Learning the principal graph of the galaxy distribution

Tony Bonnaire

Institut d'astrophysique spatiale

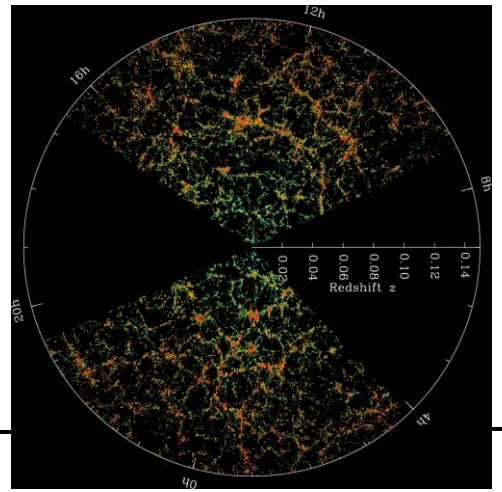
tony.bonnaire@ias.u-psud.fr

Supervisors: N. Aghanim, A. Decelle

Image from the Illustris collaboration

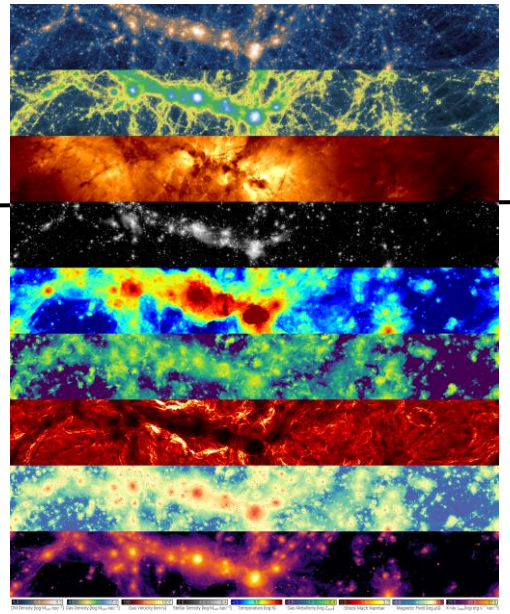
General context

- Many datasets $\mathbf{X} = \{\boldsymbol{x}_i\}_{i=1}^{N}$ are given as sets of D-dimensional datapoints. Identifying and extracting patterns in it is essential for understanding the underlying physical process that generated it and infer further properties of the process at hand.
- Most of these datasets do not span the entire $\mathbb{R}^D$ space but stand only on a lower-dimensional manifold.



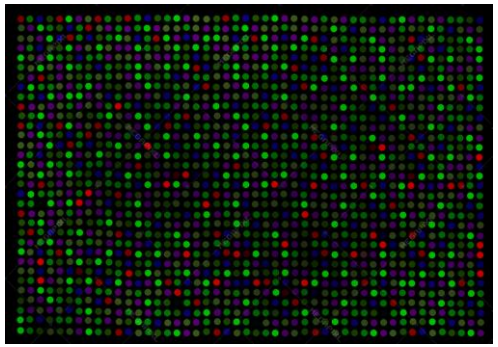
Galaxy distribution in the local Universe, $D = 3, 4$

Gas properties of the Cosmic Web, $D \sim 10$



Health status of patients, $D \sim 100$

Gene expressions, $D \sim 10^4 - 10^6$

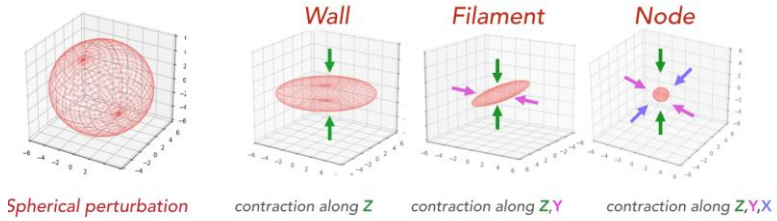


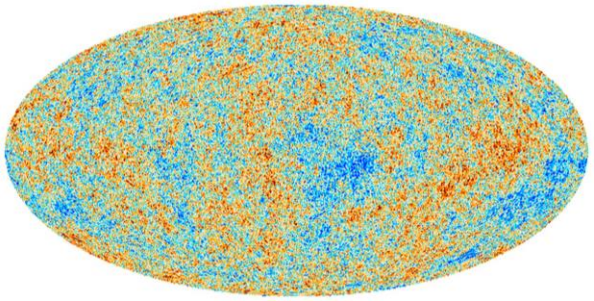
Images, $D \sim 10^3 - 10^6$

D

Cosmological context: The Cosmic Web

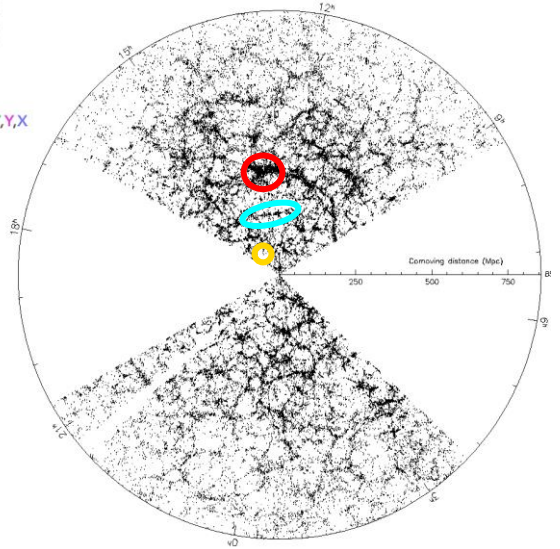
- The initial matter distribution in the universe is assumed **Gaussian and homogeneous with small perturbations**. These primordial density fluctuations gave birth, **under the effect of gravity**, to the structures observed today (Zeldovich+89).
- This spatial arrangement, called **the Cosmic Web**, falls into 4 main types of structures: Clusters, Filaments, Sheets or walls, Voids.



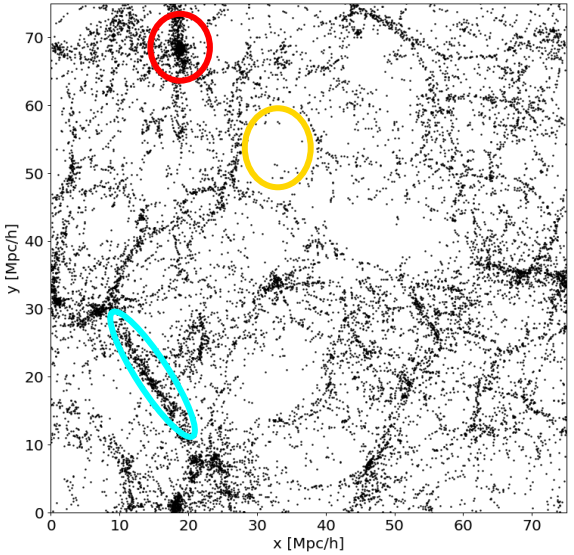


Initial density fluctuations, Gaussian field,
$t = 380\,000$ years

Gravity



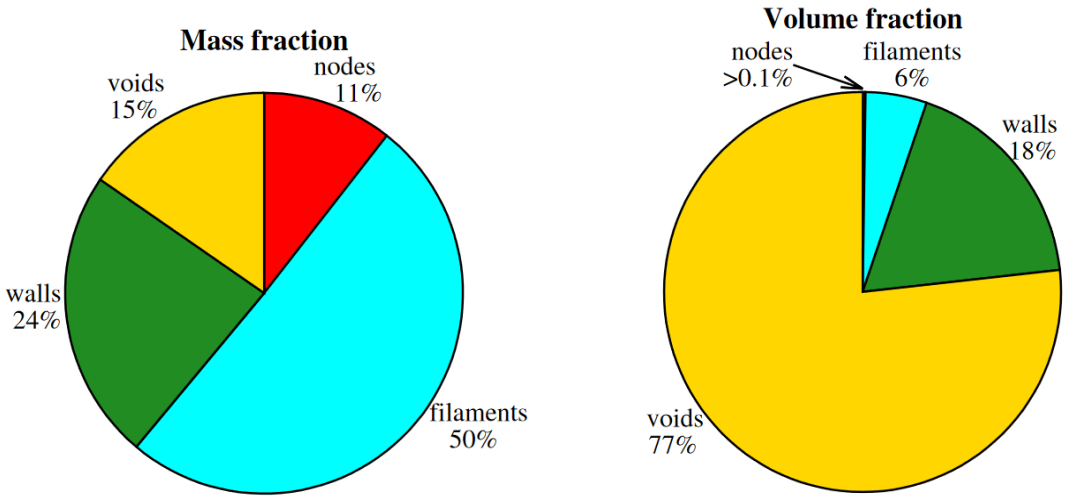
Observed local galaxy distribution from SDSS,
$t = 13.7$ billions years



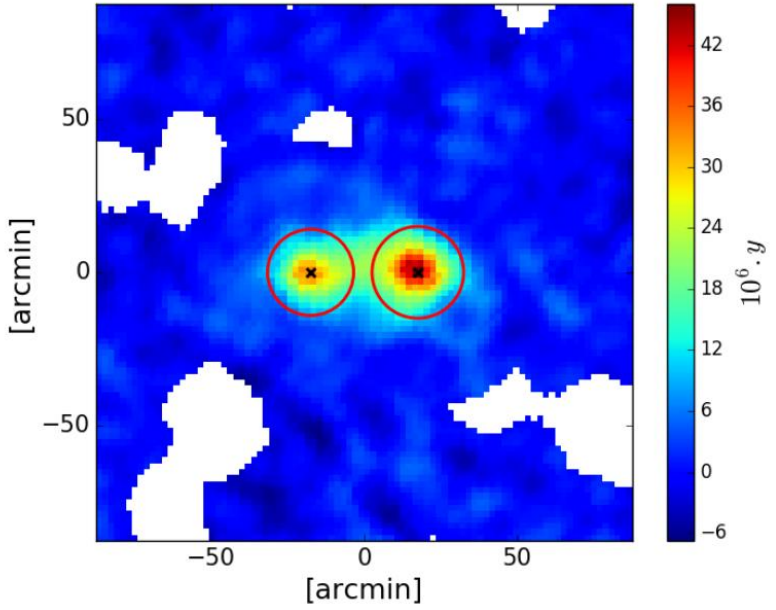
Galaxy distribution of a slice in the Illustris simulation.

Cosmological context: Filaments

- Filaments contain the **largest fraction of mass** in the Universe for a small part of the volume (Cautun+14).
- They also have been shown to **host a large part of baryons** in the form of hot and diffuse gas (Galarraga-Espinosa+20).
- Identifying filaments through the galaxy distribution is essential to (i) understand the impact of environment on the formation and evolution of galaxies and (ii) study their physical properties in other observables (SZ, X-Ray, Lensing, etc.).



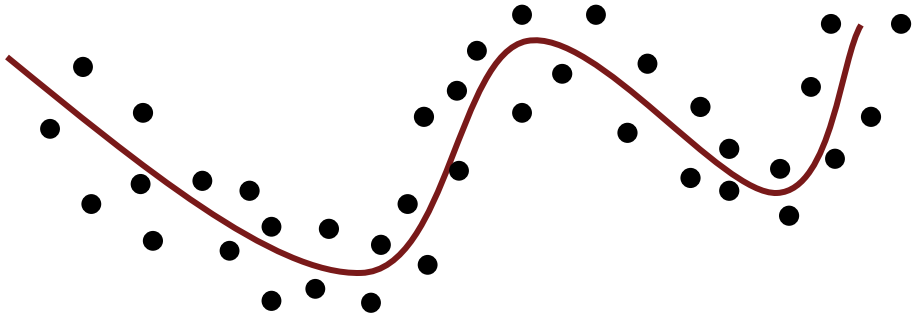
Mass and fraction volume in Cosmic Web elements (Cautun+14).



SZ Patch from the Planck tSZ map of the cluster pair A399-A401 (Bonjean+18).

Principal graph: Context

- Inspired from “principal curves” informally corresponding to **curves passing “in the middle”** (Hastie+89) of the point-cloud distribution.
- Principal graphs (Gorban+05) have been introduced to alleviate some restrictions in the curve formulation, as for instance the non self-intersecting condition.
- Here, we propose a formulation where we assume that the observed set of points are generated by the underlying one-dimensional manifold **that we model as a graph whose nodes are Gaussian clusters.**



Schematic view of the aim of principal curve or graph extraction.

Principal graph & Mixture Models

- *Method:* **Mixture Model** ➔ the probability that a datapoint is found at a position $\boldsymbol{x}_i \in \mathbb{R}^D$ given parameters $\boldsymbol{\Theta}$ of the model is

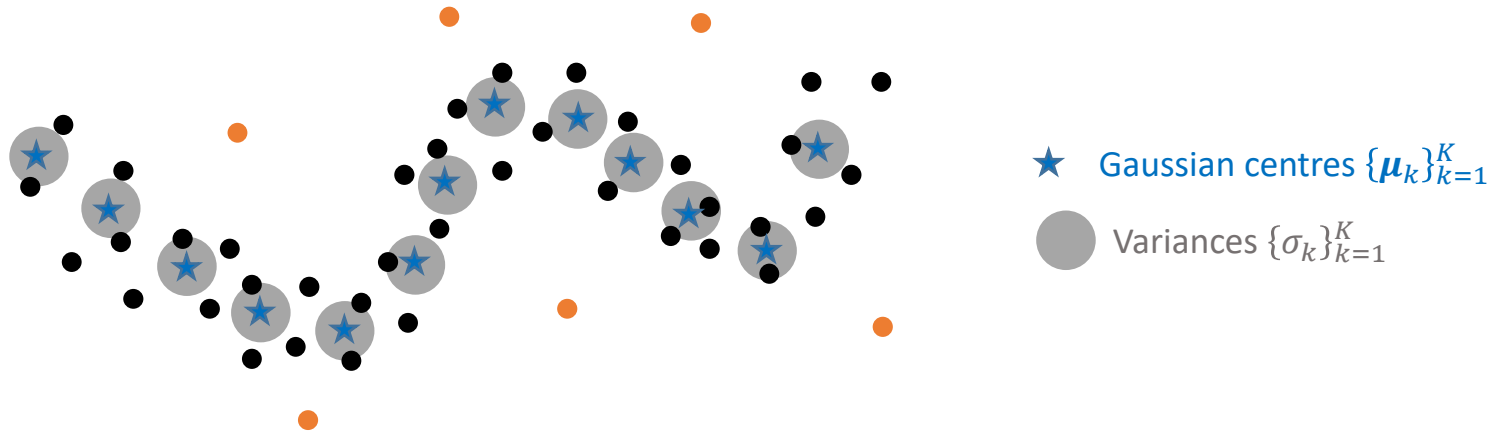
$$p(\boldsymbol{x}_i|\boldsymbol{\Theta}) = \sum_{k=1}^{K} \pi_k \, \mathcal{N}(\boldsymbol{x}_i \mid \boldsymbol{\Theta}) \; + \alpha\rho(\boldsymbol{x}_i)$$

Paves the distribution with Gaussians clusters

Assumes an additional uniform background

with $\boldsymbol{\Theta} = \{\alpha, \pi_1, \dots, \pi_K, \boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_K, \boldsymbol{\Sigma}_1, \dots \boldsymbol{\Sigma}_K\}$, $\rho(\boldsymbol{x})$ the uniform density over the data support volume $\mathcal{D} \subset \mathbb{R}^D$.

- *Assumption:* Spherical Gaussian clusters, $\forall\, k \in \{0, \dots, K\}$, $\boldsymbol{\Sigma}_k = \boldsymbol{I}_D \sigma_k^2$.



Schematic view of the aim of principal curve or graph extraction and basis of the proposed formalism.

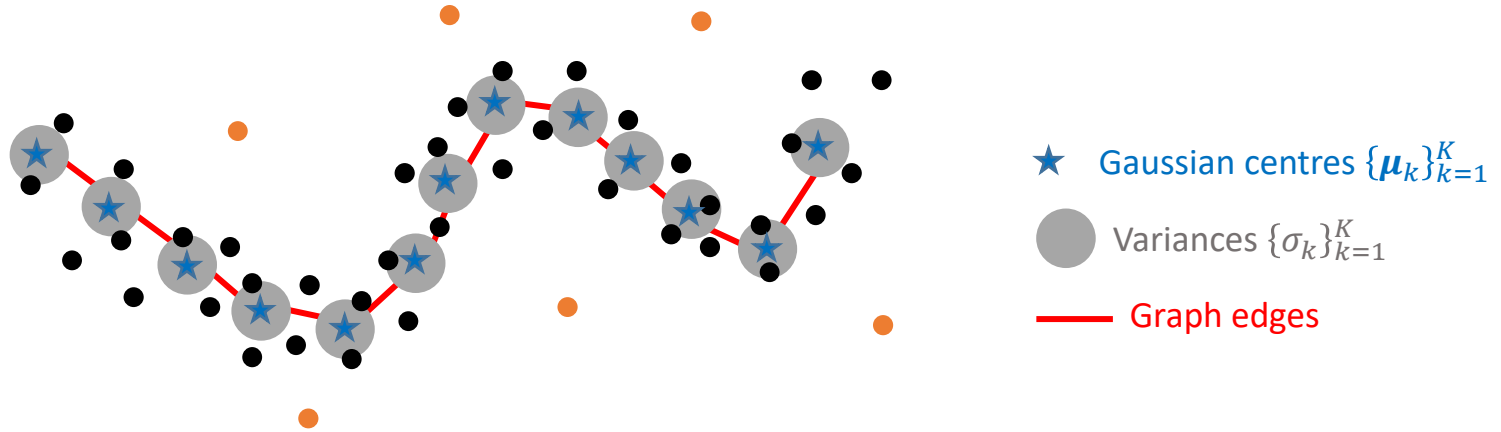
Regularised mixture models (RMM)

- The graph acts like a prior on the parameter space constraining the solution as

Encodes the graph topology

$$\log p(\boldsymbol{\mu}) \propto -\frac{\lambda_\mu}{2} \sum_{i=1}^{K} \sum_{j=1}^{K} \boldsymbol{A_{ij}} \left\| \boldsymbol{\mu}_i - \boldsymbol{\mu}_j \right\|_2^2 \qquad \text{with } A_{ij} = \begin{cases} 1 \text{ if } \mathrm{i} \sim \mathrm{j}, \\ 0 \text{ otherwise.} \end{cases}$$

➔ Gaussian form with a L_2 form **constraining the smoothness** of an estimate. Here, we do not impose the smoothness on the Euclidean space but directly using the graph structure $\mathcal{G}$.



Schematic view of the aim of principal curve or graph extraction and basis of the proposed formalism.

Full model and learning

- Optimal values of the parameters $\boldsymbol{\Theta}$ can be obtained by using **the Expectation-Maximisation algorithm** (EM, Dempster+77, Bishop+06)

$$\boldsymbol{\Theta}^{(t+1)} = \operatorname{argmin}_{\boldsymbol{\Theta}} \sum_{i=1}^{N}\sum_{k=1}^{K} p_{ik} \frac{\|\boldsymbol{x}_i - \boldsymbol{\mu}_k\|_2^2}{\sigma_k^2} + \sum_{i=1}^{N} p_i^{bkg} \log \rho(\boldsymbol{x}_i) + \lambda_\mu \sum_{i=1}^{K}\sum_{j=1}^{K} \boldsymbol{A}_{ij} \|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|_2^2 + \cdots$$

Probabilistic association terms to Gaussians or background

Data fidelity term

Regularisation parameter

Topological prior given by the graph structure

- λ_μ is a trade-off parameter between how much the graph should fit the data and how much it should be smooth.
- EM leads to the following update equations:

$$\begin{cases} p_{ik} = p(z_i = k \mid \boldsymbol{x}_i, \boldsymbol{\Theta}^{(t)}) = \dfrac{\pi_k \mathcal{N}(\boldsymbol{x}_i \mid \boldsymbol{\theta}_k)}{\sum_{k'=1}^{K} \pi_{k'} \mathcal{N}(\boldsymbol{x}_i \mid \boldsymbol{\theta}_{k'}) + \alpha\rho(x_i)} \\ p_i^{bkg} = p(z_i = K+1 \mid \boldsymbol{x}_i, \boldsymbol{\Theta}^{(t)}) = \dfrac{\alpha\rho(x_i)}{\sum_{k'=1}^{K} \pi_{k'} \mathcal{N}(\boldsymbol{x}_i \mid \boldsymbol{\theta}_{k'}) + \alpha\rho(x_i)} \end{cases}$$

$$\begin{cases} \boldsymbol{\mu}_k^{(t+1)} = \dfrac{\sum_{i=1}^{N} \frac{p_{ik}}{\sigma_k^2} \boldsymbol{x}_i + 2\lambda \sum_{j=1}^{K} \boldsymbol{A}_{kj} \boldsymbol{\mu}_j^{(t+1)}}{\sum_{i=1}^{N} \frac{p_{ik}}{\sigma_k^2} + 2\lambda \sum_{j=1}^{K} \boldsymbol{A}_{kj}} \\ \sigma_k^{(t+1)} = \left[\dfrac{1}{D \sum_{i=1}^{N} p_{ik}} \sum_{i=1}^{N} p_{ik} (\boldsymbol{\mu}_k - \boldsymbol{x}_i)^{\mathrm{T}} (\boldsymbol{\mu}_k - \boldsymbol{x}_i) \right]^{1/2} \\ \pi_k^{(t+1)} = \frac{1}{N} \sum_{i=1}^{N} p_{ik} \\ \alpha^{(t+1)} = \frac{1}{N} \sum_{i=1}^{N} p_i^{bkg} \end{cases}$$

Full model and learning

- Optimal values of the parameters $\boldsymbol{\Theta}$ can be obtained by using **the Expectation-Maximisation algorithm** (EM, Dempster+77, Bishop+06)

Probabilistic association terms to Gaussians or background

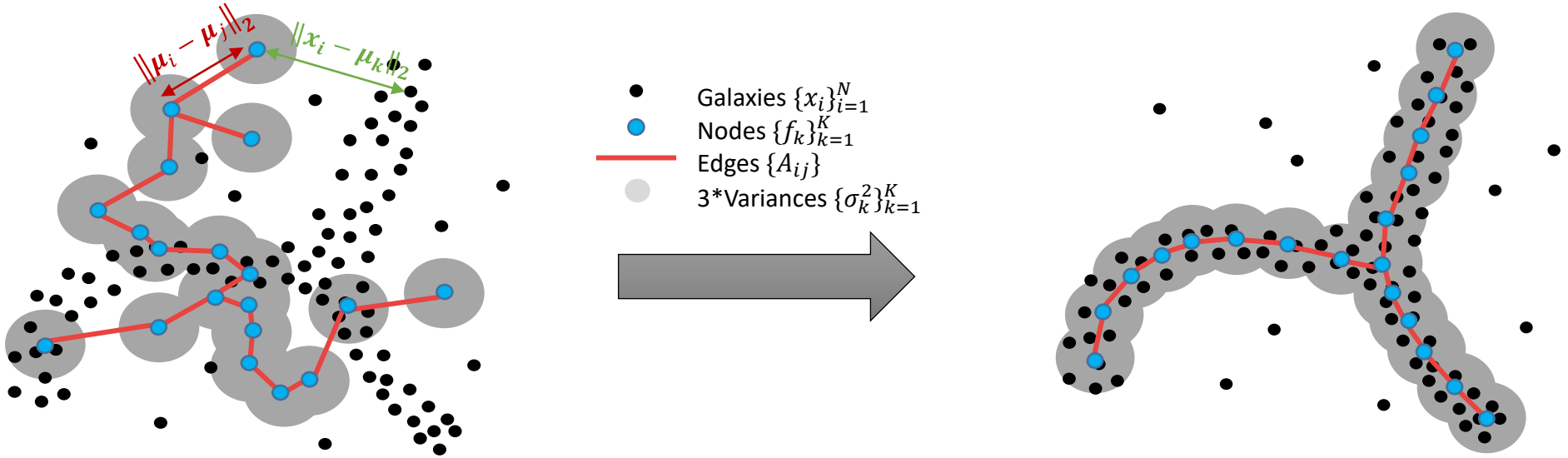
$$\boldsymbol{\Theta}^{(t+1)} = \text{argmin}_{\boldsymbol{\Theta}} \sum_{i=1}^{N}\sum_{k=1}^{K} p_{ik} \frac{\|\boldsymbol{x}_i - \boldsymbol{\mu}_k\|_2^2}{\sigma_k^2} + \sum_{i=1}^{N} p_i^{bkg} \log \rho(\boldsymbol{x}_i) + \lambda_\mu \sum_{i=1}^{K}\sum_{j=1}^{K} \boldsymbol{A}_{ij} \|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|_2^2 + \cdots$$

Data fidelity term

Regularisation parameter

Topological prior given by the graph structure

- λ_μ is a trade-off parameter between how much the graph should fit the data and how much it should be smooth.



Graph prior: Minimum spanning tree

- Historically the first method used to exhibit the filamentary structure (Barrow+85). It associates to a galaxy distribution a **unique graph with no parameter** minimising the global distance.
- The proposed approach:
  - ✓ Leads to a **smoother version** which keeps the **same definition** for filaments (Bonnaire+20).
  - ✓ Can be extended **to take into account cycles** that can be observed in the spatial distribution. (Bonnaire+20, submitted).
  - ✓ Can describe **the local size of the filamentary pattern** through the variances.

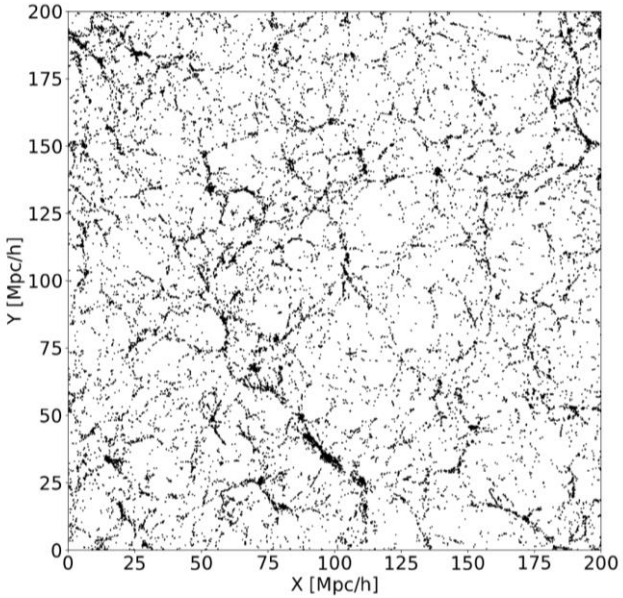


Fig. 8. Galaxy distribution of a slice of the IllustrisTNG simulation (Pillepich+18).

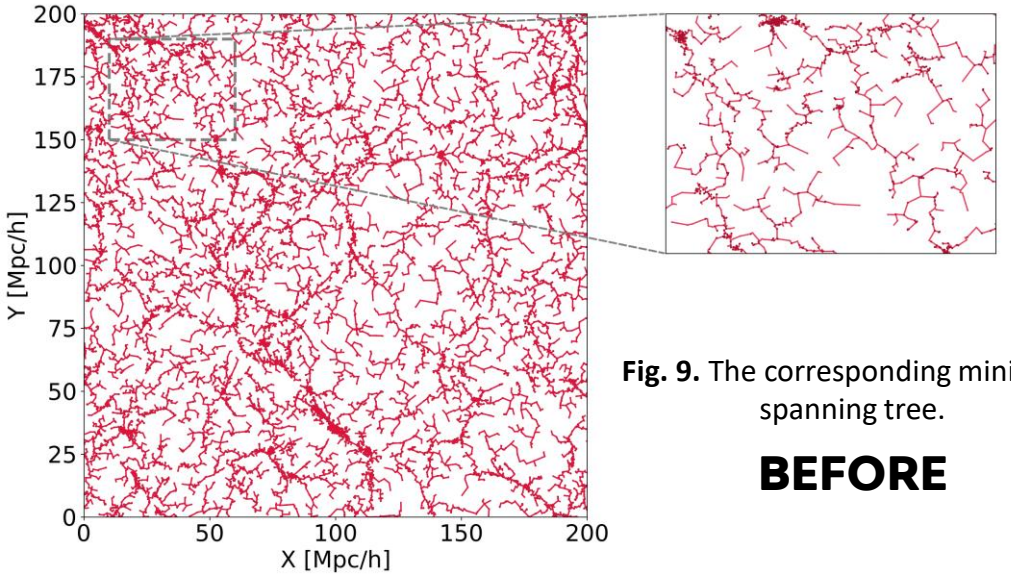


Fig. 9. The corresponding minimum spanning tree.

BEFORE

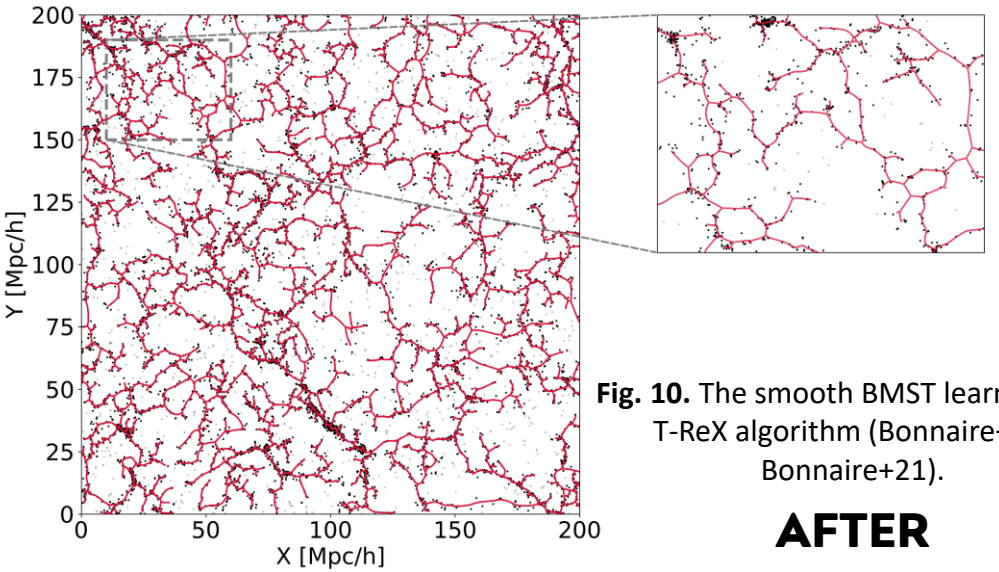
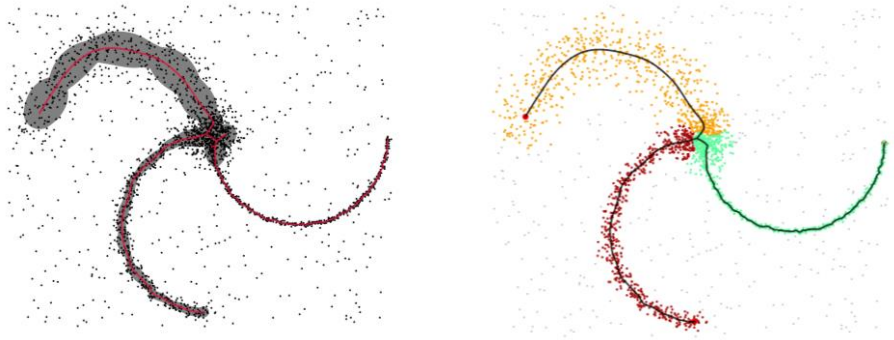


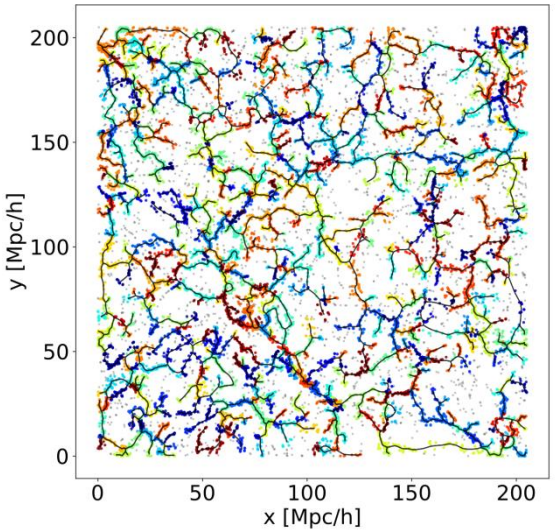
Fig. 10. The smooth BMST learnt by the T-ReX algorithm (Bonnaire+20, Bonnaire+21).

AFTER

Astrophysics with the RMM

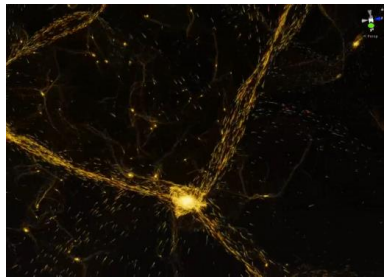


Toy dataset with 20% background noise (Black dots). The learnt RMST (red) with 3*variances of nodes (grey shaded areas).

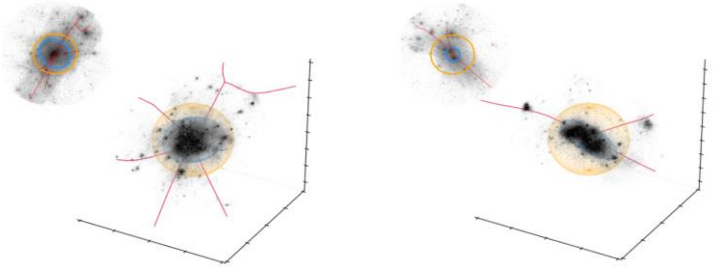


Simulation of galaxy distribution coloured by with respect to the background (grey points) or to each branch (coloured points).

Study the physical properties of galaxies (mass, star formation, etc.) and their host filaments (length, thickness, curvature)

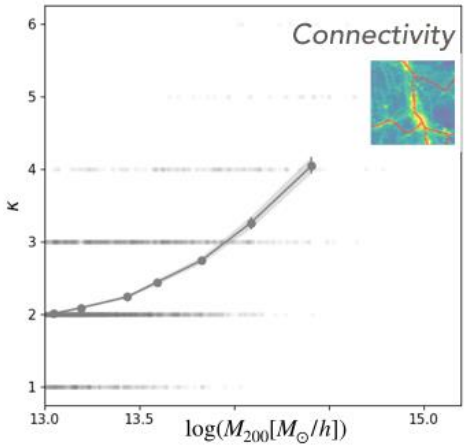


A node accreting matter through filaments in a simulation (Credit: Miguel Aragon-Calvo)



The connectivity κ is the number of detected filaments crossing a spherical volume around a massive node.

Study the physical properties of clusters (mass, shape, accretion history, etc.) with respect to their connectivity



Evolution of the connectivity κ with the mass of the node (Gouin+21).

Summary and perspectives

Summary

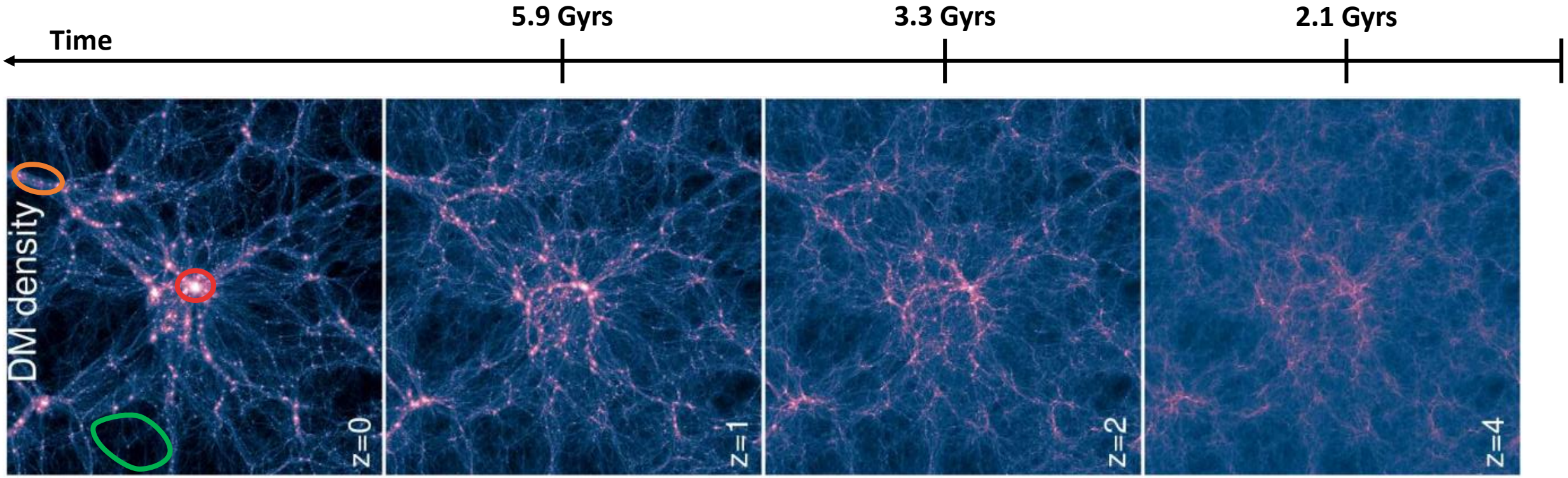
- Filaments are interesting objects to assess cosmological/astrophysical hypotheses
- Graph regularised mixture models can be used to extract a model of the 1D manifold that generated the data
- Not presented: Extensions of the tree topology (persistent homology), application of the RMM to other problematics (road network extraction), exploitation of the reformulation in statistical physics of the clustering to gain an insight on the dataset

Future works

- Information content of cosmic structures (cosmological parameters)
- Study the links between the characteristics of filaments and their galaxies
- Time evolution of the graph structure (in simulation)
- Write a manuscript…

Additional & back-up slides

The Cosmic Web



21.3Mpc/h deep slice of DM density as computed by Illustris simulation, width is 106.5Mpc/h (*Vogelsberger M. et al., 2014*)

1. **Nodes**
2. **Filaments**
3. **Walls**
4. **Voids**

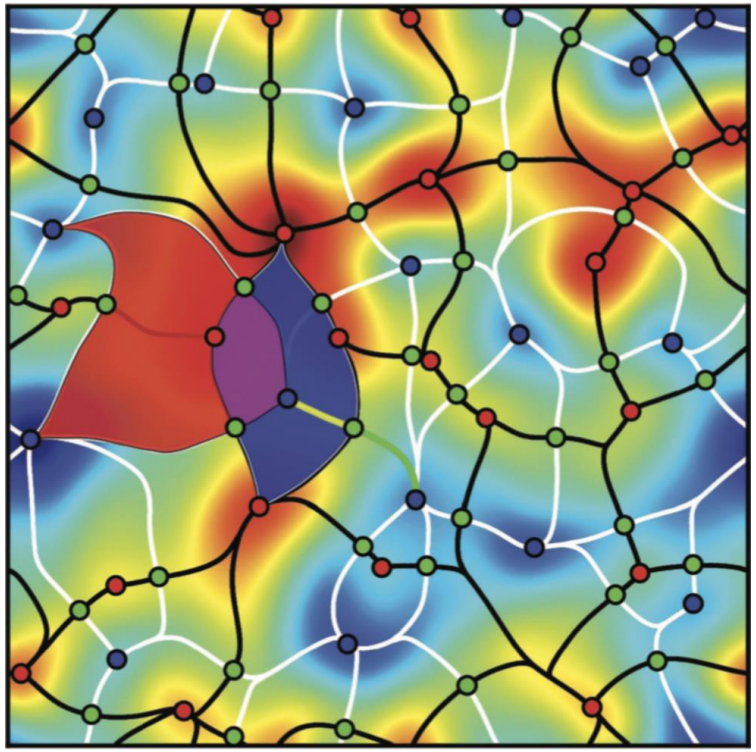
Overview of (a few) existing procedures

Name	Global methodology	Filaments definition
MST *Barrow et al., 1985*	Compute the Minimum Spanning tree over the galaxy distribution	Branches of a post-processed version of the MST
MMF *Aragon-Calvo, M., et al. 2007* **Nexus** *Cautun M. et al., 2013*	Scale-space representation of the continuous density field*	Locally defined through eigenvalues of the Hessian matrix
T-Web *Forero-Romero et al., 2009*	Compute the tidal shear tensor and extract the morphology of the field	Locally defined through eigenvalues of the tidal shear tensor
DisPerSE *Sousbie et al., 2011* **SpineWeb** *Aragón-Calvo, M., et al., 2010*	Study the topological properties of the continuous density field* through the Discrete Morse Theory	Connections between singularities (maxima, minima and saddle points)
Bisous *Tempel et al., 2014*	Stochastic fit of parametric and interactive cylinders in the galaxy distribution	Aligned and close cylinders
SCMS *Chen et al., 2015*	**S**ubspace **C**onstraint **M**ean-**S**hift algorithm is applied on galaxy distribution	Principal curves of the point cloud distribution

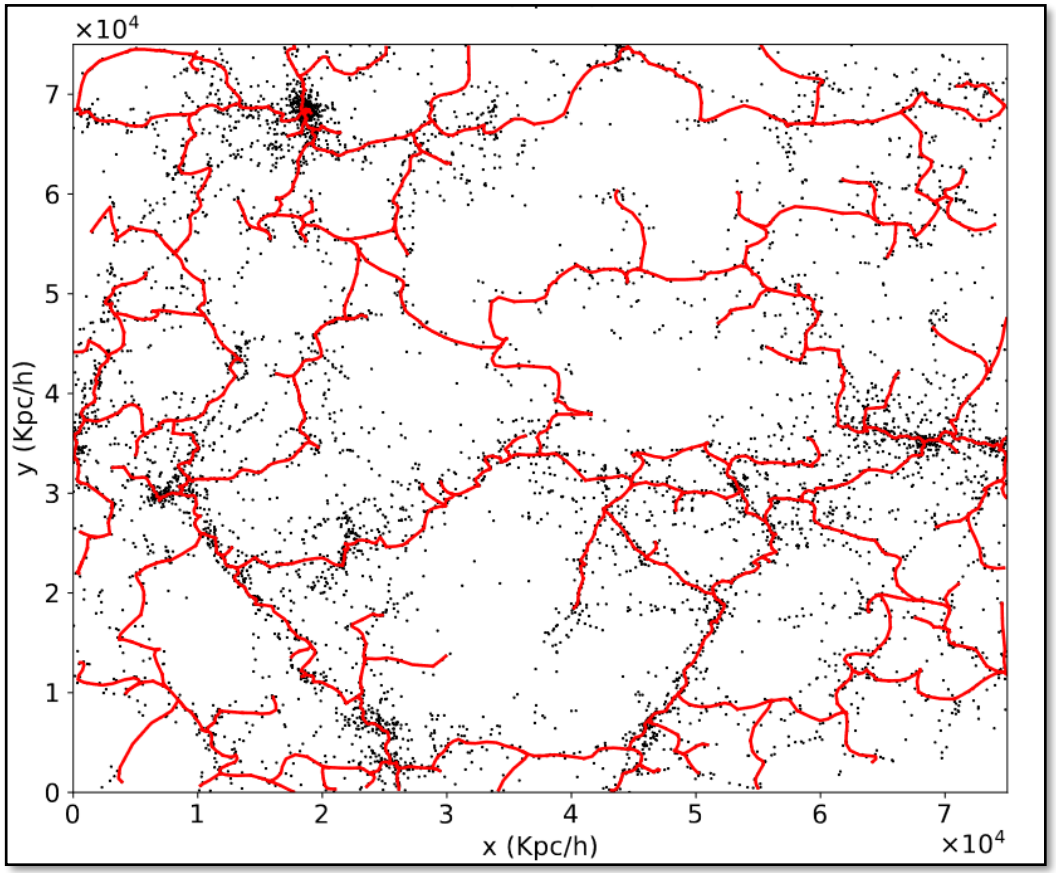
* Continuous density fields are obtained by performing a **D**elaunay **T**essellation **F**ield **E**stimation (DTFE)

DisPerSE (Discrete Persistent Structure Extractor)

- Identify topological features (peaks, voids, walls and filaments) from a N-D distribution
- The persistence measures the robustness of the extracted features to remove spurious detections



Example of critical points with the integral lines thus delimiting ascending (black) or descending (white) manifolds (*Sousbie et al., 2010*)



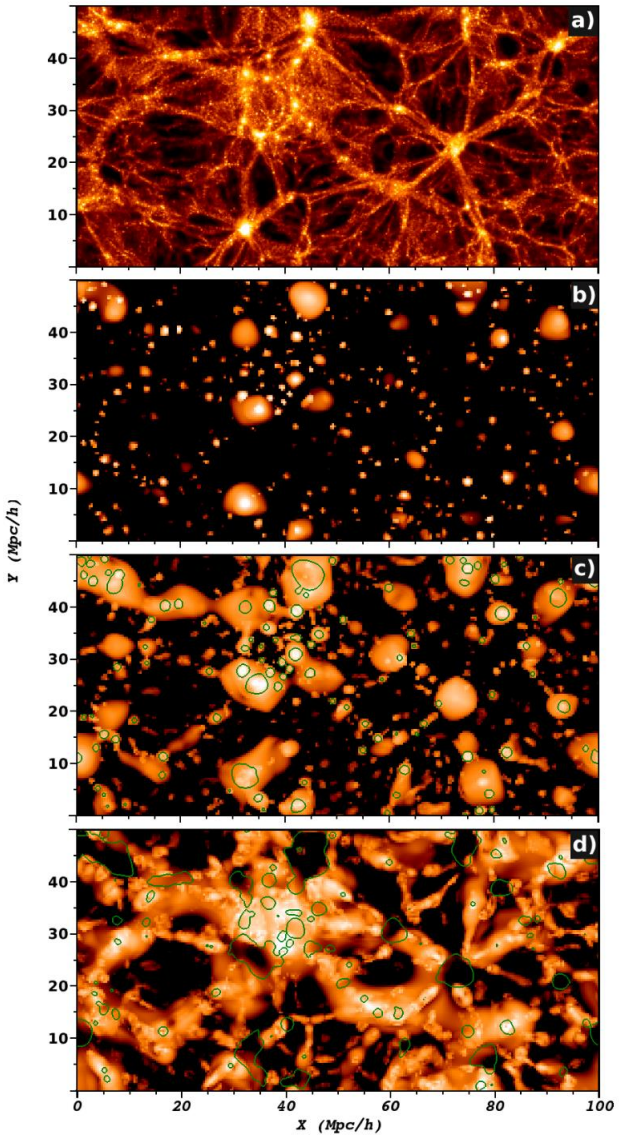
Skeleton provided by DisPerSE (2σ persistence, 1 smoothing of the DTFE) on a 2Mpc/h slice of Illustris-3

Nexus

- Uses elements from image processing to build a scale-space representation of the continuous field (obtained via DTFE)
- The strength of Nexus is to compute the signature over a range of scales
- Steps:
  1. Smooth the field with a gaussian kernel of variance R_n
  2. Perform the spectral decomposition of the Hessian at each position
  3. Compute the signature for each position and structure
  4. Update $R_{n+1} = \sqrt{2}\, R_n$ and iterate
- Physical criterion are then used to threshold the signature values

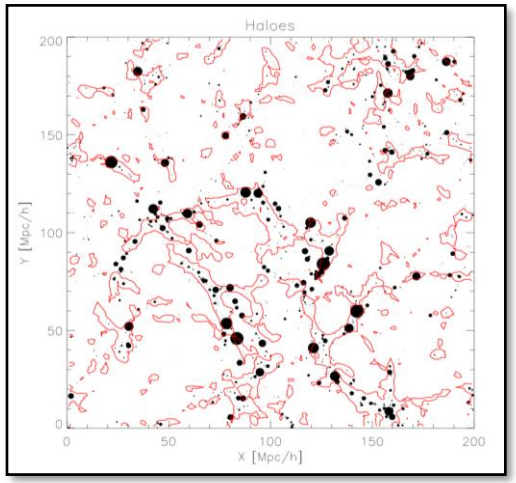
Structure	Soft constraints	Strict constraints
cluster	$\lvert\lambda_1\rvert \simeq \lvert\lambda_2\rvert \simeq \lvert\lambda_3\rvert$	$\lambda_1 < 0;\ \lambda_2 < 0;\ \lambda_3 < 0$
filament	$\lvert\lambda_1\rvert \simeq \lvert\lambda_2\rvert \gg \lvert\lambda_3\rvert$	$\lambda_1 < 0;\ \lambda_2 < 0$
wall	$\lvert\lambda_1\rvert \gg \lvert\lambda_2\rvert;\ \lvert\lambda_1\rvert \gg \lvert\lambda_3\rvert$	$\lambda_1 < 0$

Images from Cautun+14.

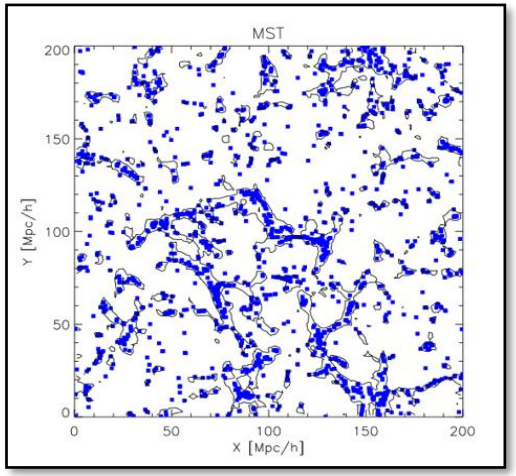


Different definitions lead to different results

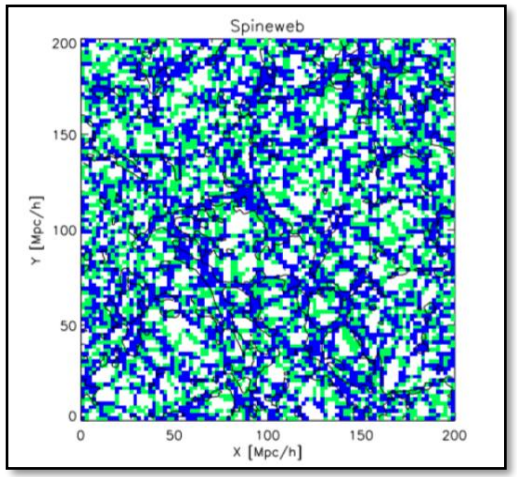
Halos



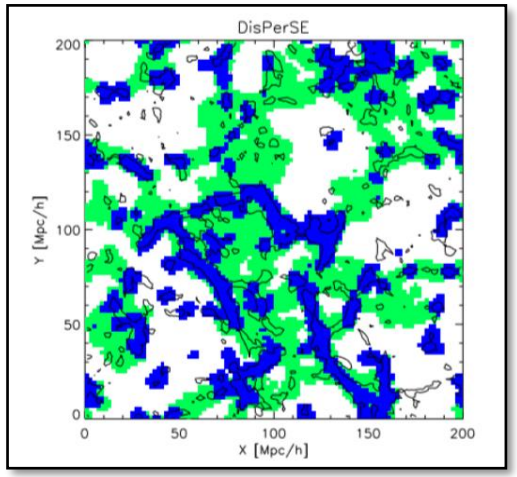
MST

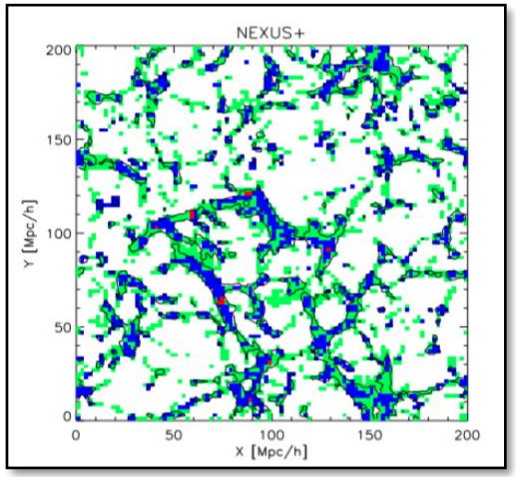


Spineweb

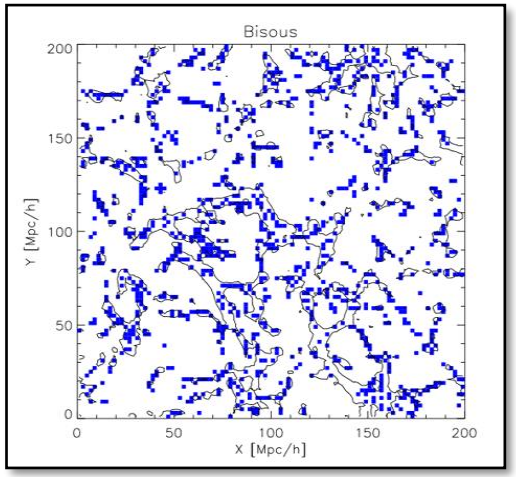


DisPerSE

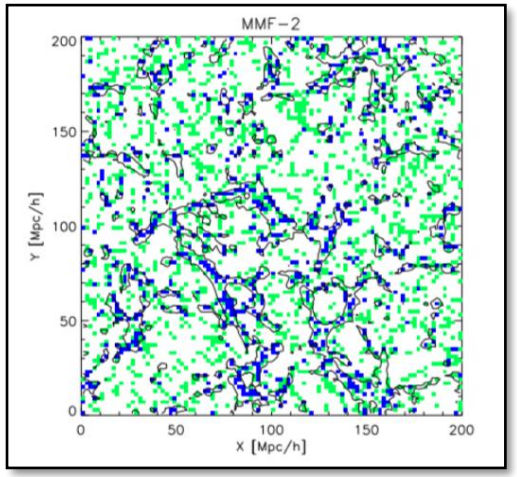




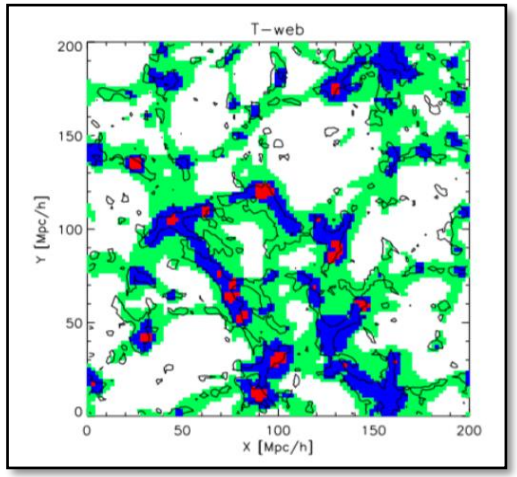
Nexus+



Bisous



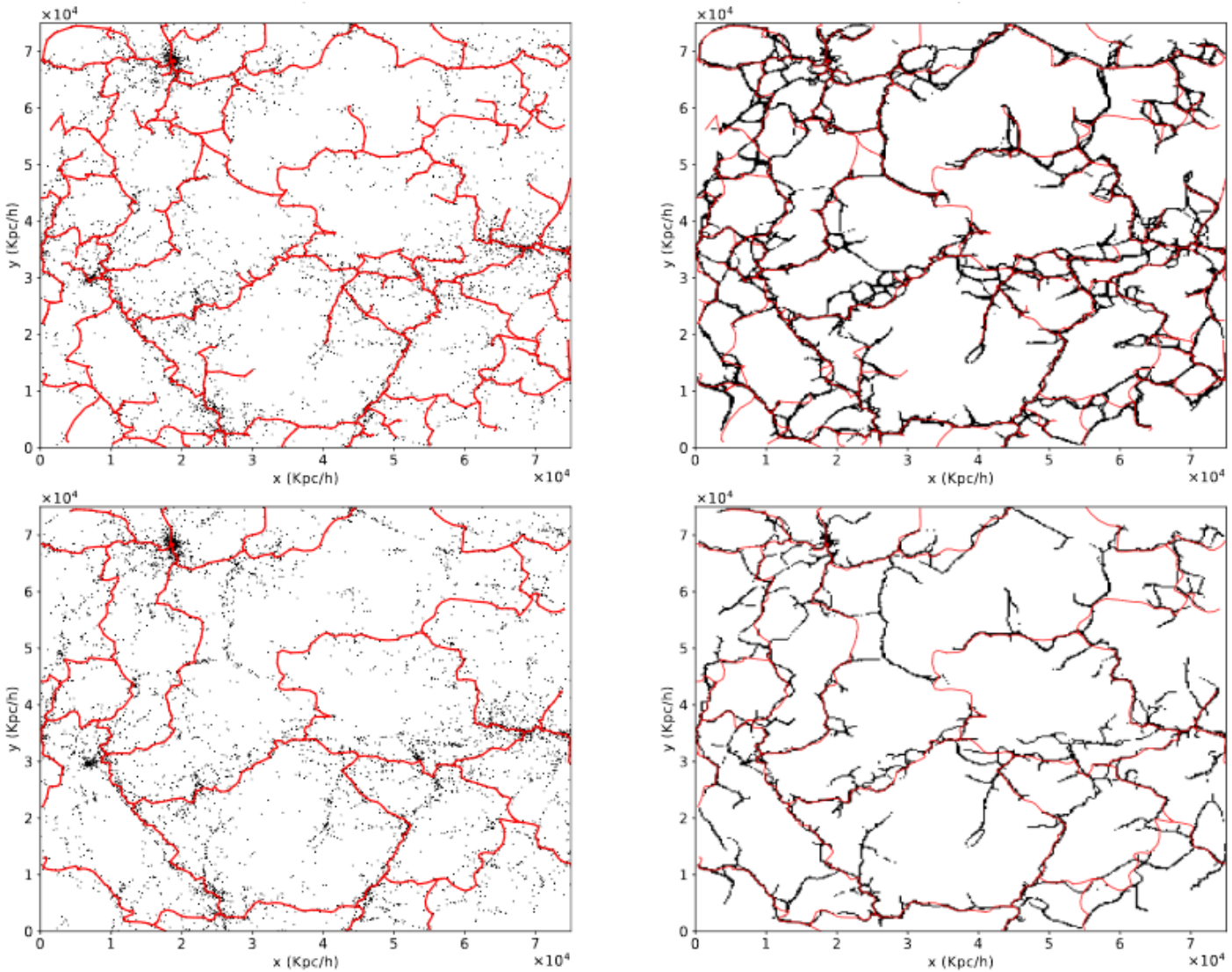
MMF



T-Web

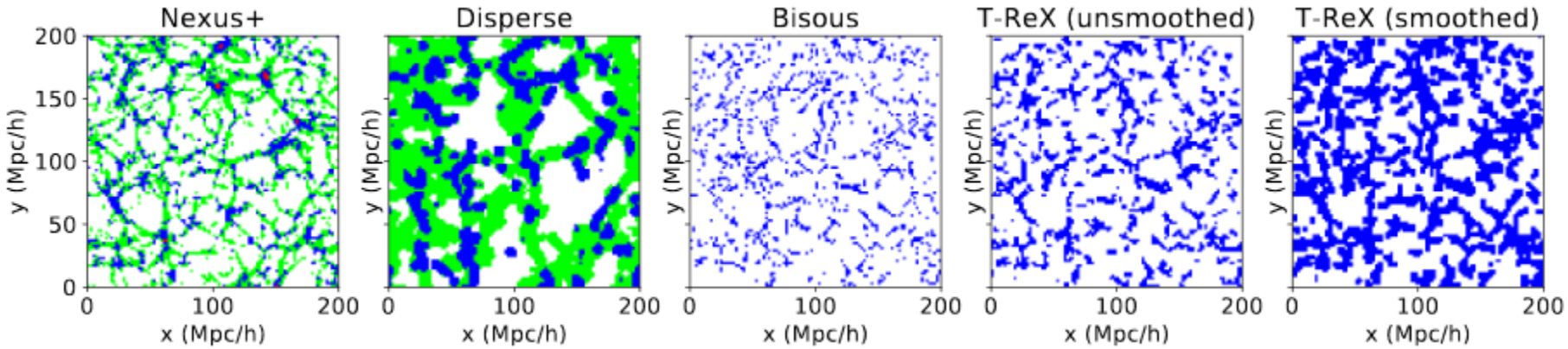
Comparison of existing procedures. Filaments are in blue. (Images from Linbeskind+17)

T-ReX vs DisPerSE (2D)



(left) A 2Mpc/h slice of Illustris-3 with the 2σ *(top)* or 5σ *(bottom)* persistence DisPerSE skeleton *(right)* Probability map thresholded at $p = 0.1$ *(top)* and $p = 0.25$ *(bottom)*

T-ReX vs Nexus+ vs Bisous vs DisPerSE (3D)



- The measurement of the similarity between results highlights differences:

$$S(H_1, H_2) = \frac{|H_1 \cap H_2|}{|H_1|}$$

where H_1 and H_2 are two detection maps and $|\cdot|$ denotes the cardinal

- $S(H_1, H_2)$ can be seen as the ratio of true positive results provided by H_1 when considering H_2 as a reference

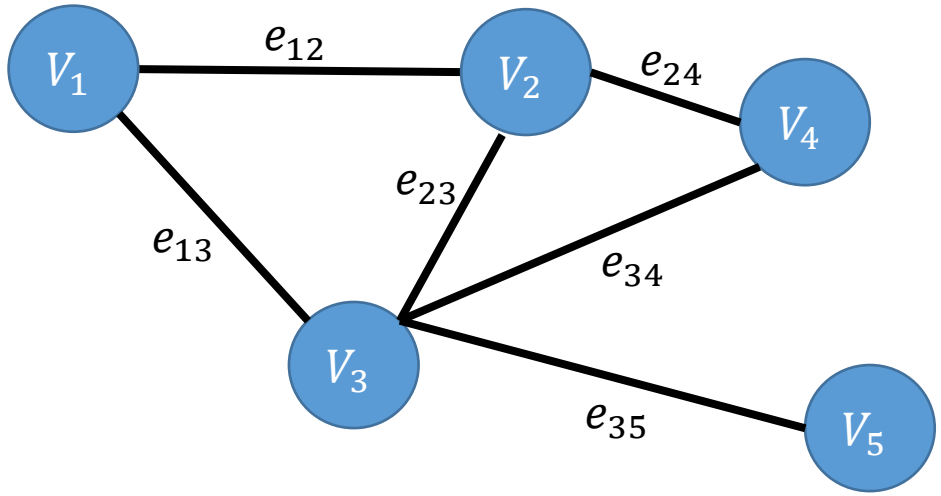
H_1 \ H_2	T_{us}	T_s	N	D	B
T_{us}	1	1	0.85	0.62	0.37
T_s	0.48	1	0.62	0.62	0.24
N	0.53	0.81	1	0.62	0.30
D	0.22	0.46	0.35	1	0.12
B	0.66	0.87	0.86	0.62	1

A quick introduction to graph theory

- A **graph** $G = (V, E, w)$ is a mathematical object made of nodes $\{V_i\}$, of edges linking nodes together $\{E_i\}$ with a weight w_e associated to each edge.

- A graph can be represented by a matrix describing interactions between nodes: **the adjacency matrix** whose value is 1 when nodes are connected and 0 otherwise.

$$A = \begin{pmatrix} 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 & 1 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \end{pmatrix}$$



- Examples of graphs:
  - Internet => Nodes are URL addresses. Two nodes are connected if there is a link allowing the passage of on to another.
  - Road or air network => Nodes are cities (or airports) and are linked if there is a path between the two.
  - Cosmic Web => Nodes are galaxies. Two galaxies are linked if they are sufficiently close with $w_{ij} = \|x_i - x_j\|_2$. We can also extend this notion to the entire CW with nodes and filaments.

Graph Laplacian and smoothness

- The **Laplacian matrix** is another way to algebraically represent a graph, with $\boldsymbol{D}$ the diagonal degree matrix, $\boldsymbol{A}$ the adjacency matrix,

$$\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{A}$$

- In the RMM, we **constrain the smoothness** of the graph through $\log p(\boldsymbol{\mu}) \propto -\frac{\lambda_\mu}{2} \sum_{i=1}^{K} \sum_{j=1}^{K} \boldsymbol{A}_{ij} \left\| \boldsymbol{\mu}_i - \boldsymbol{\mu}_j \right\|_2^2$.

Total length of the graph

- We look for the graph that is "short" AND fits the data => smoothness constraint
- The quadratic summation can be written as the norm of $\boldsymbol{\mu}$ on the graph structure can be written in terms of the Laplacian matrix as

$$\|\boldsymbol{\mu}\|_{\mathcal{G}}^2 = \sum_{i=1}^{K} \sum_{j=1}^{K} \boldsymbol{A}_{ij} \left\| \boldsymbol{\mu}_i - \boldsymbol{\mu}_j \right\|_2^2 = 2 \operatorname{Tr}\{\boldsymbol{\mu}^{\mathrm{T}} \boldsymbol{L} \boldsymbol{\mu}\}.$$

- This is equivalent to constraining the L_2 of a vector to constrain its smoothness in the Tikhonov regularisation but the norm is being computed on the graph structure $\mathcal{G}$.

Regularised mixture models

- The graph acts like a prior on the parameter space constraining the solution as

$$\log p(\boldsymbol{\mu}) \propto -\frac{\lambda_\mu}{2} \sum_{i=1}^{K} \sum_{j=1}^{K} \boldsymbol{A_{ij}} \left\| \boldsymbol{\mu}_i - \boldsymbol{\mu}_j \right\|_2^2$$

Fully encodes the graph information

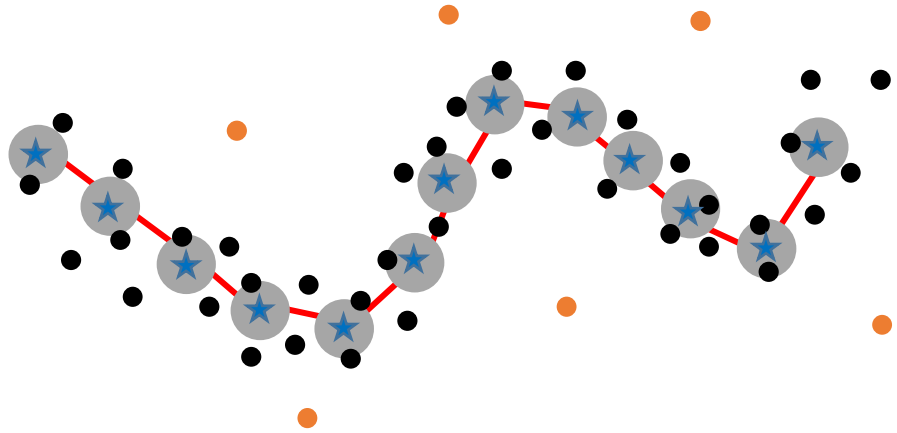
$$\text{with } A_{ij} = \begin{cases} 1 \text{ if } \mathrm{i} \sim \mathrm{j}, \\ 0 \text{ otherwise.} \end{cases}$$

- This prior has a Gaussian form with a L_2 constraint like usually done when **constraining the smoothness** of an estimate. Here the idea is the same, except that we do not impose the smoothness on the Euclidean space but directly using the graph structure $\mathcal{G}$.

- Additional embedding of the **spatially coherent evolution of the learnt variances**

$$\log p\left(\sigma_k^2\right) \propto -\lambda_\sigma [\log \sigma_k^2 + \sigma_{\mathcal{N}_k} / \sigma_k^2]$$

with $\sigma_{\mathcal{N}_k} = |\mathcal{N}_k|^{-1} \sum_{i \in \mathcal{N}_k} \sigma_i^2$ and $N_k = \{i \,|A_{ik} = 1\}$ the set of neighbouring nodes of k in the graph is **the mode of the distribution**.



Schematic view of the aim of principal graph extraction and basis of the proposed formalism.

- A prior on amplitudes is used to avoid singular solutions, $\log p(\pi_k) \propto -\frac{\lambda_\pi}{2} \left[\frac{1-\alpha}{K} - \pi_k \right]^2$.

- The full prior distribution on $\boldsymbol{\Theta}$ is $\log p(\boldsymbol{\Theta}) = \log p(\boldsymbol{\mu}) + \sum_k \log p(\sigma_k^2) + \sum_k \log p(\pi_k)$.

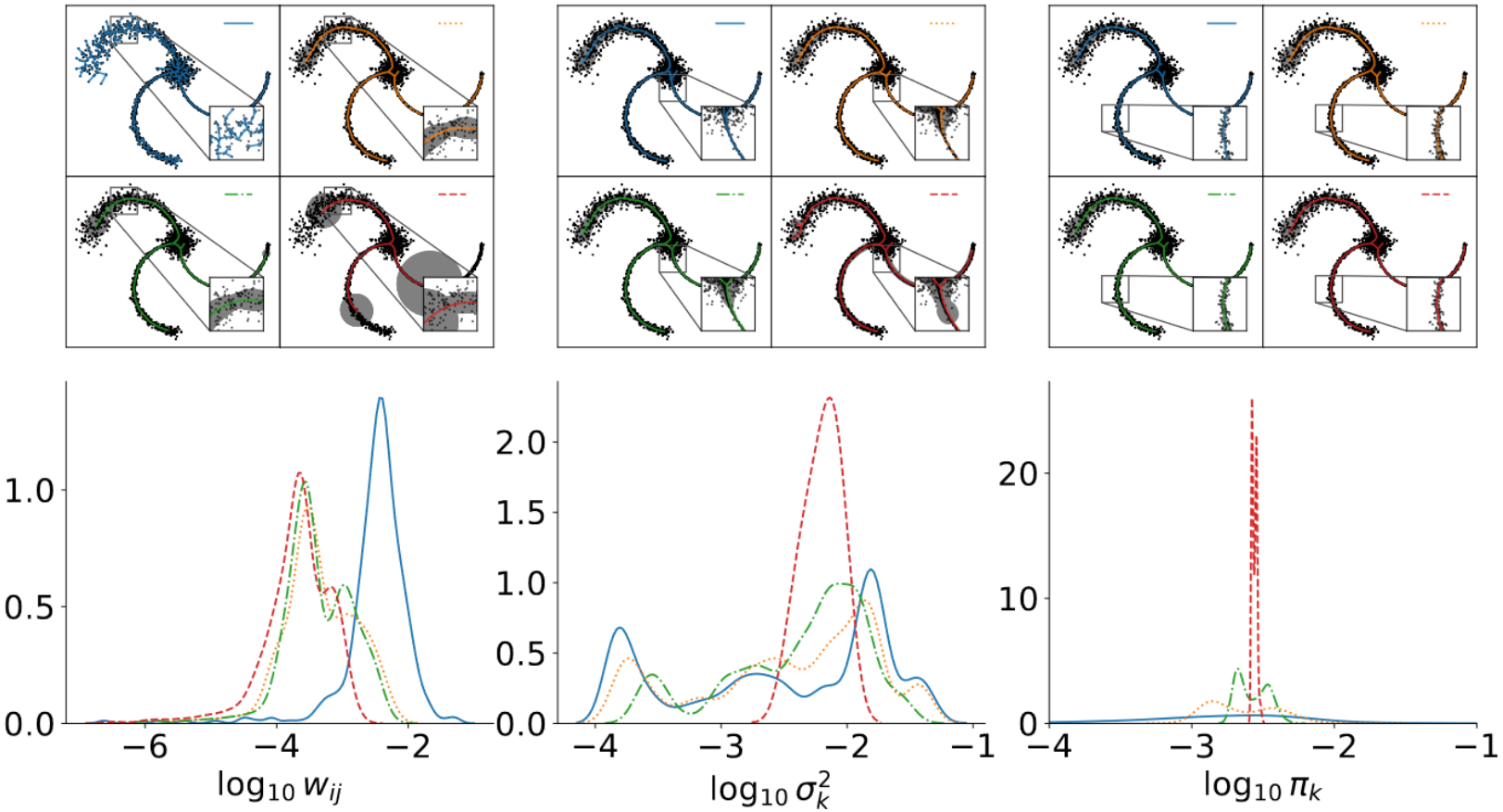
★ Gaussian centres $\{\boldsymbol{\mu}_k\}_{k=1}^K$

Variances $\{\sigma_k\}_{k=1}^K$

Graph edges

T-ReX: Effects of hyper-parameters

- Hyper-parameters of the full model are $\mathcal{Y} = (\lambda_\mu, \lambda_\sigma, \lambda_\pi)$ and K. All hyper-parameters in $\mathcal{Y}$ have an impact on the force of the prior of the corresponding parameter indicated as a subscript.
- There is a wide range of values for which the algorithm provides similar results.



T-ReX: Initialisation

- There are four parameters in the model are $\boldsymbol{\mu}$, $\{\sigma_k\}_{k=1}^K$, $\{\pi_k\}_{k=1}^K$ and α to initialise.
- $\boldsymbol{\mu}$ is initialised with datapoints $\boldsymbol{X}$, possibly with a random sub-sampling or gridding to have $K \ll N$.
- Variances can be initialised as $\forall k \in \{1, \dots, K\}, \sigma_k^2 = \sigma_0^2$ with σ_0^2 chosen in a similar way as in Chen+14 for the Subspace Constrained Mean-Shift algorithm (SCMS), using prescription from kernel-density estimation method like the **Silverman rule** (Silverman+86)

$$\sigma_0 = A_0(Nd + 2N)^{-\frac{1}{d+4}}\sigma_{\min}$$

where $\sigma_{\min}$ is the minimum variance in all dimensions of the datasets.

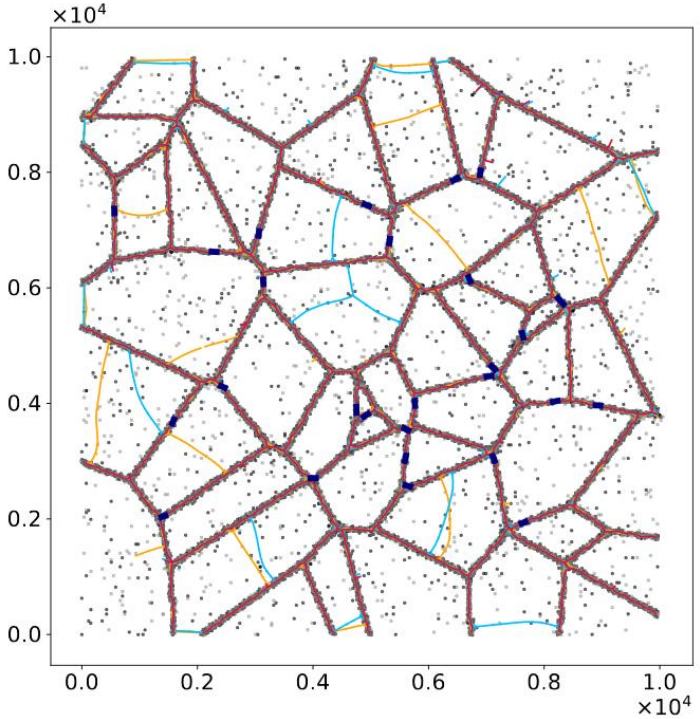
- Amplitudes of Gaussian clusters are initialised uniformly as $\forall k \in \{1, \dots, K\}, \pi_k = \frac{1-\alpha}{K}$
- α, the amplitude of uniform background should be initialised as a first guess of the outliers level in the dataset
- Although EM is known to highly depend on the initialisation because it can be trapped in local maxima near the initialisation point, there are ways to alleviate the problem **using simulated annealing** procedures (Bonnaire+20b).
- When the dataset is large, these solution comes with a large computational cost. In all our runs, we did not find this problem sufficiently significant to invoke such solutions.

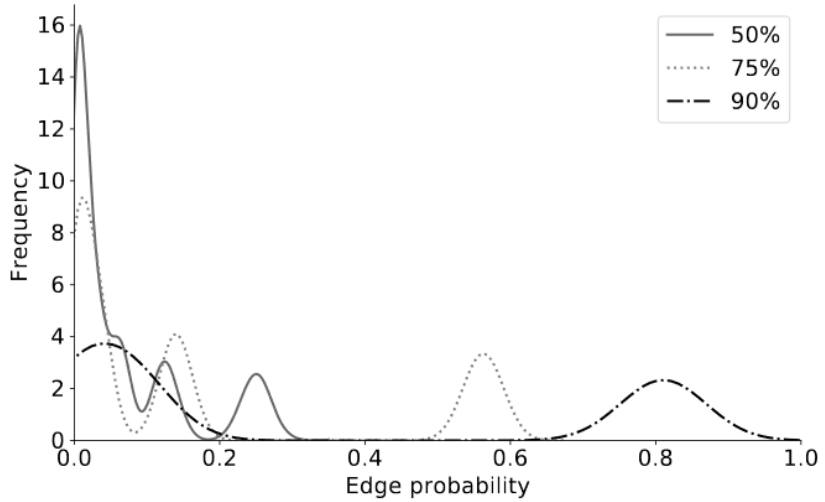
T-ReX: Average graph prior

- Even if it has a lot of convenient features, the MST topology can not represent cycles.
- In the model, the graph intervenes only through its algebraic representation (like the adjacency or Laplacian matrices).
- We are not restricted to the MST topology and **we can obtain a regularised version of any kind of graph as long as we can compute those matrices**, with $\overline{\boldsymbol{A}} = 1/B \sum_{b=1}^{B} \boldsymbol{A}_b$,

$$(\boldsymbol{A})_{ij} = \left((\boldsymbol{A}_{\mathrm{MST}})_{ij}, (\overline{\boldsymbol{A}}_{>m})_{ij}\right)$$

Turquoise and orange: Persistent homology based prior (Kurlin+15).
Red: MST.
Dark blue: Added edges.

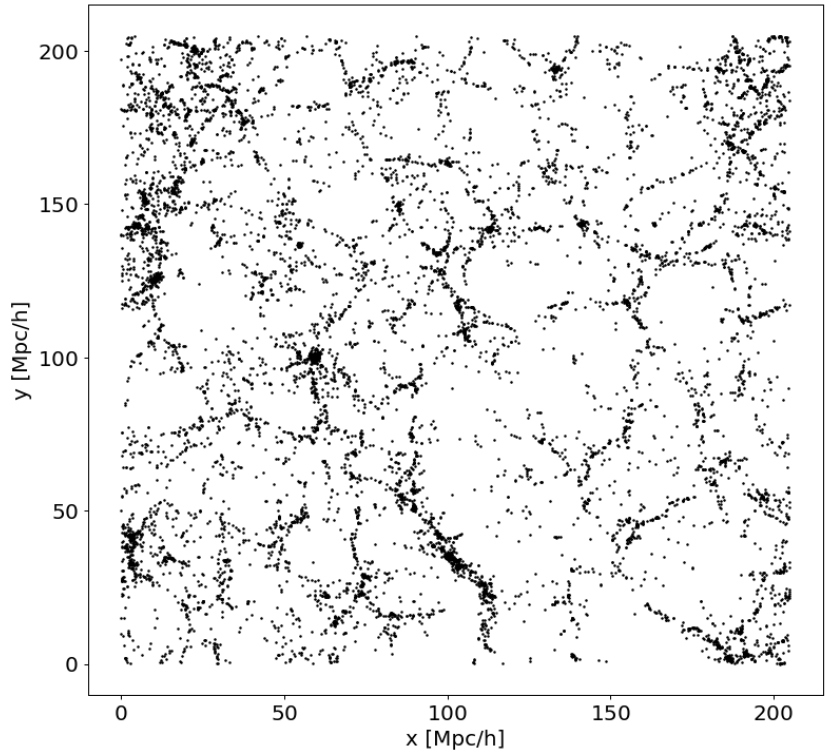


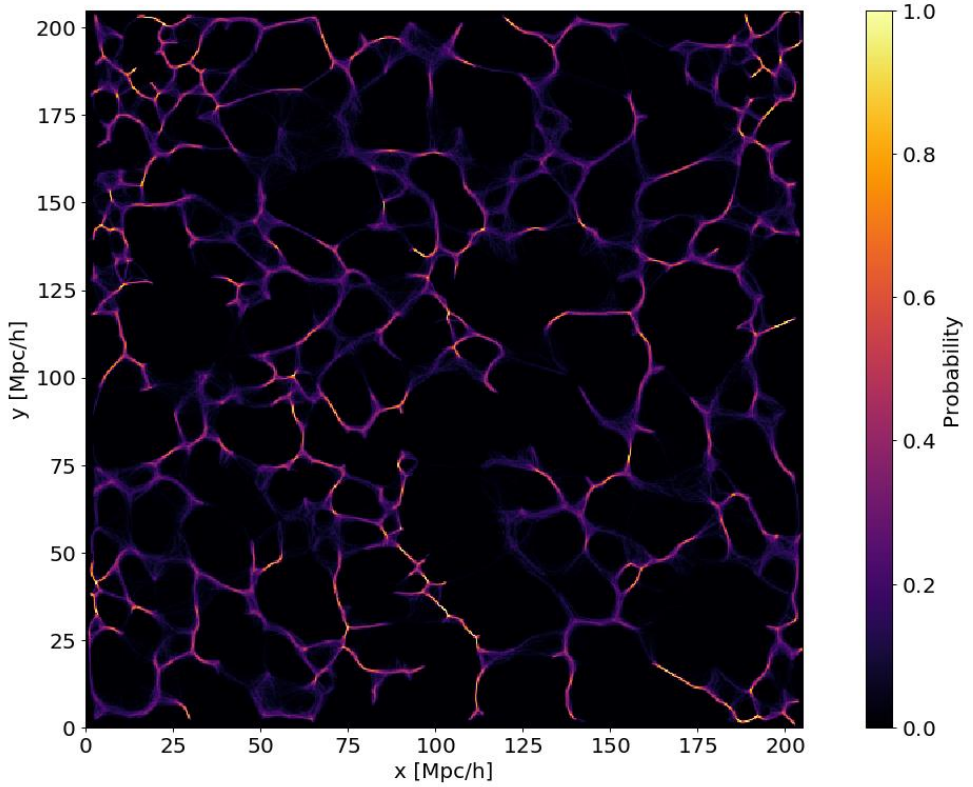


Probability distribution function of $\overline{\boldsymbol{A}}$ values.

Probability maps

- To assess the reliability of the path obtained using the learning and get rid of the tree topology (not allowing cycles), we can use a bootstrap method to get B tree estimates.
- Those trees allow to compute a probability (frequency) to cross a given region of the space during all the realisations





Expectation-Maximisation (1)

- Considering a Mixture Model with K components, we have the log-likelihood

$$\log p(\boldsymbol{X}|\boldsymbol{\Theta}) = \sum_{i=1}^{N} \log \left(\sum_{k=1}^{K} \pi_k \, f(\boldsymbol{x}_i|\boldsymbol{\Theta}) \right).$$

- This log-likelihood cannot be optimised analytically because of the summation occurring inside the logarithm function.
- To alleviate this problem, **EM introduces a set of latent variables** $\boldsymbol{Z} = \{z_i\}_{i=1}^{N}$ that describes the partition of the dataset such that $z_i \in \{1, \dots, K\}$ encodes which one of the K Gaussian component generated the observation $\boldsymbol{x}_i$.
- The new log-likelihood conditioned on $\boldsymbol{Z}$ can be written

$$\log p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\Theta}) = \mathcal{L}(\boldsymbol{\Theta}; \mathbf{X}, \mathbf{Z}) = \sum_{i=1}^{N} \log\left(\pi_{z_i} f(\boldsymbol{x}_i|\boldsymbol{\Theta})\right).$$

- This completed log-likelihood can be optimised more easily an alternating procedure in which we estimate the values of z_i and then maximise the likelihood.

Expectation-Maximisation (2)

- We can re-write the likelihood as a marginal over the latent variables

$$p(\boldsymbol{X}|\boldsymbol{\Theta}) = \sum_{\boldsymbol{Z}} p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\Theta})$$

- For any distribution over the latent variables $q(\boldsymbol{Z})$, we have

$$\log p(\boldsymbol{X}|\boldsymbol{\Theta}) = \log \sum_{\boldsymbol{Z}} q(\boldsymbol{Z}) \left[\frac{p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\Theta})}{q(\boldsymbol{Z})}\right]$$

- **Jensen's inequality** tells us that, from the concavity of the log function and considering $\sum_{\boldsymbol{Z}} q(\boldsymbol{Z}) = 1$,

$$\log p(\boldsymbol{X}|\boldsymbol{\Theta}) \geq \sum_{\boldsymbol{Z}} q(\boldsymbol{Z}) \log \frac{p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\Theta})}{q(\boldsymbol{Z})} = L(q, \boldsymbol{\Theta}).$$

- Hence, $L(q, \boldsymbol{\Theta})$ defines a **lower-bound on the log-likelihood**. Increasing the lower-bound consequently should increase the likelihood.

- Re-writing $L(q, \boldsymbol{\Theta})$ using the decomposition $p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\Theta}) = p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\Theta})p(\boldsymbol{X}|\boldsymbol{\Theta})$, we get that

$$\log p(\boldsymbol{X}|\boldsymbol{\Theta}) = L(q, \boldsymbol{\Theta}) + D_{KL}\big(q(\boldsymbol{Z})||p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\Theta})\big)$$

with $D_{KL}(q||p) = \sum q \log \frac{q}{p} \geq 0$ the **Kullback-Leibler divergence.**

Expectation-Maximisation (3)

- Hence, we know that we need $D_{KL}\big(q||p(\mathbf{Z}|\mathbf{X},\boldsymbol{\Theta})\big)$ to cancel out to maximise the lower-bound hence leading to the **E-step**

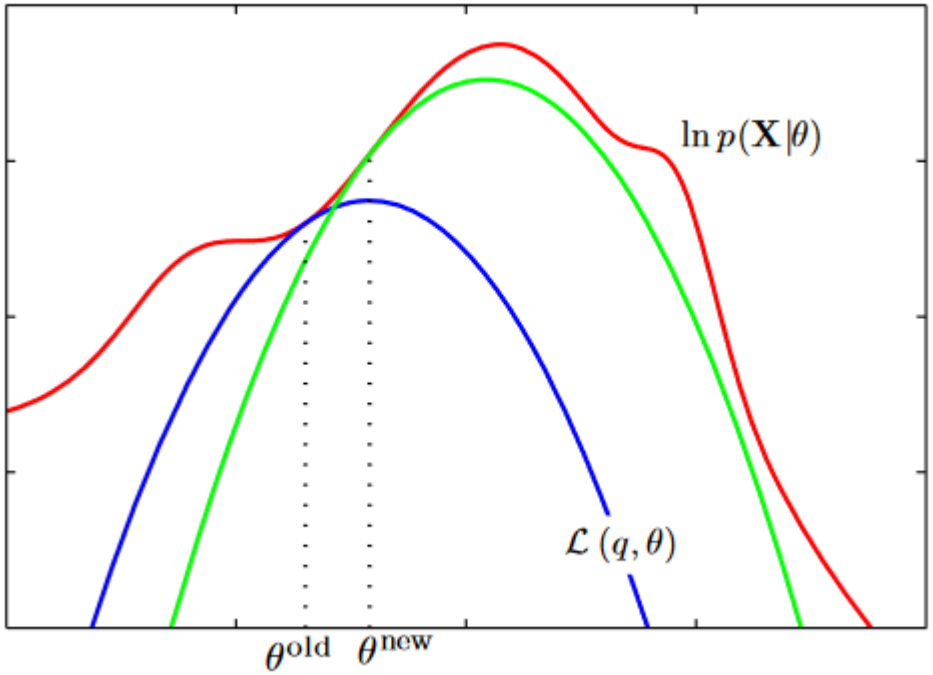
$$\hat{q}(\boldsymbol{Z}) = p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\Theta})$$

- We can then compute the lower-bound, i.e. the expectation over $\hat{q}(\boldsymbol{Z})$, $L(\hat{q},\boldsymbol{\Theta}) = \sum_Z \hat{q}(\boldsymbol{Z}) \log \frac{p(\boldsymbol{X},\boldsymbol{Z}|\boldsymbol{\Theta})}{\hat{q}(\boldsymbol{Z})}$.

- The set of new parameters is obtained by maximising $L(\hat{q},\boldsymbol{\Theta})$ in the **M-step** as

$$\boldsymbol{\Theta}^{\mathrm{t+1}} = \mathrm{argmax}_{\boldsymbol{\Theta}}\, \mathrm{L}(\hat{q},\boldsymbol{\Theta}).$$

- EM has the following theoretical properties:
  1. Monotonic increase of the log-likelihood at each iteration,
  2. Guaranteed convergence towards a local maximum of the log-likelihood (although this maximum is local and can depend on the initialisation),
  3. Computational complexity of $O(NKD)$.



Schematic view of the EM algorithm iteratively finding the maximum of the log-likelihood. Image from Bishop06.

Galaxy classification using the RMM

- The probabilistic formalism of mixture model allows the assignment of each data point (galaxy) to the component, among the $K+1$, that the most probably generated it **(background or node of the graph)**.
- In practice, we use the latent variables $z_i \in \{0, \dots, K+1\}$ the assignment of the galaxy i at position x_i. During the E-step, we compute

$$\forall\, k \in \{0, \dots, K+1\}, \qquad p_{ik} = p(z_i = k \mid \boldsymbol{x}_i, \boldsymbol{\theta}_k)$$

- Hence we can estimate $\hat{\boldsymbol{z}}_{\boldsymbol{i}} = \underset{\boldsymbol{k}}{\mathbf{argmax}}\, \boldsymbol{p}_{\boldsymbol{ik}}$ and get the most probable component.
- Together with the identification of filaments through branches, we can associate each galaxy to a given filament

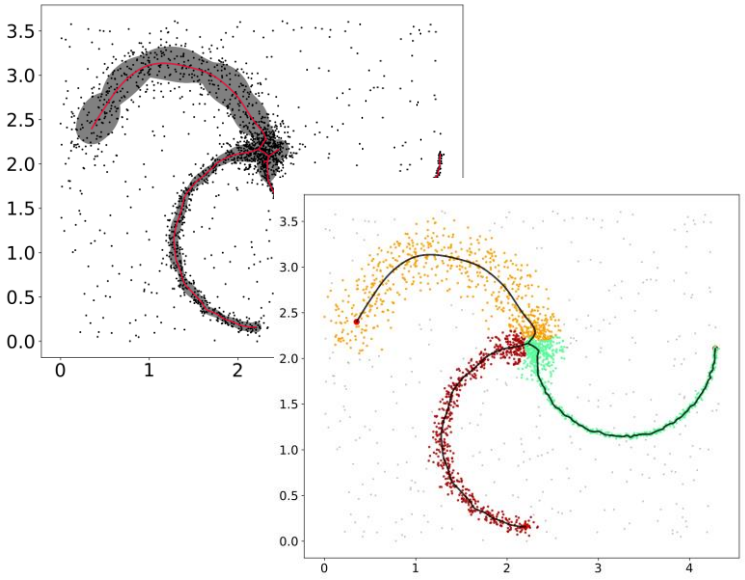


Fig. 11. Toy dataset with 20% background noise (Black dots). The learnt RMST (red) with 3*variances of nodes (grey shaded areas).

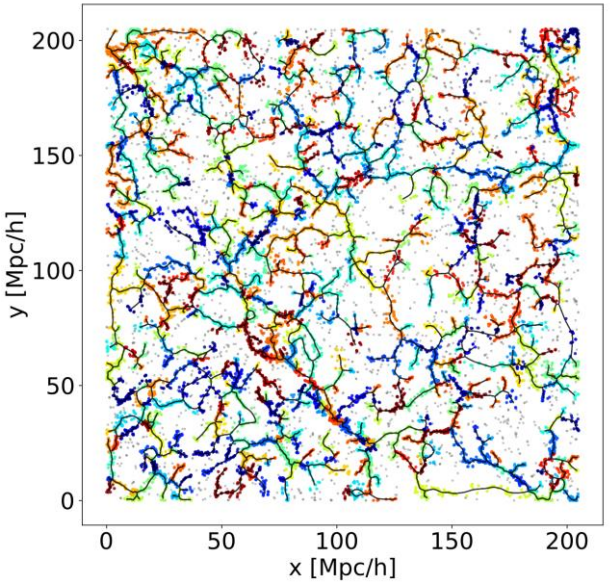


Fig. 12. Datapoints coloured by values of $\hat{z}_i$ with respect to the background (grey points) or to each branch (coloured points).

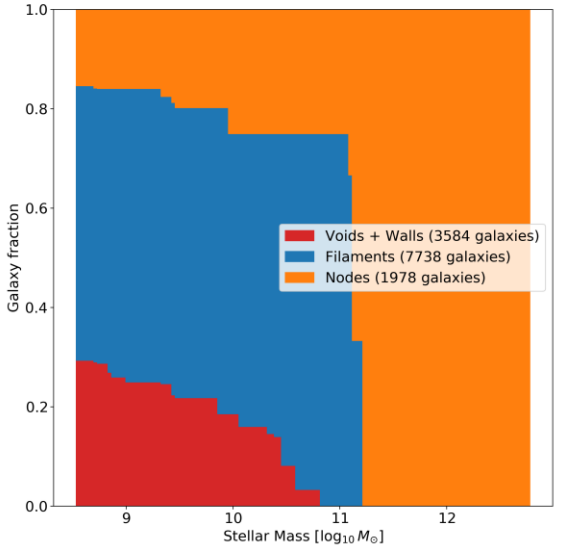


Fig. 13. Proportions of galaxies in each environment of the Eagle simulation identified by T-ReX.

Halo connectivity in the Cosmic Web

- Cluster halos are nodes of the Cosmic Web
- The **halo connectivity increases with its mass** coherent with the hierarchical formation scenario.
- Moreover, we show that **connectivity traces the dynamical state of halos** and this result can be attributed to 2 distinct assembly histories of relaxed and unrelaxed halos (Gouin+21, in prep.).

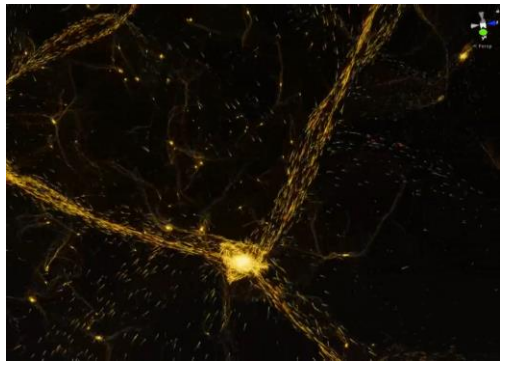


Fig. 14. A node accreting matter through filaments in simulation (Credit: Miguel Aragon-Calvo)

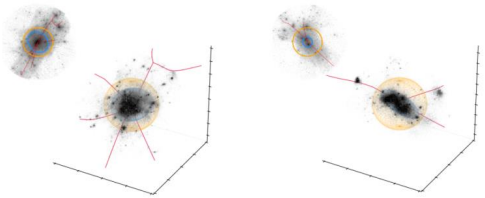


Fig. 15. Illustration of the definition of the connectivity

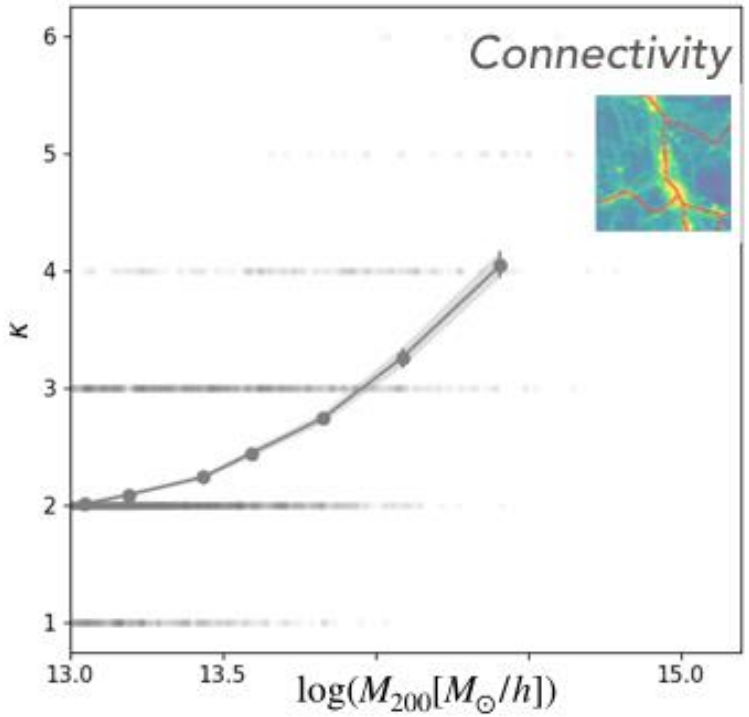


Fig. 16. Evolution of the connectivity κ with the mass of the halo.

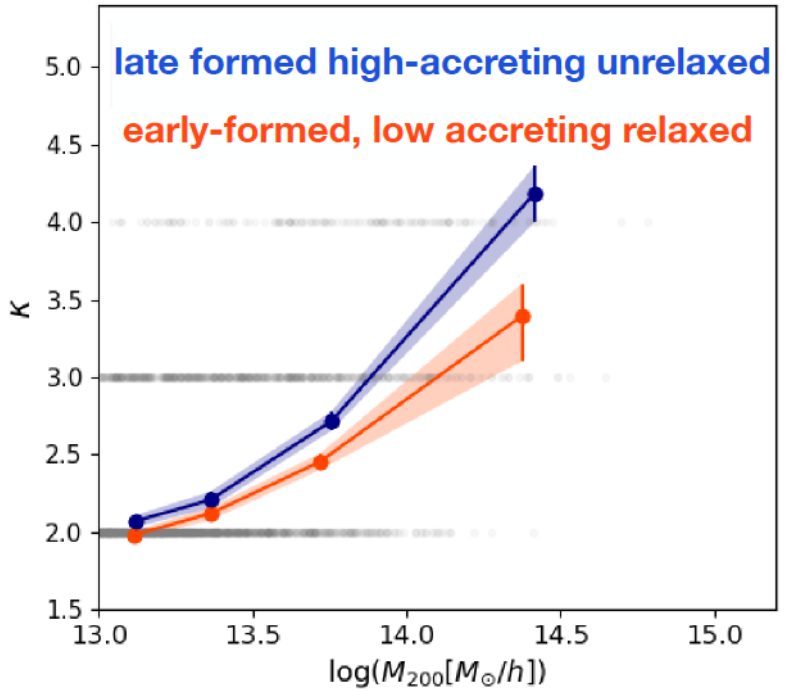


Fig. 17. Evolution of the connectivity with the mass of the cluster for different accretion histories.

Statistical physics formulation

- The clustering problem aims at partitioning a given dataset $\boldsymbol{X} = \{\boldsymbol{x}_i\}_{i=1,\dots,N}$ into K several classes.

- It can be rephrased in statistical physics terms by associating an energy cost to a given configuration (i.e., a set $\boldsymbol{Y} = \{y_i\}_{i=1,\dots,N}$ with $y_i \in \{1, \dots, K\}$ describing the association of the i^{th} datapoint) and a set of clusters' positions $\boldsymbol{\mu}$

$$< E(\boldsymbol{\mu}, \boldsymbol{Y}) > = \sum_{i=1}^{N} \sum_{k=1}^{K} p(y_i = k)\, E_k(\boldsymbol{\mu}_k, \boldsymbol{x}_i)\,.$$

- The free energy can be written, assuming a quadratic cost $E_k(\boldsymbol{\mu}_k, \boldsymbol{x}_i) = \|\boldsymbol{x}_i - \boldsymbol{\mu}_k\|_2^2$, as

$$F = -\frac{1}{\beta} \sum_{i=1}^{N} \ln \sum_{k=1}^{K} e^{-\beta \|\boldsymbol{x}_i - \boldsymbol{\mu}_k\|_2^2}\,.$$

- We see that $F \propto -\log p(\boldsymbol{X}|\boldsymbol{\Theta})$, the log-likelihood of a Gaussian Mixture Model with centre positions $\boldsymbol{\mu}$ and variances $\sigma_k^2 = 1/2\beta$

➔ Maximizing $\log p(\boldsymbol{X}|\boldsymbol{\Theta})$ is equivalent to minimizing the free energy

Reformulation of the problem

- The clustering problem aims at partitioning a given dataset $\boldsymbol{X} = \{\boldsymbol{x}_i\}_{i=1,\dots,N}$ into K several classes.
- It can be rephrased in statistical physics terms by associating an energy cost to a given configuration (i.e., a set $\boldsymbol{Z} = \{z_i\}_{i=1,\dots,N}$ with $z_i \in \{1, \dots, K\}$ describing the association of the i^{th} datapoint) and a set of clusters' parameters $\boldsymbol{\mu}$

$$E(\boldsymbol{\mu}, \boldsymbol{Z}) = \sum_{i=1}^{N} \sum_{k=1}^{K} p(z_i = k)\, E_k(\boldsymbol{\mu}_k, \boldsymbol{x}_i)\,.$$

- The principle of maximum entropy tells us that, among all the probabilities that maximizes E, the Boltzmann distribution is the least informative

$$p(\boldsymbol{\mu}, \boldsymbol{Z}) = \frac{e^{-\beta E(\boldsymbol{\mu}, \boldsymbol{Z})}}{\sum_{\boldsymbol{Z}} e^{-\beta E(\boldsymbol{\mu}, \boldsymbol{Z})}}\,.$$

- Marginalizing to obtain the most probable set of parameters

$$p(\boldsymbol{\mu}) = \sum_{\boldsymbol{Z}} p(\boldsymbol{\mu}, \boldsymbol{Z}) = \frac{Z(\boldsymbol{\mu})}{\sum_{\boldsymbol{\mu}} Z(\boldsymbol{\mu})}$$

- And, in term of free energy F

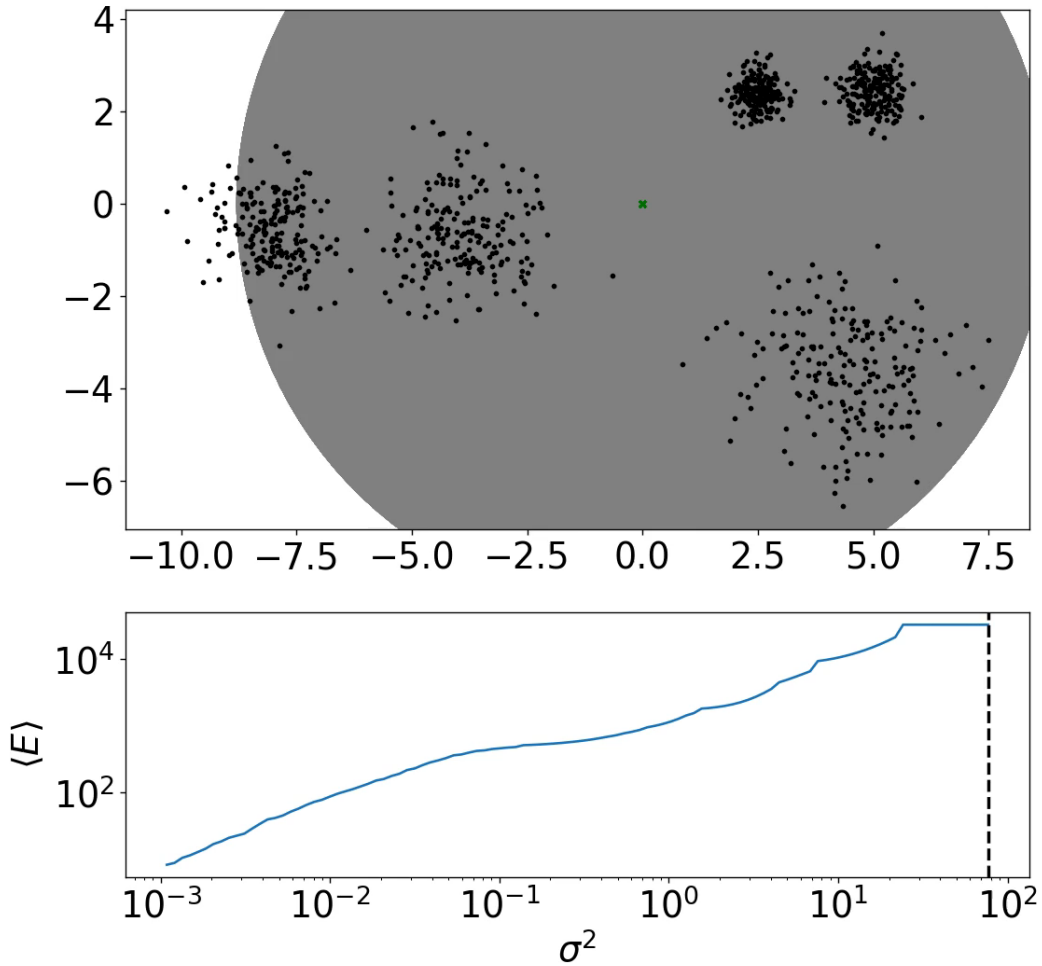
$$p(\boldsymbol{\mu}) = \frac{e^{-\beta F(\boldsymbol{\mu})}}{\sum_{\boldsymbol{\mu}} e^{-\beta F(\boldsymbol{\mu})}}$$

Phase transitions in hard annealing

- Simulated annealing in EM was introduced by Ueda+98 to overcome the problem of dependency on the initialisation.
- The variance plays the role of the temperature.
- When $\boldsymbol{\sigma^2 > T_c^{\mathrm{hard}}}$, all clusters are collapsed at the centre of mass of the dataset.

  ➔ "paramagnetic" phase
- When $\boldsymbol{\sigma^2 < T_c^{\mathrm{hard}}}$, clusters are hierarchically moving towards the centre of mass of sub-datasets.

  ➔ "ferromagnetic" phase
- The study of the linear stability of the fixed-point equations teaches us that $\boldsymbol{T_c^{\mathrm{hard}} = \max \mathrm{eig}\, C}$, with $\boldsymbol{C}$ the data covariance matrix.



Exploiting the phase transitions

Can we learn something from this formulation and more particularly from these successive transitions?

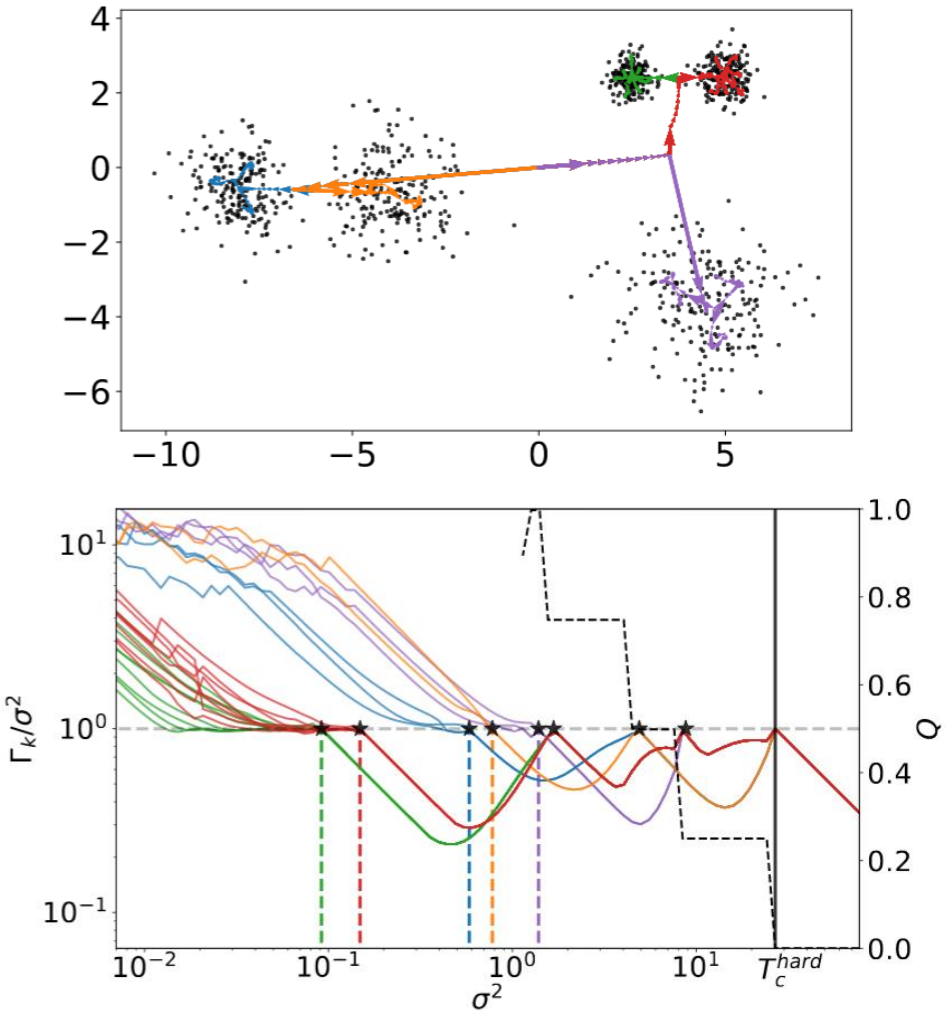
➔ By following, during the annealing, the local size of the represented features, i.e. the maximum eigenvalue Γ_k of the covariance matrix, namely

$$\Sigma_k = \frac{1}{N_k} \sum_{i=1}^{N} p_{ik} \, (\boldsymbol{x}_i - \boldsymbol{\mu}_k)^{\mathrm{T}} (\boldsymbol{x}_i - \boldsymbol{\mu}_k).$$

- The successive transitions can be traced during the annealing and provides **information on the structure of the dataset (number, size and hierarchy of clusters) independently of the number of components K**.
- Interestingly, the overlap between the inferred configuration and the *true* one

$$Q(\boldsymbol{Z}, \widehat{\boldsymbol{Z}}) = \frac{\max_{\pi \in \Pi} \frac{1}{N} \sum_i \delta_{\hat{y}_i, \pi(y_i)} - \frac{1}{q}}{1 - \frac{1}{q}},$$

acts like an order parameter taking increasing values in each phase.



Colors indicate in which final cluster the center ends.

Evolution of the ratio $\frac{\Gamma_k}{\sigma^2}$ as a function of σ^2.

Phase transitions in other settings

Soft annealing for multi-scale clustering

- Relaxing the constraint that, at a given time, all clusters have the same variance leads to a *soft* version of the annealing where we put a prior distribution on the variances (namely an inverse Gamma distribution) whose mode becomes the annealing parameter.
- The threshold temperature T_c^{soft} can still be computed from the (modified) update equations as the maximum eigenvalue of the squared block matrix of order $D+1$ defined as

$$\boldsymbol{M} = \begin{pmatrix} \boldsymbol{C}/\sigma_0^2 & \boldsymbol{a}^{\mathrm{T}} \\ \boldsymbol{b} & c \end{pmatrix},$$

with $\boldsymbol{a}, \boldsymbol{b}, c$ data-related quantities.

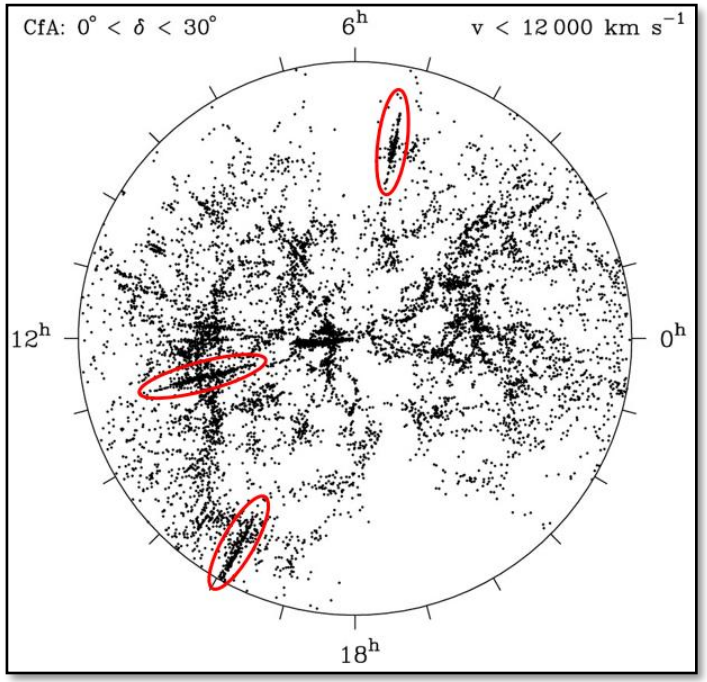
Principal graphs learning

- In the principal graph setting, update equation for $\boldsymbol{\mu}_k$ is modified by the interactions between centres on the graph.
- The temperature T_c^{graph} can be obtained by studying the linear stability of this new update and is given by the spectral radius of

$$\boldsymbol{M} = \left[\left(\boldsymbol{I}_K - \frac{1}{K}\boldsymbol{J}_K\right) \otimes \boldsymbol{C}\right]\left[\sigma^2 \boldsymbol{I}_{KD} + \frac{2\lambda_\mu K \sigma^4}{N} \boldsymbol{L} \otimes \boldsymbol{I}_D\right]^{-1}.$$

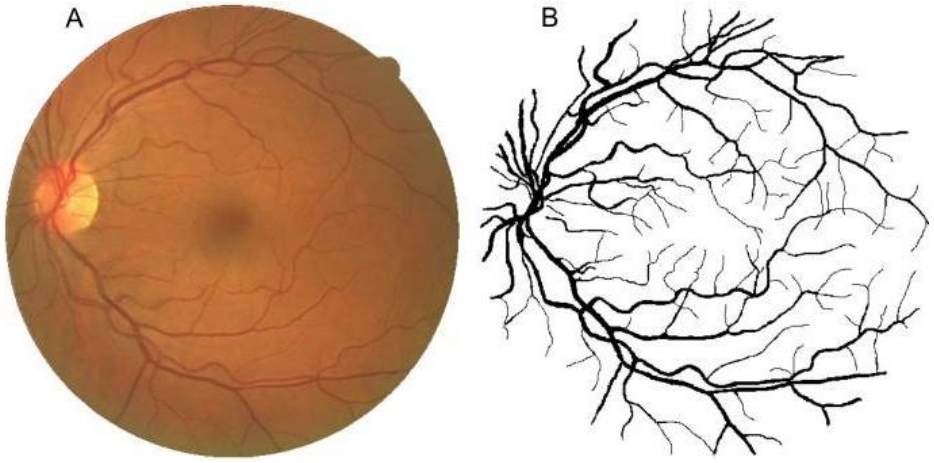
Sources of errors in galaxy surveys

- With real datasets come the usual issues of noise and outliers but in astronomy we add:
  - ✓ **Observational effects:** Redshift-space distortions (Finger-of-God effect), redshift estimation (photo-Z vs spectro), etc.
  - ✓ **Selection effects:** Parts of the sky scanned multiple times, masked region.
- Often considered in pre-processing operations and not by algorithms themselves: fill masked regions with Poisson noise for voids finder, incorporate selection function in DTFE estimate, etc.



Related works in biomedical field

A similar problem: Blood vessels detection



Eye vessels network from DRIVE database (a) Original image (b) Corresponding manual segmentation

Method	Year	Anatomical region	Imaging technique	Image processing method
Feng et al. [20]	2010	Brain	MRA	Unsupervised machine learning (Sec. V-A)
Hassouna et al. [21]	2006	Brain	MRA	
Oliveira et al. [22]	2011	Liver	CT	
Goceri et al. [23]	2017	Liver	MRI	
Bruyninckx et al. [24]	2010	Liver	CT	
Bruyninckx et al. [25]	2009	Lung	CT	
Asad et al. [26]	2017	Retina	CFP	
Mapayi et al. [27]	2015	Retina	CFP	
Sreejini et al. [28]	2015	Retina	CFP	
Cinsdikici et al. [29]	2009	Retina	CFP	
Al-Rawi et al. [30]	2007	Retina	CFP	
Hanaoka et al. [31]	2015	Brain	MRA	Supervised machine learning (Sec. V-B)
Sironi et al. [32]	2014	Brain	Microscopy	
Merkow et al. [33]	2016	Cardiovascular and Lung	CT and MRI	
Sankaran et al. [34]	2016	Coronary	CTA	
Schaap et al. [35]	2011	Coronary	CTA	
Zheng et al. [36]	2011	Coronary	CT	
Nekovei et al. [37]	1995	Coronary	CT	
Smistad et al. [38]	2016	Femoral region, Carotid	US	
Chu et al. [39]	2016	Liver	X-ray fluoroscopic	
Orlando et al. [40]	2017	Retina	CFP	
Dasgupta et al. [41]	2017	Retina	CFP	
Mo et al. [42]	2017	Retina	CFP	
Lahiri et al. [43]	2017	Retina	CFP	
Annunziata et al. [44]	2016	Retina	Microscopy	
Fu et al. [45]	2016	Retina	CFP	
Luo et al. [46]	2016	Retina	CFP	
Liskowski et al. [47]	2016	Retina	CFP	
Li et al. [48]	2016	Retina	CFP	
Javidi et al. [49]	2016	Retina	CFP	
Maninis et al. [50]	2016	Retina	CFP	
Prentasvic et al. [51]	2016	Retina	CT	
Wu et al. [52]	2016	Retina	CFP	
Annunziata et al. [53]	2015	Retina	Microscopy	
Annunziata et al. [54]	2015	Retina	Microscopy	
Vega et al. [55]	2015	Retina	CFP	
Wang et al. [56]	2015	Retina	CFP	
Fraz et al. [57]	2014	Retina	CFP	
Ganin et al. [58]	2014	Retina	CFP	
Orlando et al. [59]	2014	Retina	CFP	
Becker et al. [60]	2013	Retina	CFP	
Rodrigues et al. [61]	2013	Retina	OCT	
Fraz et al. [62]	2012	Retina	CFP	
Zhang et al. [63]	2012	Retina	CFP	
Marin et al. [64]	2011	Retina	CFP	
Lupascu et al. [65]	2010	Retina	CFP	
Salem et al. [66]	2007	Retina	CFP	
Soares et al. [67]	2006	Retina	CFP	
Staal et al. [68]	2004	Retina	CFP	
Lee et al. [69]	2015	Aorta & mesenteric artery	CTA	Edge-based deformable models (Sec. VI-A)
Valencia et al. [70]	2007	Artery	MRA	
Law et al. [71]	2009	Brain & Coronary	MRA & CTA	
Moreno et al. [72]	2013	Coronary	CTA	
Wang et al. [73]	2012	Coronary	CTA	
Cheng et al. [74]	2015	Carotid,Coronary Liver, & Lung		
Zhu et al. [75]	2009	Lung	CTA	
Zhang et al. [76]	2015	Retina	CFP	
Patwardhan et al. [77]	2012	—	US	

Method	Year	Anatomical region	Imaging technique
Klepaczko et al. [78]	2016	Brain	MRA
Tian et al. [79]	2014	Abdomen, Brain, Heart, Lung & Retina	CT, DSA Infrared, US & MRA
Law et al. [80]	2007	Brain	MRA
Wang et al. [81]	2009	Carotid	US
Liang et al. [82]	2015	Liver	Microscopy
Zhao et al. [83]	2015	Retina	CFP & FA
Zhao et al. [84]	2015	Retina	CFP
Wang et al. [85]	2015	Retina	CFP
Xiao et al. [86]	2013	Retina	CFP
Law et al. [87]	2006	Retina	CFP
Robben et al. [88]	2016	Brain	MRA
Rempfler et al. [89]	2015	Brain	MRA
Yureidini et al. [90]	2012	Brain	3DRA
Cetin et al. [91]	2015	Brain	MRA
		Coronary	CTA
Cetin et al. [92]	2013	Brain	MRA
		Coronary	CTA
Shim et al. [93]	2006	Brain	CTA
Cherry et al. [94]	2015	Colon	CTA
Shin et al. [95]	2016	Coronary	FA
Carrillo et al. [96]	2007	Carotid, aorto-iliac	MRA
		Coronary, pulmonary arteries	CTA
Amir-Khalili et al. [97]	2015	Carotid	US
Benmansour et al. [98]	2011	Carotid	CTA
Biesdorf et al. [99]	2015	Coronary	CTA
Lugauer et al. [100]	2014	Coronary	CTA
Tang et al. [101]	2012	Coronary	MR
Wang et al. [102]	2012	Coronary	CTA
Friman et al. [103]	2010	Coronary & Liver	CTA
Li et al. [104]	2009	Coronary	CTA
Wink et al. [105]	2002	Coronary	MRA
Zeng et al. [106]	2017	Liver	CTA
Bauer et al. [107]	2010	Liver	CT
Amir-Khalili et al. [108]	2015	Kidney	Endoscopy images
Amir-Khalili et al. [105]	2002	Kidney	Endoscopic video
Chen et al. [109]	2016	Retina	CFP
Chen et al. [110]	2014	Retina	CFP
Bhuiyan et al. [111]	2013	Retina	CFP
Liao et al. [112]	2013	Retina	CFP
Rouchdy et al. [113]	2013	Retina	CFP
Stuhmer et al. [114]	2013	Retina	CFP
Turetken et al. [115]	2013	Retina	Microscopy
Liao et al. [116]	2012	Retina	CFP
Kaul et al. [117]	2012	Retina	CFP
Delibasis et al. [118]	2010	Retina	CFP
Breitenreicher et al. [119]	2013	—	—
Benmansour et al. [120]	2009	—	—
Wink et al. [121]	2004	—	X-ray

Review of existing procedures in the biomedical field for automatic vessels segmentation *(Moccia+18)*