and orientation maps locally as defined above. The advantage of this approach is that all required quantities can be computed for every space and orientation in the sample volume and we are free of gridding. The chosen 1600 cylinder configuration is large enough for visit map and orientation map estimation and at the same time it requires reasonable amount of computational resources.

Using previous definitions, for every point k we have three values: the filament density $\mathcal{D}(k)$, the orientation strength $\mathcal{D}_{\mathcal{G}}(k)$ and the filament orientation in that location $\omega_{\mathcal{G}}(k)$. To extract a single filament using these three quantities, we do the following.

(i) We start at a point of the highest density $\mathcal{D}(k)$ that is not yet masked out (we will discuss masking later). We designate this point as k_0. The initial density map is calculated on a $0.5\,h^{-1}$ Mpc grid, which is sufficiently fine (compared with cylinder size) for global maxima. After maximum is found, the density map is calculated locally on a $0.01\,h^{-1}$ Mpc grid. Initially, all the regions where $\mathcal{L}(k) < \mathcal{L}_{\text{lim}} = 0.05$ are masked out: e.g. we are searching for the filaments in the regions that have been covered at least in 5 per cent of the realizations. We remind that all the detected structures are filaments by definition, and this is only the detection probability that depends on the model parameters.

(ii) If the orientation at that point is defined, we start extracting a filament. We say that the orientation is defined if $\mathcal{D}_{\mathcal{G}}(k_0) > \mathcal{D}_{\mathcal{G}\text{lim}} = 0.75$. Otherwise, we mask out the region around this point and continue with the step (i). The size of the masked region is taken as $1.0\,h^{-1}$ Mpc.

(iii) We look from the point k_0 to both sides along $\omega_{\mathcal{G}}(k_0)$.

(iv) To extract the filament, we move from the point k_0 in the direction $\omega_{\mathcal{G}}(k_0)$ by $\delta x = 0.5\,h^{-1}$ Mpc. The step size is arbitrary, but a smaller step size gives smoother filaments. The step size $\delta x = r$ is good enough (r is the cylinder radius). We designate the new point as k_i.

(v) We calculate the density map $\mathcal{D}_{\omega_{\mathcal{G}}}(k_i)$ that is perpendicular to the direction $\omega_{\mathcal{G}}(k_i)$. The density map $\mathcal{D}_{\omega_{\mathcal{G}}}(k_i)$ is two-dimensional and from that map we find the maximum that is closest to the point k_i: the location of this maximum is marked as $k_{i'}$.

(vi) If the orientation is not defined at $k_{i'}$, we stop the filament extracting algorithm and continue with the step (x).

(vii) If the orientation is defined ($\mathcal{D}_{\mathcal{G}}(k_{i'}) > \mathcal{D}_{\mathcal{G}\text{lim}}$), we go forward by δx and find a new point as previously explained. This point is needed to perform two additional checks.

(viii) First, to avoid breaks in the filament, we calculate the curvature of the filament at the point $k_{i'}$ using this point and its neighbours. The curvature $\kappa = 1/R$, where R is the radius of the sphere that these three points touch. We use the limiting value $\kappa > \kappa_{\text{lim}} = 1.0$ to stop the filament finding algorithm.

(ix) Secondly, we require that the orientation at the point $k_{i'}$ and at the neighbouring points is roughly the same:

$$\max |\omega_{\mathcal{G}}(k_{i'}) \cdot \omega_{\mathcal{G}}(k_{i\pm1})| > \tau_{\text{lim}} = 0.95. \qquad (33)$$

If the tests are not satisfied, we stop the filament finding algorithm. Otherwise, we move in the direction $\omega_{\mathcal{G}}(k_{i'})$ by δx and continue with the step (v).

(x) If all the filament points from both sides of k_0 have been found, we mask out the region that this filament covers. The radius of the masked out region is taken $1.0\,h^{-1}$ Mpc (twice the filament radius). We save the extracted points as a single filament.

(xi) We return to the point (i) until all the volume is masked out.

Basically, this algorithm walks along the mountain chain in the filament density map and tests if the orientation is defined and the orientation is the same as the walking direction. There are only four parameters that define the filaments: $\mathcal{L}_{\text{lim}} = 0.05$ defines the limiting visit map density and the strength of a filament, $\mathcal{D}_{\mathcal{G}\,\text{lim}} = 0.75$ defines the orientation and estimates the strength of orientation for a filament, $\kappa_{\text{lim}} = 1.0$ defines the limiting curvature and $\tau_{\text{lim}} = 0.95$ defines the angle between the filament and the orientation field. All these criteria are unimportant for strong filaments, but they influence the regions where filaments intersect or the regions where filaments are poorly defined.

5 RESULTS AND DISCUSSION

Fig. 8 illustrates the detected filaments and their axes. In left-hand panel are shown the luminosity density field smoothed with $1\ h^{-1}$ Mpc B_3-spline kernel. Right-hand panel shows the corresponding visit map ($\mathcal{L}$) for detected filaments. Extracted filament spines are shown with lines. Qualitatively, the filament axes plotted in these figures appear to closely trace the underlying large-scale filaments. This is not surprising because, by definition, our filament finder is based on cylindrical shapes and the filament detection probability field should trace the filamentary structures. Tempel et al. (2014) used the Bisous process to detect the filaments from dark matter simulations and showed that the detected filaments are very

well aligned with the underlying velocity field. This shows that the detected filaments are also dynamical structures.
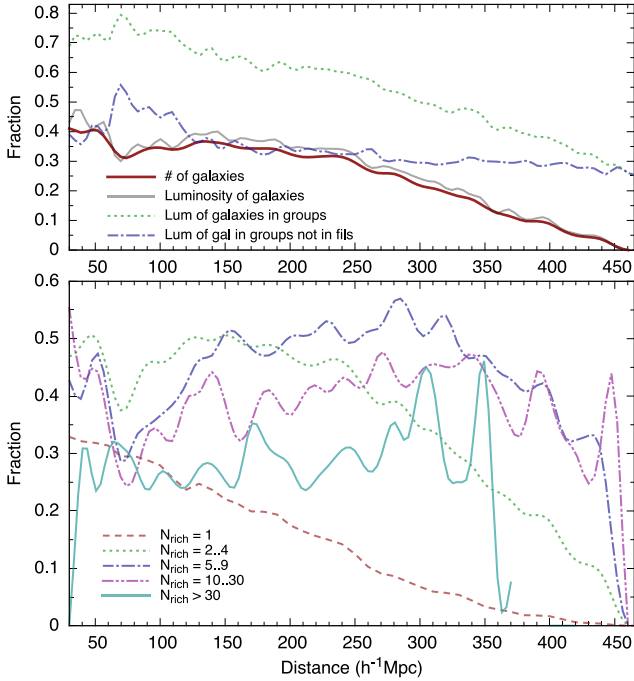
The upper panel in Fig. 9 shows the fraction of galaxies in filaments as a function of distance: the red solid line shows the number of galaxies and the grey dashed line shows the fraction of luminosity in filaments. We note that there is slightly more luminosity in filaments than the number fraction predicts – meaning that in filaments the luminosity density of galaxies is higher than in average. The green dotted line in the upper panel of Fig. 9 shows the fraction of luminosity that is in groups (we use the groups as defined in Tempel et al. 2012). Since we use a flux-limited sample, the number density of groups (and their luminosity) is higher for nearby regions. The blue dot–dashed line shows the fraction of luminosity in groups that are not in filaments. We see that this fraction is almost constant. The lower panel in Fig. 9 gives an explanation for that. Most of the galaxies that are in filaments are also in small groups and since filaments have a chain-like inner structure, the nearby filaments are made of smaller groups that are aligned. Further away, the number density of smaller groups is lower and the filament detection probability is also lower (the red solid line in the upper panel). The lower panel in Fig. 9 also shows that isolated galaxies are preferentially not located in filaments and also the galaxies in large clusters are mostly not in filaments: only galaxies in the outskirts of large clusters are in filaments.

All this implies that filaments are far from being smooth uniform structures. Visual inspection of the density field and the spines of
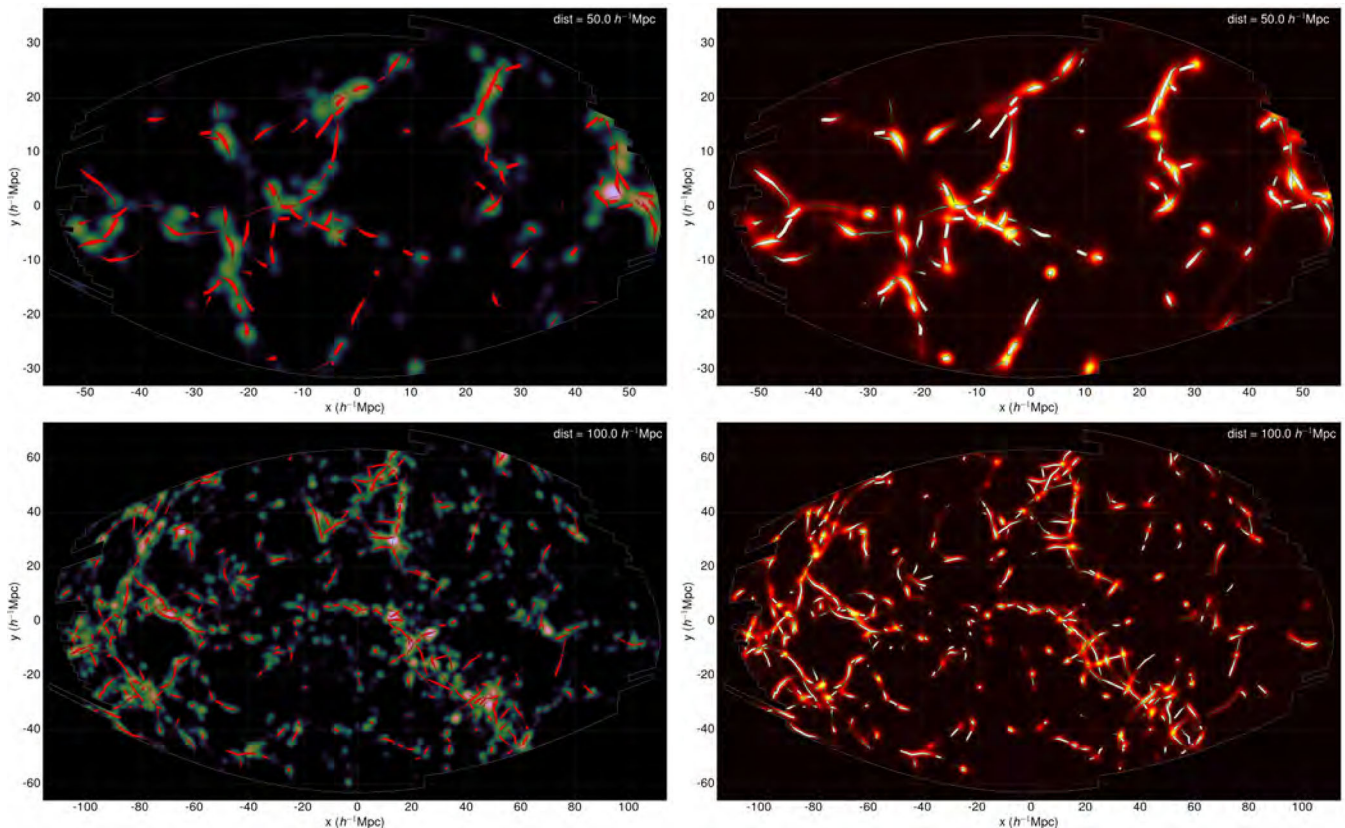
Figure 8. Sky projections of luminosity density field (left-hand panels) and the visit map (right-hand panels) at distances 50 h^{-1} Mpc (top row) and 100 h^{-1} Mpc (bottom row). Luminosity density field is smoothed with $1\ h^{-1}$ Mpc B_3 spline kernel. For better visualization, both images are created by summing up projected densities on several planes within range of $-4\ldots+4\ h^{-1}$ Mpc from the indicated distance (using $1\ h^{-1}$ Mpc step) and presented in logarithmic scale. Extracted filaments in the same distance interval are drawn with red and white lines, the width of line denotes the distance between filament and the plane of the image. There is good correspondence between the structures in the luminosity density field and detected filaments. The fly-through movie, showing the full observed volume is available at http://www.aai.ee/~elmo/sdss-filaments/.

Figure 9. The upper panel shows the fraction of galaxies in filaments (red solid line) and the fraction of observed luminosity in filaments (grey dashed line) as a function of distance. The green dotted line shows the fraction of luminosity in groups and the blue dot–dashed line shows the fraction of luminosity in groups that are not in filaments. The filament radius is taken to be $0.5 \, h^{-1}$ Mpc. The lower panel shows the fraction of galaxies in filaments for different group richness bins.

filaments (Fig. 8) show that filaments are populated by small galaxy groups and large clusters are in intersection of those filaments, as already shown in Bond et al. (1996). The same impression is quantitatively confirmed in Fig. 9.

The fraction of galaxies (or luminosity) in filaments gives us roughly the mass filling fraction of filaments. Up to distance $250 \, h^{-1}$ Mpc, the fraction is 35–40 per cent, which is in very good agreement with *N*-body simulations (Forero-Romero et al. 2009; Aragón-Calvo et al. 2010a; Hoffman et al. 2012). Using the SDSS data, the same filament mass filling fraction has been measured by Jasche et al. (2010). We note that after $250 \, h^{-1}$ Mpc, the number of detected structures decreases. Smith et al. (2012) search structures in the SDSS in the same scale as we, and their number of detected structures follows the same behaviour with distance. This is logically expected, since the number density of objects (that decreases with distance in flux-limited survey) is strictly related to the number of detected structures.

Although filaments have been studied extensively in general, there are only few studies addressing their radial density profile (Colberg, Krughoff & Connolly 2005; Dolag et al. 2006; Aragón-Calvo et al. 2010a). However, in these papers, larger filaments are considered, so a strict comparison is not possible. Fig. 10 shows the filament radial profile in the current study: the upper panel shows the distribution of the number of galaxies per radius, the lower panel shows the number density profile for filaments. We see that most of the galaxies in filaments are closer than $0.5 \, h^{-1}$ Mpc to the filament axis. This is because our defined filament radius is $0.5 \, h^{-1}$ Mpc. We also see that there is a break in the density profile around $r = 0.2 \, h^{-1}$ Mpc. Using weak lensing, Jauzac et al. (2012) studied a single filament that connects two clusters and they also
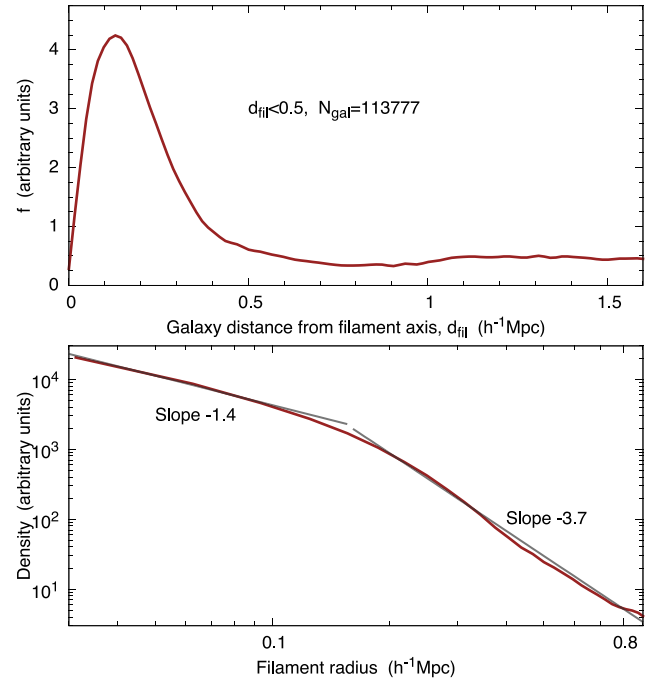
Figure 10. Upper panel: the distribution of galaxy distances from the nearest filament axis. After $0.5 \, h^{-1}$ Mpc, the distribution is dropping since our defined filament radius is $0.5 \, h^{-1}$ Mpc. The number of galaxies closer than $0.5 \, h^{-1}$ Mpc to the filament axis is shown in the figure. Lower panel: the number density of galaxies as a function of the distance from the axis.

see the break in the density profile roughly at the same distance. Direct comparison with other studies is possible, once we extract thicker filaments: this is planned for future work.

Another interesting quantity is the filament length distribution. Fig. 11 shows the filament length distribution in the linear
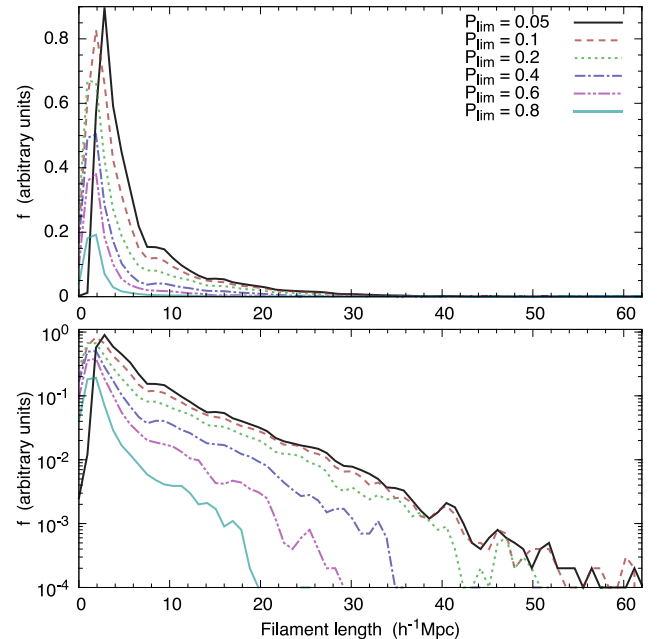
Figure 11. The filament volume distribution as a function of the filament length. The distribution is shown for various limiting covering probability values ($\mathcal{L}_{\mathrm{lim}}$). The lower panel emphasizes the differences for long filaments.
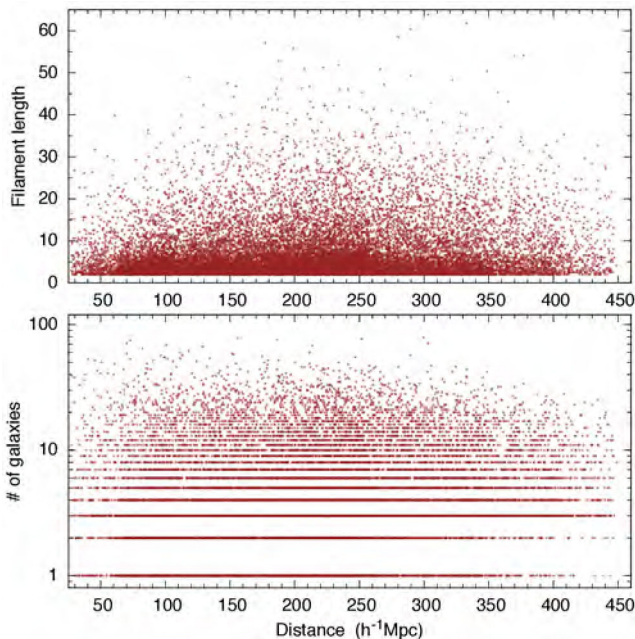
(upper panel) and logarithmic (lower panel) scale. The distribution is shown for different minimum detection probability values ($\mathcal{L}_{lim}$). The black solid line shows the filaments as given in our catalogue. Increasing the limiting detection probability, the filaments start to fragment and small (weak) filaments disappear. We see that increasing the detection probability up to 0.2 practically does not change the long-end of the distribution. The longest filaments are strong and dominant filaments. We emphasize that the value of the detection probability is model dependent and it does not reflect the probability of the filamentary structure (all detected structures in our model are filaments based on the definition).

The filament length distribution has been also studied in Bond et al. (2010b) using *N*-body simulations. Compared with our length distribution, the longest filaments are roughly the same, but we have more short filaments. This is probably because the spatial distribution of observed galaxies is sparse and in many cases we only found a piece of a filament.

The maximum length of our filaments is $\sim 60\,h^{-1}$ Mpc. This is in very good agreement with other measured values (Bharadwaj, Bhavsar & Sheth 2004; Bond et al. 2010b; Pandey et al. 2011).

Fig. 12 shows the filament lengths and the number of galaxies in filaments as a function of distance. We note that both distributions are quite uniform. There is lack of long filaments in the nearby region because its volume is small. Further away, the longest filaments are missing because the number density of galaxies is too low. However, there exist filaments with lengths up to $30\,h^{-1}$ Mpc further than $400\,h^{-1}$ Mpc. The number of galaxies in filaments (lower panel) decreases with distance because we used a flux-limited sample: the faintest galaxies are missing further away.

Fig. 13 shows the luminosity of a filament (upper panel) and the luminosity per unit length (lower panel) as a function of distance. We notice that the faintest filaments are missing further away, because of the flux-limited survey. However, the upper limit is distance independent and it shows that the brightest filaments nearby and further away are practically the same. We also note that the scatter

Figure 13. The upper panel shows the filament luminosity as a function of distance. The lower panel shows the change of filament luminosity per unit length with distance. The distribution for the luminosity per unit length is much more tight. Red dots are for all filaments, blue dots are for long filaments (at least $10\,h^{-1}$ Mpc long).

in the lower panel is quite small, indicating that luminosity per unit length in filaments does not vary much. Blue points in the figure show the longest filaments (at least $10\,h^{-1}$ Mpc long). We see that the longest filaments are also the most luminous filaments, however, their luminosity per unit length lies within the average. This indicates that short and long filaments have on average the same luminosity density.

One important quantity that describes filaments is their volume filling factor. Fig. 14 shows the volume filling factor as a function of distance. Since the number density of galaxies decreases with distance, the volume filling factor also decreases. The filament
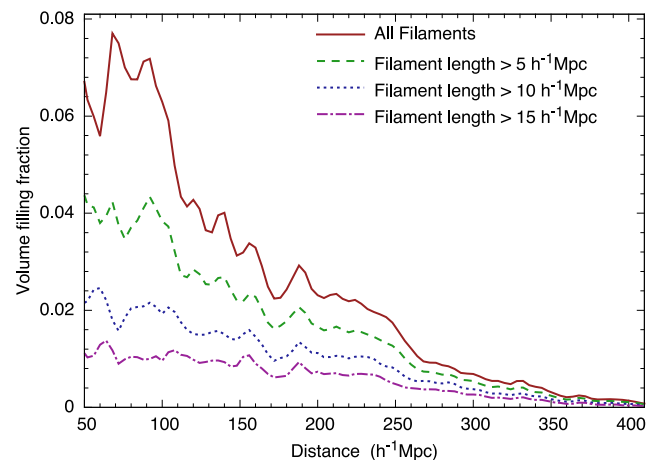

Figure 12. The upper panel shows the filament length as a function of distance from the observer. The lower panel shows the number of galaxies in a filament as a function of distance.

Figure 14. The volume filling fraction of filaments as a function of distance. Red solid line shows the volume filling fraction for all detected filaments. Other lines show the volume filling fraction, when only longer filaments are considered. The filament volume is calculated using the spines of the filaments and the filament radius is taken to be $r = 1\,h^{-1}$ Mpc.

volume is calculated around the detected spines, using the filament radius $r = 1.0\,h^{-1}$ Mpc; in the nearby region, the filling factor is $\sim$7 per cent and it decreases with distance (due to the flux-limited survey). From N-body simulations, the volume filling factor has been measured by Forero-Romero et al. (2009) and Aragón-Calvo et al. (2010a). In these papers, the volume filling factor depends on the used threshold, but it is of the order of 10 per cent, which is in good agreement with our results. Based on the SDSS data, the filaments have been extracted by Jasche et al. (2010). They use a novel Bayesian sampling algorithm, which permits precise recovery of poorly sampled objects in a non-linear density field. Based on their analysis, the filament volume filling factor is 10–20 per cent, which is slightly larger than ours. However, since they use a $3\,h^{-1}$ Mpc grid their filaments are thicker than ours and these results cannot be directly compared. On the contrary, Hoffman et al. (2012) showed that the filament volume filling fraction is 4–5 per cent using the velocity shear tensor. This is slightly lower than we found in the current study (considering the fact that we do not detect all structures). Since the mass filling fractions in both studies are comparable, the filaments found in the velocity field are probably located mostly in higher density environments and are somewhat thicker (Tempel et al. 2014).

5.1 Robustness of the filamentary pattern detection

Our method is sensitive to the galaxy density. If a data set contains too few galaxies, no filaments will be detected since the filamentary pattern is not observable. If the data set contains many galaxies, the filamentary network is already defined, hence our method will always work. Indeed, intuitively, there is an optimal range for the galaxy number density so that our model delineates correctly the filaments.

In general, the model parameters depend on the minimum number density. The model parameters of our 'machinery' were designed studying the SDSS data set. After several trials and errors, we found the parameter values that give the best results (see Table 1). To reliably determine filaments, two (preferentially three or more) cylinders have to be aligned and connected. With the present model parameters, this leads to the minimum number density (inside a filament) of six galaxies within a cylindrical volume of the radius 0.5 and the length 6–10 h^{-1} Mpc.

As the data energy in our model is determined by the ratio of densities in the cylinder and its shadow, it does not depend on the local number density of galaxies (for fixed model parameters, and if the minimum number density condition is satisfied). This allows us to detect physically similar filaments regardless of the environmental density, and our method can recover structures of relatively sparsely sampled objects (filaments in lower density environments).

In a flux-limited survey, the number density of galaxies decreases with distance. For our model, it means that the filament detection probability decreases (we do not detect all filaments further away and/or we only detect parts of the filamentary network), but the reliability of the detected filaments, determined by the visit map value, is largely unaffected, due to the robustness of the model.

As an example, Fig. 14 shows the volume filling fraction of filaments as a function of distance. Including all filaments (red line), the sample is not homogeneous. However, it is possible to construct a statistically homogeneous sub-sample of filaments, when the sample is limited by distance and filament length. Fig. 14 shows that when using only longer filaments, the volume filling fraction is roughly constant with distance up to 240 h^{-1} Mpc. Further away, the galaxy number density decreases rapidly, and the full filamen-

tary network for the scale used here ($r = 0.5\,h^{-1}$ Mpc) is not clearly outlined.

Another important point to be raised is the question whether the optimal choice of the model parameters should depend on the galaxy number density. Stoica et al. (2010) addressed this issue and concluded that simple choices (e.g. increasing the cylinder size) do not produce good results: different cylinder sizes detect different structures. To reduce the incompleteness in filament detection, larger cylinders should be used everywhere to detect the same structures.

How can shot noise affect the filament detection in low number density regions? Like for any other methods, shot noise affects the results. The main advantage of our method is in its probabilistic nature. Individual realizations of the solution may be sensitive to noise. Still, averaging these realizations reduces the noise influence, and allows computation of robust statistical quantities (Stoica et al. 2010). These quantities are the estimates of the sufficient statistics, the level sets (visit maps) and the local detection probability.

It is also possible to think about the reverse formulation of the detection problem. That is, knowing the topological structure of the filamentary pattern, we may wonder what is the density range within the observed volume, that still outlines the given filamentary network. However, the optimal model choice for filament detection, is at the moment an open mathematical and data analysis problem.

6 CONCLUSIONS AND FUTURE WORK

This paper uses and develops an object point process with interactions (the Bisous process) to trace the filamentary network in the flux-limited SDSS data. This method works directly on the galaxy distribution and does not require any additional smoothing, it only requires fixing the scale of structures. For the current work, we fixed the radius of a filament as $r = 0.5\,h^{-1}$ Mpc, which is close to the scale of galaxy groups/clusters; such filaments should have the largest impact for galaxy formation and evolution.

Our filament finder is probabilistic in the sense that it gives us the filament detection probability field together with the filament orientation field. Using these two fields, we define the spines of the filaments and extract single filaments from the data. We showed that the detected filaments fit well with the visible large-scale structure. The composed catalogue of filaments for the SDSS is made publicly available (see Appendix A).

We showed that the filament mass and volume filling factors are in good agreement with structures found in N-body simulations and in previous observational studies. The mass filling fraction of our filaments is 35–40 per cent, and the volume filling fraction is $\sim$8 per cent and decreases with distance due to the flux-limited data. Consequently, filaments contain the largest fraction of mass in the Universe and they represent the most salient component of the cosmic web: they form the bridges between all structural features at the group/cluster scale.

Our catalogue of filaments is not the first attempt to extract filaments from SDSS data. Filaments from SDSS have been extracted by Sousbie et al. (2008), Jasche et al. (2010) and Smith et al. (2012). In the following studies, we plan to compare how various filament finders work and how the filaments detected using different methods differ.

In our method, we have to define the filament scale (radius). In the current study, it is fixed at 0.5 h^{-1} Mpc. This scale was chosen to find the bridges between galaxy groups and because it was known that this scale affects the galaxy evolution (e.g. Tempel & Libeskind 2013; Tempel et al. 2013). Since filaments are hierarchical by

nature (Aragón-Calvo et al. 2010a; Smith et al. 2012) it is interesting to search for filaments at many scales. The number density of galaxies in the SDSS does not allow us to search for smaller filaments. We are preparing our filament finder to search for thicker filaments, and the catalogue and comparison of multiscale filaments will be our next step. The multiscale filaments will allow us to better determine the filament scale that affects galaxy evolution.

Fig. 9 shows that galaxies in large clusters are not in filaments. This is expected since filaments are the bridges between clusters and large clusters are in intersection of many filaments (Aragón-Calvo et al. 2010a). In our following work, we plan to study how filaments and groups/clusters are connected and how this connection depends on cluster/filament properties.

The Bisous process can also be applied to other structure elements, as clusters, sheets and voids. This is also one of the future directions of our work.

REFERENCES

Aihara H. et al., 2011, ApJS, 193, 29
Alpaslan M. et al., 2013, MNRAS, preprint (arXiv:1311.1211)
Aragón-Calvo M. A., 2013, preprint (arXiv:1303.1590)
Aragón-Calvo M. A., Szalay A. S., 2013, MNRAS, 428, 3409
Aragón-Calvo M. A., van de Weygaert R., Jones B. J. T., van der Hulst J. M., 2007a, ApJ, 655, L5
Aragón-Calvo M. A., Jones B. J. T., van de Weygaert R., van der Hulst J. M., 2007b, A&A, 474, 315
Aragón-Calvo M. A., van de Weygaert R., Jones B. J. T., 2010a, MNRAS, 408, 2163
Aragón-Calvo M. A., Platen E., van de Weygaert R., Szalay A. S., 2010b, ApJ, 723, 364
Beygu B., Kreckel K., van de Weygaert R., van der Hulst J. M., van Gorkom J. H., 2013, AJ, 145, 120
Bharadwaj S., Bhavsar S. P., Sheth J. V., 2004, ApJ, 606, 25
Blanton M. R., Roweis S., 2007, AJ, 133, 734
Blanton M. R., Eisenstein D., Hogg D. W., Schlegel D. J., Brinkmann J., 2005, ApJ, 629, 143
Bond J. R., Kofman L., Pogosyan D., 1996, Nature, 380, 603
Bond N. A., Strauss M. A., Cen R., 2010a, MNRAS, 406, 1609
Bond N. A., Strauss M. A., Cen R., 2010b, MNRAS, 409, 156
Brunino R., Trujillo I., Pearce F. R., Thomas P. A., 2007, MNRAS, 375, 184
Cautun M., van de Weygaert R., Jones B. J. T., 2013, MNRAS, 429, 1286
Cen R., Ostriker J. P., 1999, ApJ, 514, 1
Cervantes-Sodi B., Hernandez X., Park C., 2010, MNRAS, 402, 1807
Codis S., Pichon C., Devriendt J., Slyz A., Pogosyan D., Dubois Y., Sousbie T., 2012, MNRAS, 427, 3320
Colberg J. M., Krughoff K. S., Connolly A. J., 2005, MNRAS, 359, 272
Davis M., Peebles P. J. E., 1983, ApJ, 267, 465
Dietrich J. P., Werner N., Clowe D., Finoguenov A., Kitching T., Miller L., Simionescu A., 2012, Nature, 487, 202
Dolag K., Meneghetti M., Moscardini L., Rasia E., Bonaldi A., 2006, MNRAS, 370, 656
Doroshkevich A. G., Gottlöber S., Madsen S., 1997, A&AS, 123, 495
Dressler A., 1980, ApJ, 236, 351
Einasto J., Saar E., Kaasik A., Chernin A. D., 1974, Nature, 252, 111
Forero-Romero J. E., Hoffman Y., Gottlöber S., Klypin A., Yepes G., 2009, MNRAS, 396, 1815
Fukugita M., Hogan C. J., Peebles P. J. E., 1998, ApJ, 503, 518
Geyer C. J., 1999, in Barndorff-Nielsen O., Kendall W. S., van Lieshout M. N. M., eds, Stochastic Geometry, Likelihood and Computation. CRC Press/Chapman and Hall, Boca Raton, p. 79
Geyer C. J., Møller J., 1994, Scan. J. Stat., 21, 359
González R. E., Padilla N. D., 2010, MNRAS, 407, 1449
Green P. J., 1995, Biometrika, 82, 711
Hahn O., Porciani C., Carollo C. M., Dekel A., 2007, MNRAS, 375, 489
Hahn O., Teyssier R., Carollo C. M., 2010, MNRAS, 405, 274
Heinrich P., Stoica R. S., Tran V. C., 2012, Spatial Stat., 2, 47
Heß S., Kitaura F.-S., Gottlöber S., 2013, MNRAS, 435, 2065
Hoffman Y., Metuki O., Yepes G., Gottlöber S., Forero-Romero J. E., Libeskind N. I., Knebe A., 2012, MNRAS, 425, 2049
Hunter J. D., 2007, Comput. Sci. Eng., 9, 90
Jackson J. C., 1972, MNRAS, 156, 1P
Jasche J., Kitaura F. S., Li C., Enßlin T. A., 2010, MNRAS, 409, 355
Jauzac M. et al., 2012, MNRAS, 426, 3369
Jõeveer M., Einasto J., 1978, in Longair M. S., Einasto J., eds, Proc. IAU Symp. 79, Large Scale Structures in the Universe. Reidel, Dordrecht, p. 241
Jõeveer M., Einasto J., Tago E., 1978, MNRAS, 185, 357
Jones B. J. T., van de Weygaert R., Aragón-Calvo M. A., 2010, MNRAS, 408, 897
Kaiser N., 1987, MNRAS, 227, 1
Kendall W. S., Møller J., 2000, Adv. Appl. Prob., 32, 844
Kirshnamoorty K., Thomson J., 2004, J. Stat. Plan. Inference, 119, 23
Kitaura F. S., Enßlin T. A., 2008, MNRAS, 389, 497
Komatsu E. et al., 2011, ApJS, 192, 18

Lee J., Erdoğdu P., 2007, ApJ, 671, 1248
Lee J., Lee B., 2008, ApJ, 688, 78
Libeskind N. I., Hoffman Y., Knebe A., Steinmetz M., Gottlöber S., Metuki O., Yepes G., 2012, MNRAS, 421, L137
Libeskind N. I., Hoffman Y., Forero-Romero J., Gottlöber S., Knebe A., Steinmetz M., Klypin A., 2013, MNRAS, 428, 2489
Liivamägi L. J., Tempel E., Saar E., 2012, A&A, 539, A80
Martínez V. J., Saar E., 2002, Statistics of the Galaxy Distribution. Chapman & Hall, Boca Raton
Martínez V. J., Arnalte-Mur P., Stoyan D., 2010, A&A, 513, A22
Murphy D. N. A., Eke V. R., Frenk C. S., 2011, MNRAS, 413, 2288
Møller J., Waagepetersen R. P., 2004, Statistical Inference and Simulation for Spatial Point Processes. Chapman and Hall/CRC, Boca Raton
Navarro J. F., Abadi M. G., Steinmetz M., 2004, ApJ, 613, L41
Novikov D., Colombi S., Doré O., 2006, MNRAS, 366, 1201
Pandey B., Kulkarni G., Bharadwaj S., Souradeep T., 2011, MNRAS, 411, 332
Pimbblet K. A., Drinkwater M. J., Hawkrigg M. C., 2004, MNRAS, 354, L61
Platen E., van de Weygaert R., Jones B. J. T., 2007, MNRAS, 380, 551
Przyborowski J., Wilenski H., 1940, Biometrika, 31, 313
Rieder S., van de Weygaert R., Cautun M., Beygu B., Portegies Zwart S., 2013, MNRAS, 435, 222
Robert C. P., Casella G., 2004, Stat. Sci., 19, 1
Shandarin S. F., 2011, J. Cosmol. Astropart. Phys., 5, 15
Shandarin S., Habib S., Heitmann K., 2012, Phys. Rev. D, 85, 083005
Shen J., Abel T., Mo H. J., Sheth R. K., 2006, ApJ, 645, 783
Smith A. G., Hopkins A. M., Hunstead R. W., Pimbblet K. A., 2012, MNRAS, 422, 25
Sousbie T., 2011, MNRAS, 414, 350
Sousbie T., Pichon C., Courtois H., Colombi S., Novikov D., 2008, ApJ, 672, L1
Sousbie T., Pichon C., Kawahara H., 2011, MNRAS, 414, 384
Stoica R. S., Gregori P., Mateu J., 2005a, Stoch. Process. Appl., 115, 1860
Stoica R. S., Martínez V. J., Mateu J., Saar E., 2005b, A&A, 434, 423
Stoica R. S., Gay E., Kretzschmar A., 2007a, Biometrical J., 49, 1
Stoica R. S., Martínez V. J., Saar E., 2007b, J. R. Stat. Soc. Ser. C, 56, 459
Stoica R. S., Martínez V. J., Saar E., 2010, A&A, 510, A38
Stoyan D., Kendall W., Mecke J., 1995, Stochastic Geometry and its Applications. Wiley, New York
Strauss M. A. et al., 2002, AJ, 124, 1810
Tago E., Einasto J., Saar E., Tempel E., Einasto M., Vennik J., Müller V., 2008, A&A, 479, 927
Tago E., Saar E., Tempel E., Einasto J., Einasto M., Nurmi P., Heinämäki P., 2010, A&A, 514, A102
Tegmark M. et al., 2004, Phys. Rev. D, 69, 103501
Tempel E., Libeskind N. I., 2013, ApJ, 775, L42
Tempel E., Saar E., Liivamägi L. J., Tamm A., Einasto J., Einasto M., Müller V., 2011, A&A, 529, A53
Tempel E., Tago E., Liivamägi L. J., 2012, A&A, 540, A106
Tempel E., Stoica R. S., Saar E., 2013, MNRAS, 428, 1827
Tempel E., Libeskind N. I., Hoffman Y., Liivamägi L. J., Tamm A., 2014, MNRAS, 437, L11
Trowland H. E., Lewis G. F., Bland-Hawthorn J., 2013, ApJ, 762, 72
Trujillo I., Carretero C., Patiri S. G., 2006, ApJ, 640, L111
Tully R. B., Fisher J. R., 1978, in Longair M. S., Einasto J., eds, Proc. IAU Symp. 79, Large Scale Structures in the Universe. Reidel, Dordrecht, p. 31
van de Weygaert R., Bond J. R., 2008a, in Plionis M., López-Cruz O., Hughes D., eds, Lecture Notes in Physics, Vol. 740, A Pan-Chromatic View of Clusters of Galaxies and the Large-Scale Structure. Springer, Heidelberg, p. 335
van de Weygaert R., Bond J. R., 2008b, in Plionis M., López-Cruz O., Hughes D., eds, Lecture Notes in Physics, Vol. 740, A Pan-Chromatic View of Clusters of Galaxies and the Large-Scale Structure. Springer, Heidelberg, p. 409
van Lieshout M. N. M., 2000, Markov Point Processes and Their Applications. Imperial College Press, London

van Lieshout M. N. M., Stoica R. S., 2003, Stat. Neerlandica, 57, 177
van Lieshout M. N. M., Stoica R. S., 2006, Comput. Stat. Data Anal., 51, 679
Viel M., Branchini E., Cen R., Ostriker J. P., Matarrese S., Mazzotta P., Tully B., 2005, MNRAS, 360, 1110
Wang H., Mo H. J., Yang X., van den Bosch F. C., 2012, MNRAS, 420, 1809
Wang H., Szalay A., Aragon-Calvo M. A., Neyrinck M. C., Eyink G. L., 2013, preprint (arXiv:1309.5305)
York D. G. et al., 2000, AJ, 120, 1579
Zhang Y., Yang X., Faltenbacher A., Springel V., Lin W., Wang H., 2009, ApJ, 706, 747
Zhang Y., Yang X., Wang H., Wang L., Mo H. J., van den Bosch F. C., 2013, ApJ, 779, 160

APPENDIX A: DESCRIPTION OF THE CATALOGUE

The catalogue of filaments consists of three tables. The first table lists the extracted filaments and the general properties (e.g., length) of the filaments. The second table gives all the filament points with their properties, every filament from the first table consists of a point set with a spacing of $\approx 0.5\,h^{-1}$ Mpc. The third table lists the galaxies that we used to generate the filaments, the galaxies are extracted from Tempel et al. (2012). The galaxy table lists the basic galaxy properties (for more properties, please see the table in Tempel et al. 2012) together with the info on the filament where the galaxy belongs to.

The catalogues are accessible at http://www.aai.ee/~elmo/sdss-filaments/ with a complete description in the README.TXT file. We give these catalogues as ascii files as well as a FITS table with three extensions, one for each table. We will also upload the catalogues to the Strasbourg Astronomical Data Center (CDS).

A1 Description of the filament catalogue

The filament catalogue (see Table A1) contains the following information (the column numbers are given in square brackets):

1. [1] `id` – unique identification number for a filament;
2. [2] `npts` – number of points in the filament with a spacing $\sim 0.5\,h^{-1}$ Mpc (the filament consists of these points);
3. [3] `len` – filament length in units of h^{-1} Mpc, measured along the filament from point to point;
4. [4–5] `ngal1`, `ngal2` – numbers of galaxies in the filament that are closer than 0.5, 1.0 h^{-1} Mpc to filament axis;
5. [6–7] `lum1`, `lum2` – luminosity of the filament: the sum of luminosities of observed galaxies that are closer than 0.5, 1.0 h^{-1} Mpc to filament axis (in units of $10^{10}\,h^{-2}\,L_{\odot}$);
6. [8–10] `xmin`, `ymin`, `zmin` – filament minimum coordinate in x-, y-, z-axis; the coordinates are defined by equation (1);
7. [11–13] `xlen`, `ylen`, `zlen` – filament range in x-, y-, z-axis.

A2 Description of the filament points table

The table of filament points (see Table A2) contains the following information (the column numbers are given in square brackets):

1. [1] `id` – filament identification number;
2. [2] `idpts` – unique identification number for a filament point, shared for all filaments;
3. [3] `npts` – number of filament points in the filament the point belongs to;

Table A1. A sample of filament table. The full table is available online. See Section A1 for detailed description.

Id	npts	len (Mpc h^{-1})	ngal1	ngal2	lum1 (10^{10} L$_\odot$ h^{-2})	lum2 (10^{10} L$_\odot$ h^{-2})	x_{min} (Mpc h^{-1})	y_{min} (Mpc h^{-1})	z_{min} (Mpc h^{-1})	x_{len} (Mpc h^{-1})	y_{len} (Mpc h^{-1})	z_{len} (Mpc h^{-1})
1	50	24.79	17	22	24.57	40.07	− 129.64	270.94	− 158.41	17.97	13.90	6.47
2	91	45.27	21	25	30.94	36.40	− 128.90	170.87	− 120.65	14.51	38.65	11.60
3	45	22.22	20	28	5.85	8.98	49.34	87.33	− 55.79	6.29	10.46	17.61
4	76	37.63	27	29	21.99	25.18	− 46.78	134.59	41.37	7.24	30.41	9.61
5	33	16.22	18	25	7.33	11.01	− 32.24	75.18	− 15.86	1.79	15.40	3.31

Table A2. A sample of filament points table. The full table is available online. See Section A2 for detailed description.

Id	idpts	npts	len (Mpc h^{-1})	x (Mpc h^{-1})	y (Mpc h^{-1})	z (Mpc h^{-1})	Dist (Mpc h^{-1})	d_x	d_y	d_z	v_{map}	f_{den}	f_{ori}
1	1	50	24.79	−111.66	270.94	−155.93	331.96	−0.5467	0.4396	−0.7126	0.434	0.095	0.855
1	2	50	24.79	−111.98	271.13	−156.31	332.40	−0.6738	0.5447	−0.4991	0.530	0.139	0.836
1	3	50	24.79	−112.33	271.36	−156.61	332.85	−0.7001	0.5836	−0.4113	0.586	0.176	0.788
1	4	50	24.79	−112.68	271.61	−156.87	333.29	−0.7015	0.5523	−0.4503	0.507	0.148	0.915
1	5	50	24.79	−113.04	271.85	−157.13	333.73	−0.7198	0.5147	−0.4657	0.451	0.131	0.974

Table A3. A sample of galaxy table. The full table is available online. See Section A3 for detailed description.

Id	nrich	redshift	RA (deg)	Dec. (deg)	distcor (Mpc h^{-1})	mag_r (mag)	lumr (10^{10} L$_\odot$ h^{-2})	w	edgedist (Mpc h^{-1})	fil_dist (Mpc h^{-1})	fil_id	fil_idpts
16	2	0.1044	251.16	28.22	308.57	16.73	1.61	1.81	10.79	2.59	6778	177 435
17	2	0.1062	251.17	28.13	309.10	17.40	0.91	1.81	11.05	2.54	6778	177 435
18	3	0.1324	251.34	28.46	387.90	17.68	1.13	2.55	11.37	0.12	218	10 252
19	3	0.1337	251.35	28.48	388.02	17.56	1.25	2.55	11.29	0.06	218	10 253
22	3	0.1331	251.33	28.49	387.96	17.66	1.11	2.55	11.32	0.07	218	10 253

4. [4] len – length of the filament (in units of h^{-1} Mpc) the point belongs to;

5. [5–7] *x, y, z* – the comoving coordinates (*x, y, z*) in units of h^{-1} Mpc as defined by equation (1);

6. [8] dist – distance to the filament point in units of h^{-1} Mpc;

7. [9–11] dx, dy, dz – orientation of the filament at that point as defined by $\boldsymbol{\omega}_\mathcal{G}$, the orientation is given as a unit vector;

8. [12] vmap – visit map (level set) value ($\mathcal{L}$);

9. [13] fden – weighted visit map value, filament density ($\mathcal{D}$);

10. [14] fori – strength of orientation as defined by $\mathcal{D}_\mathcal{G}$.

A3 Description of the galaxies table

The table of galaxies (see Table A3) contains the following information (the column numbers are given in square brackets):

1. [1] id – unique identification number for a galaxy, as used in Tempel et al. (2012);

2. [2] nrich – richness of the group the galaxy belongs to;

3. [3] redshift – redshift, corrected to the CMB rest frame;

4. [4–5] Ra, Dec. – right ascension and declination (deg);

5. [6] distcor – comoving distance of the galaxy when the finger-of-god effect is suppressed (as used in filament extraction);

6. [7–11] mag_*x* – Galactic extinction corrected Petrosian magnitude ($x \in ugriz$ filters);

7. [12] lumr – absolute luminosity in the *r* band in units of $10^{10} h^{-2}$ L$_\odot$, where M$_\odot$ = 4.64 (Blanton & Roweis 2007);

8. [13] *w* – weight factor for the galaxy ($w\cdot$lumr was used to calculate the luminosity density field);

9. [14] edgedist – comoving distance of the galaxy from the border of the survey mask;

10. [15] fil_dist – distance from the nearest filament axis (or from filament end point) in units of h^{-1} Mpc;

11. [16] fil_id – id of the nearest filament;

12. [17] fil_idpts – id of the nearest filament point.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article:

Table A1. The filament catalogue.
Table A2. The filament points table.
Table A3. The galaxies table. (http://mnras.oxfordjournals.org/lookup/suppl/doi:10.1093/mnras/stt2454/-/DC1).

Please note: Oxford University Press are not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.

This paper has been typeset from a TeX/LaTeX file prepared by the author.