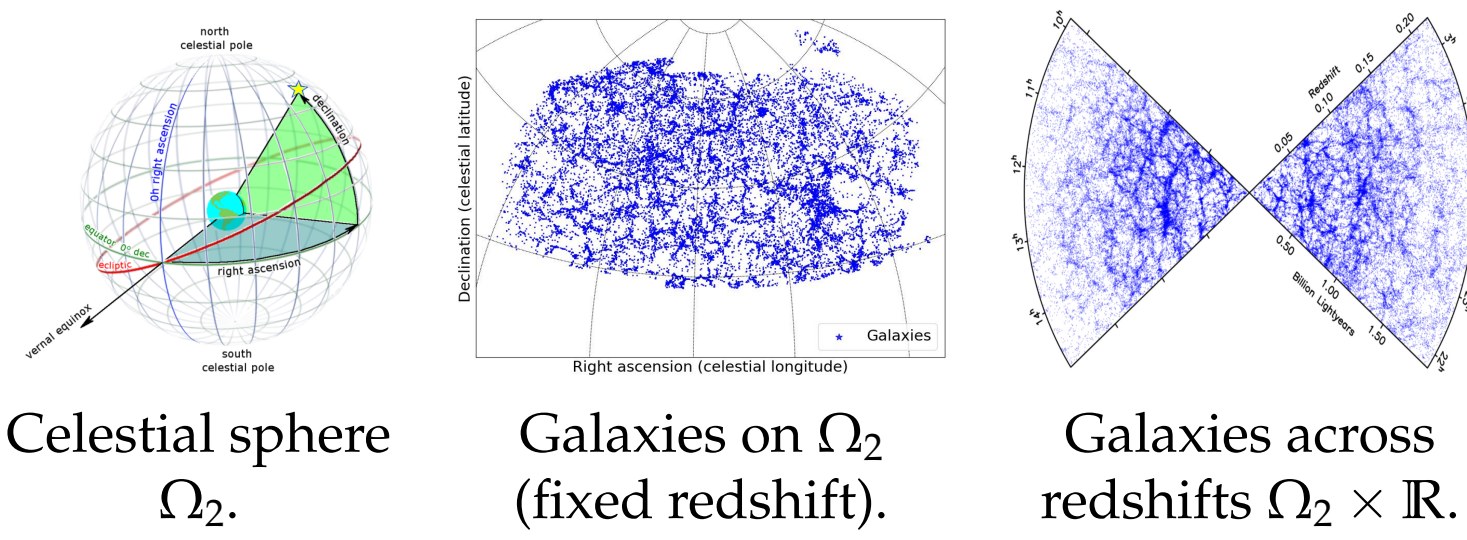


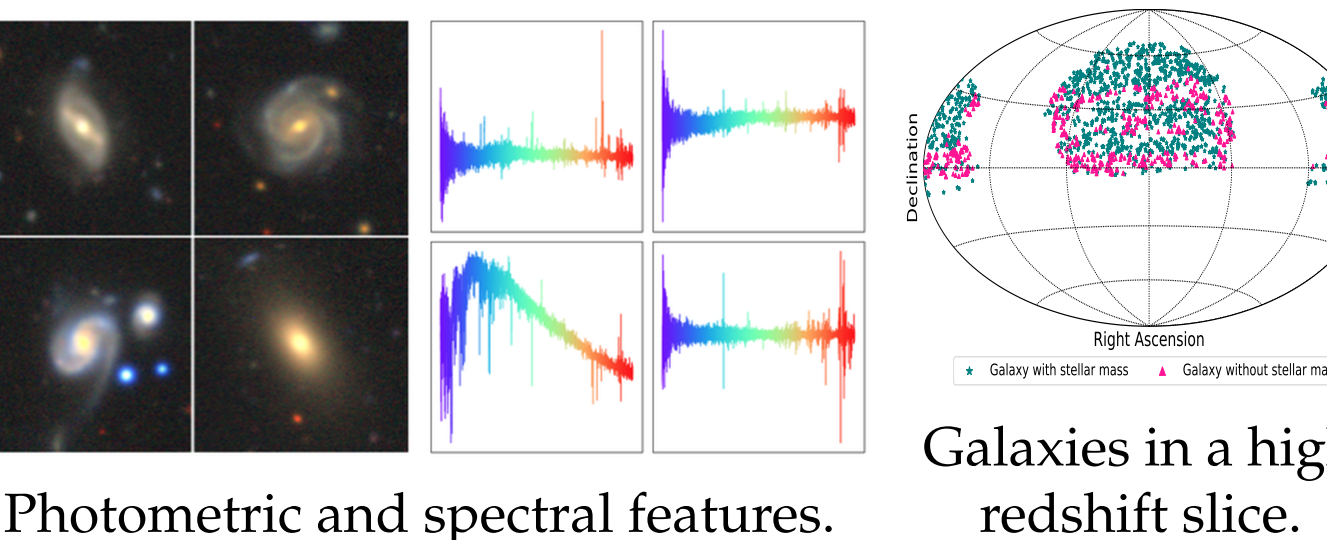


Three Statistical Challenges Behind Cosmic Web Analysis

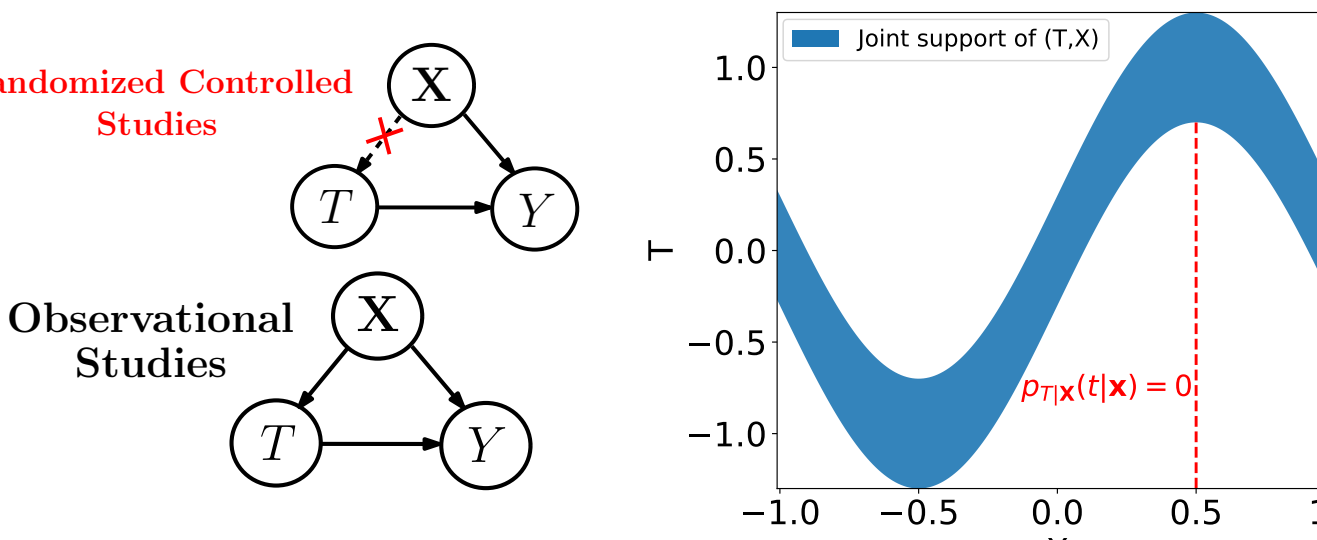
1. Nonlinear Geometry and Low-Dimensional Structures: $\Omega_q = \{x \in \mathbb{R}^{q+1} : \|x\|_2 = 1\}$.



2. High Dimensionality and Missingness: $X \in \mathbb{R}^d$, $d \gg n$, and $Y \in \mathbb{R} \cup \{NA\}$.



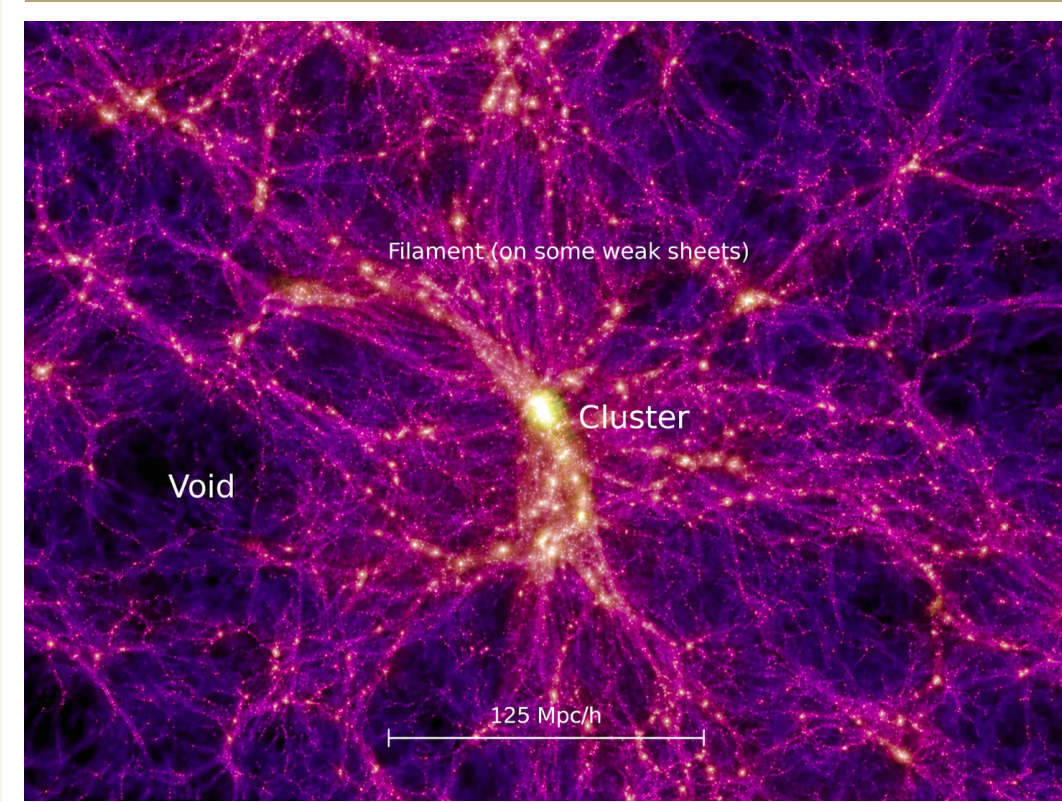
3. Confounding and Positivity Violations: $p_{T|X}(t | x) = 0$ for some $(t, x) \in \mathcal{T} \times \mathcal{X}$.



Main Theme: Develop **geometry-aware** statistical learning and causal inference methods to address these challenges.

Geometry-Aware Cosmic Web Detection

Background and Data



Cosmic Web: The large-scale network structure revealing the non-uniformity of matter distribution in the Universe.

- 0D galaxy clusters (*cosmic nodes*);
- 1D filaments connecting clusters;
- 2D tenuous sheets or walls;
- 3D nearly empty cosmic voids.

We focus on **cosmic nodes and filaments**: key tracers of matter concentration, galaxy evolution, and weak lensing.

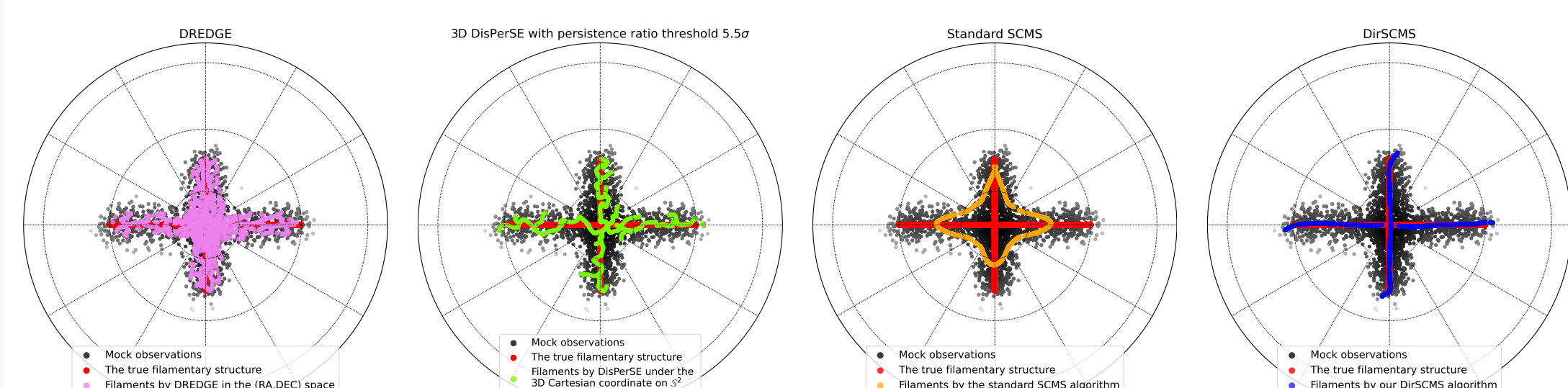
Observed Galaxy Data: Angular coordinates and redshifts $\{(\alpha_i, \delta_i, z_i)\}_{i=1}^n \subset \Omega_2 \times \mathbb{R}^+$.

- $\alpha_i \in [0^\circ, 360^\circ]$: *right ascension* (RA; celestial longitude);
- $\delta_i \in [-90^\circ, 90^\circ]$: *declination* (DEC; celestial latitude);
- $z_i \geq 0$: *redshift*, which provides distance information.

Existing Approaches and Limitations

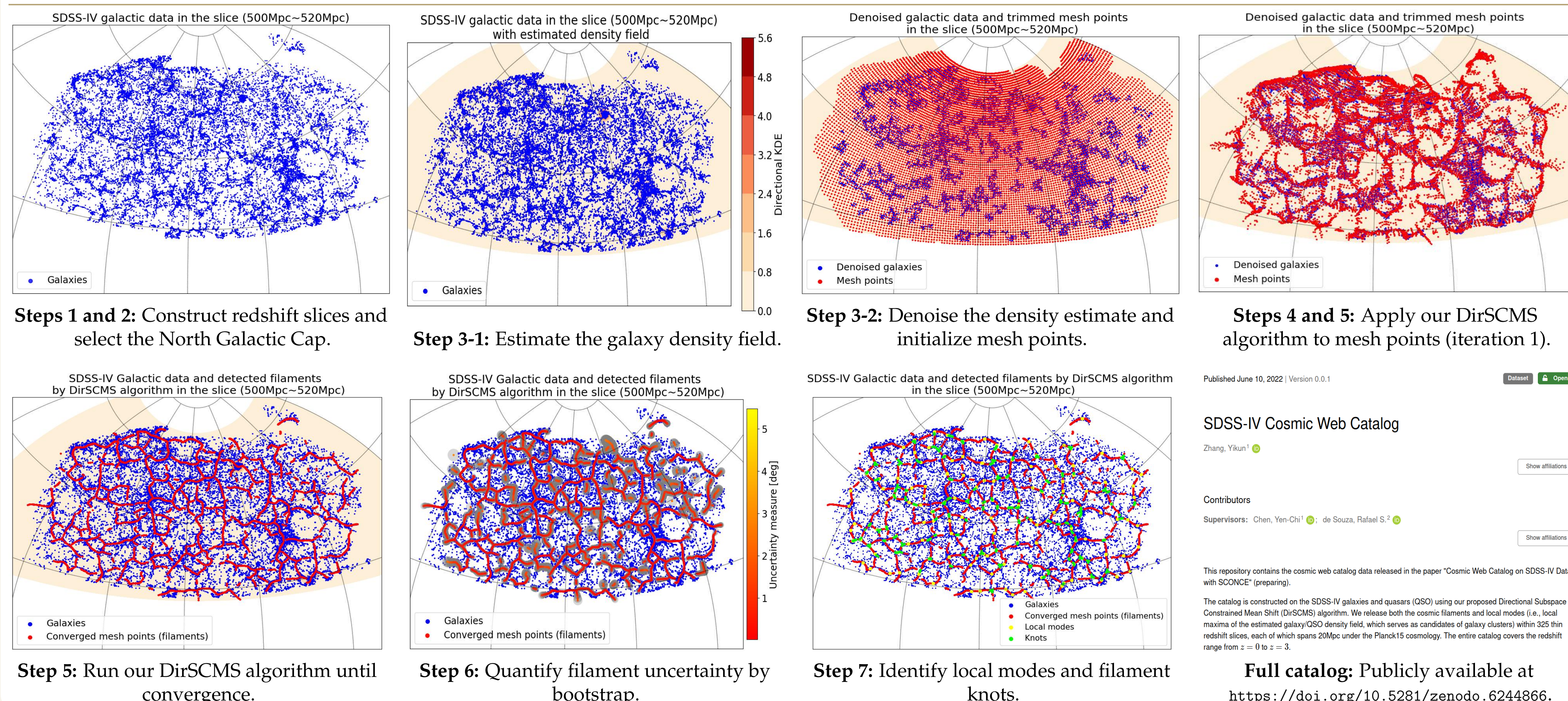
3D Approach: Convert redshifts to comoving distances and detect filaments in 3D.
► Requires a cosmological distance model and is sensitive to “Finger-of-God” distortions.

2D Approach: Divide the survey by redshift and detect filaments within each 2D slice.
► Existing methods rely on a flat-sky approximation and neglect cosmic expansion.



Our method supports both filament detection on 2D spherical slices and an extension to full 3D light-cone analysis.

Cosmic Web Detection in an Example Redshift Slice



Proposed Model, Estimation, and Algorithm

Step 1 (Redshift Slices): Partition the Sloan Digital Sky Survey (SDSS-IV) into 325 non-overlapping slices of comoving thickness $\Delta L = 20$ Mpc (*tomographic analysis*).

Step 2 (Survey Regions): Retain the selected *North Galactic Cap* and *South Galactic Cap* regions in each slice.

Step 3 (Galaxy Density): Map each galaxy’s angular coordinates $(\alpha_i, \delta_i) \in [0, 360^\circ] \times [-90^\circ, 90^\circ]$, $i = 1, \dots, n$ to Ω_2 :

$$X_i = (\cos \delta_i \cos \alpha_i, \cos \delta_i \sin \alpha_i, \sin \delta_i), \quad i = 1, \dots, n.$$

Fit a directional kernel density estimate (DirKDE):

$$f(x) = \frac{C(b)}{n} \exp \left(\frac{x^T X_i}{b^2} \right),$$

where $b > 0$ is the bandwidth and $C(b) > 0$ normalizes the density.

Step 4 (Filament Model): Represent the 1D cosmic filaments on Ω_2 as directional density ridges of f :

$$R_1(f) = \{x \in \Omega_2 : v_2(x)^T \text{grad } f(x) = 0, \lambda_2(x) < 0\},$$

where $\text{grad } f(x)$ is the *Riemannian gradient* of f at $x \in \Omega_2$, and $v_2(x)$ is the tangent-space eigenvector of the *Riemannian Hessian* $\mathcal{H}f(x)$ corresponding to its smaller eigenvalue $\lambda_2(x)$.

Step 5 (SCONCE Algorithm): Compute $R_1(f)$ using directional subspace constrained mean shift (DirSCMS), implemented in our filament finder SCONCE, until convergence ($t = 0, 1, \dots$):

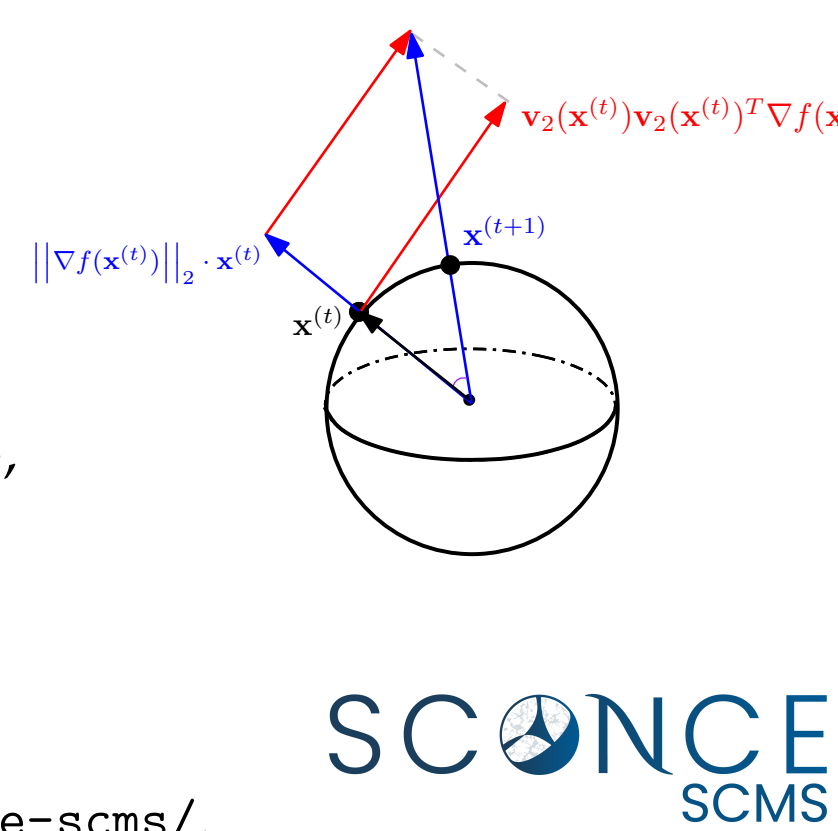
$$\tilde{x}^{(t+1)} \leftarrow x^{(t)} + v_2(x^{(t)}) v_2(x^{(t)})^T \frac{\sum_{i=1}^n X_i \exp \left(\frac{X_i^T x^{(t)}}{b^2} \right)}{\left\| \sum_{i=1}^n X_i \exp \left(\frac{X_i^T x^{(t)}}{b^2} \right) \right\|_2},$$

$$x^{(t+1)} \leftarrow \frac{\tilde{x}^{(t+1)}}{\|\tilde{x}^{(t+1)}\|}.$$

► Python package: <https://pypi.org/project/sconce-scms/>.

Step 6 (Uncertainty): Quantify the local filament uncertainties by nonparametric bootstrap.

Step 7 (Nodes and Filament Knots): Identify local modes of f using directional mean shift algorithm, and detect filament intersections via metric graph reconstruction.

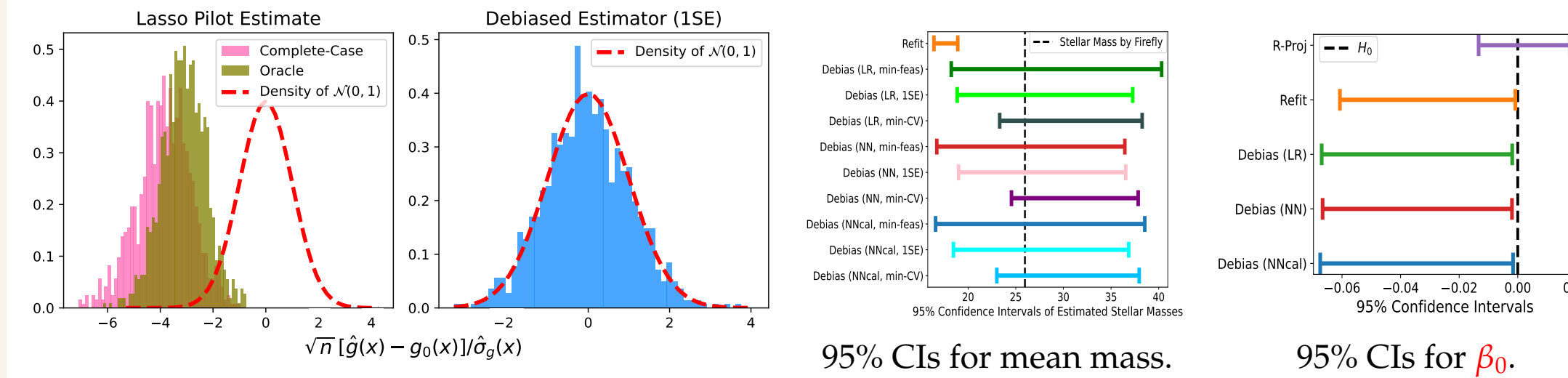


High-Dimensional Inference on the Cosmic Web

Question: Are galaxies farther from filaments *significantly less massive* after adjustment for observed features?

Challenges and Assumptions

- Missing Outcomes:** Stellar mass is unobserved for some galaxies.
► *Missing at random:* $Y \perp\!\!\!\perp R \mid (D_{\text{fila}}, X)$, where $R = 1$ if Y is observed and $R = 0$ otherwise.
- High-dimensional Covariates:** Each galaxy has many observed features and nonlinear transformations.
► *Sparse linear model:* $Y = \beta_0 D_{\text{fila}} + \beta_1^T X + \epsilon$, with $\|\beta_1\|_0 = s_\beta \ll d$, $\mathbb{E}(\epsilon \mid D_{\text{fila}}, X) = 0$, and $\text{Var}(\epsilon \mid D_{\text{fila}}, X) = \sigma_\epsilon^2$.



Finding: The proposed 95% confidence intervals (CIs) for β_0 lie below zero: galaxies farther from nearby filaments tend to have lower stellar masses after adjustment.

Causal Inference for Continuous Treatments Under Positivity Violations

Causal Setup and Estimand

- **Treatment** $T \in \mathcal{T} \subset \mathbb{R}$: Local environmental density of a galaxy.
- **Covariates** $X \in \mathcal{X} \subset \mathbb{R}^d$: Observed treatment–outcome confounders.
- **Outcome** $Y \in \mathcal{Y} \subset \mathbb{R}$: Stellar mass or another galaxy property.
- **Potential Outcome** $Y^{(u_0)} \in \mathcal{Y}$: Outcome if treatment were set to t_0 .

Causal Estimands:

- Dose-response curve $t_0 \mapsto m(t_0) := \mathbb{E}[Y^{(t_0)}]$;
- Derivative effect curve $t_0 \mapsto \theta(t_0) = m'(t_0) = \frac{d}{dt} \mathbb{E}[Y^{(t_0)}]$.

Issue: When positivity fails at t_0 with $p_{T|X}(t_0|x) = 0$ for units with covariates x , $m(t_0)$ and $\theta(t_0)$ need not be *identifiable* without additional assumptions.

Route A: Impose Structural Assumptions

Assume **consistency** ($Y = Y^{(T)}$) and **conditional ignorability** ($Y^{(t)} \perp\!\!\!\perp T | X$), and the following condition:

$$\begin{aligned} \theta(t) &= \frac{d}{dt} \mathbb{E}[Y^{(t)}] = \mathbb{E}[\partial_t \mathbb{E}[Y^{(t)} | X]] \\ &\stackrel{+}{=} \mathbb{E}[\partial_t \mathbb{E}[Y^{(t)} | X] | T = t] \\ &\stackrel{**}{=} \mathbb{E}[\partial_t \mathbb{E}(Y | T = t, X) | T = t], \end{aligned}$$

where “ $\stackrel{+}{=}$ ” is the key restriction; (**) uses ignorability and consistency.

By the fundamental theorem of calculus, $m(t_0) = \mathbb{E}[Y + \int_{t_0}^t \theta(u) du]$.

► **Regression Adjustment (RA) Estimator:** Write $\mu(t, x) = \mathbb{E}(Y \mid T = t, X = x)$ and $\beta(t, x) = \partial_t \mu(t, x)$. Then,

$$\hat{\theta}_{\text{C,RA}}(t) = \int \hat{\beta}(t, x) d\hat{F}_{X|T}(x|t) \quad \text{and} \quad \hat{m}_{\text{C,RA}}(t) = \frac{1}{n} \sum_{i=1}^n \left[Y_i + \int_{t_0}^{t_1} \hat{\theta}_{\text{C,RA}}(\bar{t}) d\bar{t} \right].$$

► **Inverse Probability Weighting (IPW) Estimator:**

$$\hat{\theta}_{\text{C,IPW}}(t) = \frac{1}{nh^2} \sum_{i=1}^n \frac{Y_i \cdot \left(\frac{T_i - t}{h} \right) K \left(\frac{T_i - t}{h} \right) \hat{p}_\zeta(X_i | t)}{\kappa_2 \cdot \hat{p}(T_i, X_i)},$$

where $\hat{p}_\zeta$ estimates the ζ -interior conditional density function $p_\zeta(x|t)$.

► **Doubly Robust (DR) Estimator:** Write $K : \mathbb{R} \rightarrow [0, \infty)$ as a smoothing kernel, and $\kappa_2 = \int u^2 K(u) du$. Then,

$$\hat{\theta}_{\text{C,DR}}(t) = \frac{1}{nh^2} \sum_{i=1}^n \frac{\left(\frac{T_i - t}{h} \right) K \left(\frac{T_i - t}{h} \right) \hat{p}_\zeta(X_i | t)}{\kappa_2 \cdot \hat{p}(T_i, X_i)} \left[Y_i - \hat{\mu}(t, X_i) - (T_i - t) \cdot \hat{\beta}(t, X_i) \right] + \int \hat{\beta}(t, x) \cdot \hat{p}_\zeta(x|t) dx.$$

Limitation of Route A: The sufficient model for “ $\stackrel{+}{=}$ ” implicitly assumes a *homogeneous* treatment effect: $Y^{(t)} = \bar{m}(t) + \xi(X) + \epsilon$, with independent, mean-zero ϵ .

Route B: New Estimand Under a Feasible Intervention

For the nonempty and closed conditional support $\mathcal{T}(x)$ of $p_{T|X}(t|x)$, assume a unique projection almost surely:

$$t_0^*(x) := \arg \min_{t \in \mathcal{T}(x)} |t - t_0|.$$

The *Nearest feasible treatment policy (NFTP)* estimand is defined as

$$\bar{m}_0(t_0) := \mathbb{E}[Y^{(t_0^*)}] \stackrel{\Delta}{=} \int_{\mathcal{X}(t_0)} \mu(t_0, x) dP_X(x) + \int_{\mathcal{X} \setminus \mathcal{X}(t_0)} \mu(t_0^*(x), x) dP_X(x)$$

► $Y^{(t_0^*)}$: potential outcome under the dynamic treatment regime $x \mapsto t_0^*(x)$.

► (Δ) : identification without global positivity, and $\mathcal{X}(t_0)$: the support of $p_{X|T}(x|t_0)$.

► $\bar{m}_0(t_0)$ equals the classical dose-response curve $t_0 \mapsto m(t_0)$ under positivity.

Debiased Inference Procedure

Write $Z_i = (D_{\text{fila},i}, X_i^T)^T$, $z = (d_{\text{fila}}, x^T)^T$, and $\beta = (\beta_0, \beta_1^T)^T \in \mathbb{R}^{d+1}$.

Step 1: Compute the **Lasso pilot estimate** with complete-case data:

$$\hat{\beta} \in \arg \min_{\beta \in \mathbb{R}^{d+1}} \left\{ \frac{1}{n} \sum_{i=1}^n R_i (Y_i - Z_i^T \beta)^2 + \lambda \|\beta_1\|_1 \right\}, \quad \lambda > 0.$$

Step 2: Obtain **consistent estimates of the observation probabilities** $\hat{\pi}_i = \hat{P}(R_i = 1 | Z_i)$ for $i = 1, \dots, n$ by any machine learning method fitted to $\{(Z_i, R_i)\}_{i=1}^n$.

Step 3: Solve the convex **debiasing program**:

$$\hat{w}(z) \in \arg \min_{w \in \mathbb{R}^{d+1}} \left\{ \sum_{i=1}^n \hat{\pi}_i \hat{w}_i^T : \left\| z - \frac{1}{\sqrt{n}} \sum_{i=1}^n w_i \hat{\pi}_i Z_i \right\|_\infty \leq \frac{\gamma}{n} \right\}, \quad \gamma > 0.$$

Step 4: Construct the **debiased estimator** of $g_0(z) = z^T \beta$ as:

$$\hat{g}^{\text{debias}}(z) = z^T \hat{\beta} + \frac{1}{\sqrt{n}} \sum_{i=1}^n \hat{w}_i(z) R_i (Y_i - Z_i^T \hat{\beta}).$$

Taking $z = e_1 = (1, 0, \dots, 0)^T \in \mathbb{R}^{d+1}$ targets β_0 .

Theory: Under regularity conditions, if $\sigma_\zeta^2(z) = \lim_{n \rightarrow \infty} \sigma_\epsilon^2 z^T [\mathbb{E}(RZZ^T)]^{-1} z \in (0, \infty)$ exists,

$$\sqrt{n} [\hat{g}^{\text{debias}}(z) - g_0(z)] \xrightarrow{d} \mathcal{N}(0, \sigma_\zeta^2(z)).$$

Stochastic Tilted Intervention to NFTP

Approach the NFTP intervention through a **positivity-lean point-targeted** (PLPT) tilted intervention:

$$q_{h,t_0}(t|x) := \frac{\eta(t; t_0, h) p_{T|X}(t|x)}{\int_T \eta(u; t_0, h) p_{T|X}(u|x) du}.$$

- $t \mapsto \eta(t; t_0, h) \propto \exp \left[-\left| \frac{t - t_0}{h} \right|^\gamma \right]$ for some $\gamma > 0$;
- $h > 0$ governs concentration around t_0 .
- $Q_{h,t_0}(\cdot | x) \rightarrow \delta_{t_0^*(x)}(\cdot)$ as $h \rightarrow 0$.

Tilted Intervention Estimand: $t_0 \mapsto \bar{m}_h(t_0) := \mathbb{E}[Y^{(q_{h,t_0})}]$.

► $Y^{(q_{h,t_0})}$: the outcome under treatment drawn from $q_{h,t_0}(t|x)$ for each unit $X = x$.

► $\bar{m}_h(t_0) \rightarrow \bar{m}_0(t_0)$ recovers the NFTP estimand as $h \rightarrow 0$.

One-Step Estimator Based on the Efficient Influence Function:

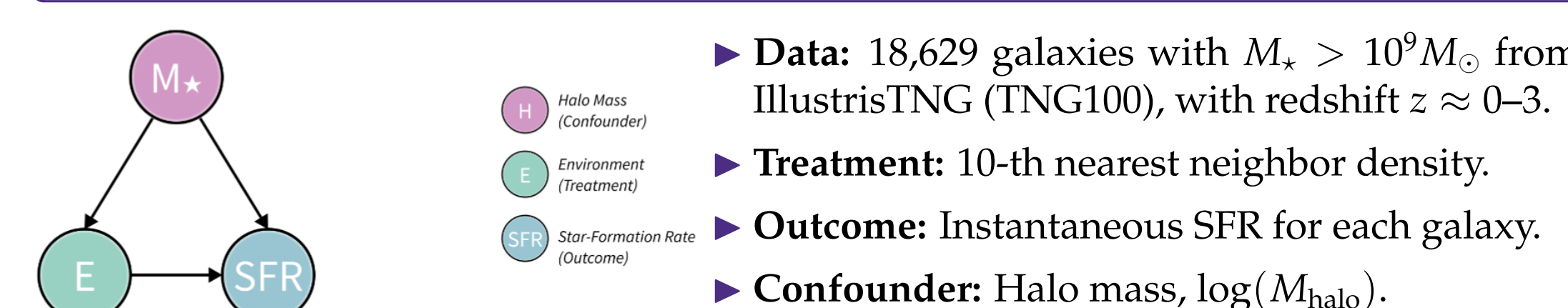
$$\hat{m}_h(t_0) = \frac{1}{n} \sum_{i=1}^n \left\{ \frac{\hat{q}_{h,t_0}(T_i | X_i)}{\hat{p}_{T|X}(T_i | X_i)} \left[Y_i - \int_T \hat{\mu}(t, X_i) \hat{q}_{h,t_0}(t | X_i) dt \right] + \int_T \hat{\mu}(t, X_i) \hat{q}_{h,t_0}(t | X_i) dt \right\}.$$

Asymptotic Normality: Under regularity conditions with $n \rightarrow \infty$ and $h \rightarrow 0$,

$$\sqrt{\frac{n}{V_{\text{eff}, \bar{m}_h(t_0)}}} [\hat{m}_h(t_0) - \bar{m}_0(t_0)] \xrightarrow{d} \mathcal{N}(0, 1).$$

“Nature vs. Nurture” in Galaxy Evolution

Question: How does local environment (“nurture”) affect a galaxy’s star formation rate (SFR) after adjustment for its halo mass (“nature”)?



- **Data:** 18,629 galaxies with $M_h > 10^9 M_\odot$ from IllustrisTNG (TNG100), with redshift $z \approx 0-3$.
- **Treatment:** 10-th nearest neighbor density.
- **Outcome:** Instantaneous SFR for each galaxy.
- **Confounder:** Halo mass, $\log(M_{\text{halo}})$.

