Topological Data Analysis for Cosmology and String Theory

Gary Shiu

University of Wisconsin-Madison

Big Data in Big Sciences

Cosmology is marching into a big data era:

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

	data volume	schedule
SDSS	40 TB [3]	2000-2020
DESI	2 PB [5]	2019-2027
LSST	> 60 PB [1, 7]	2020-2030
Euclid	>10 PB [9]	2020-2027
WFIRST	>2 PB [12]	2023-2030
CMB-S4	$10^4 \times$Planck [13, 5]	2020-2027(?) [14]
SKA	4.6 EB [3, 15]	2019-2030(?)

Big Data in Big Sciences

Cosmology is marching into a big data era:

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

	data volume	schedule
SDSS	40 TB [3]	2000-2020
DESI	2 PB [5]	2019-2027
LSST	> 60 PB [1, 7]	2020-2030
Euclid	>10 PB [9]	2020-2027
WFIRST	>2 PB [12]	2023-2030
CMB-S4	$10^4 \times$Planck [13, 5]	2020-2027(?) [14]
SKA	4.6 EB [3, 15]	2019-2030(?)

~ 200PB of ***raw data*** are collected in the first 7 years of the **LHC.**

Big Data in Big Sciences

Cosmology is marching into a big data era:

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

	data volume	schedule
SDSS	40 TB [3]	2000-2020
DESI	2 PB [5]	2019-2027
LSST	> 60 PB [1, 7]	2020-2030
Euclid	>10 PB [9]	2020-2027
WFIRST	>2 PB [12]	2023-2030
CMB-S4	$10^4 \times$Planck [13, 5]	2020-2027(?) [14]
SKA	4.6 EB [3, 15]	2019-2030(?)

~ 200PB of ***raw data*** are collected in the first 7 years of the **LHC.**

In terms of sheer volume, nothing trumps the volume of ***theoretical data of string vacua***. A rough estimate gives:

10^{500} (Type IIB flux vacua) [Ashok-Denef-Douglas]

$10^{272,000}$ (F theory flux vacua) [Taylor-Wang]

The Shape of Data

United States presidential election, 2016

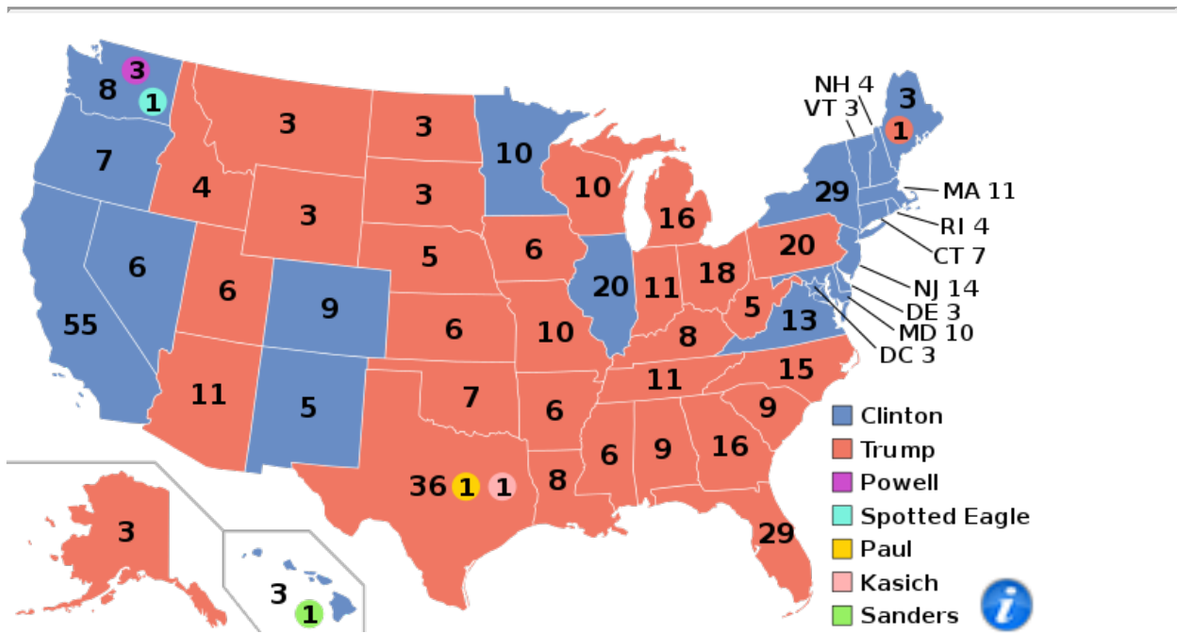
November 8, 2016

538 members of the Electoral College

270 electoral votes needed to win

Turnout	55.7% (estimated)[1] ▲ 0.8 pp

Nominee	**Donald Trump**	Hillary Clinton
Party	Republican	Democratic
Home state	New York	New York
Running mate	**Mike Pence**	Tim Kaine
Electoral vote	**304**[a][2]	227[a][2]
States carried	**30 + ME-02**	20 + DC
Popular vote	62,984,825[3]	**65,853,516**[3]
Percentage	46.1%	**48.2%**



Presidential election results map. Red denotes states won by Trump/Pence, blue denotes those won by Clinton/Kaine. Numbers indicate electoral votes allotted to the winner of each state. Faithless votes: Colin Powell 3 (WA), John Kasich 1 (TX), Ron Paul 1 (TX), Bernie Sanders 1 (HI), Faith Spotted Eagle 1 (WA)

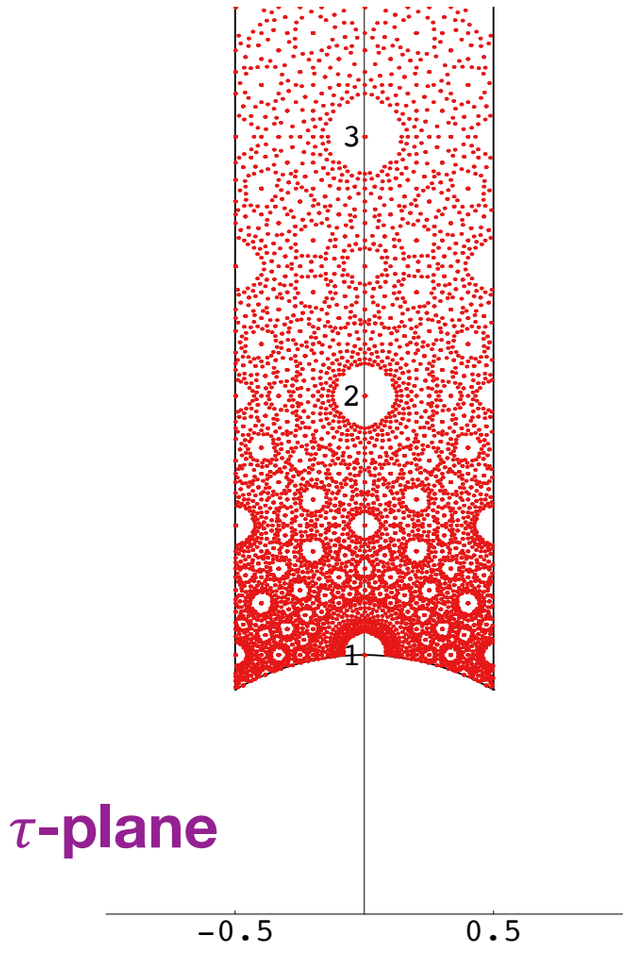
President before election	**Elected President**
Barack Obama	Donald Trump
Democratic	Republican

(from Wikipedia)

Distribution of String Vacua

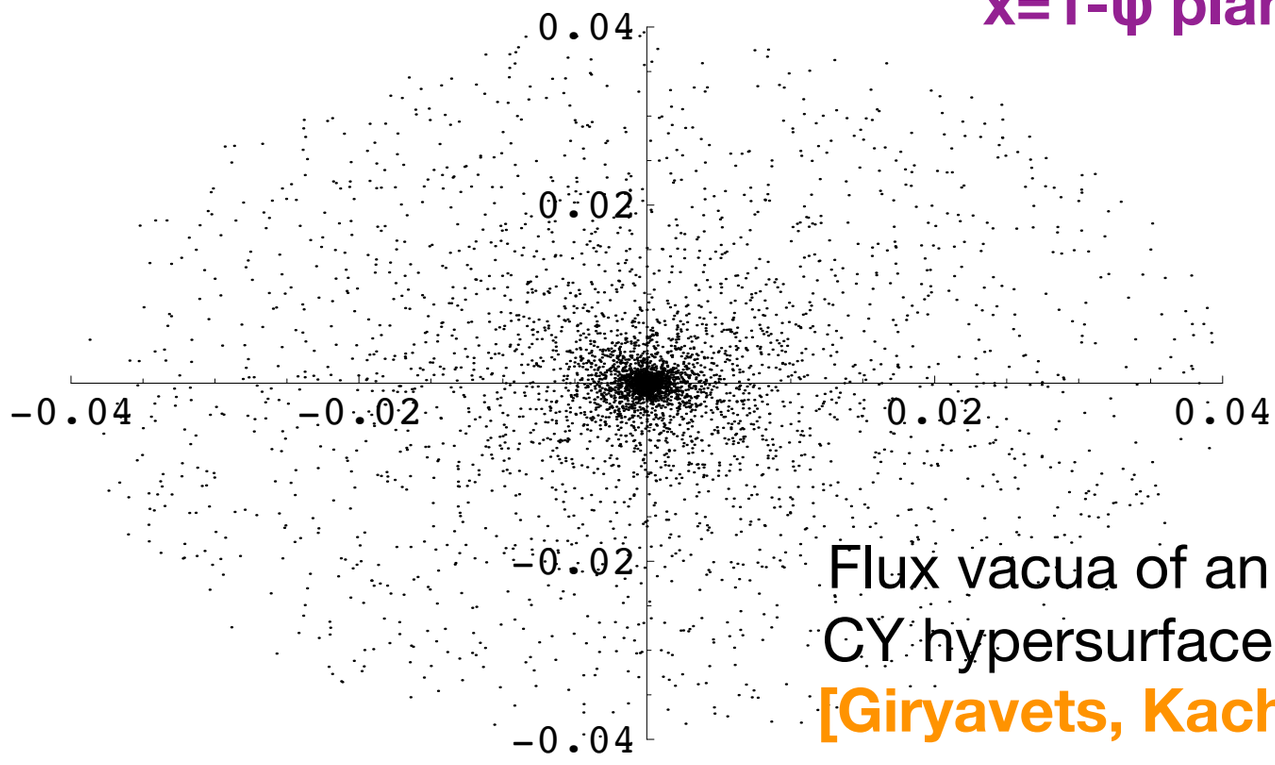
Flux vacua on rigid CY

[Denef-Douglas]



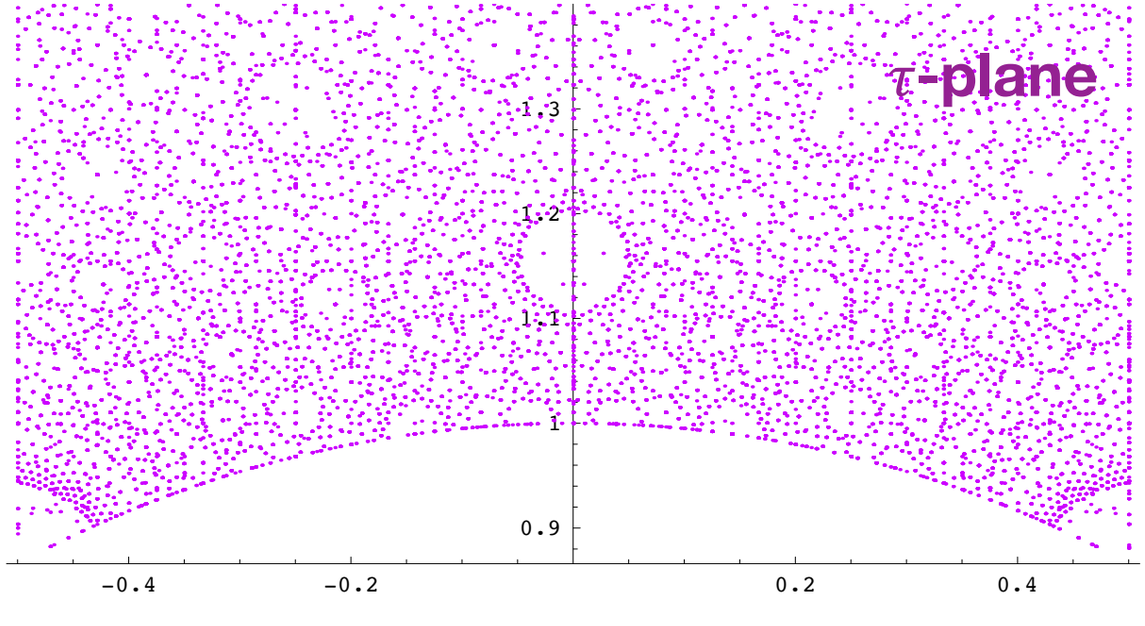
τ-plane

x=1-ψ plane



Flux vacua of an orientifold of CY hypersurface in $WP^4_{1,1,1,1,4}$

[Giryavets, Kachru, Tripathy]



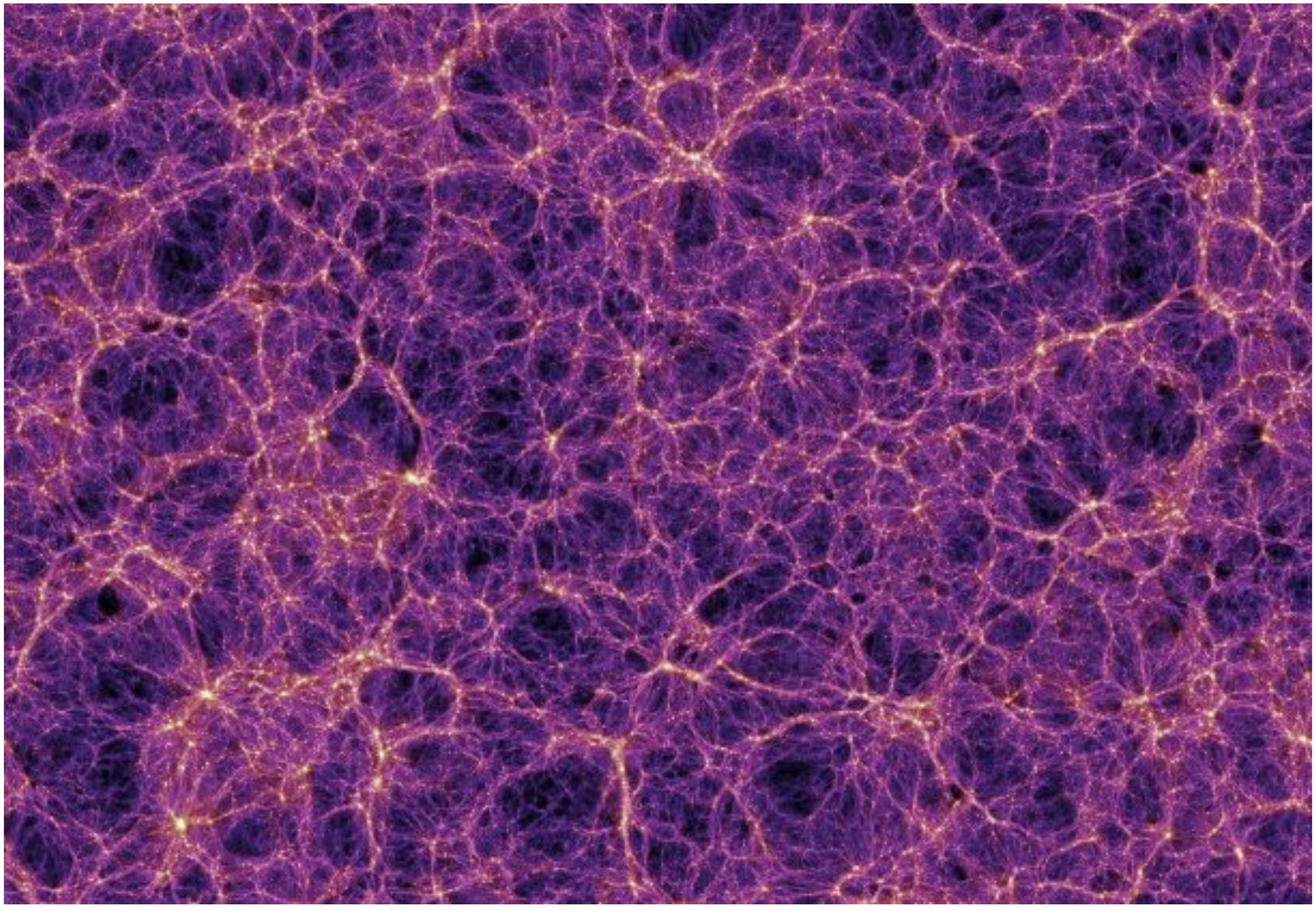
τ-plane

Toroidal Flux vacua with W=0

[DeWolfe, Giryavets, Kachru, Taylor]

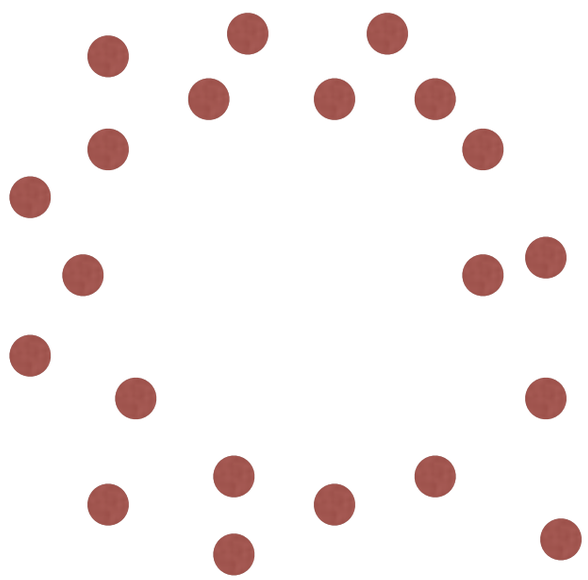
Distribution of Large Scale Structure

Similar **clustering** and **void** features also appear in LSS:



Topological Data Analysis

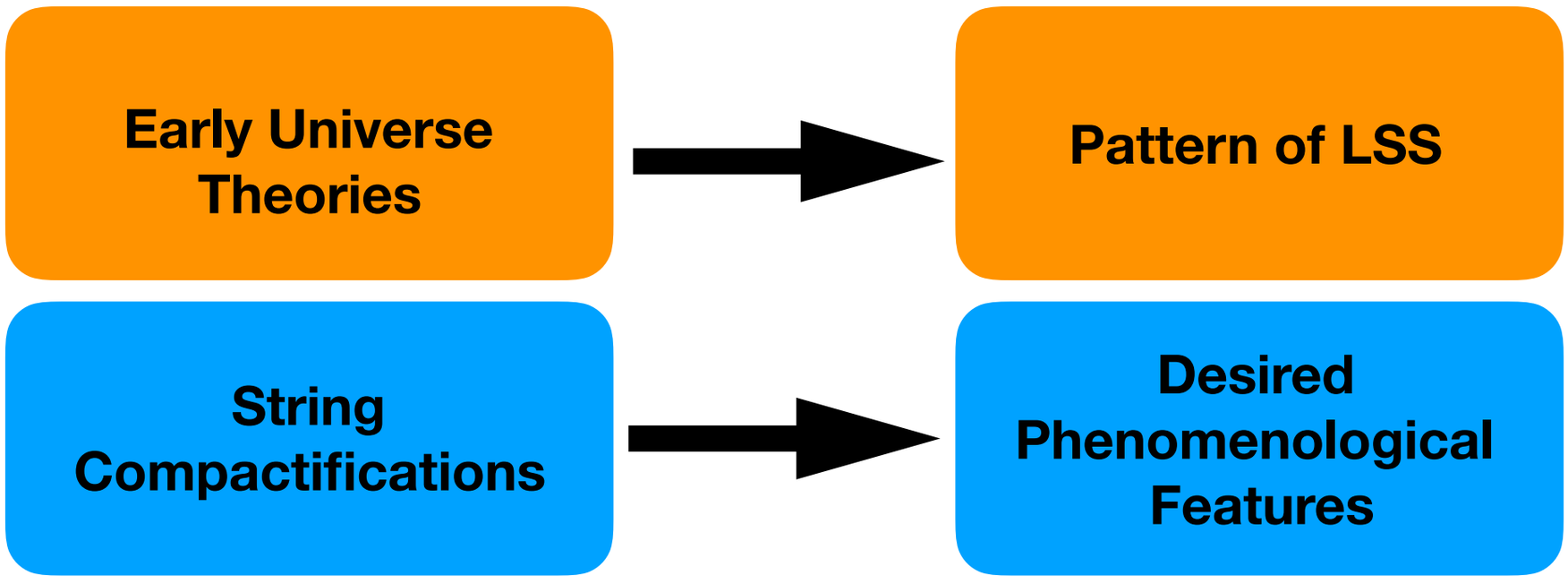
- When the space of data is huge, we cannot simply “visualize” the structure of data. We need a systematic diagnostic tool.
- Topological data analysis (TDA) is a systematic tool in applied topology to diagnose the “shape” of data.
- To turn a discrete set of data points (point cloud) into a topological space, we need a notion of ***persistence***.



Vary simplicial complexes formed by the point cloud with continuous parameters (filtration parameters)

Topological Data Analysis

- TDA is widely used in other fields, e.g., imaging, neuroscience, and drug design. It is well suited for machine learning.
- From the persistent homology of the point cloud, we can test e.g., the effectiveness of drugs. Similarly, we can test:



- A **selector algorithm** is often used due to the huge volume of data. We can test these algorithms on cosmological datasets and the string vacua.

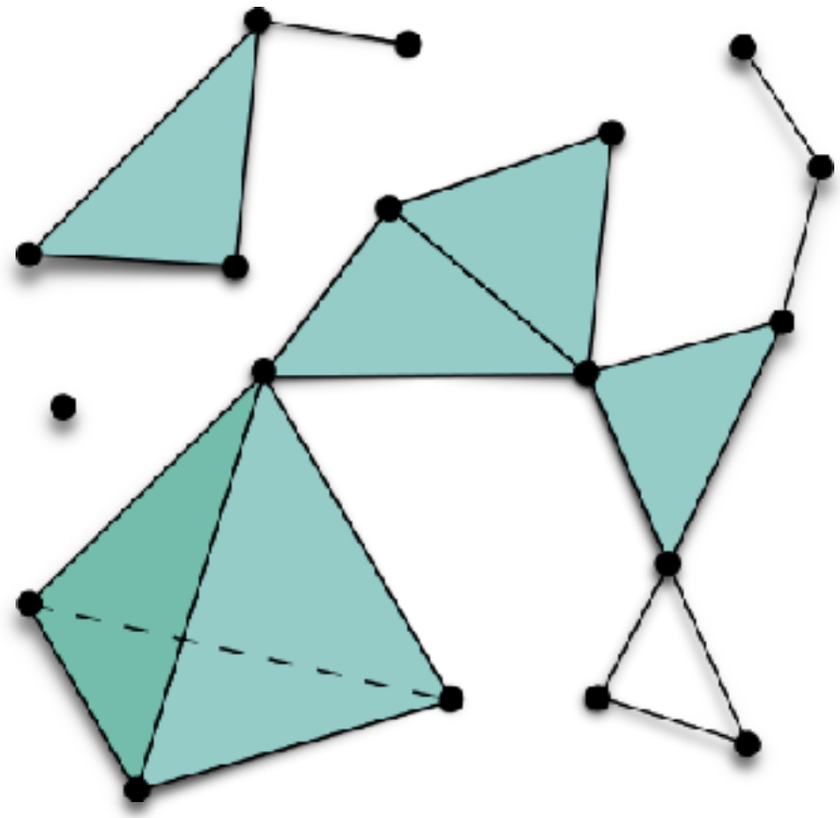
Plan of this talk

- Introduce the basic concepts of **topological data analysis**: persistent homology, barcodes, and persistent diagrams.
- Applying TDA to constrain **primordial non-Gaussianities**.
- Back to the **String Data Project**. Computing the persistent homology of string vacua to analyze their structure.
- This talk is based on several projects done in collaboration with

- "Persistent Homology and Local Non-Gaussianity", A. Cole, GS, MAD-TH-17-11, to appear.
- "Topological Data Analysis for the String Landscape", A. Cole, GS, MAD-TH-17-12, work in progress.
- **...**

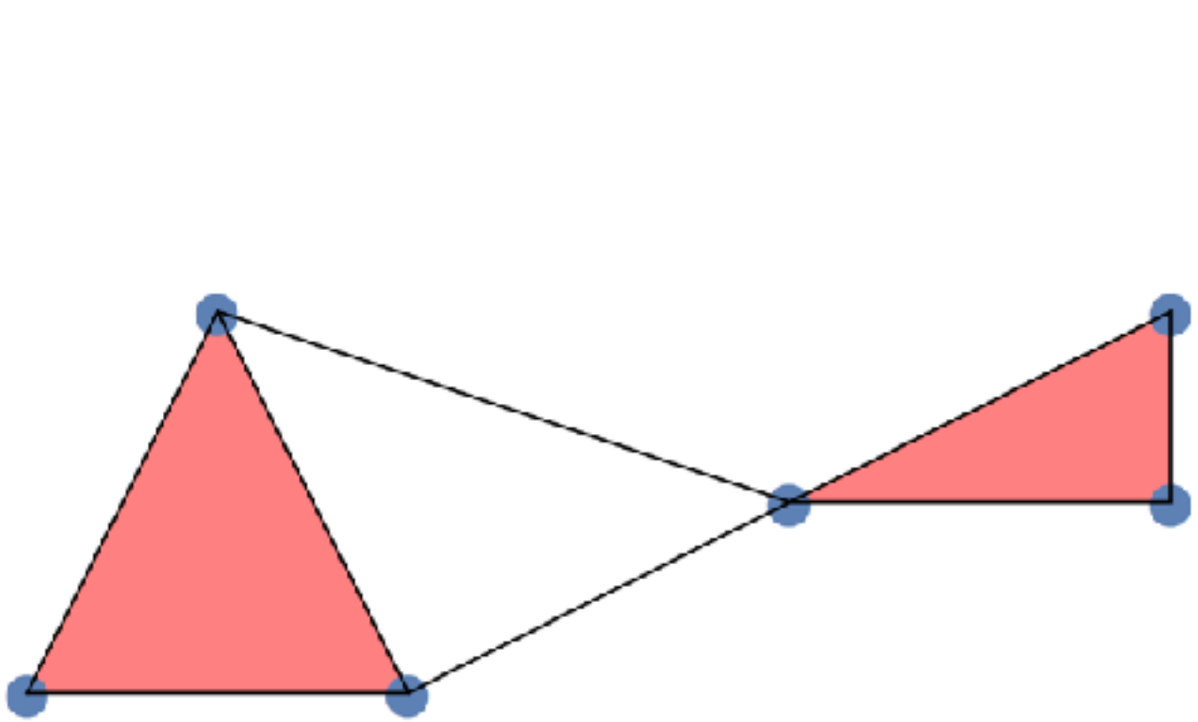
Simplicial Complexes

- In $\mathbb{R}^3$, simplices are vertices, edges, triangles, and tetrahedra
- Simplicial complexes are collections of simplices that are:
  - Closed under intersection of simplices
  - Closed under taking faces of simplices
- Combinatorial representations — easy calculations for computers

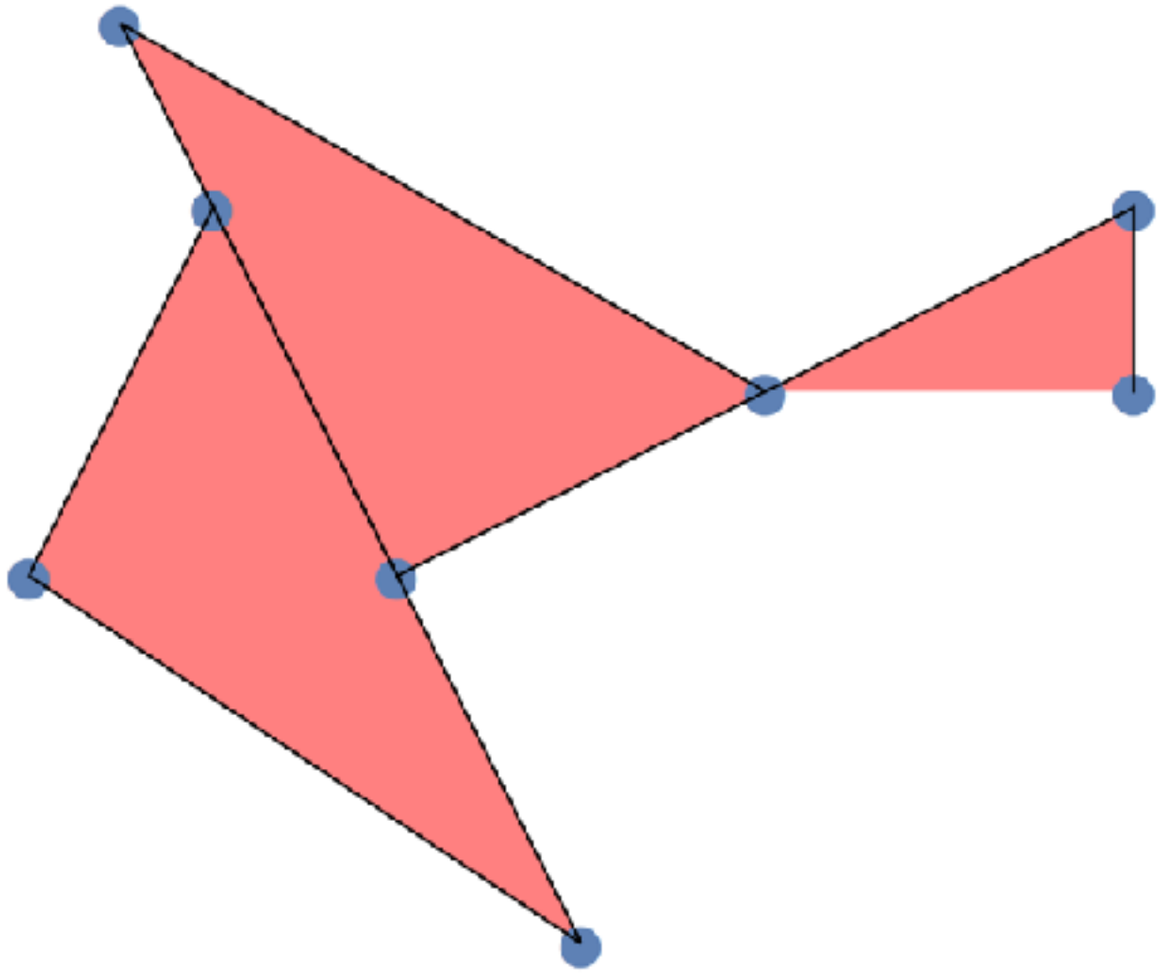


Source: Wikipedia, “Simplicial Complex”

Simplicial Complexes



A Simplicial Complex



Not a Simplicial Complex

Simplicial Homology

- Given a simplicial complex, define a boundary operator ∂_p that maps p-simplices to (p-1)-simplices

  - We want to count independent p-cycles (i.e. p-loops) that are not boundaries of higher-dimensional objects

- Group theoretic: $Z_p = \ker \partial_p$, $B_p = \text{im } \partial_{p+1}$,

$$H_p \equiv Z_p / B_p$$

- Betti numbers: $\beta_p \equiv \text{rank} H_p$

	$\beta_0 = 1$	$\beta_0 = 1$
	$\beta_1 = 1$	$\beta_1 = 0$

  - 0-th Betti number is number of connected components

  - p-th Betti number is number of independent p-loops

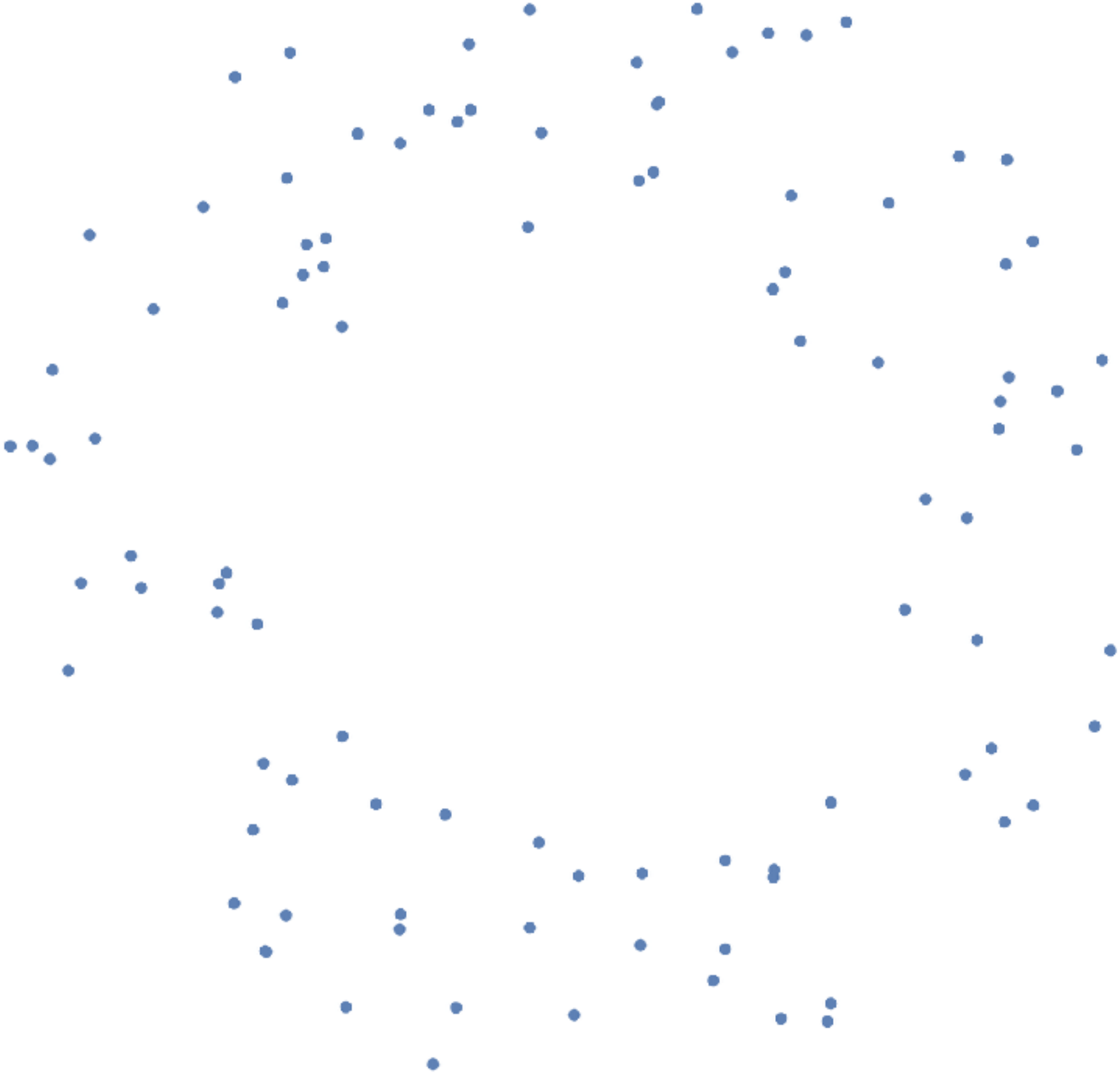
- In practice, homology calculation is a matrix reduction

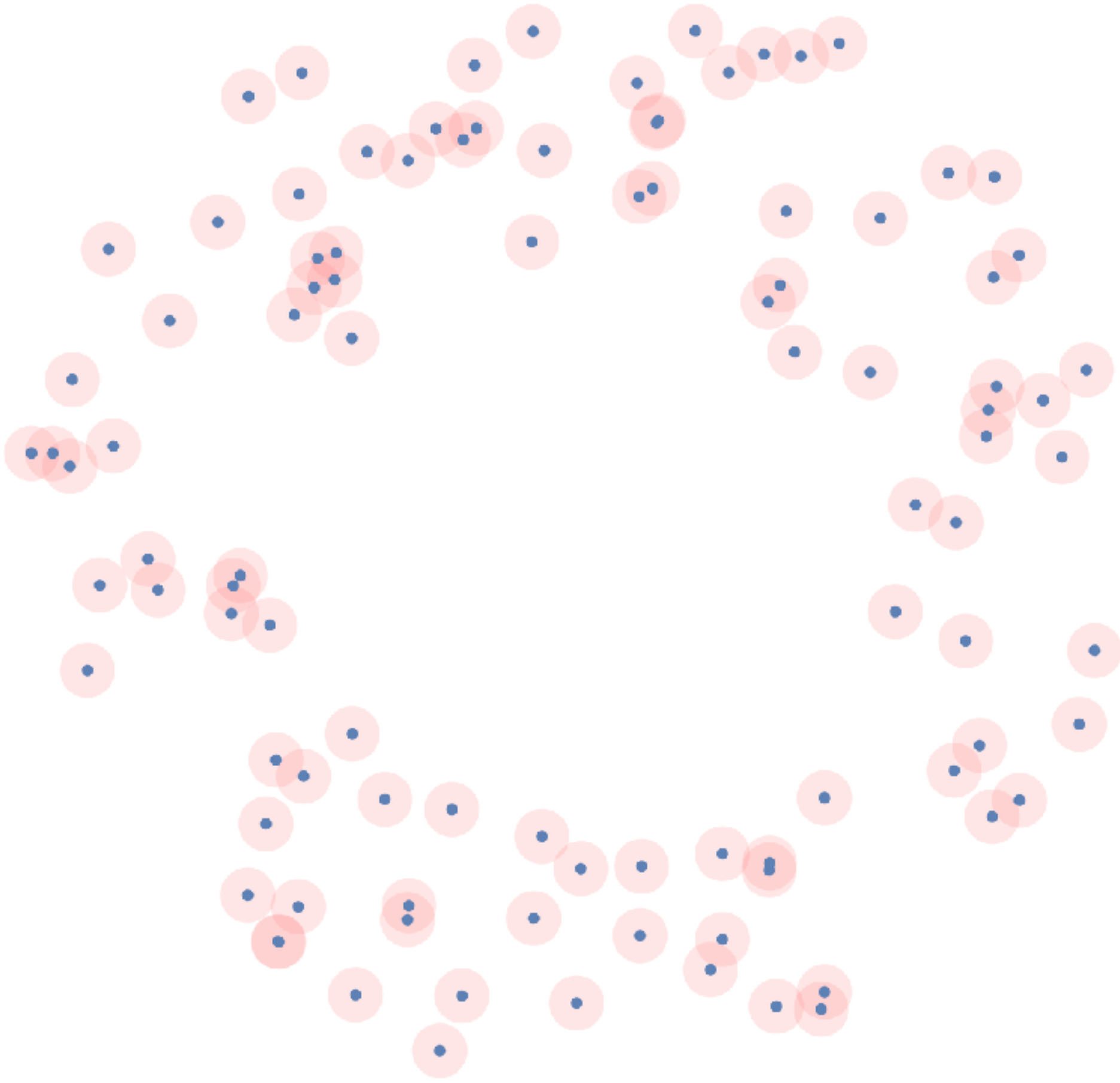
Persistence

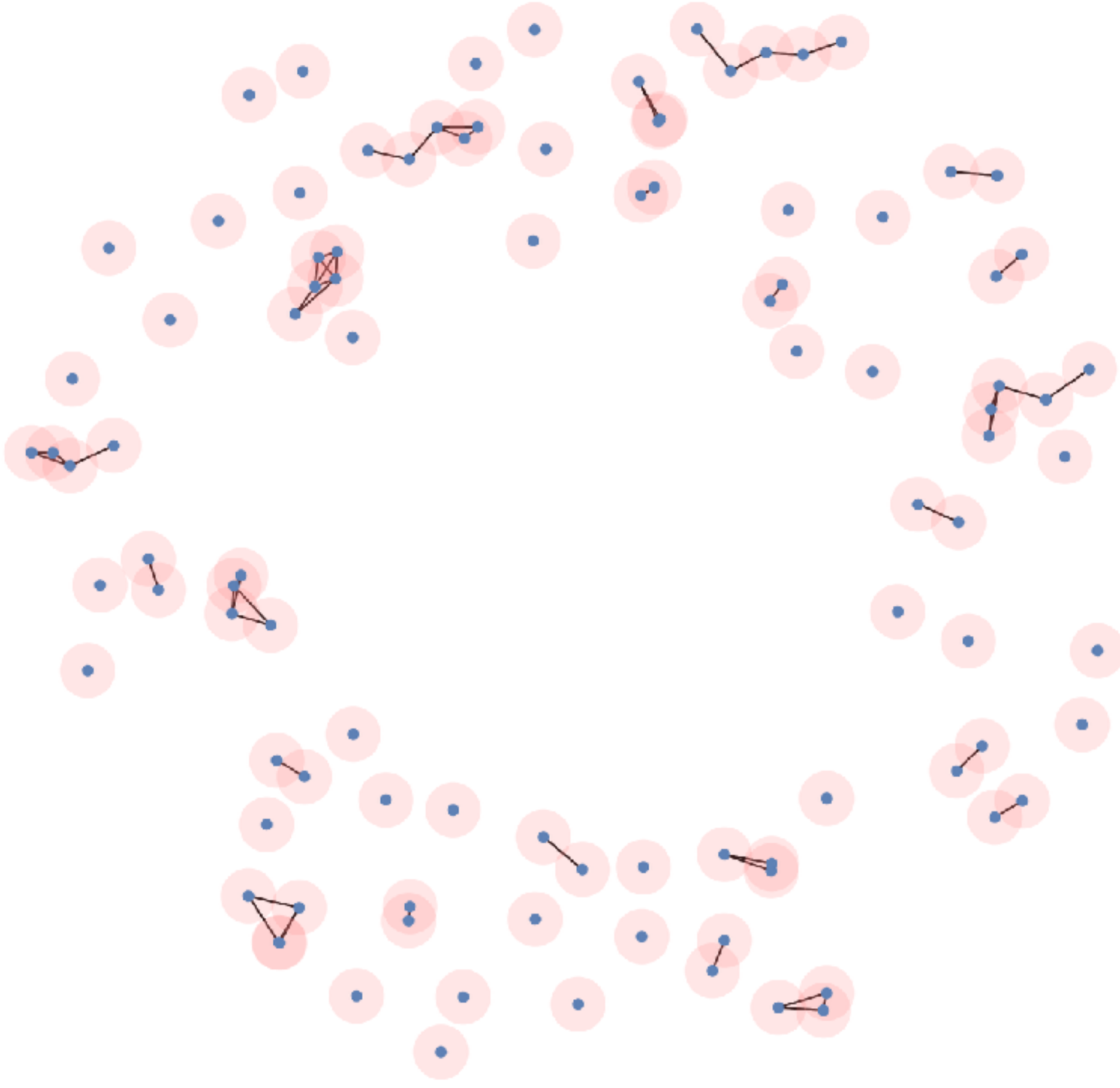
- How to choose simplicial representation of our data?
- *Persistent* homology: vary simplicial representation Σ_ν of data with some *filtration parameter* ν such that

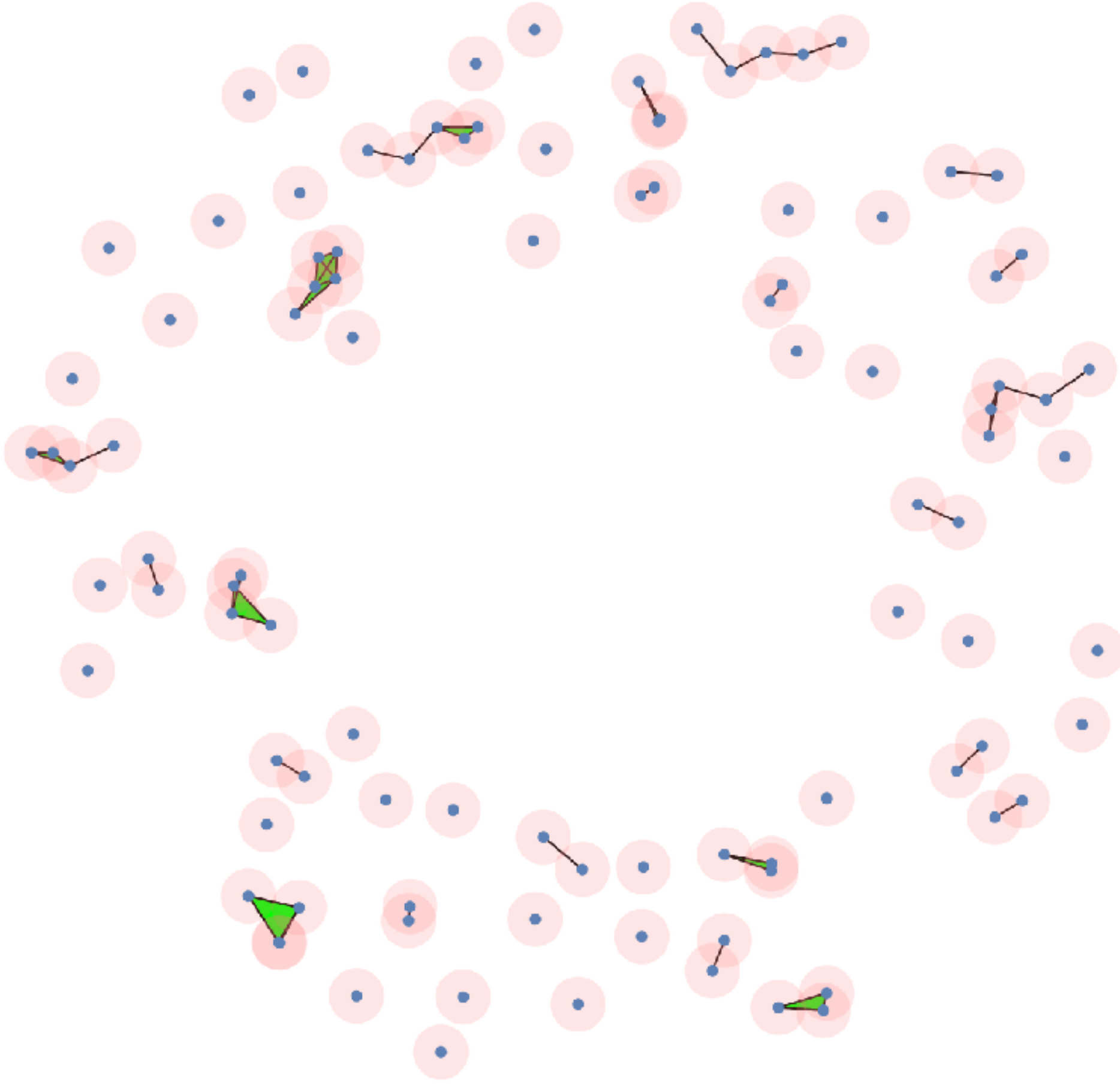
$$\nu_1 \leq \nu_2 \implies \Sigma_{\nu_1} \subseteq \Sigma_{\nu_2}$$

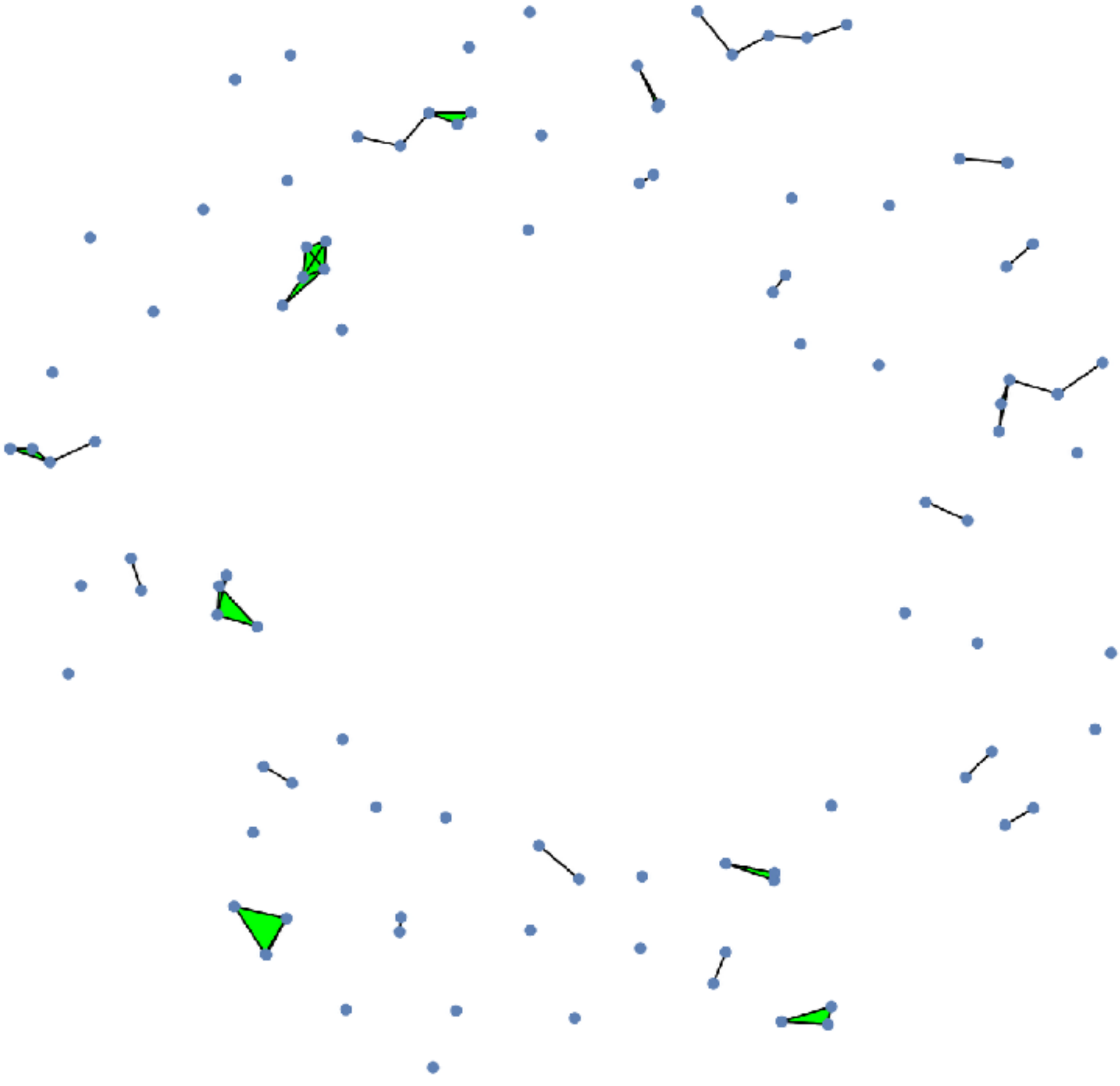
  - Track each distinct feature's lifetime (birth and death)
- Intuition: "real" topological features *persist*, short-lived features are noise
- Procedure is stable against perturbations to data **[Cohen-Steiner 2005]**

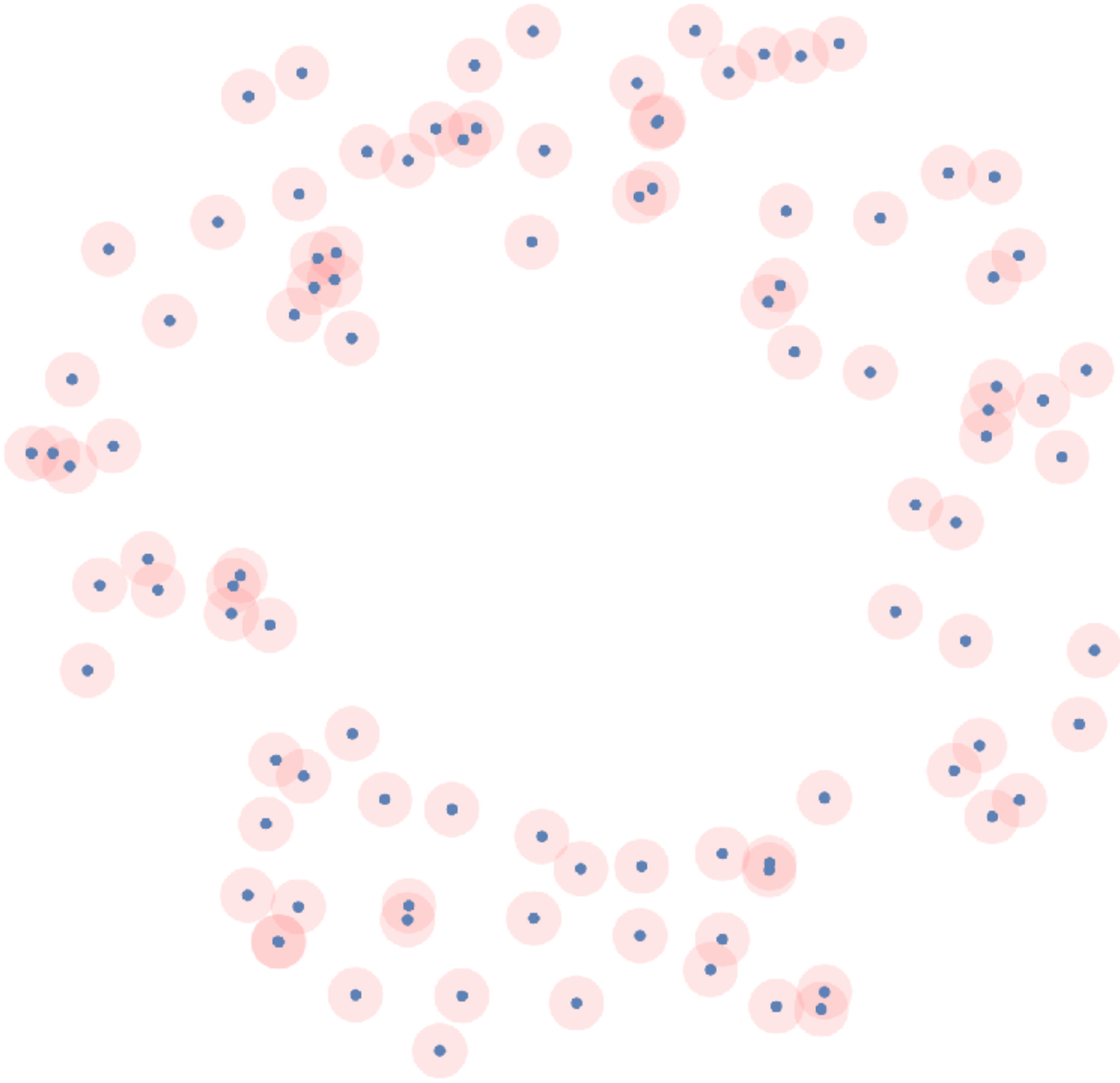


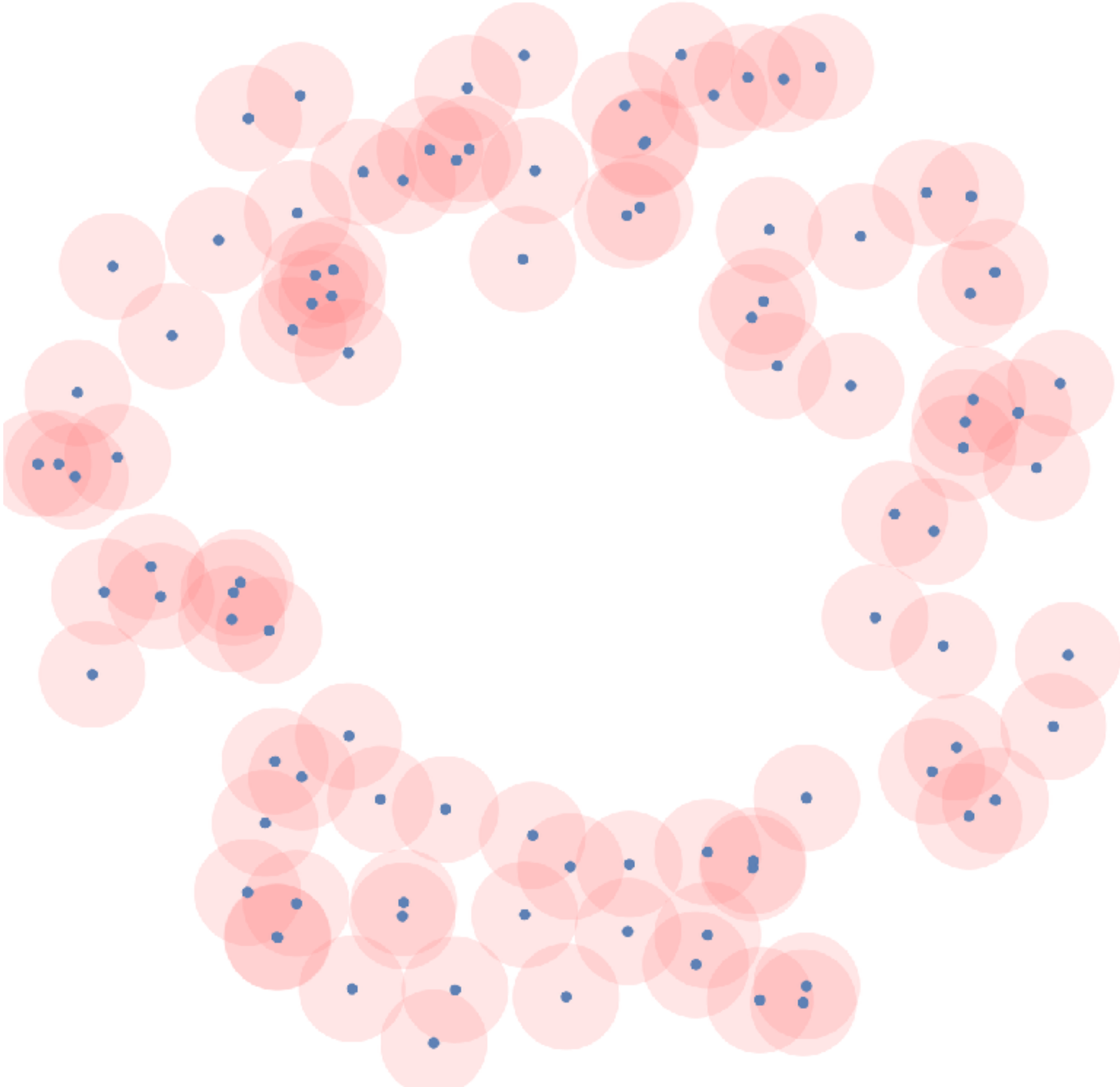


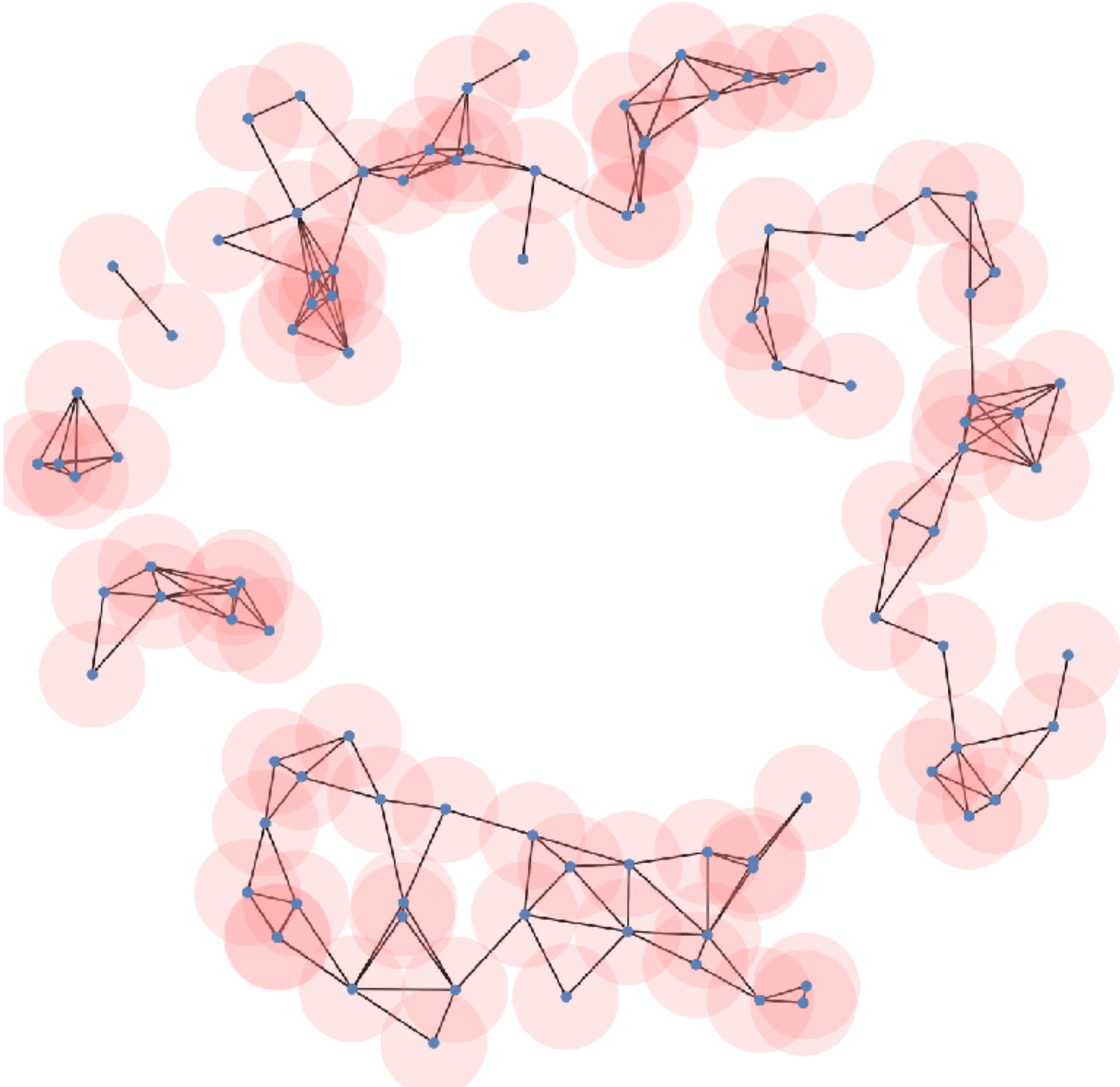


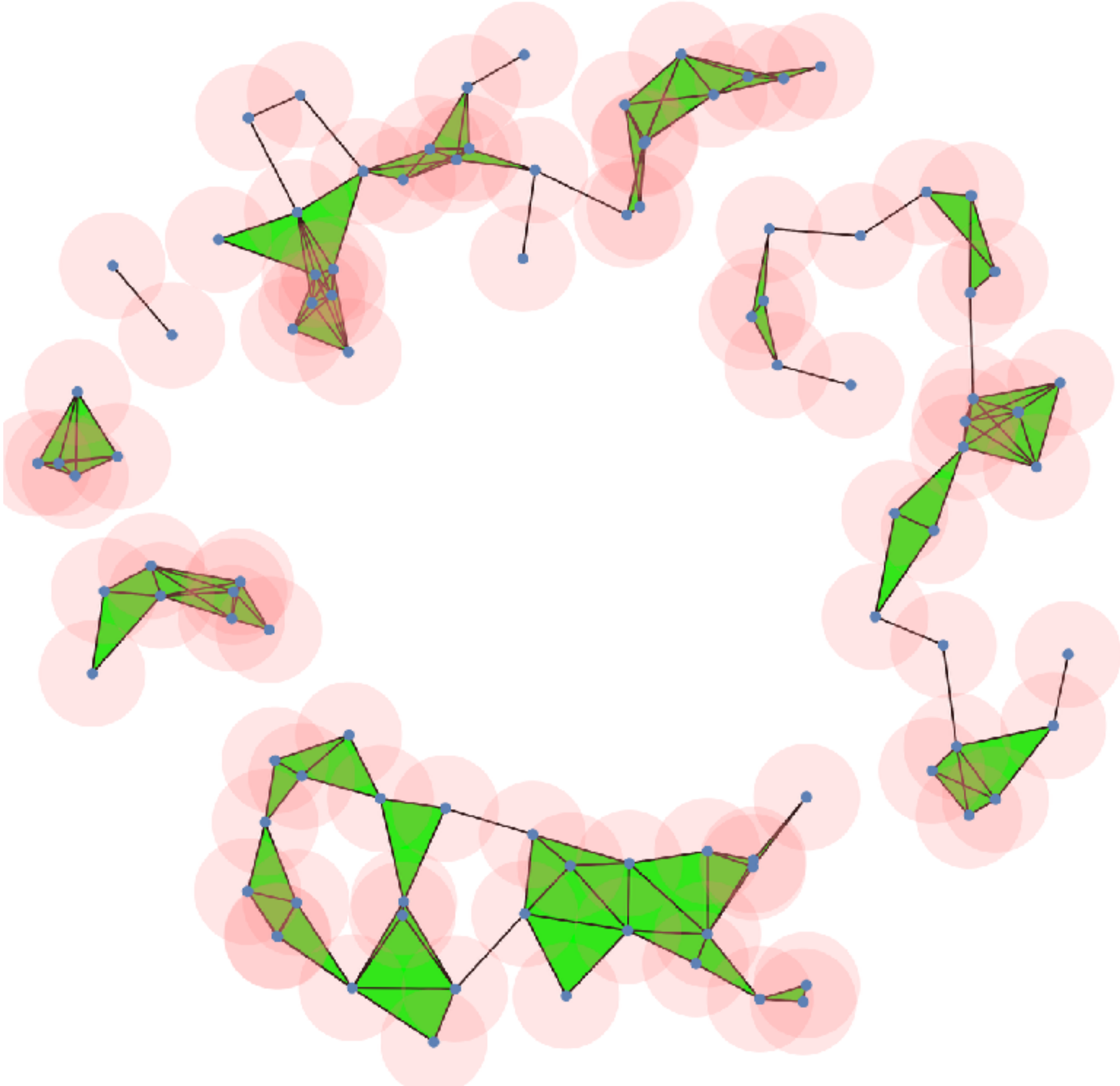


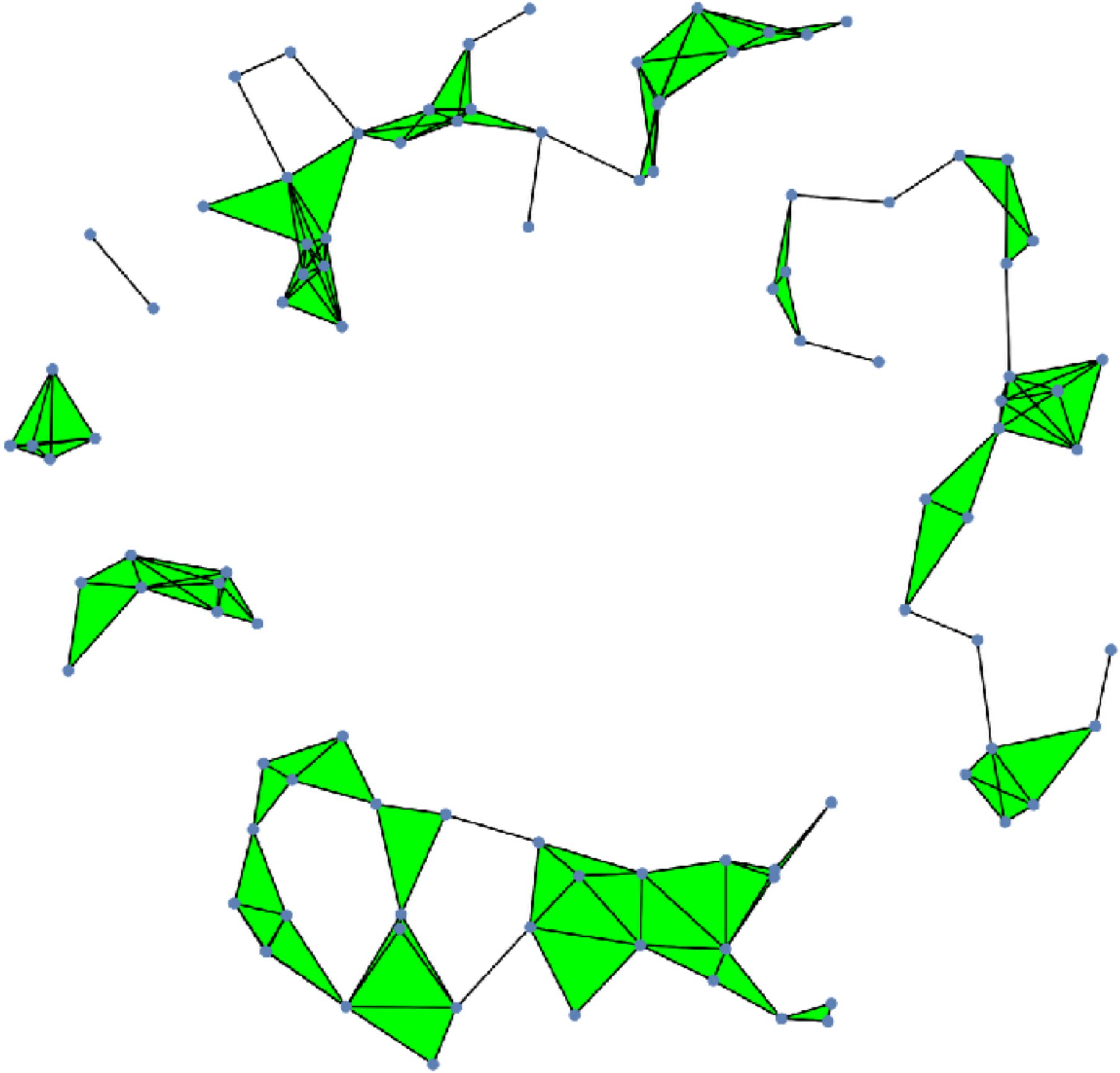


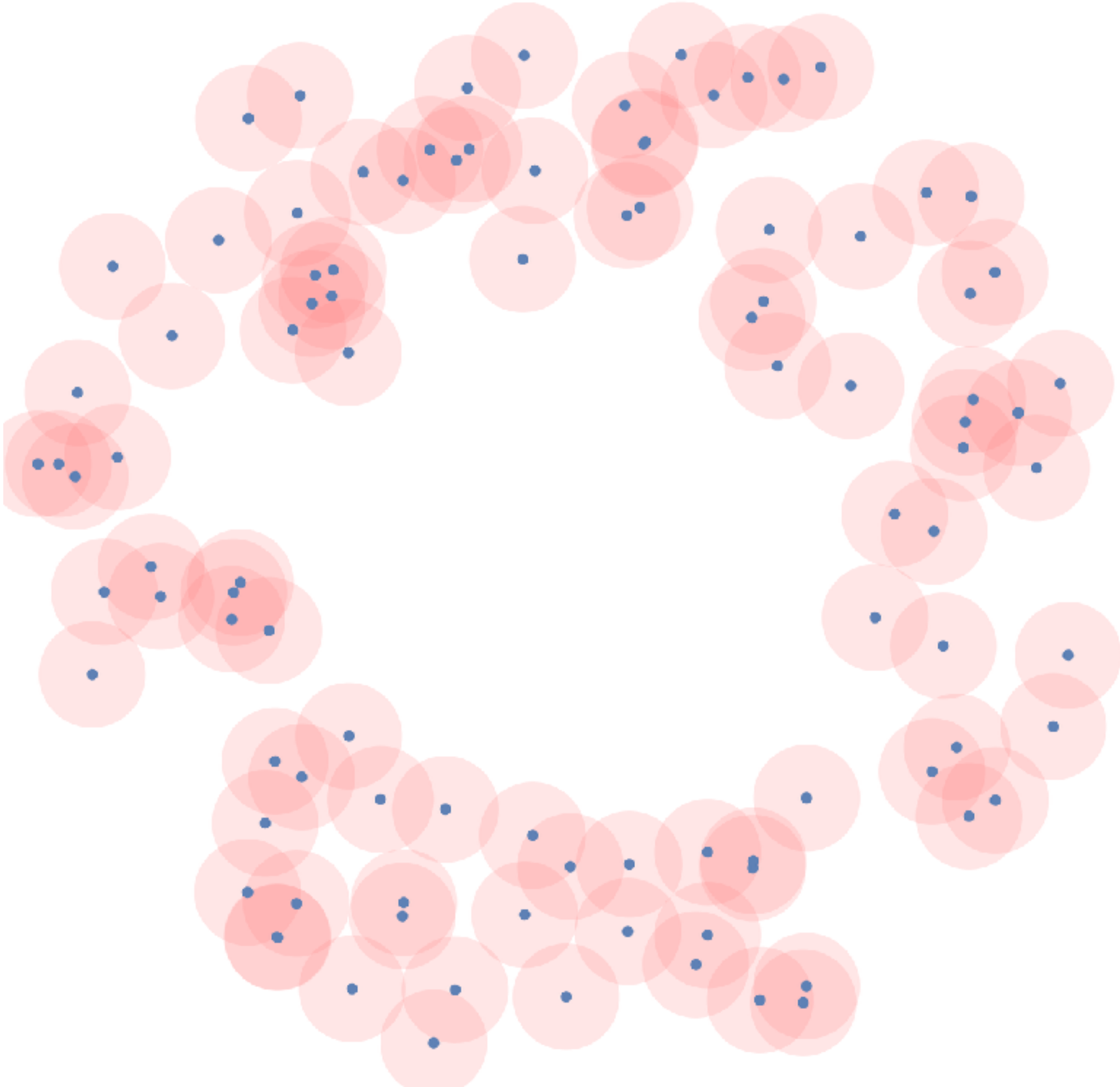


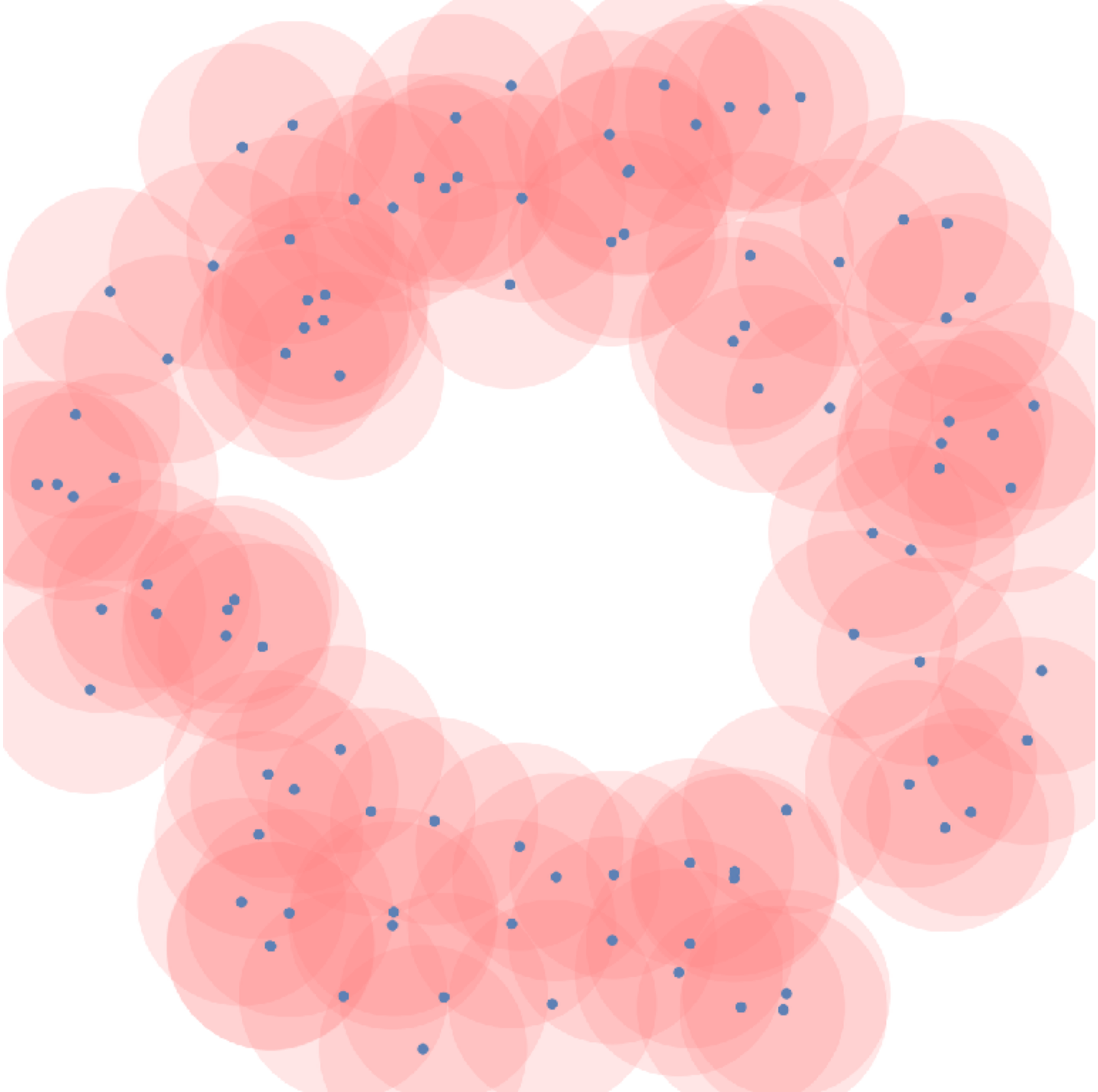


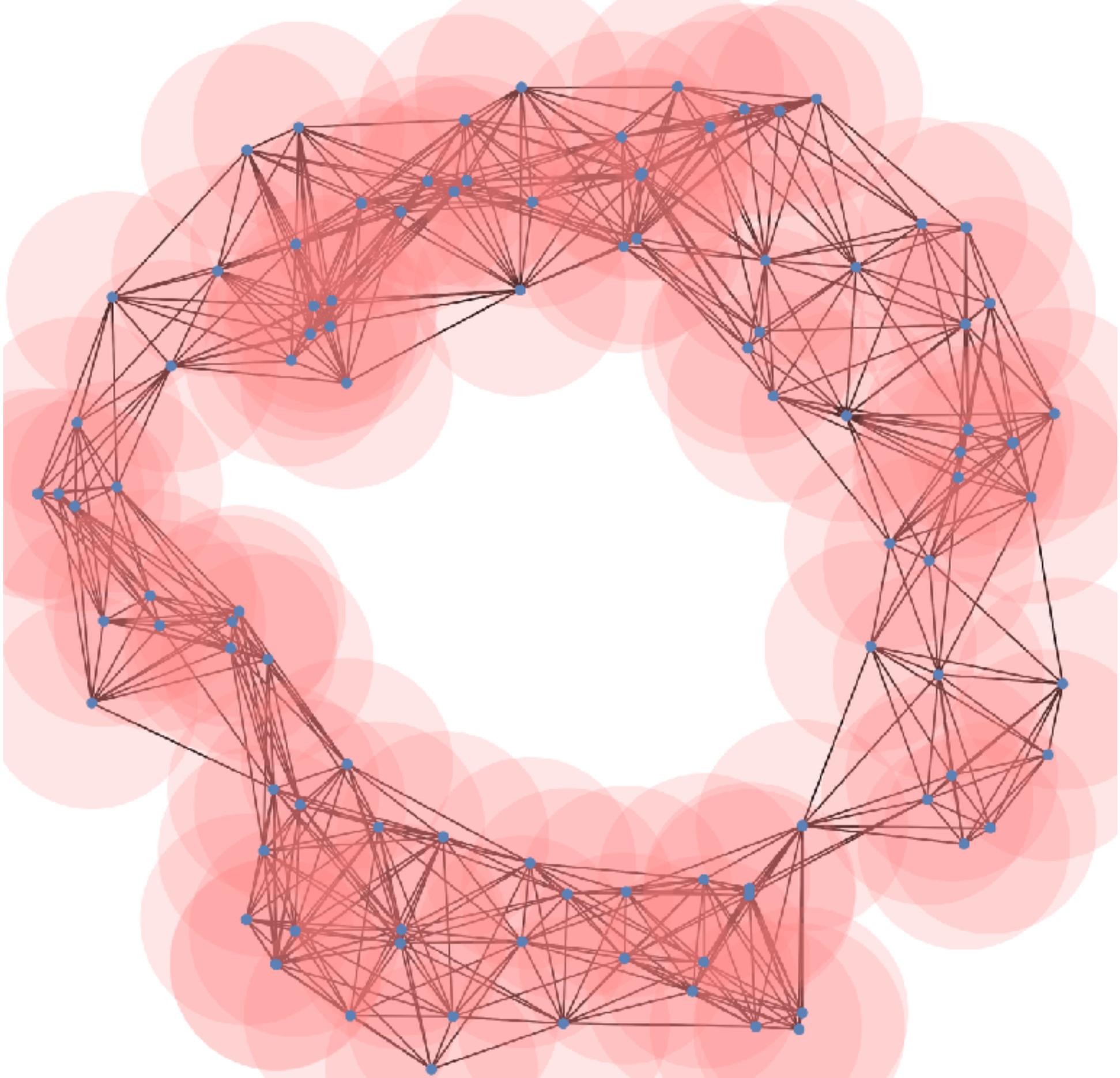


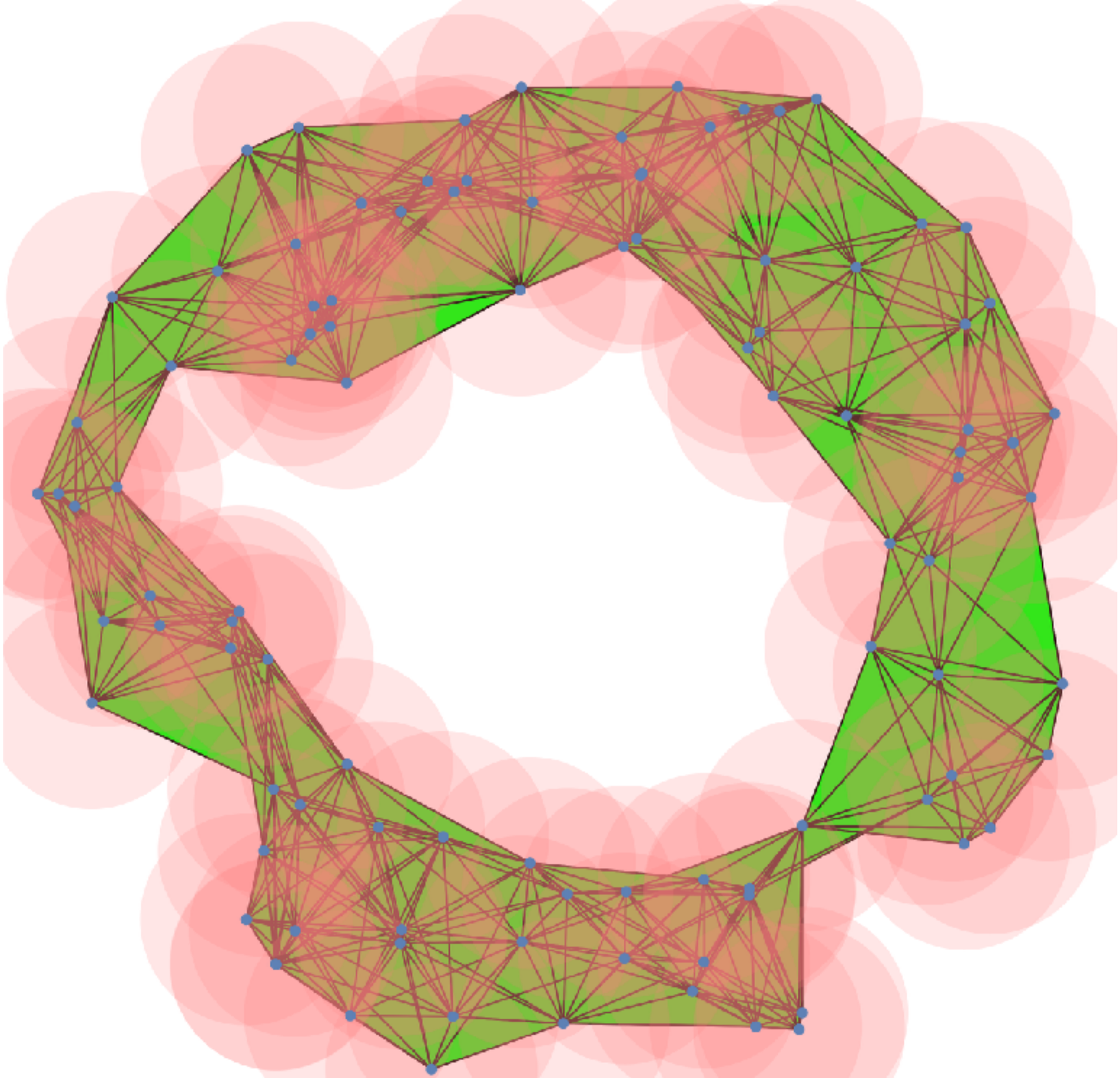


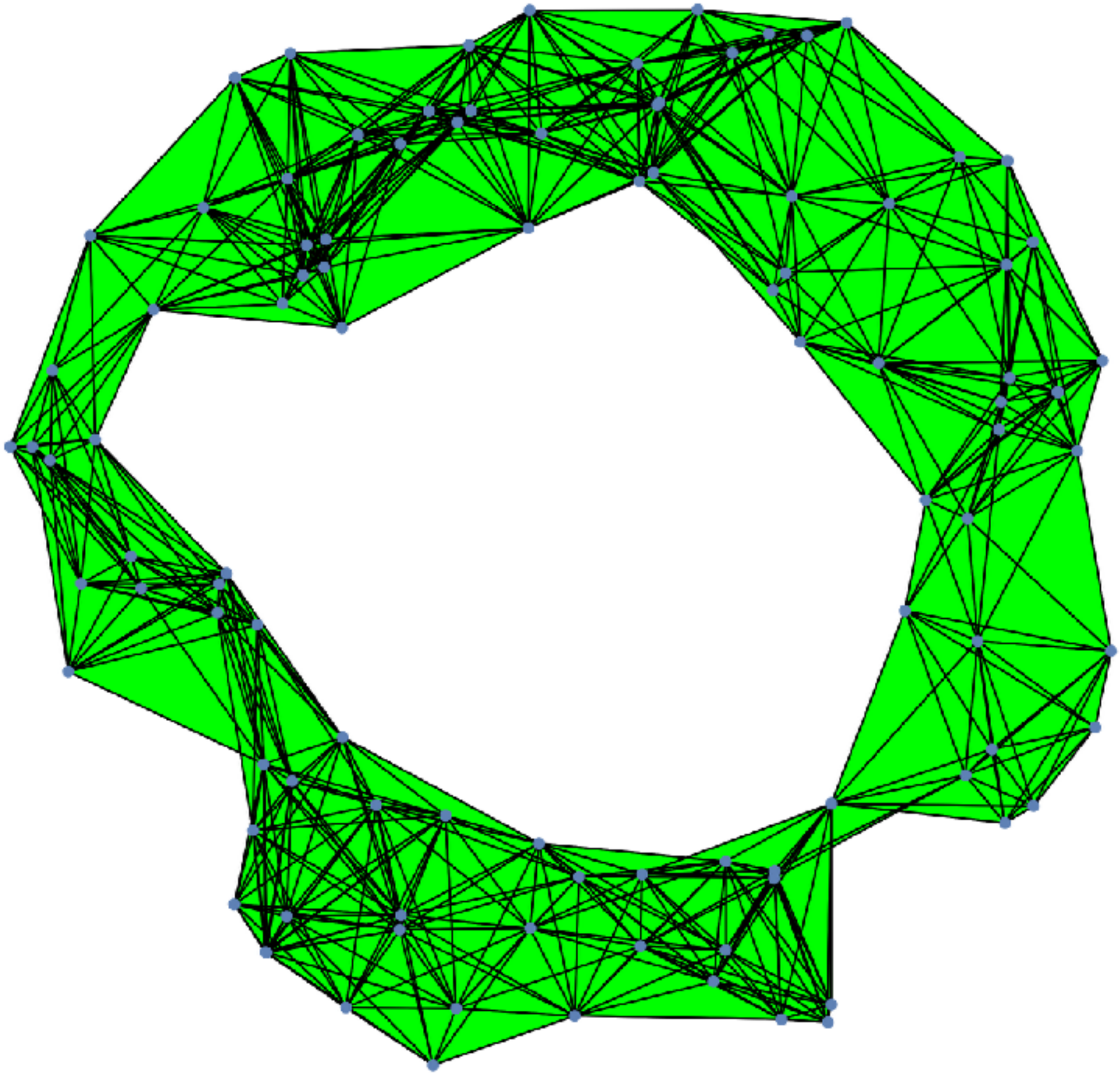






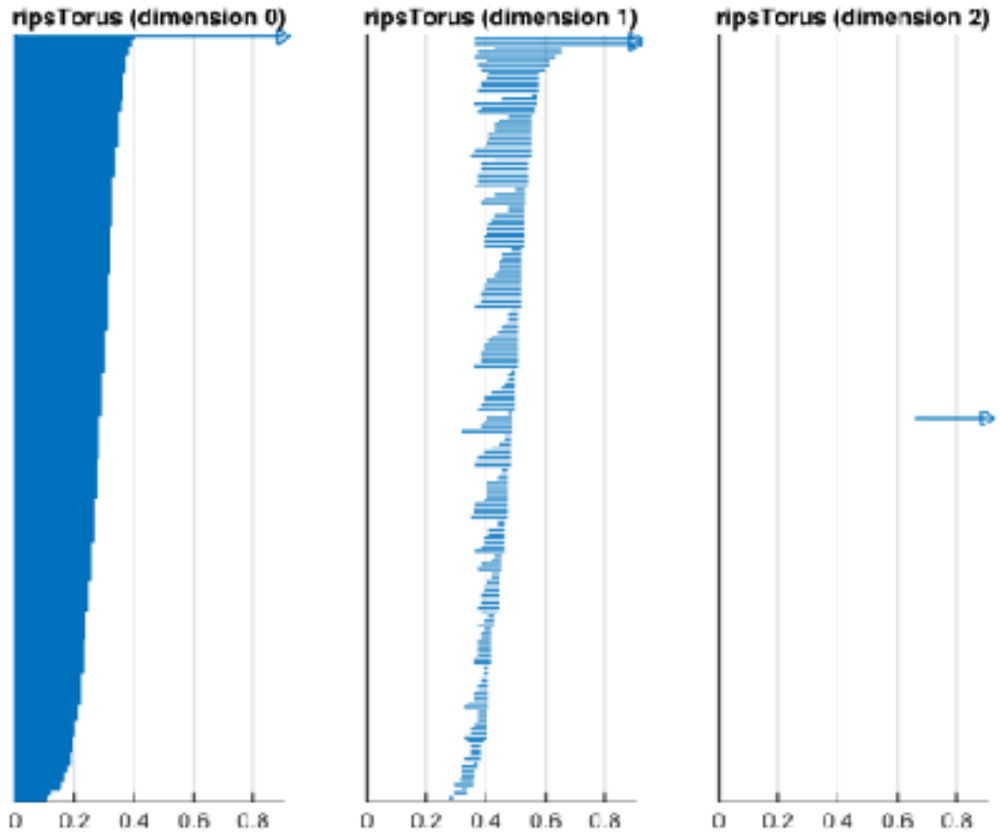




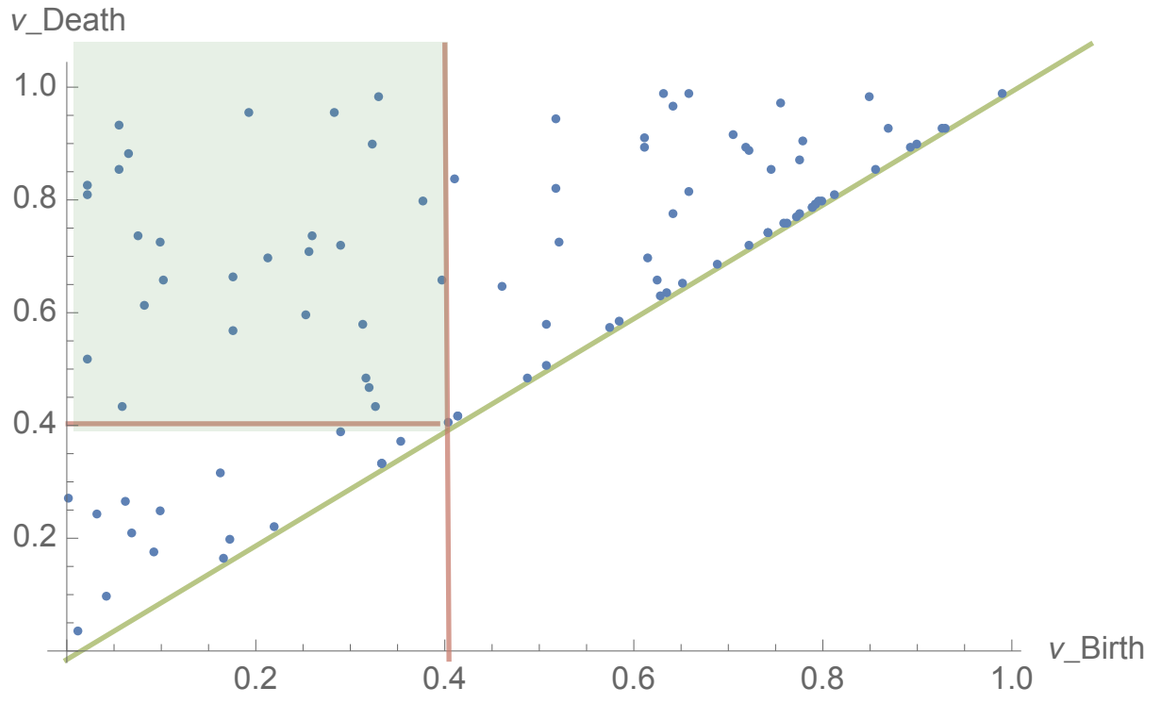


Visualizing Persistent Homology

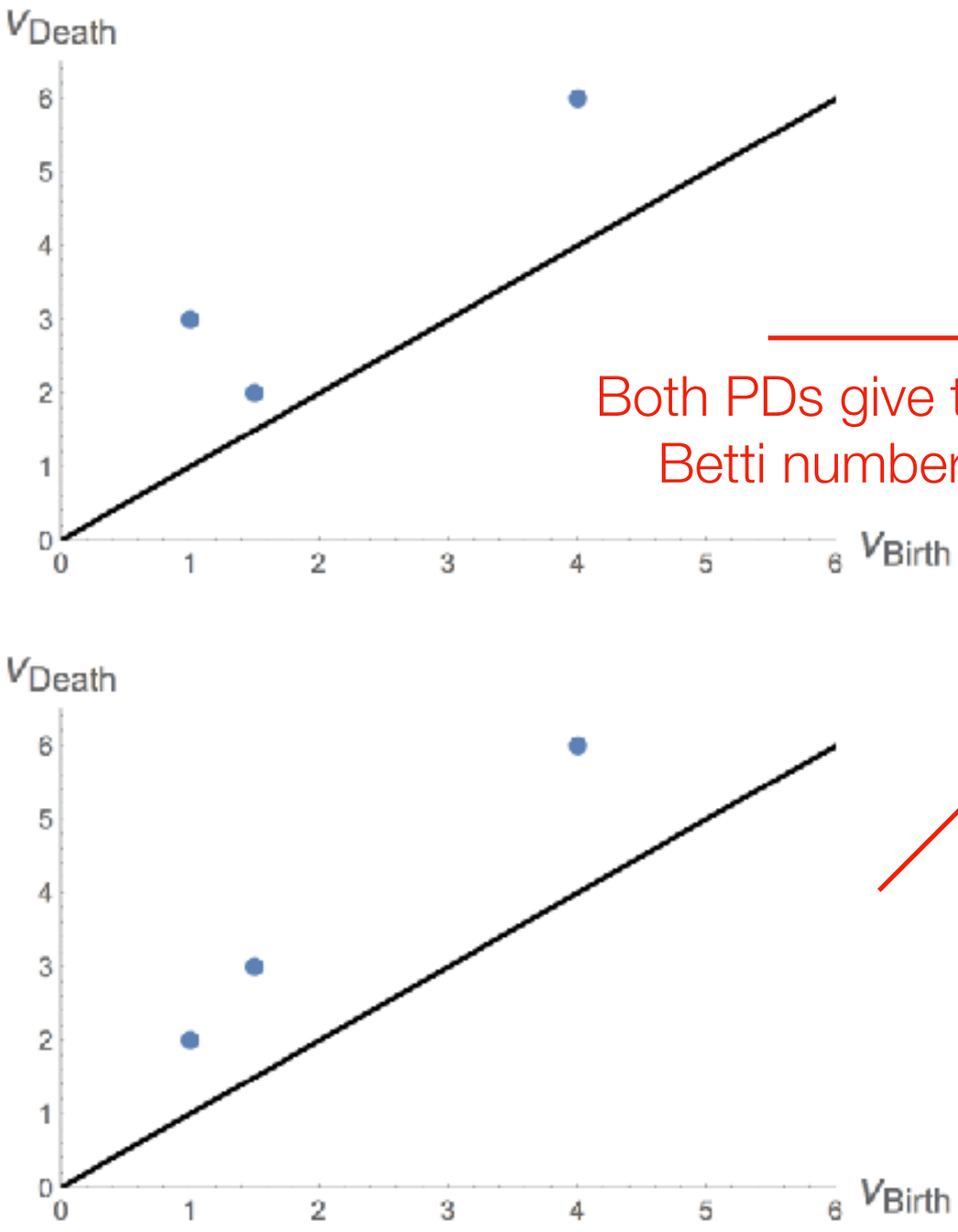
- **Barcodes**:
  - Each horizontal line represents an independent cycle contributing to a particular Betti number (i.e. a connected component, loop, void…)
  - Lines start at birth and end at death
  - To calculate Betti number, make vertical slice and count intersections



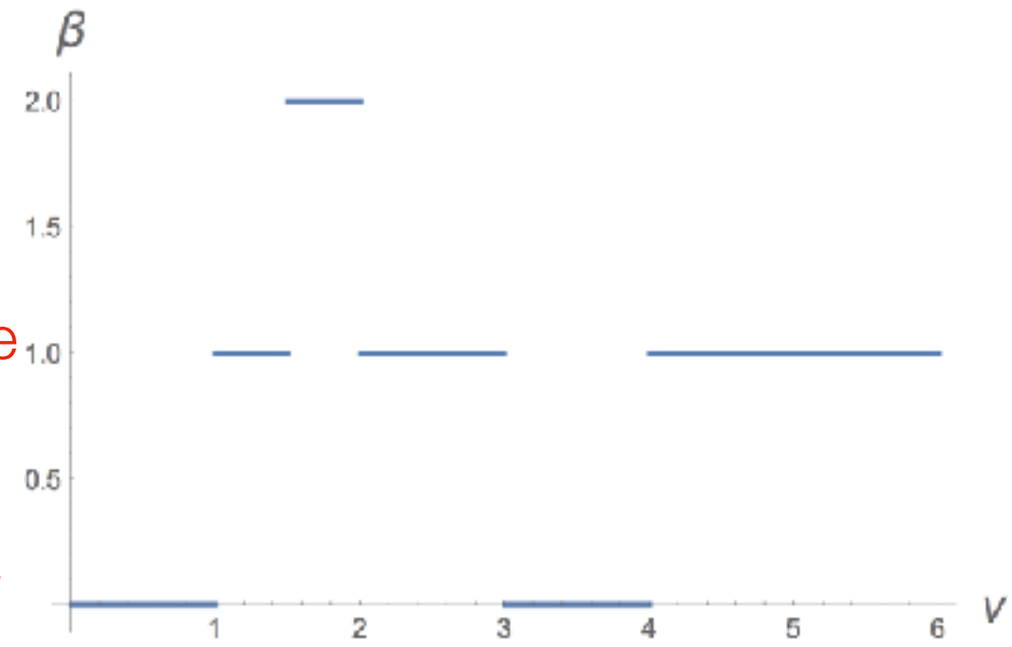
- **Persistence diagrams**:
  - Scatter plot, each point representing an independent cycle
  - Calculate Betti number by counting “living” cycles



Persistence diagrams contain more information than Betti number curves!



Both PDs give the same Betti number curve

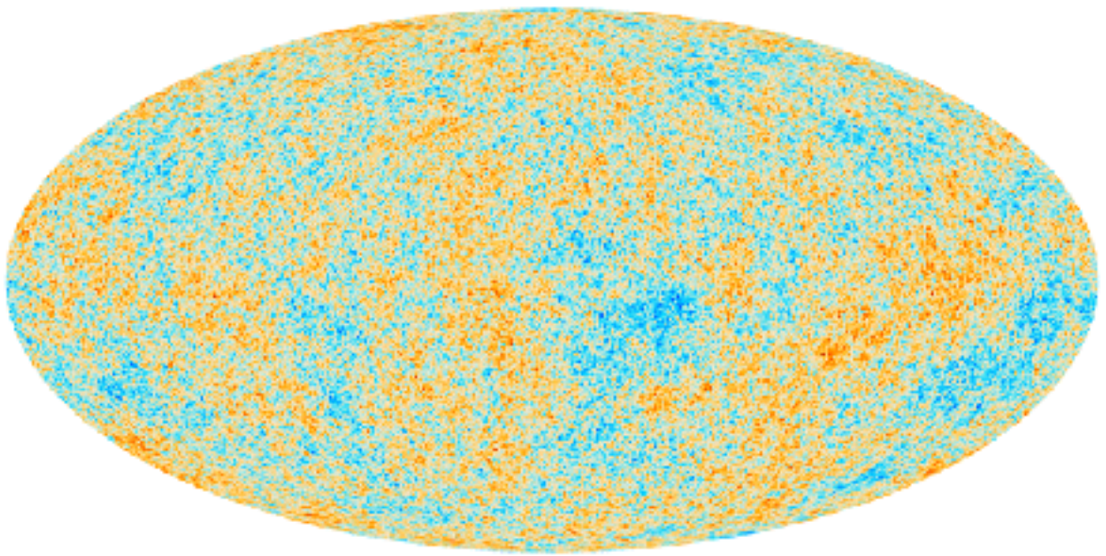


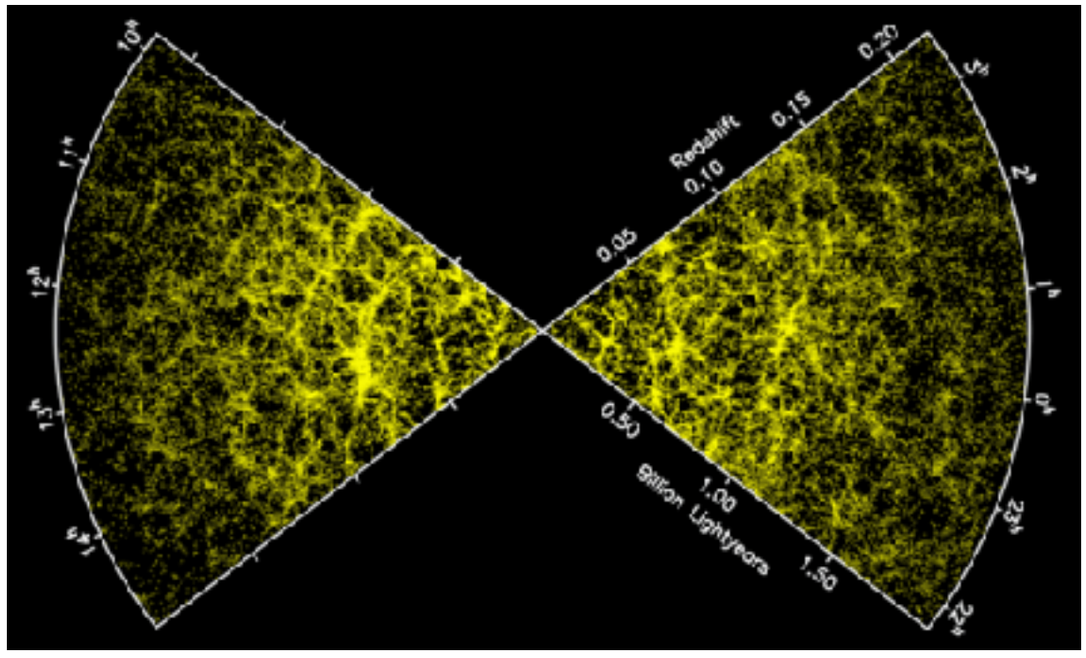
We can exploit this to improve CMB data analysis

Inflation

[Starobinsky];[Guth];[Linde];[Albrecht, Steinhardt];…

- Period of **accelerated expansion** in early universe
  - Solves flatness, horizon, and monopole problems
  - Predicts **nearly scale-invariant**, **Gaussian** curvature fluctuations
    - Source anisotropies in CMB, inhomogeneities in LSS





- A myriad of models. Taxonomy done mostly through their observables (n_s, r)

Anisotropies

- The lowest order correlation we can extract from the anisotropies is the **power spectrum**

$$\left\langle 0 \left| \hat{\mathcal{R}}_{\mathbf{k_1}} \hat{\mathcal{R}}_{\mathbf{k_2}} \right| 0 \right\rangle = (2\pi)^3 P_{\mathcal{R}}(k_1) \delta(\mathbf{k}_1 + \mathbf{k}_2) \qquad \Delta^2_{\mathcal{R}} = \left(\frac{k^3}{2\pi^2} \right) P^2_{\mathcal{R}} \propto k^{n_s - 1}$$

- For a Gaussian theory, the power spectrum dictates all higher-pt correlations.
- However, the inflationary fluctuations are not perfectly Gaussian.
- The leading **non-Gaussianity** is the **bispectrum**:

$$\langle 0 | \hat{\mathcal{R}}_{\mathbf{k_1}} \hat{\mathcal{R}}_{\mathbf{k_2}} \hat{\mathcal{R}}_{\mathbf{k_3}} | 0 \rangle = (2\pi)^3 \delta^3(\mathbf{k_1} + \mathbf{k_2} + \mathbf{k_3}) F(\mathbf{k_1}, \mathbf{k_2}, \mathbf{k}_3)$$

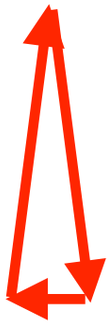
- Scaling and symmetries imply that F(**k_1**, **k_2**, **k_3**) is fixed by an overall **size** ~ f_{NL} and its ‘’**shape”** F(1, k_2/k_1, k_3/k_1).
- More **powerful discriminator** of inflationary models.

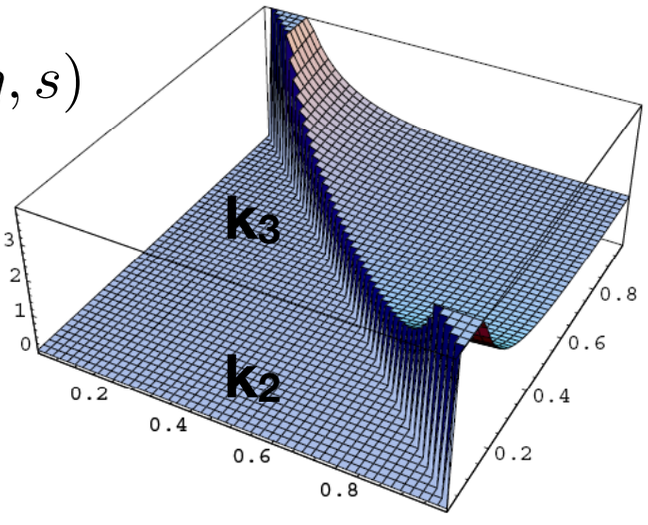
Non-Gaussianities

- The bispectrum for **single field slow-roll** inflation was computed in **[Maldacena, '02];[Acquaviva et al, '02]**; its size is $f_{NL} \sim O(\varepsilon,\eta)$:
- The bispectrum for **general single field inflation** was found to be parametrized by 5 parameters **[Chen, Huang, Kachru, GS, '06]**:

Local shape

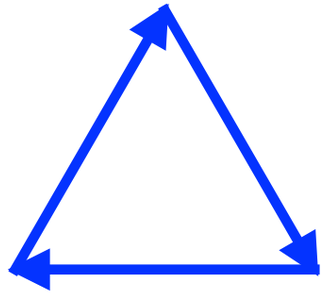
$$f_{NL}^{local} \sim \mathcal{O}(\epsilon, \eta, s)$$

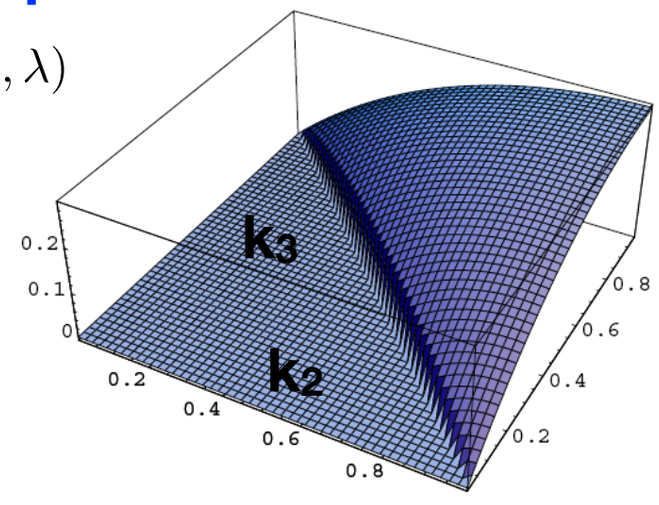




Equilateral shape

$$f_{NL}^{equil} \sim \mathcal{O}(\frac{1}{c_s^2} - 1, \lambda)$$





- More shapes if the inflationary vacuum is **non-Bunch Davis** or the potential is **oscillatory** (e.g. axion monodromy).

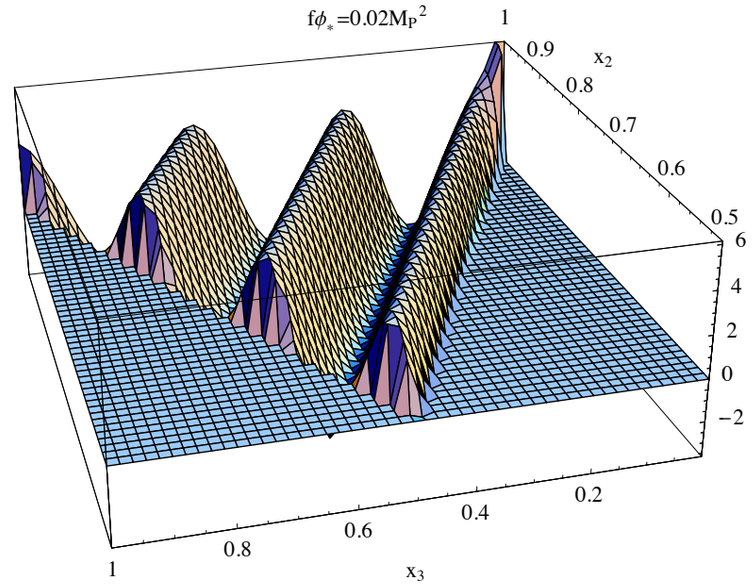
Measuring Non-Gaussianity

- **Harmonic space:** fits with *templates* of bispectrum, trispectrum, etc. One can define a "cosine" between distributions:

$$\cos(F_1, F_2) = \frac{F_1 \cdot F_2}{(F_1 \cdot F_1)^{1/2} (F_2 \cdot F_2)^{1/2}}$$

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

Resonant shape (axion monodromy)

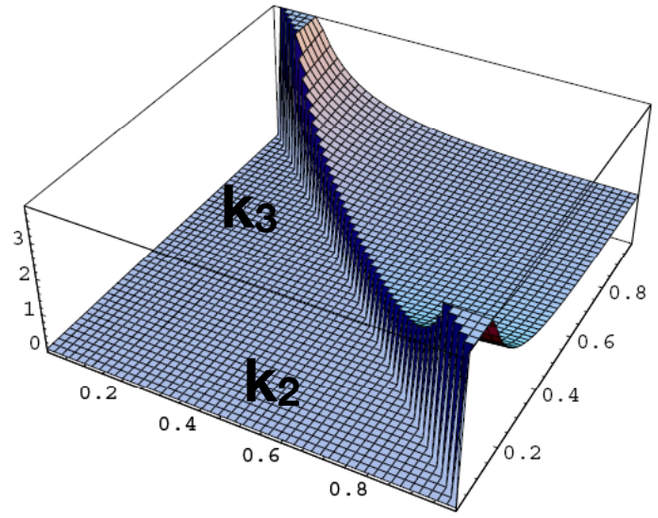


- Geometrical/topological: **Minkowski functionals** (for CMB: area fraction, length of boundaries, and genus of excursion sets)

- Current bound on non-Gaussianity (Planck '15):

$$f_{NL}^{local} = 2.5 \pm 5.7 \qquad f_{NL}^{equil} = -16 \pm 70$$

Local Non-Gaussianity



- Local shape of non-Gaussianity is generated by a local ansatz for primordial gravitational potential **[Komatsu-Spergel]**

$$\Phi = \Phi_G + f_{NL}\left(\Phi_G^2 - \left\langle \Phi_G^2 \right\rangle\right)$$

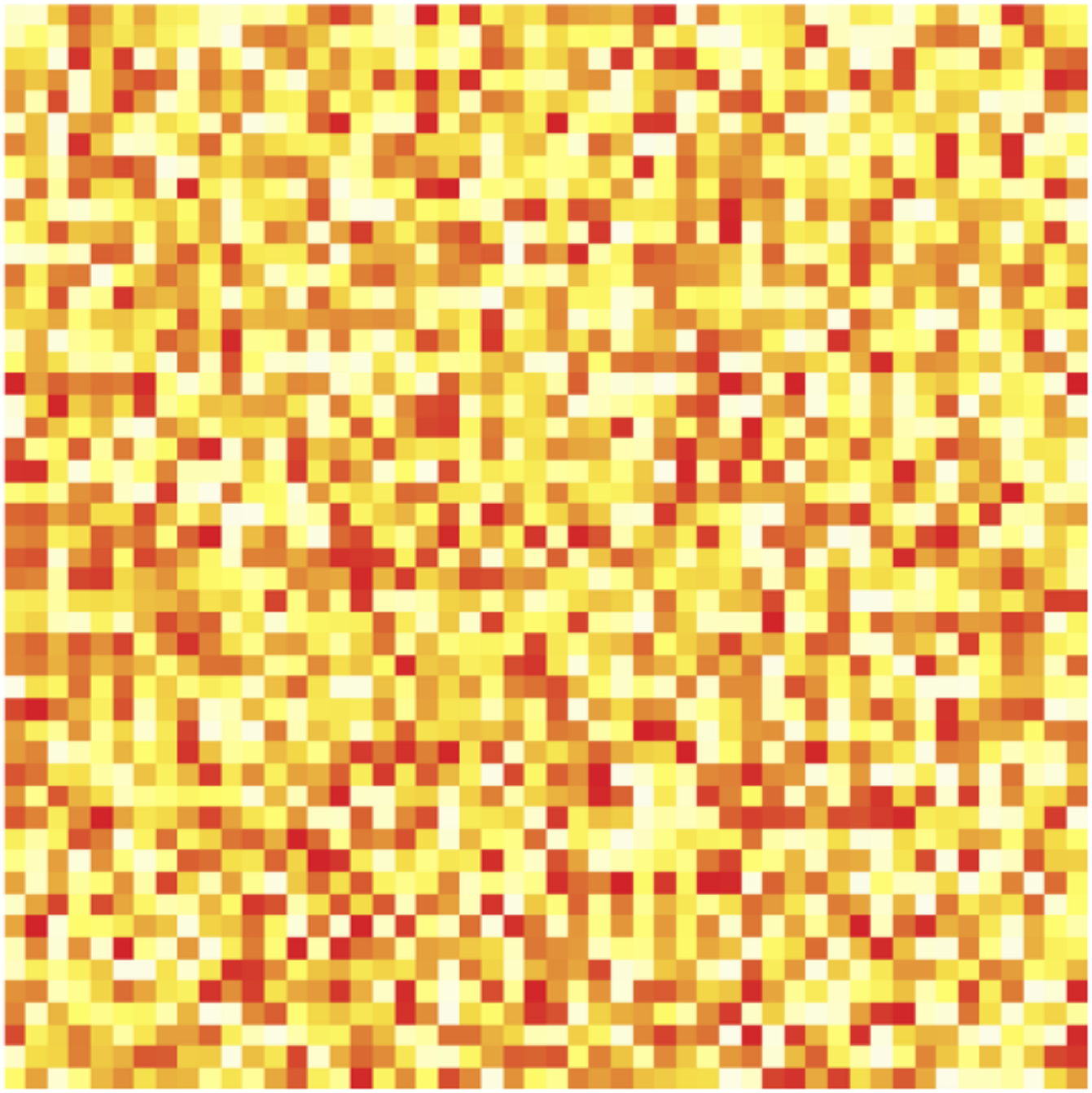
- One among many shapes, but simulations available **[Wandelt-Elsner]** with results stored in spherical harmonics coefficients:

$$\ell_{max} = 1024 \qquad a_{\ell m} = a_{\ell m}^{G} + f_{NL} a_{\ell m}^{NG}$$

- We first carried out TDA for this shape as a warmup, more in our pipeline.

Sublevel Filtration

(Hotter points are deeper red)

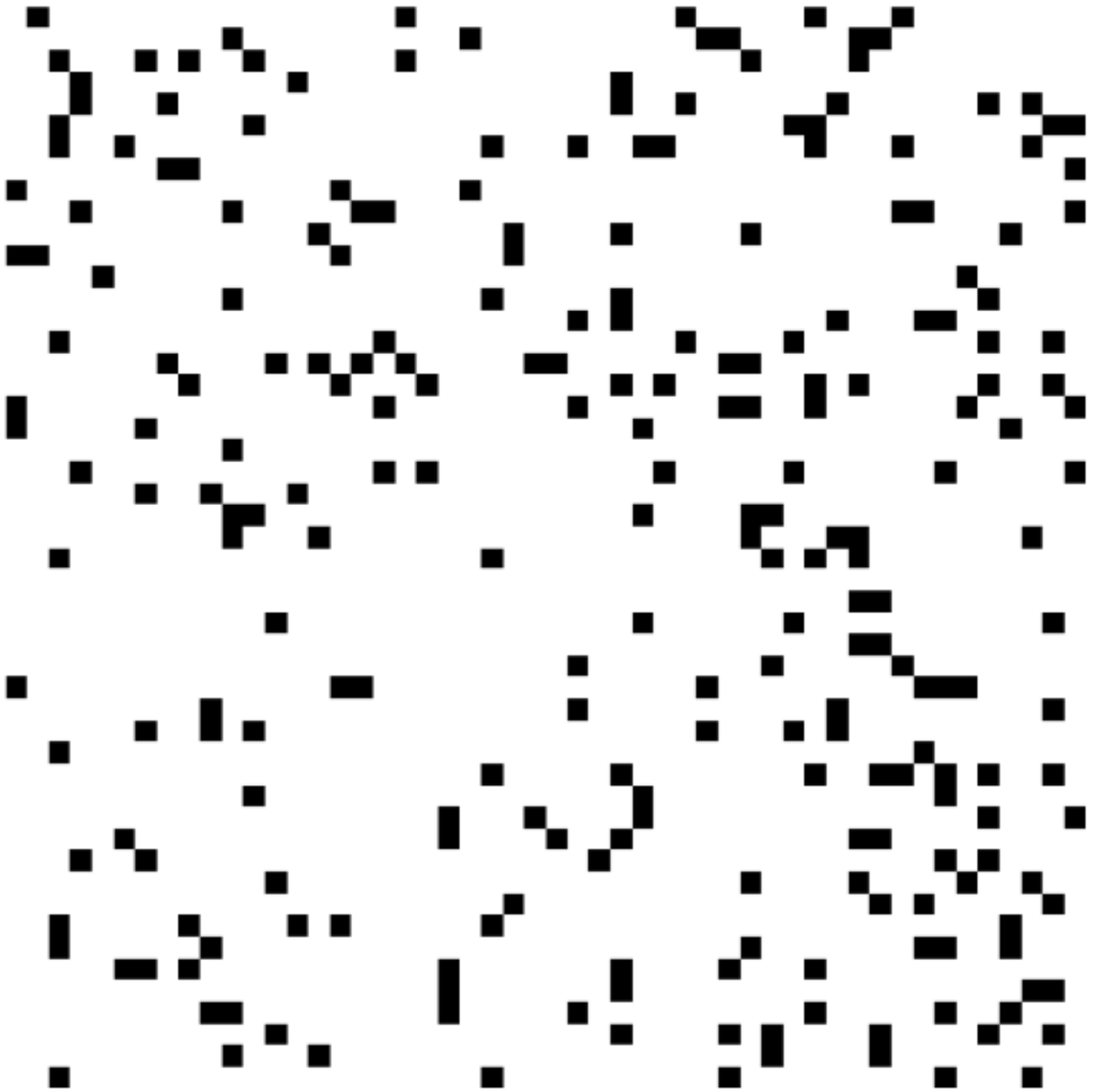


Sublevel Filtration

$\nu = -1$

Many distinct components, no loops

(Sublevel set in black)



Sublevel Filtration

$\nu = 0$

Many loops, fewer distinct components

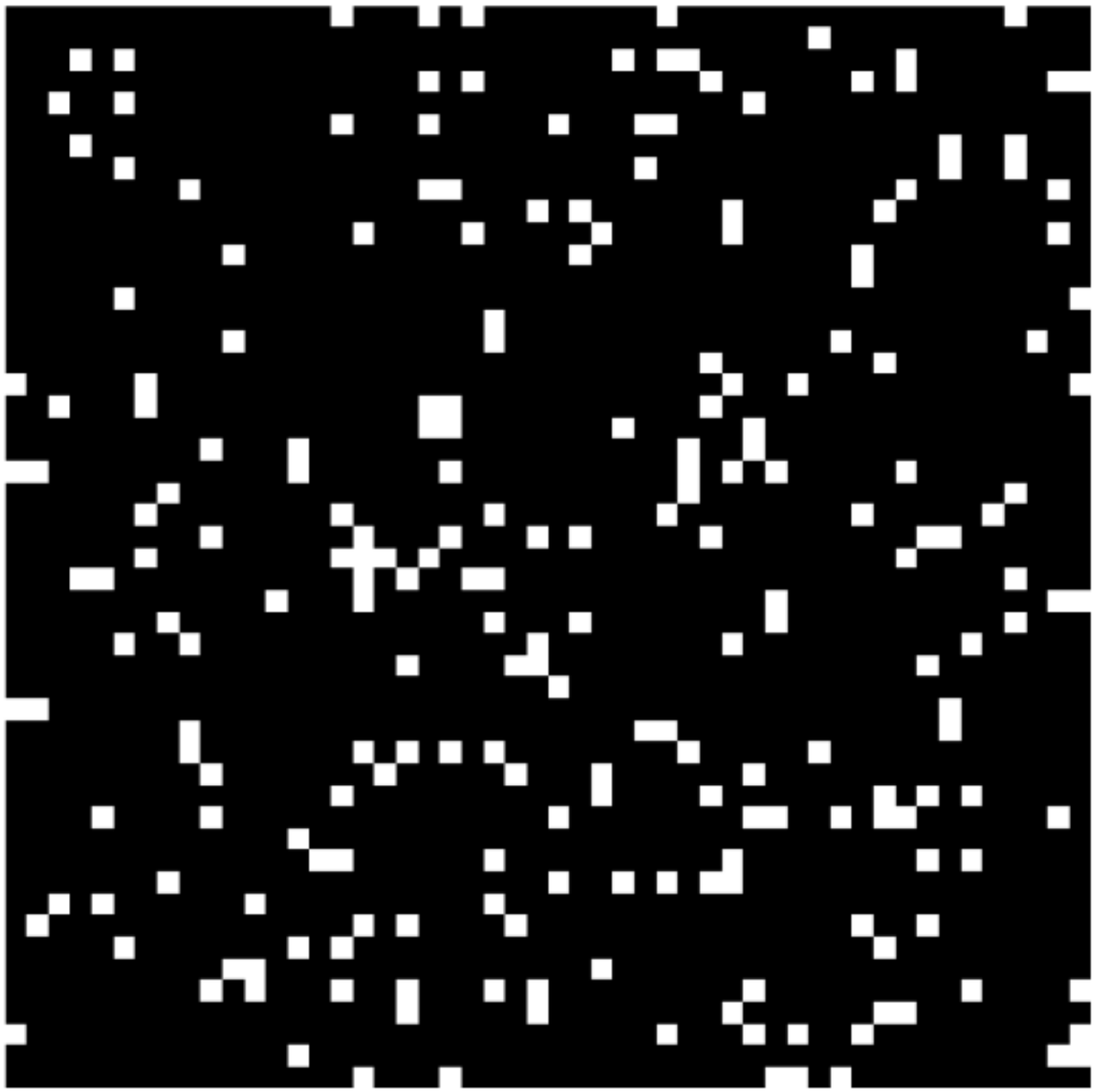
(Sublevel set in black)

Sublevel Filtration

$\nu = 1$

One connected component, many loops have been filled in

(Sublevel set in black)



Topological Measures

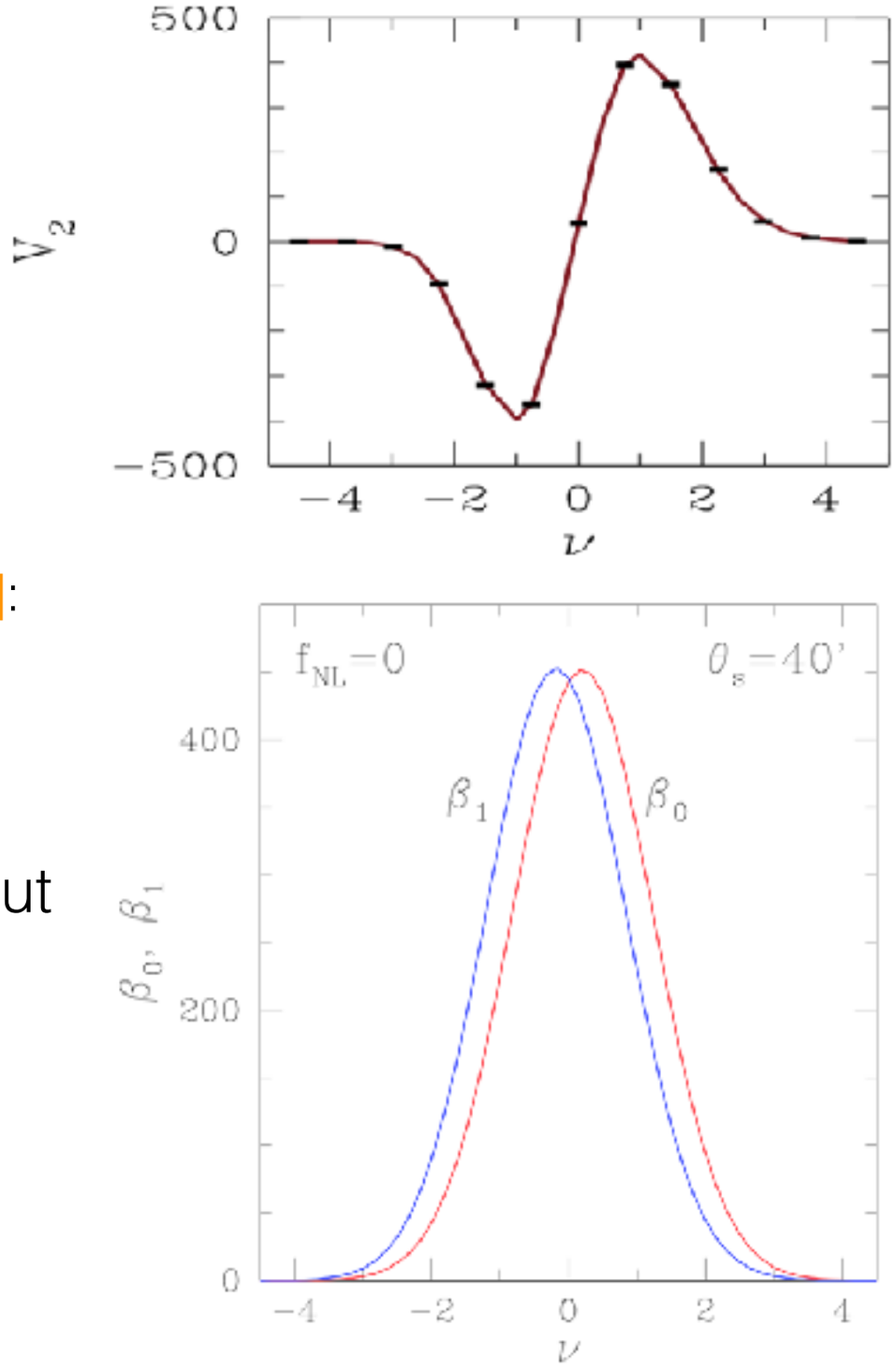
- **Genus (Euler characteristic)**:

$$\chi(\nu) = \beta_0(\nu) - \beta_1(\nu)$$

- Analytic formula for **Gaussian fields**:

$$\chi_G(\nu) \propto \nu e^{-\nu^2/2}$$

- **[Chingabam, Park, Yogendran, van de Weygaert]**: consider Betti numbers instead
  - Demonstrated explicitly that Betti numbers contain more information about NG using simulations
- Our project: make topological analysis stronger using **persistence diagrams**



Sensitivity to Non-Gaussianity

- **Likelihood function**: probability of data given non-Gaussianity

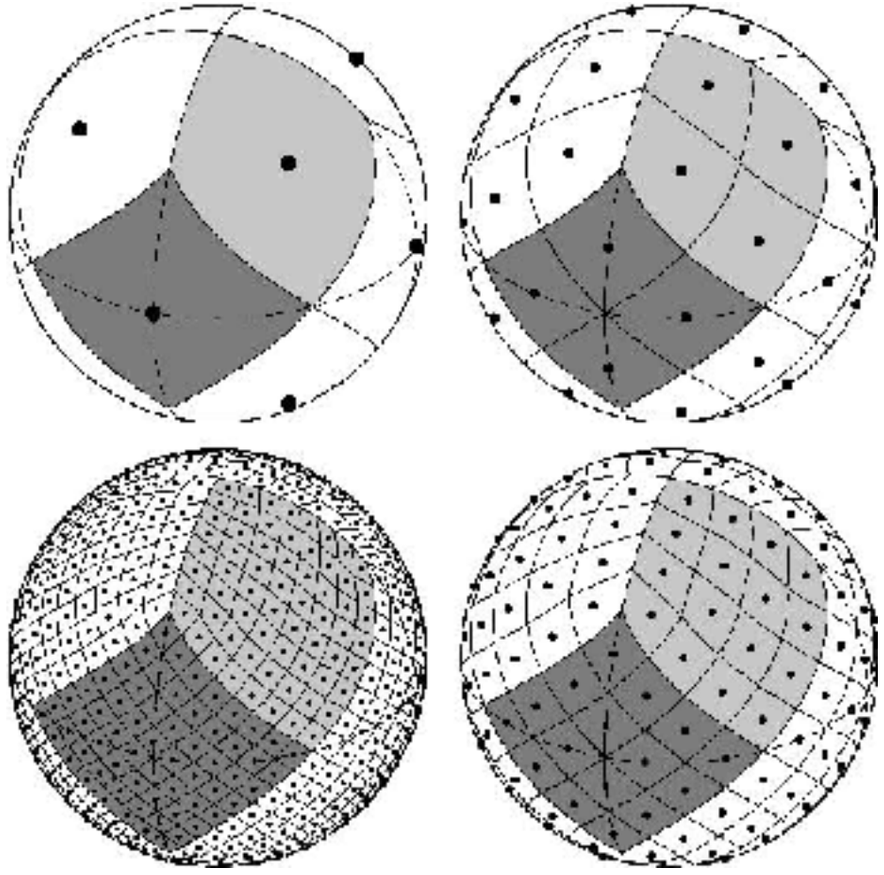
$$P(\mathbf{d}|f_{NL}) \propto \exp\left(-\frac{1}{2}(\mathbf{d}-\boldsymbol{\mu}(f_{NL}))^T\mathbf{C}^{-1}(\mathbf{d}-\boldsymbol{\mu}(f_{NL}))\right)$$

$$C_{ij} = \sum_{k}^{N_{\rm sim}} (d_i^{0,k}-\mu_i)(d_j^{0,k}-\mu_j)$$

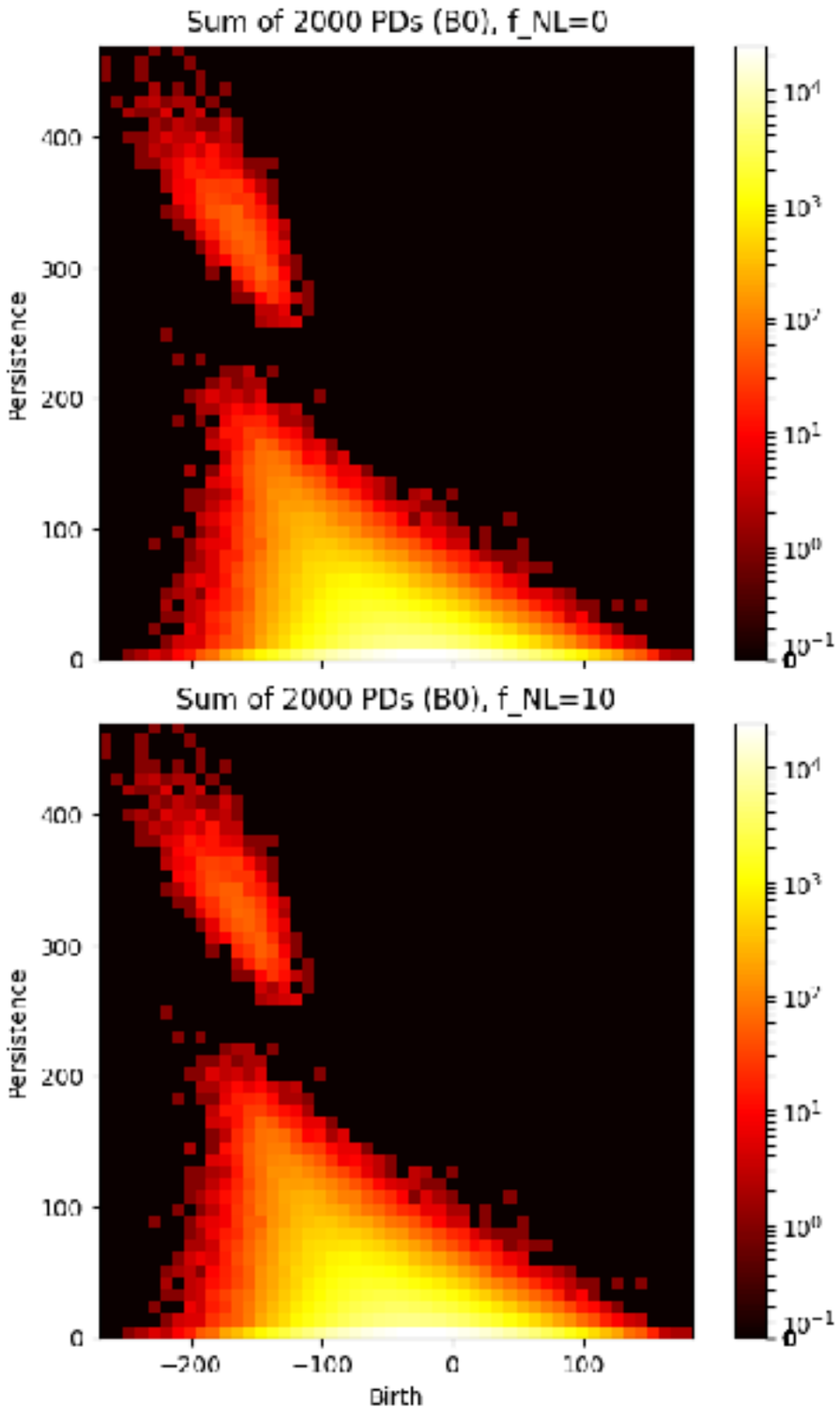
- Sharpness of peak determines σ
- Strategy: compute $A \equiv \frac{1}{2}(\mathbf{d}-\boldsymbol{\mu}(f_{NL}))^T\mathbf{C}^{-1}(\mathbf{d}-\boldsymbol{\mu}(f_{NL}))$
  - d from $f_{NL}=10$, μ, C from $f_{NL}=0$
  - Larger A corresponds to more sensitive statistic

Numerical Pipeline

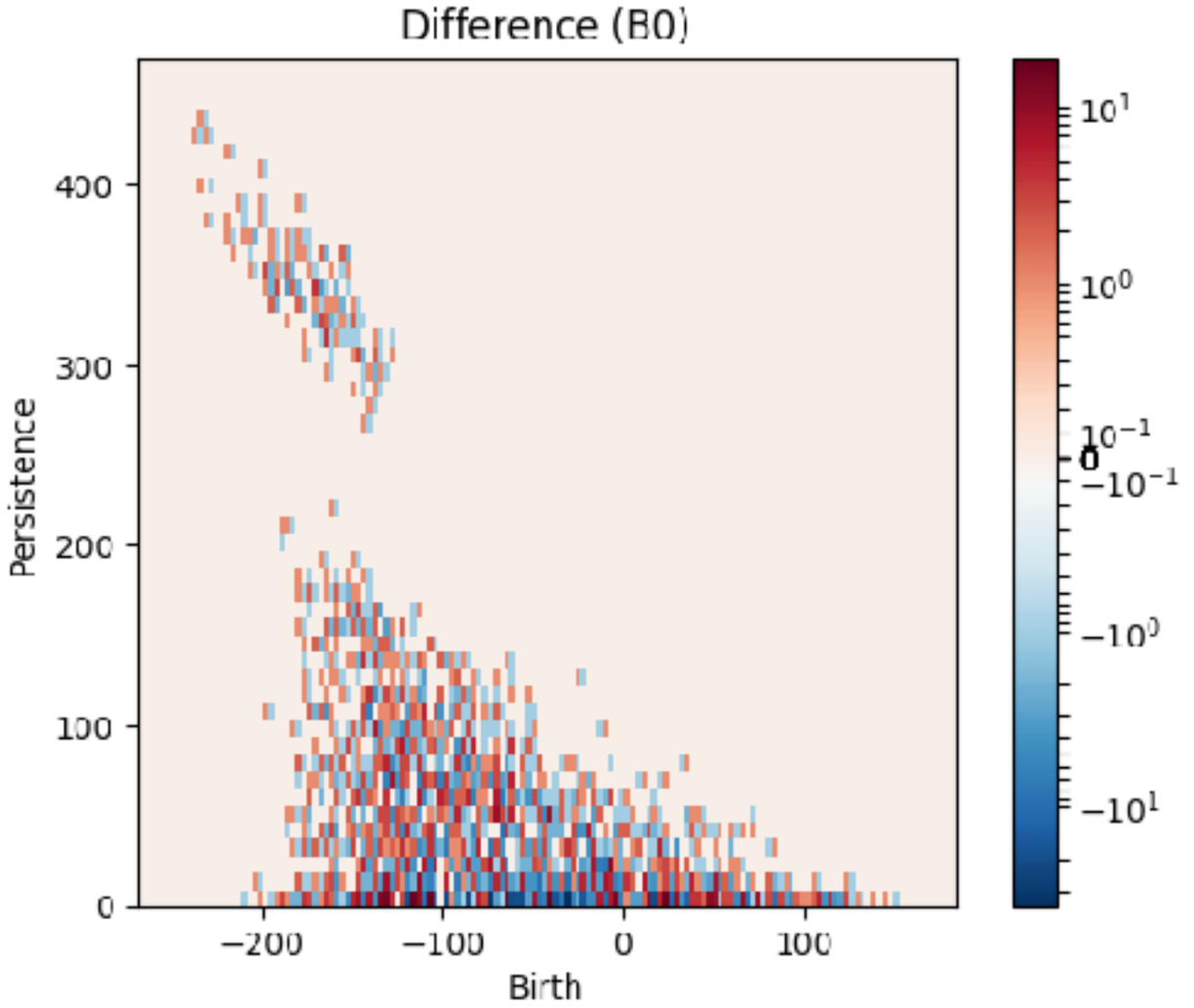
- Generate maps and select rectangular regions using HEALPIX
  - 2,000 maps for each level of NG, 1024x1024 grids
- Compute sublevel filtrations and persistent homology using R package TDA and Persistent Homology Algorithms Toolbox (PHAT)
- Bin the histograms and perform likelihood analysis in Python



Results

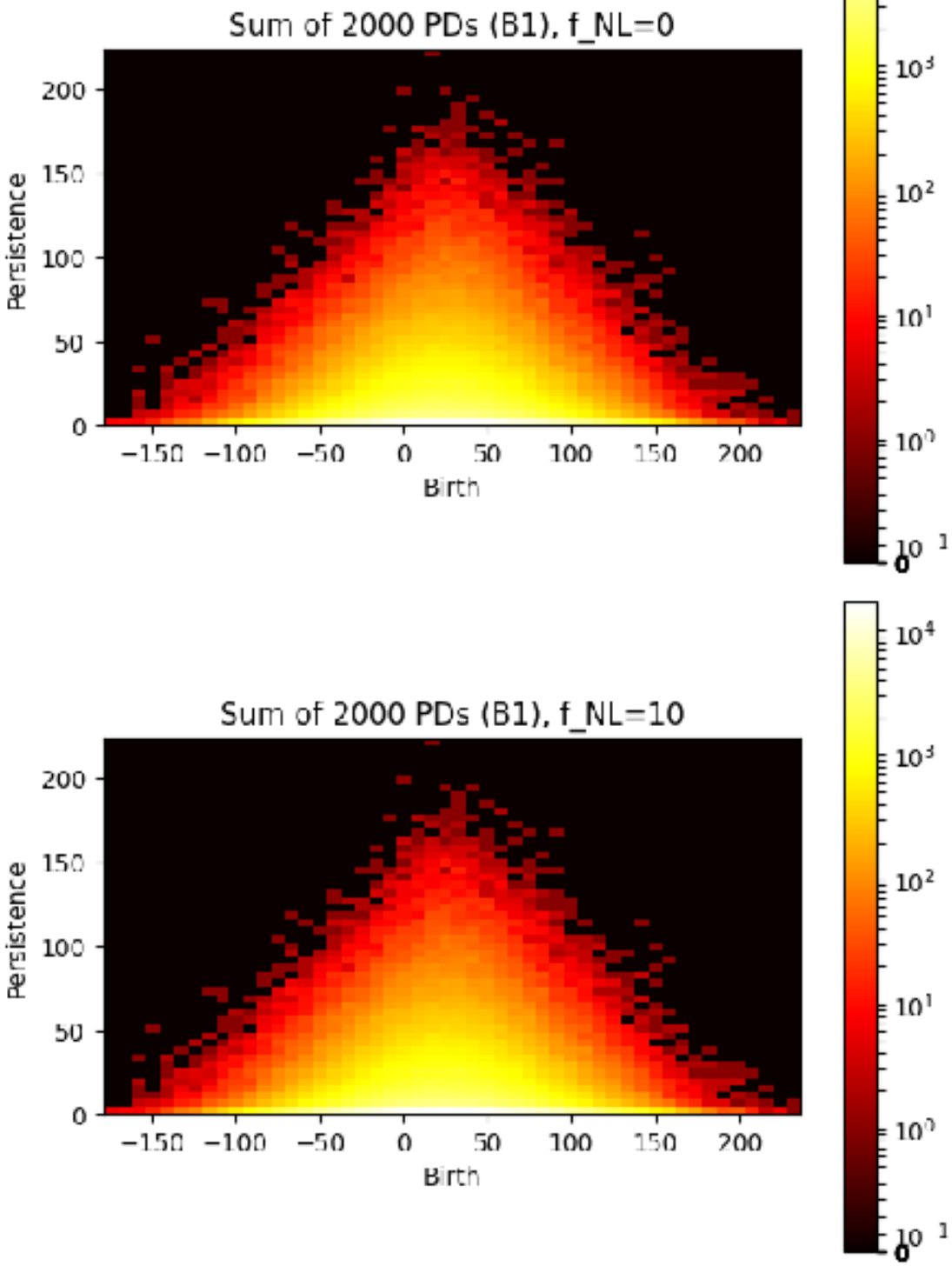


Input for likelihood analysis

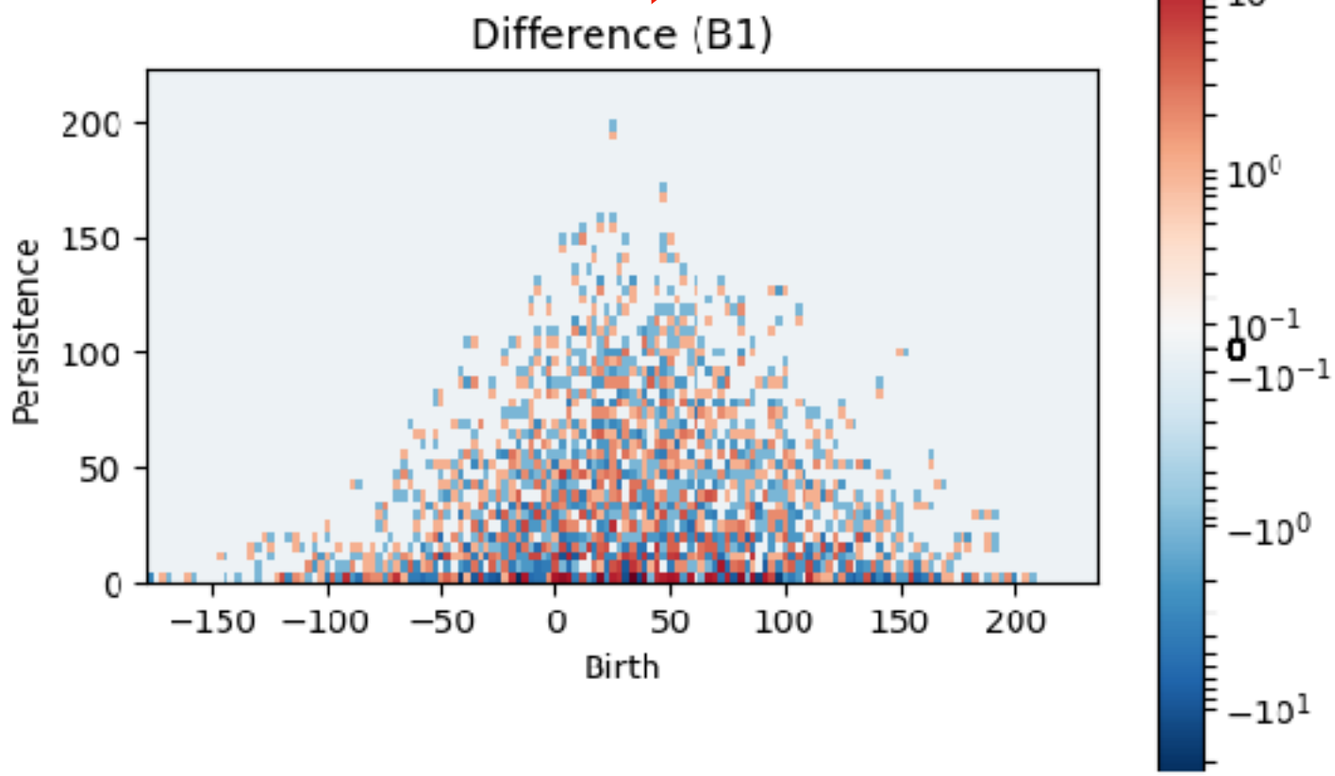


~1,500 bins

Results

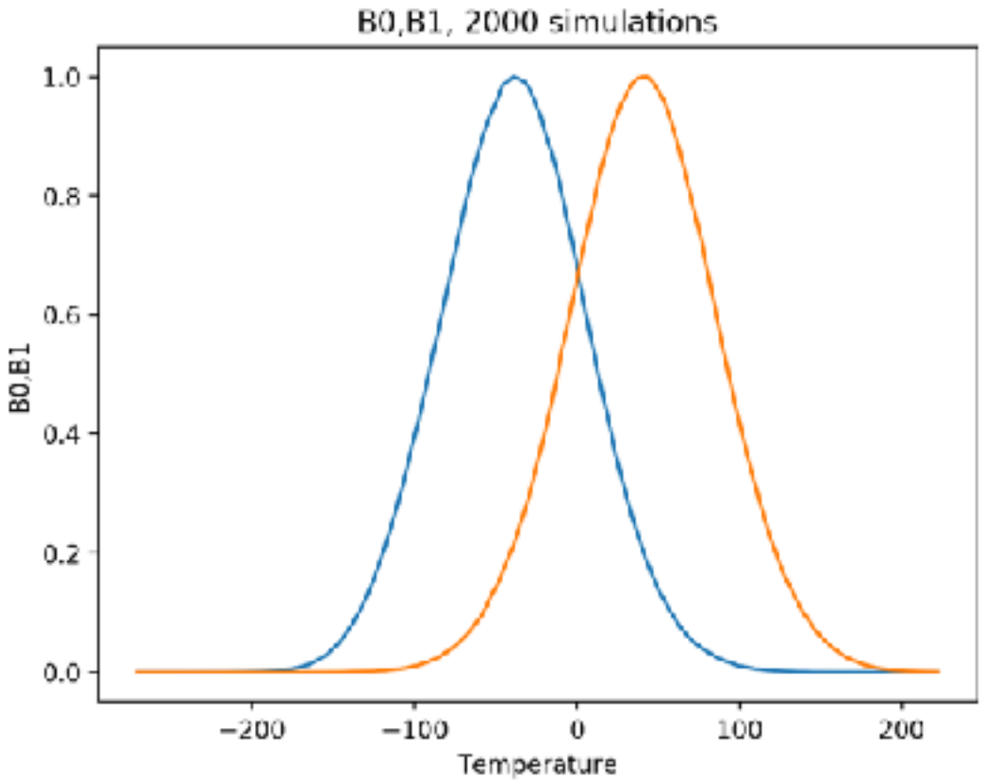


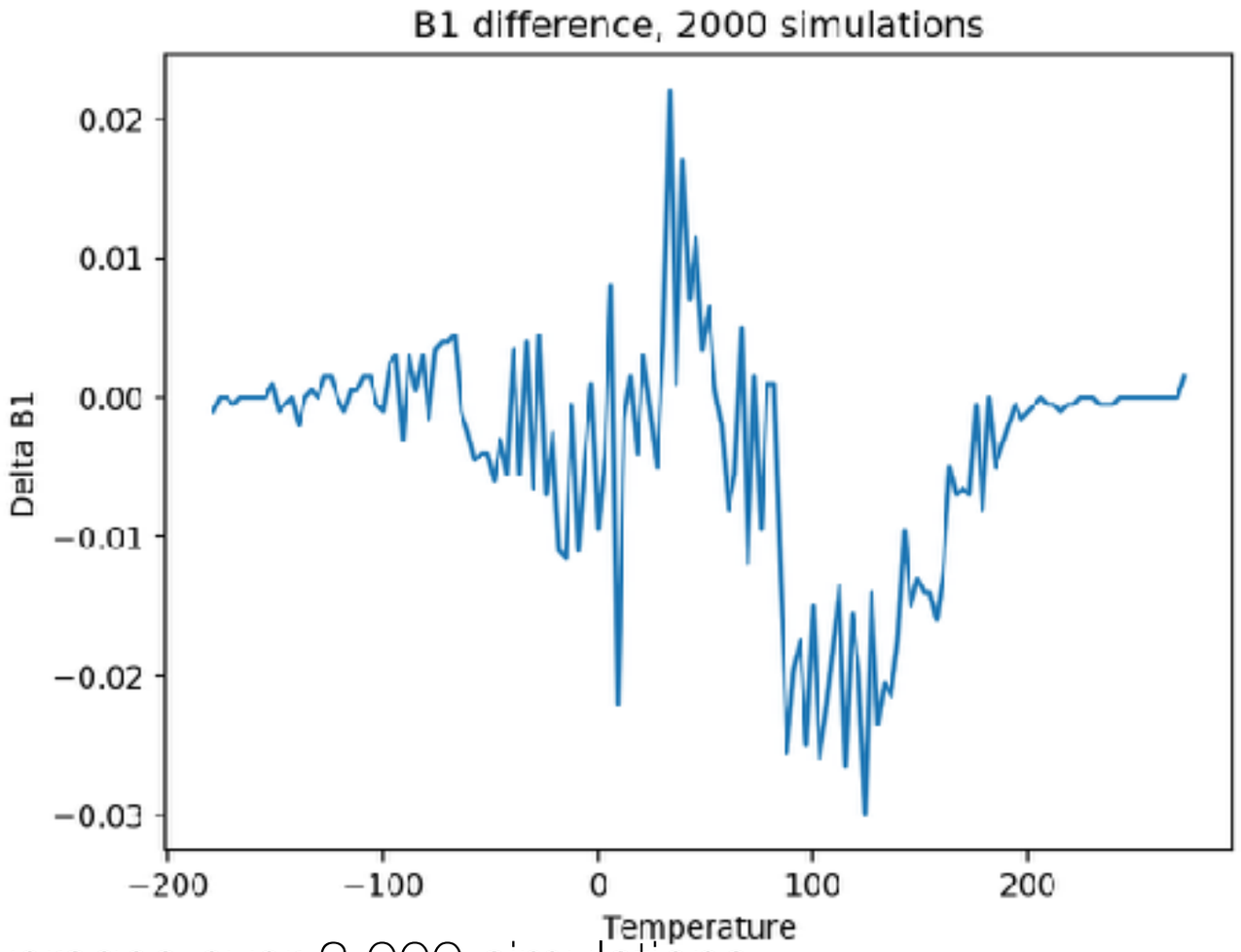
Input for likelihood analysis



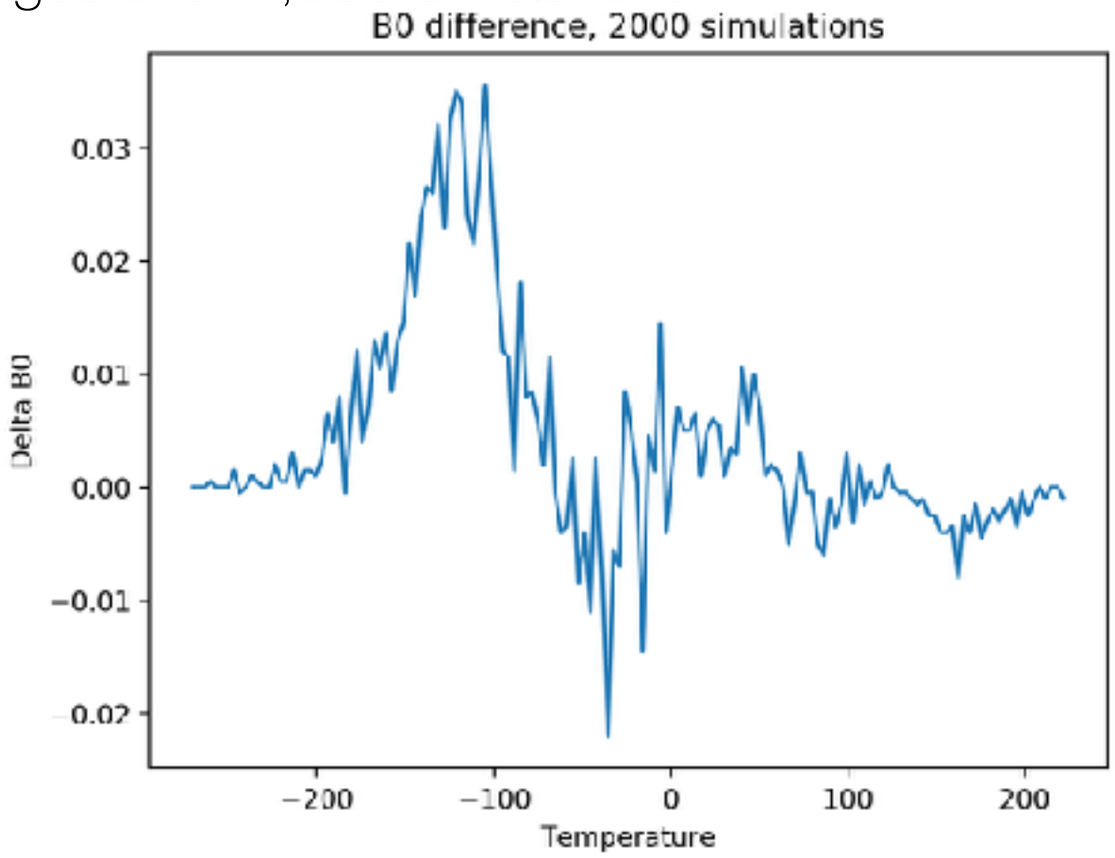
~1,500 bins

Betti number curves





Averages over 2,000 simulations



Normalized so $\beta_{\max} = 1$

Sensitivity to Non-Gaussianity

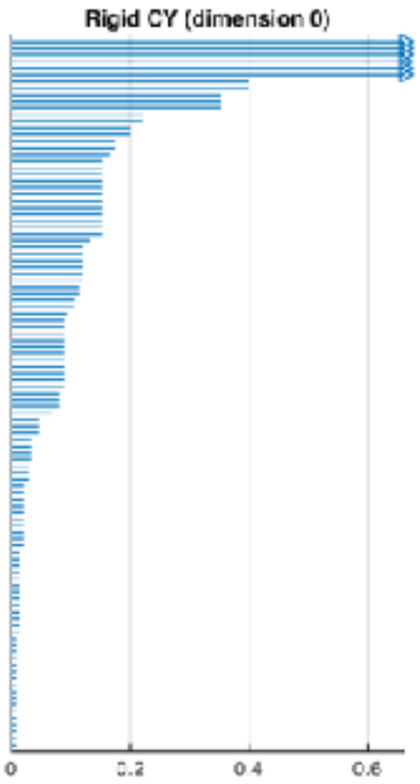
$$P(\mathbf{d}|f_{NL}) \propto \exp\left(-\frac{1}{2}(\mathbf{d}-\boldsymbol{\mu}(f_{NL}))^T\mathbf{C}^{-1}(\mathbf{d}-\boldsymbol{\mu}(f_{NL}))\right)$$

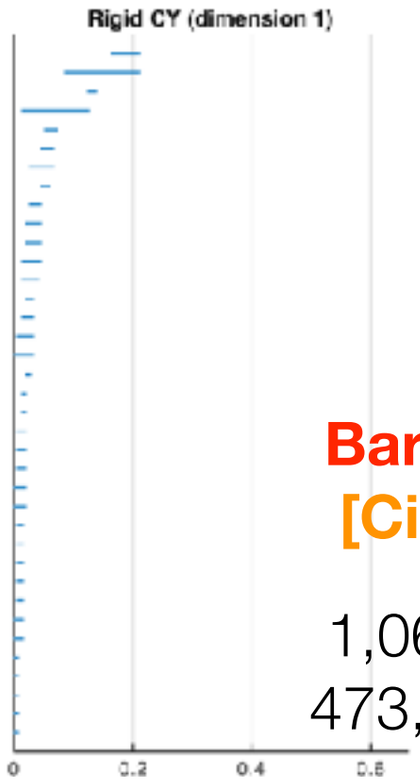
$$A \equiv \frac{1}{2}(\mathbf{d}-\boldsymbol{\mu}(f_{NL}))^T\mathbf{C}^{-1}(\mathbf{d}-\boldsymbol{\mu}(f_{NL}))$$

$A_{\beta_0} = 0.00461$	$A_{PD_0} = 0.02416$
$A_{\beta_1} = 0.00172$	$A_{PD_1} = 0.02225$

- More than a factor of five increase! $A \sim \sigma^{-1}$
- Correspondingly better constraints for full data analysis
  - Persistence diagrams strengthen topological analysis significantly

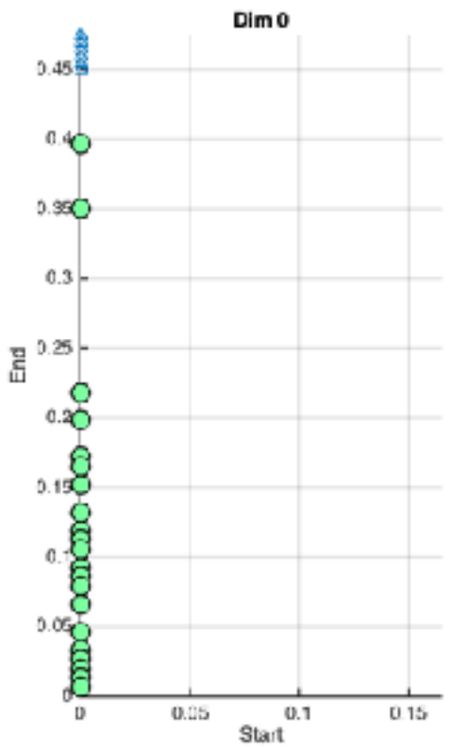
TDA for String Vacua

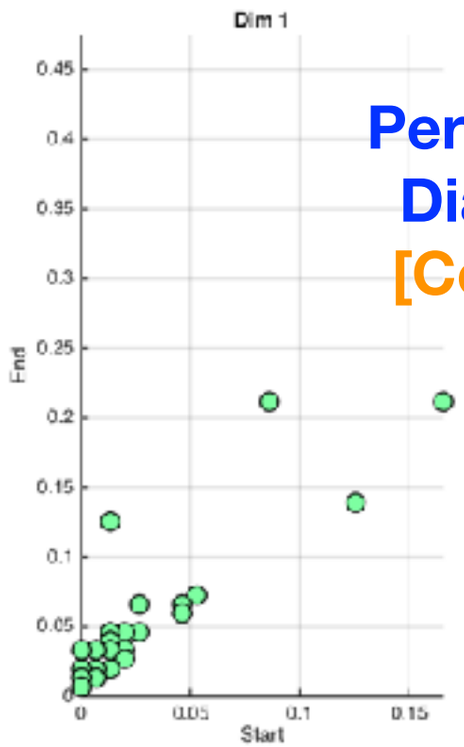




Barcodes
[Cirafici]

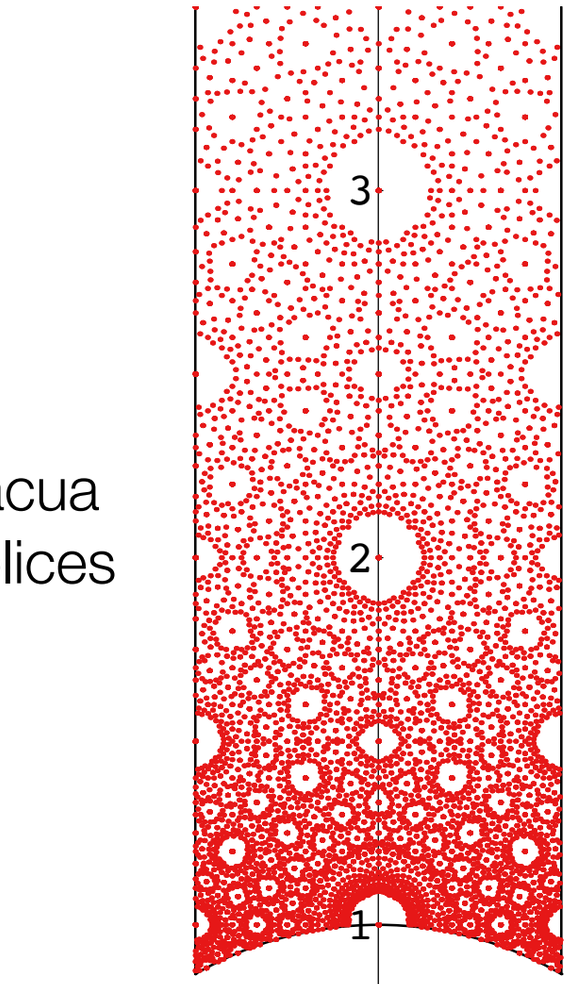
1,064,598 vacua
473,801 simplices

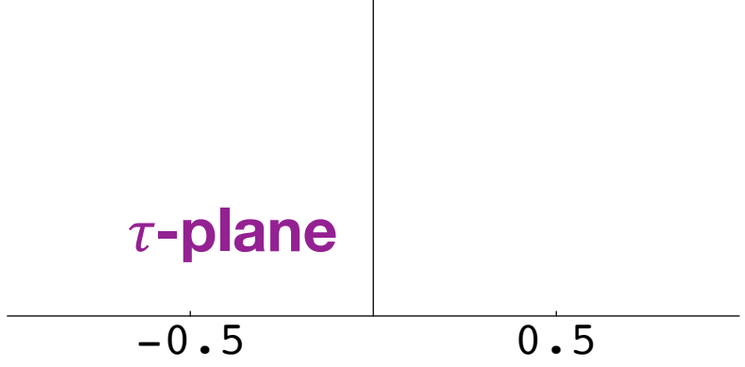




Persistence Diagrams
[Cole, GS]

"Topological Complexity"

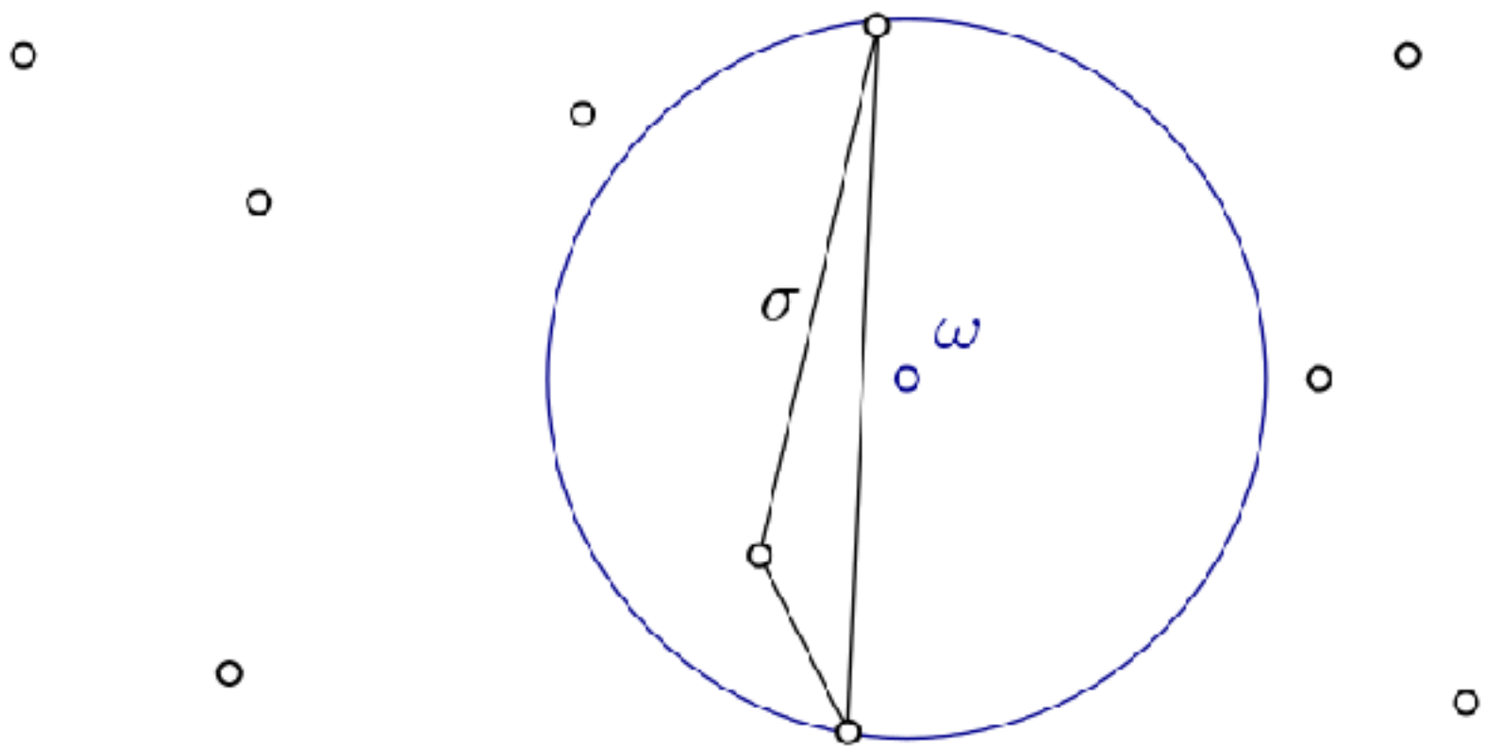


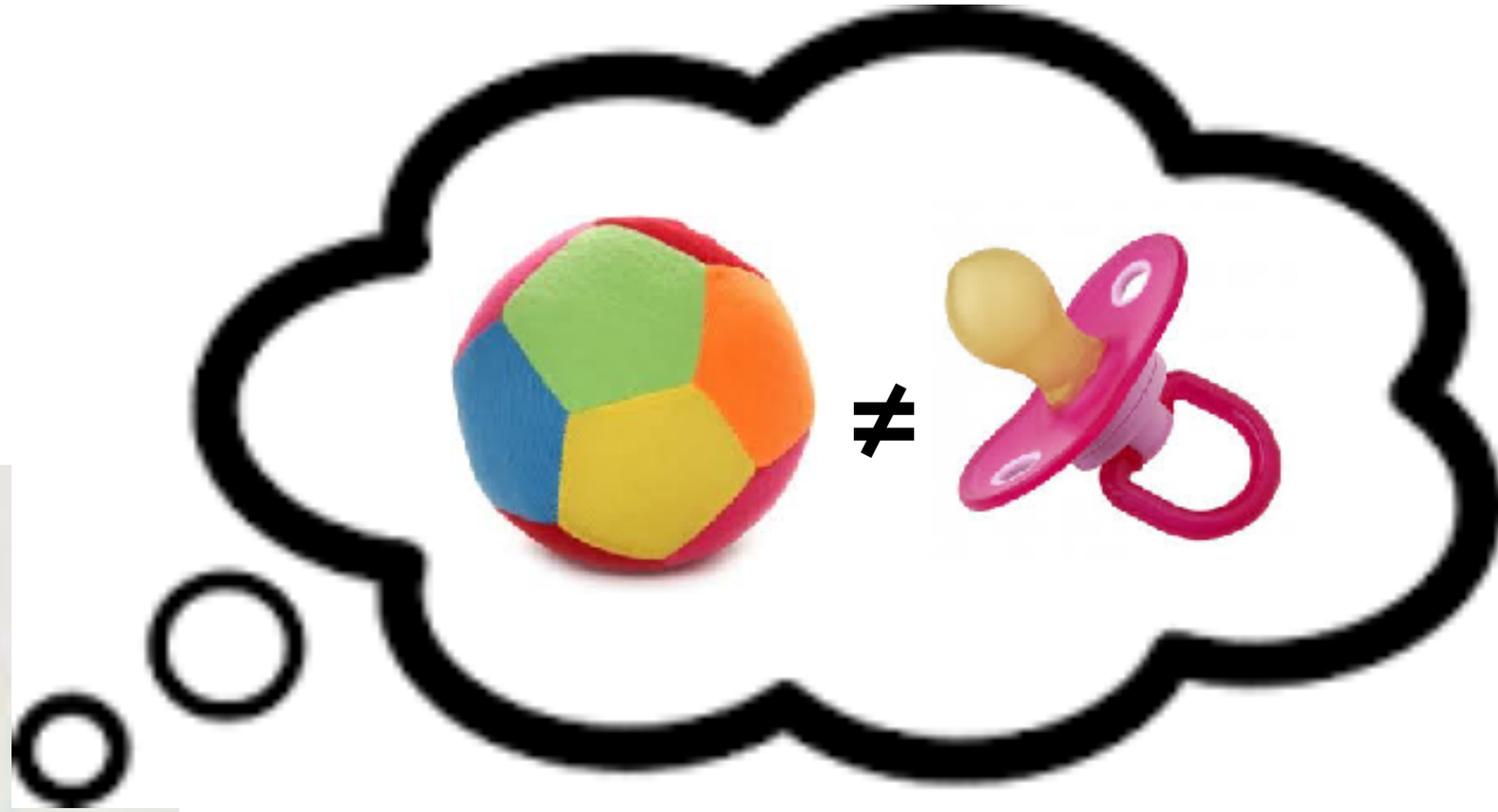


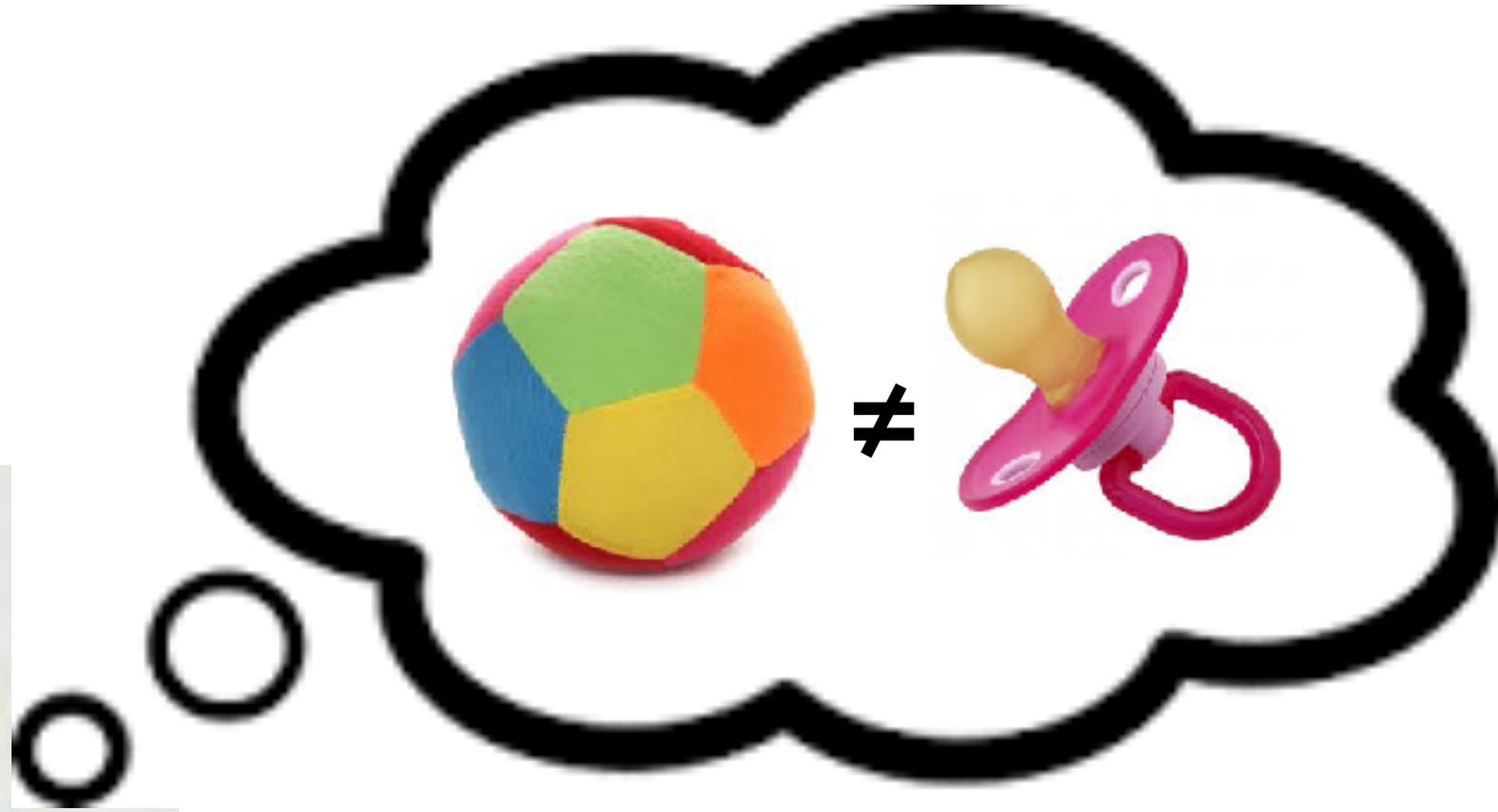
e.g., for rigid CY voids correspond to degeneracy of vacua — relationship between **topology of distribution** and **physics**

Sampling in TDA

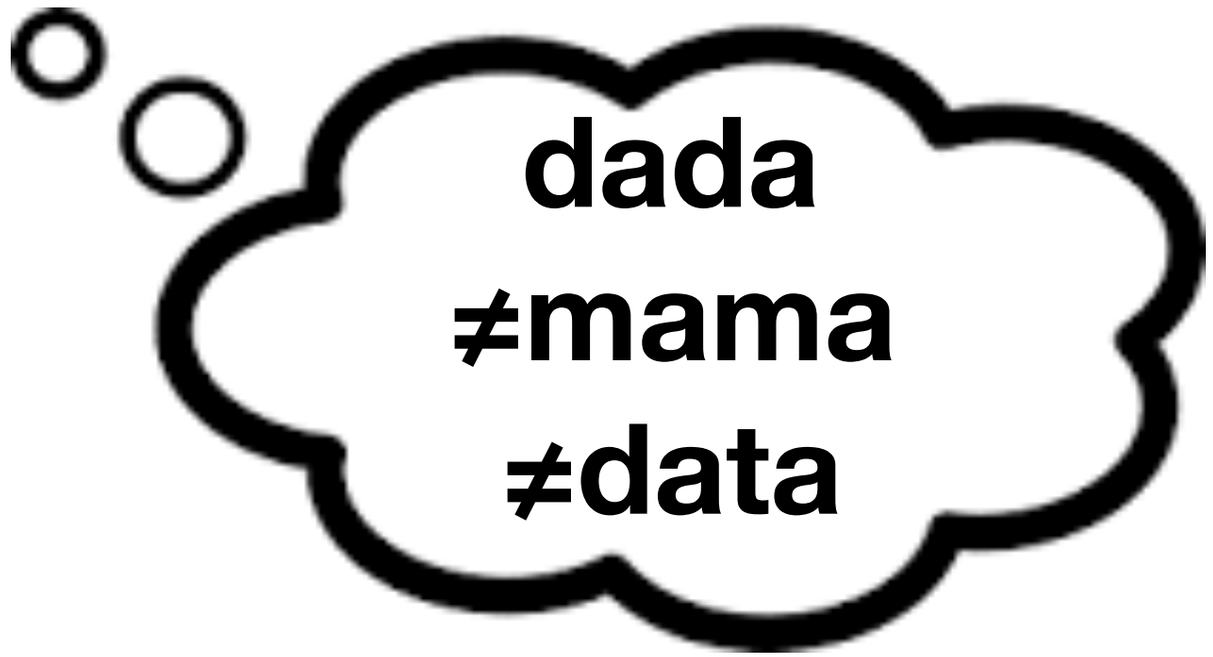
- We can't realistically include all 10^{500} vacua as vertices
- Can sample the topology via the **witness complex**:
  - From the entire point cloud Z, choose a ***landmark set L*** as the complex's vertices. Often chosen randomly or via sequential maxmin algorithm
  - Let $m_k(z)$ be the distance 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 $\max\{d(l_0,z), d(l_1,z), \dots, d(l_k,z)\} \leq \nu + m_k(z)$







≠


dada
≠mama
≠data

Conclusions

Conclusions

- Applications of TDA to **cosmological datasets** and **string vacua**.
- Persistent diagrams strengthen constraints on local **non-Gaussianities**, and potentially other shapes & other observables.
- Techniques we developed can be applied to analyze the structure of string vacua. We performed initial study of “featureless vacua”.
- Next step is to examine the **topology** of string vacua point clouds with desired features, supplementing earlier work on ***statistics***:
  - Enhanced symmetries **[DeWolfe, Giryavets, Kachru, Taylor], …**
  - Particle physics features **[Marchesano, GS, Wang];[Dienes];[Gmeiner, Blumenhagen, Honecker, Lust, Weigand], [Douglas, Taylor], [AbdusSalam, Conlon, Quevedo Suruliz], …**
- String Landscape vs the Swampland? **[see Vafa’s talk]**

Conclusions

- Applications of TDA to **cosmological datasets** and **string vacua**.
- Persistent diagrams strengthen constraints on local **non-Gaussianities**, and potentially other shapes & other observables.
- Techniques we developed can be applied to analyze the structure of string vacua. We performed initial study of “featureless vacua”.
- Next step is to examine the **topology** of string vacua point clouds with desired features, supplementing earlier work on ***statistics***:
  - Enhanced symmetries **[DeWolfe, Giryavets, Kachru, Taylor], …**
  - Particle physics features **[Marchesano, GS, Wang];[Dienes];[Gmeiner, Blumenhagen, Honecker, Lust, Weigand], [Douglas, Taylor], [AbdusSalam, Conlon, Quevedo Suruliz], …**
- String Landscape vs the Swampland? **[see Vafa’s talk]**