Identifying and Exploiting Structure in the Landscape

Alex Cole (University of Wisconsin)
June 11, 2020
String Phenomenology 2020

Based on 1712.08159 (JCAP),
1812.06960 (JHEP),
1907.10072 (JHEP),
2006.xxxx, 200x.xxxx

with collaborators:
Matteo Biagetti,
Andreas Schachner,
Gary Shiu

Outline

1. Motivation: Big Data in cosmology and string theory
2. Topological data analysis: persistent homology
    1. Cosmology: CMB, LSS

} "identify" structure

3. Stochastic optimization: genetic algorithms
    1. Flux vacua
    2. Fitness-distance correlation and encodings

} "exploit" structure

4. Summary, future work

1. Motivation

Big Data in Cosmology

Big Data in Cosmology

Experimental Data	2013	2020	2030+
Storage	1PB	6PB	100-1500PB
Cores	10^3	70K	300+K
CPU hours	3x10^6 hrs	2×10^8 hrs	$\sim 10^9$ hrs
Simulations	2013	2020	2030+
Storage	1-10 PB	10-100PB	> 100PB - 1EB
Cores	0.1-1M	10-100M	> 1G
CPU hours	200M	>20G	> 100G

Table I: Estimated compute and storage needs for the next 10-20 years of Cosmic Frontiers simulations and experiments..

1311.2841

	data volume	schedule
SDSS	40 TB	2000-2020
DESI	2 PB	2019-2027
LSST	> 60 PB	2020-2030
Euclid	>10 PB	2020-2027
WFIRST	>2 PB	2023-2030
CMB-S4	O(1) PB	2020-2027(?)
SKA	4.6 EB	2019-2030(?)

- This data lives in **high-dimensional** phase space
  - N particles: $\mathbf{dim} \sim N_{\mathrm{particles}}$
  - Function on sphere: $\mathbf{dim} \sim N_{\mathrm{pix}} \sim \ell^2_{\mathrm{max}}$
- Theorist's job: **predict** and **recover** signals to constrain theory space

Fluctuations

- The lowest order correlation in primordial fluctuations is the **power spectrum**

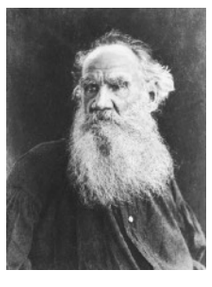
$$\langle 0 \left| \, \zeta_{\mathbf{k_1}} \zeta_{\mathbf{k_2}} \, \right| 0 \rangle = (2\pi)^3 \delta^3(\mathbf{k_1} + \mathbf{k_2}) P_\zeta(k_1)$$

- For a **Gaussian** theory, the power spectrum $P_\zeta(k)$ dictates all higher-point correlations. **Huge compression** of data to the power spectrum: $\ell_{\rm max}$ d.o.f.

- The leading **non-Gaussianity** is the **bispectrum**:

$$\langle 0 \left| \, \zeta_{\mathbf{k_1}} \zeta_{\mathbf{k_2}} \zeta_{\mathbf{k_3}} \, \right| 0 \rangle = (2\pi)^3 \delta^3(\mathbf{k_1} + \mathbf{k_2} + \mathbf{k_3}) B_\zeta(k_1, k_2, k_3)$$

- If scale-invariant, we can characterize a bispectrum by its size $\sim f_{\rm NL}$ and shape $\sim B_\zeta\left(1, \frac{k_2}{k_1}, \frac{k_3}{k_2}\right)$



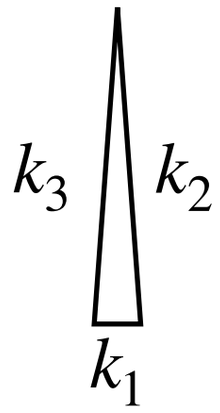
Tolstoy: "Gaussian fields* are all alike; every non-Gaussian field is non-Gaussian in its own way"

Some Non-Gaussianities

- Take **local** ansatz **[Komatsu, Spergel]**

$$\zeta(\mathbf{x}) = \zeta_g(\mathbf{x}) + \frac{3}{5} f_{\rm NL} \left(\zeta_g(\mathbf{x})^2 - \left\langle \zeta_g(\mathbf{x})^2 \right\rangle \right)$$

  - Bispectrum peaks in **squeezed limit** $k_1 \ll k_2 \sim k_3$



  - Single-field consistency relation: $B_{k_1 \ll k_2 \sim k_3} \propto (1 - n_s)$

[Creminelli, Zaldarriaga]

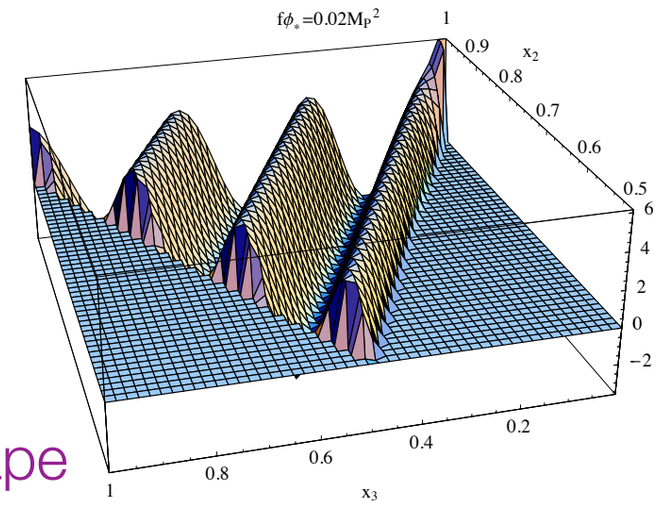
    - Detection in squeezed limit would rule out all single-field inflationary models!

- Other shapes: equilateral $\sim \mathcal{O}\left(c_s^{-2} - 1\right)$, orthogonal, **resonant**…

Measuring Non-Gaussianity

- **Harmonic space**: fit with *templates* of bispectrum. What is your template is wrong? How well you do depends on "cosine" between distributions:

$$C(S_1, S_2) = \frac{S_1 \cdot S_2}{\sqrt{S_1 \cdot S_1}\sqrt{S_2 \cdot S_2}}$$



- Some shapes are harder to find, e.g.,

Resonant shape
(axion monodromy)

progress: **[Münchmeyer, Meerburg, Wandelt]**

- Current 1σ bounds on NG (Planck '18):

$$f_{\rm NL}^{\rm loc} = -0.9 \pm 5.1; f_{\rm NL}^{\rm equil} = -26 \pm 47; f_{\rm NL}^{\rm ortho} = -38 \pm 24$$

This talk: real-space topological observables via **persistent homology**

Big Data in String Theory

Type IIB flux vacua

- **[Giddings,Kachru,Polchinski]** type IIB on CY_3.
  Flux superpotential **[Gukov,Vafa,Witten]** stabilizes axiodilaton and complex structure moduli.

$$G_3 = F_3 - \phi H_3$$

$$W = \int_M G_3 \wedge \Omega$$

$$D_a W = 0 \rightarrow \langle \phi \rangle, \langle z^a \rangle$$

- Flux quantization and tadpole cancellation: finite number of solutions.
  - "Typical" number $\sim 10^{500}$

$$F_3 = \begin{pmatrix} f_1 \\ \vdots \\ f_{2(h^{2,1}+1)} \end{pmatrix} \quad H_3 = \begin{pmatrix} h_1 \\ \vdots \\ h_{2(h^{2,1}+1)} \end{pmatrix}$$

- A hard problem: what is the smallest $L_{\max}$ such that there exists a vacuum with vevs, masses, etc. in a certain range?

$$0 < N_{\text{flux}} = f^T \cdot \Sigma \cdot h \leq L_{\max}$$

$$\Sigma \equiv \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}$$

Big Data in String Theory

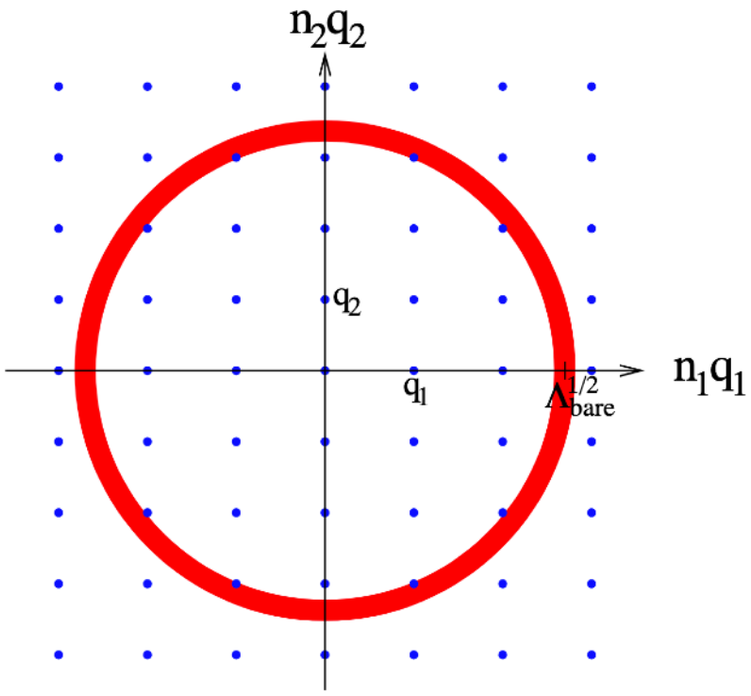
- The space of string compactifications is **large** and **complex**.
  - Flux vacua: $10^{500} \to 10^{272,000}$ flux choices per geometry
    [Ashok,Denef,Douglas],[Taylor,Wang], $\gtrsim 10^{755}$ geometries
    [Halverson,Long,Sung]

- We'd like a useful and efficient map of the global structure.

- Understanding the **boundary** of the landscape is logically equivalent to the **swampland** program [Vafa];…
  - **Correlations** as an obstruction/guide to model-building. Can't simply "pick and choose" multiple desirable features.

Computational Complexity in String Theory

- In this regime, **computational complexity** is important.
  - Difficult to **solve for a vacuum** (moduli stabilization, Diophantine equations, cohomology) [Halverson,Rühle] and to **find specific vacua** (e.g. with small cosmological constant in toy models) [Denef,Douglas]



[Bousso,Polchinski]

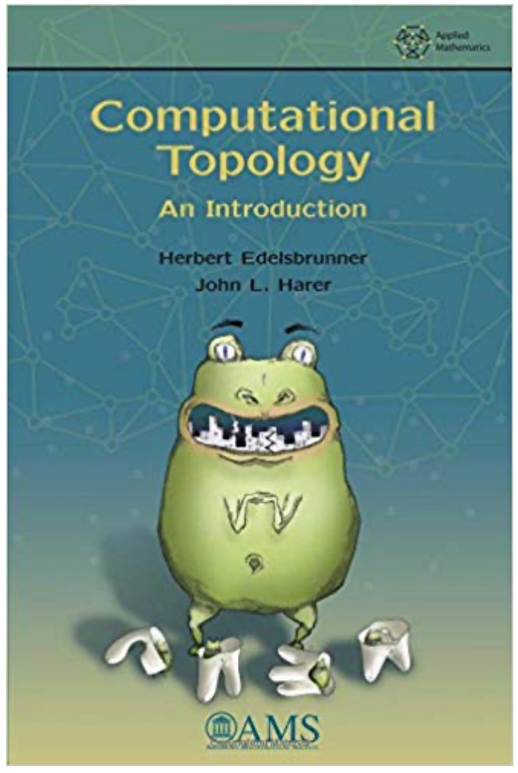
  - We'd benefit from shortcuts and reliable approximations. **Opportunity for data science/ML!**

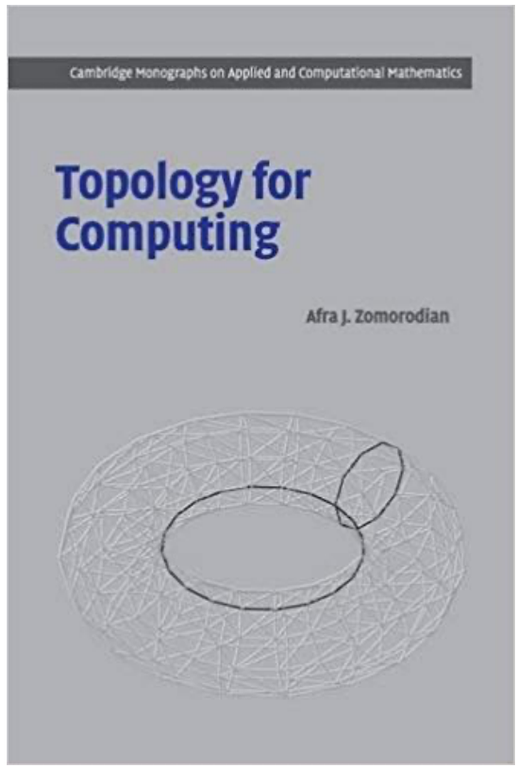
Summer 2017: [He];[Rühle]; [Carifio,Halverson,Krioukov,Nelson]; …[lots of recent work]

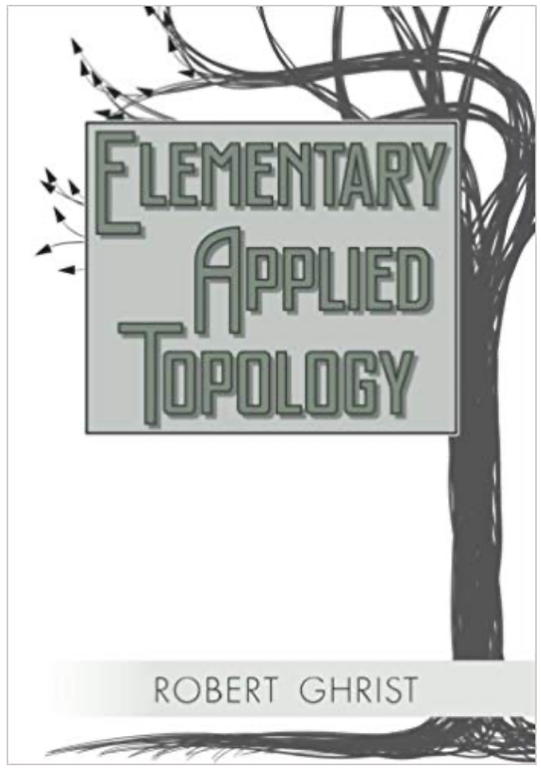
This talk: search for specific vacua with **genetic algorithms**

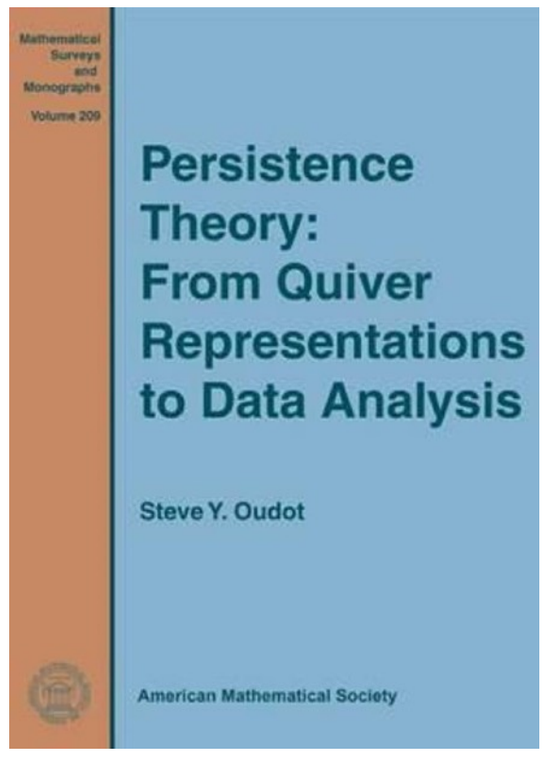
review: "Data science applications to string theory," F. Rühle

2. Topological Data Analysis







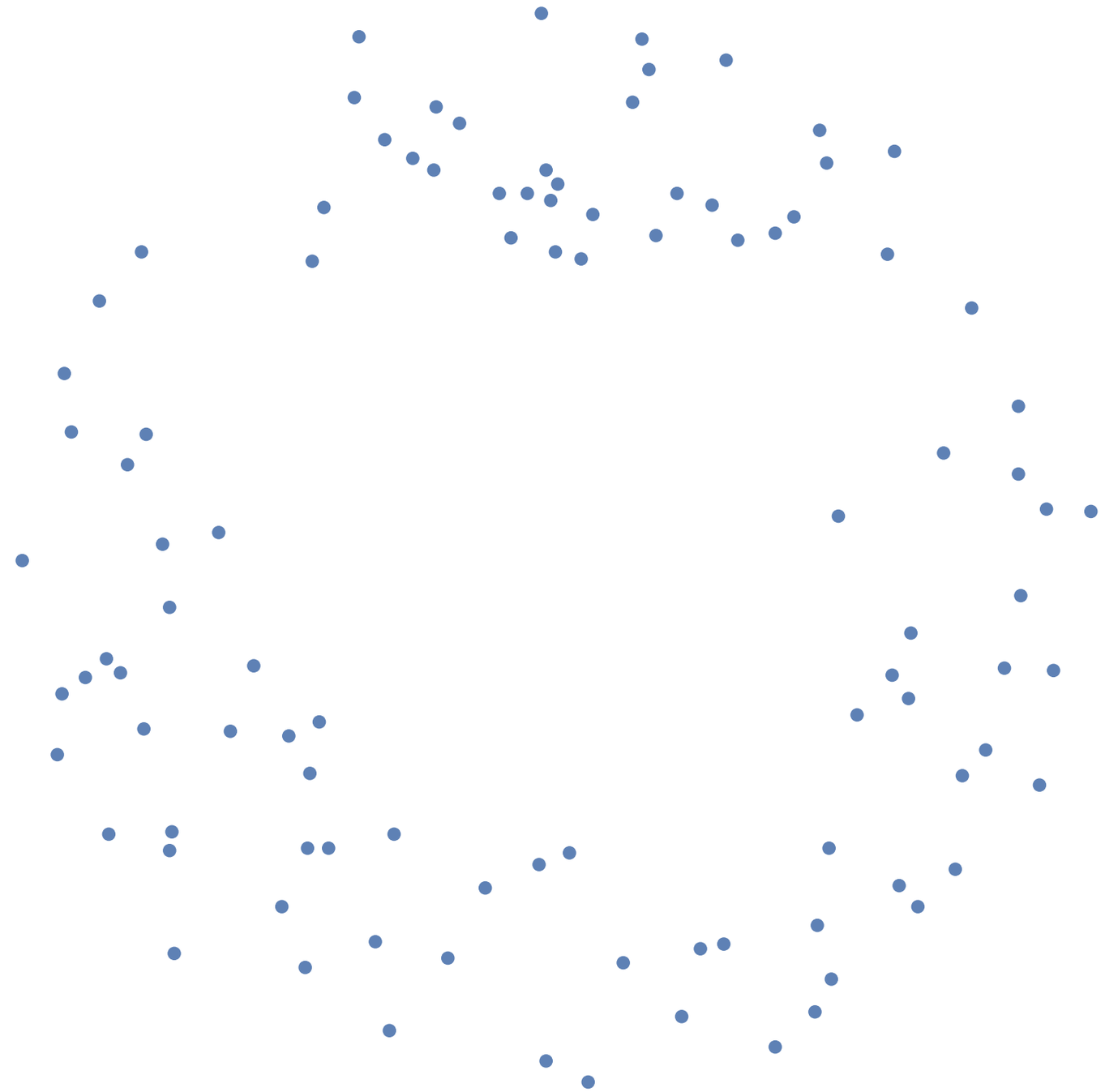


See also the introductory article “Topology and Data,” G. Carlsson

Persistent Homology

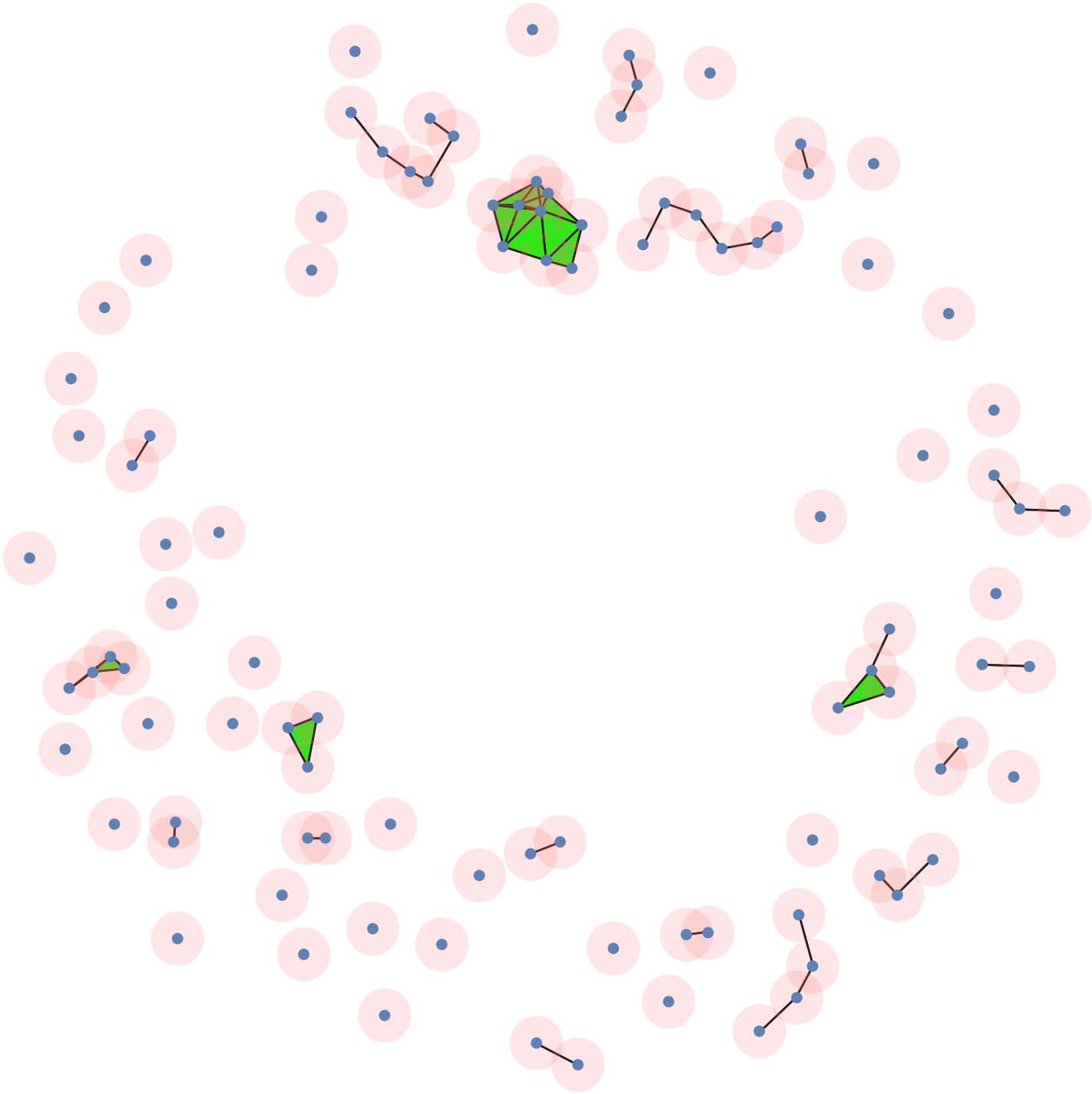
- Quantify shape in its most basic sense — topology.
- **Persistent homology:**
  - represent a data set with a **filtration** (growing sequence) of **simplicial complexes**.
  - filtration often parameterized by **coarse-graining** scale.
  - track **individual topological features** as they are **created** and **destroyed** in the filtration. (More refined than **Betti numbers**, which just **count** features.)

Example: Vietoris-Rips filtration



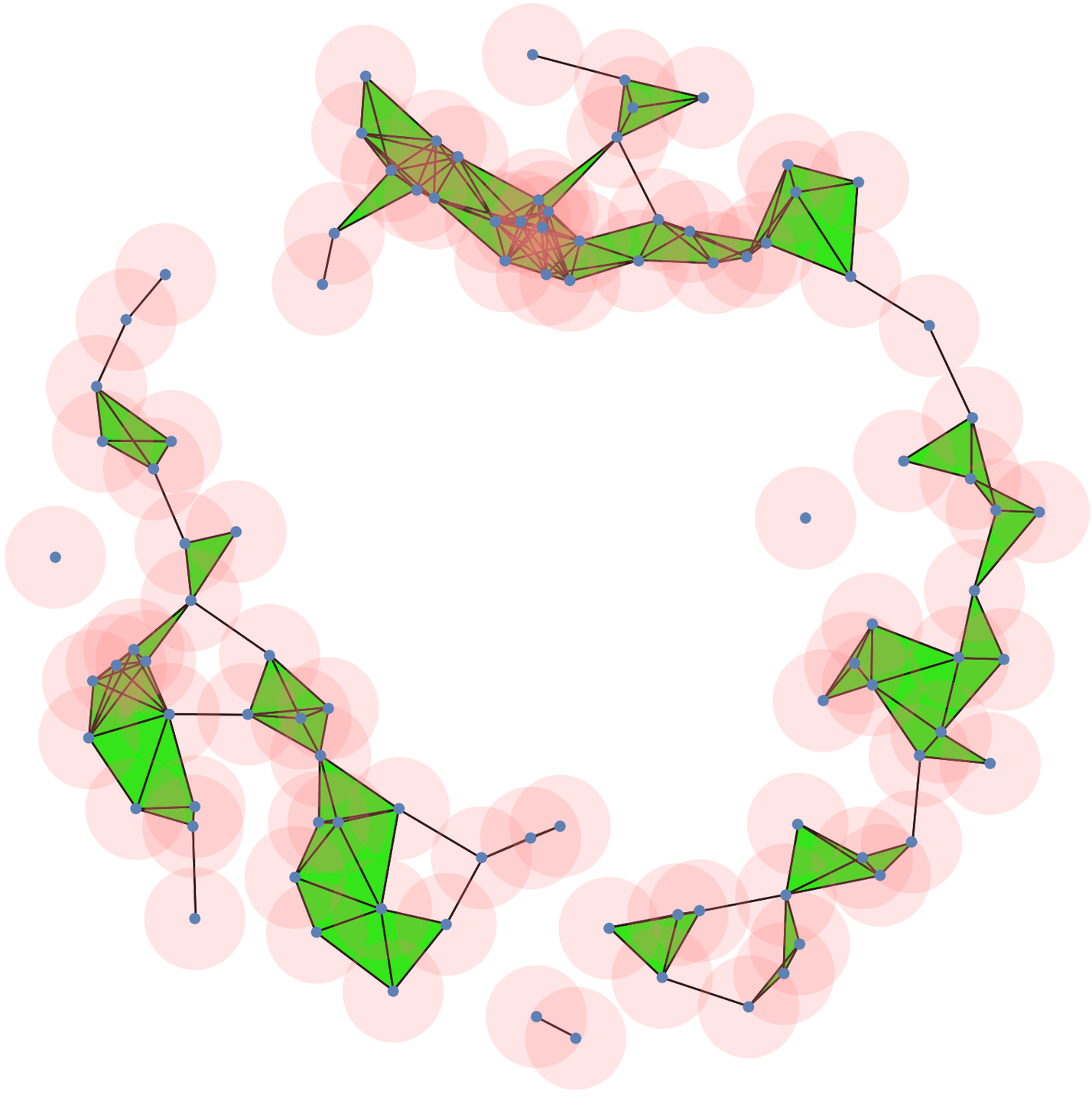
Example: Vietoris-Rips filtration

$\nu = 1$



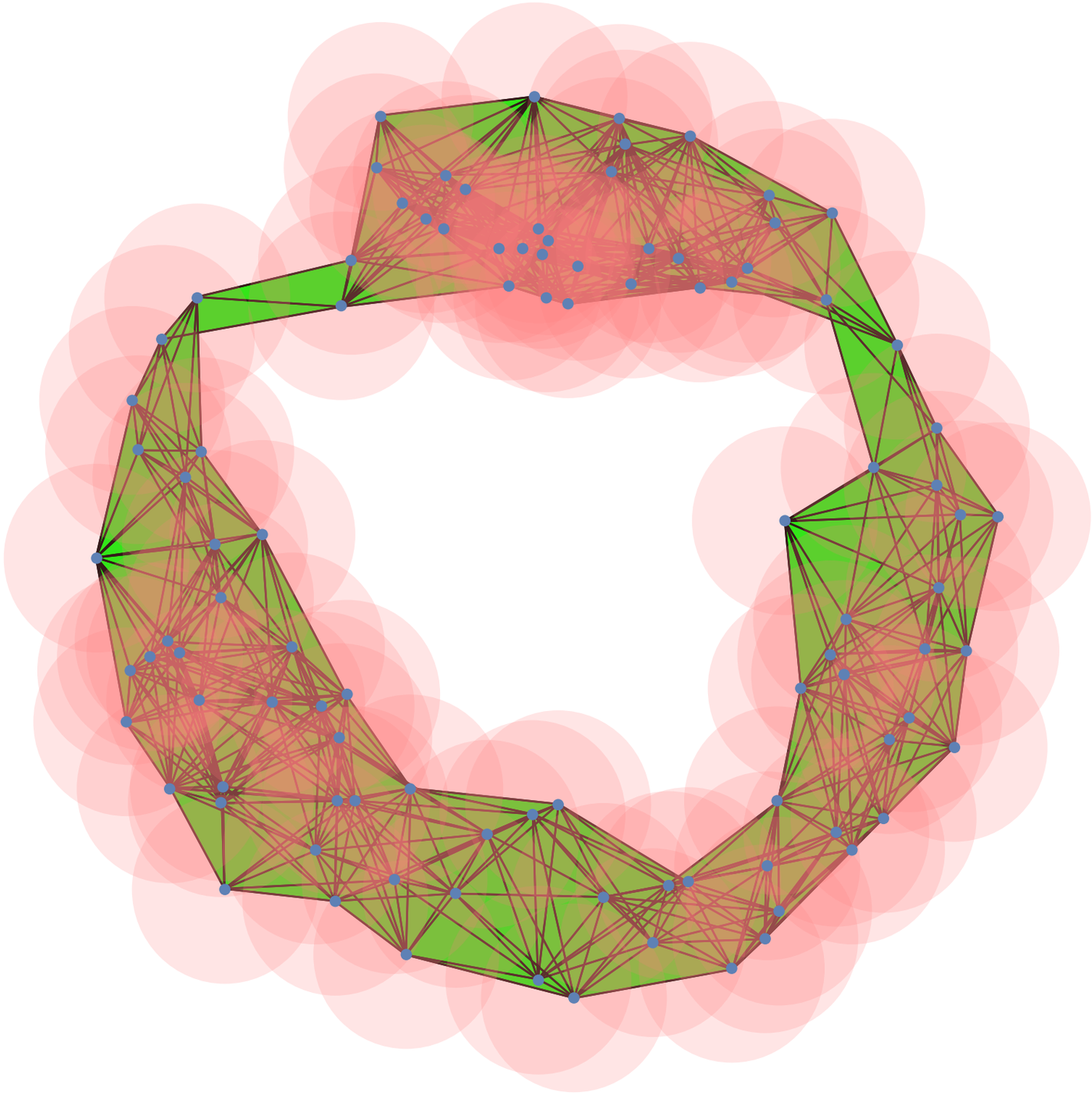
Example: Vietoris-Rips filtration

$\nu = 2$



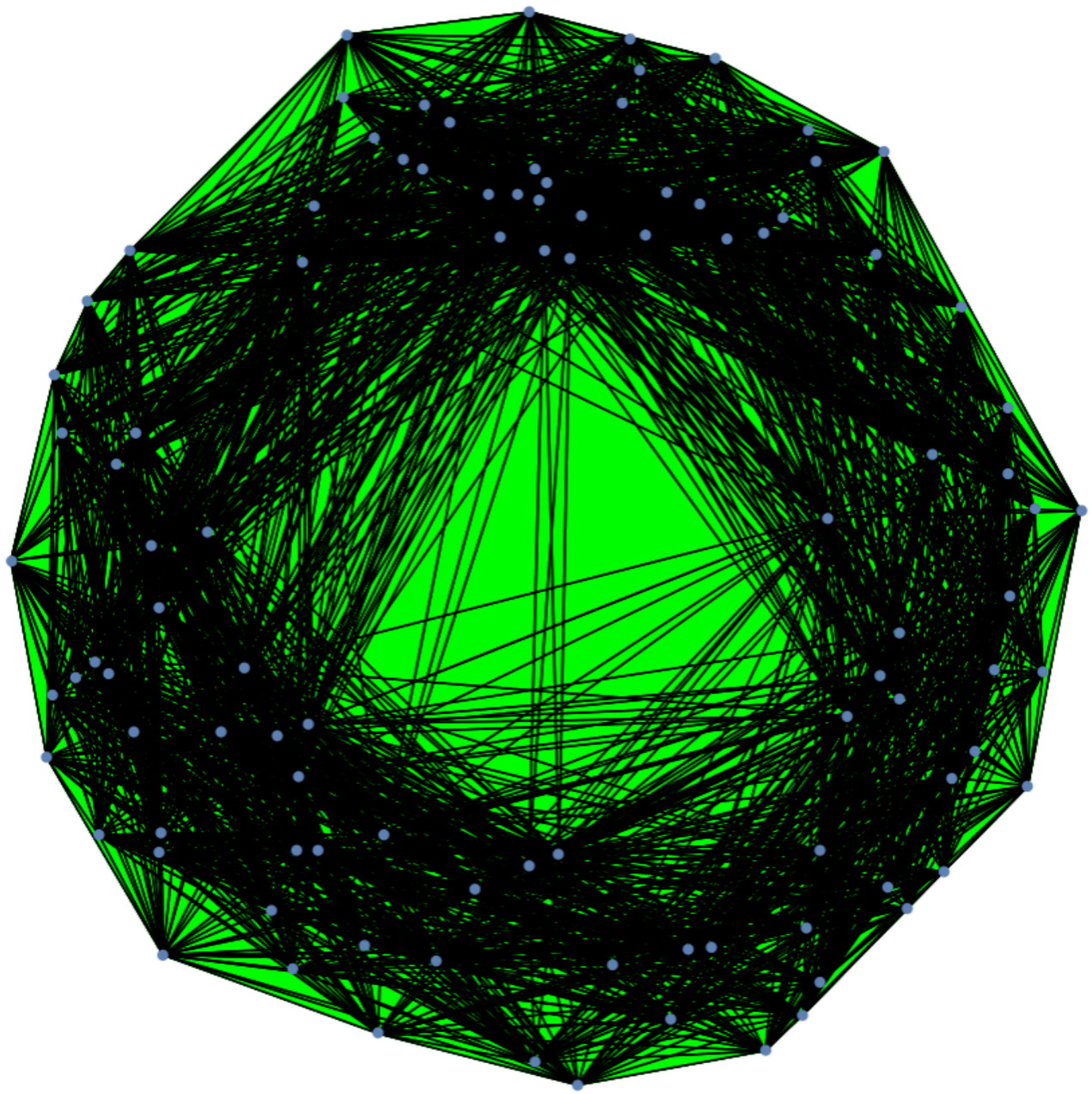
Example: Vietoris-Rips filtration

$\nu = 3$



Example: Vietoris-Rips filtration

$\nu = 5$

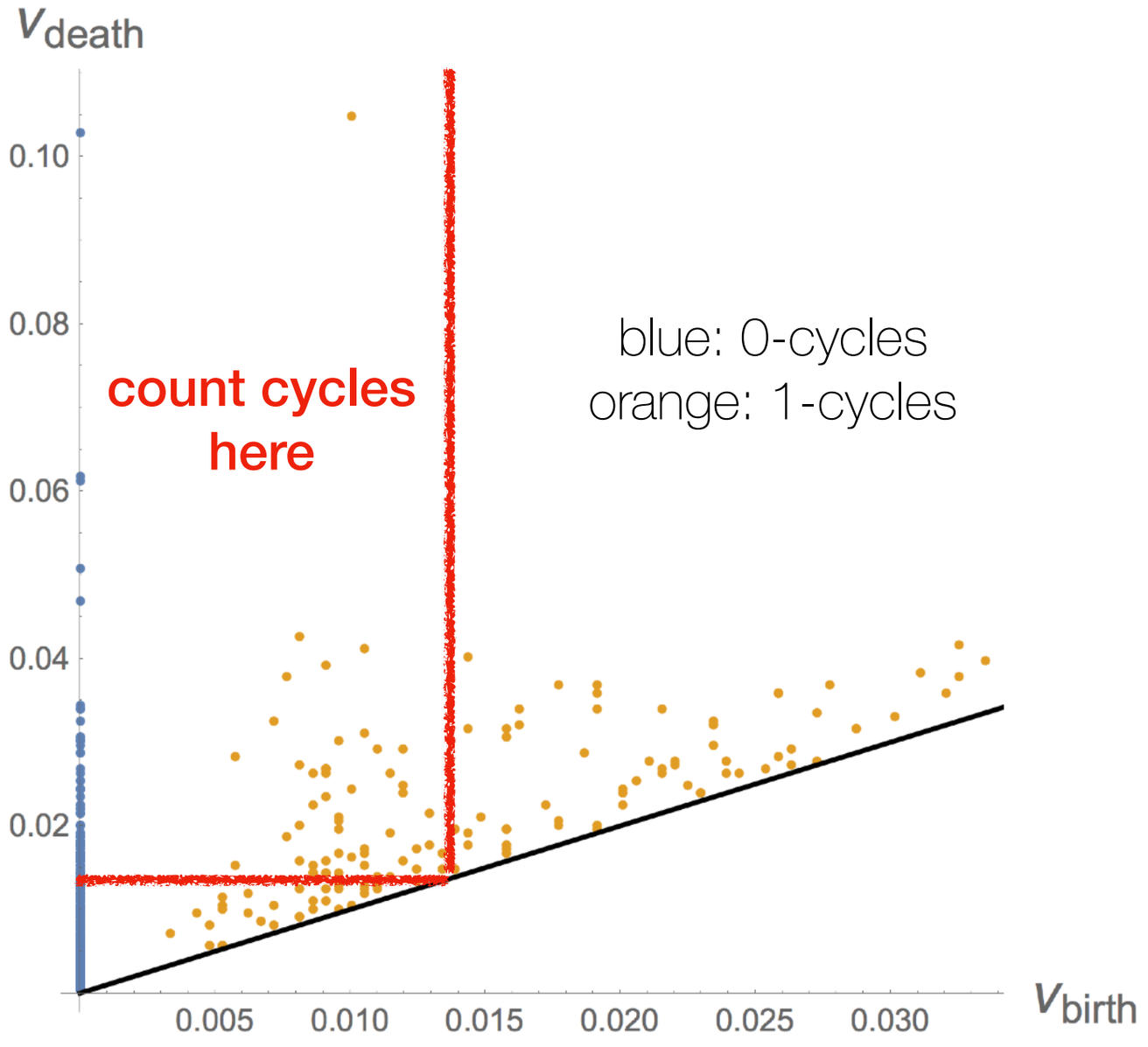


Visualizing Persistent Homology

- **Persistence diagrams** are scatter plots of birth and death times for **individual** homology generators

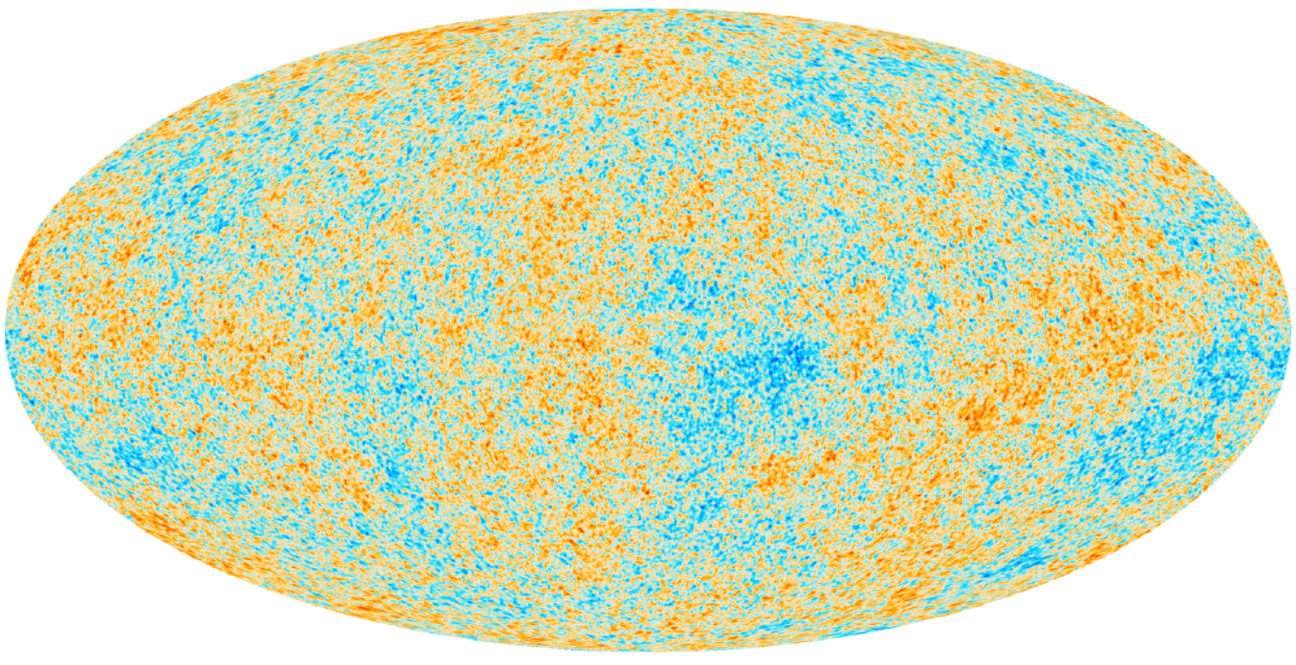
- Intuition*: long-lived features are “real,” short-lived features are “noise.”

- To calculate Betti numbers at a specific filtration time, count **“living cycles”**



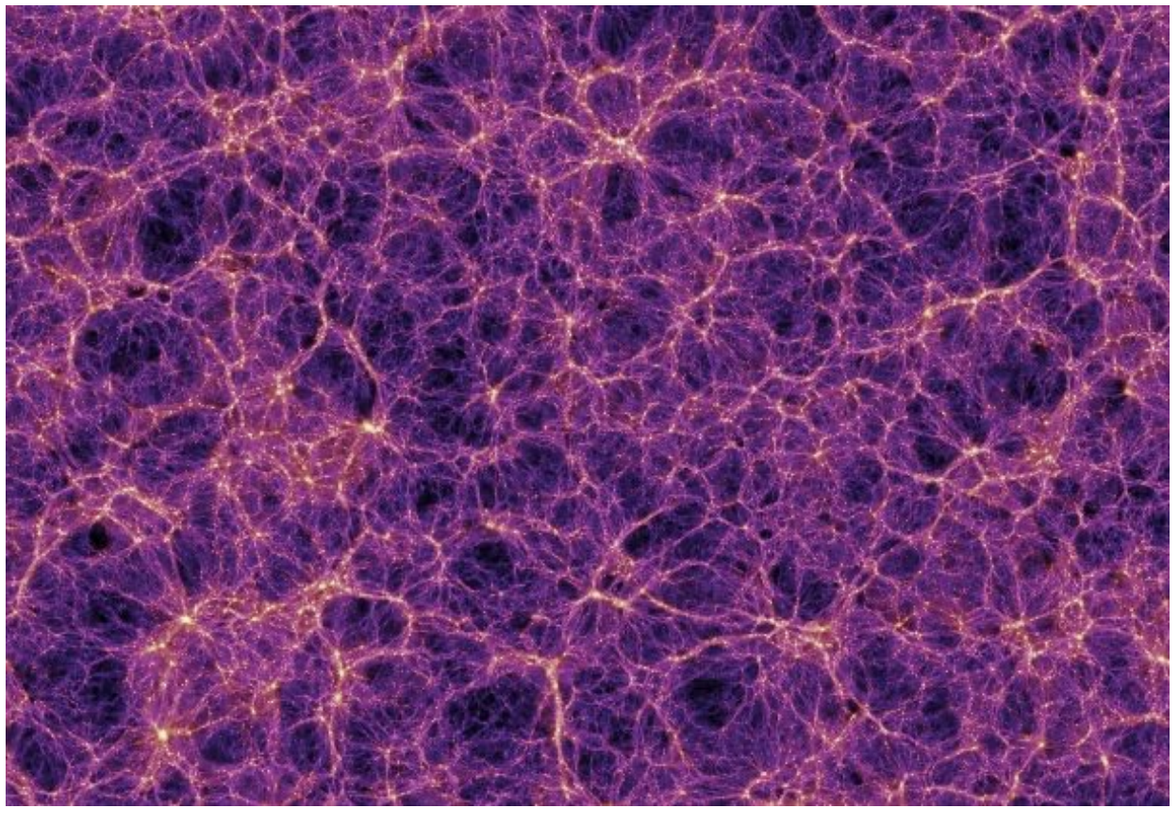
NB: will sometimes use $\nu_{persist} \equiv \nu_{death} - \nu_{birth}$ for vertical axis

2.1: TDA for Cosmology



[AC, Shiu]

1712.08159 (JCAP)



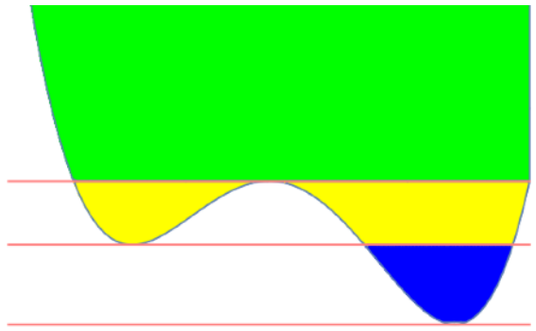
[WIP w/ Biagetti, Shiu]

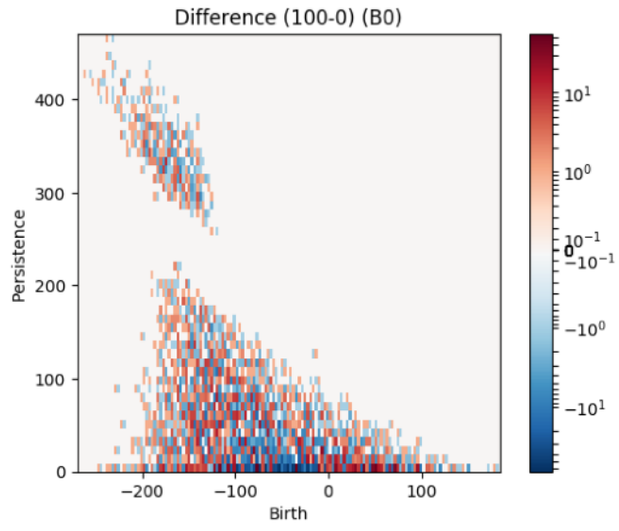
2006.xxxxx, 20xx.xxxx

[AC,Shiu]

TDA for CMB

- Take $\frac{\Delta T}{T}$ as Morse function on sphere, compute **sublevel persistence**. Topological features correspond to cold/hot spots.





- (low-resolution) CMB simulations with varying $f_{\mathrm{NL}}^{\mathrm{loc}}$ [Elsner,Wandelt]

- Take binned persistence diagram as statistic, perform toy likelihood analysis.

  - Sensitivity of topological statistics improved by factor of ~2

Statistic	Δf_{NL}
β_0	67.4
β_1	66.1
$\beta_0 + \beta_1$	60.6
PD_0	39.1
PD_1	37.4
$PD_0 + PD_1$	35.8

improvement

68% confidence constraints

NB: WMAP-resolution simulations, what's important here is the factor of ~2

Primordial NG in LSS

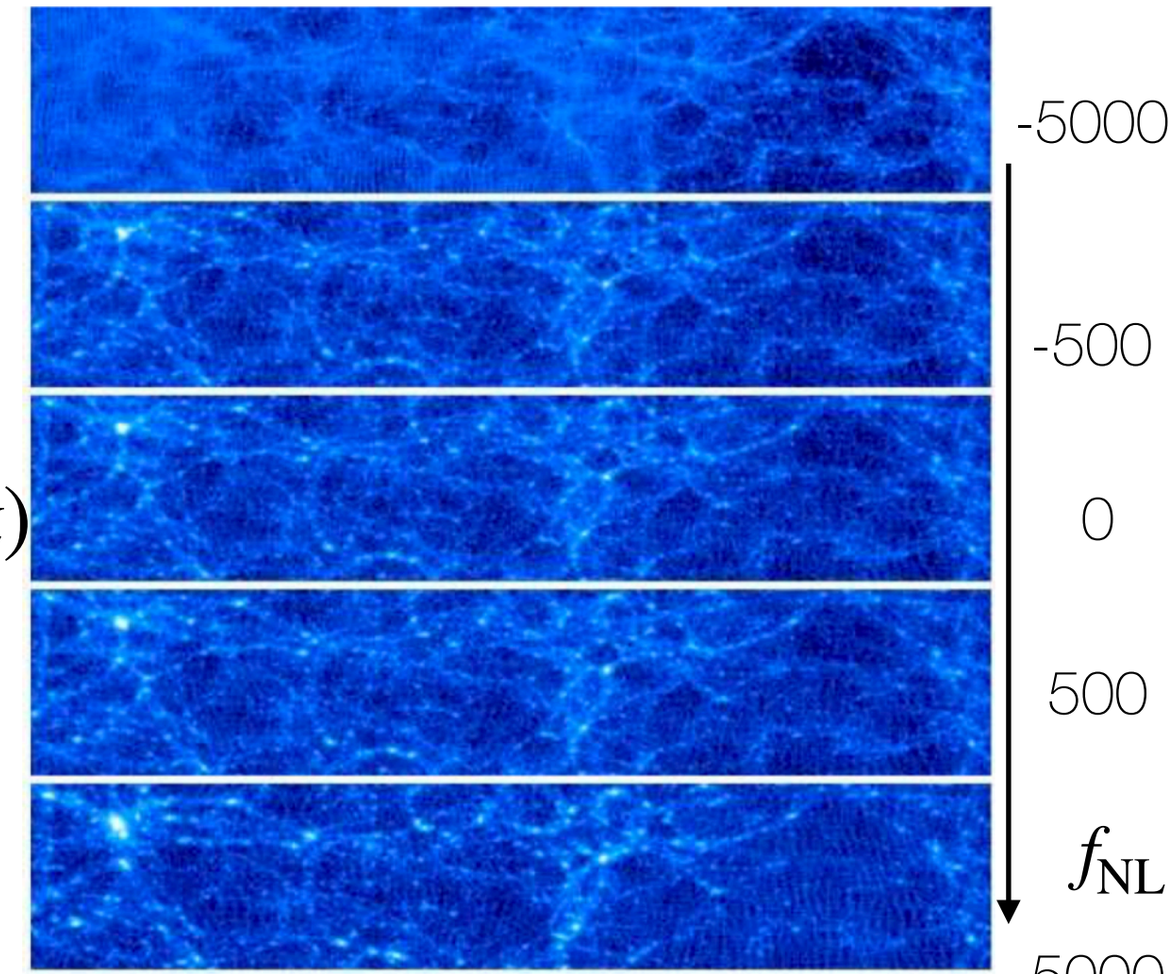
- Upcoming galaxy surveys will constrain $f_{\rm NL}^{\rm loc} \sim O(1-5)$ (e.g. SphereX [Dore et al. '16])

- Local NG: **scale-dependent bias**

$$P_{hh}(k) = \left(b_g + \frac{12}{5} \frac{f_{\rm NL}^{\rm loc}}{\mathscr{M}(k)} b_\zeta \right)^2 P_{mm}(k)$$

[Dalal et al.], [Matarrese,Verde], [Slosar et al.]

$\mathscr{M}(k) \sim k^2$ at small k



0710.4560

- [WIP w/ Biagetti, Shiu]: **persistent homology** pipeline for detecting local NG in LSS, tested on N-body simulations.

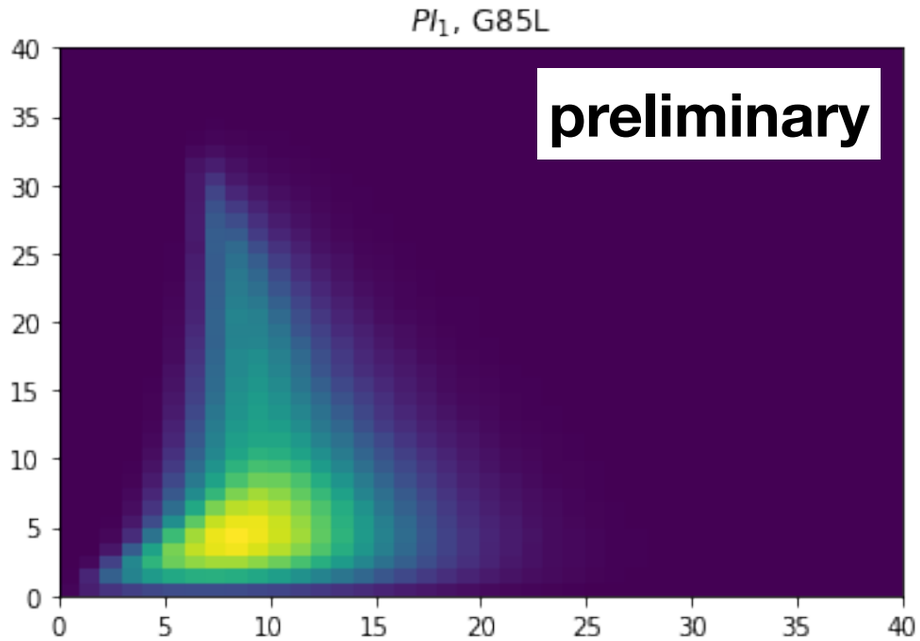
[WIP w/ Biagetti, Shiu]

TDA for LSS: Pipeline

- N-body *EoS dataset*:
  - https://mbiagetti.gitlab.io/cosmos/nbody/eos/

- Compute **alpha-filtration**, persistence diagrams for subsampled N-body simulations. Process these into **topological curves** and **persistence images**.

- **Subsampling** accounts for observational unknowns and allows for use of templates with large $f_{\rm NL}^{\rm loc}$



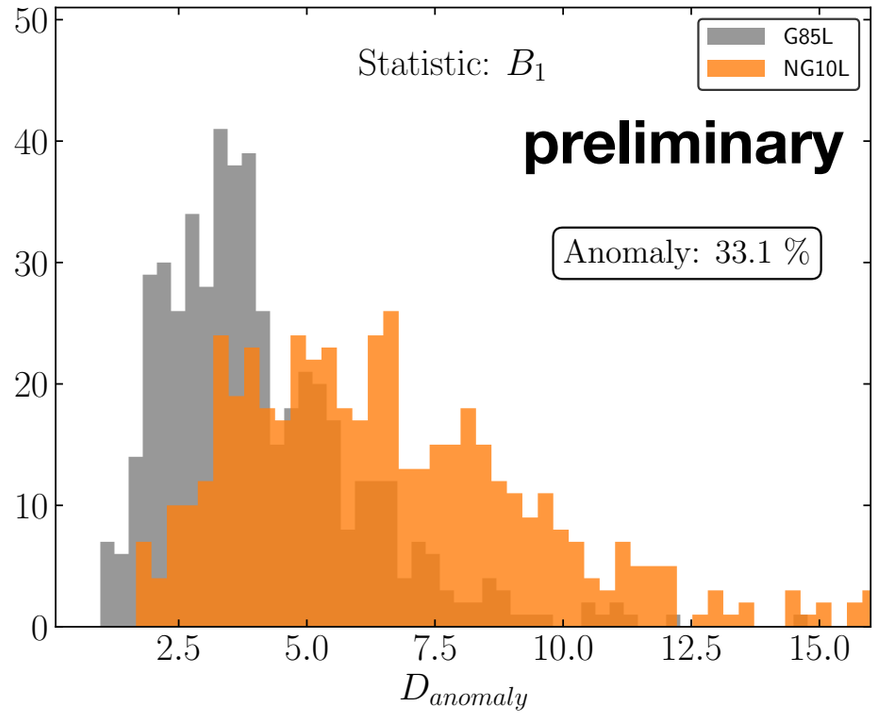
[WIP w/ Biagetti, Shiu]

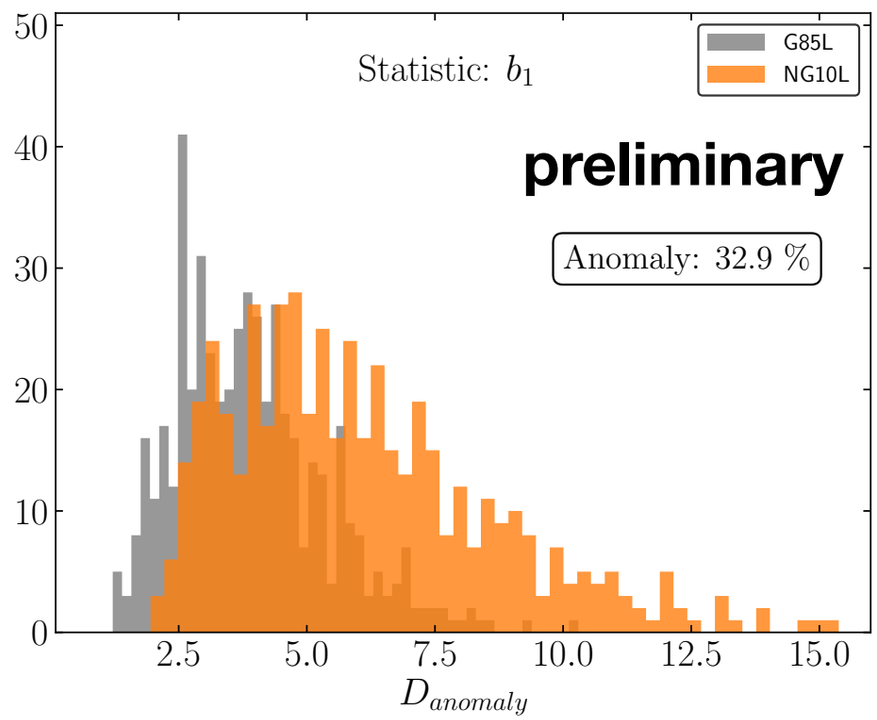
TDA for LSS: Anomalies

- **Anomaly detection**: given a statistic X from a simulation with $f_{\rm NL}^{\rm loc} = 10$, compute probability that it arises when $f_{\rm NL}^{\rm loc} = 0$.

$$D^X_{\rm anomaly} = \frac{1}{N}\sum_{i=1}^{N}\frac{(d_i - \mu_i)^2}{\sigma_i}$$

- Compare to results from fiducial cosmology to account for cosmic variance. Use to set **threshold for anomaly**, e.g. at 95 percentile.





[WIP w/ Biagetti, Shiu]

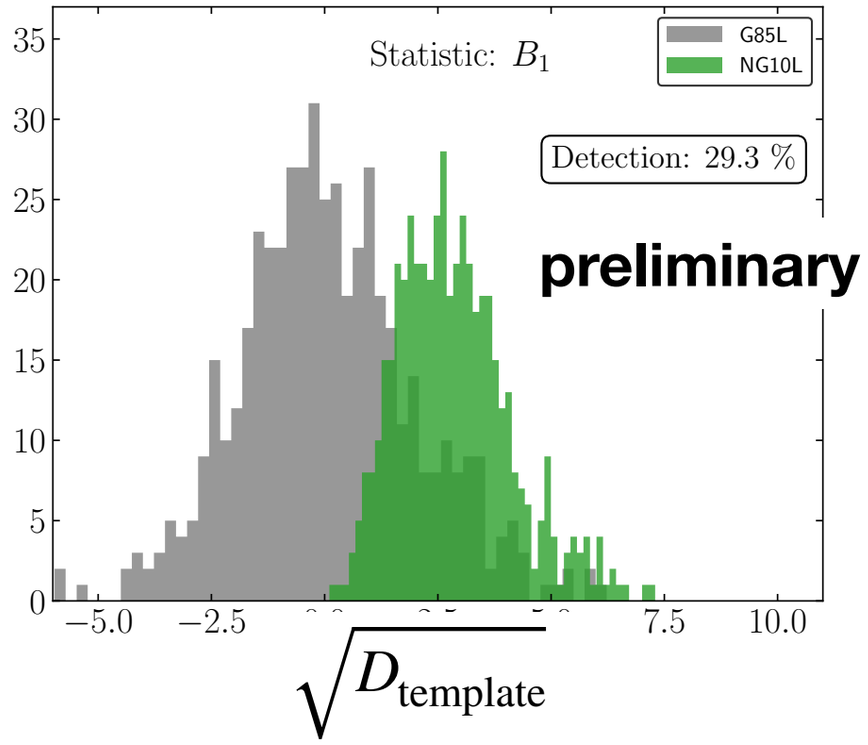
TDA for LSS: Templates

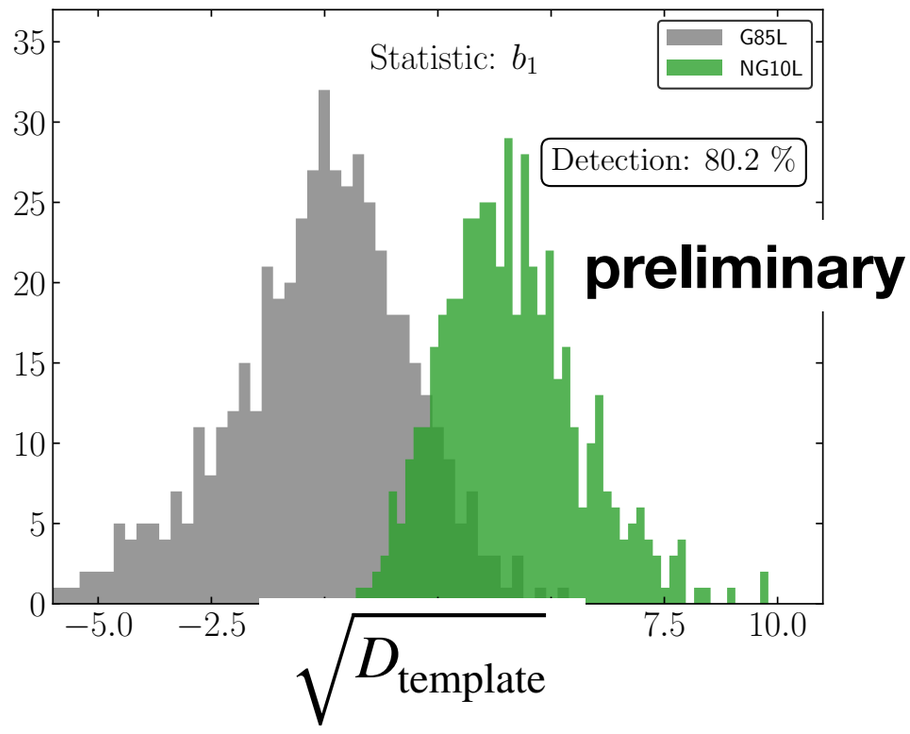
- Template method: compute templates corresponding to deviations from the fiducial cosmology.

$$\overrightarrow{T^X} \equiv \frac{1}{N_r} \sum_{i=1}^{N_r} \overrightarrow{S}^X_{NG_i} - \overrightarrow{S}^X_{G_i}$$

$$D_{\text{template}} = \frac{\left(\overrightarrow{S}_{\text{survey}} \cdot \overrightarrow{T} - \overrightarrow{S}_{\text{mock,avg}} \cdot \overrightarrow{T} \right)^2}{\sigma^2}$$

- Compare D_{template} to results from fiducial cosmology to account for cosmic variance. Set **threshold for detection** at e.g. 95%.





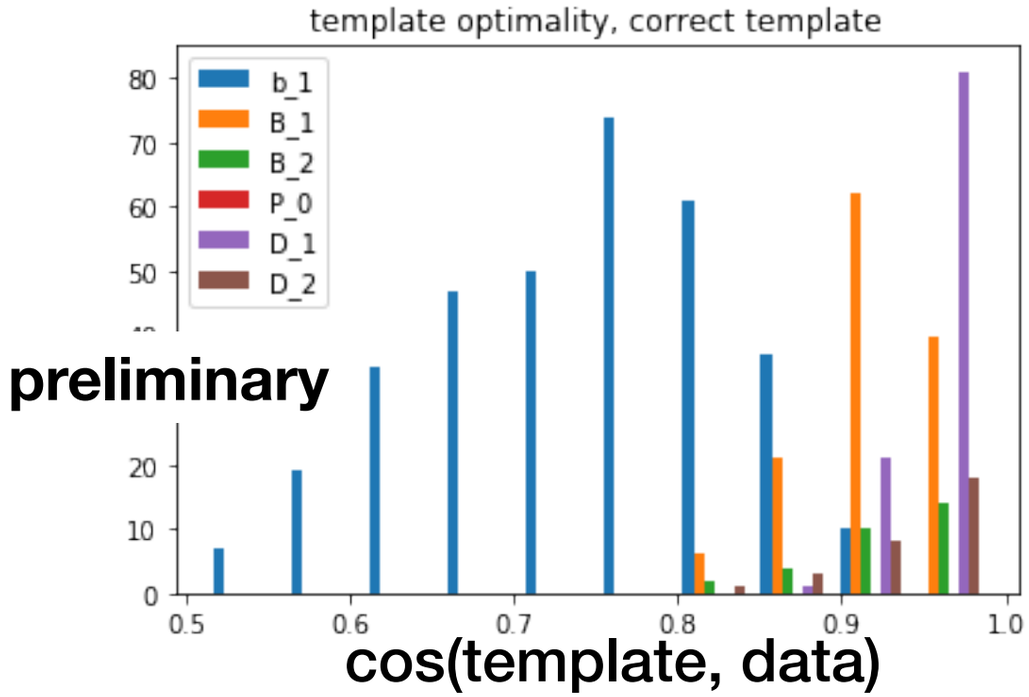
[WIP w/ Biagetti, Shiu]

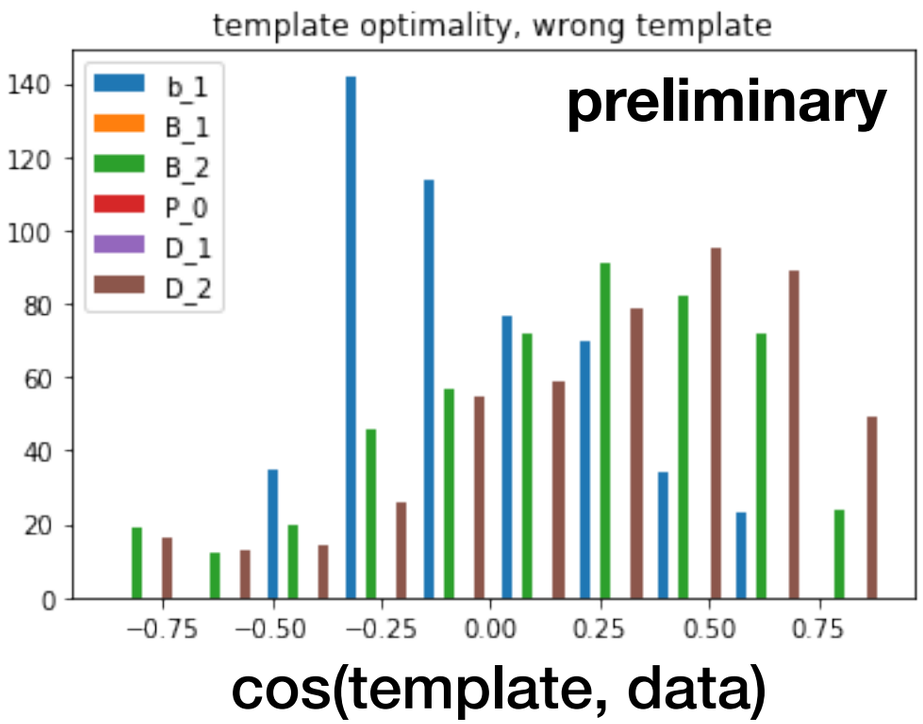
TDA for LSS: Degeneracies

- Degeneracy test: to avoid false positives, compute **template optimality** via "cosine"

- Included in *EoS*: variation of σ_8. **Cutoff for template optimality removes false detections** of $f_{\rm NL}^{\rm loc} \neq 0$.

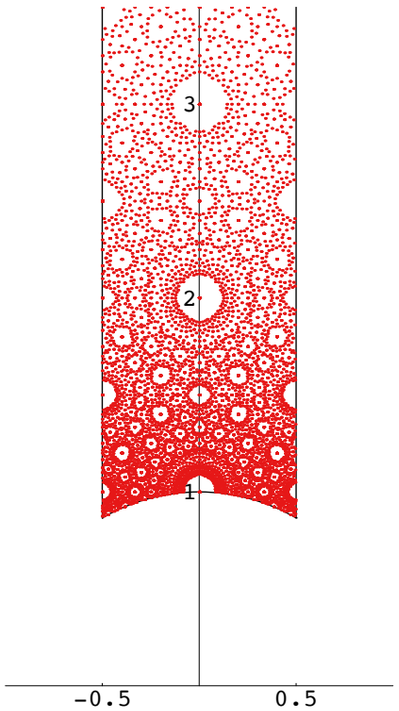
- Ongoing: compute possible degeneracies for wider range of cosmological parameters.





Persistent homology gives *explicit* computation of data's shape.
We used these to introduce **new observables for cosmology**.

See also **[AC,Shiu]**, where these techniques were applied to collections of flux vacua. In that context, topological features reflect correlations and finiteness; these prevent arbitrary fine-tuning.



[Denef, Douglas]

3. Genetic Algorithms

Stochastic Optimization

- Computational complexity affect both the computation of vacua (e.g. fluxes→ vacuum properties) and the **identification of particular vacua** (vacuum properties→fluxes).

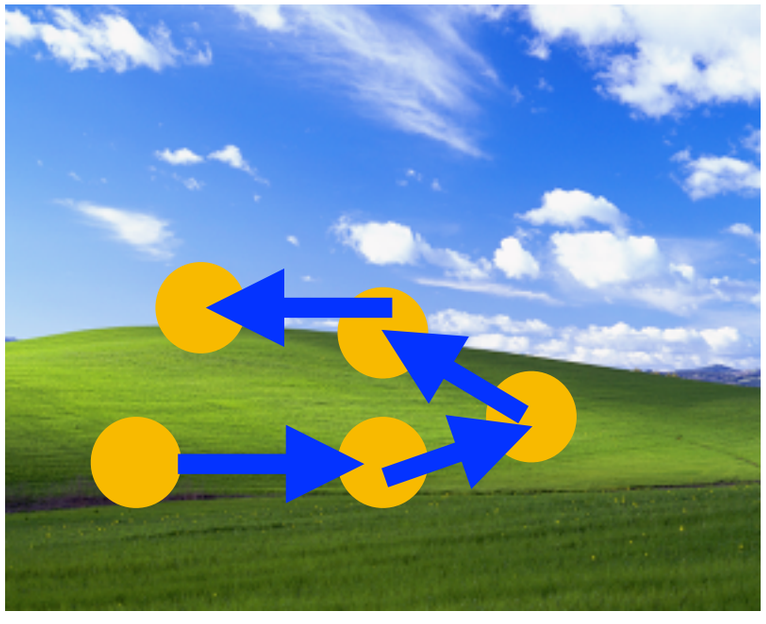
$$(F_3, H_3) \rightarrow_{DW=0} (\langle\phi\rangle, \langle z^a\rangle, \dots)$$

$$\boxed{(\langle\phi\rangle, \langle z^a\rangle, \dots) \rightarrow_{DW=0} (F_3, H_3)}$$

our mission

  - We'll focus on the latter issue. For nice progress on the former, see recent work [AbdusSalam,Abel,Cicoli,Quevedo,Shukla]

- Stochastic optimization: hill-climbing, simulated annealing, **genetic algorithm**…



optimization in a landscape

Genetic Algorithms

[Abel,Rizos],[Rühle], [AC,Schachner,Shiu], [AbdusSalam et al.] (refs)

- Genetic algorithm [Hollands]: model dynamics after **natural selection**.

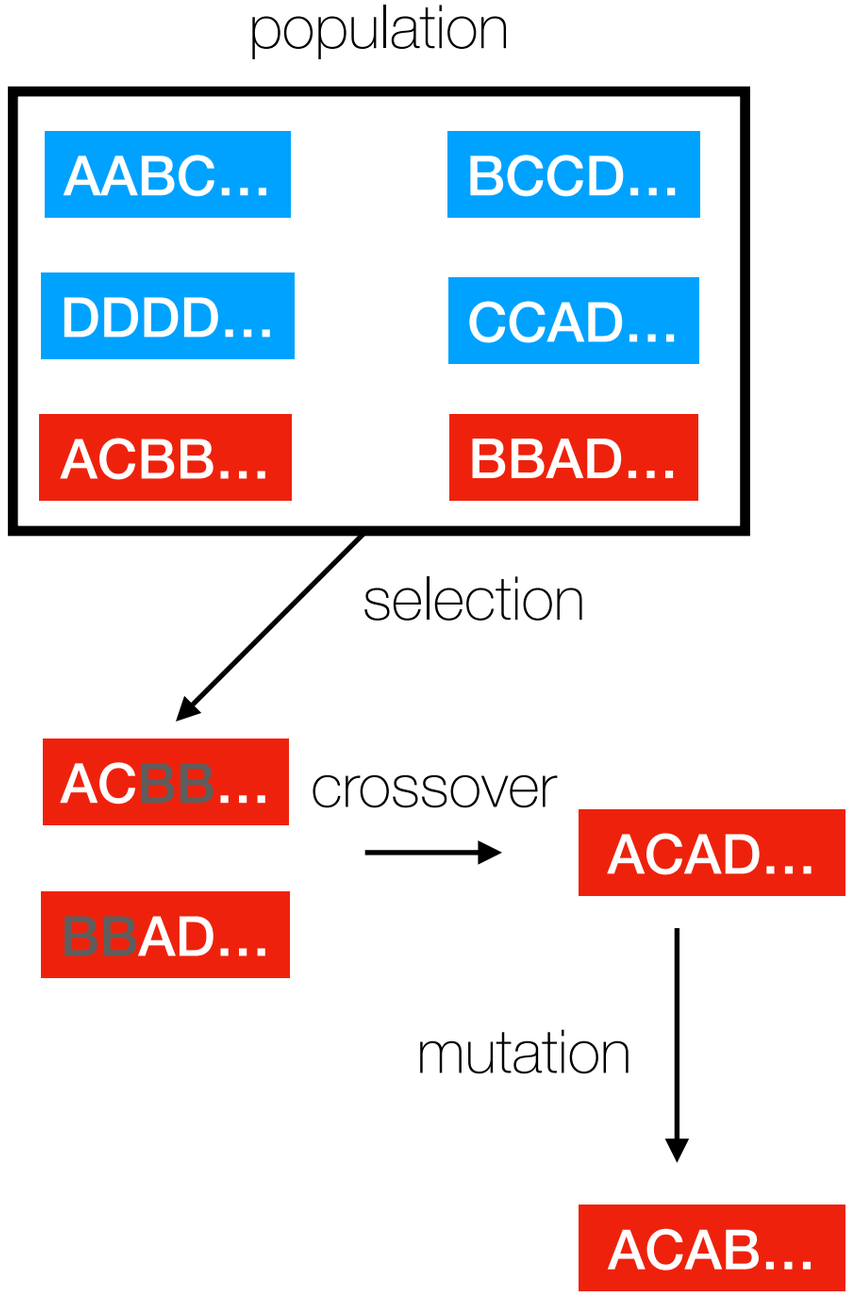
1. Generate a **population** of candidate solutions.

2. Parents are **selected** according to their fitness.

3. Parents **breed**: their genotypes are combined according to some predefined set of operators.

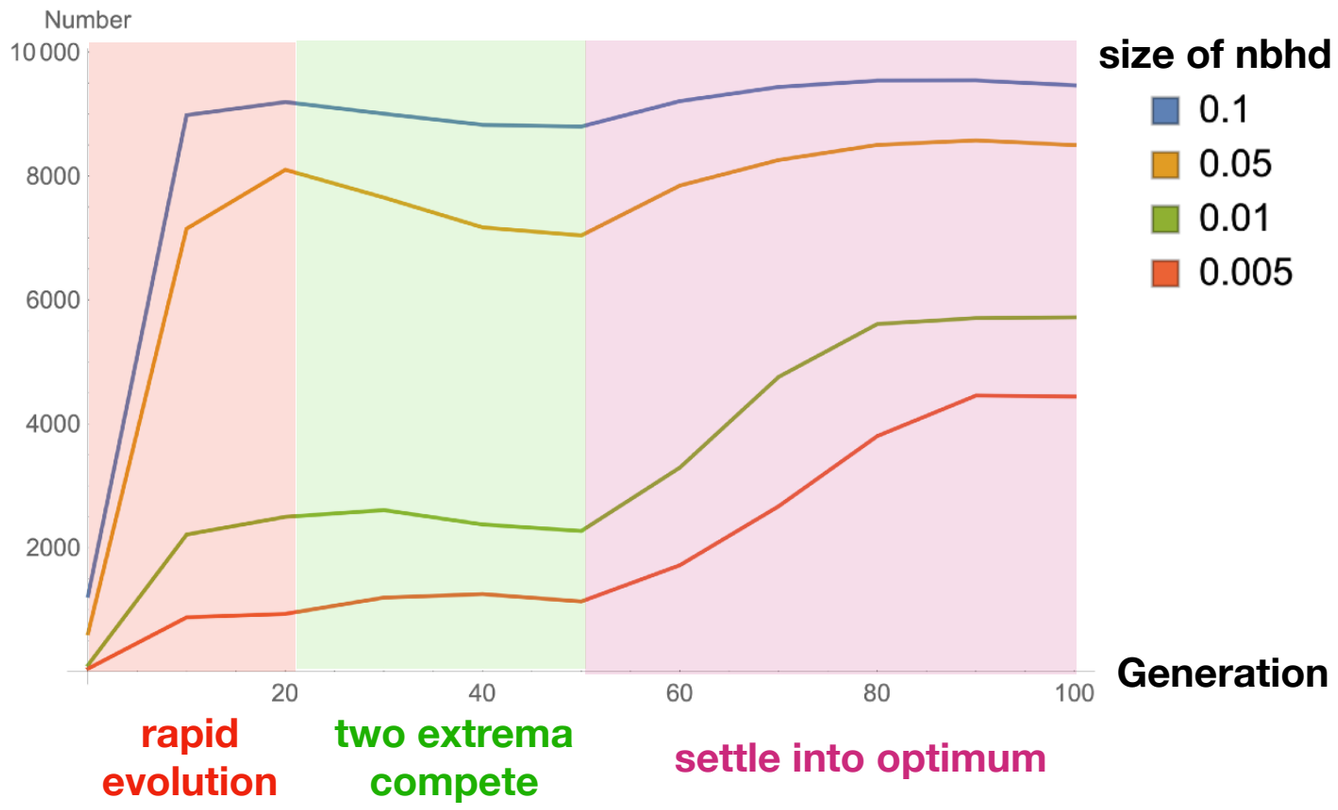
4. Children **mutate** with some probability.

Repeat 2-4 with children replacing parents.



[AC,Schachner,Shiu]

Genetic Algorithm: Example

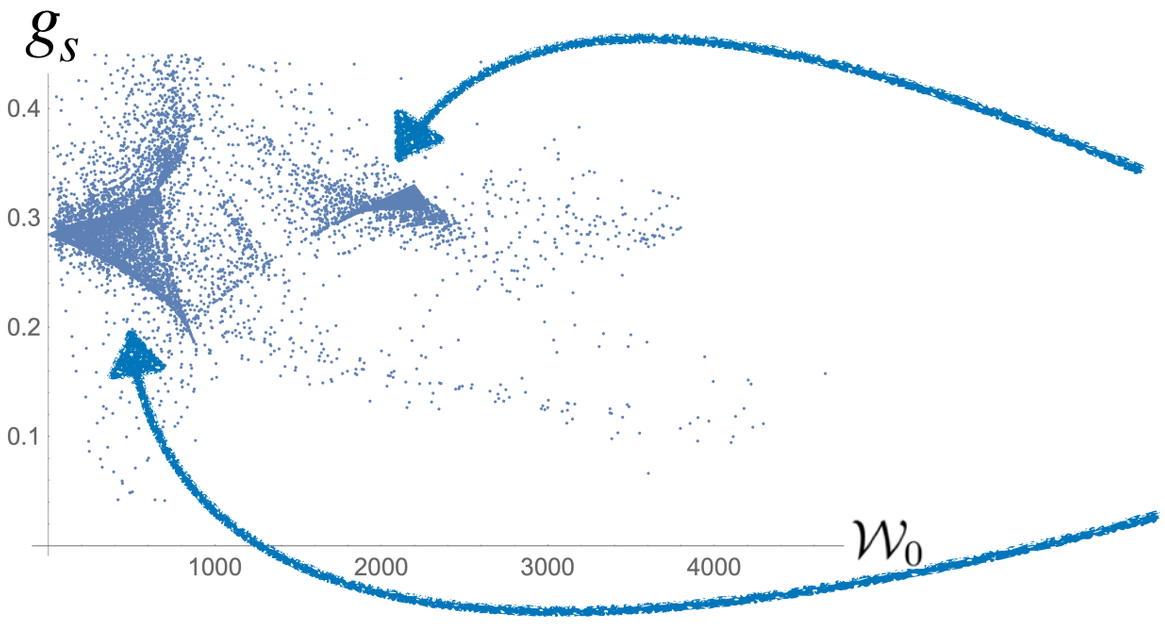


symmetric $T^6 = (T^2)^3$

Task: search for $g_s^* = 0.3$

Population size

$p = 10000$



Two competing optima, cf. "schema"

$\mathbf{N}_A = (3, -1, *, *, *, -8, *, 10)$

arbitrary

$\mathbf{N}_B = (*, 4, *, -7, 41, -2, -33, *)$

Genetic Algorithm: Applicability

- When should a local search be feasible? Some intuition via **fitness-distance correlation (FDC)**:

$$FDC = \frac{1}{N}\sum_{i}^{N}\frac{(f_i - \bar{f})(d_i - \bar{d}_i)}{\sigma_f \sigma_d}$$

f_i: fitness

d_i: distance to optimum

- Fitness is easy to max/minimize if f_i and d_i are anti/correlated, or $|FDC| \sim 1$.

- Note that FDC depends on encoding, or **representation**.

  - Connection to **dualities** (see [Betzler,Krippendorf]). For nearest-neighbor Ising model:

$$FDC_{\text{neighbor}} \sim -0.3$$

$$FDC_{\text{domain}} = -1 \longrightarrow \text{easier to minimize energy via local operations}$$

Summary

- Data has **structure**.
- We can explicitly compute this structure using tools like **persistent homology**.
  - Persistent homology provides new real-space observables for **non-Gaussianity** in CMB and LSS.
- Efficient **search strategies** should exploit this structure.
  - **Genetic algorithms** promising as tool for searching for particular string vacua.

Future work

- LSS: more non-Gaussian shapes, degeneracies of cosmological observables, analytical foothold

- Genetic algorithms: higher-dimensional moduli spaces, testing swampland conjectures

  - ML to discover nice encodings for search algorithms?

  - Compare/contrast/combine with NN-based approaches, other optimization approaches

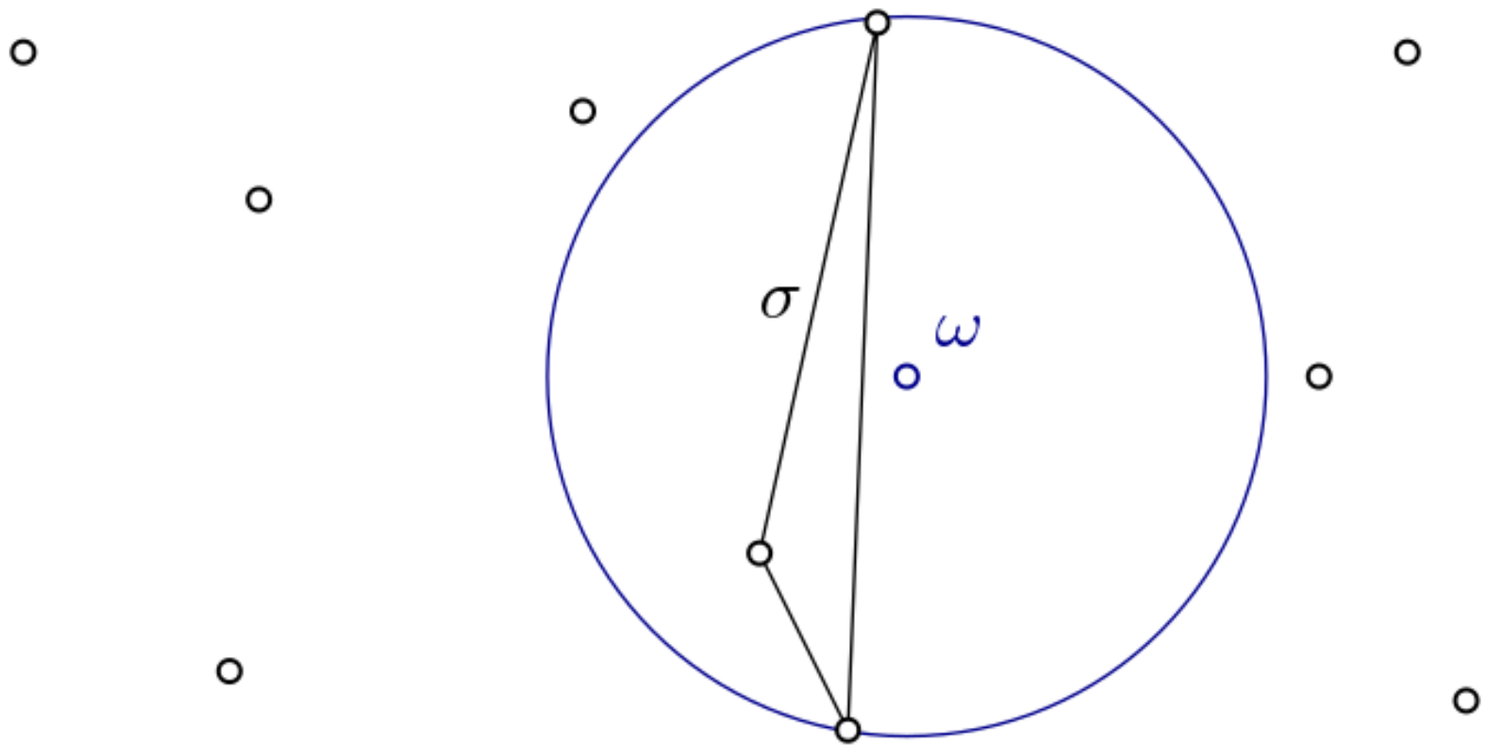
- Relationship between search algorithms, encodings, and dynamics of vacuum selection **[Bao,Bousso,Jordan,Lackey], [Denef,Douglas,Greene,Zukowski], [Khoury et al.]**

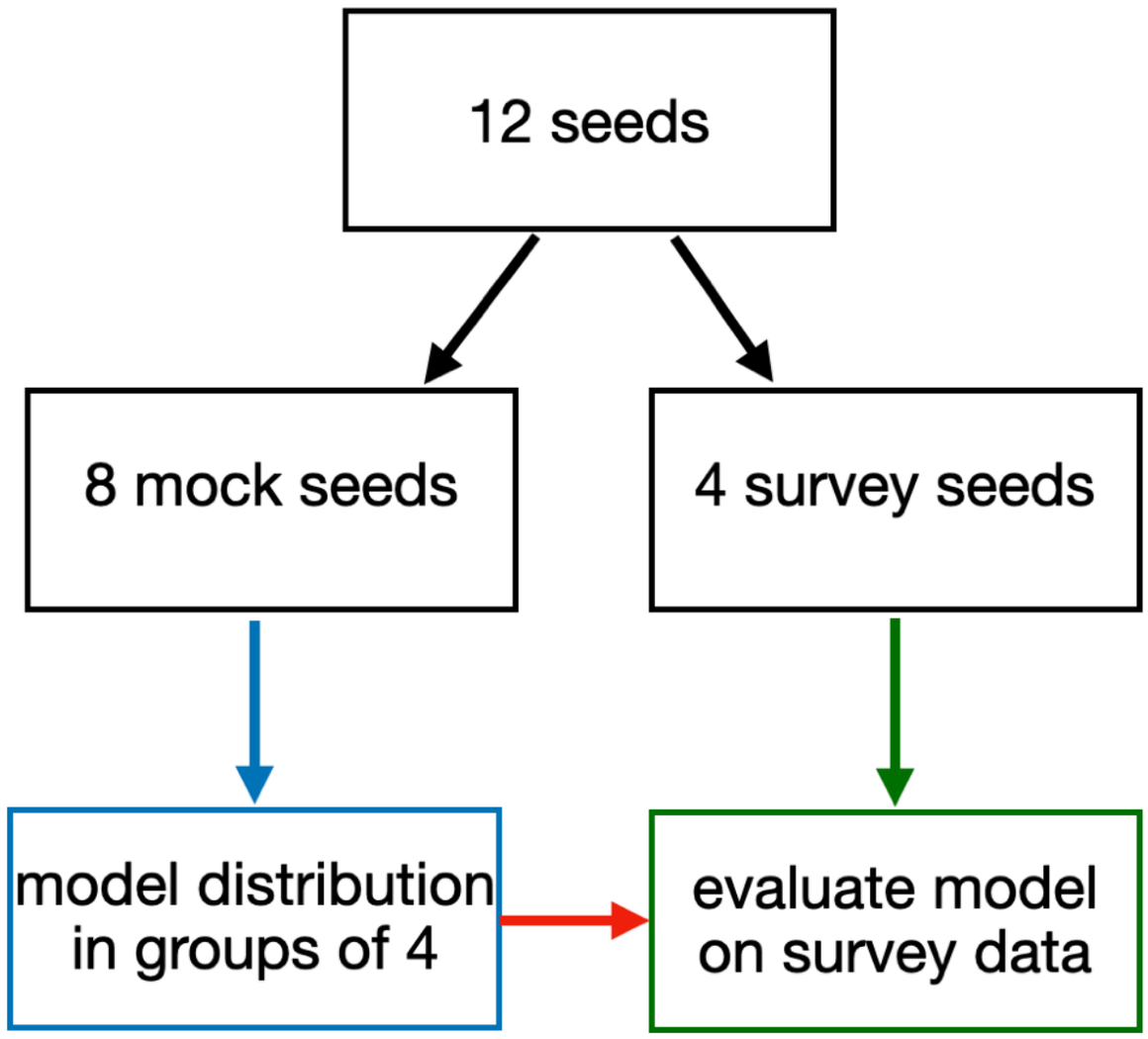
Thanks!

Extra Slides

Sampling in TDA

- We can't realistically include all 10^{500} vacua as vertices
- Can sample the topology via the witness complex: [de Silva,Carlsson]
  - From the entire point cloud Z, choose a landmark set L as the complex's vertices. Often chosen randomly or via sequential maxmin algorithm
  - Let $m_k(z)$ be the distance from some z∈Z to the (k+1)-nearest landmark point. Then, given filtration parameter ν, the simplex $[l_0 l_1 \dots l_k]$ is included in the witness complex if $\max\{d(l_0,z), d(l_1,z), \dots, d(l_k,z)\} \leq \nu + m_k(z)$




12 seeds
8 mock seeds
4 survey seeds
model distribution
in groups of 4
evaluate model
on survey data