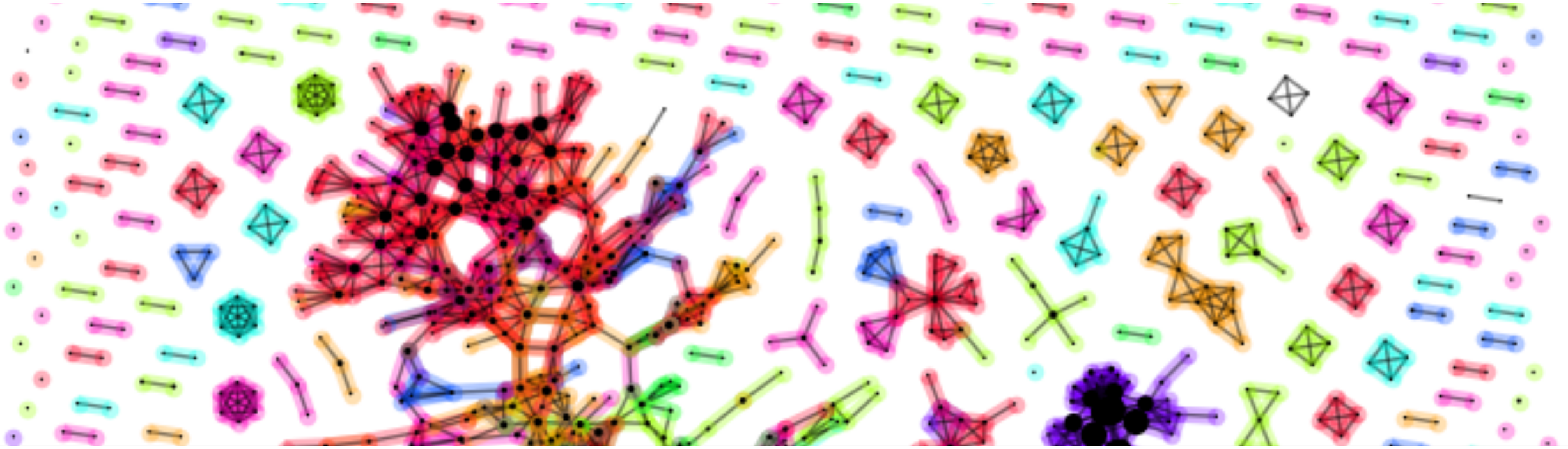
TDA Crash Course



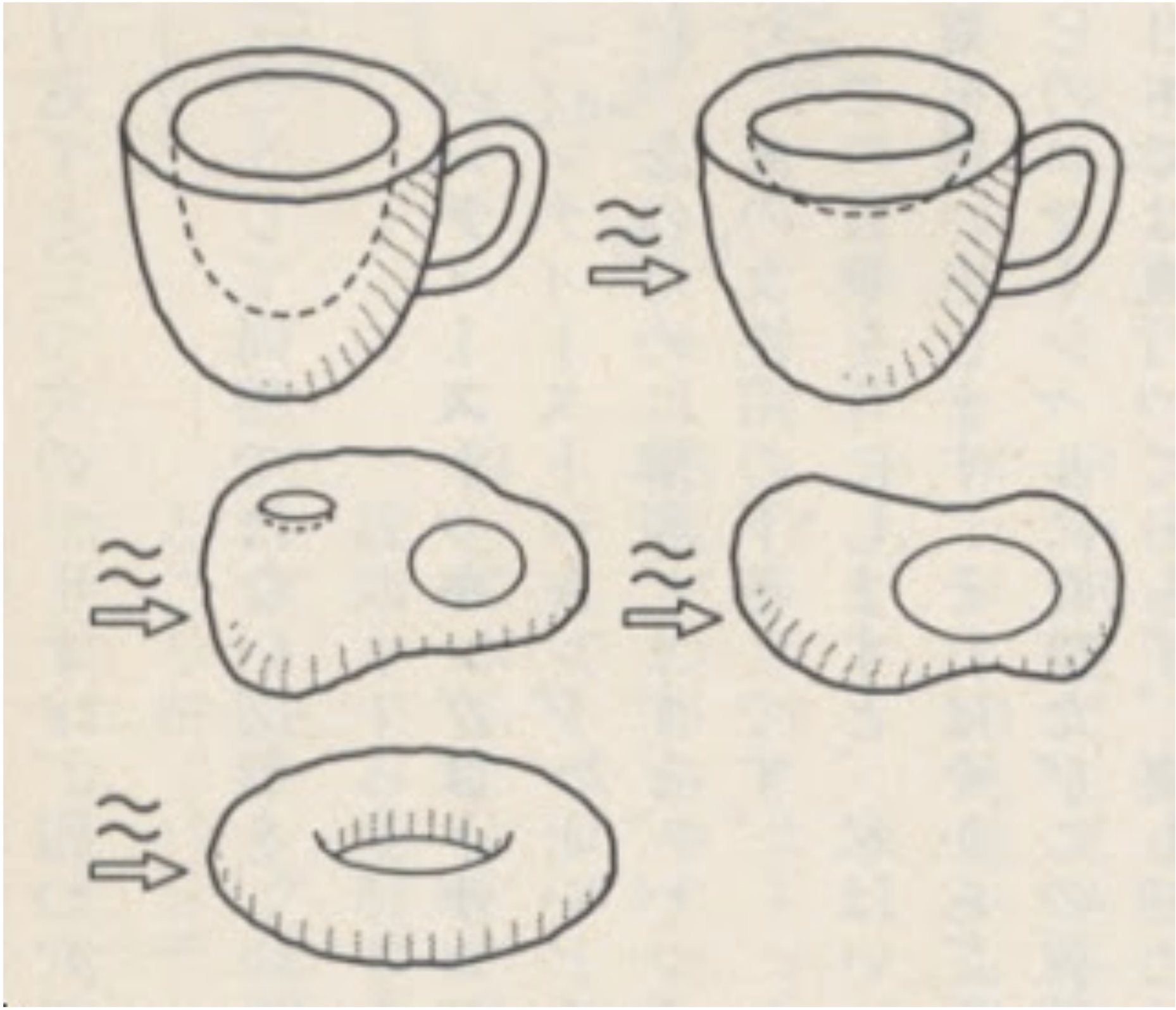
https://github.com/lordgrilo/AML-days-TDA-tutorial/

G. Petri

AML, Jan 2019

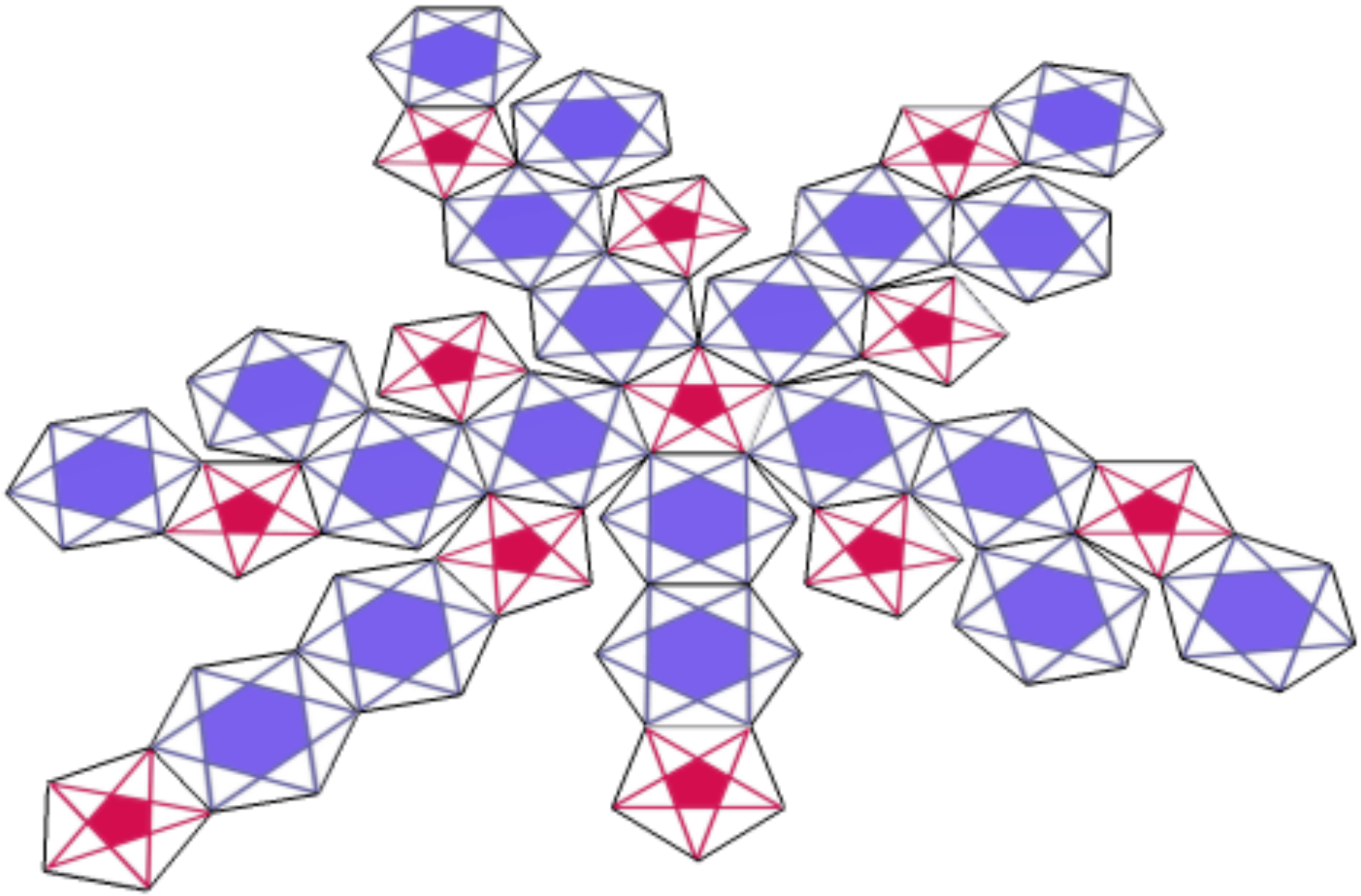
ISI Foundation
& ISI Global Science
Foundation

disclaimer for the network practitioner

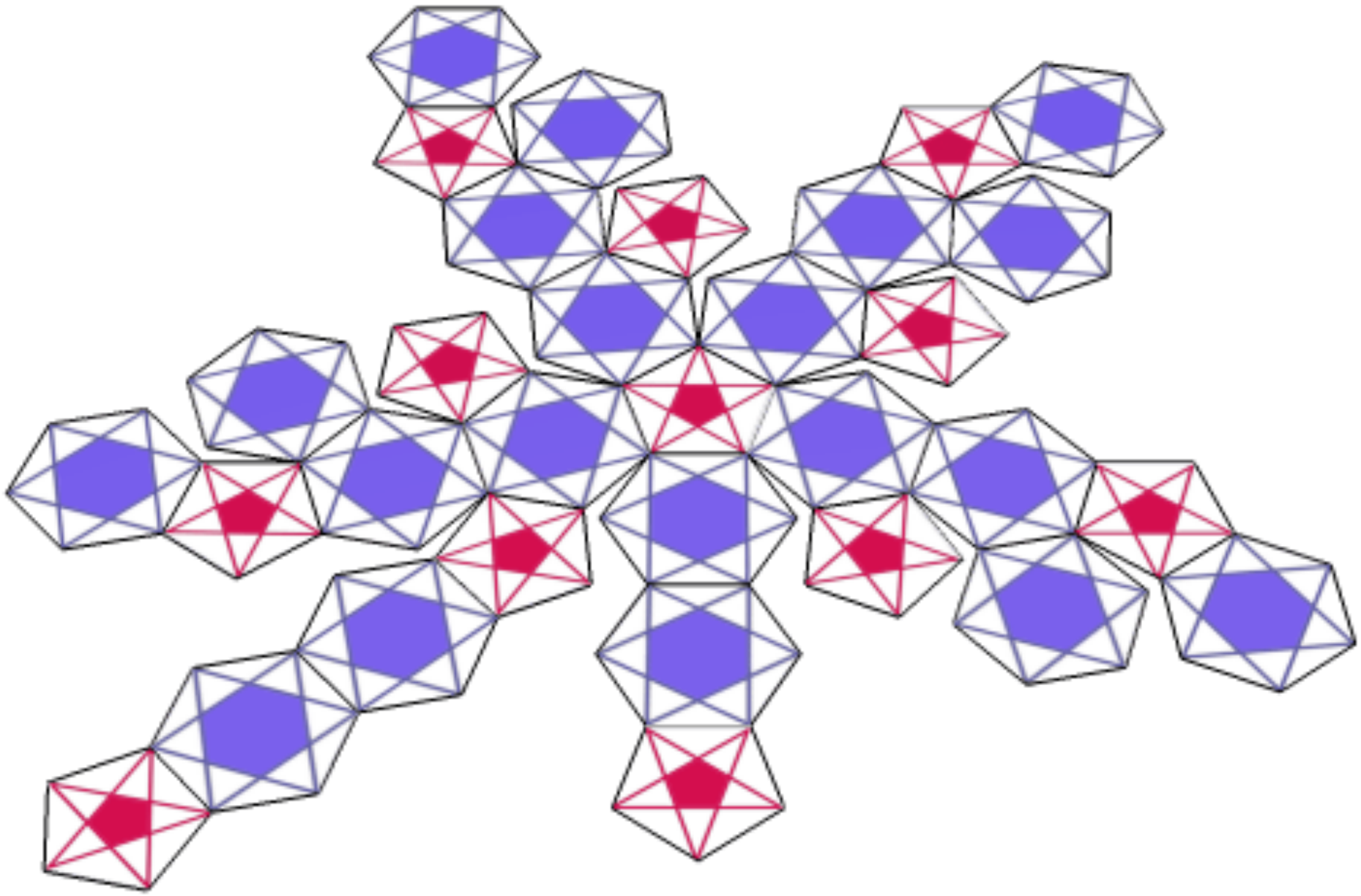


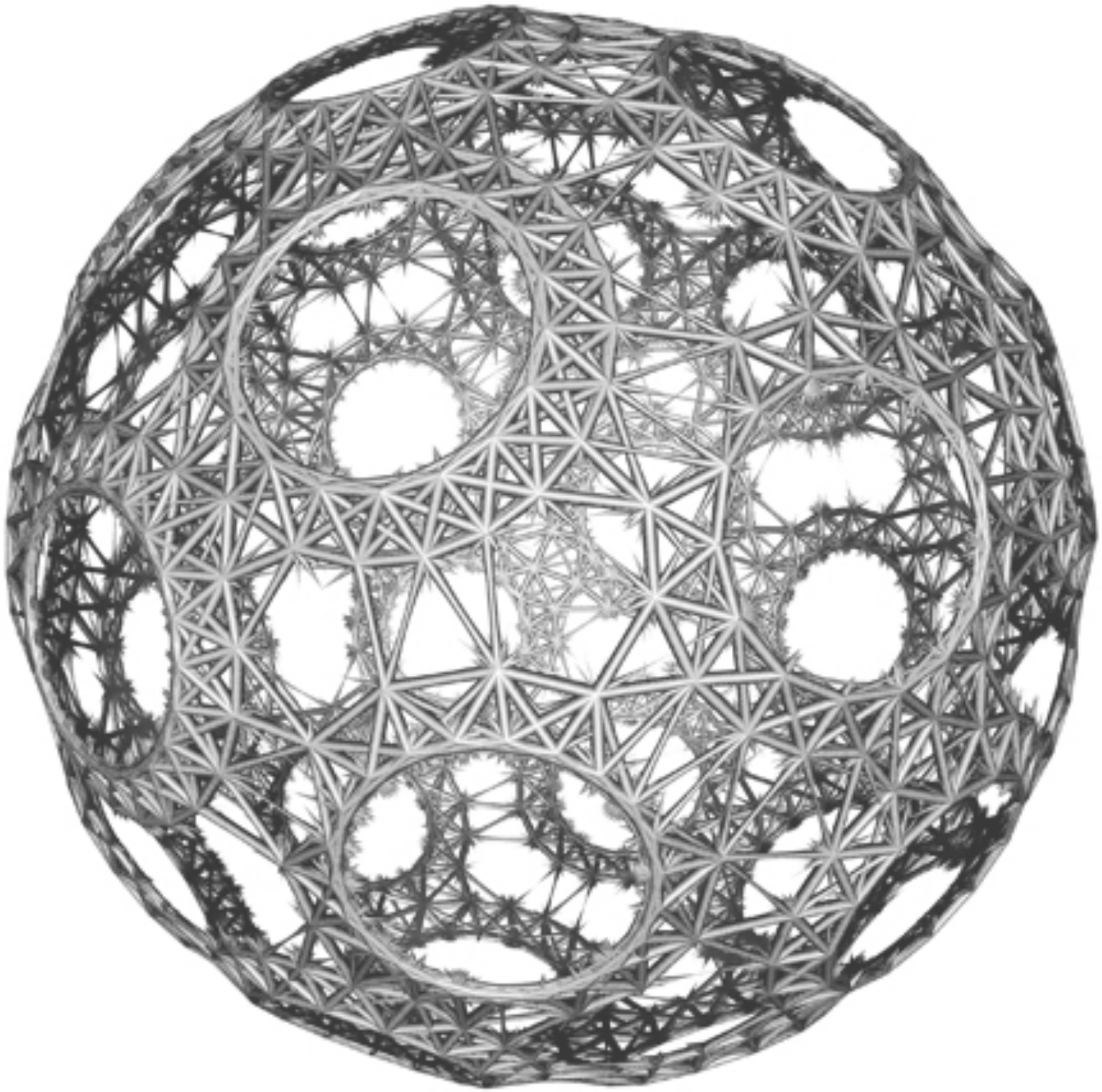
disclaimer for the network practitioner

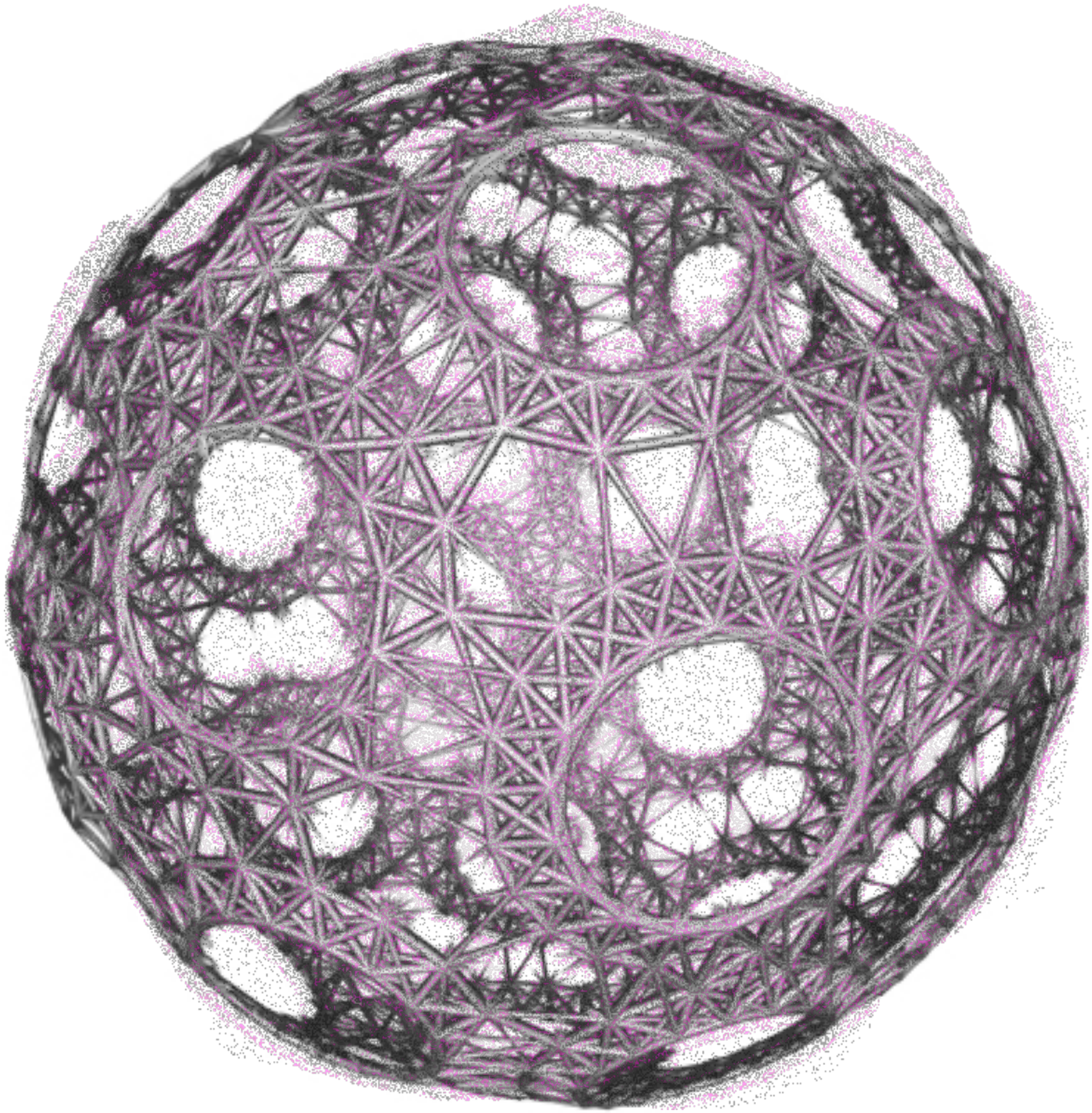
Why topology?



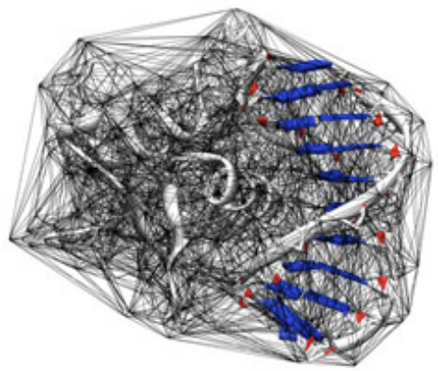
Why topology?

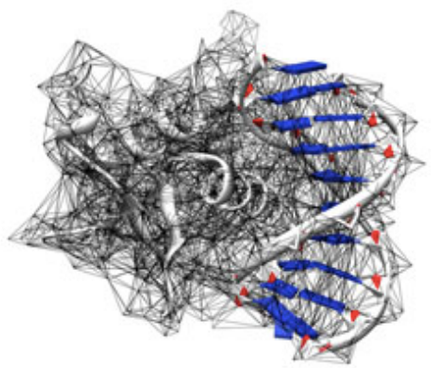


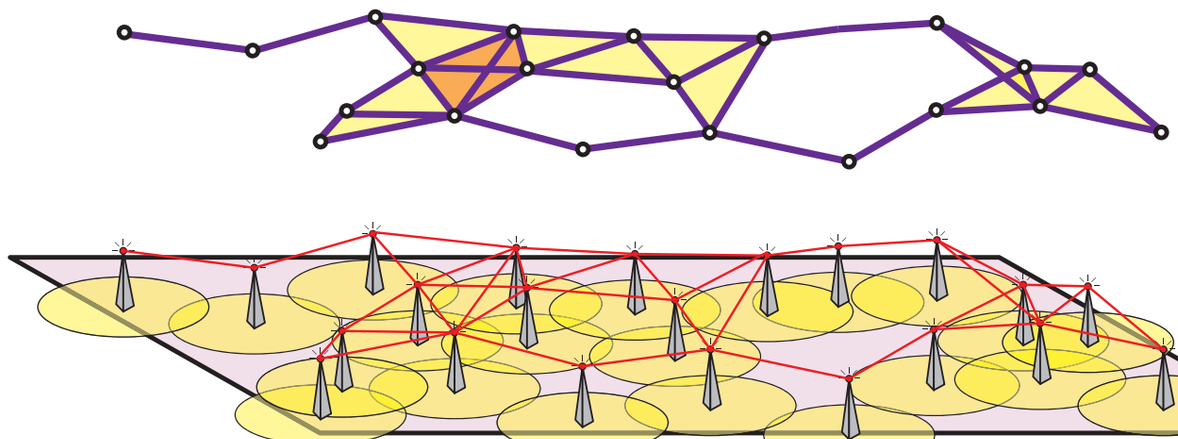


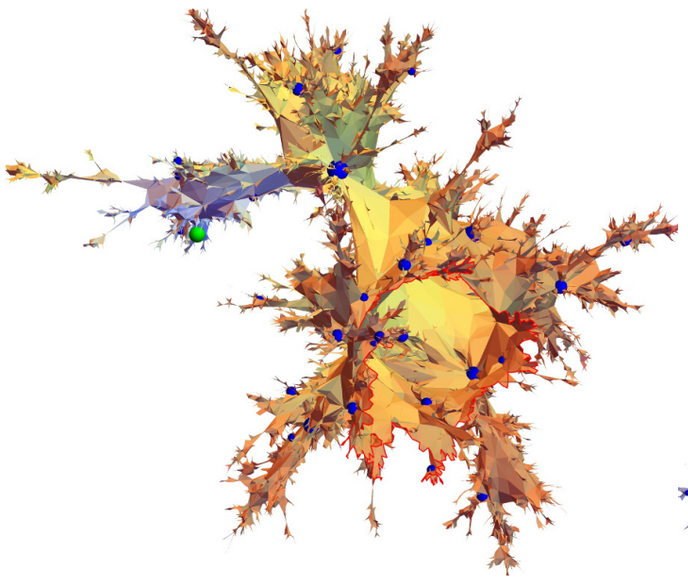


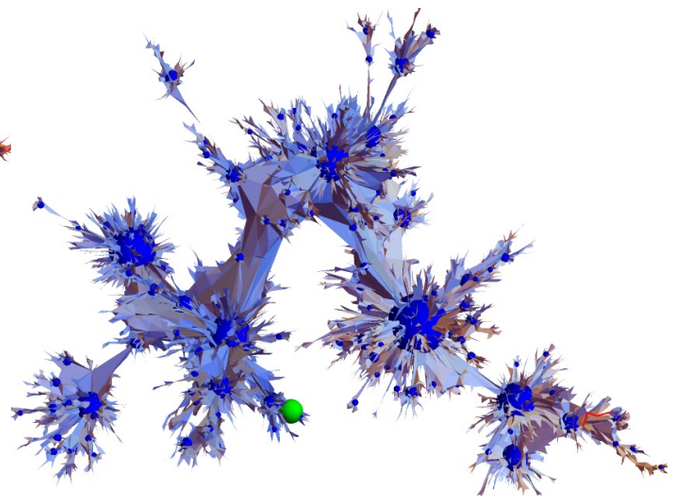
Examples of simplicial complexes

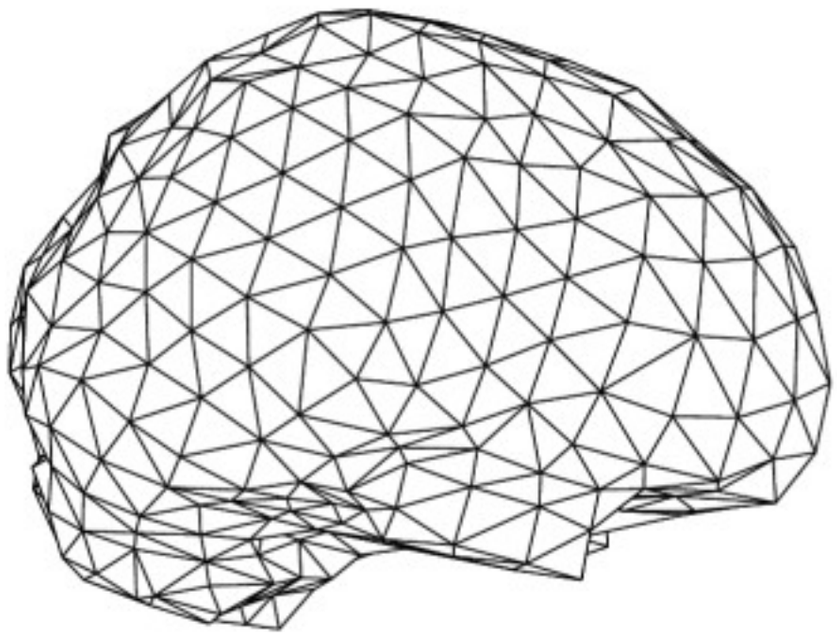






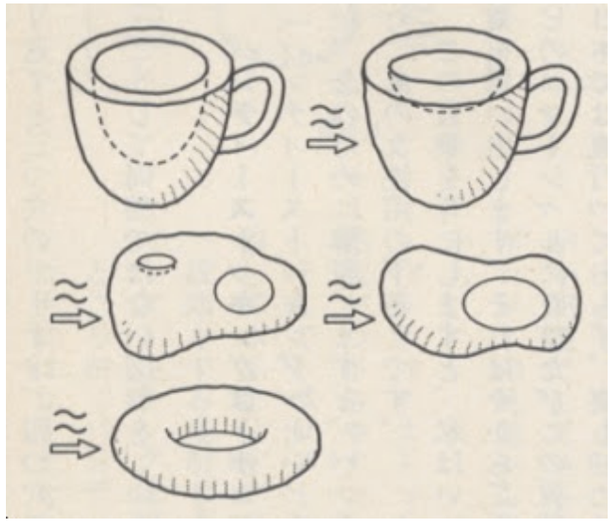


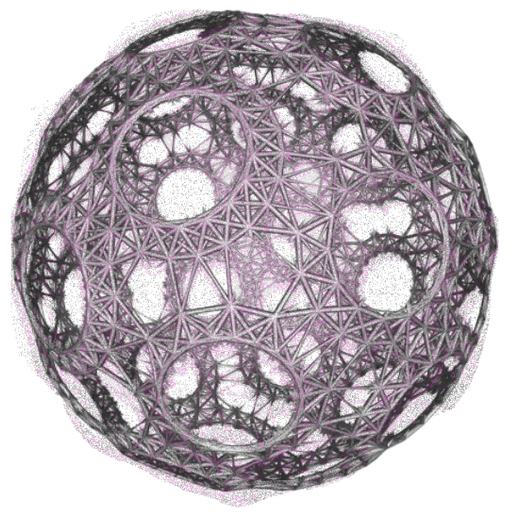


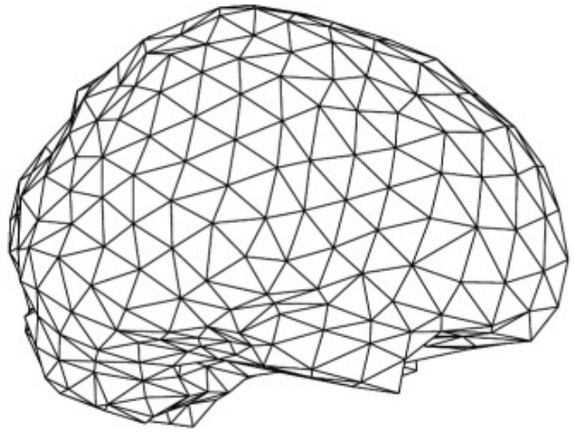


Three main advantages:

- robust
- (multiscale) notion of shape
- allows for higher order interactions





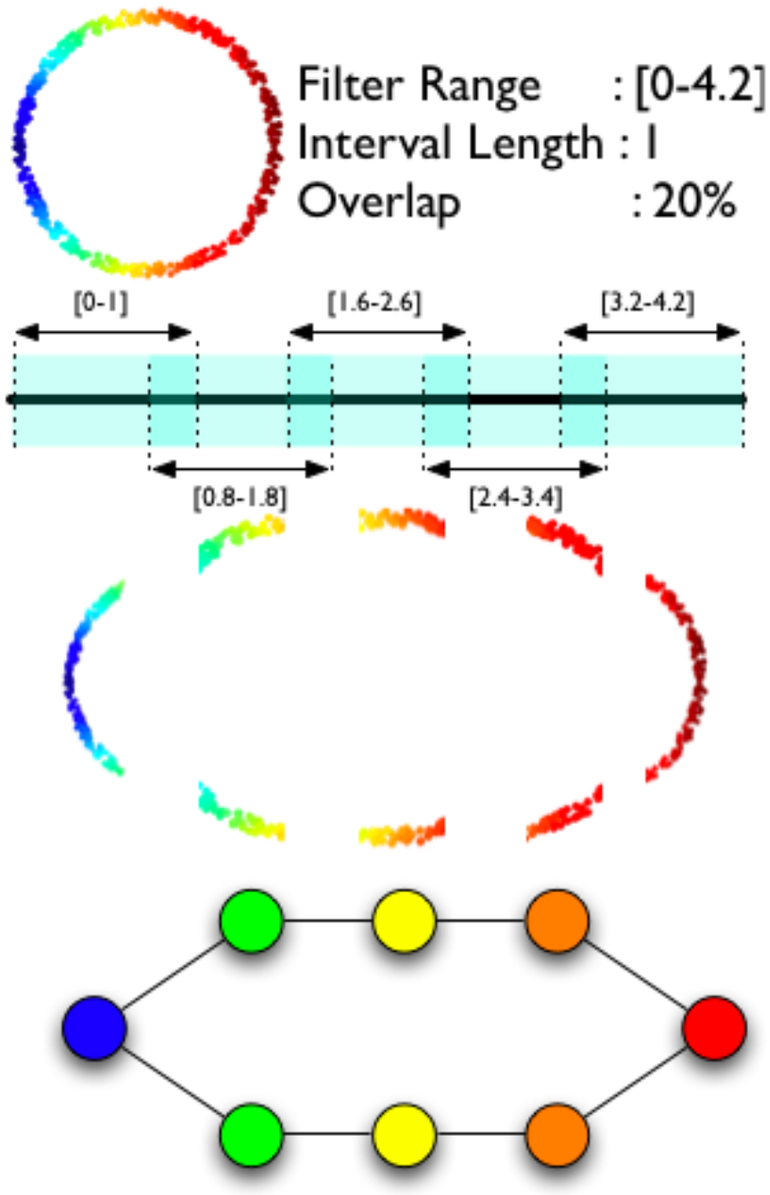


Topological methods

Two main techniques:

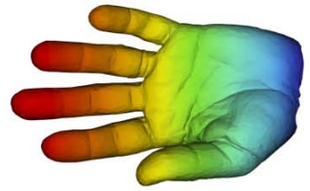
1. Topological simplification
    1. Interpretation of unstructured data
    2. Dimensionality reduction
2. Quantification of topological information
    1. Persistent homology
    2. Distances between homological structures

TDA as topological simplification

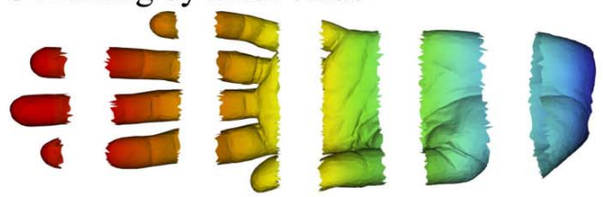


A Original Point Cloud

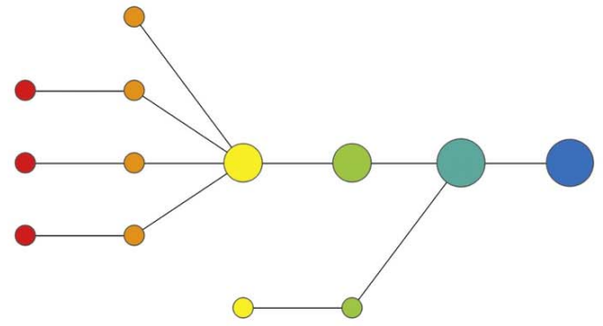
B Coloring by filter value



C Binning by filter value



D Clustering and network construction



Topological Methods for the Analysis of High Dimensional Data Sets and 3D Object Recognition

Gurjeet Singh[1] , Facundo Mémoli[2] and Gunnar Carlsson[†2]

SCIENTIFIC REPORTS

OPEN

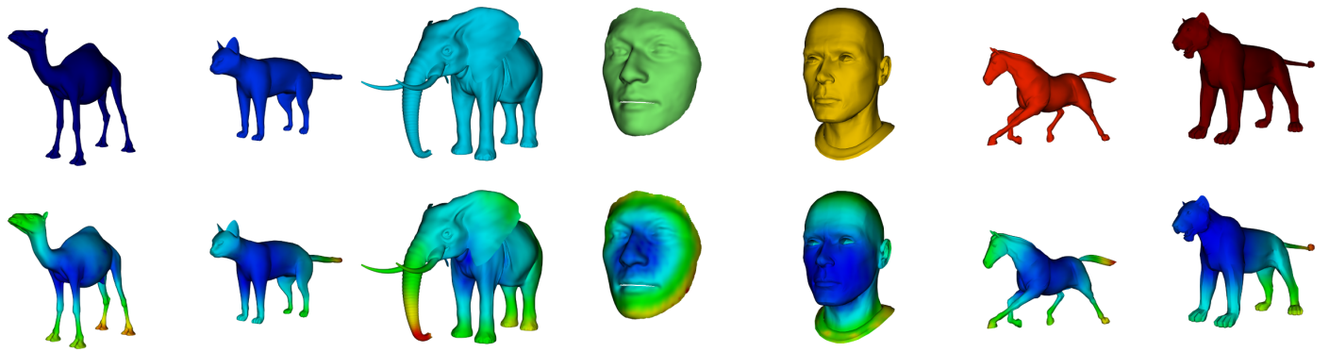
Extracting insights from the shape of complex data using topology

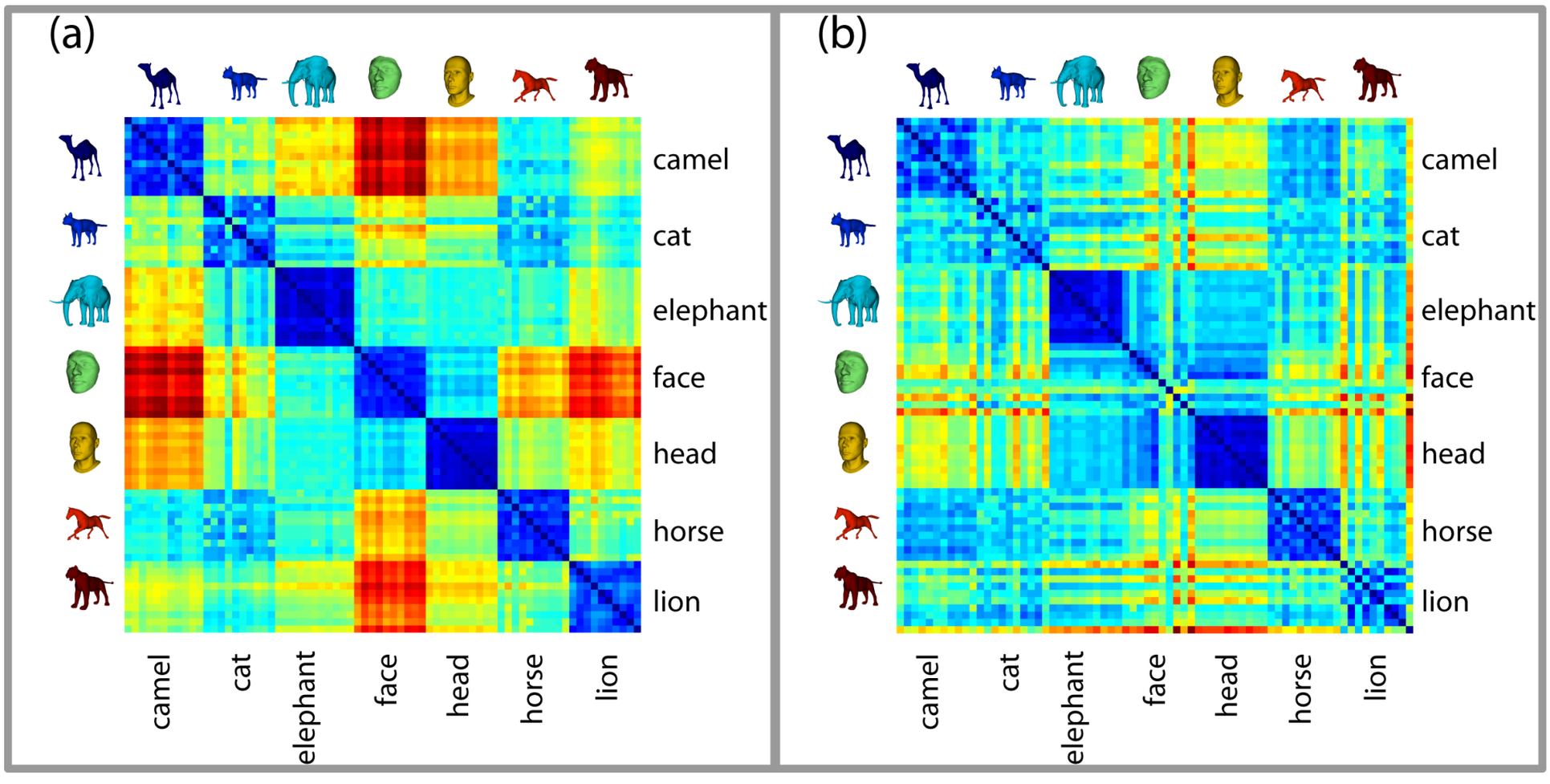
P. Y. Lum[1], G. Singh[1], A. Lehman[1], T. Ishkanov[1], M. Vejdemo-Johansson[2], M. Alagappan[1], J. Carlsson[3] & G. Carlsson[1,4]

[1]Ayasdi Inc., Palo Alto, CA, [2]School of Computer Science, Jack Cole Building, North Haugh, St. Andrews KY16 9SX, Scotland, United Kingdom, [3]Industrial and Systems Engineering, University of Minnesota, 111 Church St. SE, Minneapolis, MN 55455, USA, [4]Department of Mathematics, Stanford University, Stanford, CA, 94305, USA.

SUBJECT AREAS:
APPLIED MATHEMATICS
COMPUTATIONAL SCIENCE
SCIENTIFIC DATA
SOFTWARE

TDA as topological simplification



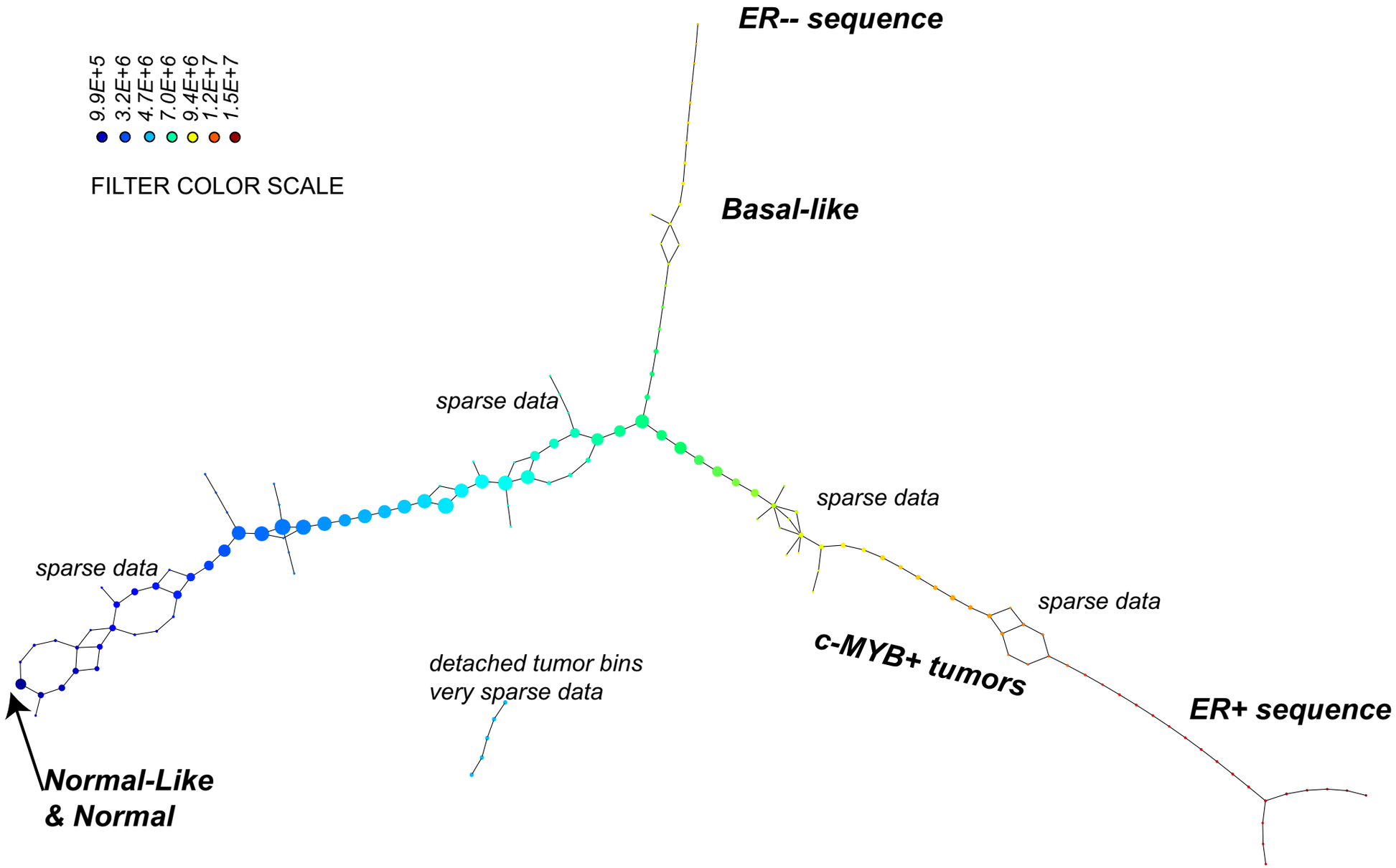


Topological Methods for the Analysis of High Dimensional Data Sets and 3D Object Recognition

Gurjeet Singh[1] , Facundo Mémoli[2] and Gunnar Carlsson[12]

Notebook 01

Classical application: breast cancers



PNAS

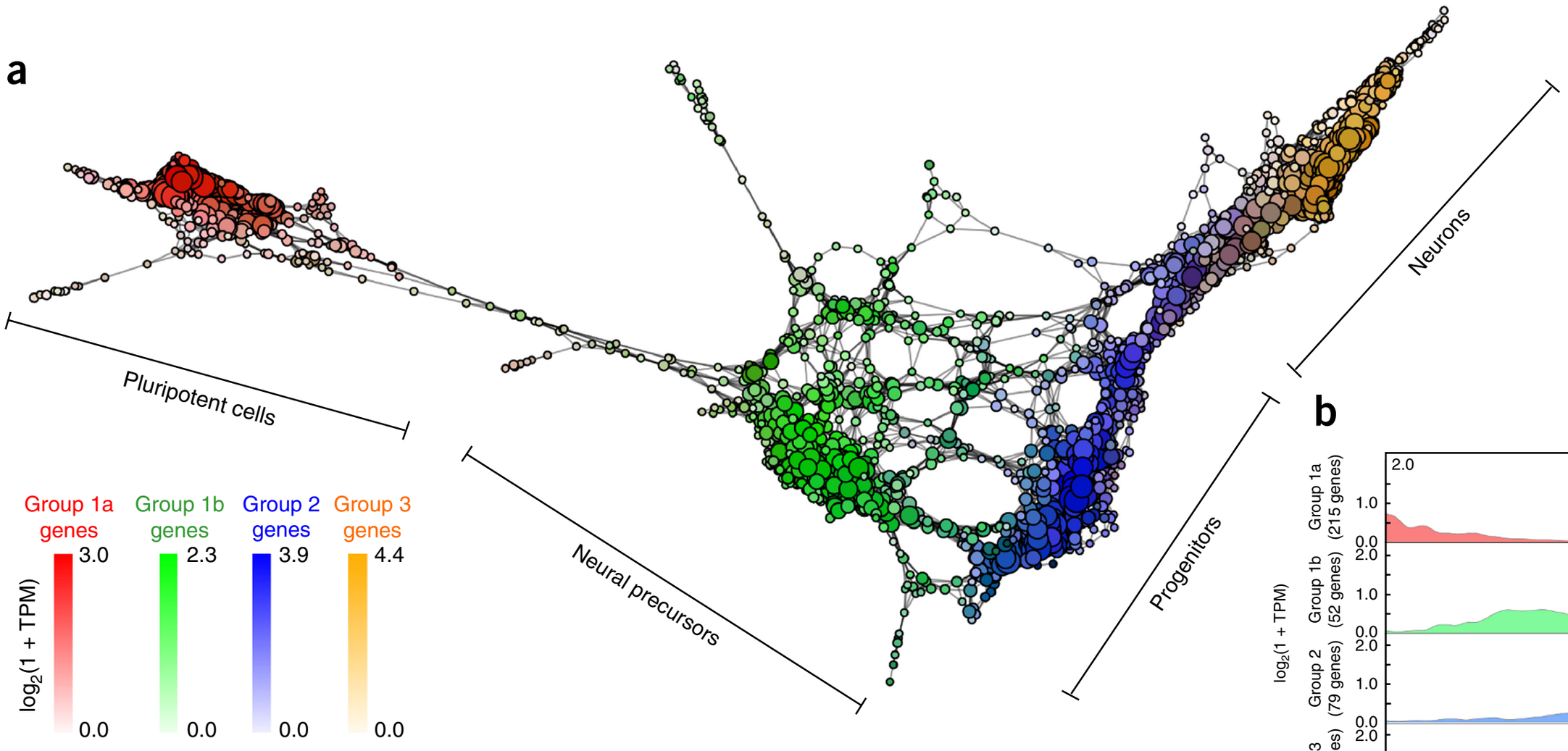
Topology based data analysis identifies a subgroup of breast cancers with a unique mutational profile and excellent survival

Monica Nicolau[a], Arnold J. Levine[b,1], and Gunnar Carlsson[a,c]

[a]Department of Mathematics, Stanford University, Stanford, CA 94305; [b]School of Natural Sciences, Institute for Advanced Study, Princeton, NJ 08540; and [c]Ayasdi, Inc., Palo Alto, CA 94301

Contributed by Arnold J. Levine, February 25, 2011 (sent for review July 23, 2010)

More recent: cellular differentiation



ARTICLES

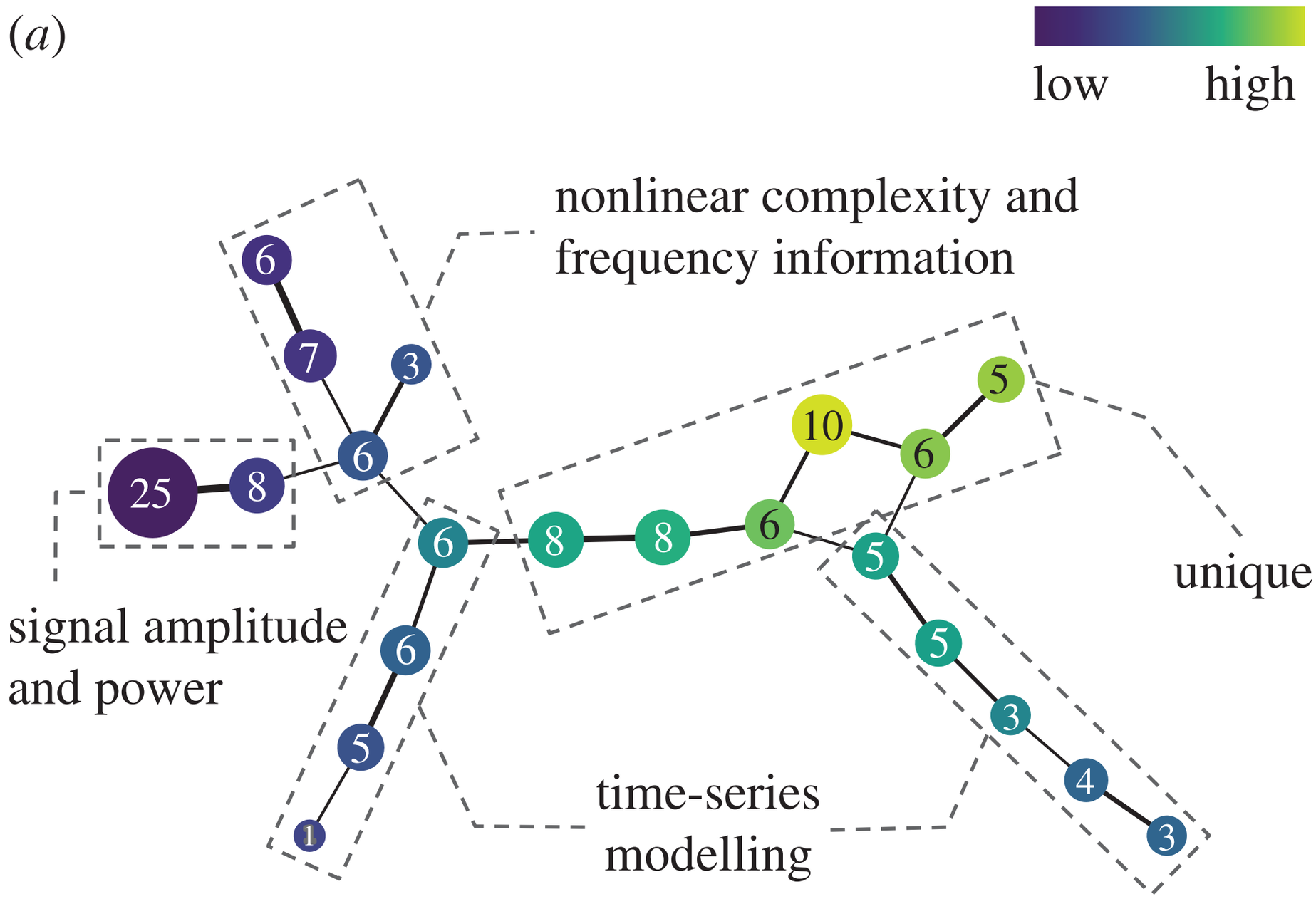
nature biotechnology

Single-cell topological RNA-seq analysis reveals insights into cellular differentiation and development

Abbas H Rizvi[1,2,6], Pablo G Camara[3,4,6], Elena K Kandror[1,2], Thomas J Roberts[1,2,4], Ira Schieren[2,5], Tom Maniatis[1,2] & Raul Rabadan[3,4]

Notebook 02

TDA as topological feature map



INTERFACE

rsif.royalsocietypublishing.org

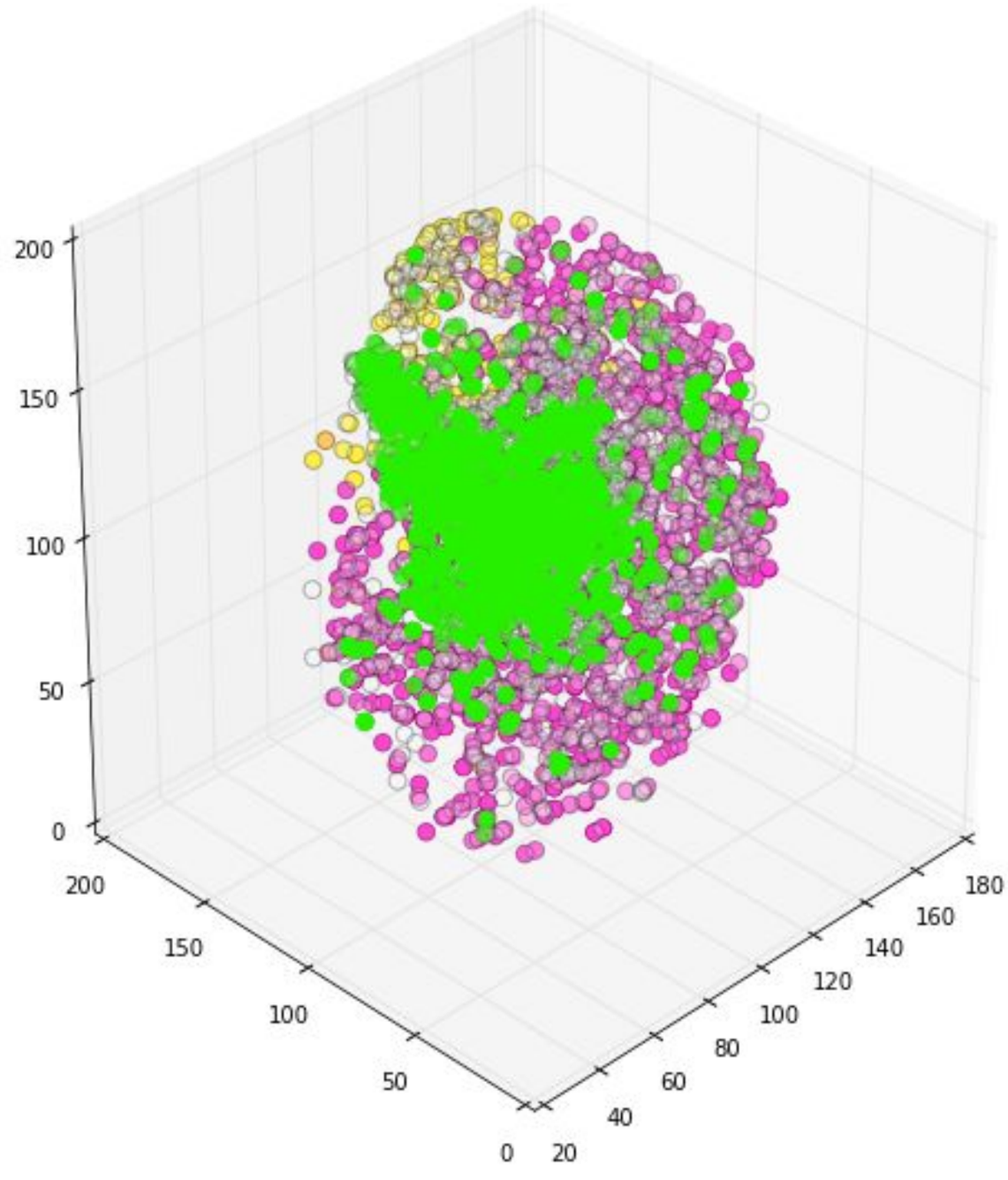
Research

Navigating features: a topologically informed chart of electromyographic features space

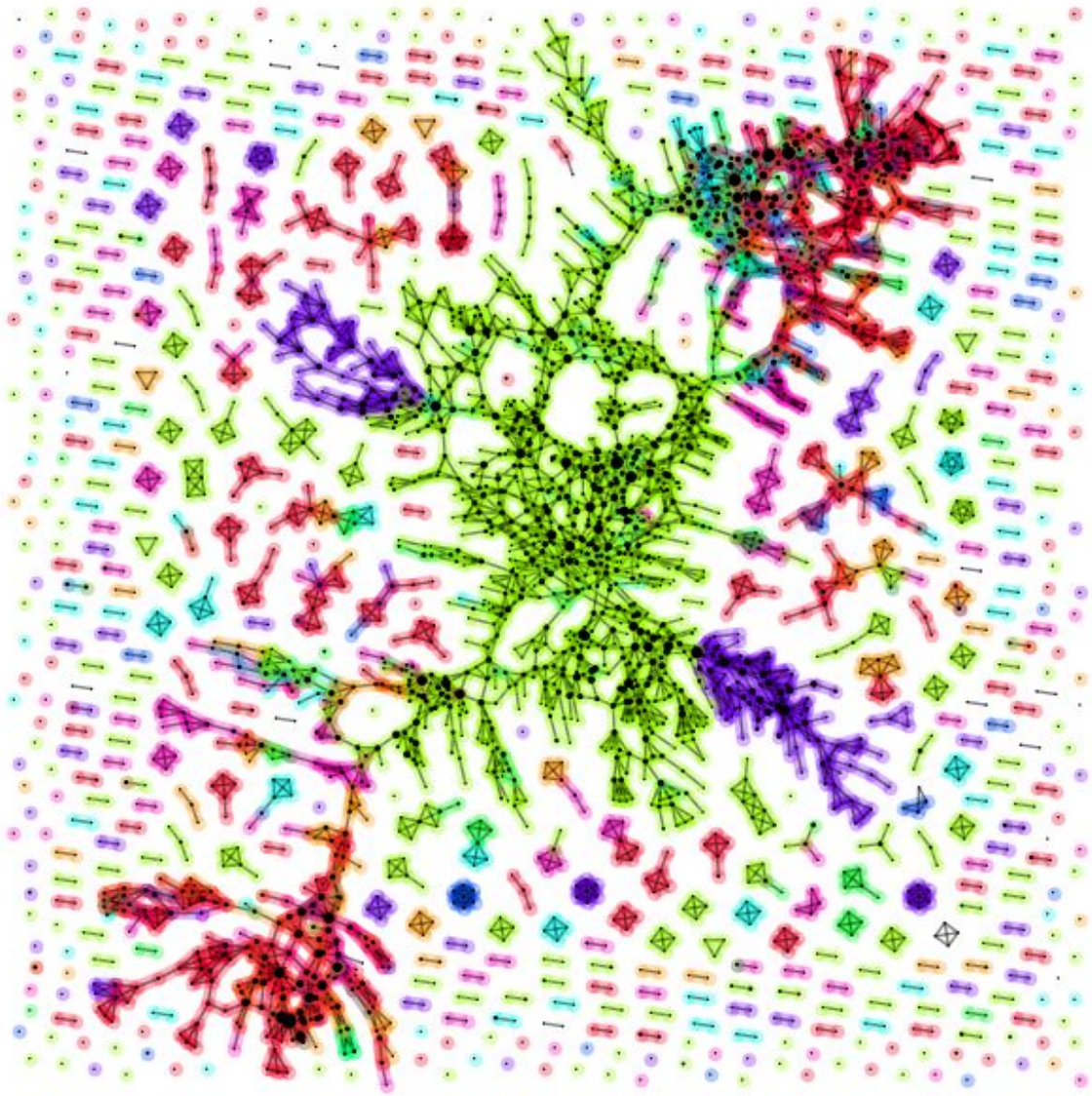
Angkoon Phinyomark[1,2], Rami N. Khushaba[3], Esther Ibáñez-Marcelo[1], Alice Patania[4], Erik Scheme[2] and Giovanni Petri[1]

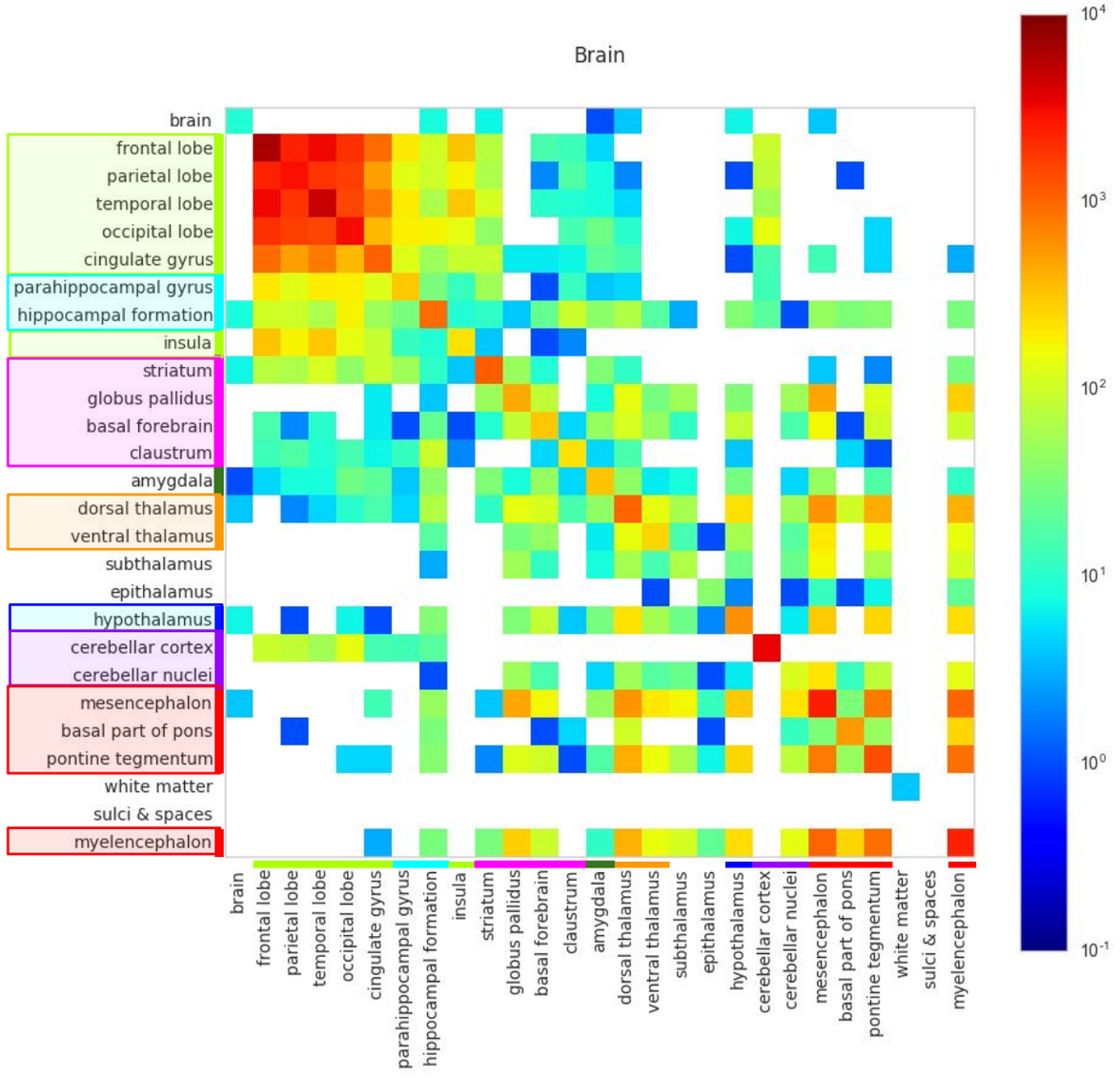
Notebook 03

Gene topological networks



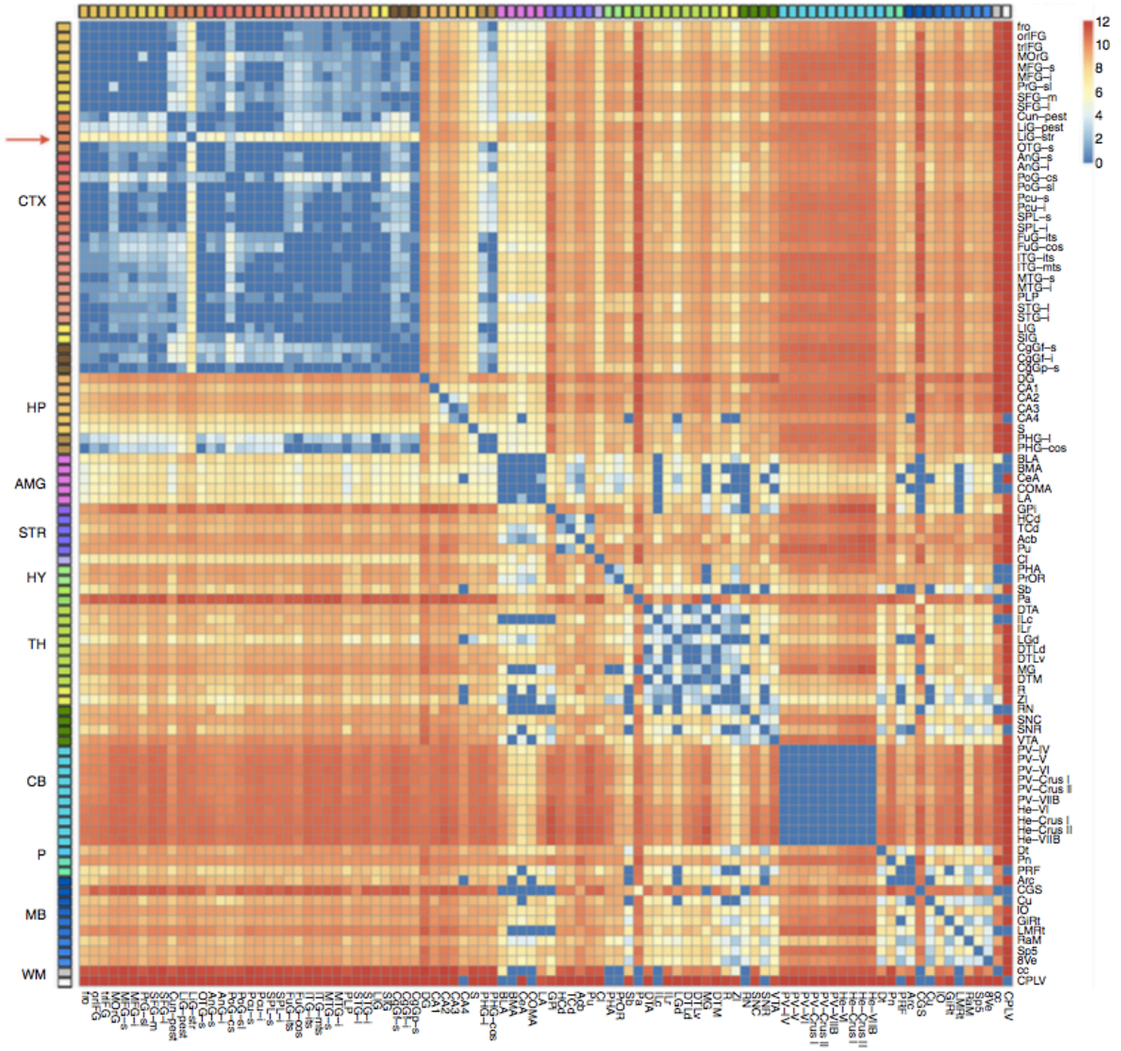
Gene topological networks

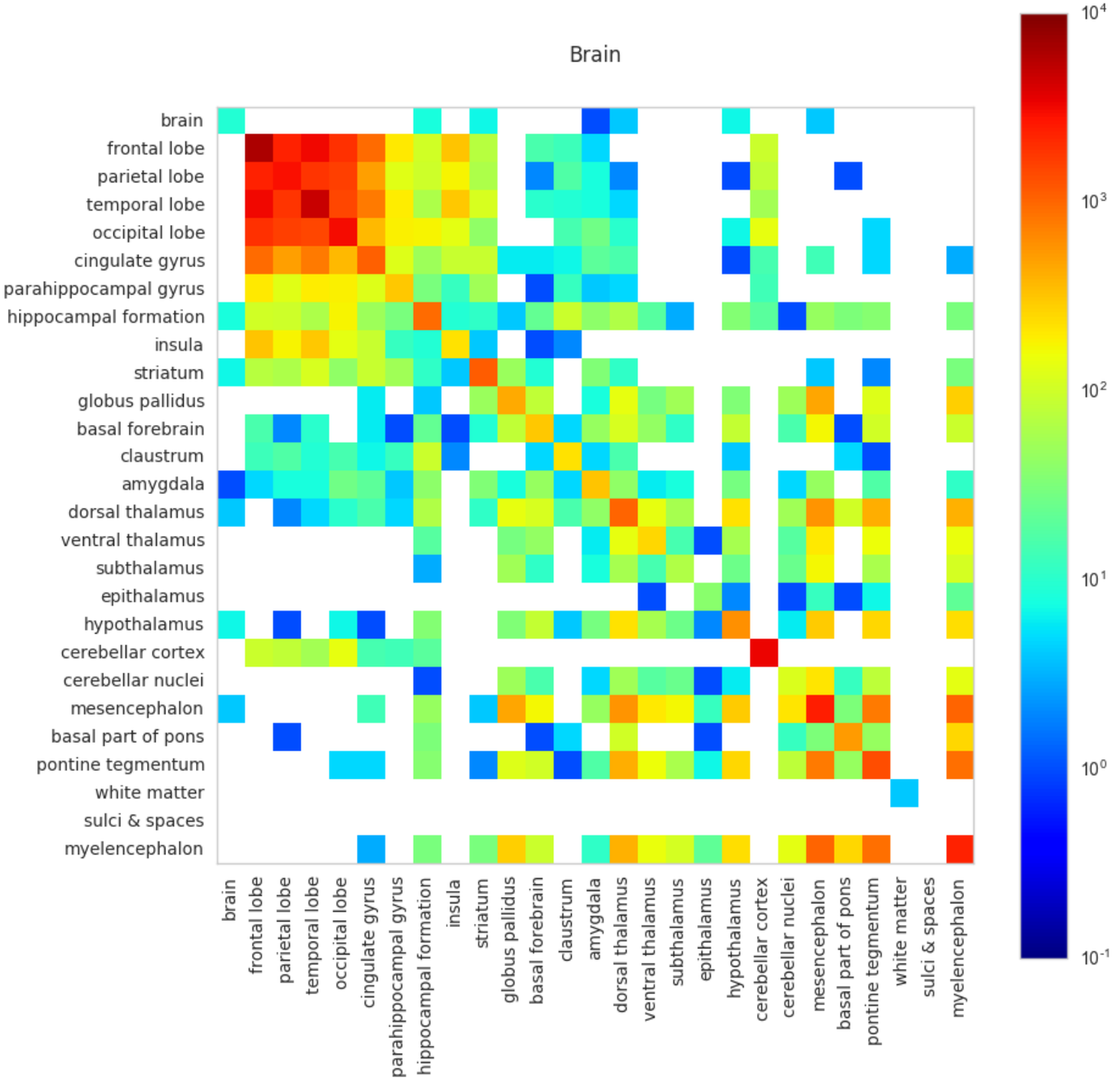




basal_ganglia
hypothalamus
amygdala
thalamus
cerebellum
hippocampus
neocortex
brainstem

Gene topological networks

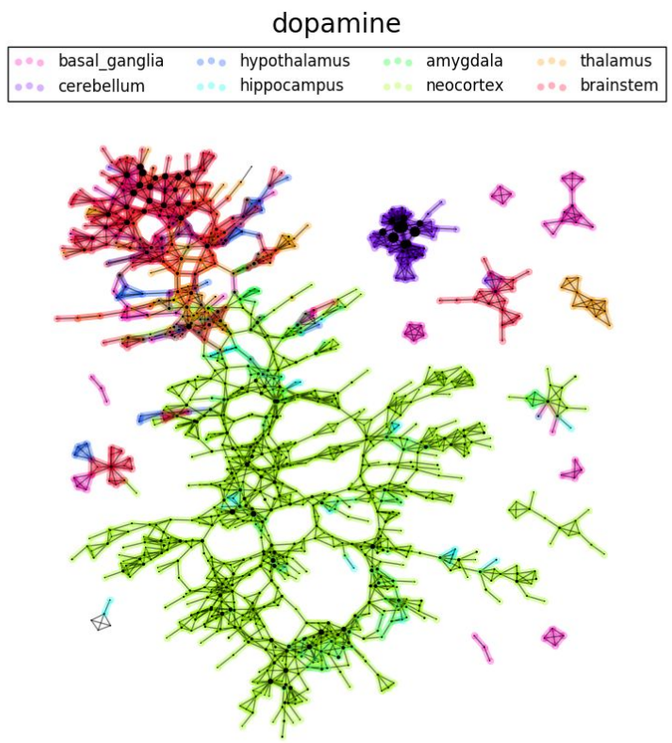




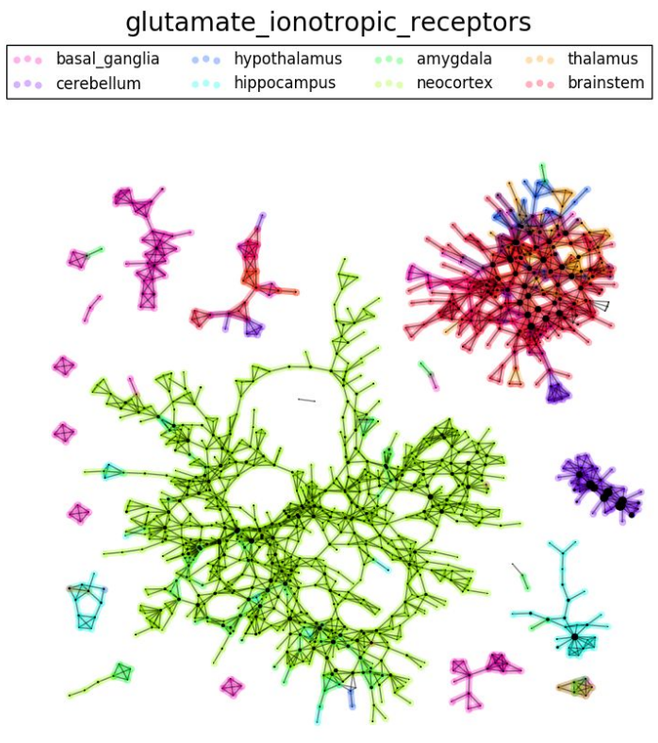
Consensus map of all genes differentially expressed between any pair of 96 regions in at least five of six specimens.

Hawrylycz, Michael, et al. "Canonical genetic signatures of the adult human brain." *Nature neuroscience* 18.12 (2015): 1832-1844.

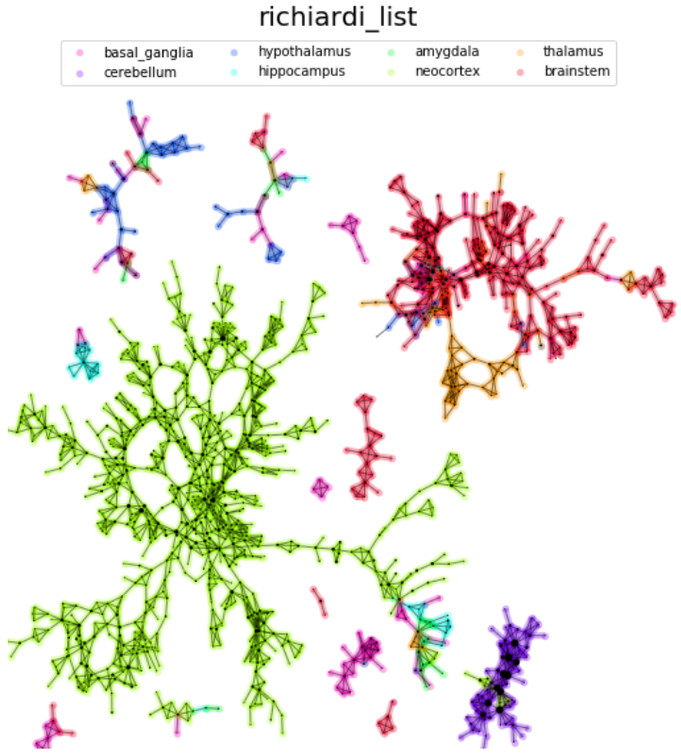
Pathway-specific simplicial networks



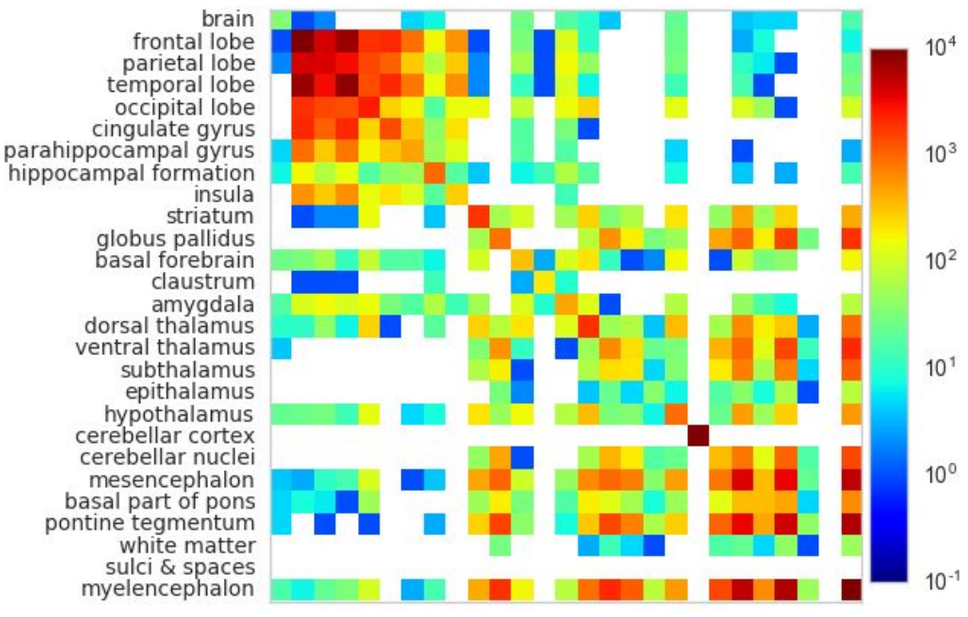
Dopamine

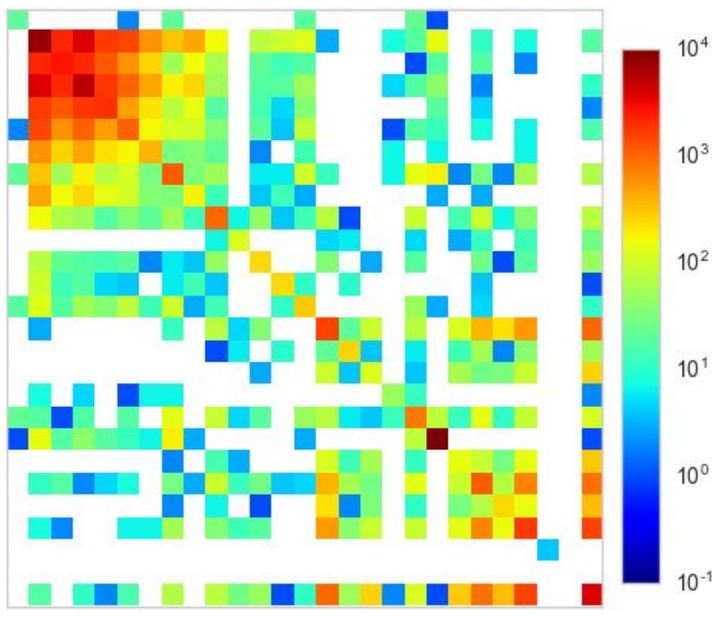


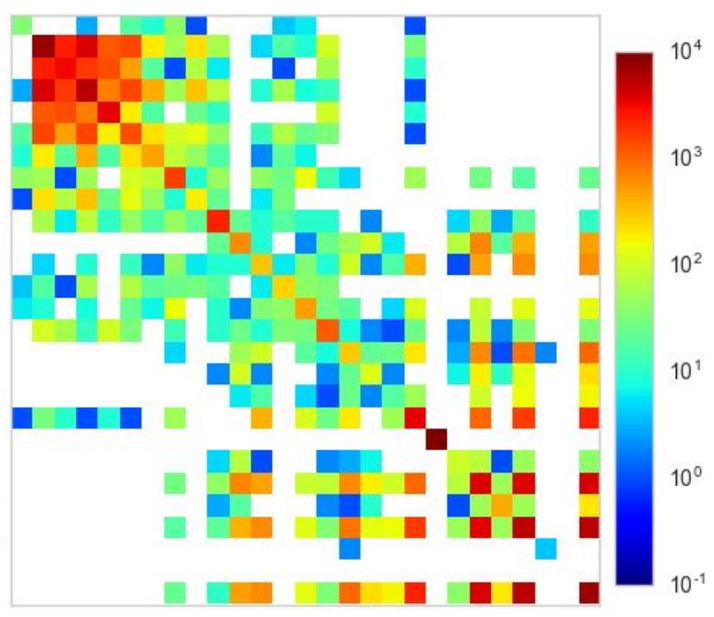
Glutamate receptors



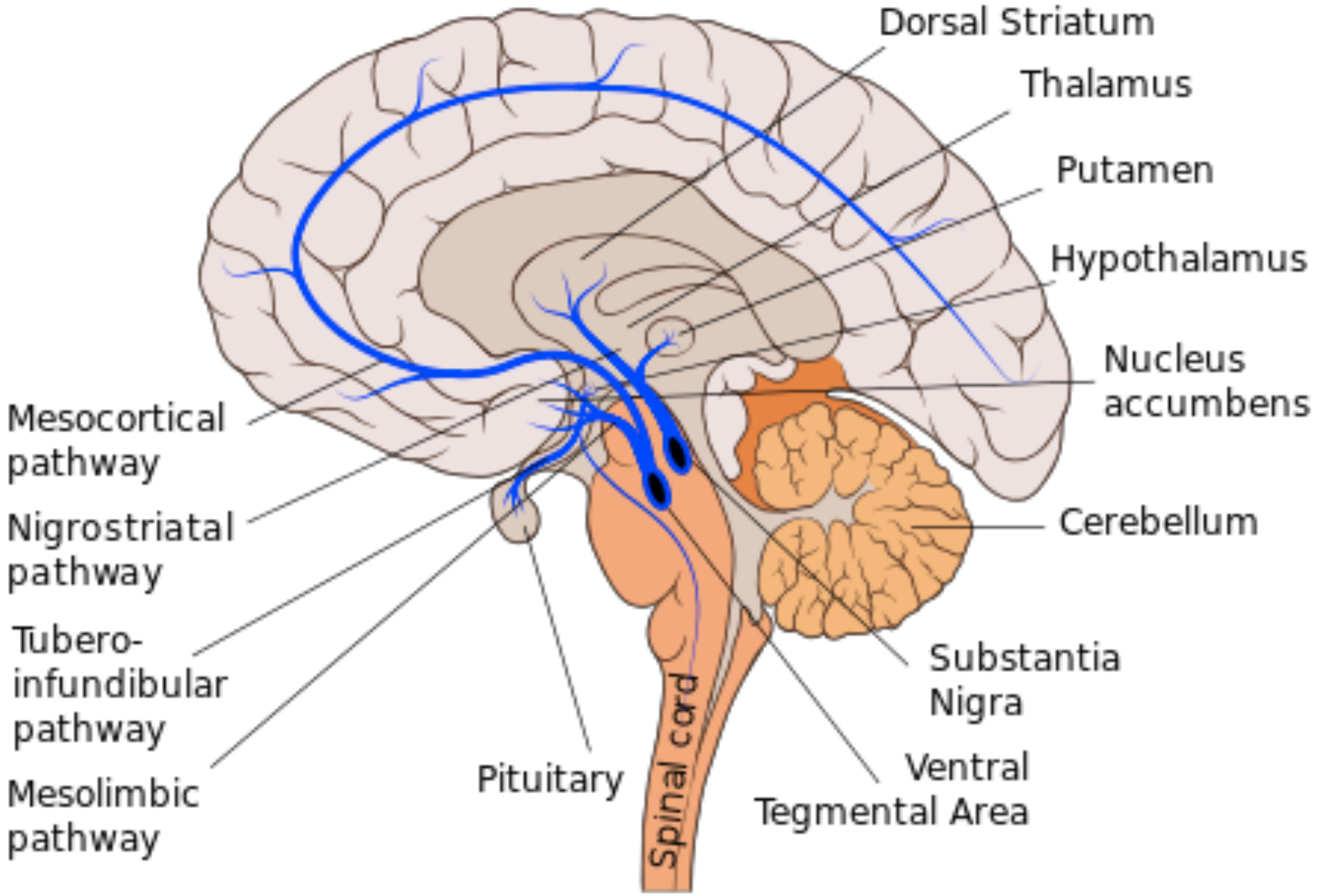
GABA receptors







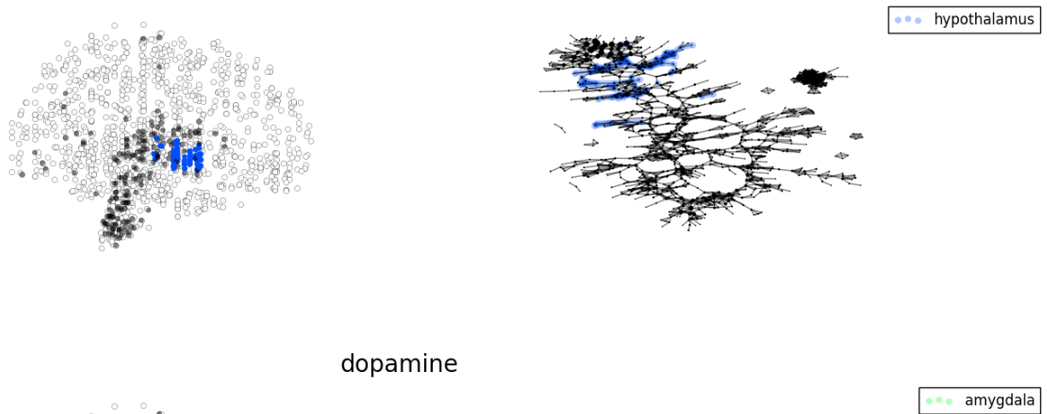
Dopamine loop



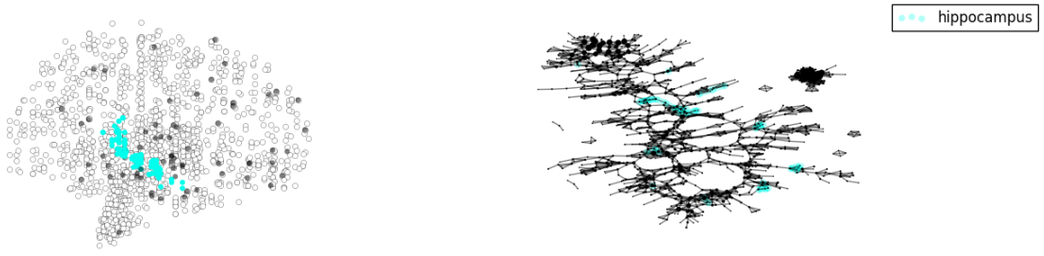
Dopamine loop

dopamine

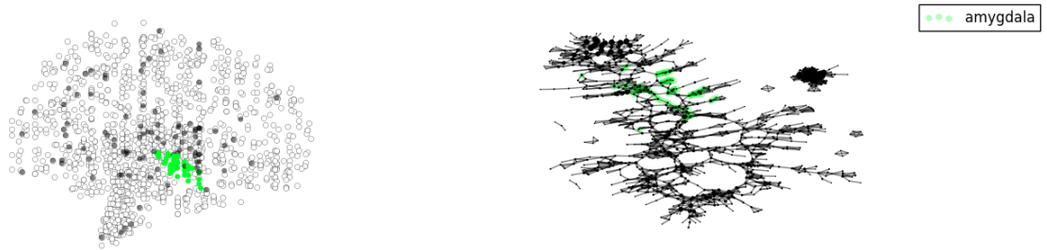
dopamine



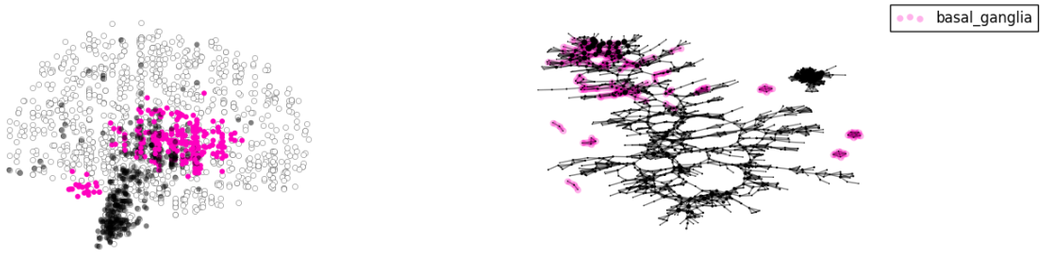
dopamine



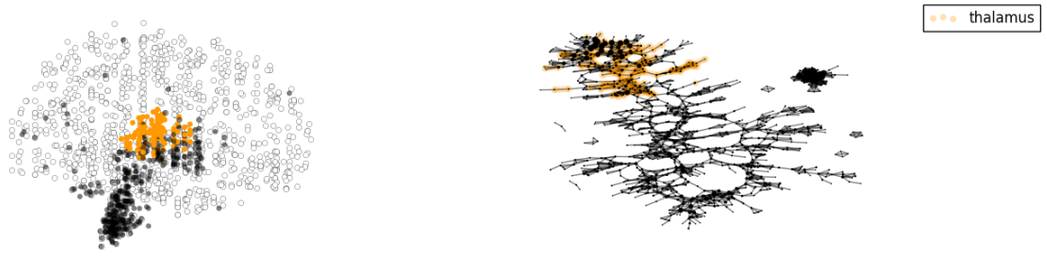
dopamine



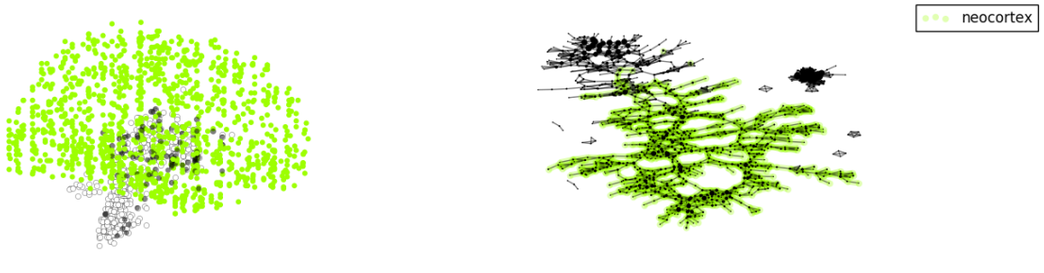
dopamine



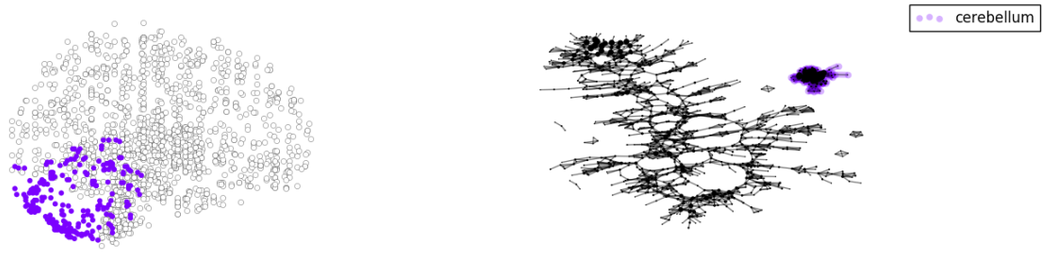
dopamine



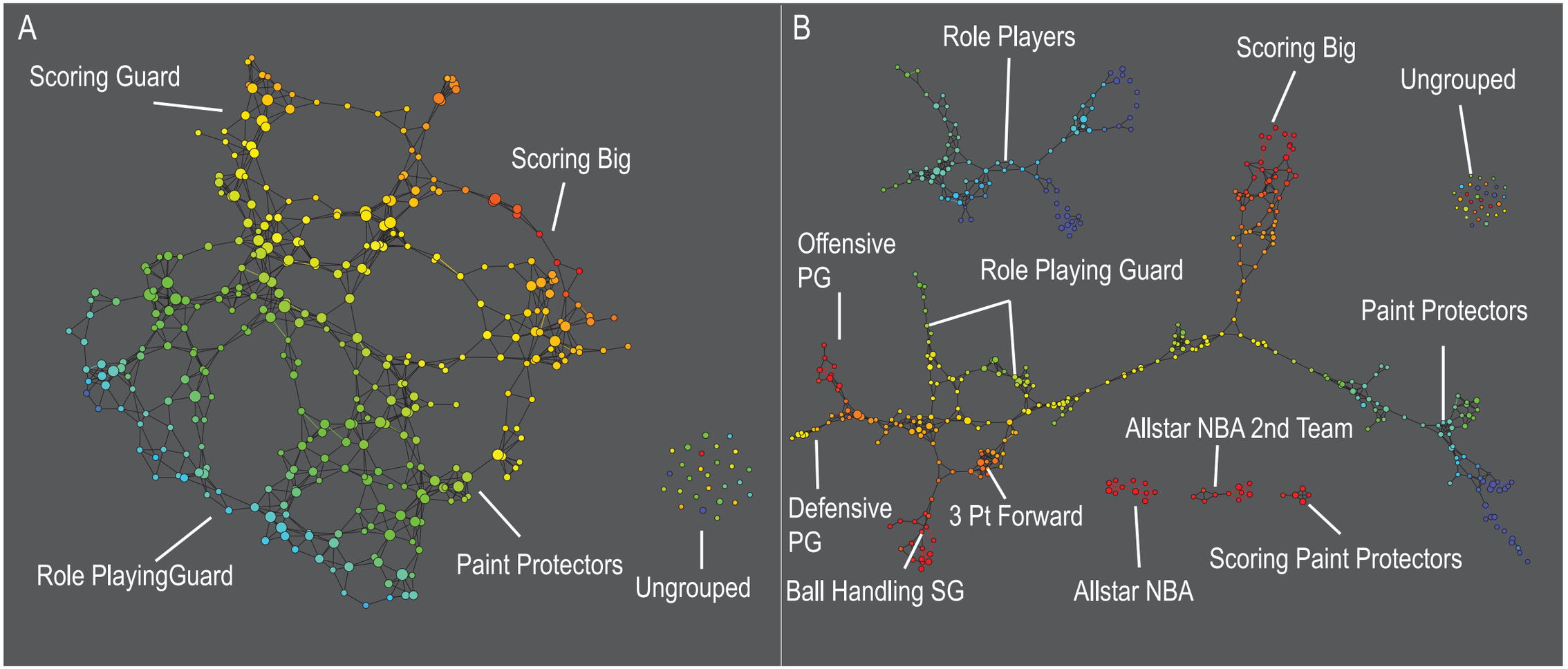
dopamine



dopamine

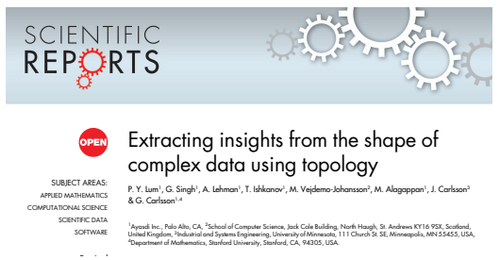


Mapper: quantification?



- Resolutions, parameter, choices..

- We are doing this by eye.


SCIENTIFIC REPORTS

OPEN

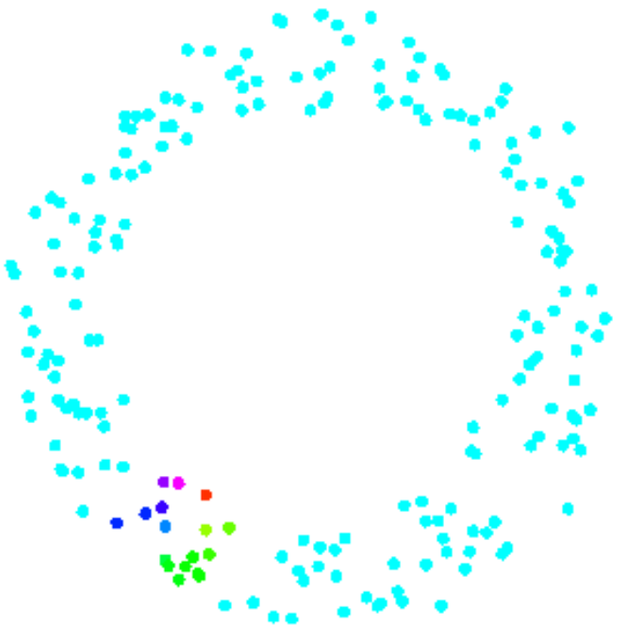
Extracting insights from the shape of complex data using topology

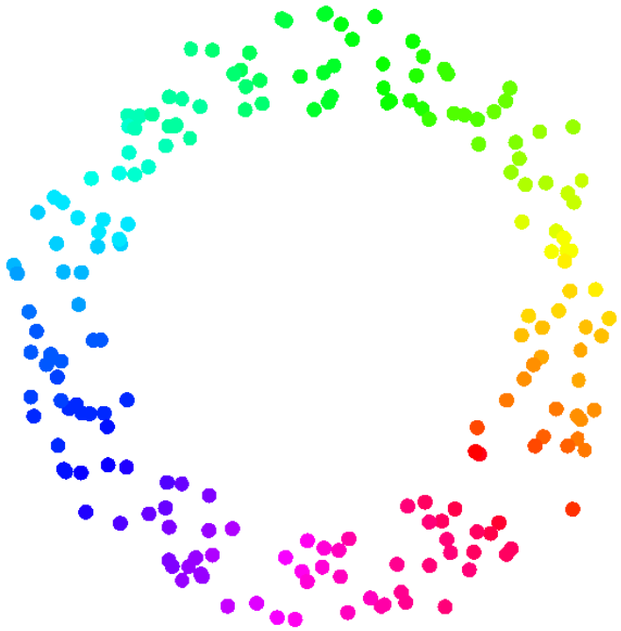
SUBJECT AREAS:
APPLIED MATHEMATICS
COMPUTATIONAL SCIENCE
SCIENTIFIC DATA
SOFTWARE

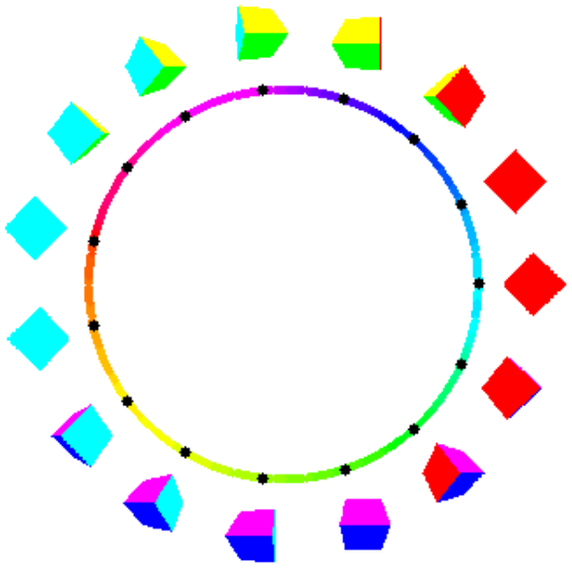
P. Y. Lum[1], G. Singh[1], A. Lehman[1], T. Ishkanov[2], M. Vejdemo-Johansson[3], M. Alagappan[1], J. Carlsson[3] & G. Carlsson[1,4]

[1]Ayasdi Inc., Palo Alto, CA, [2]School of Computer Science, Jack Cole Building, North Haugh, St. Andrews KY16 9SX, Scotland, United Kingdom, [3]Industrial and Systems Engineering, University of Minnesota, 111 Church St. SE, Minneapolis, MN 55455, USA, [4]Department of Mathematics, Stanford University, Stanford, CA, 94305, USA.

Ok, so what about circular coordinates?



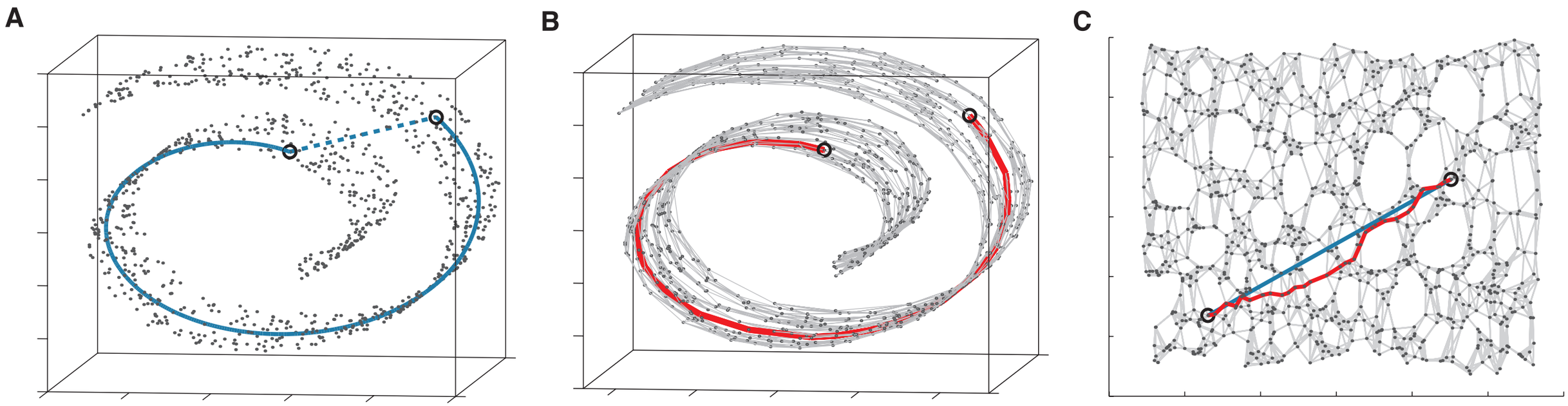




Dimensional reduction

- Standard clustering or MDS techniques assume linear dimensions

- What about non linear ones?



- Substitute distance matrix with local graph.

- Travel across the graph.

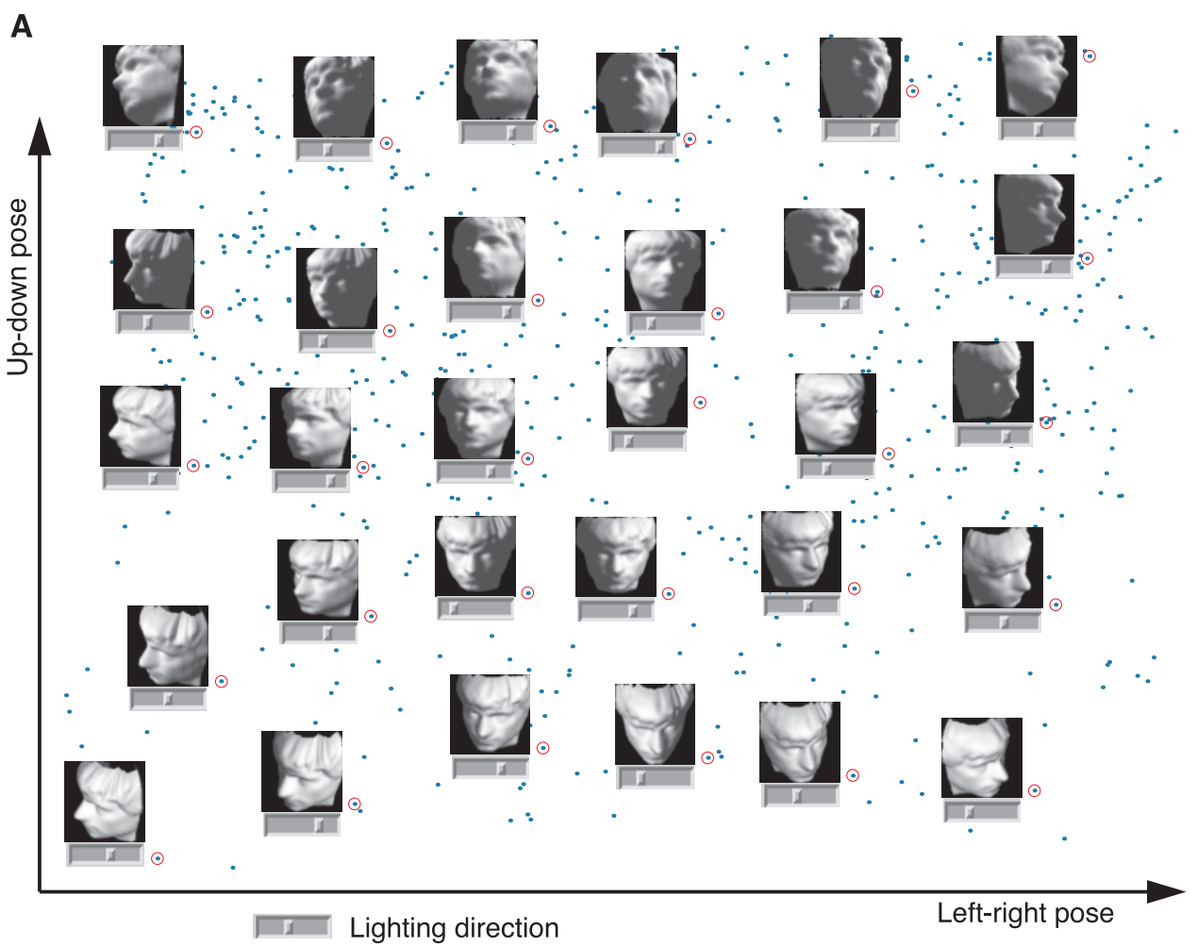
Tenenbaum et al., Science 2000

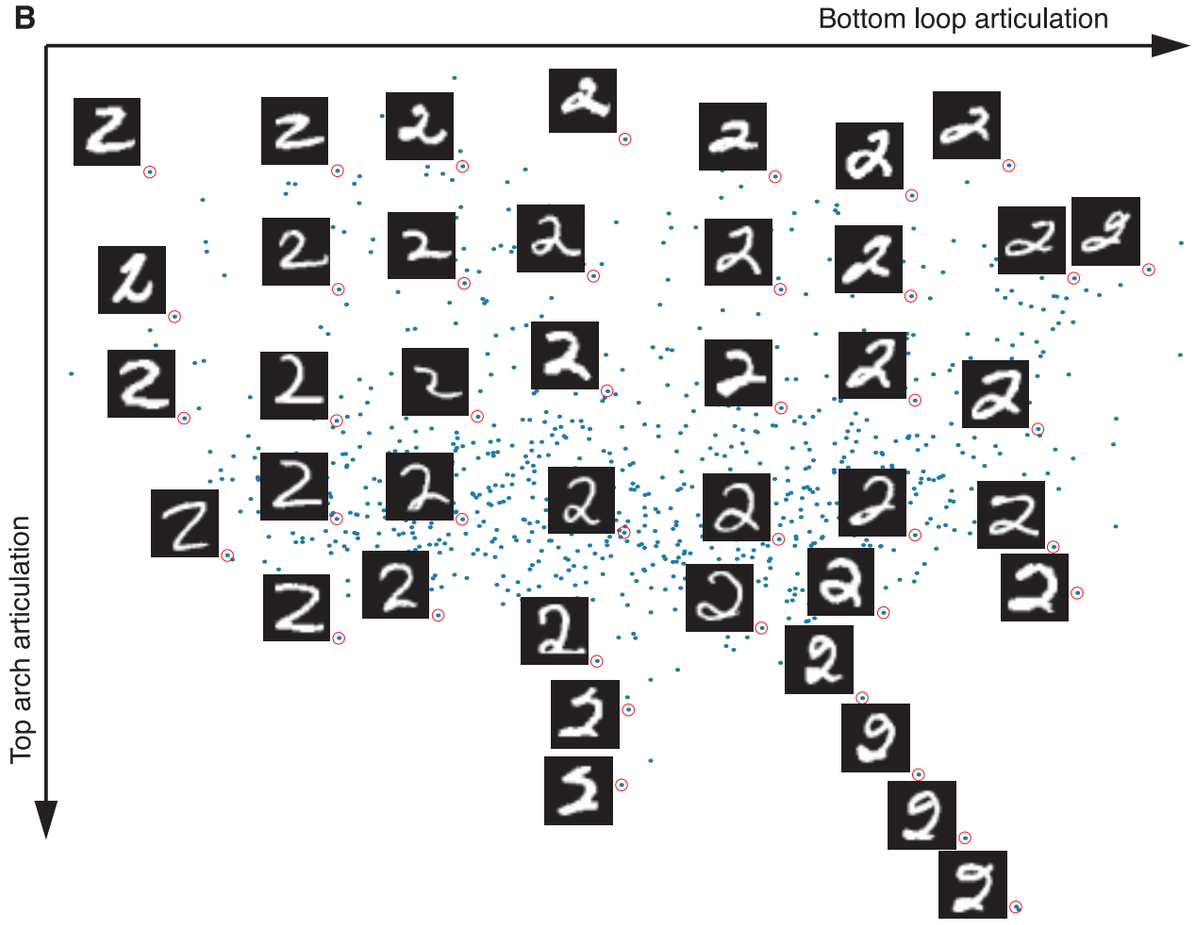
Tenenbaum et al., Science 2000

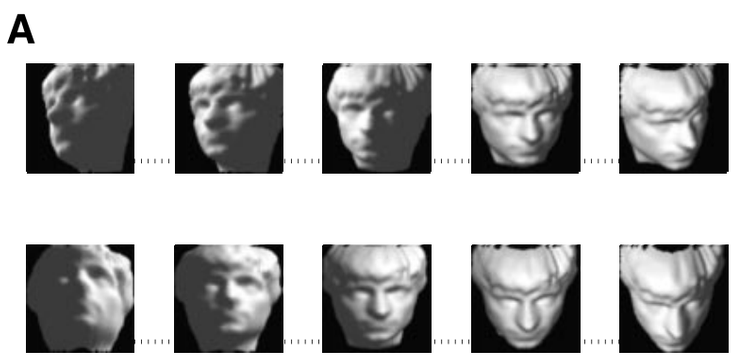
Dimensional reduction: Isomap

- It works extremely well.

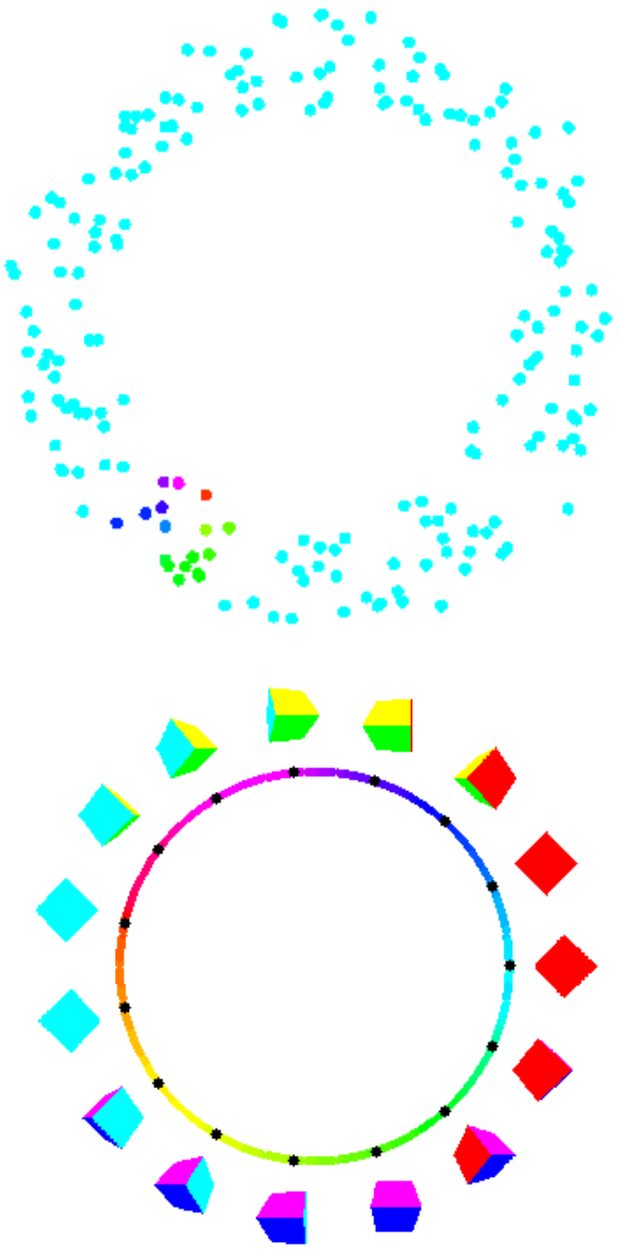
- Uncovers meaningful “lines”.

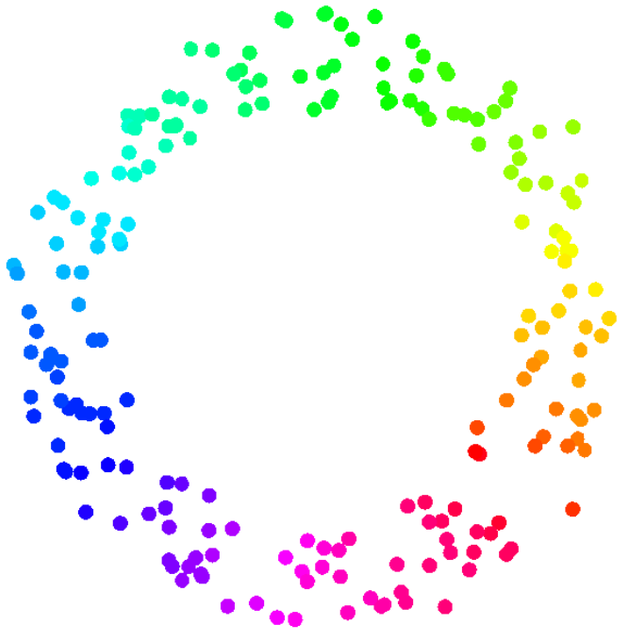






Ok, so what about circular coordinates?

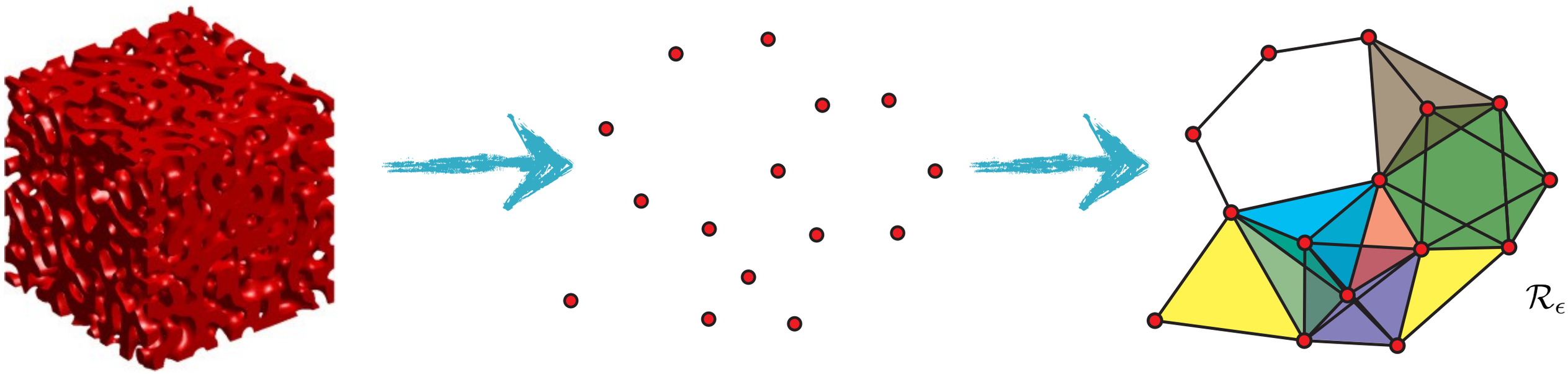




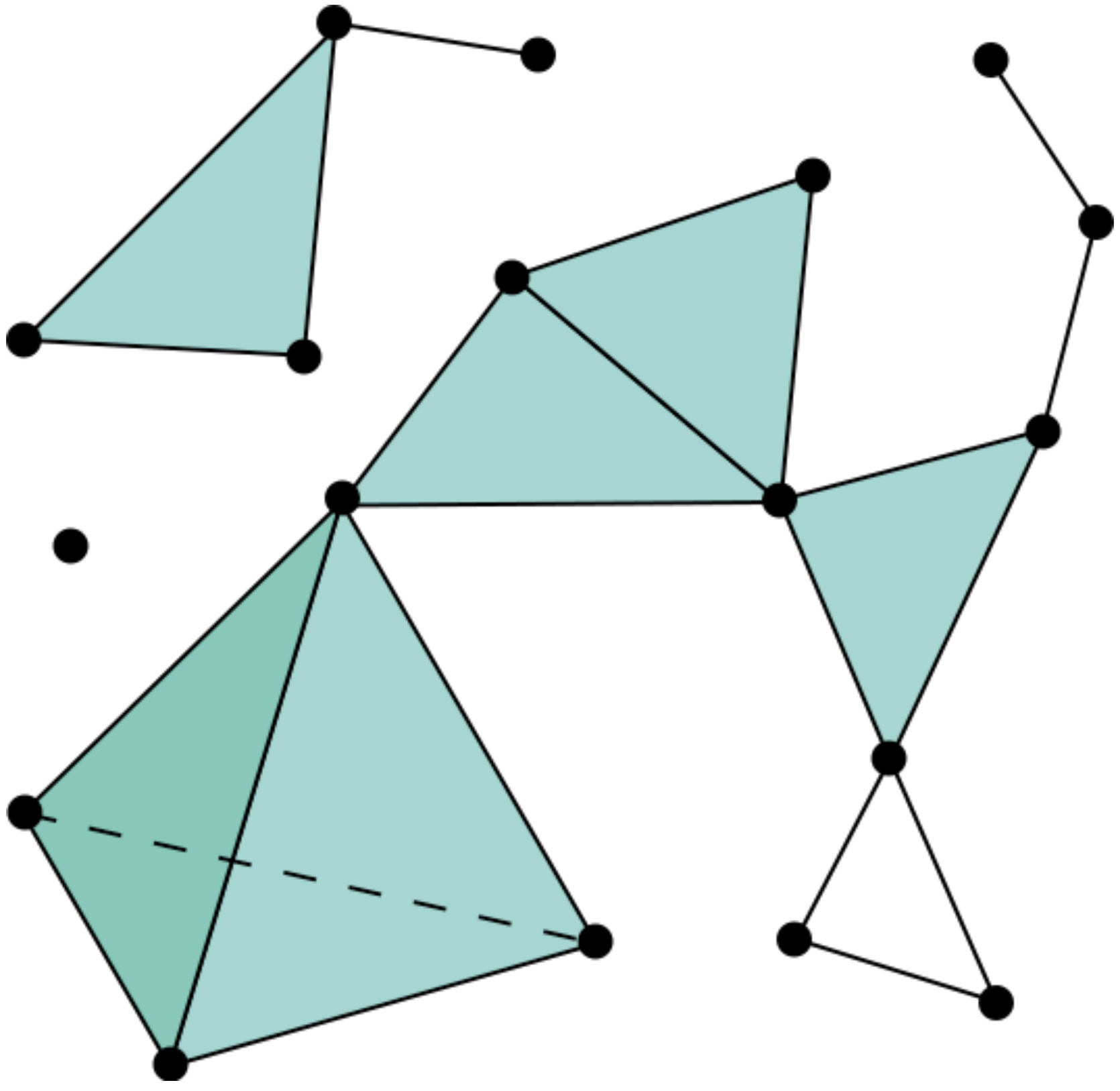
Need **persistent homology**

The shape of data: homology

- old concept
- persistent version introduced by Carlsson, Zomorodian, Ghrist, De Silva, Edelsbrunner, ...('02)
- implemented and used to understand the shape of data (Javaplex, Dionysus, Holes, Ripser, Phat,Gudhi...)



The shape of data: homology II



The shape of data: homology III

Definition

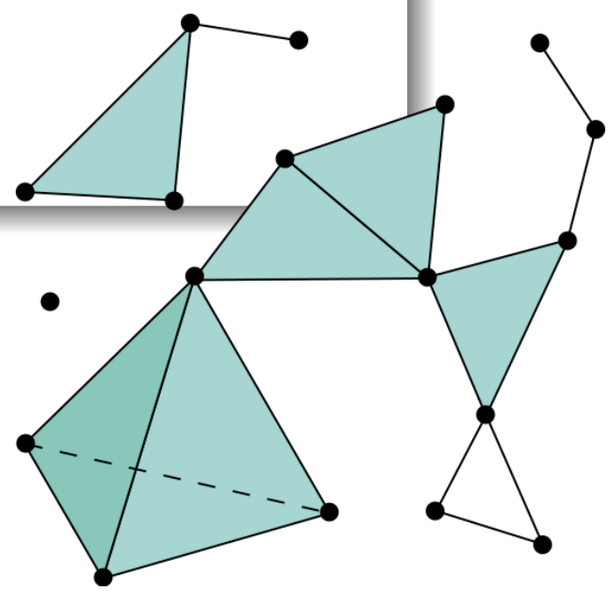
A $n-$dimensional *SIMPLEX* in $\mathbb{R}^m$ is the convex hull of $n+1$ points in general position in $\mathbb{R}^m$. $(m \geq n)$

In $\mathbb{R}^3$: 

Definition

A *SIMPLICIAL COMPLEX* in $\mathbb{R}^m$ is a family X of simplices in $\mathbb{R}^m$ such that:

- $\sigma \in X$ and $\tau \subset \sigma$ then $\tau \in X$
- $\sigma^1 \cap \sigma^2 = \emptyset$ or $\sigma^1 \cap \sigma^2 \subseteq \sigma^1$ and $\sigma^1 \cap \sigma^2 \subseteq \sigma^2$.



The shape of data: homology IV

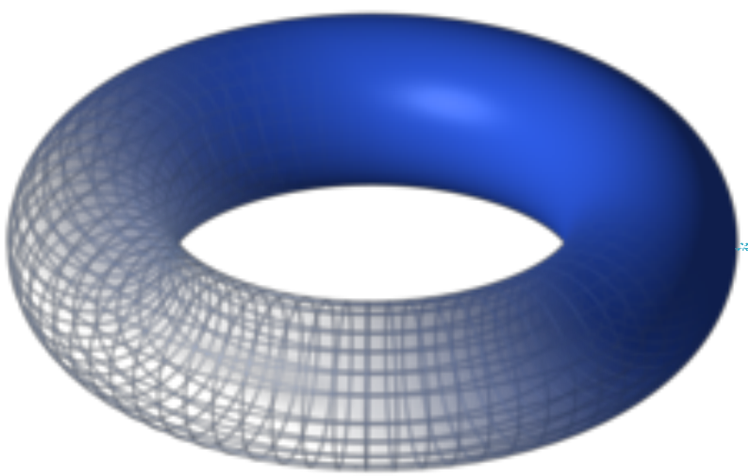
C_n the k-vector space with basis n-dimensional faces of X and

$\partial_n : \quad C_n \to C_{n-1}$ the boundary operator

$$\partial_n[p_1, \ldots, p_n] = \sum_{i=1}^{n} (-1)^i [p_1, \ldots, \hat{p}_i, \ldots, p_n]$$

fit in $C(X) : 0 \to C_t \to \ldots \to C_n \to \ldots C_0 \to 0$

and finally $H_n = \dfrac{\ker \partial_n}{\mathrm{Im} \partial_{n+1}}$



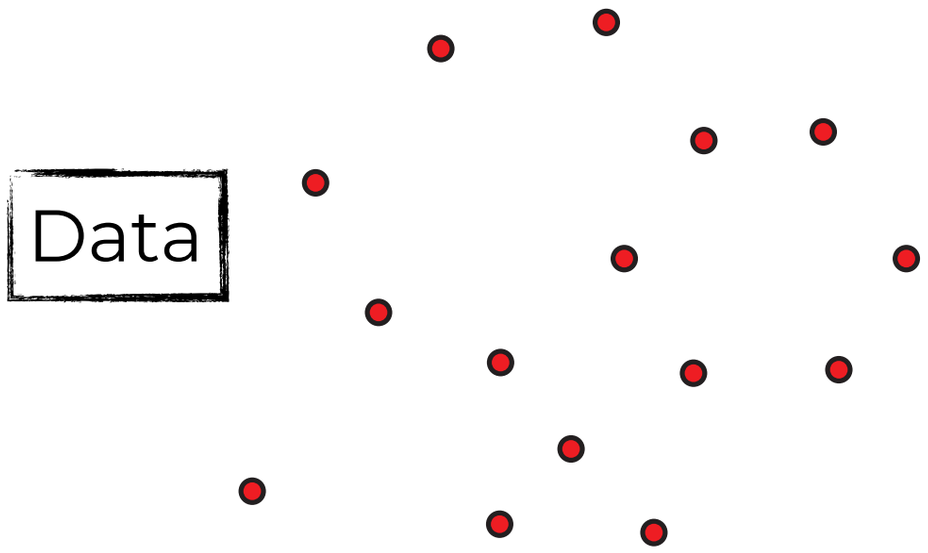
H_0 Conn. components (1)

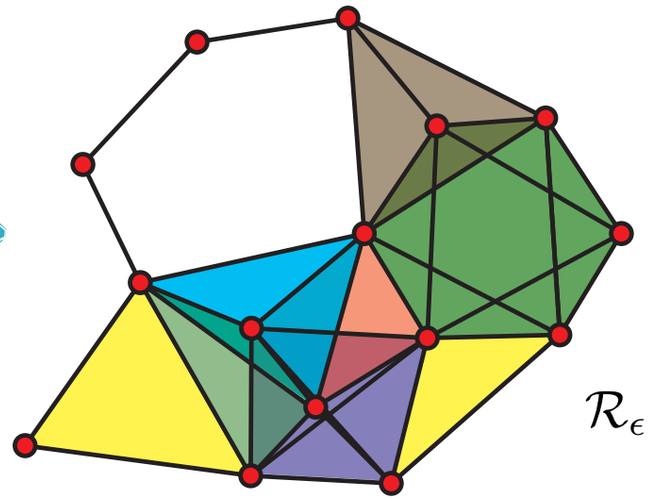
H_1 Holes (2)

H_2 Voids (1)

The shape of data: what simplicial complex?

Data



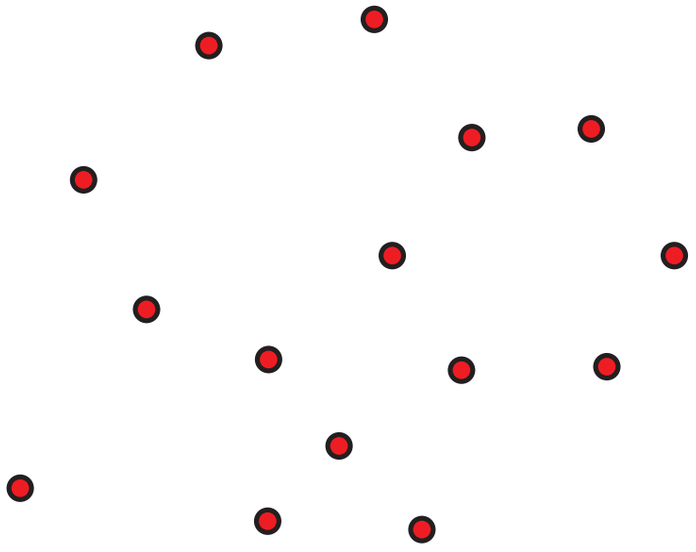


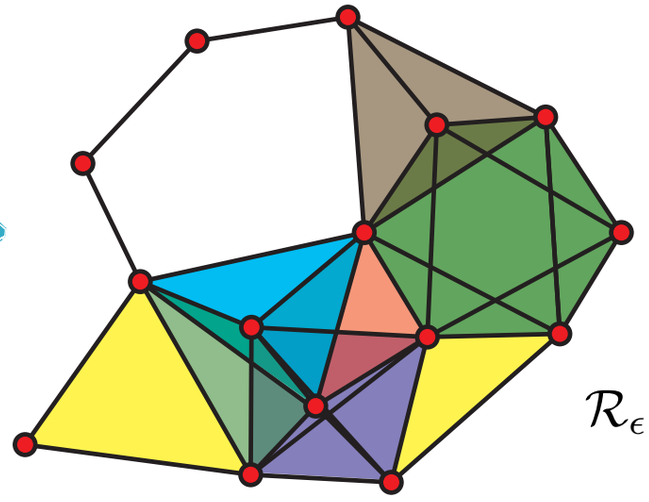
Simplicial complex

Cech complex
Alpha Complex
Lower-star complex
Dowker complex
Clique Complex
....

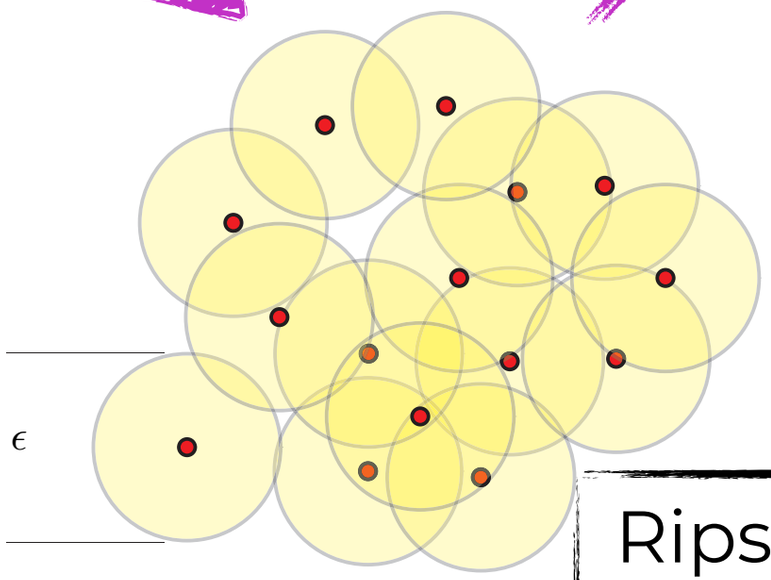
The shape of data: what simplicial complex?

Data





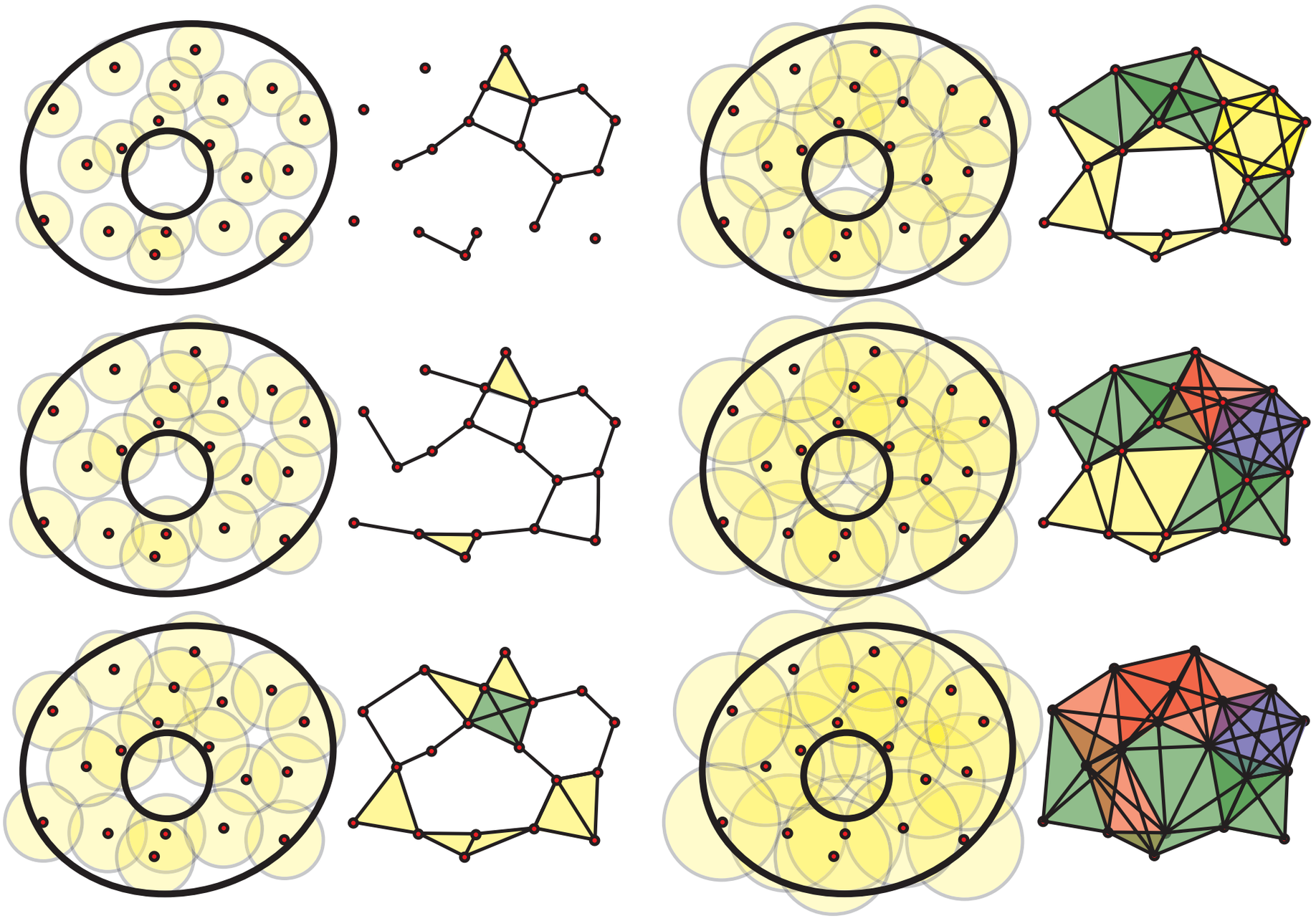
Simplicial complex



Rips-Vietoris complex

Cech complex
Alpha Complex
Lower-star complex
Dowker complex
Clique Complex
....

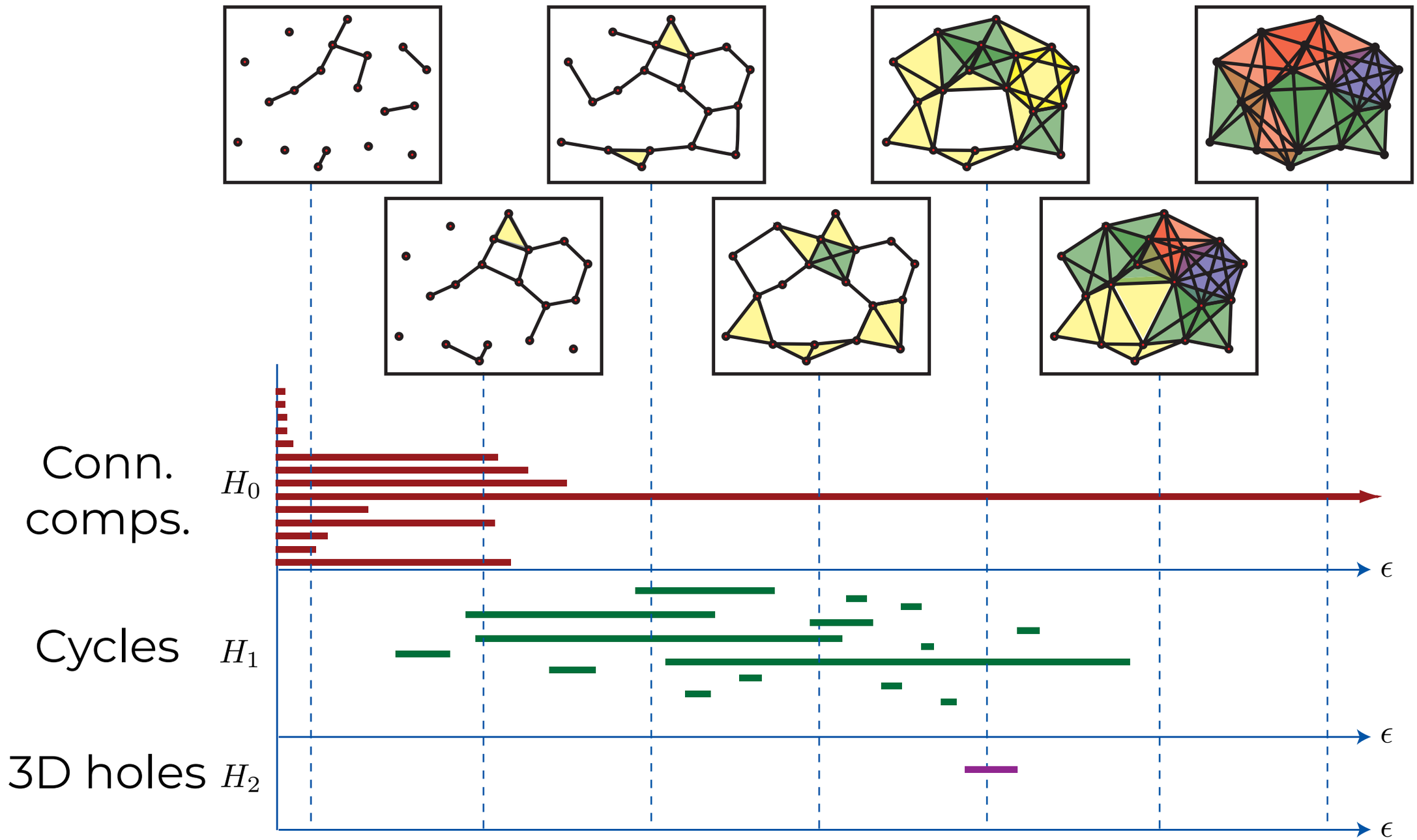
Persistent homology in two slides.. and a donut!



What is the appropriate **radius** for the balls?

Ghrist, R. (2008). Barcodes: The persistent topology of data. *Bulletin-American Mathematical Society*, *45*(1), 61.

Persistent homology: Barcode



Ghrist, R. (2008). Barcodes: The persistent topology of data. *Bulletin-American Mathematical Society*, *45*(1), 61.

Persistent homology in two slides.. and a donut!

The set of vector spaces and linear maps $\{H_n(X_t), H_n(i_t)\}_{t\in\mathbb{N}}$ can be represented as a finitely generated k[x]-module:

$$H_n = \bigoplus_{t\in\mathbb{N}} H_n(X_t) \qquad \text{with} \qquad \begin{aligned} \cdot x = H_n(X_t) &\to H_n(X_{t+1}) \\ m &\to H_n(i_t)(m) \end{aligned}$$

Structure theorem

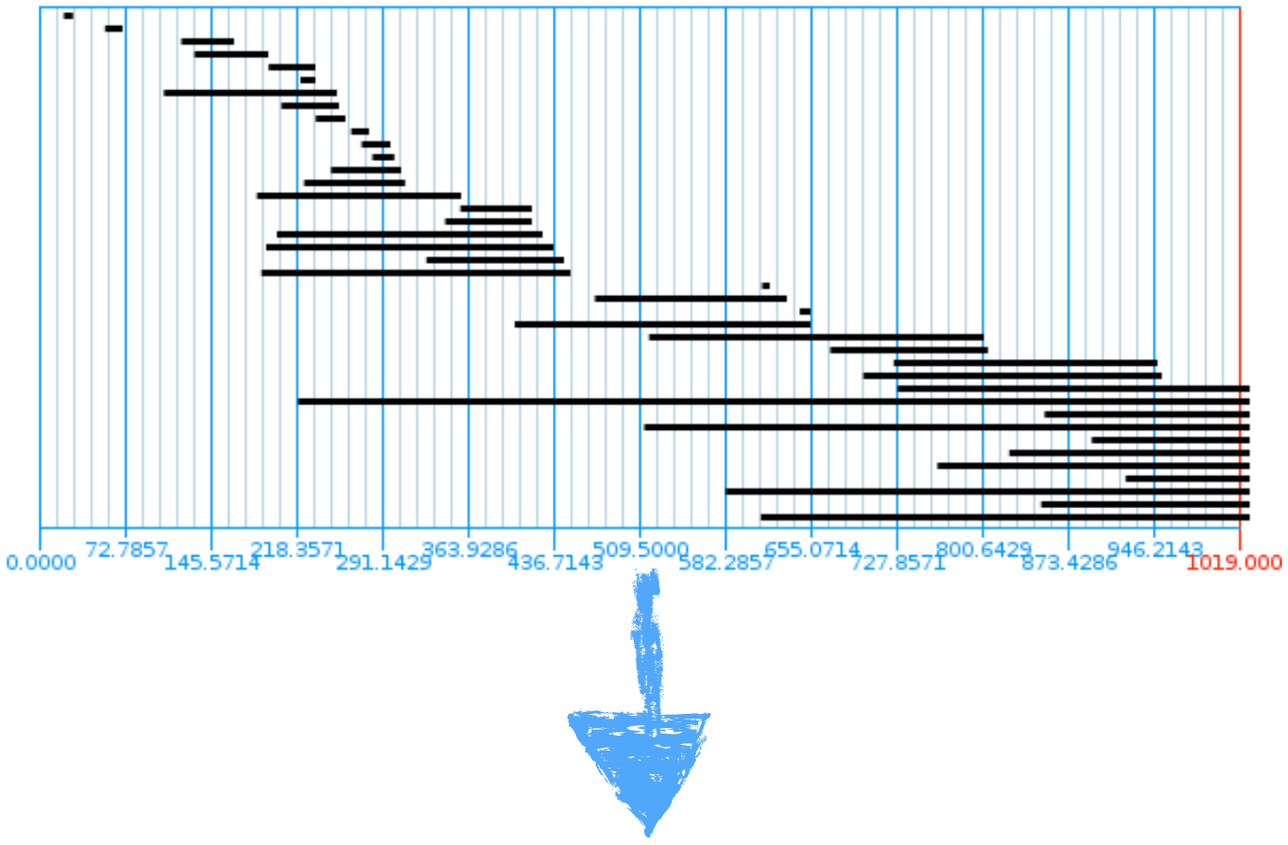
$$M \simeq \bigoplus_{j=1}^{n} k[x](-a_j) \bigoplus_{i=1}^{m} k[x](-c_i)/x_i^{d_i}$$

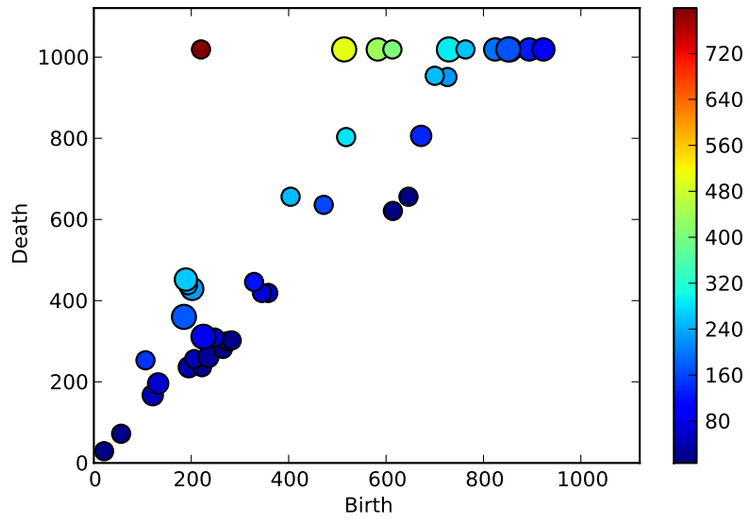
Remark

The degrees of the generators and the torsion degrees $\{d_i\}$ completely determine the module

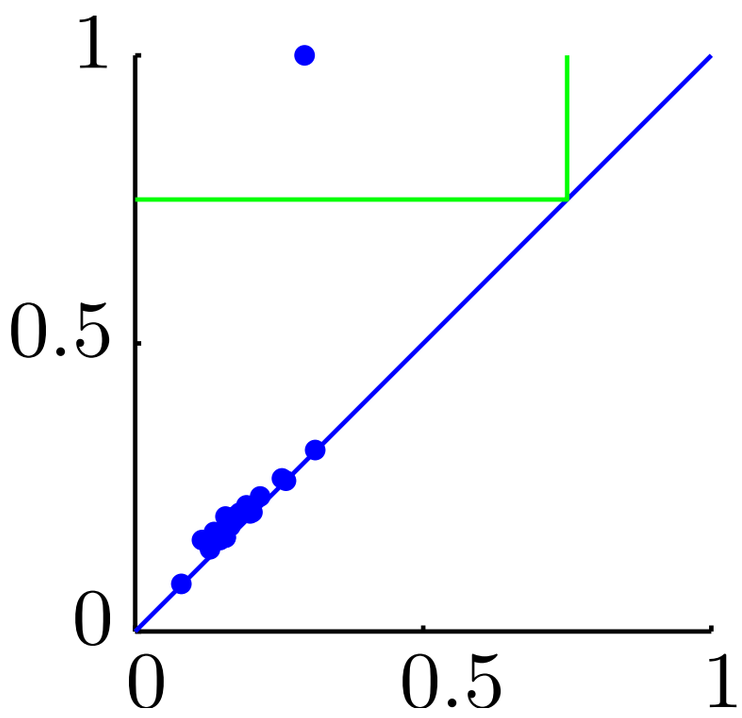
Barcode m finite intervals $[c_i, d_i)$ and n half intervals $[a_j, \infty)$

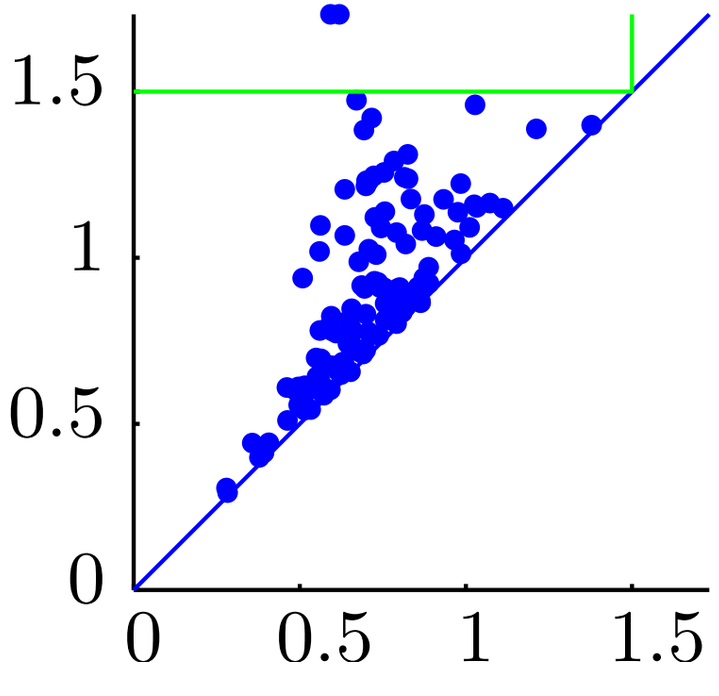
Barcode to Persistence Diagram

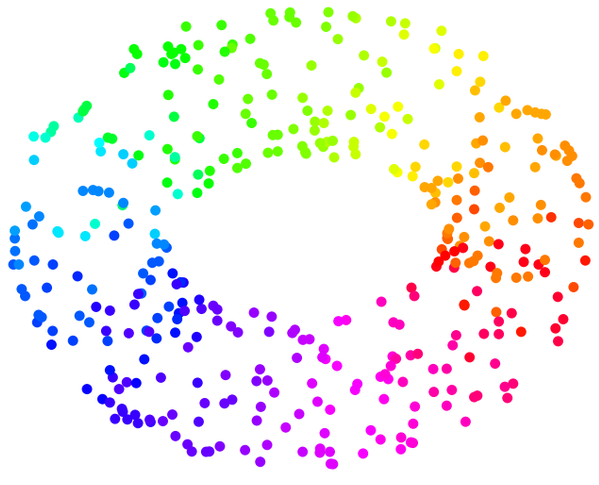


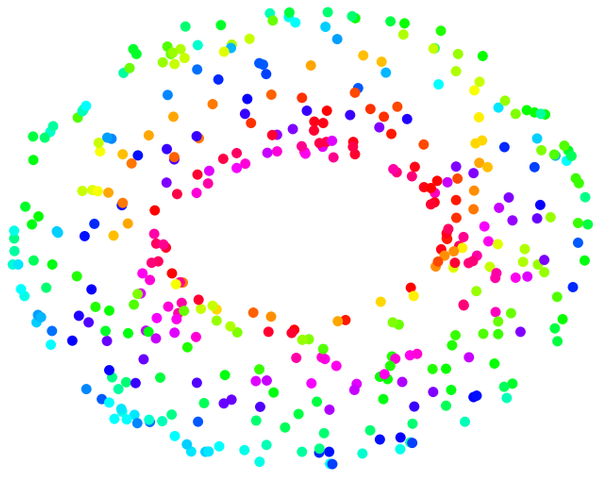


Examples



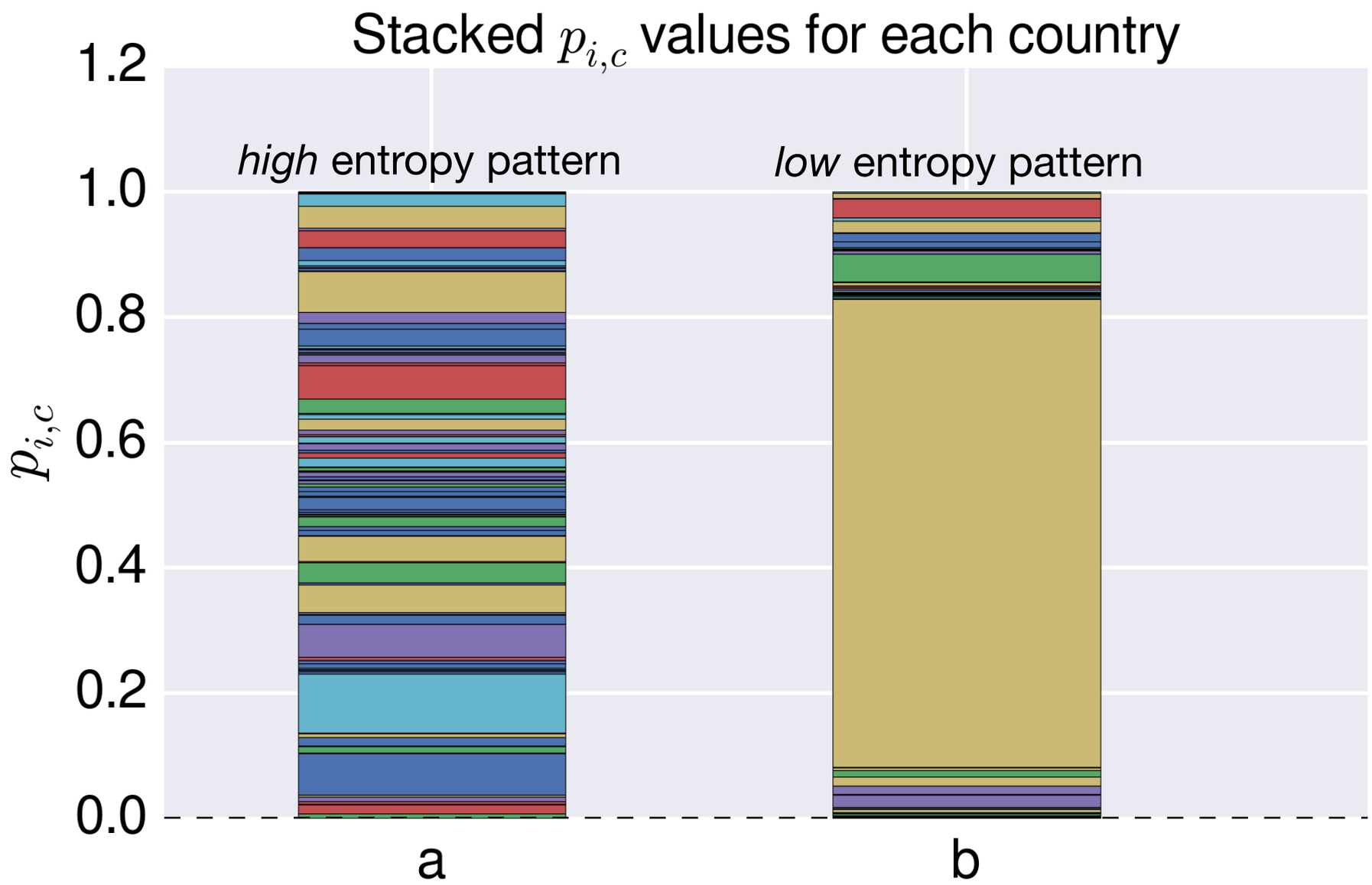






Notebook 04

Phom app: ENTROPY



$A_{i,c}(t)$ i = cell, c = country

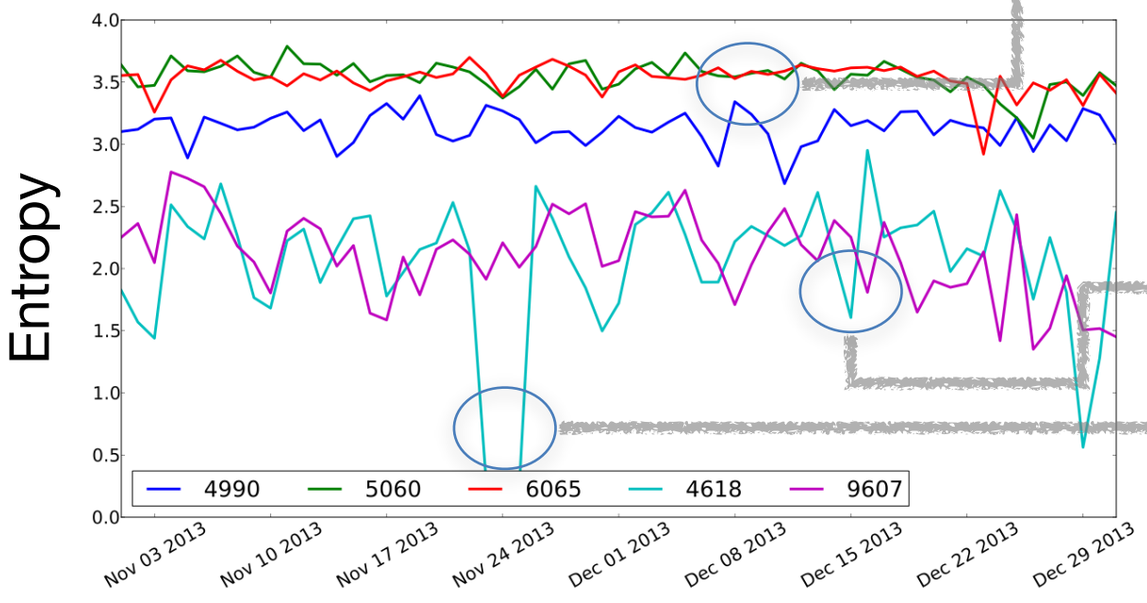
cell level $A_{i,c} = \sum_t A_{i,c}(t)$

T-slices $A_{i,c}(T) = \sum_{t \in T} A_{i,c}(t)$

district level $A_{D,c} = \sum_t \sum_{i \in D} A_{i,c}(t)$

$$\begin{cases} p_{i,c} = A_{i,c} / \sum_c A_{i,c} \\ e_i = \sum_c -p_{i,c}\, log(p_{i,c}) \end{cases}$$

ENTROPY: single cells



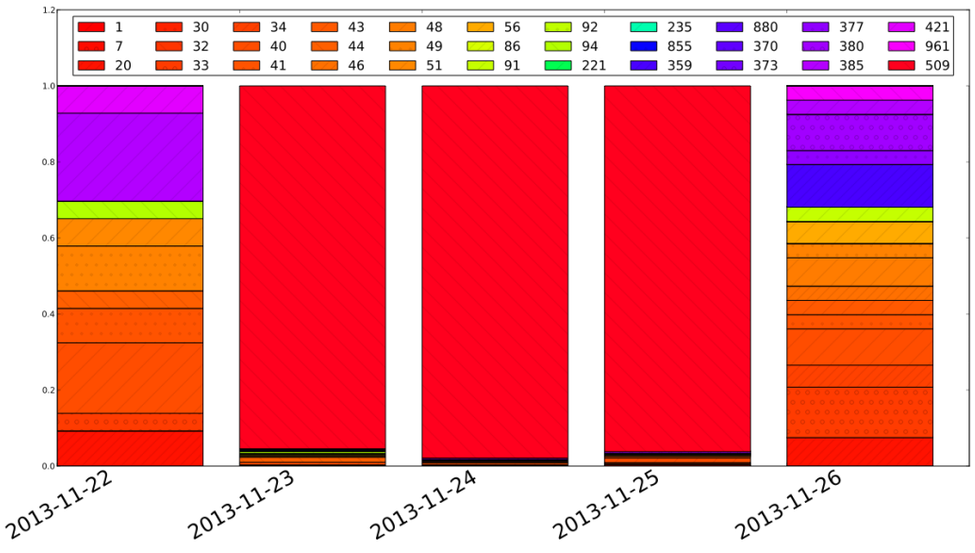
High steady entropy areas (Milano's Dome, Central railway station, Linate airport, Milano's stock exchange, etc..)

Low entropy areas with weekly patterns (University, peripheral areas, etc..)

The cell displays very low entropy and a large Haitian activity.

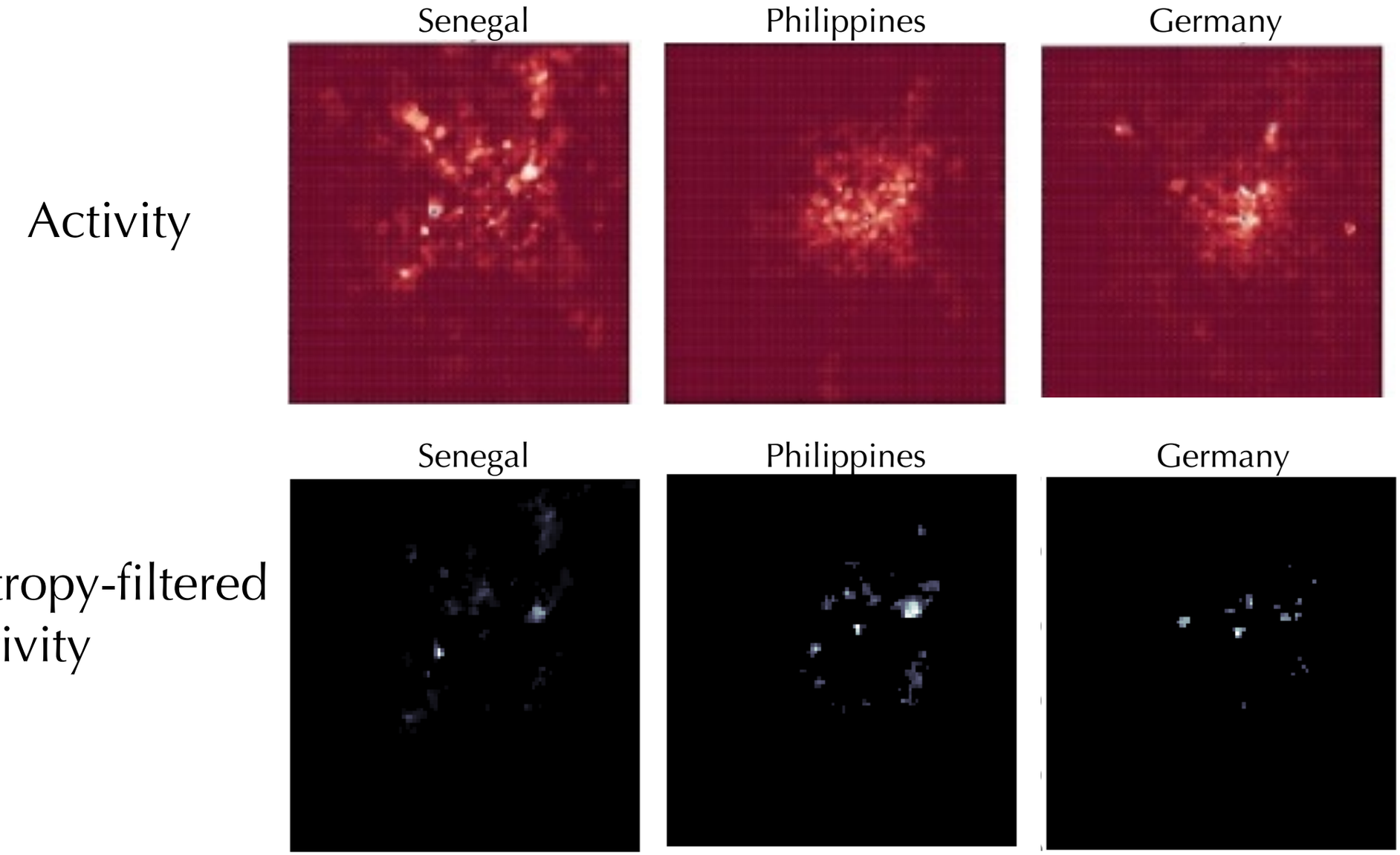
- Haitian Movie Festival Award?
- Big wedding?

Activity stack chart for cell 4618



ENTROPY: country maps

Country-specific patterns should correspond with low entropy and large country-activity.

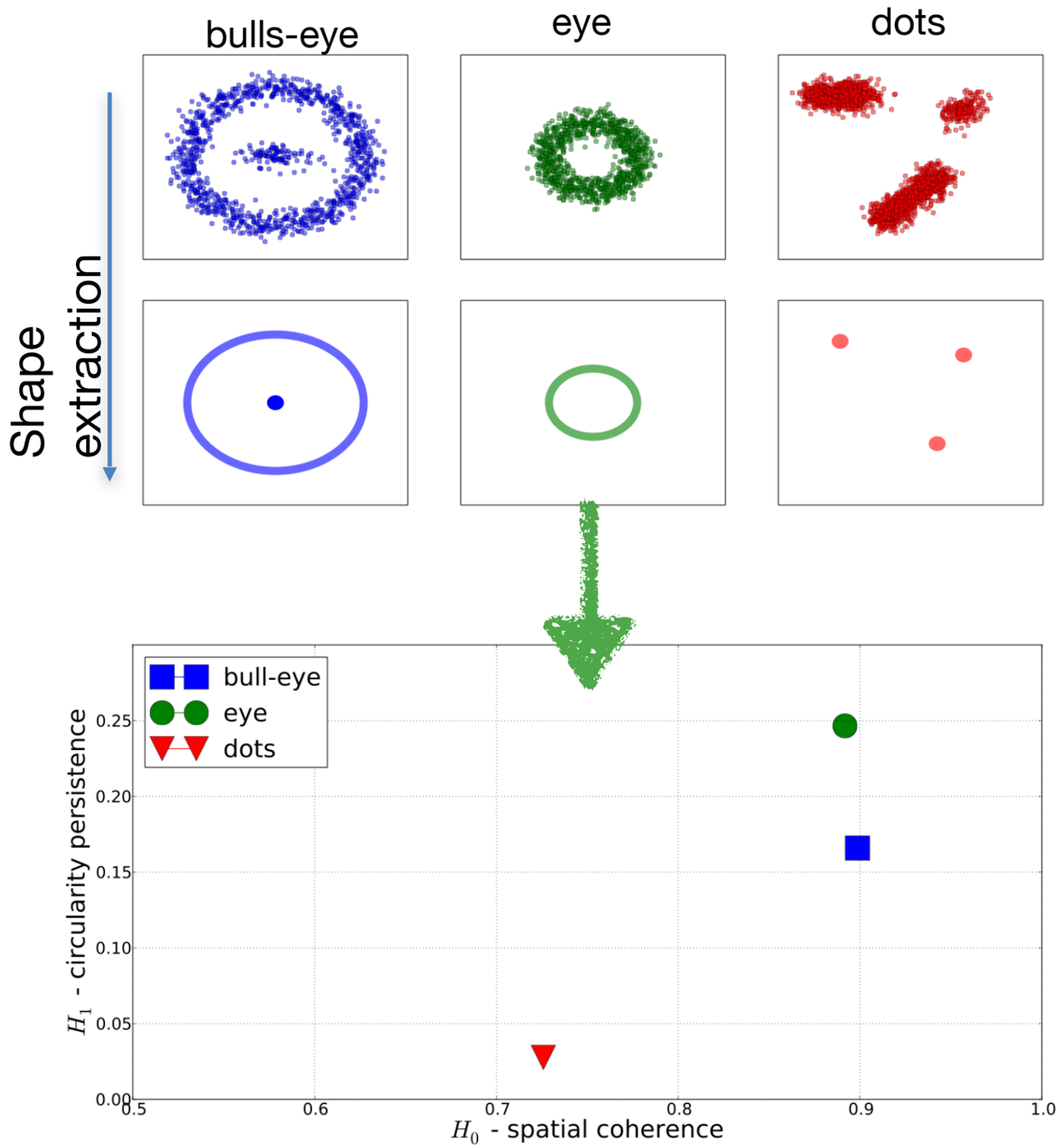


ENTROPY: country maps

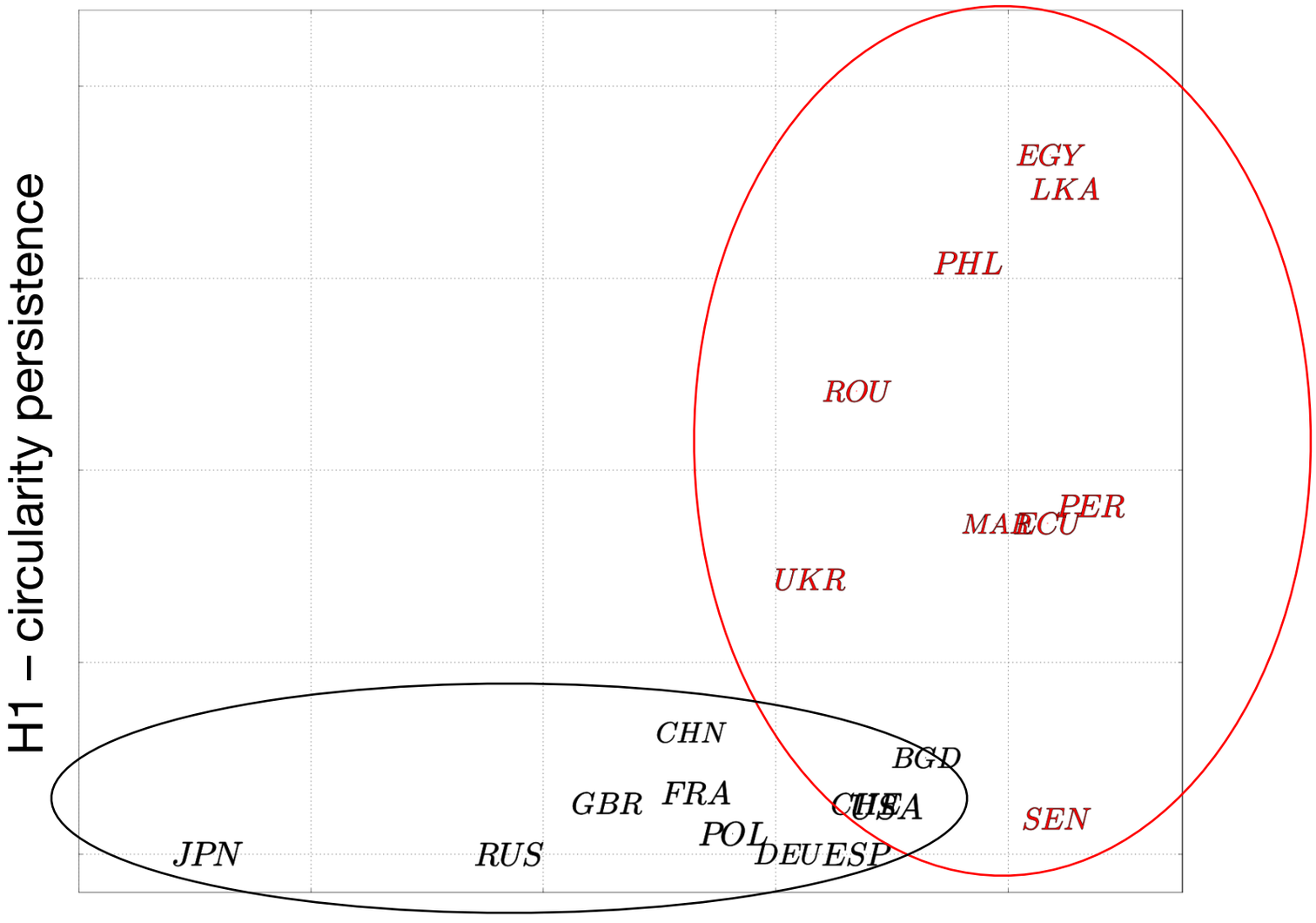
Persistent homology: classifying spatial patterns of international communities

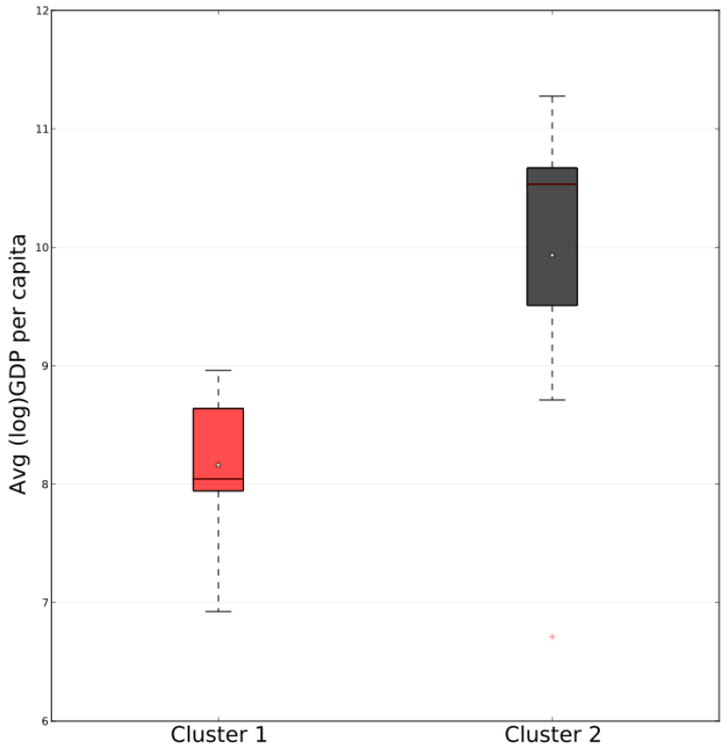
1. Recurring noisy patterns:
    1. disconnected clusters
    2. rings surrounding the city center

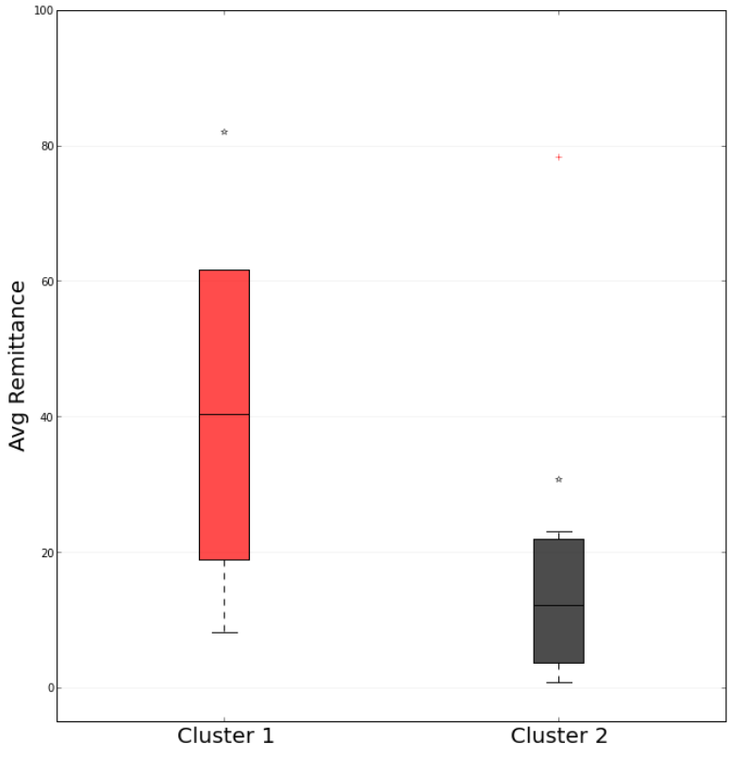
2. indicator based on their topological features
    1. robust to stretching and noise
    2. multiscale shape recognition method (Ghrist, Bull.AMS 45 (2008)).
    3. focus on the spatial coherence of country clusters:
        1. **connected components** (H0),
        2. **circularity features** (H1)



Persistent homology: entropy-activity maps

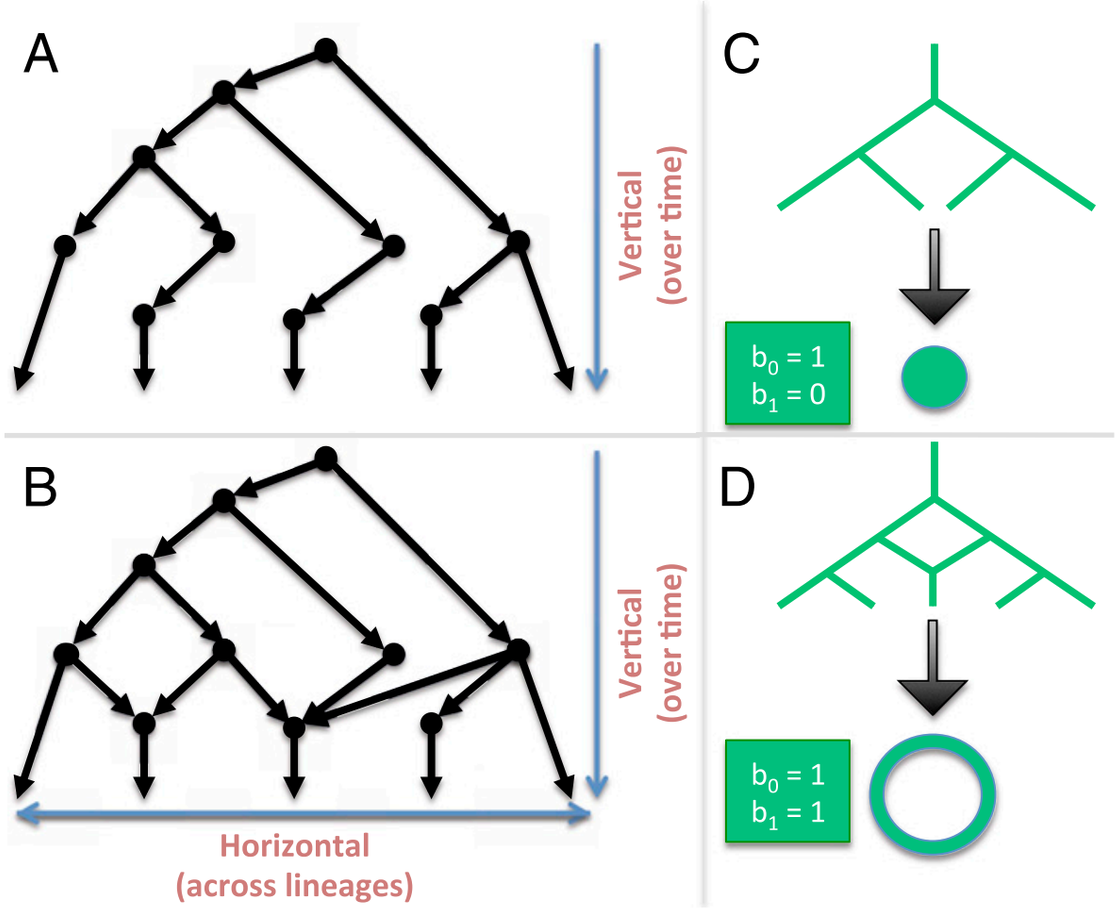


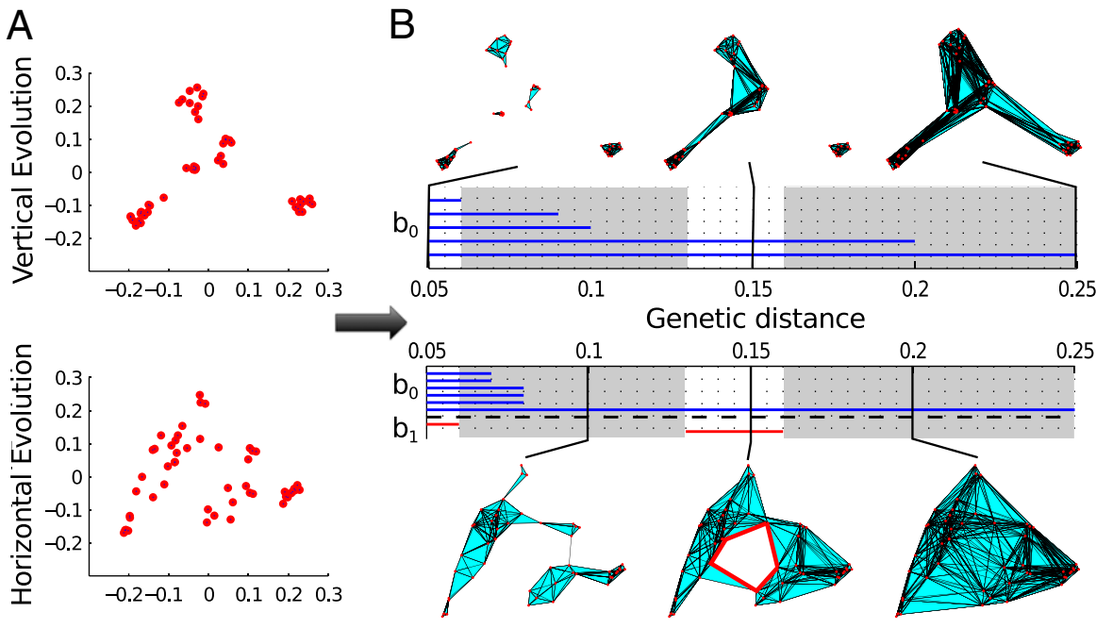


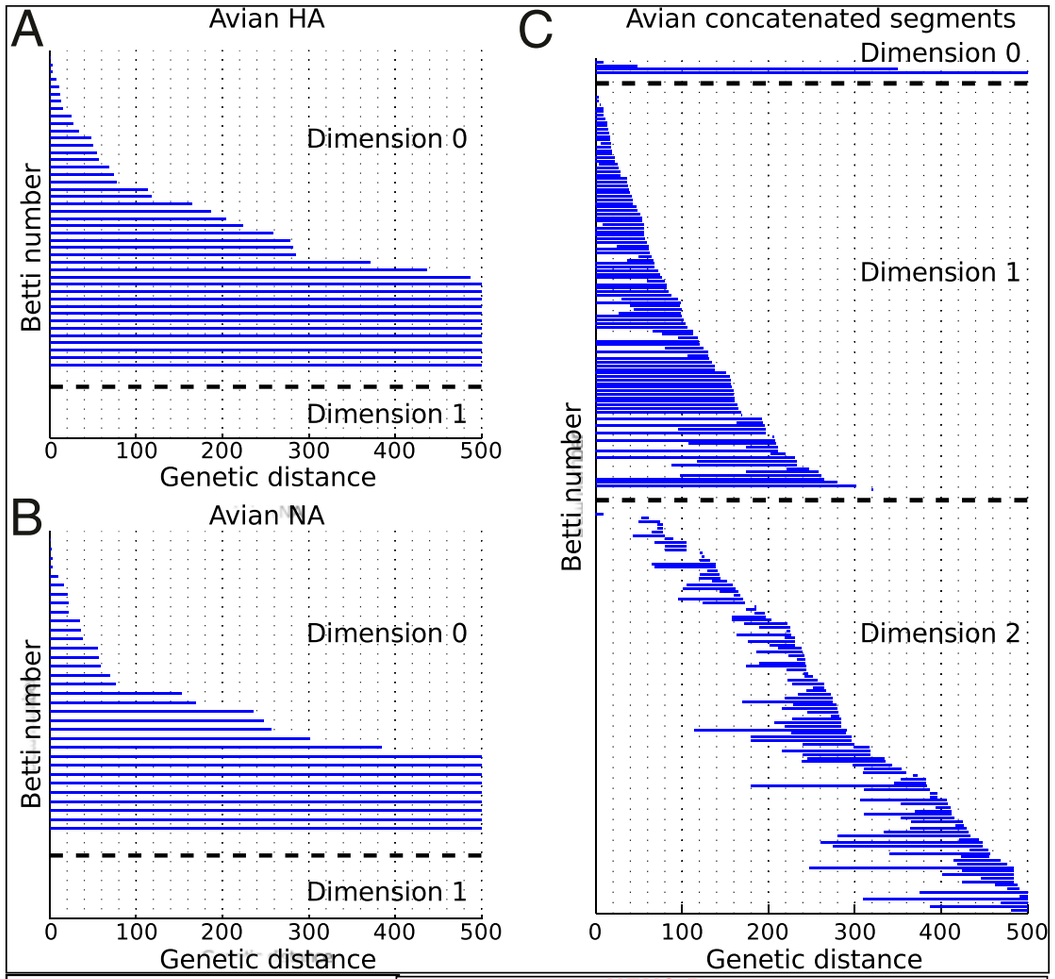


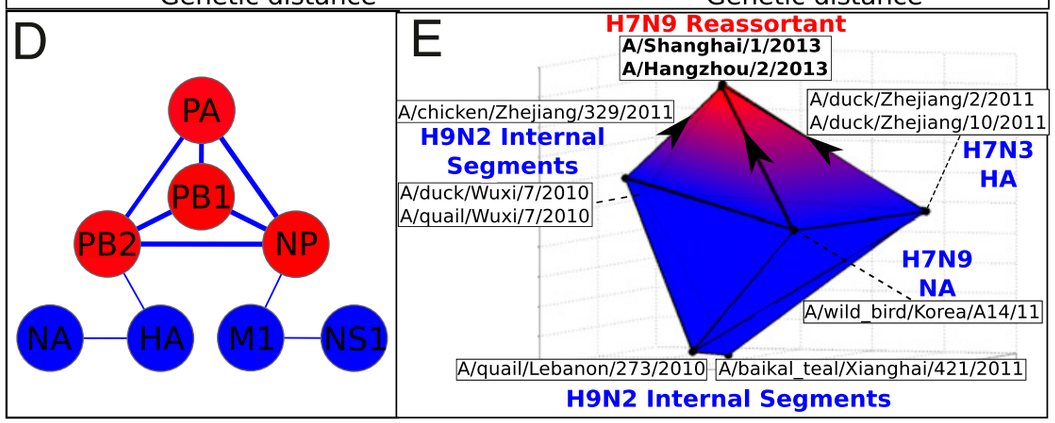
1. Clusters in 2D-scatter plot are simple k-means.
2. GDP per capita and remittance of countries in the two clusters clearly differentiate between richer migrant communities with respect people from developing countries.

Previous work: phylogenetic trees









Topology of viral evolution

Joseph Minhow Chan[a,b], Gunnar Carlsson[c], and Raul Rabadan[a,b,d,1]

[a]Center for Computational Biology and Bioinformatics and Departments of [b]Biomedical Inform
Physicians and Surgeons, New York, NY 10032; and [c]Department of Mathematics, Stanford Uni

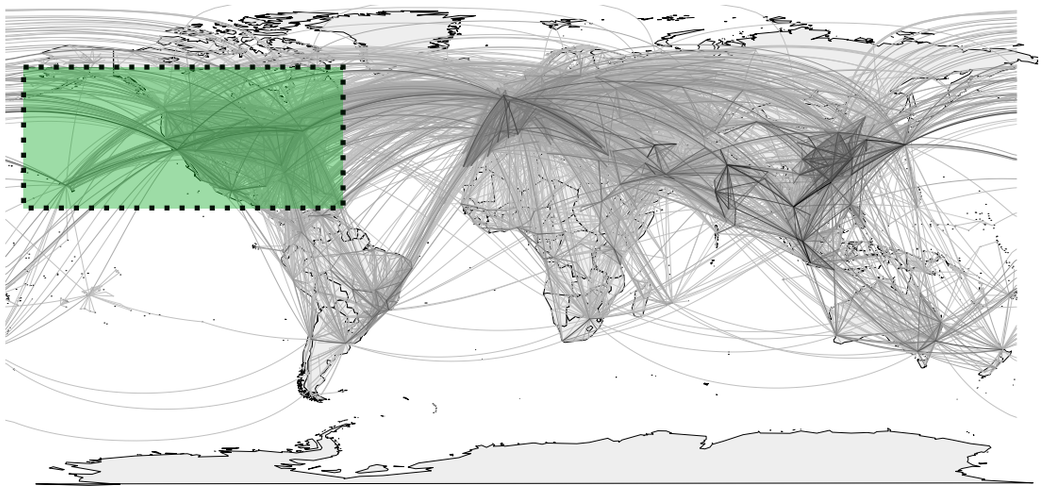
Edited* by Arnold J. Levine, Institute for Advanced Study, Princeton, NJ, and approved Octobe

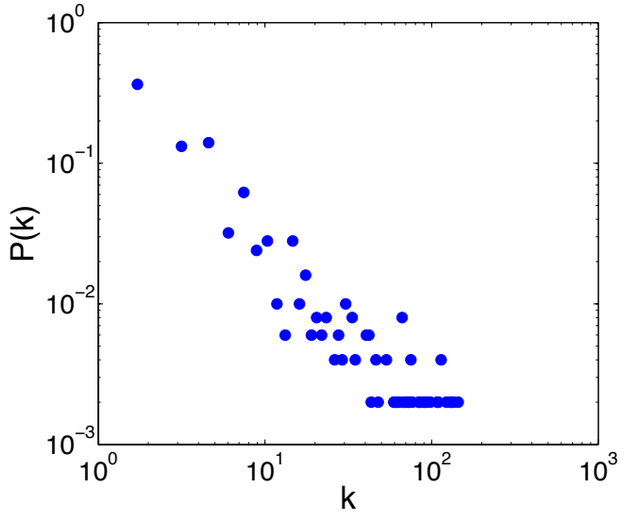
What about networks?

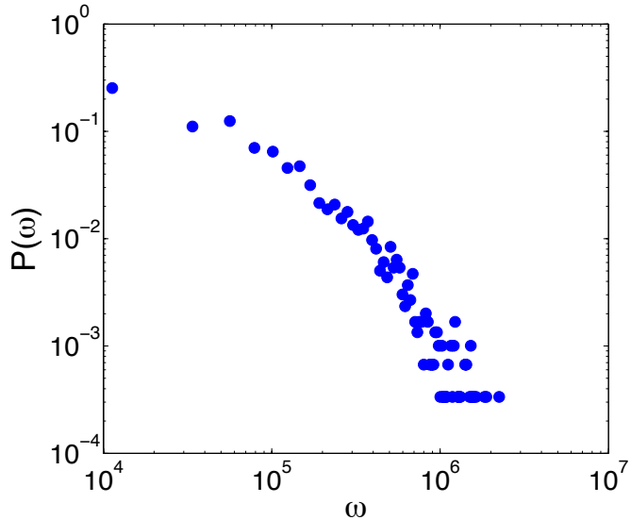
Metrical persistent homology

No metric space, but one can define a few: shortest path,
commute time
other Kernel distances..

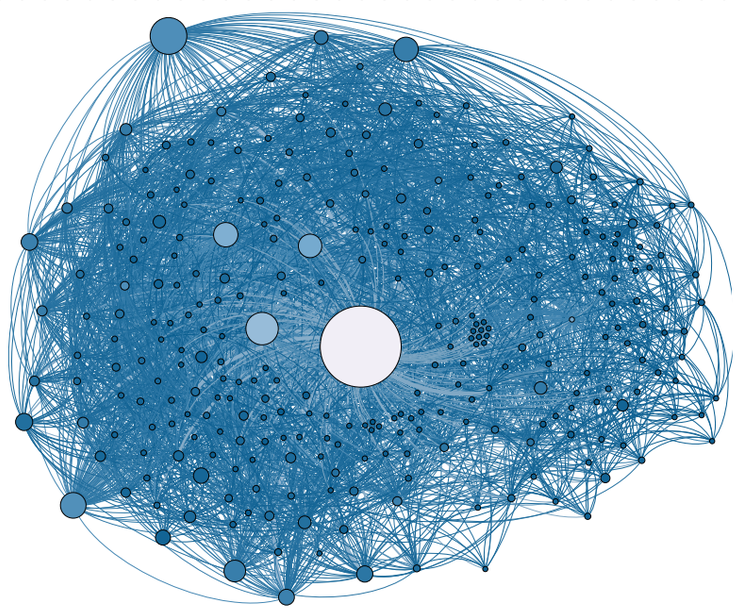
US Top-500 airports network

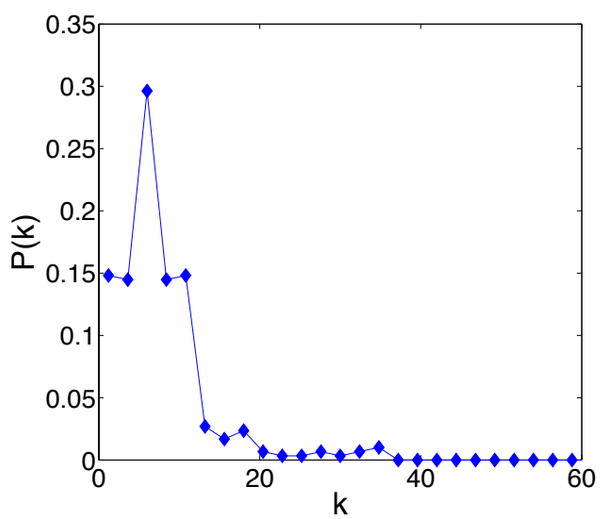


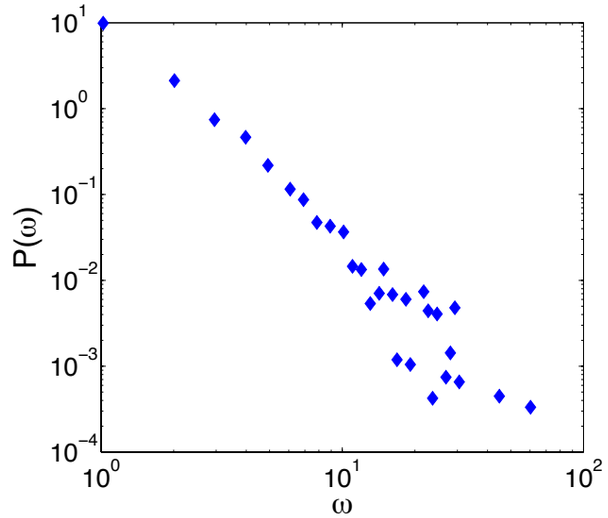




C. Elegans neural network

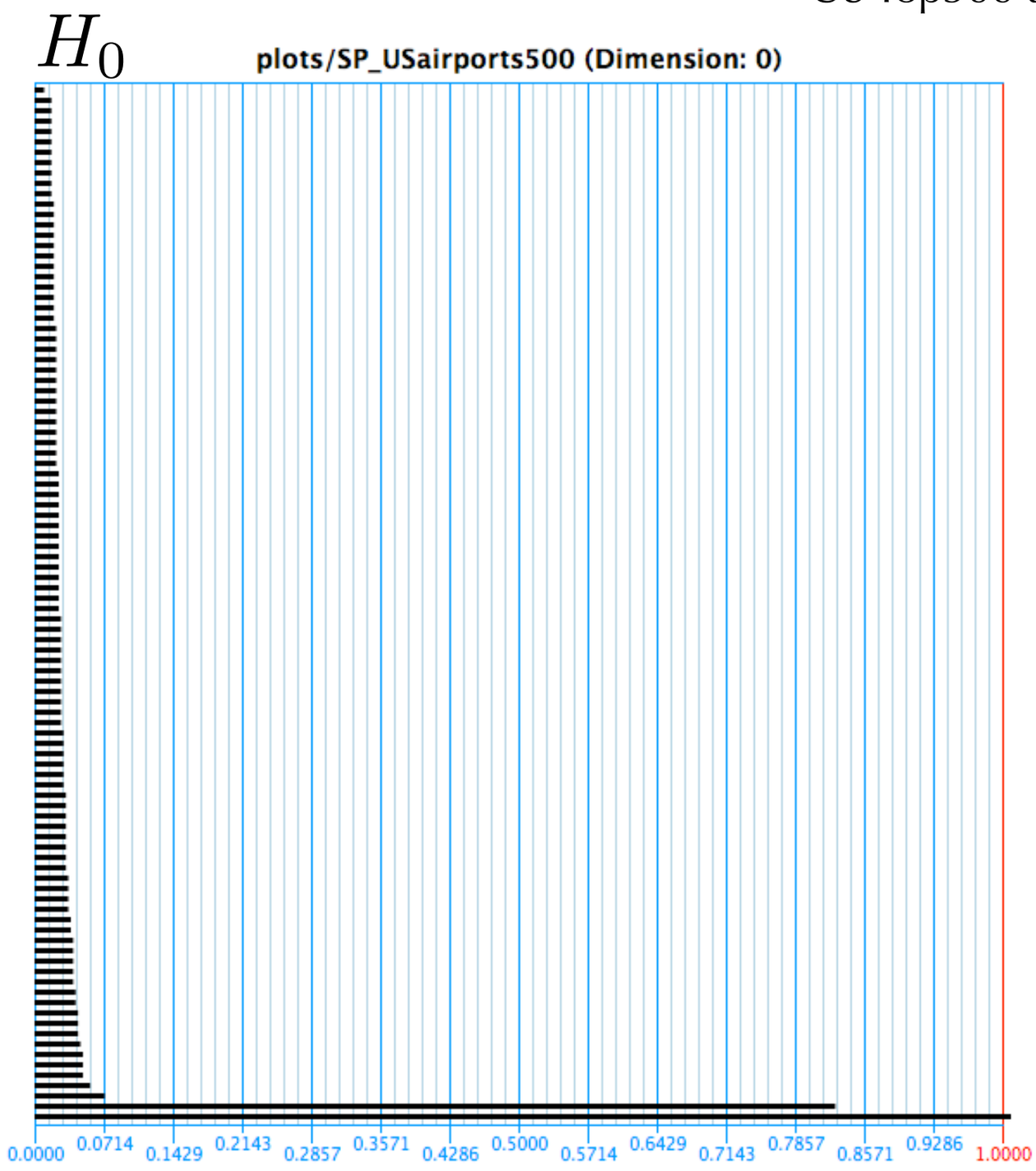




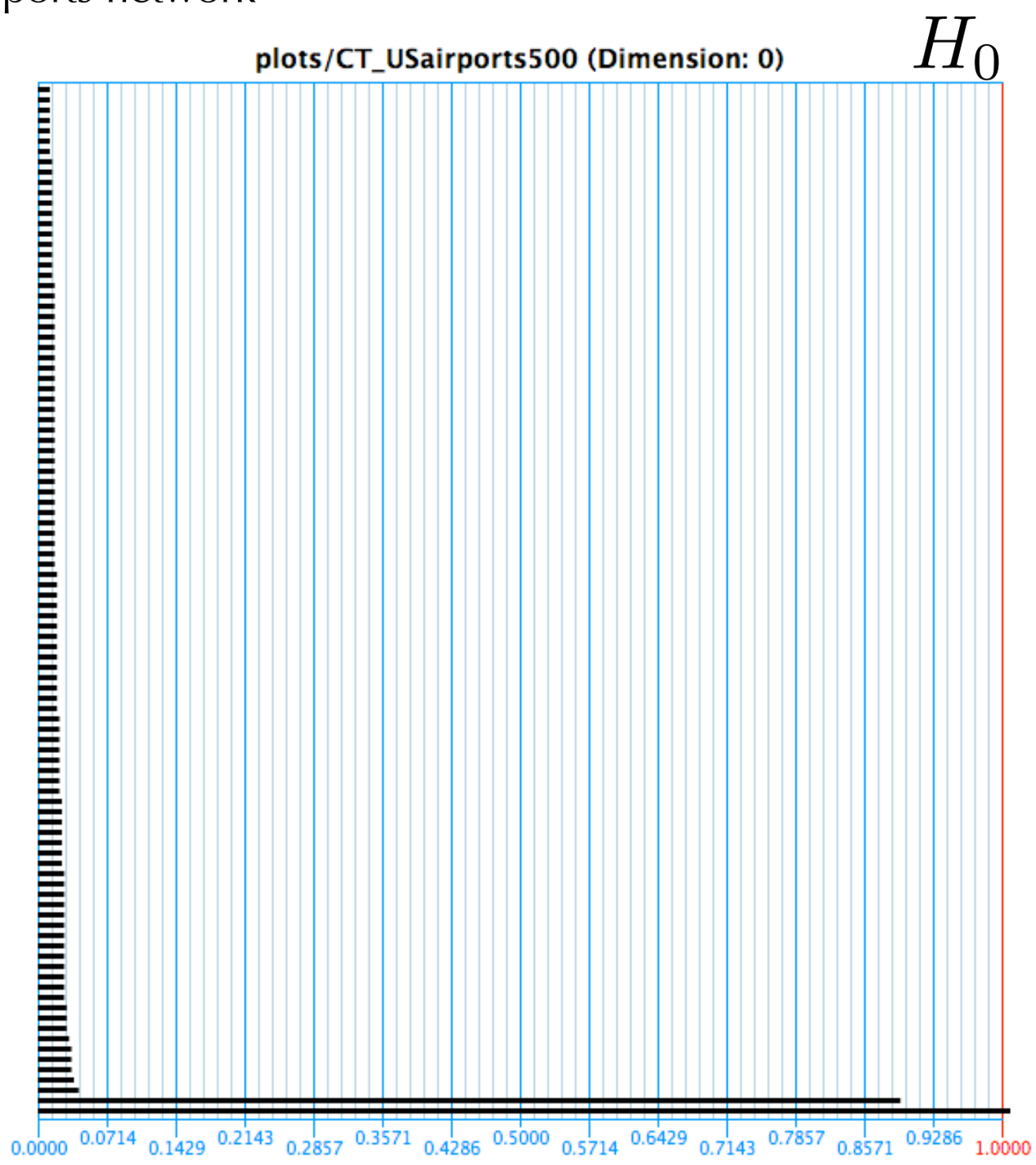


Metrical persistent homology II

US Top500 airports network



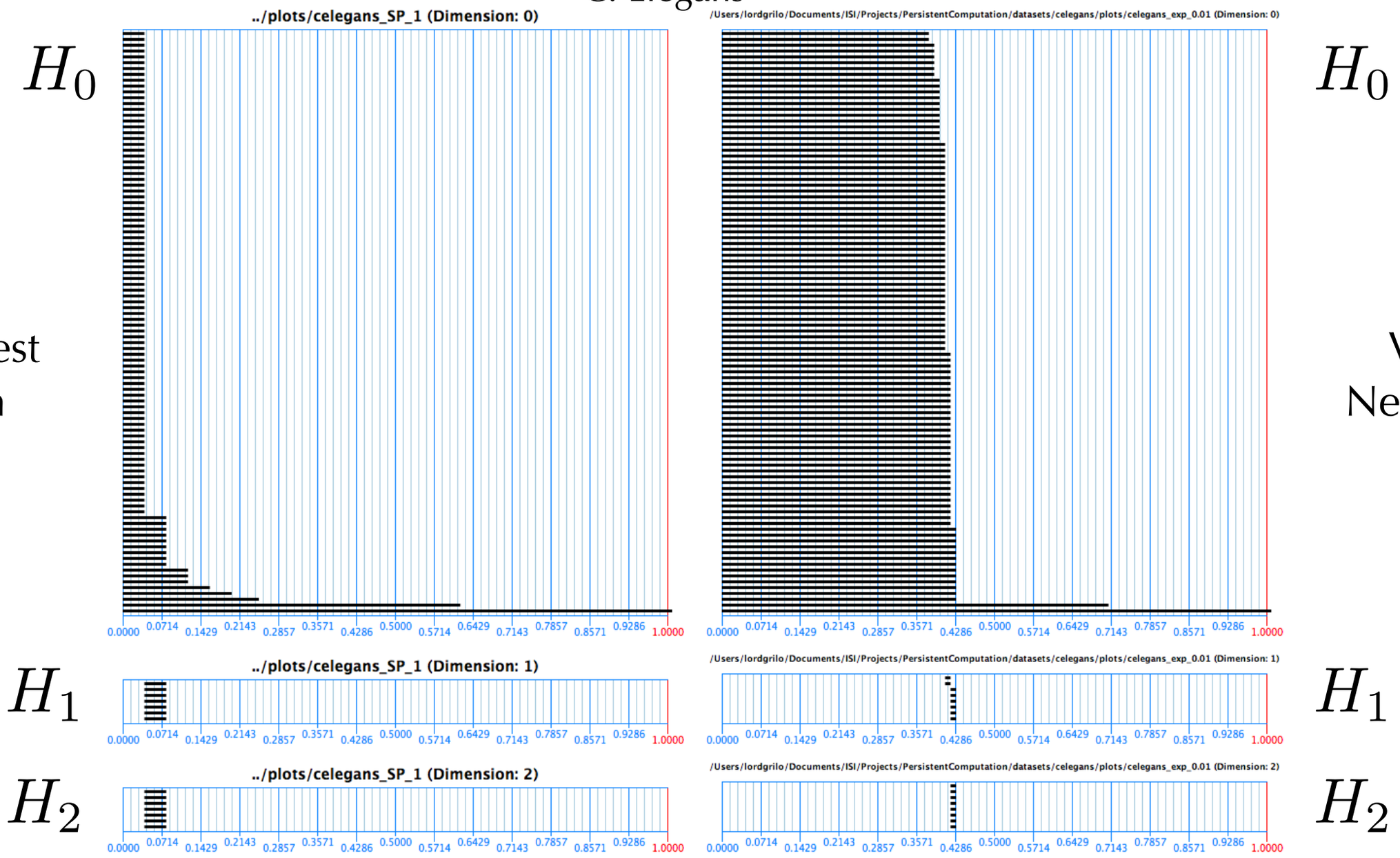
Shortest path



Commute time

Metrical persistent homology II$_{bis}$

C. Elegans



H_0 H_0

Shortest path

Von Neuman

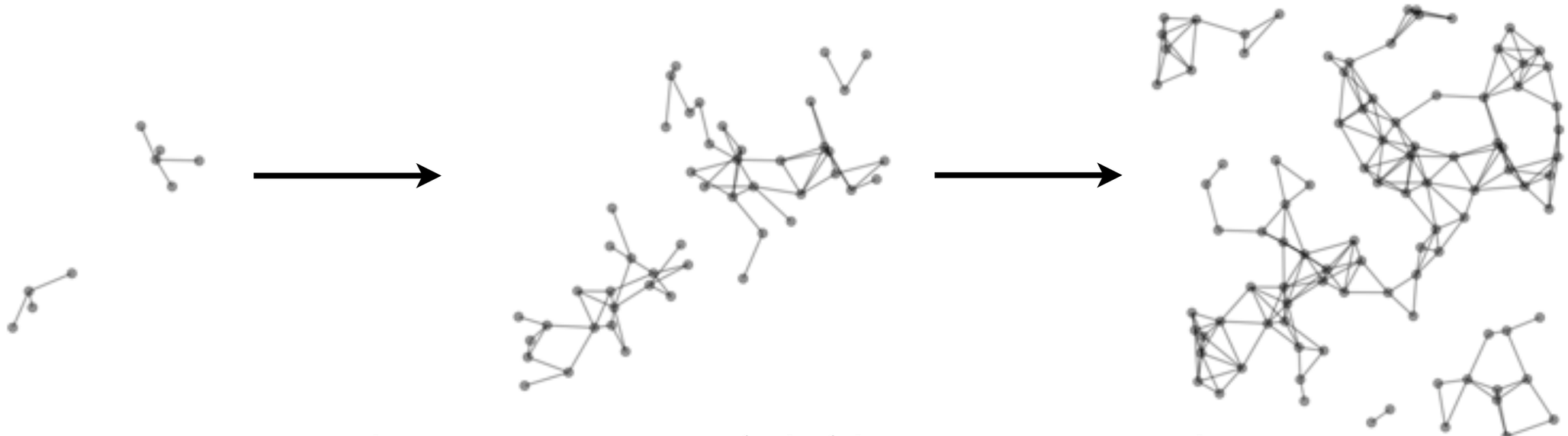
H_1 H_1

H_2 H_2

1. Different metrics ~ same information
2. Cycles appear at emergence of GCC
3. No need of homology to do this!!

Metrical persistent homology III

Metrics do not convey much information.
What now? The important ingredient is the **filtration**.



Are there meaningful filtrations combining:

Network "linking patterns"

&

Weight structure

?

Remark

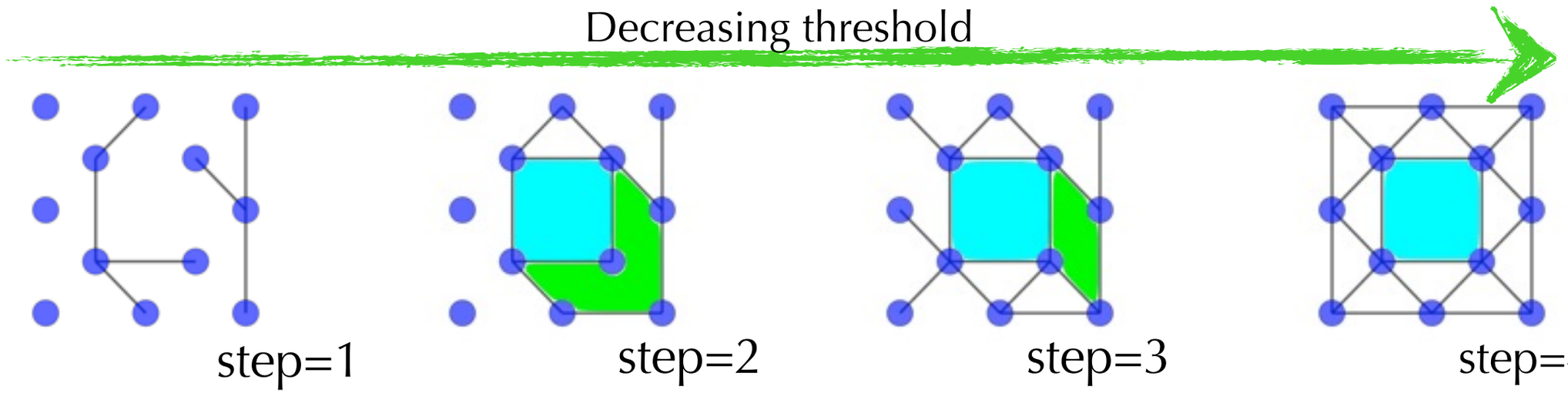
The Rips complex of points in a metric space is the clique complex of a metric graph.

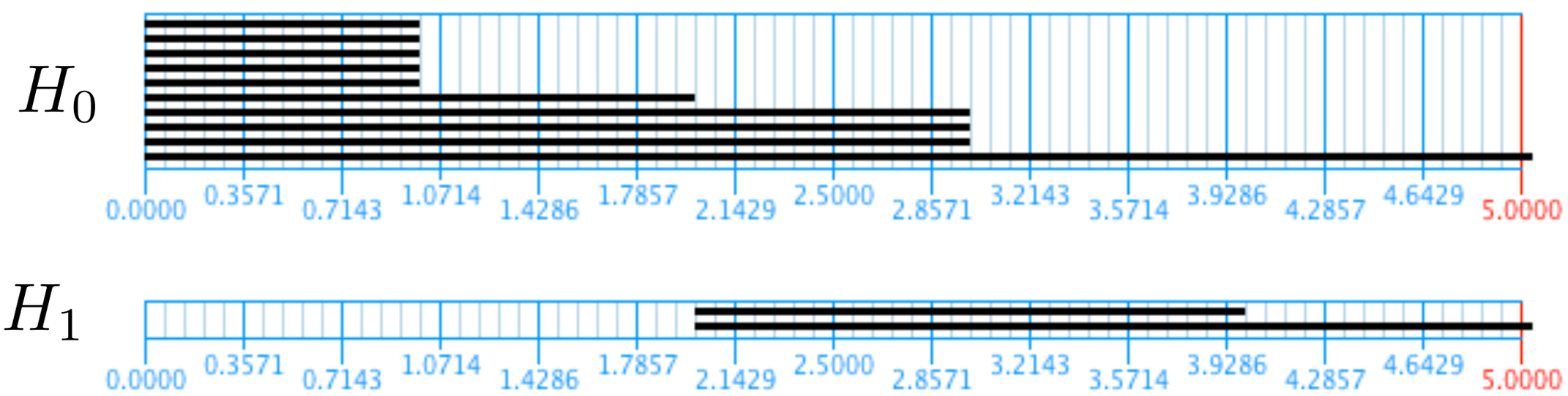
Weighted clique filtration I

1) Given a real network G=(V,E),
   consider another one with the same number of node N=|V|
   and no edges.

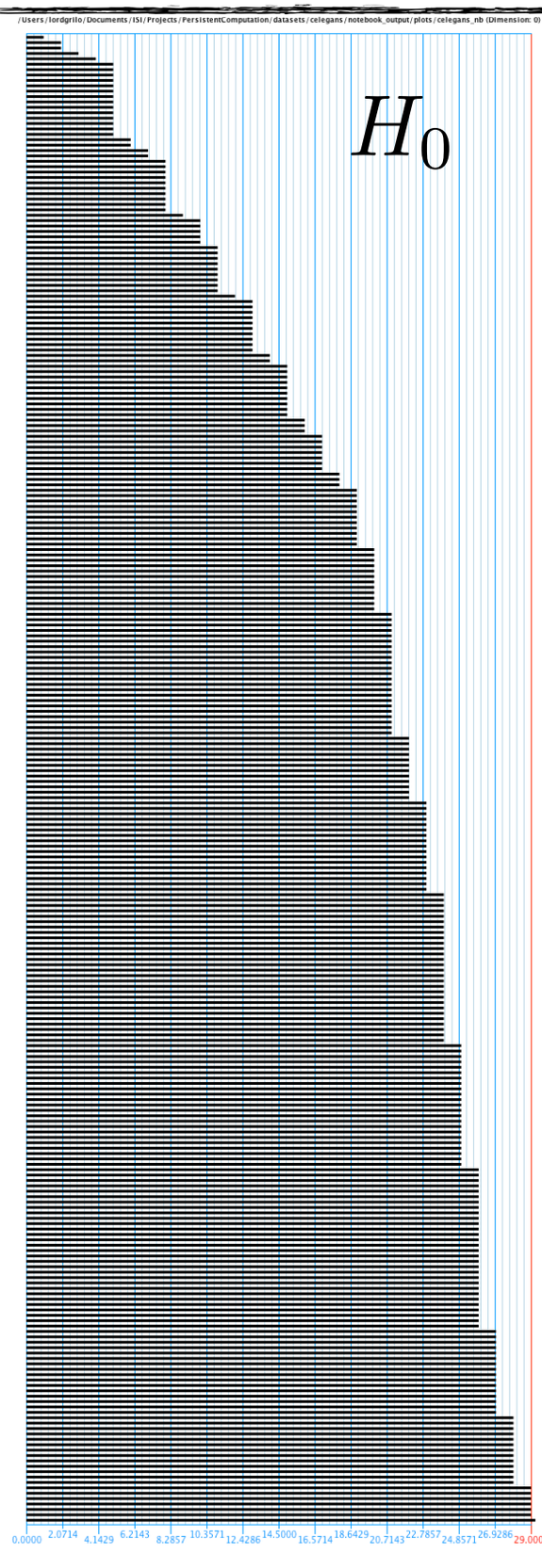
2) Start adding edges from G, in decreasing order of weight.

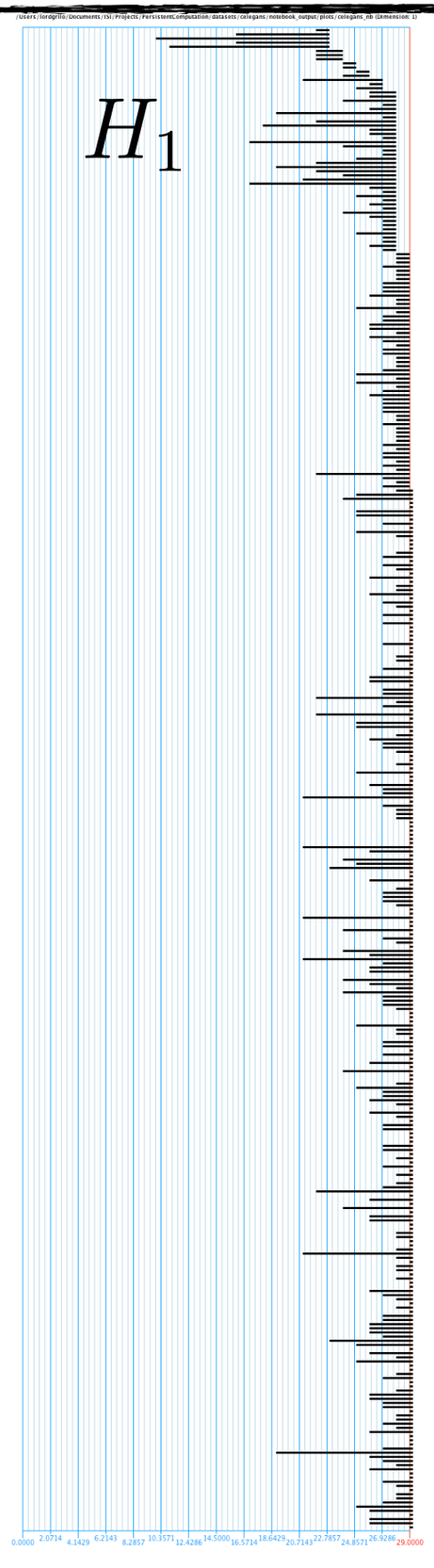
3) At each step, calculate the associated clique complex.

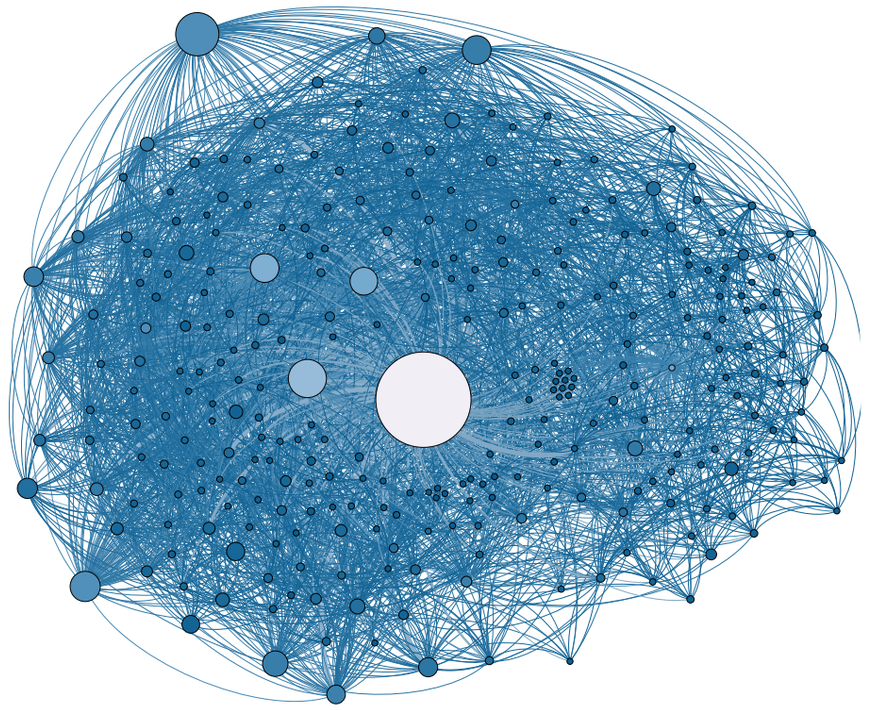




Weighted clique filtration II$_{\text{bis}}$

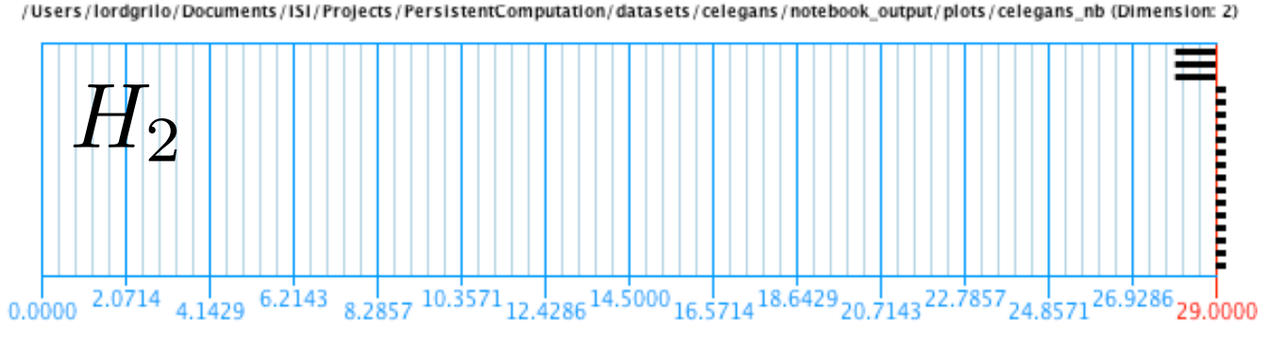




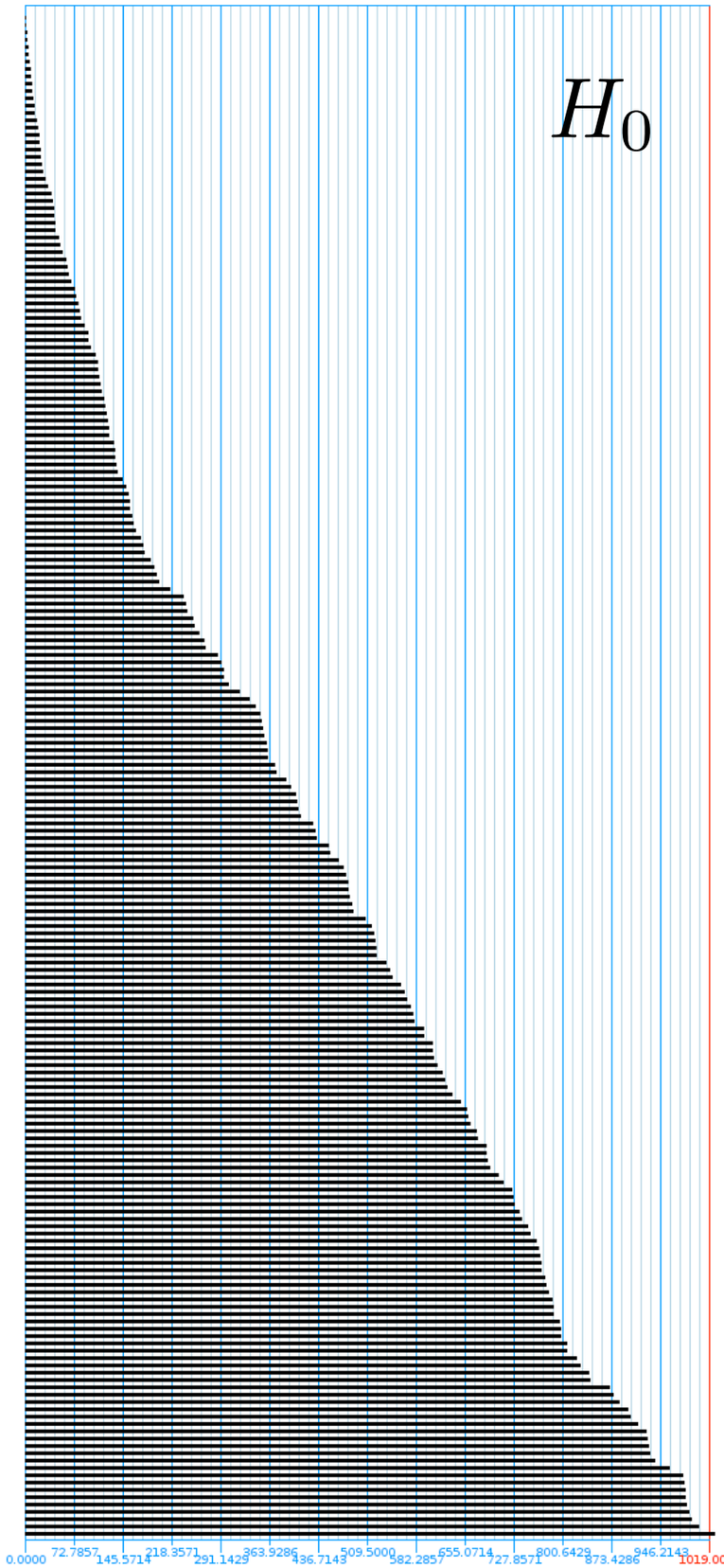


Much more structure appears!

- Conn. comps. smoothly coalesce.
- A large number of loops.
- Even some “3D”-holes for low weights.



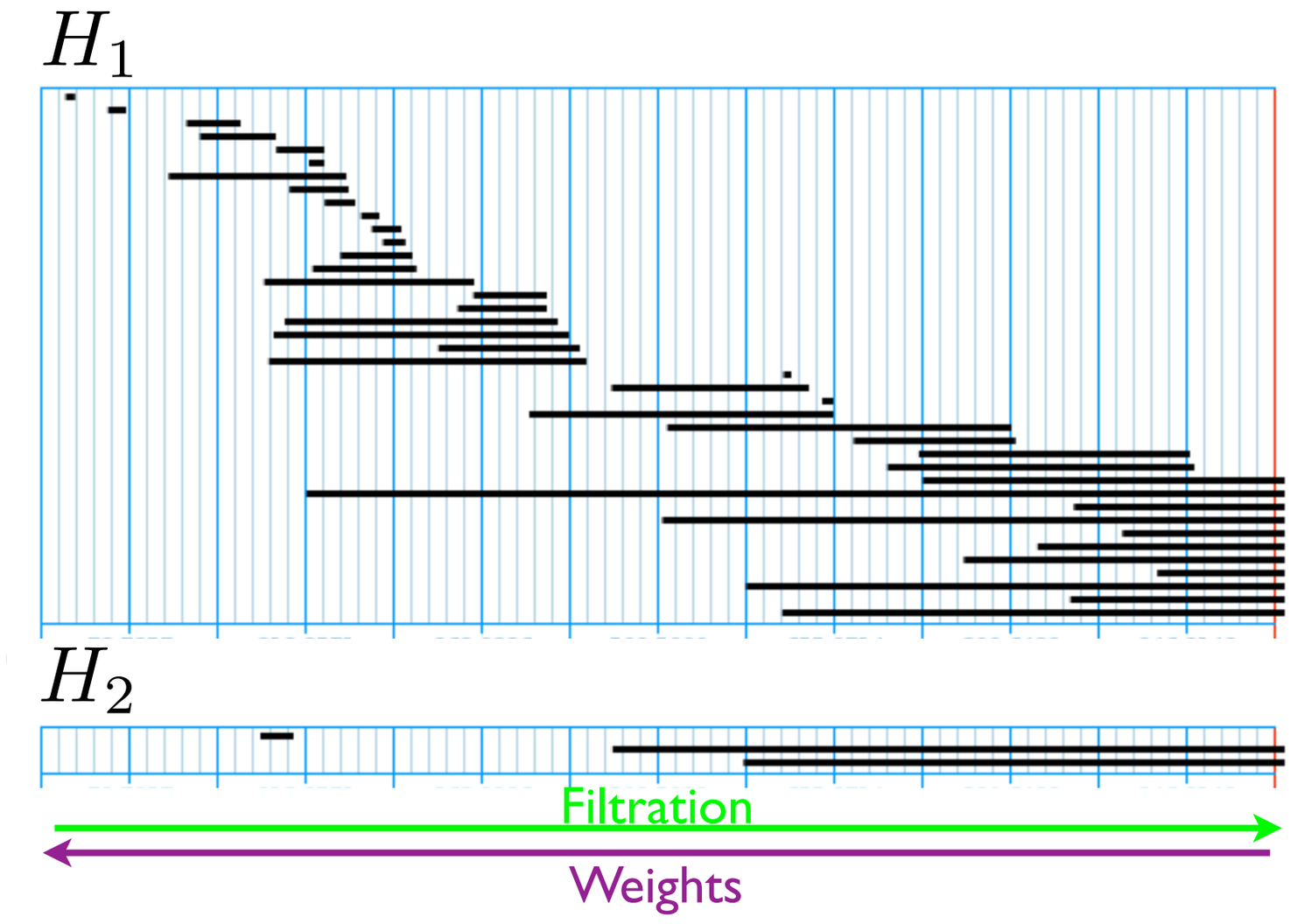
Weighted clique filtration II



H_0 Connected components

H_1 1D Cycles (2-simplexes)

H_2 3D holes (3-simplexes)

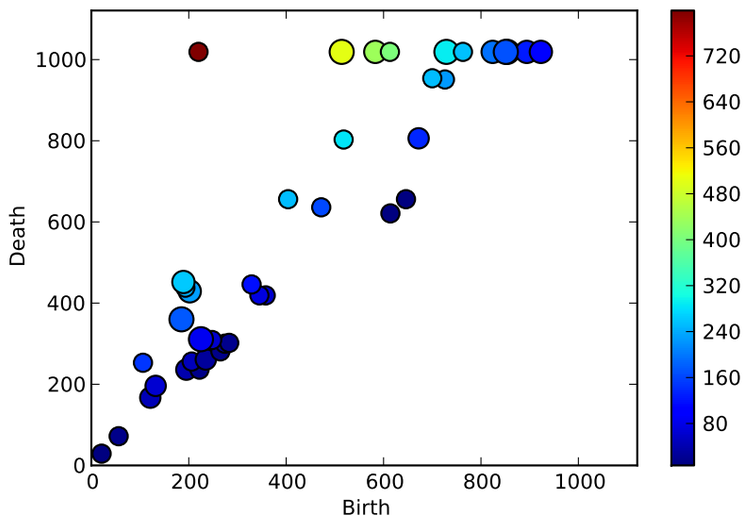


Datasets I

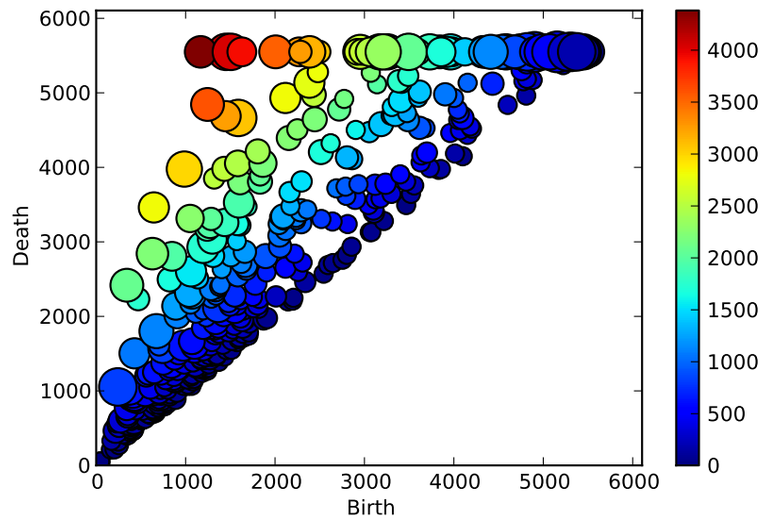
Focus on H_1, map each generator to a point in the plane:

Cycle → (Birth,Death)

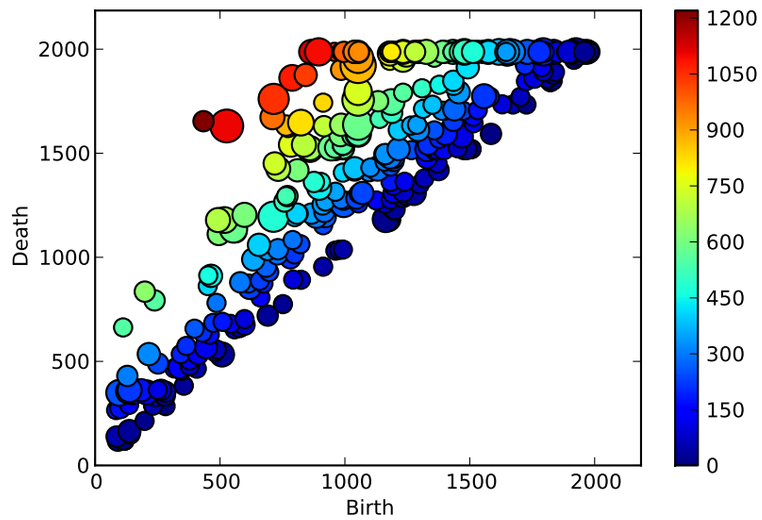
US2000 air passenger network



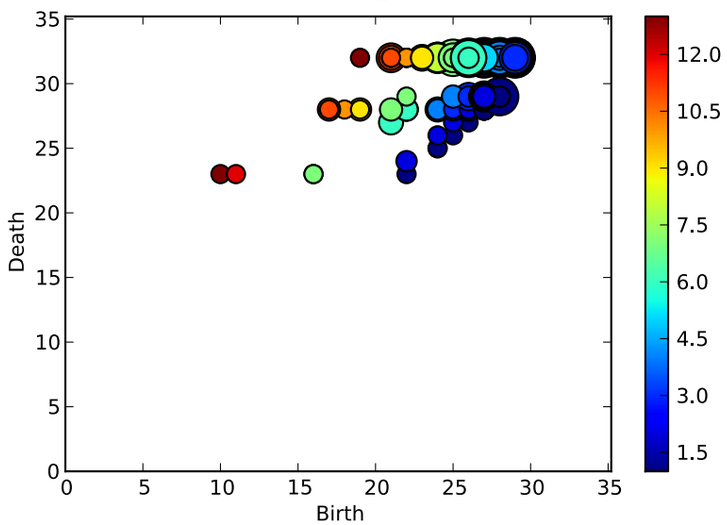
Human gene (sample)



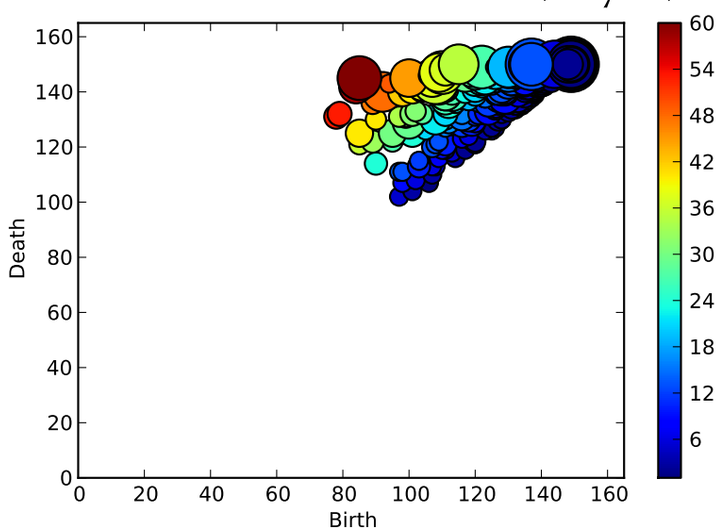
FB-like social network



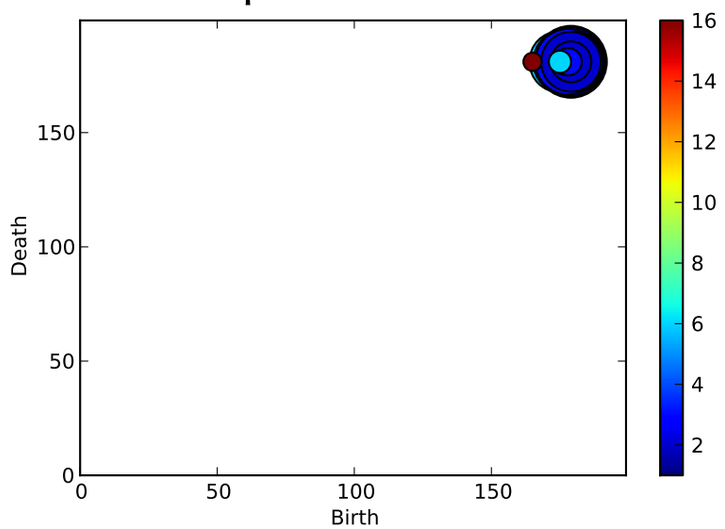
C. elegans



Kids contact duration (day II)



Kumpula et al. model



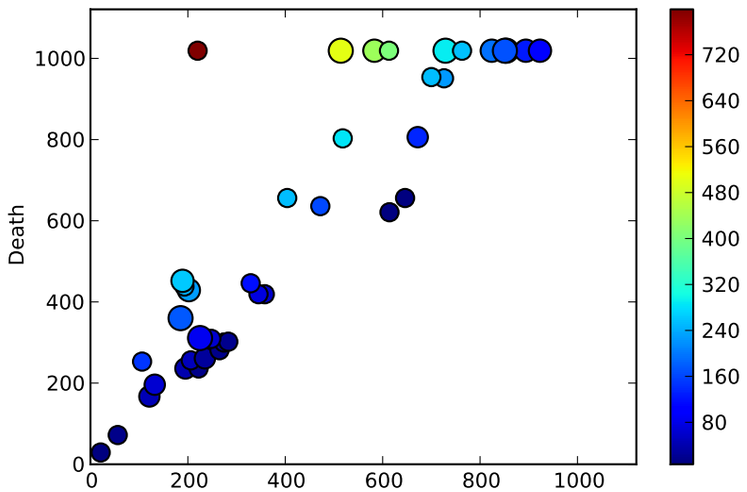
H_1

What is significant?

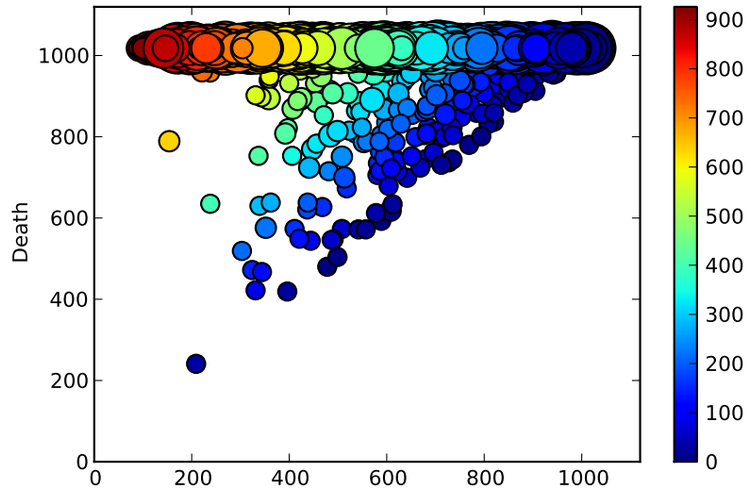
Datasets II: null models?

US 2000 air passenger network

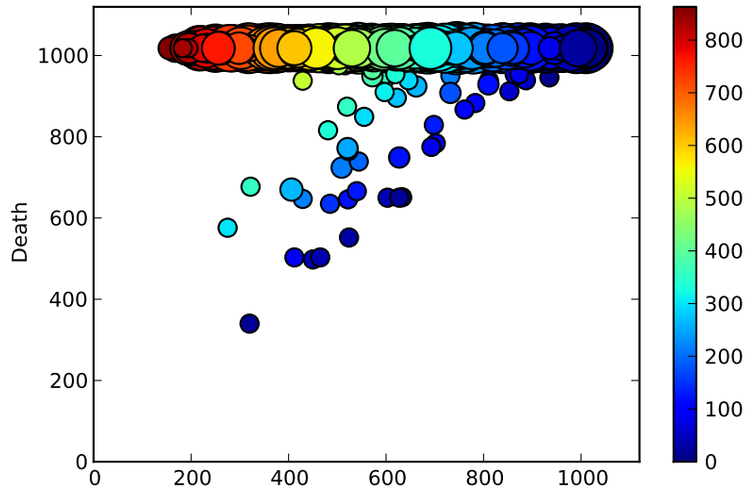
Data



Weight shuffled

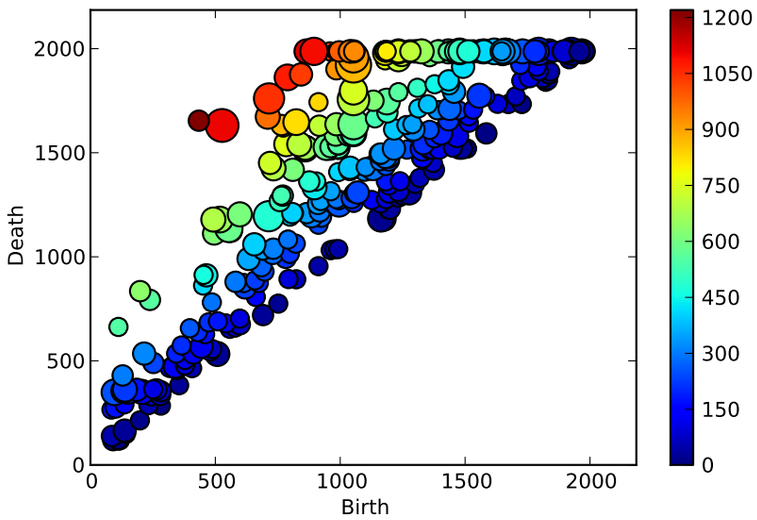


Degree-preserving

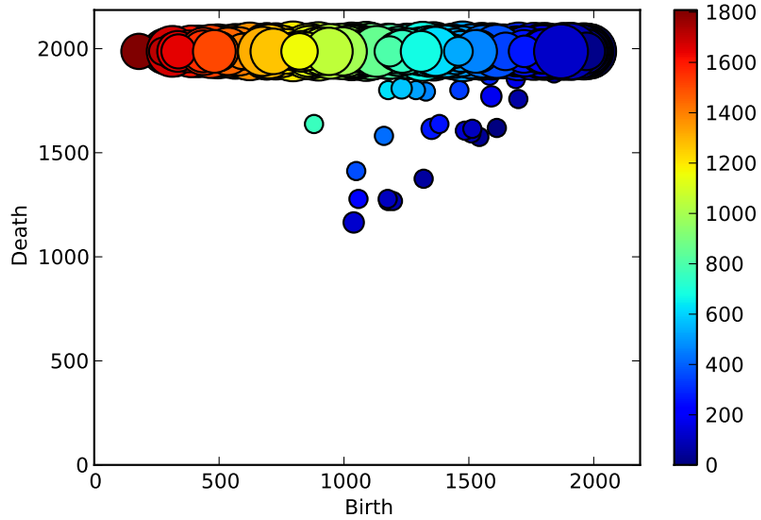


FB-like social network

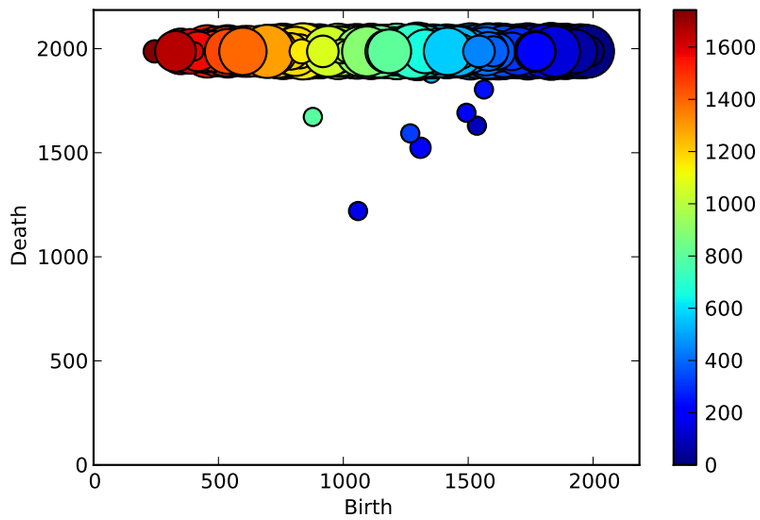
Data



Weight shuffled



Degree-preserving

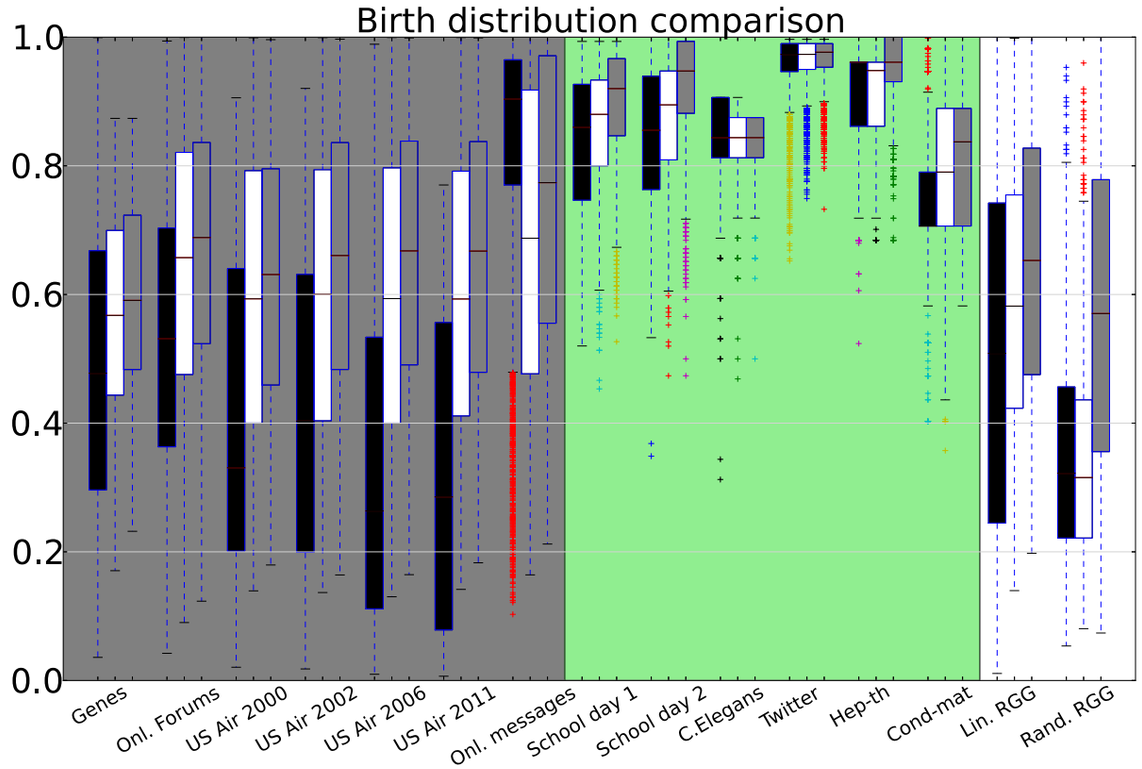


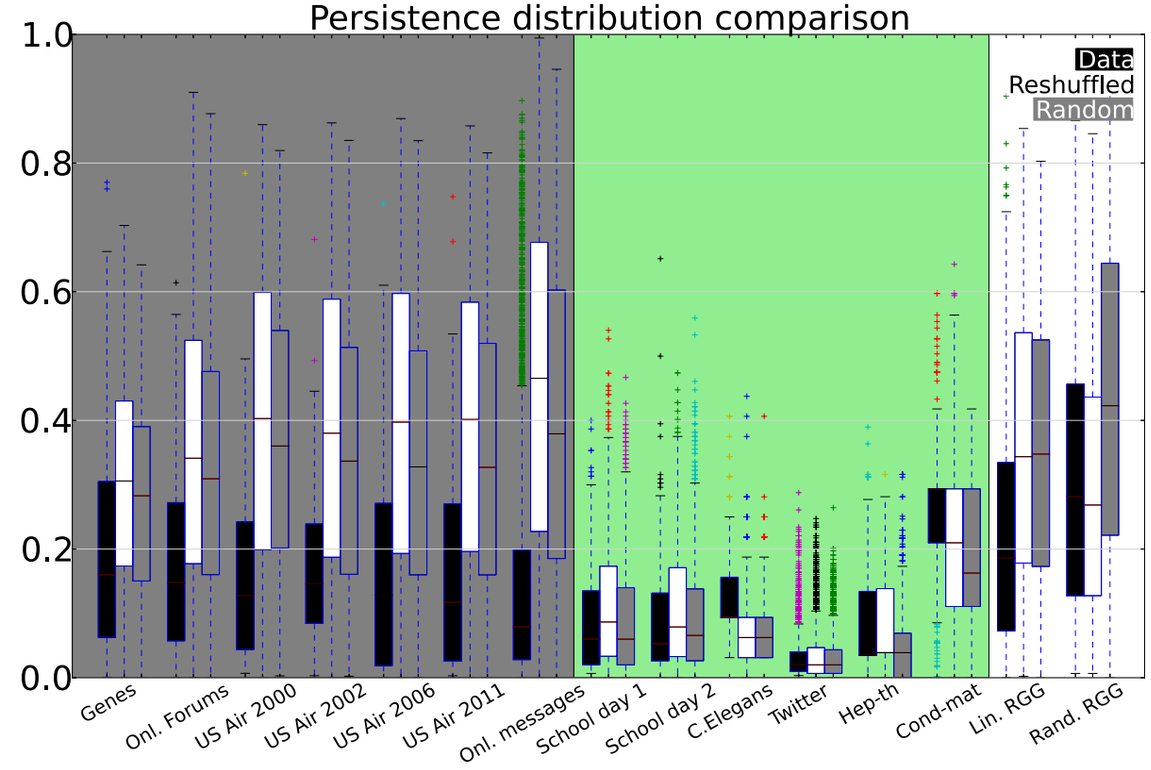
Topological cycles are not simply reproduced by statistical network properties

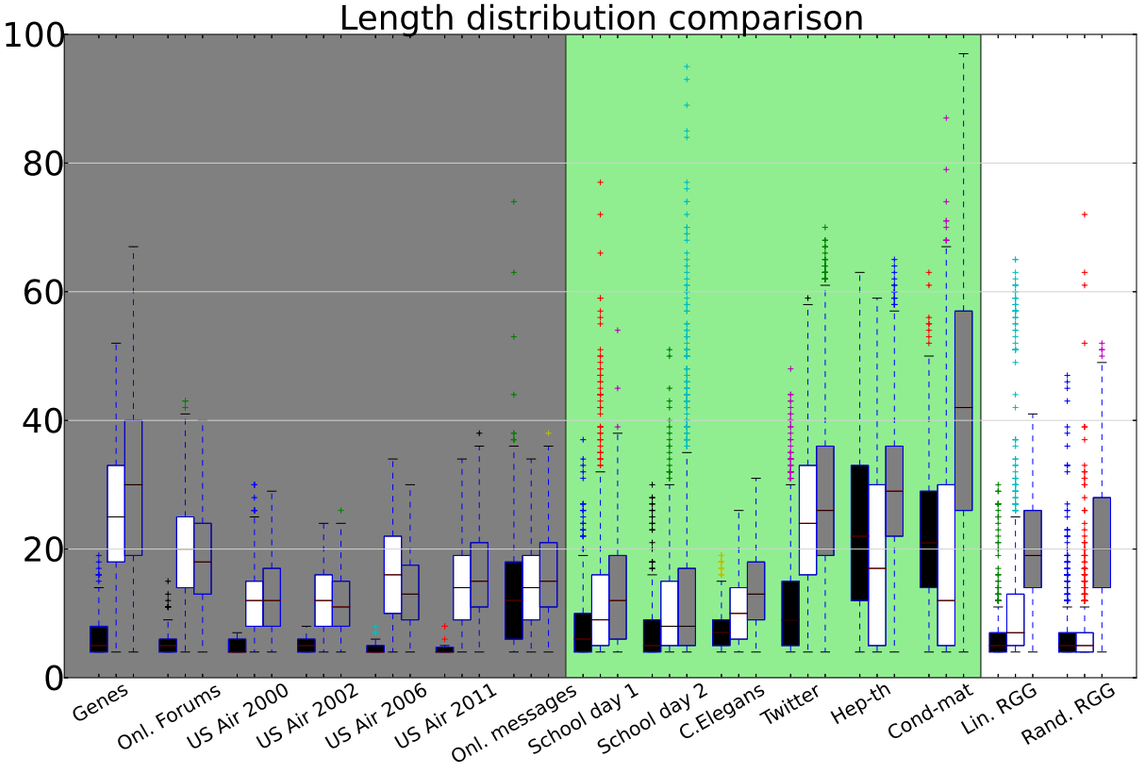
Datasets III: statistics of generators

Generators are characterized by:

1. Birth
2. Persistence
3. Length (of the cycle)







GP, M. Scolamiero, I. Donato, F.Vaccarino,
Topological strata of weighted complex networks, , PLoS ONE 8(6): e66506.

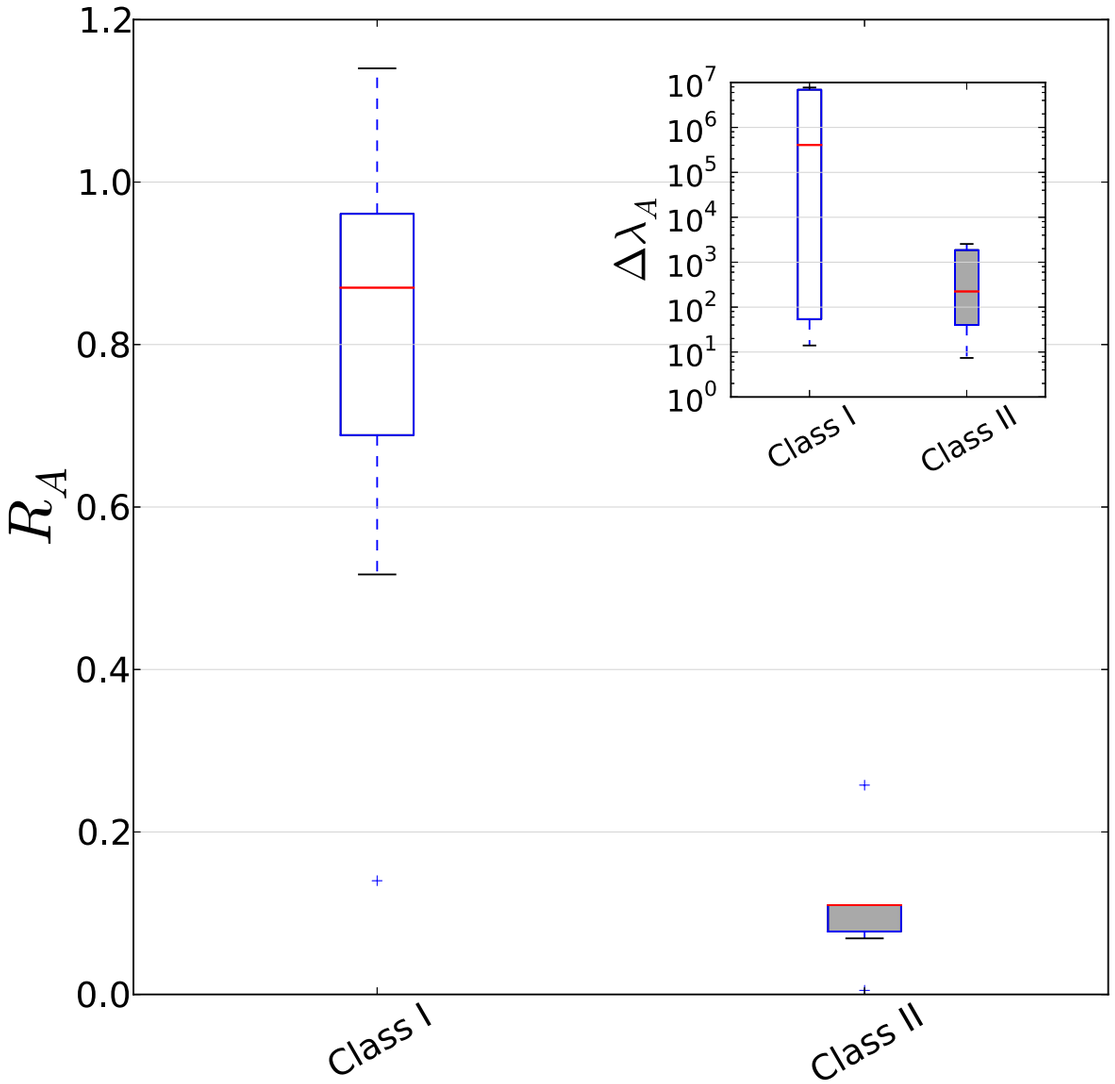
Datasets IV: spectral correlates?

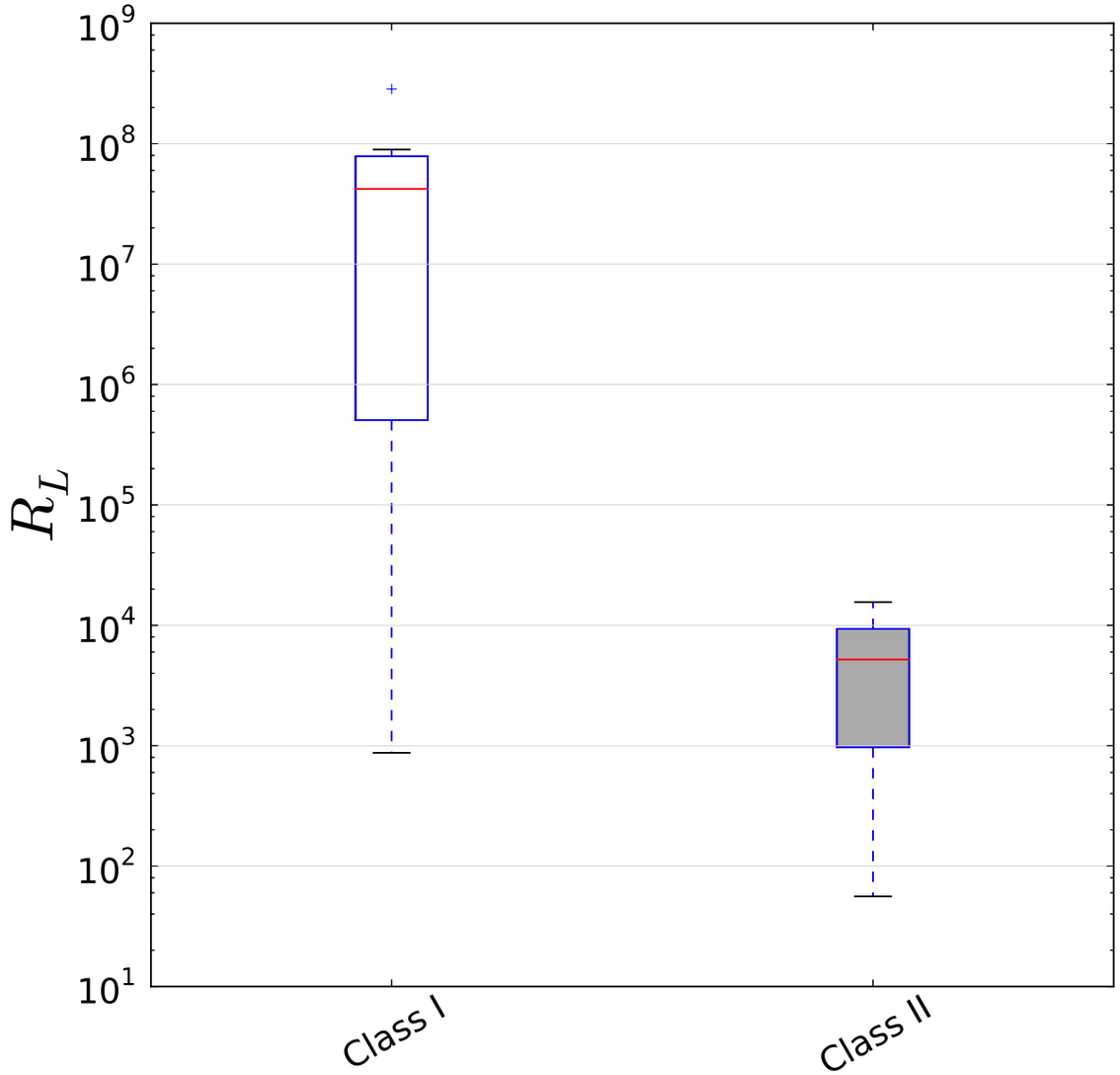
$$R_A = \frac{\lambda_1 - \lambda_2}{\lambda_2 - \lambda_N}$$

Related to graph expansion properties

$$R_L = \frac{\lambda_N^L}{\lambda_2^L}$$

Related to synchronisation and dyn. processes

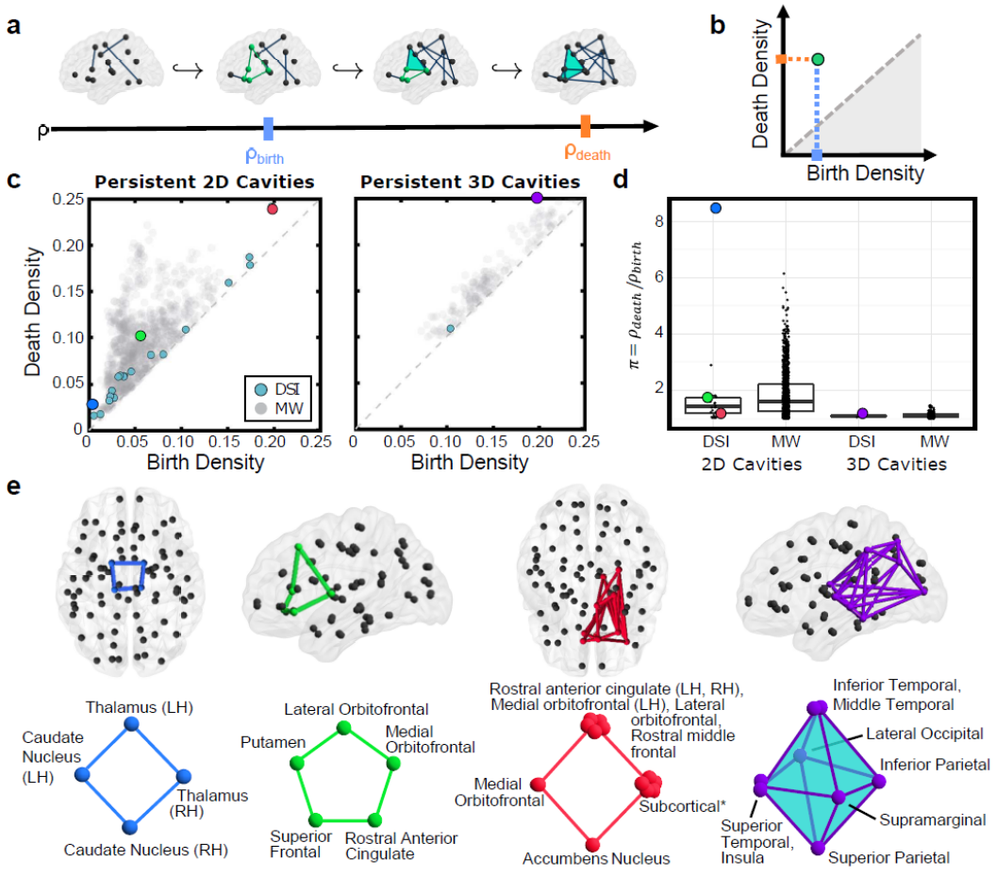


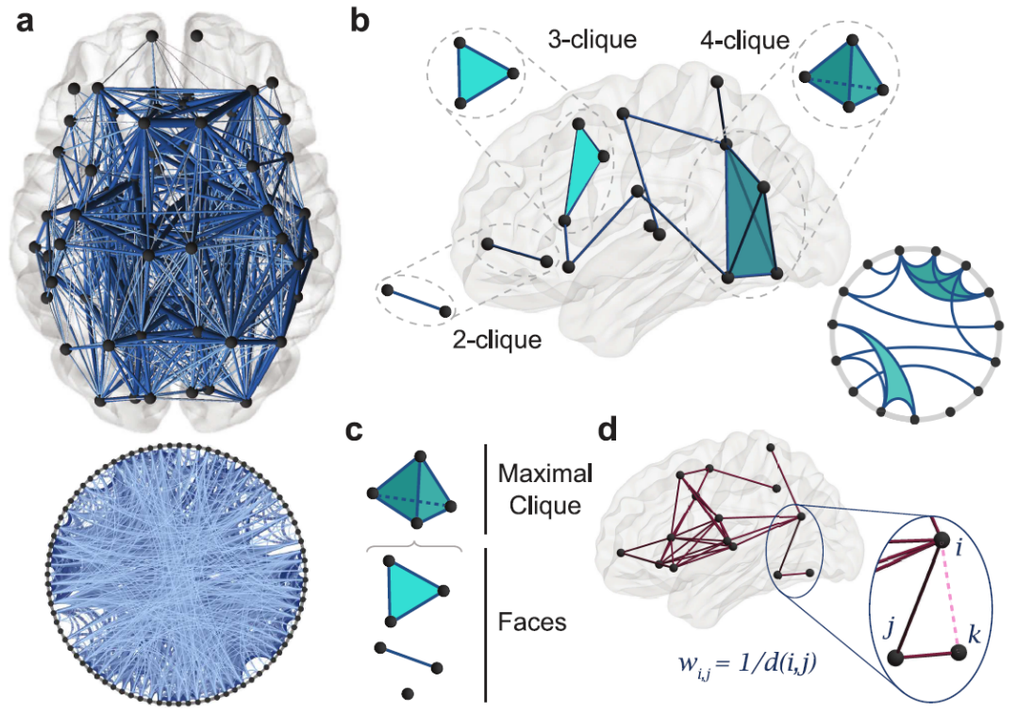


GP, M. Scolamiero, I. Donato, F.Vaccarino,
Topological strata of weighted complex networks, , PLoS ONE 8(6): e66506.

Previous work: *ctome topology

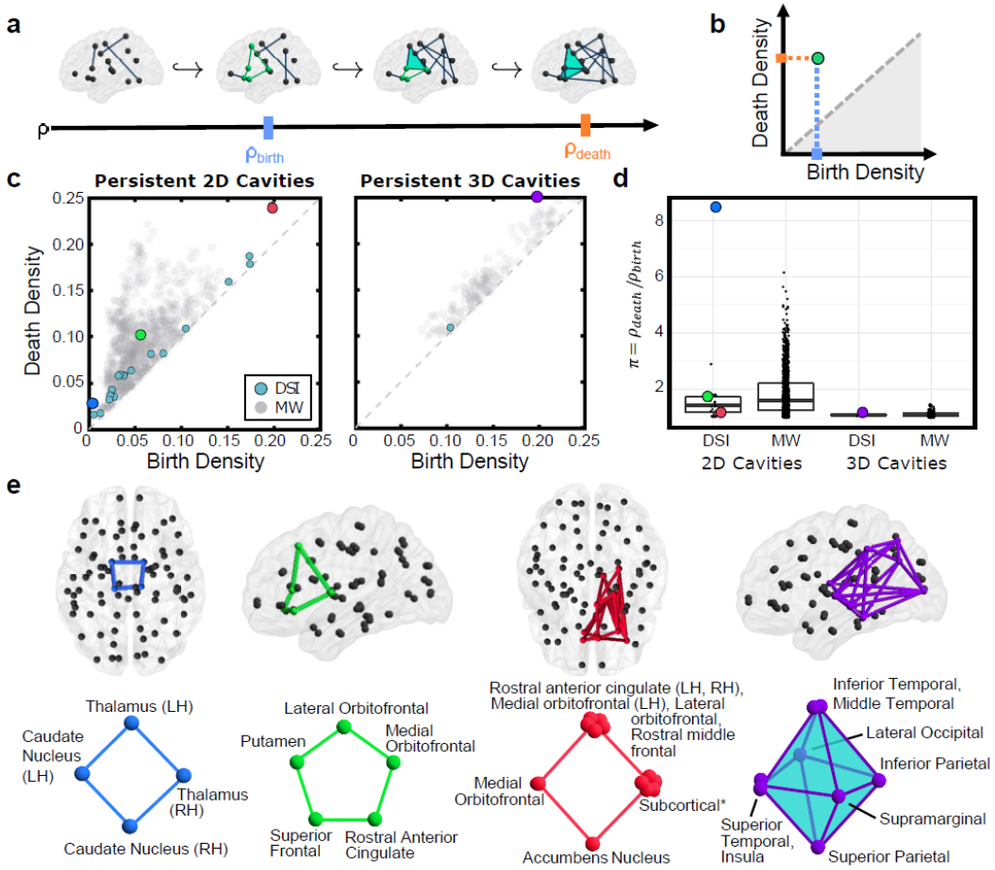
Previous work: *ctome topology

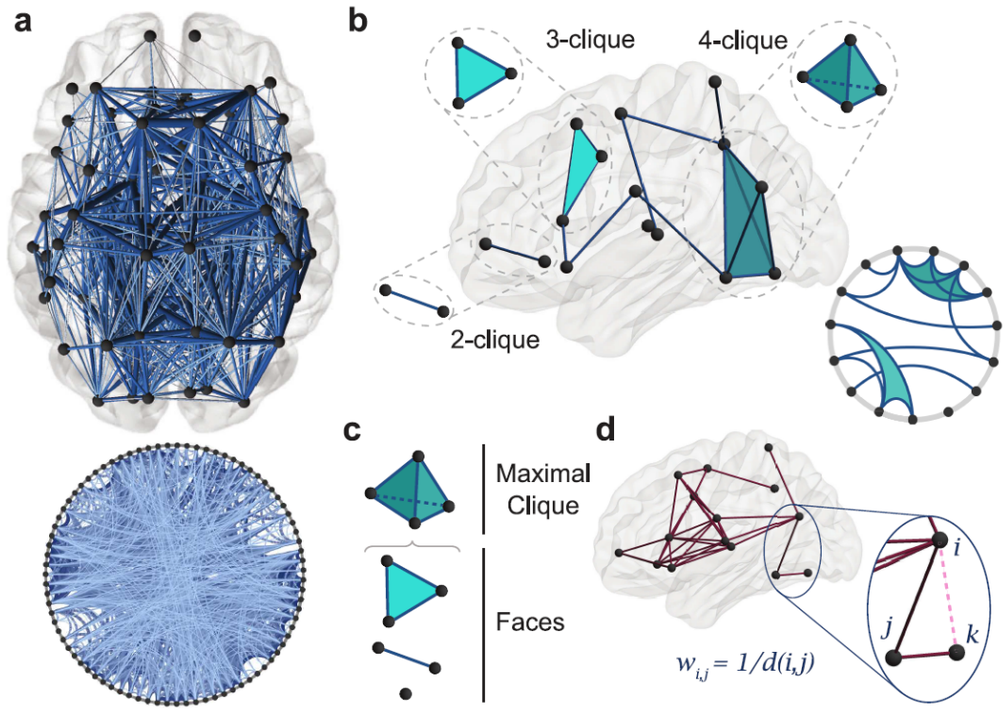




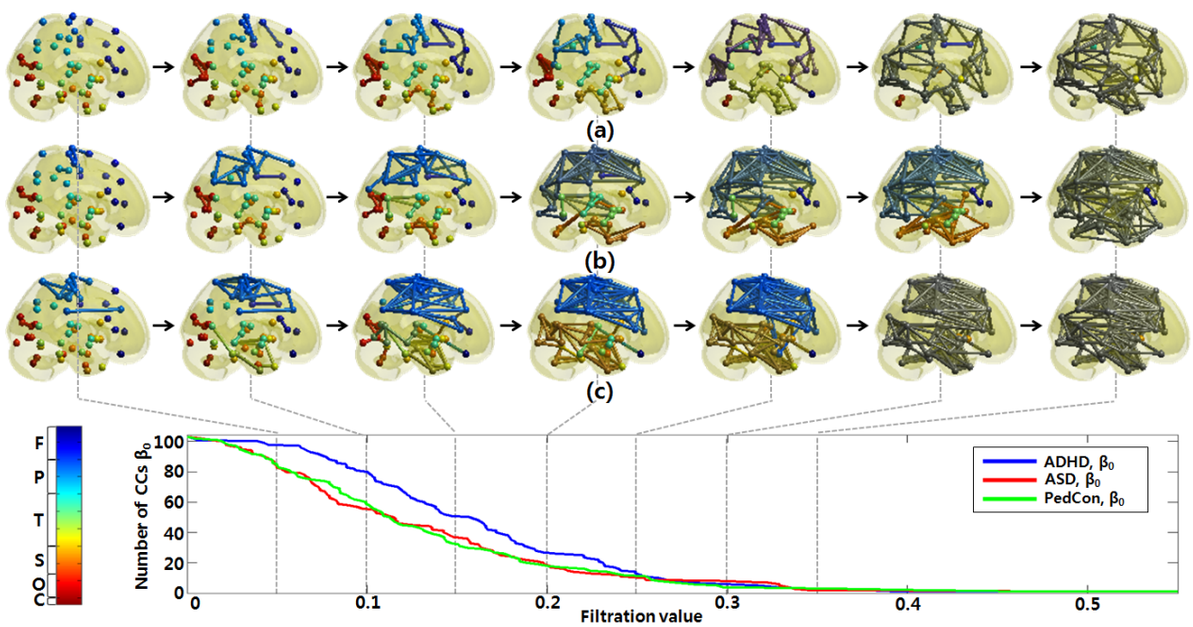
Sizemore, Ann, et al. arXiv:1608.03520 (2016).

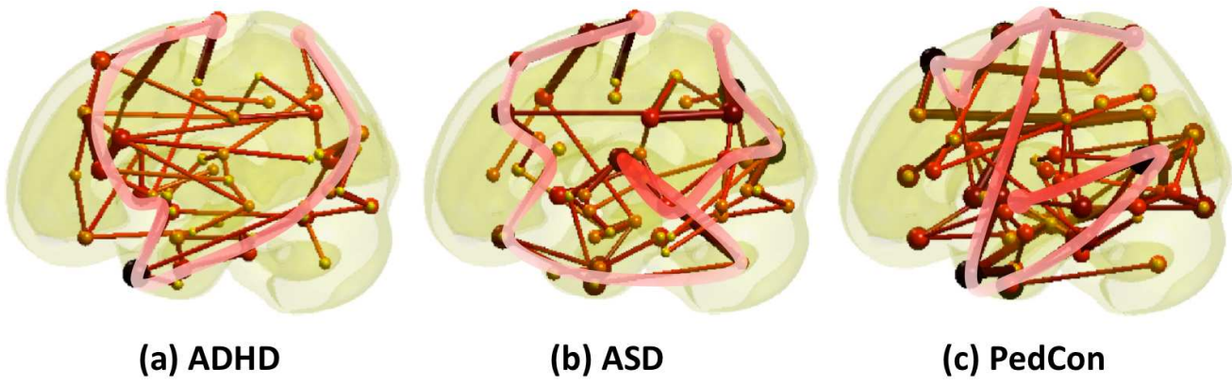
Previous work: *ctome topology





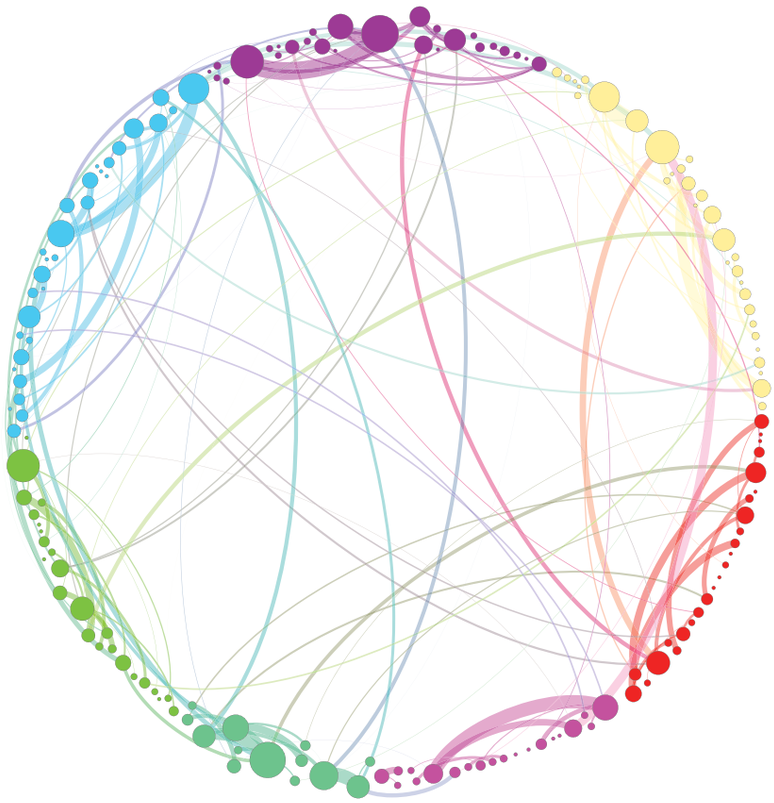
Sizemore, Ann, et al. arXiv:1608.03520 (2016).

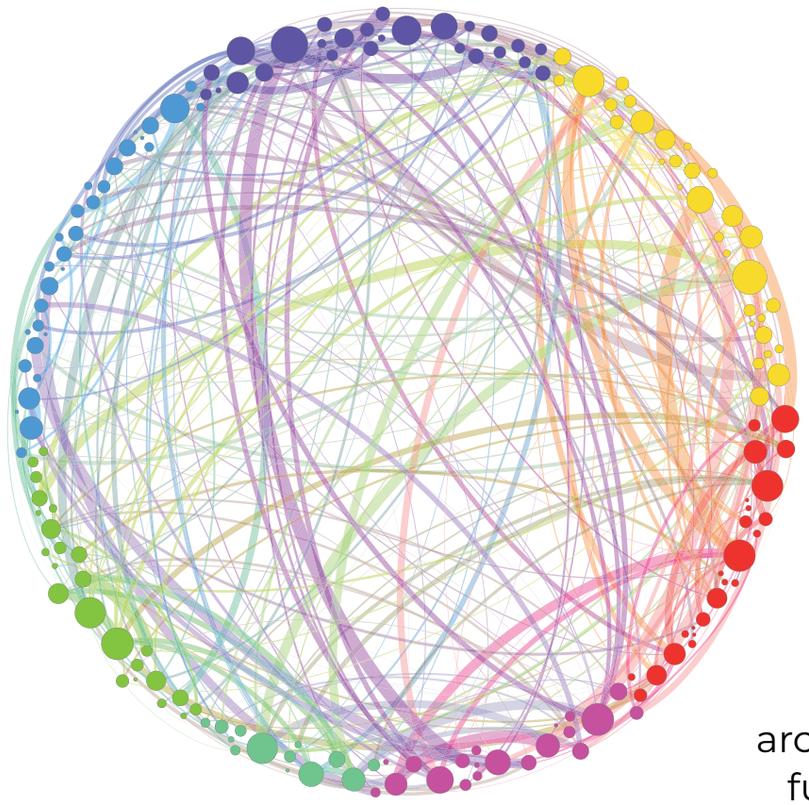




Lee et al. (2011) MICCAI2011
Lee et al. (2012) IEEE Trans. on Medical Imaging.

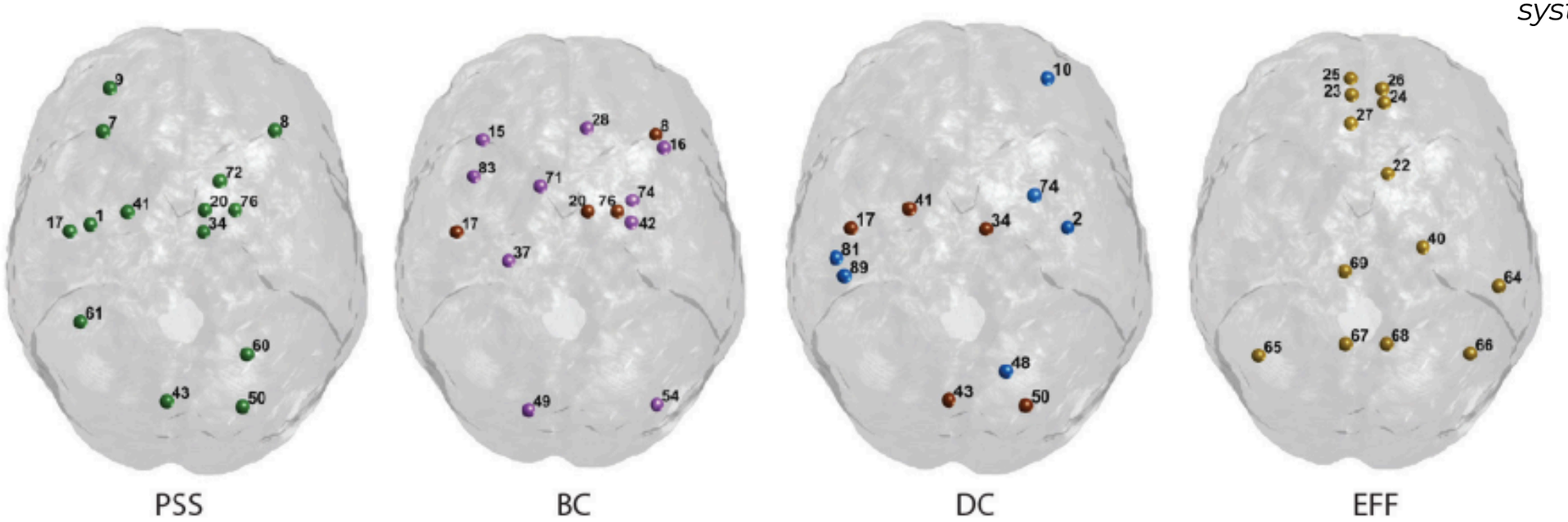
Previous work: homological backbones





Petri, Giovanni, et al. "Homological scaffolds of brain functional networks." *Journal of The Royal Society Interface* 11.101 (2014): 20140873.

Lord, Louis-David, et al. "Insights into brain architectures from the homological scaffolds of functional connectivity networks." *Frontiers in systems neuroscience* 10 (2016).

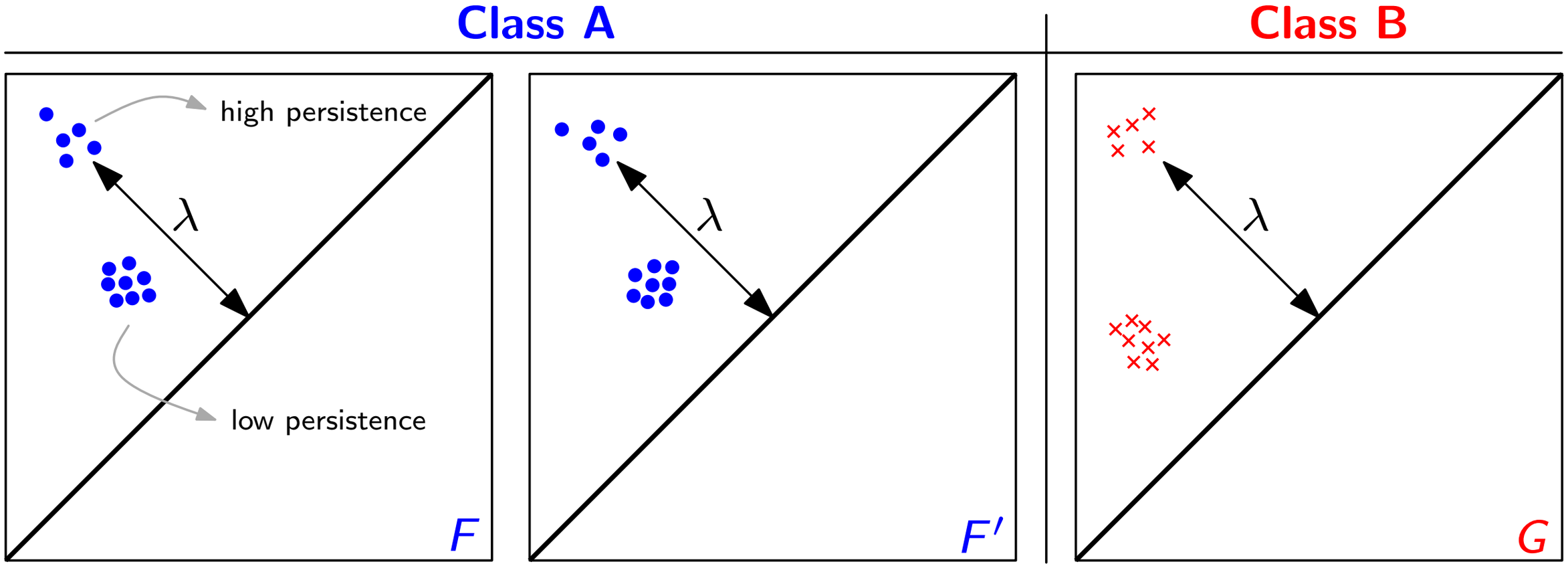


Notebook 05

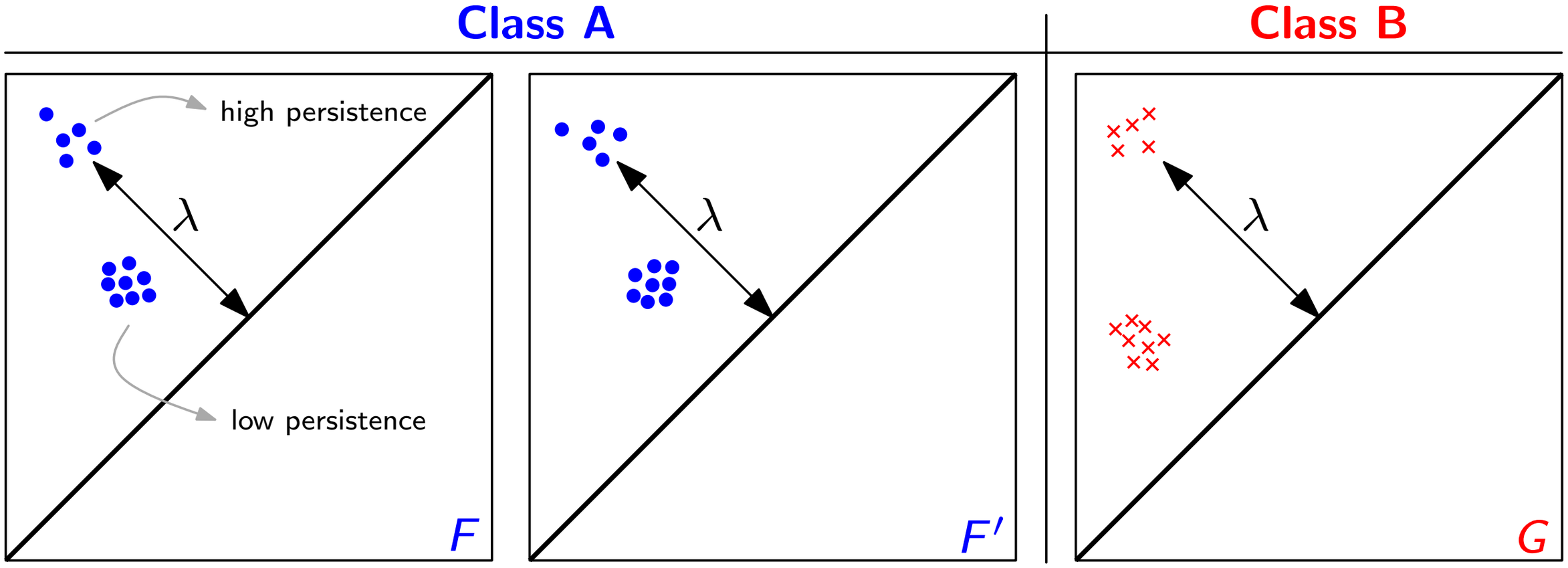
How do we compare these things?

Distances

Distances



Distances



Distances and Kernels:

- Bottleneck Distance
- (sliced) Wasserstein distance
- Persistence Scale Kernel (Reininghaus et al 2015)
- Weighted Persistence Kernel (Kusano et al, 2016)

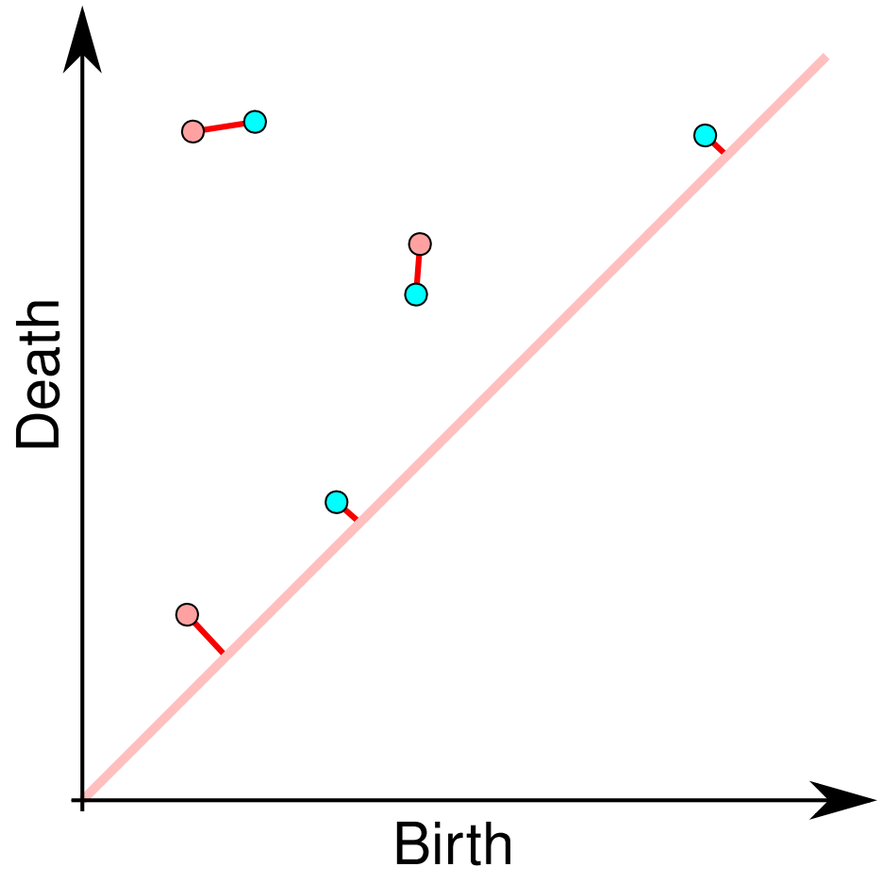
Distances

Bottleneck Distance

$$\delta_\infty(D_1, D_2) = \inf_\gamma \sup_{z \in D_1} ||z - \gamma(z)||_\infty$$

Wasserstein Distance

$$W_p\left(D_1, D_2\right) \;=\; \inf_\gamma \Big(\sum_{u \in D_1} \|u - \gamma(u)\|_\infty^p \Big)^{1/p}$$



Sliced Wasserstein Distance

$$\mathrm{SW}(\mathrm{Dg}_1, \mathrm{Dg}_2) \overset{\text{def.}}{=} \frac{1}{2\pi} \int_{\mathbb{S}_1} \mathcal{W}(\mu_1^\theta + \mu_{2\Delta}^\theta, \mu_2^\theta + \mu_{1\Delta}^\theta) \mathrm{d}\theta.$$

Persistence Scale Kernel

$$k_{\mathrm{SW}}(\mathrm{Dg}_1, \mathrm{Dg}_2) \overset{\text{def.}}{=} \exp\left(-\frac{\mathrm{SW}(\mathrm{Dg}_1, \mathrm{Dg}_2)}{2\sigma^2}\right).$$

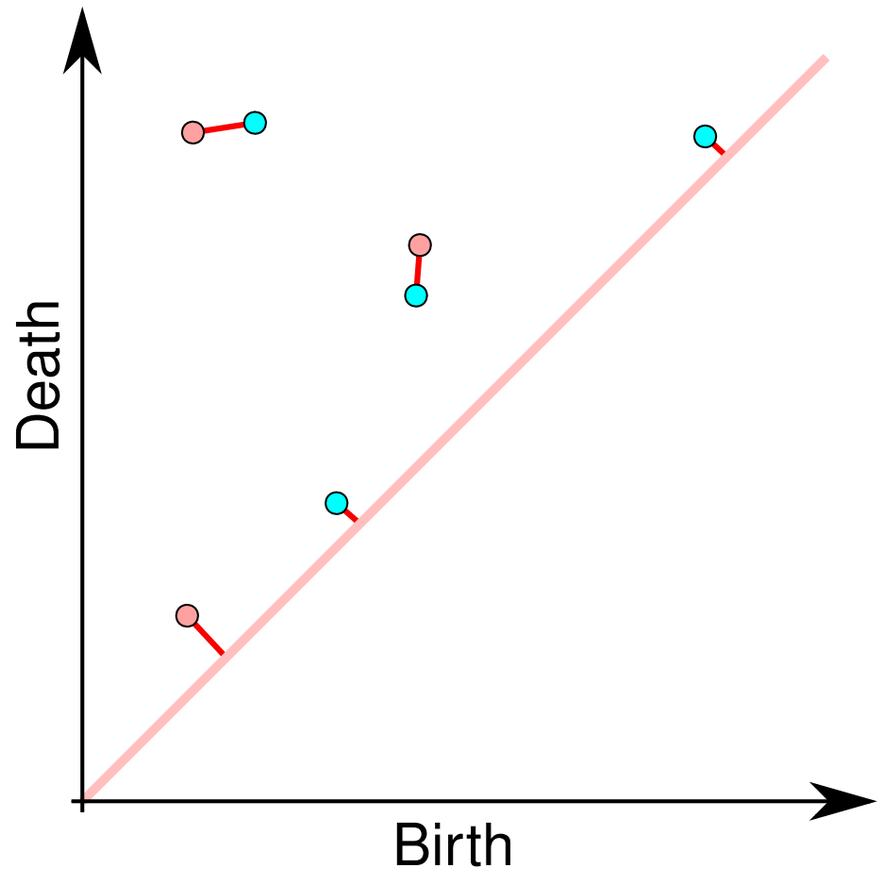
Distances

Bottleneck Distance

$$\delta_\infty(D_1, D_2) = \inf_\gamma \sup_{z \in D_1} ||z - \gamma(z)||_\infty$$

Wasserstein Distance

$$W_p\left(D_1, D_2\right) \;=\; \inf_\gamma \Big(\sum_{u \in D_1} \|u - \gamma(u)\|_\infty^p \Big)^{1/p}$$



Sliced Wasserstein Distance

$$\mathrm{SW}(\mathrm{Dg}_1, \mathrm{Dg}_2) \overset{\text{def.}}{=} \frac{1}{2\pi} \int_{\mathbb{S}_1} \mathcal{W}(\mu_1^\theta + \mu_{2\Delta}^\theta, \mu_2^\theta + \mu_{1\Delta}^\theta) \mathrm{d}\theta.$$

$$k_{\mathrm{SW}}(\mathrm{Dg}_1, \mathrm{Dg}_2) \overset{\text{def.}}{=} \exp\left(-\frac{\mathrm{SW}(\mathrm{Dg}_1, \mathrm{Dg}_2)}{2\sigma^2}\right).$$

Persistence Scale Kernel

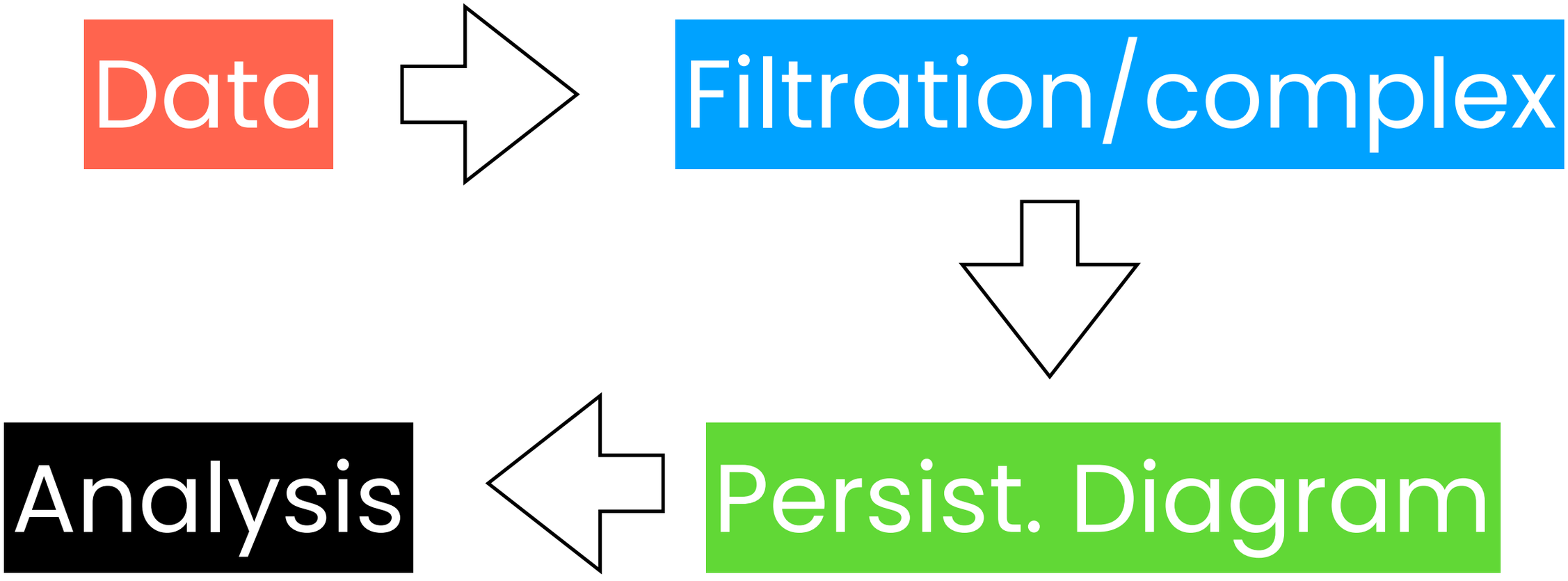
$$k_\sigma(F, G) = \frac{1}{8\pi\sigma} \sum_{\substack{p \in F \\ q \in G}} e^{-\frac{\|p-q\|^2}{8\sigma}} - e^{-\frac{\|p-\bar{q}\|^2}{8\sigma}}.$$

Notebook 06

Notebook

07-08-09

Randomization
+
validation

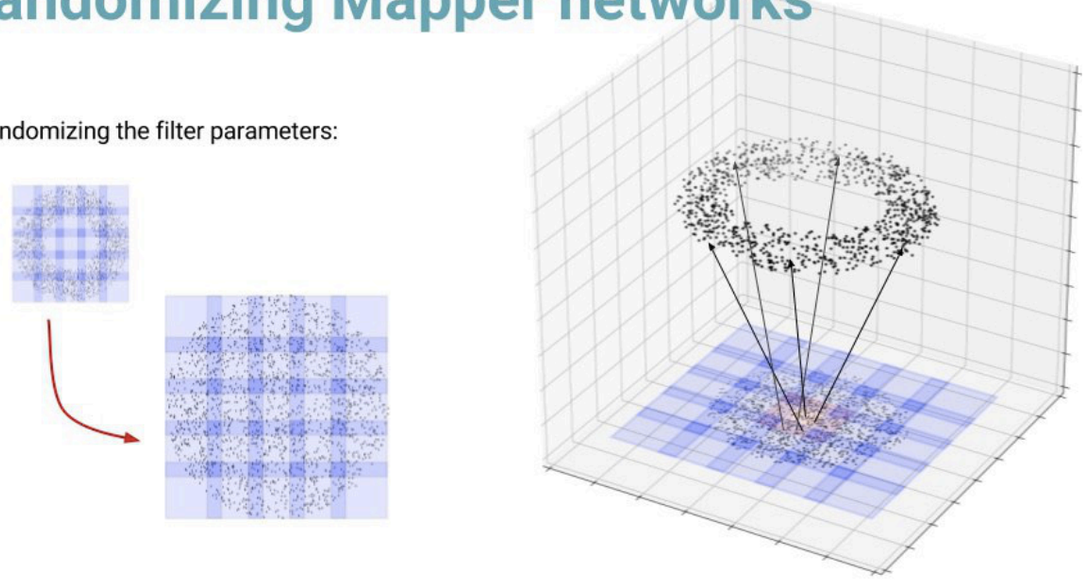


Higher-order randomizations

Filtration/complex

Randomizing Mapper networks

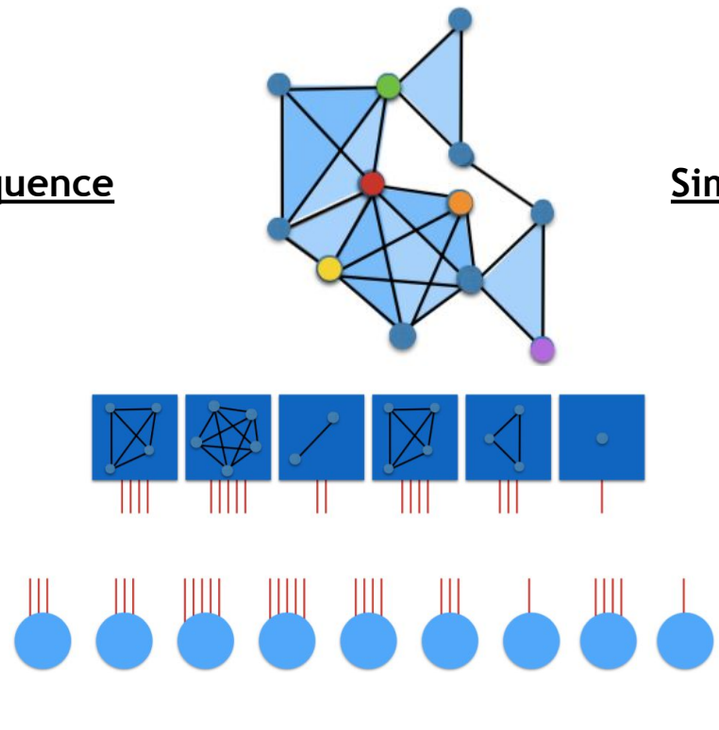
Randomizing the filter parameters:



Filter randomization

Facets size sequence

s_j
4
5
3
3
3
2

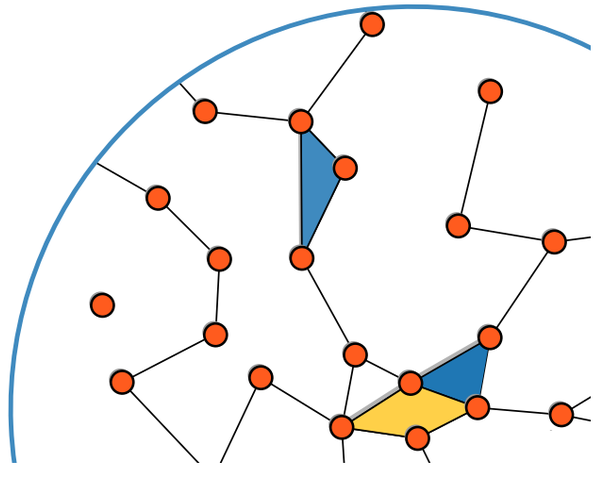


Simplicial degree seq

	k_i
●	1
●	3
●	1
●	2
●	2

Simplicial randomization

Joint work with JG Young and A. Patania
Reference : arxiv.org/1705.10298
Software : github.com/jg-you/scm



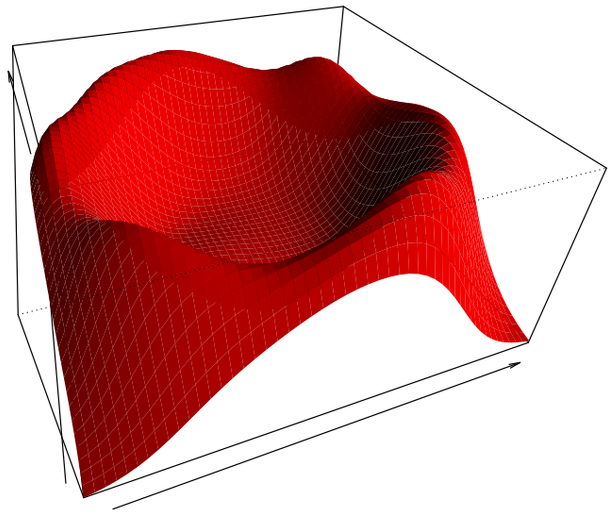
Confidence sets of PD

Persist. Diagram

Analysis

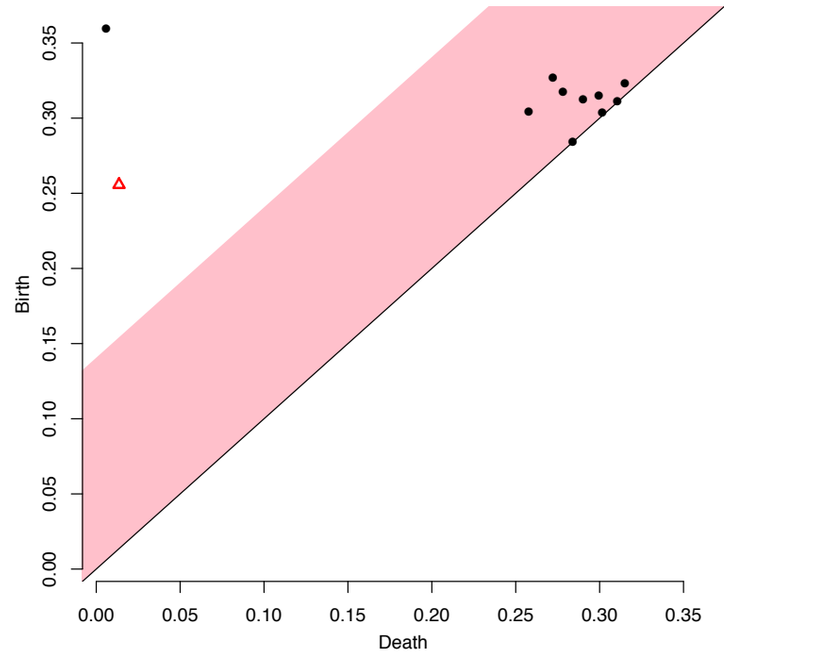
Fasy et al. (2014b); Chazal et al. (2014a) suggest the following method. Let

$$F(t) = P(\sqrt{n}\,\delta_\infty(\widehat{D}, D) \leq t)$$



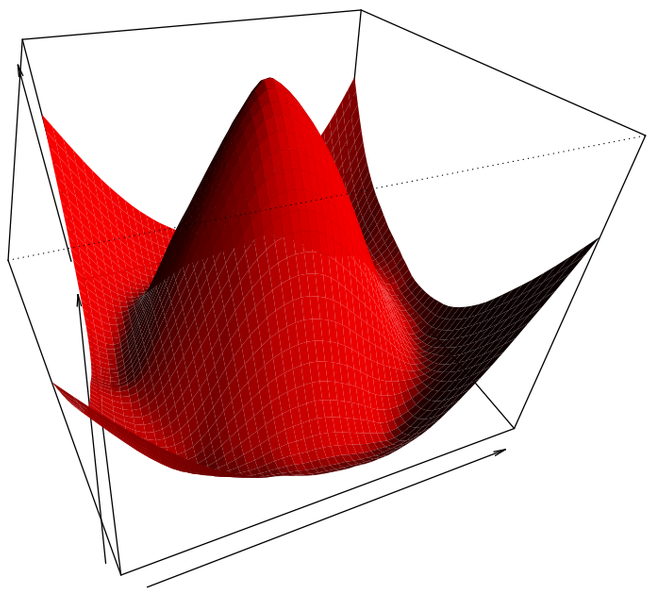
Estimated by bootstrap

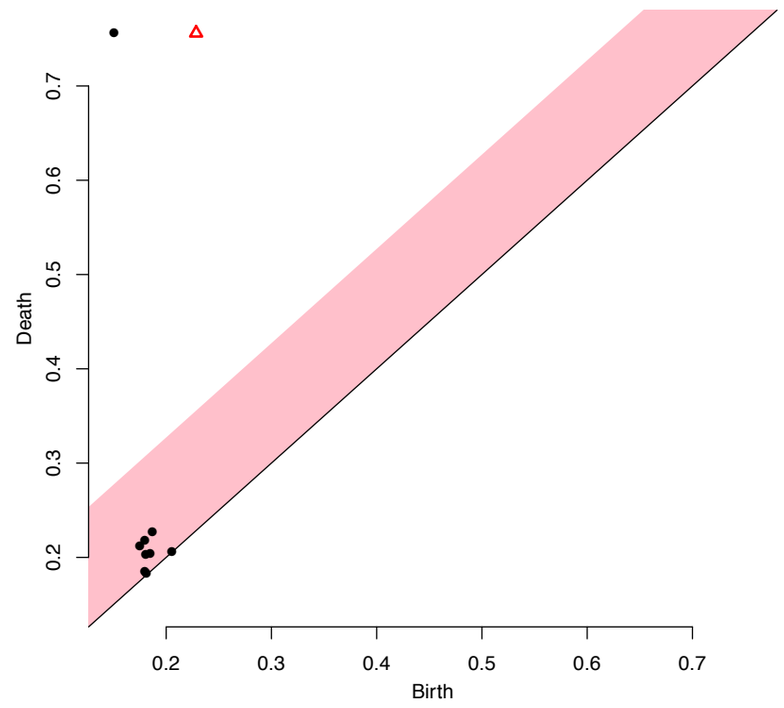
$$\widehat{F}(t) = \frac{1}{B}\sum_{j=1}^{B} I(\sqrt{n} d_\infty(\widehat{D}_j^*, \widehat{D}) \leq t)$$



Significance threshold

$$t_\alpha = F^{-1}(1-\alpha)$$





Note: Works for metrical cases.

Tip: try this in the TDA r-library

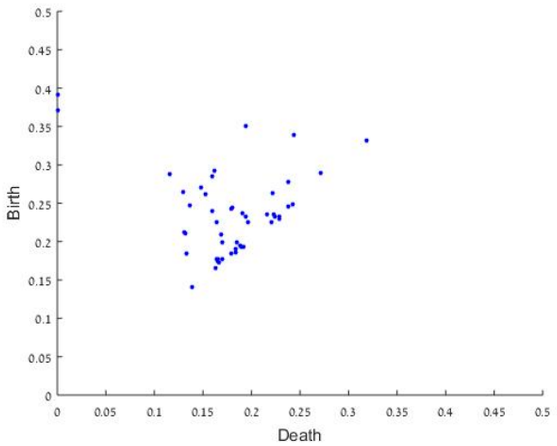
PD Multiplication

Persist. Diagram

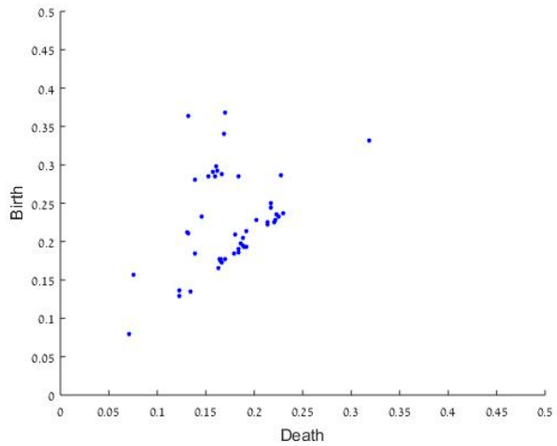
- Compute your PD
- Fit a Gibbs Model with Hamiltonian
- Use it to produce synthetic PDs

$$\varphi_\Theta(\tilde{x}_N) = \frac{1}{Z_\Theta}\exp(-H_\Theta(\tilde{x}_N)),$$

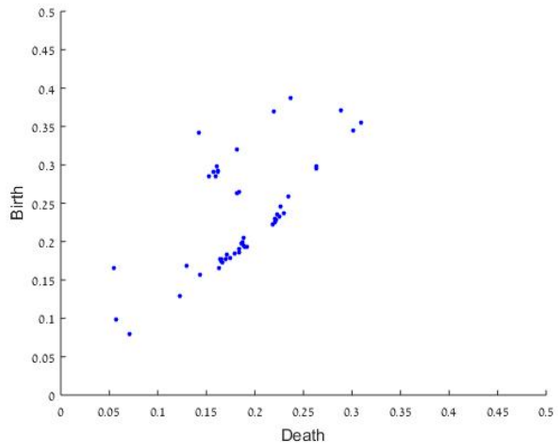
$$H^2_{\delta,\Theta}(\tilde{x}_N) = \theta_H\sigma^2_H + \theta_V\sigma^2_V + \sum_{k=1}^{3}\delta^{-2}\theta_k\mathcal{L}_{\delta,k}(\tilde{x}_N),$$



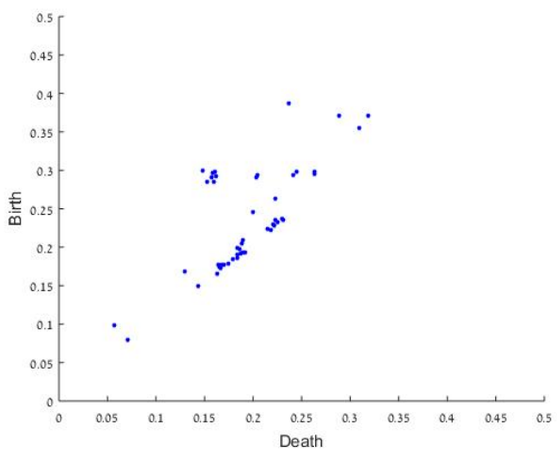
(a)



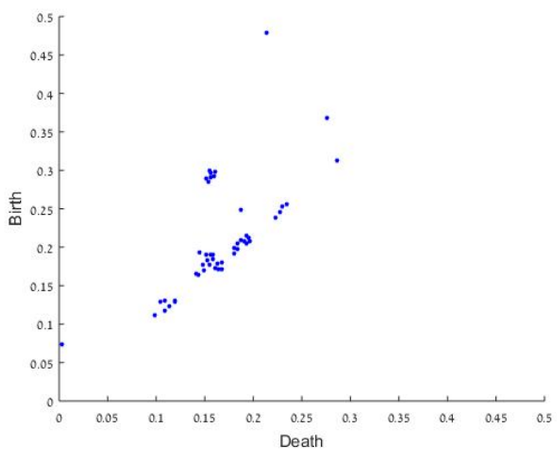
(b)



(c)



(d)



(e)

Modeling and replicating statistical topology and evidence for CMB nonhomogeneity

Robert J. Adler[a,1], Sarit Agami[a], and Pratyush Pranav[a]

The end