University of Wisconsin
SCHOOL OF MEDICINE
AND PUBLIC HEALTH

Topological Data Analysis

Moo K. Chung
Department of Biostatistics and Medical Informatics
Waisman Laboratory for Brain Imaging and Behavior
University of Wisconsin-Madison

www.stat.wisc.edu/~mchung

Topological Data Analysis

Topological data analysis (TDA) is an approach to the analysis of datasets using techniques from topology. Extraction of information from datasets that are often high-dimensional and noisy is generally challenging.

TDA provides a general framework to analyze such data in a manner that is insensitive to the particular metric and scale and provides dimensionality reduction and robustness to noise.

Limitations of TDA

1) May not work on simple data.

2) Other methods can easily beat the method in terms of performance.

3) Difficult to design the pipeline that works. TDA is not like linear regression.

Limitations of TDA

4) TDA is easy to understand but very hard to apply.

5) The framework for statistical inference is lacking.

6) TDA is too slow to compute

Advantage of TDA

1)When it works, performance is beyond incremental gain.

2) The method can easily beat deep learning or any advanced methods if underlying data/task is topological.

Topological Data Analysis

↑ Statistics

Statistics is the discipline that concerns the collection, organization, analysis, interpretation and presentation of data.

Unfortunately, traditional TDA is not statistical

Homology

Persistent homology computes the topological features of a space at different spatial resolutions.

More persistent features are detected over a wide range of spatial scales and are deemed more likely to represent true features of the underlying space rather than artifacts of sampling, noise, or particular choice of parameters.

Persistent homology

Powerful algebraic framework for characterizing topological signals at various spatial and temporal scales. Robustness to the choice of scale and parameters → Topological multiscale approach

References

Edelsbrunner et al. 2008 :Eearly survey paper

Carlsson & de Silva, 2010

Edelsbrunner & Harer, 2009

Chung et al. 2020 arXiv:2102.08623 :Review paper (*focused on topological distances between networks*)

n-simplex

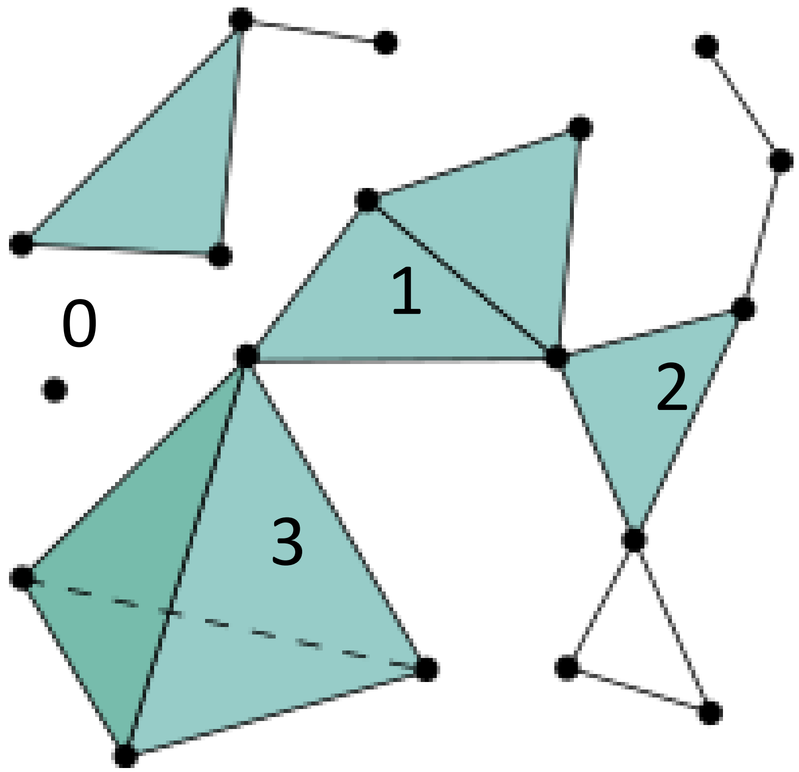
The basic building block of persistent homology

The smallest convex set containing *n+1* points

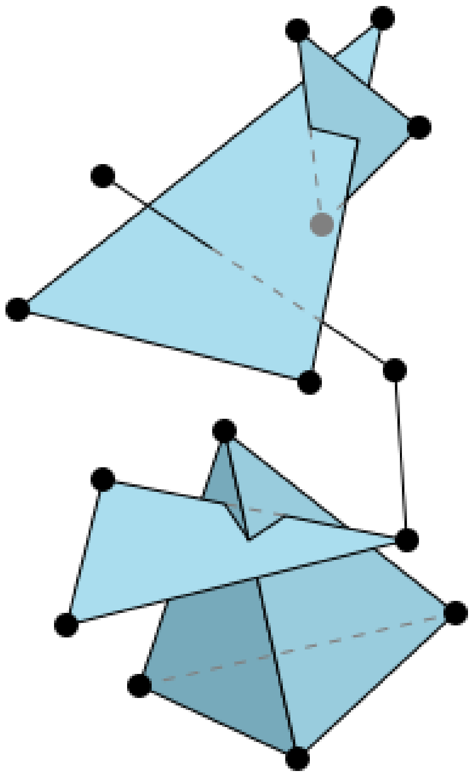
$$\sum_{i=0}^{n} x_i = 1, x_i \geq 0$$

Simplicial complex

A simplicial complex is a set composed of points, line segments, triangles, and their n-dimensional counterparts.



Simplicial complex

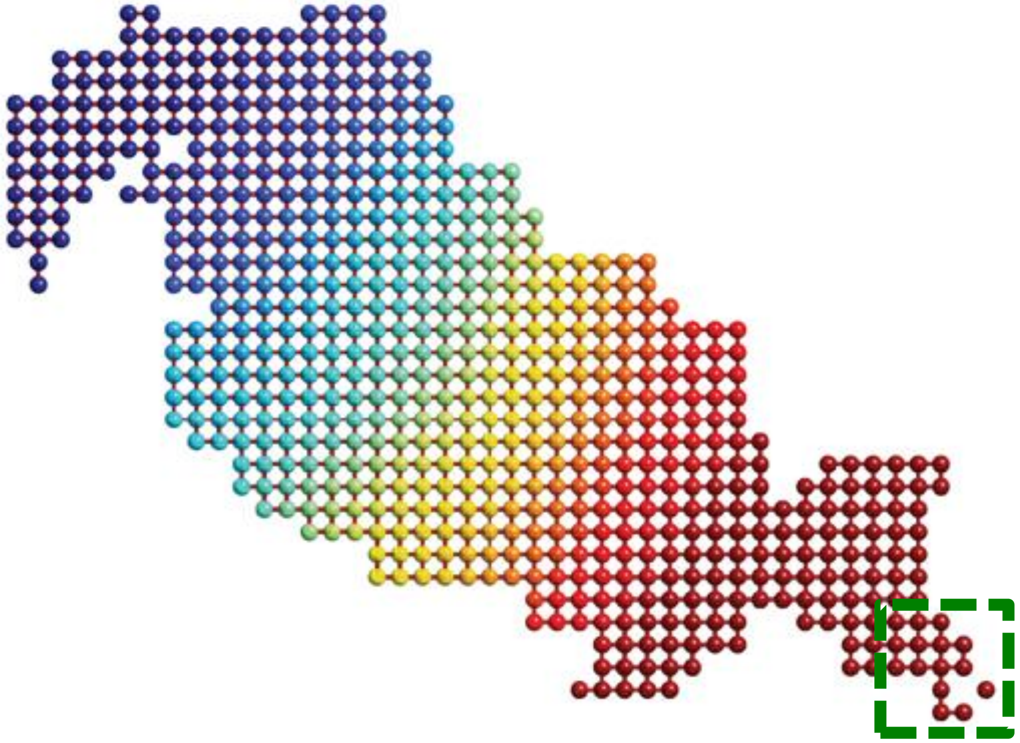


Not simplicial complex

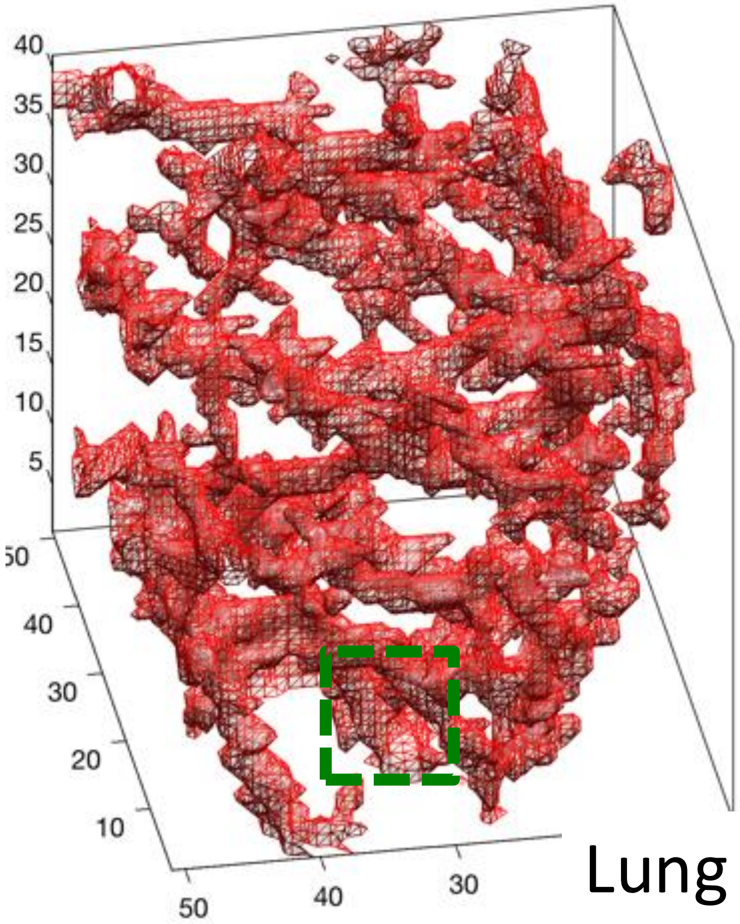
Ex. surface mesh, graphs (including hyperg

networks

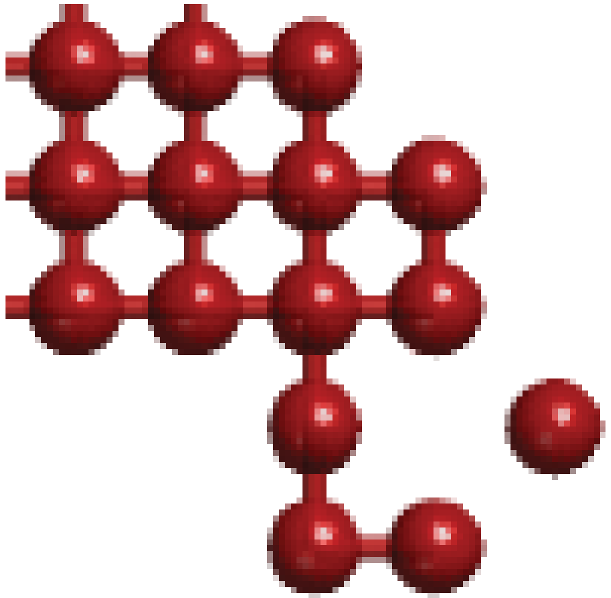
(2D or 3D images)



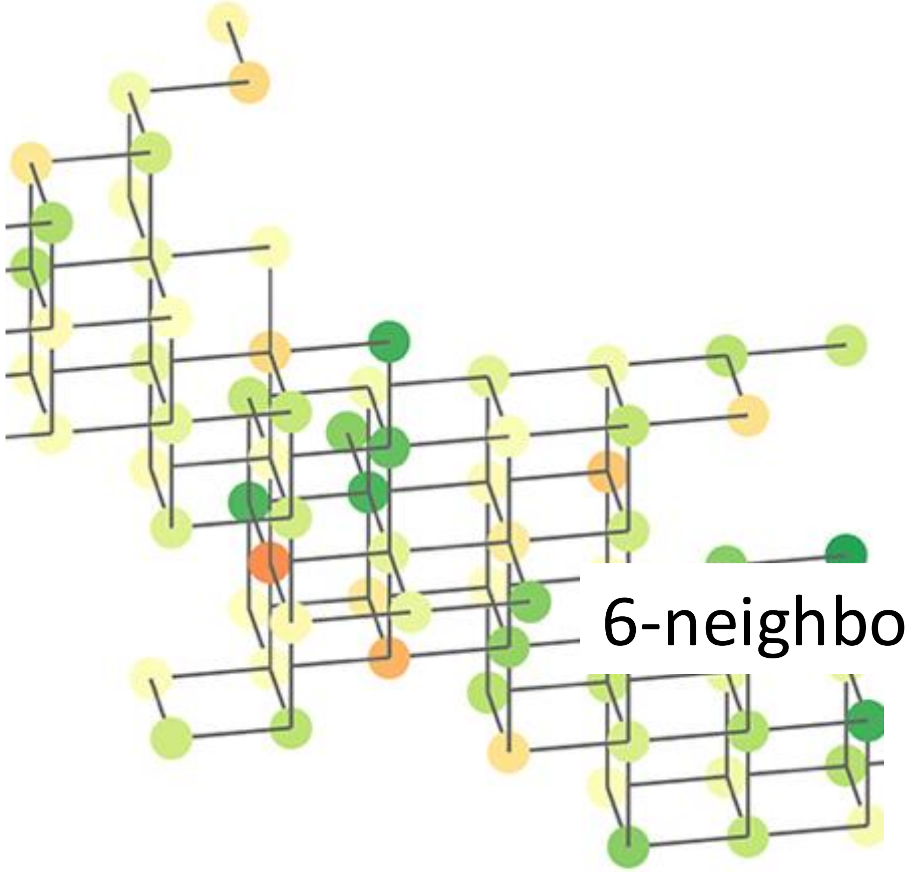
Left central gyrus



Lung blood vessel



4-neighbor connectivity



6-neighbor connectivity

Filtrati
on

$$\mathcal{G}_1 \subset \mathcal{G}_2 \subset \mathcal{G}_3 \subset \cdots$$

Sequence of nested objects or vector spaces

↓

Extract persistent homological features

Persistent diagram, barcodes

Hiearchical nestness does not imply robustness

$$\mathcal{G}_1 \subset \mathcal{G}_2 \subset \mathcal{G}_3 \subset \cdots$$

Sequence of nested objects or vector spaces

To have robustness, you need monotone feature

$$\beta_i(\mathcal{G}_1) < \beta_i(\mathcal{G}_2) < \beta_i(\mathcal{G}_3) < \cdots$$

Robust approach   Chung et al. 2019 Network Neuroscience

$$\mathcal{G}_1 \subset \mathcal{G}_2 \subset \mathcal{G}_3 \subset \cdots$$

Use more structured filtration

$$\tilde{\mathcal{G}}_1 \subset \tilde{\mathcal{G}}_2 \subset \tilde{\mathcal{G}}_3 \subset \cdots$$

Betti numbers will be monotonic

$$\beta_i(\tilde{\mathcal{G}}_1) < \beta_i(\tilde{\mathcal{G}}_1) < \beta_i(\tilde{\mathcal{G}}_1) < \cdots$$

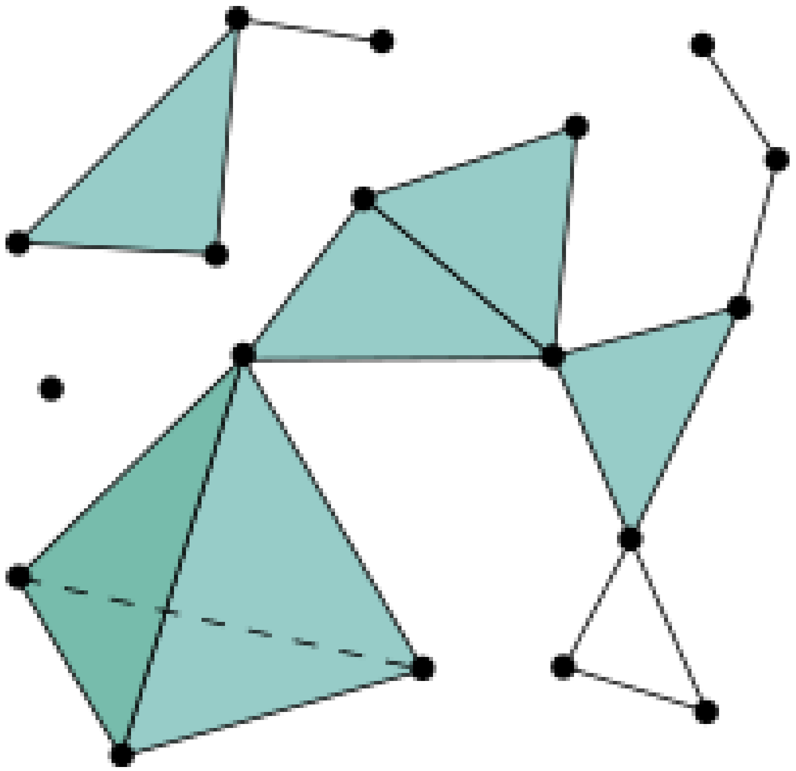
Graph filtrations

Baseline filtration for brain networks first introduced in

Lee et al. 2011 MICCAI 302-309
Lee et al. 2012 IEEE Transactions on Medical Imaging 31:2267-2277

1-skeleton

A simplicial complex consisting of points and line segments only → graphs & network



For networks, we may not really need filled-in triangles.1-skeleton is often more than enough.

Question: What is the biological interpretation of a filled triangle?

Rips filtration vs. graph filtration

Metric space

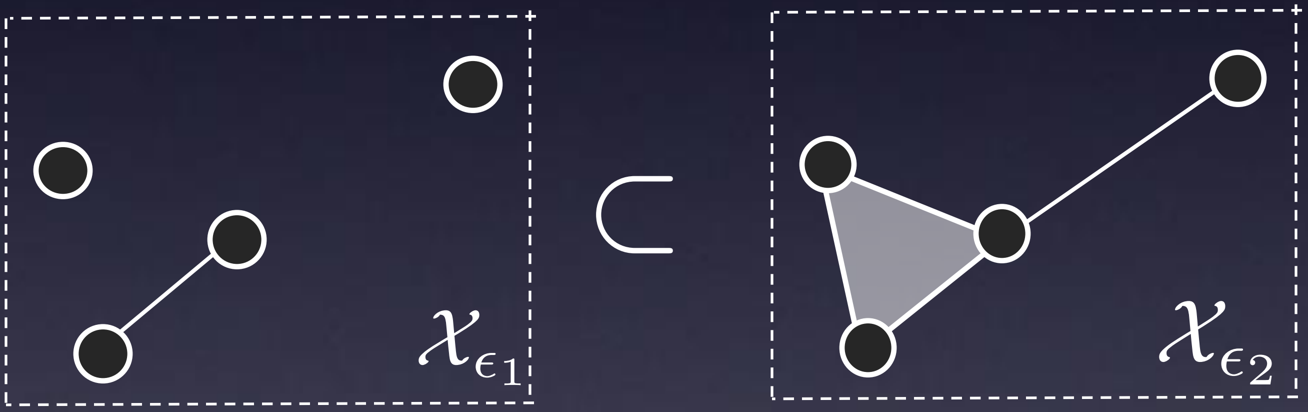
$$\mathcal{X} = (V, w) \quad w = (w_{ij})$$

Node set

Metric

$$w_{ik} < w_{ij} + w_{jk}$$

Rips complex = Simplicial complex



Rips filtration

$$\mathcal{X}_{\epsilon_0} \subset \mathcal{X}_{\epsilon_1} \subset \mathcal{X}_{\epsilon_2} \subset \cdots$$

for increased radius

$$\epsilon_0 < \epsilon_1 < \epsilon_2 < \cdots$$

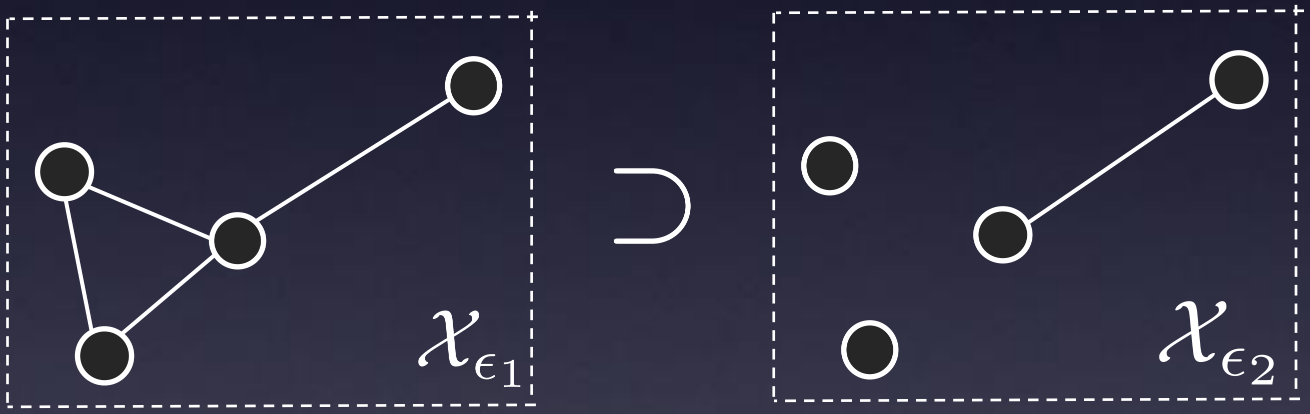
Weighted graph

$$\mathcal{X} = (V, w) \quad w = (w_{ij})$$

Node set

Edge weight

Binary graph: 1-skeleton



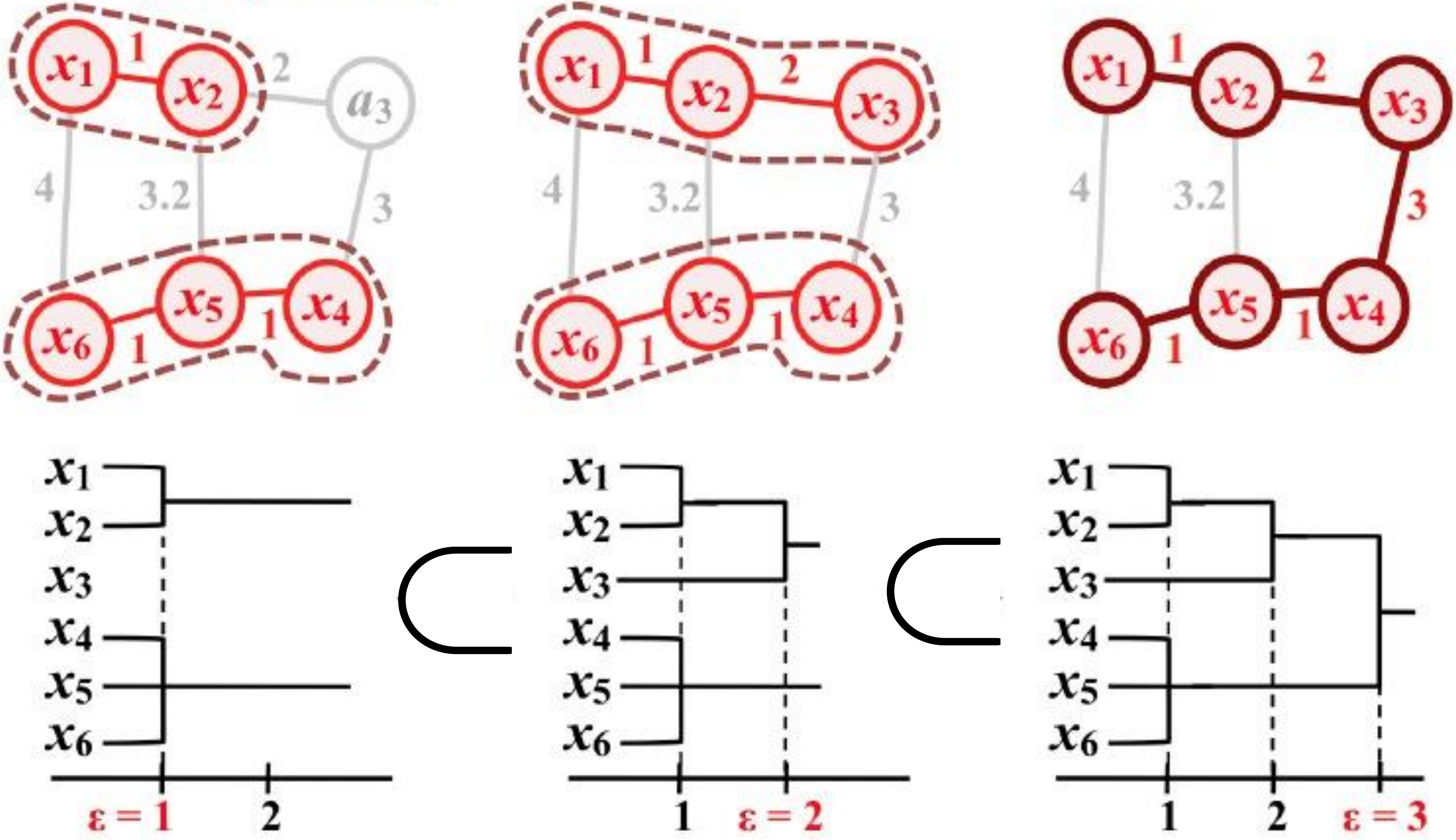
Graph filtration

$$\mathcal{X}_{\epsilon_0} \supset \mathcal{X}_{\epsilon_1} \supset \mathcal{X}_{\epsilon_2} \supset \cdots$$

for increased edge weights

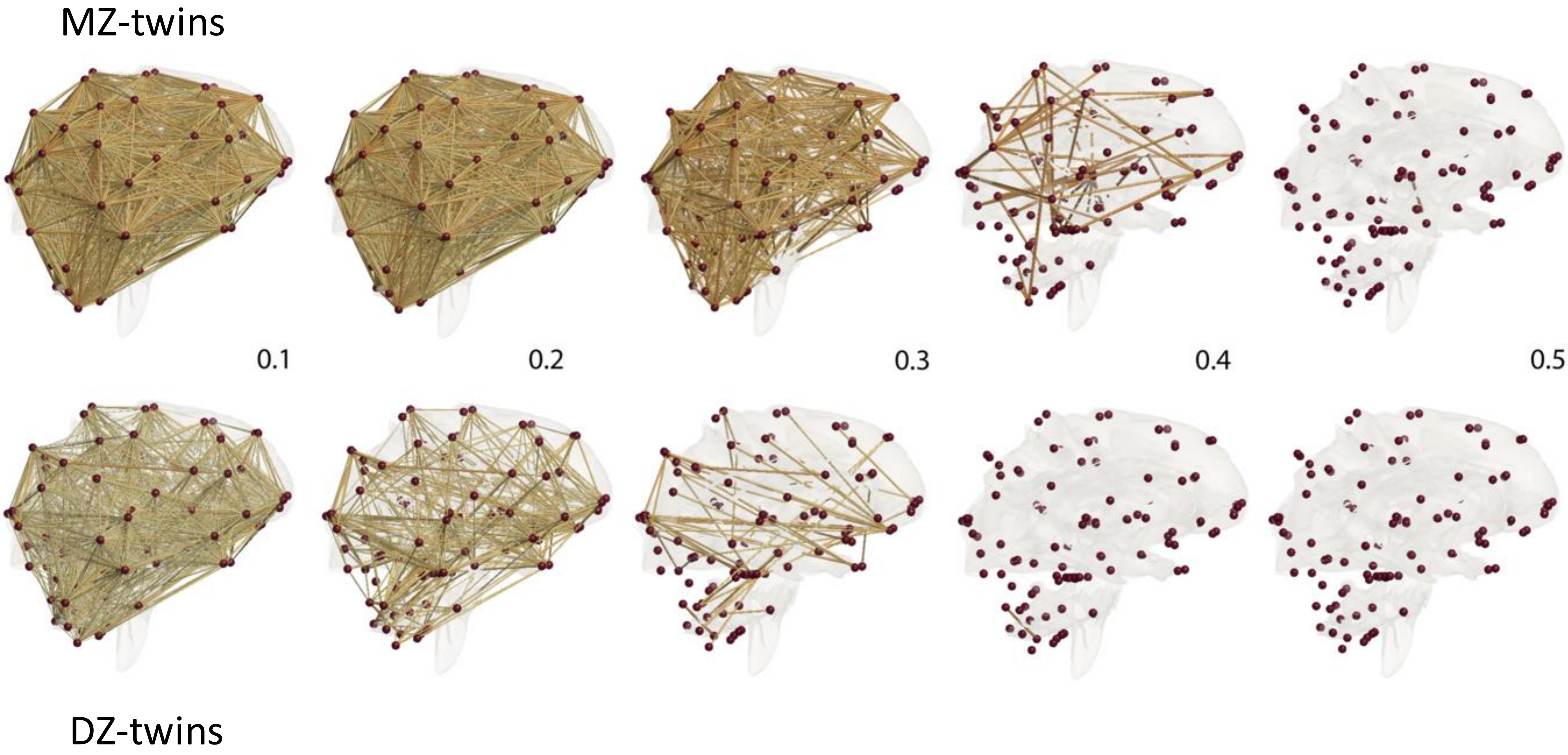
$$\epsilon_0 < \epsilon_1 < \epsilon_2 < \cdots$$

Graph filtration=single linkage clustering



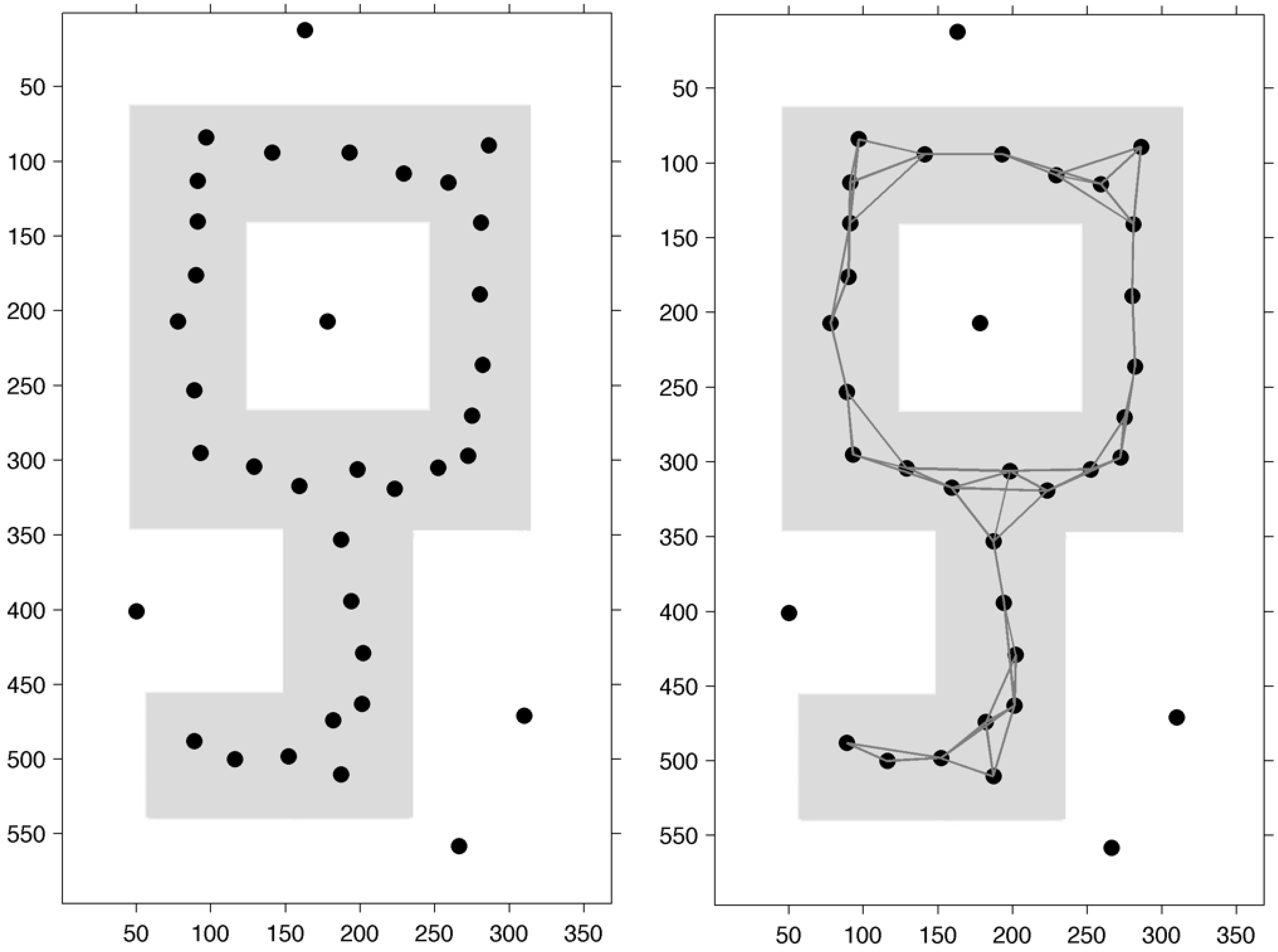
Lee et al. 2011 MICCAI 302-309

Graph filtrations on resting-state fMRI



1-skeleton of point cloud data

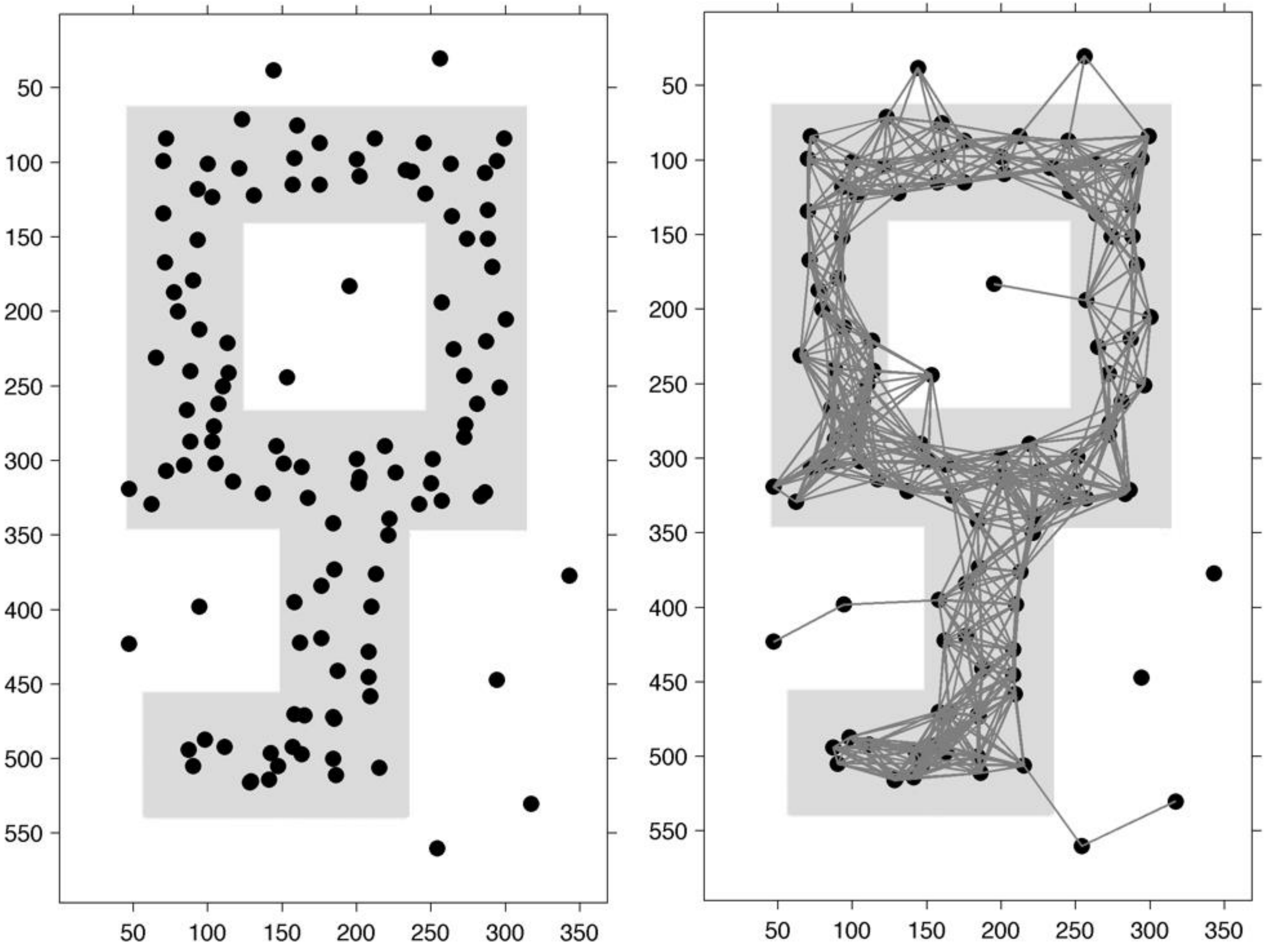
ε = 70mm



Recovering underlying topology

1-skeleton of point cloud data

ε = 70mm



Better approach: perform kernel smoothing and then Morse filtration

Advantage of 1-skeleton over Rips complex

Easy biological interpretation

Betti numbers are monotone over filtration

Robustness

Easier statistical inference

Graph filtration (filtration on 1-skeleton)

Rips filtration is computationally expensive:
For *n*-nodes, O(n^{3k+3}) for the *k*-th Betti number.

For 1-skeleton, graph filtration is O(*n* log *n*) for both 0-th and 1-st Betti number.

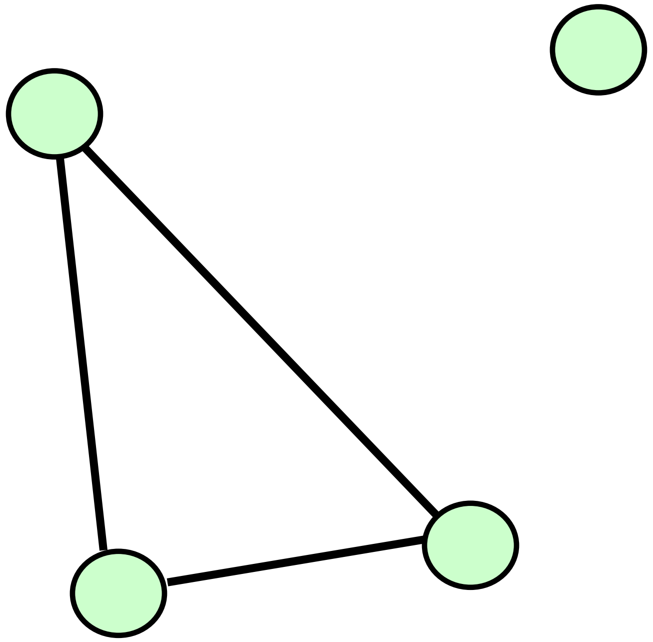
Exercise: Check the run time

Monotonicity of Betti-0 plot

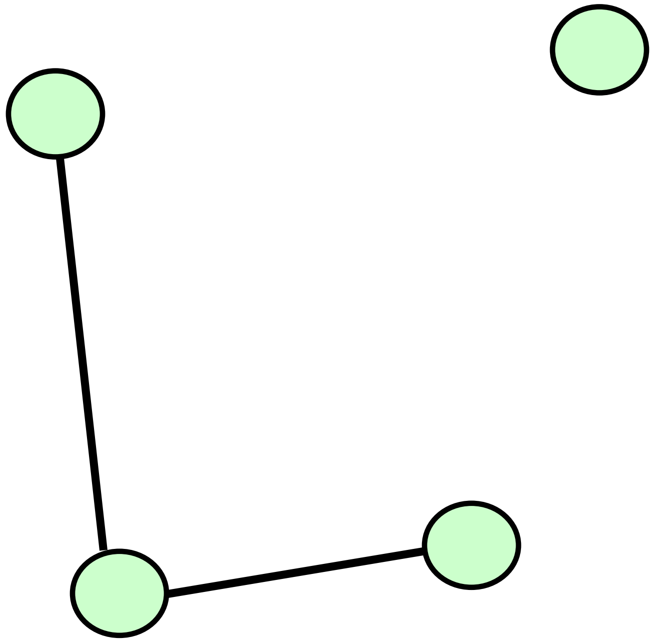
Monotonicity of β_0

The deletion of edge increases the the number of connected components by at most 1.
β_0 increases by 0 or 1.

Case 1

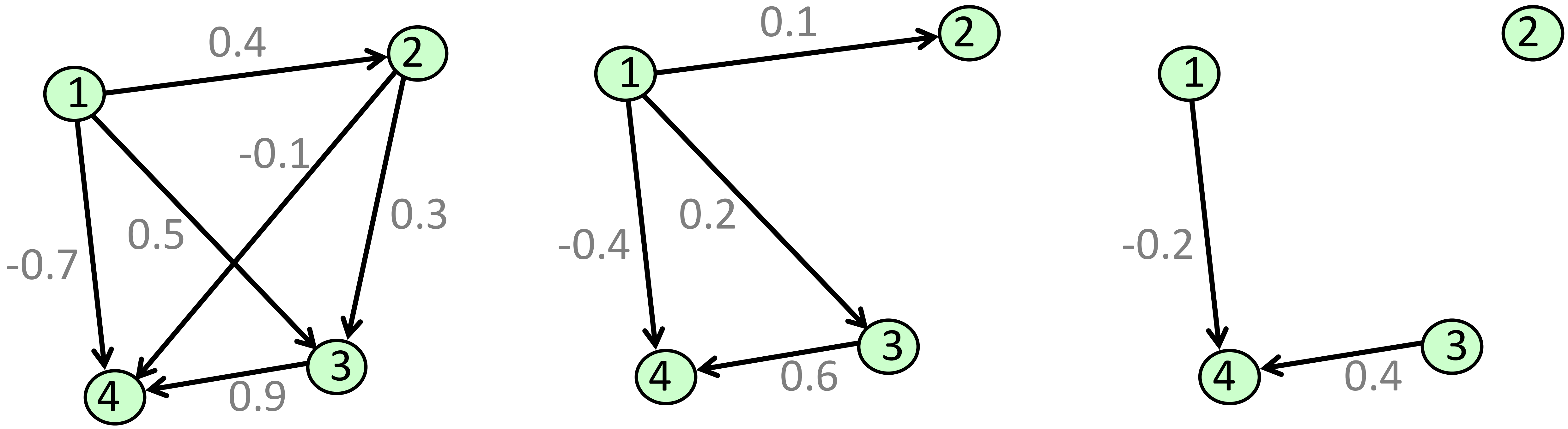


Case 2



Proof: Chung et al. 2019 Network Neuroscience

Graph filtration on directed graphs



$G(0)$ É $G(0.3)$ É $G(0.5)$

Building persistent homology on directed graphs is not trivial and important → Research project

How to compute the number of cycles in big network data?

Radeon Vega64 eGPU

How many cycles in the network?

Monotonicity of Betti-1 plot

Monotonicity of β_1:

The deletion of edge (in the filtration download) decreases the the number of cycles by at most 1.

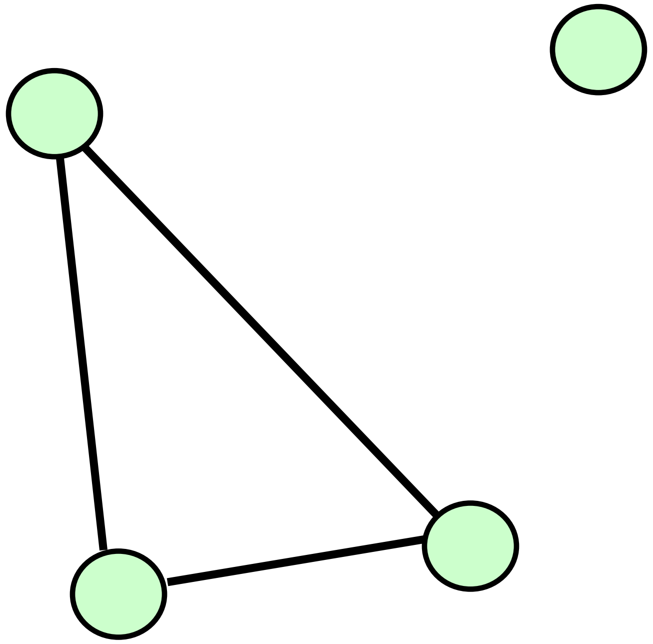
β_1 decreases by 0 or 1.

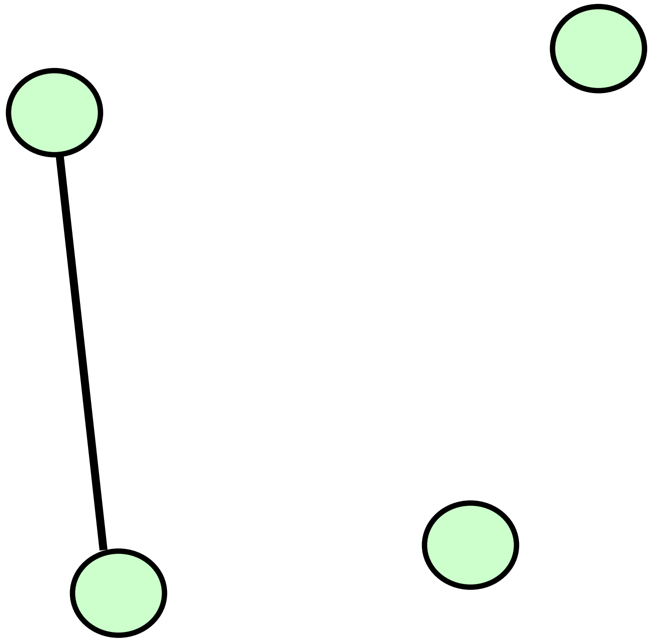
Euler characteristic for 1-skeleton:

$$\chi = \beta_0 - \beta_1 = \underset{\substack{\uparrow \\ \text{nodes}}}{p} - \underset{\substack{\nwarrow \\ \text{edges}}}{q}$$

$$\underset{\substack{\uparrow \\ -1,\ 0}}{\beta_1} = \underset{\substack{\uparrow \\ 0,\ +1}}{\beta_0} - \underset{\substack{\uparrow \\ \text{fixed}}}{p} + \underset{\substack{\uparrow \\ -1}}{q}$$

How Betti numbers change over downward graph filtration





$\beta_0 = 2$

$$\beta_0 - \beta_1 = 1$$

$\beta_1 = 1$

$\beta_0 = 3$

$$\beta_0 - \beta_1 = 3$$

$\beta_1 = 0$

$p = 4$

$$p - q = 1$$

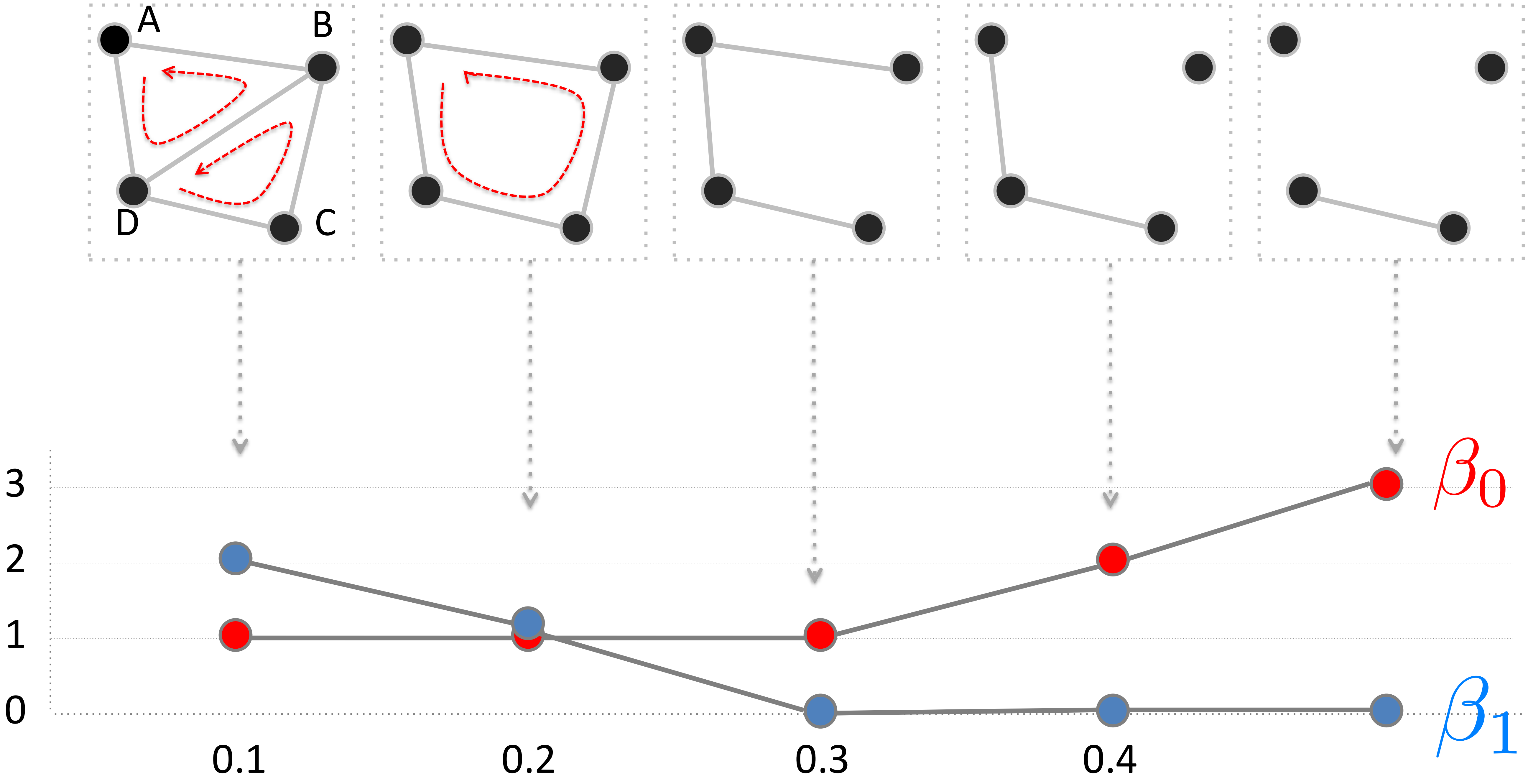
q = 3

$p = 4$

$$p - q = 3$$

q = 1

Betti numbers over graph filtration



Theorem (monotonicity of Betti numbers) Betti plots over graph filtration are monotone.

Chung et al. 2019 ISBI

Computation of Betti-plots in practice

Computation of β_0: Many existing algorithms. Can use a built-in function in MATLAB.

```
                [beta_0, S] =
graphconncomp(adj)
```

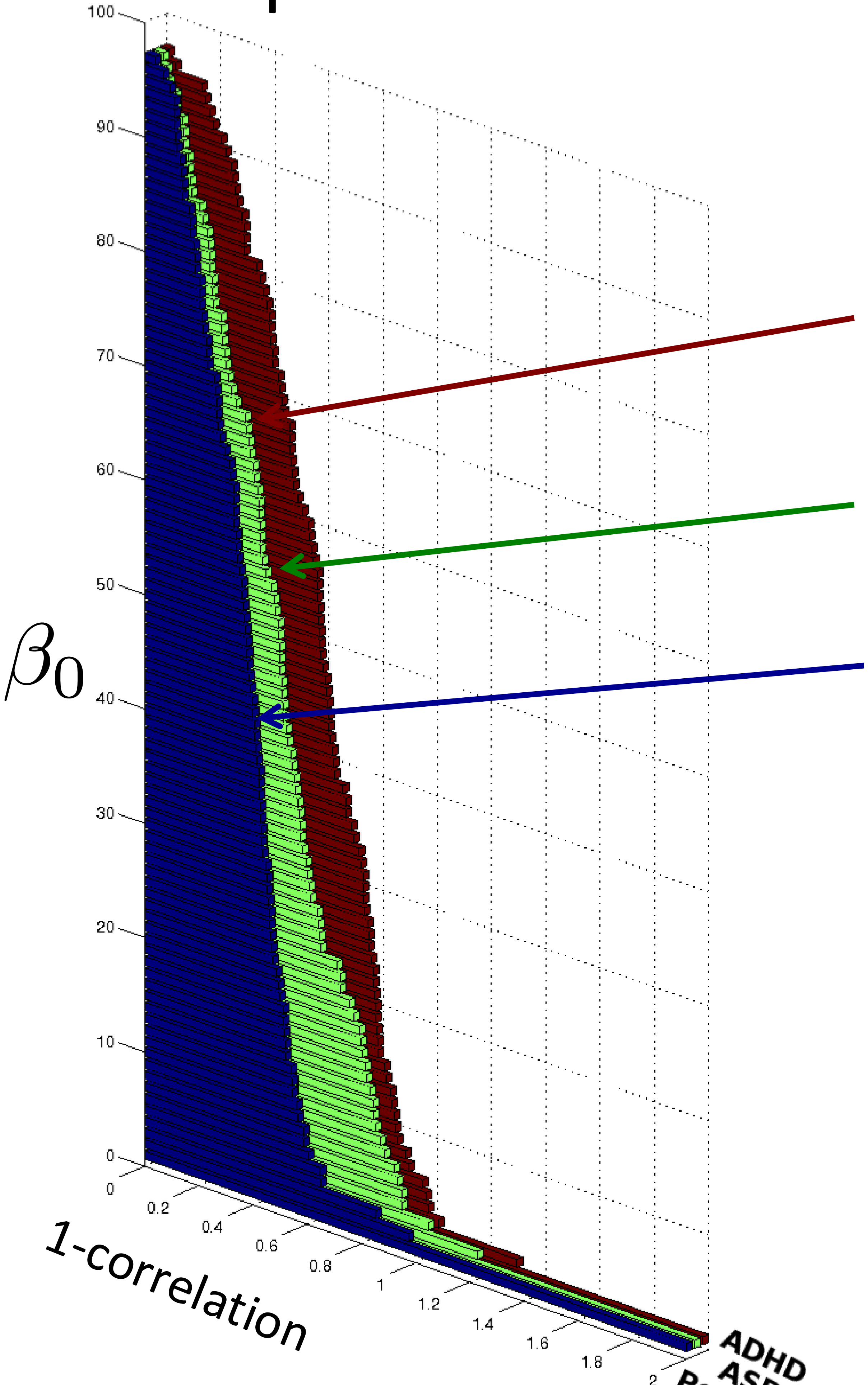
Computation of β_1: As a function of β_0

$$\beta_1 = \beta_0 - p + q$$

```
               q=sum(sum(adj))/2;
                  beta_1 = beta_0 - p + q;
```

<u>This is not efficient</u>. Need an incremental algorithm that updates as we delete one edge at a time.

0-th Betti plot on PET correlation network



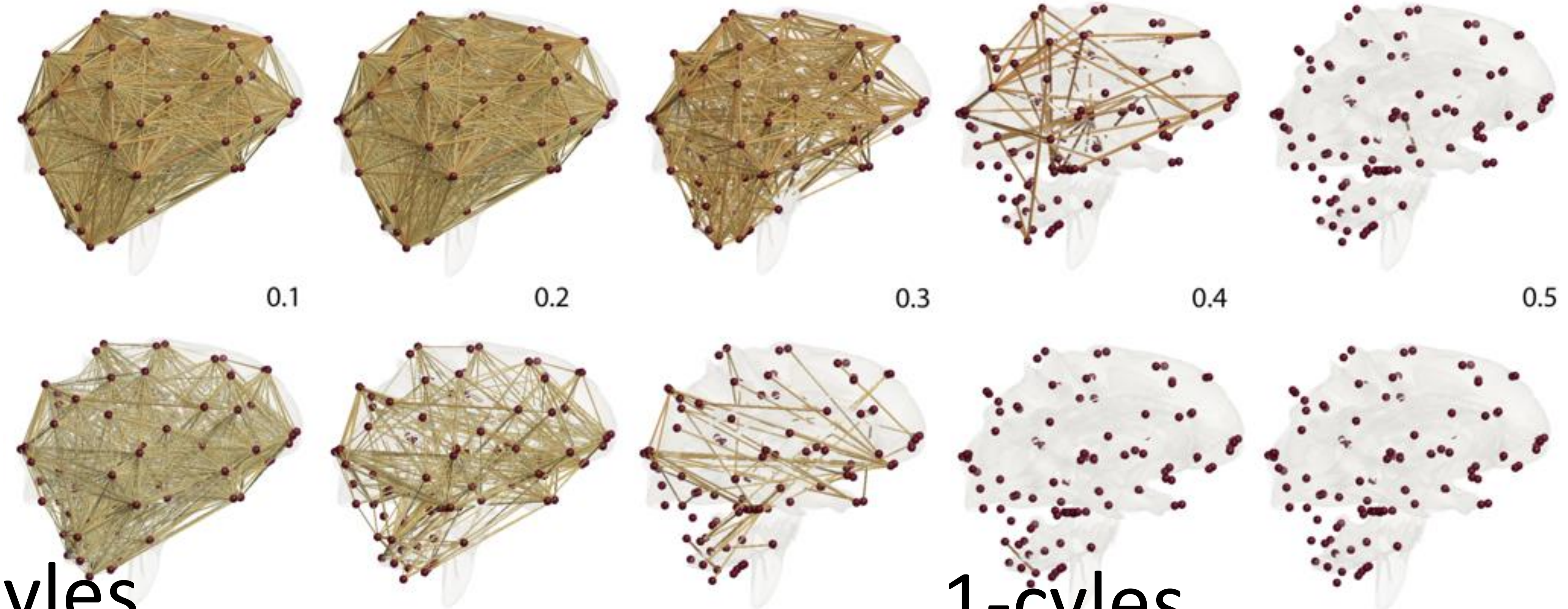
24 attention deficit hyperactivity disorder (ADHD) children

26 autism spectrum disorder (ASD) children

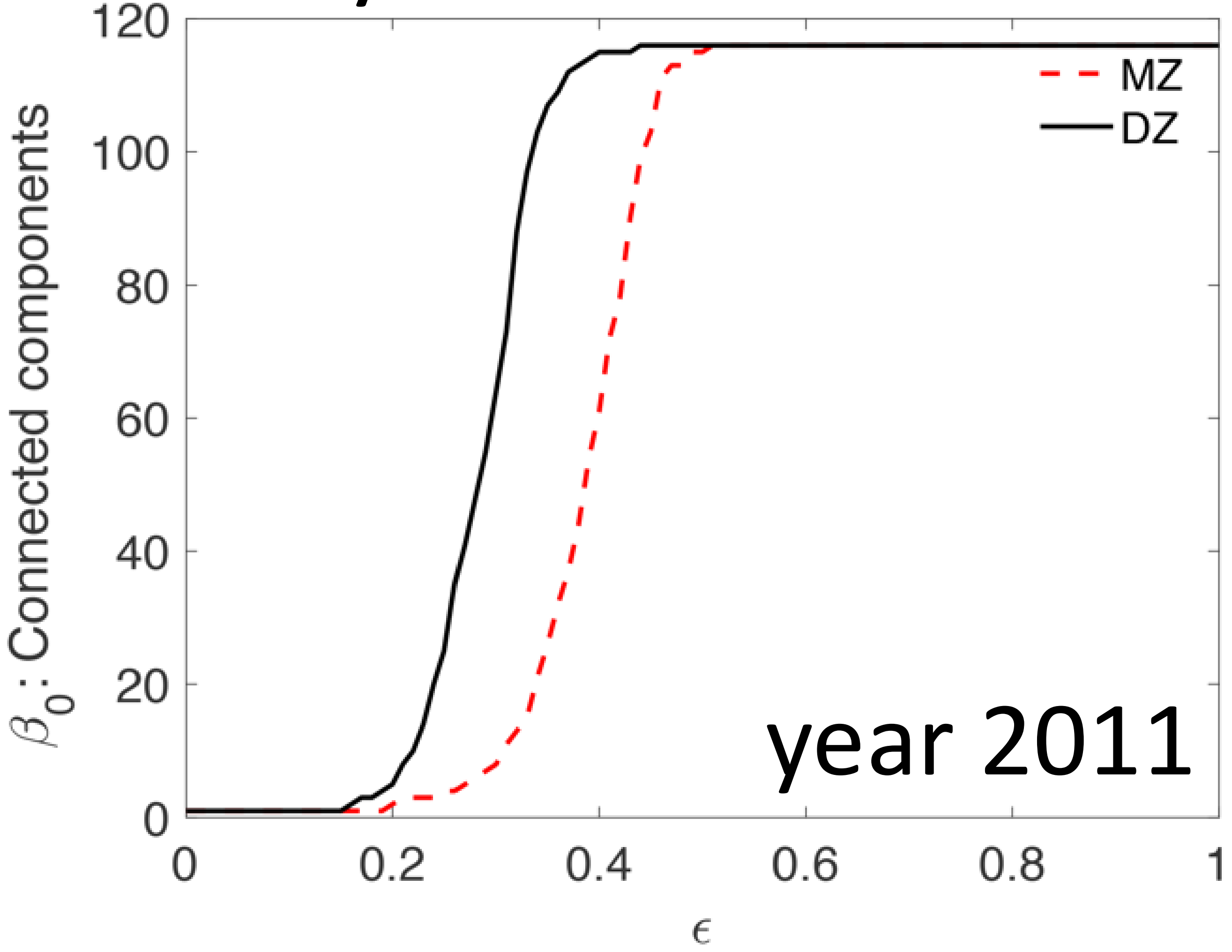
11 pediatric control subjects

Lee et al. (2011) ISBI

Betti-plots in 116 nodes network

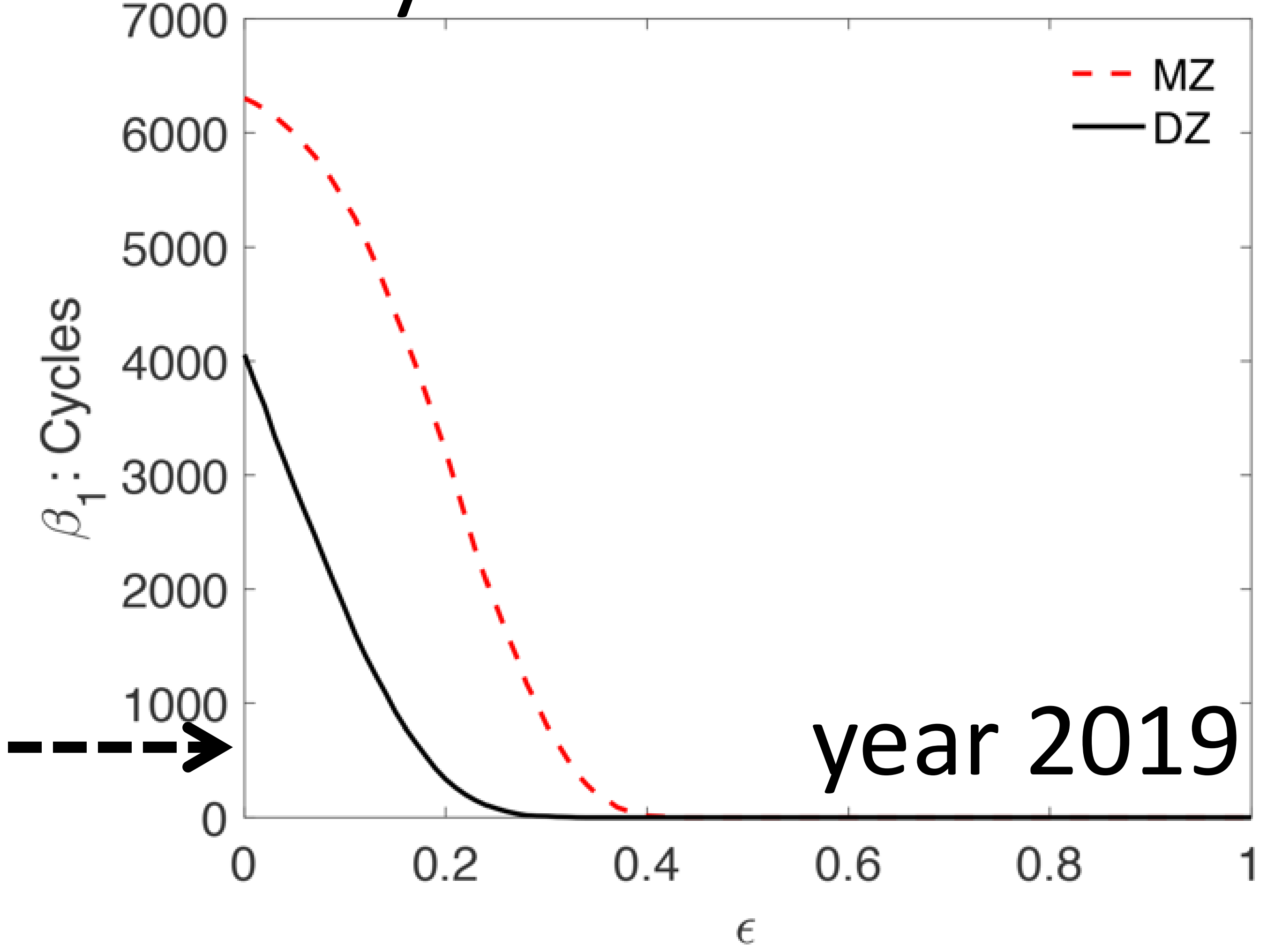


0-cyles



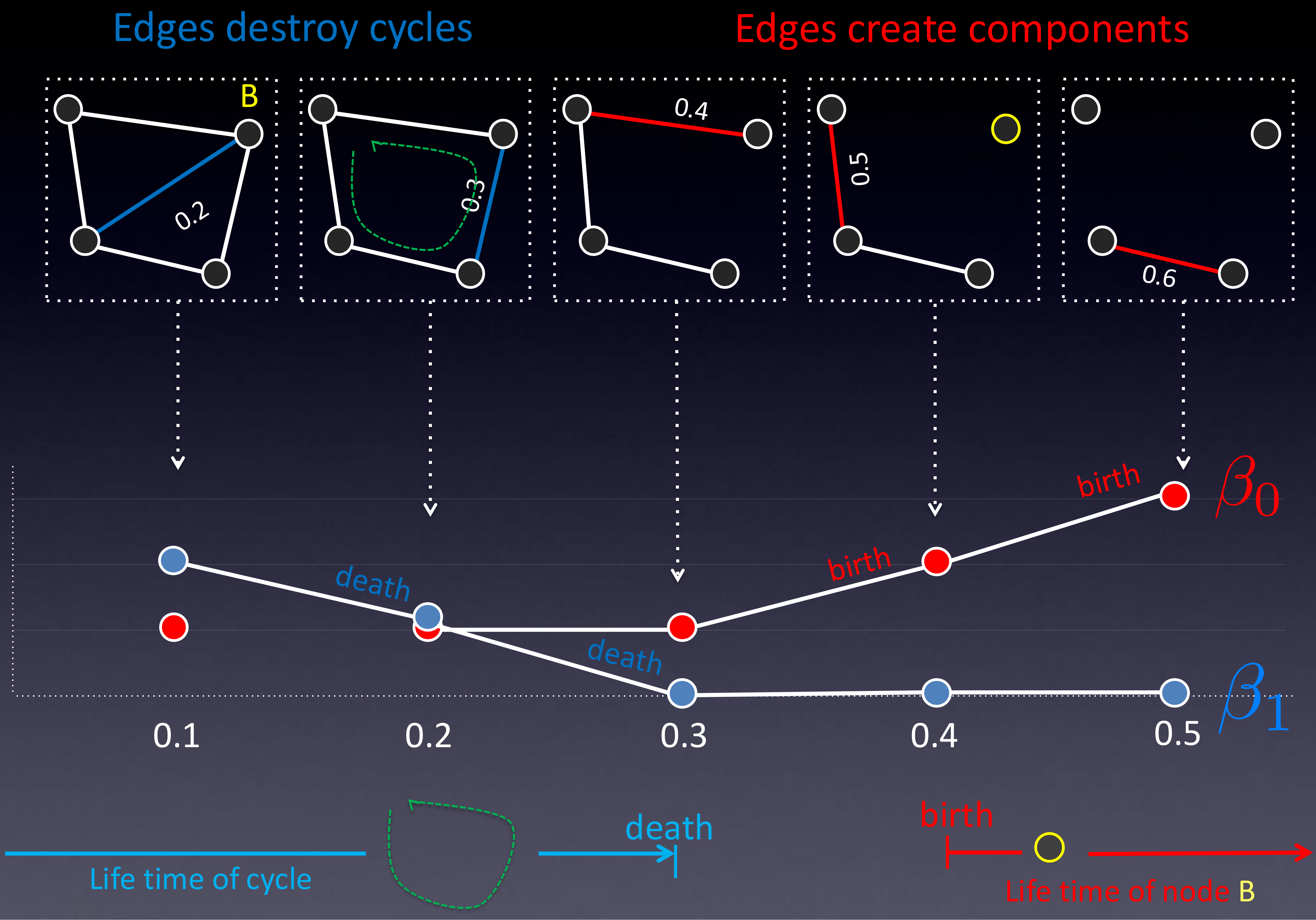
year 2011

1-cyles



year 2019

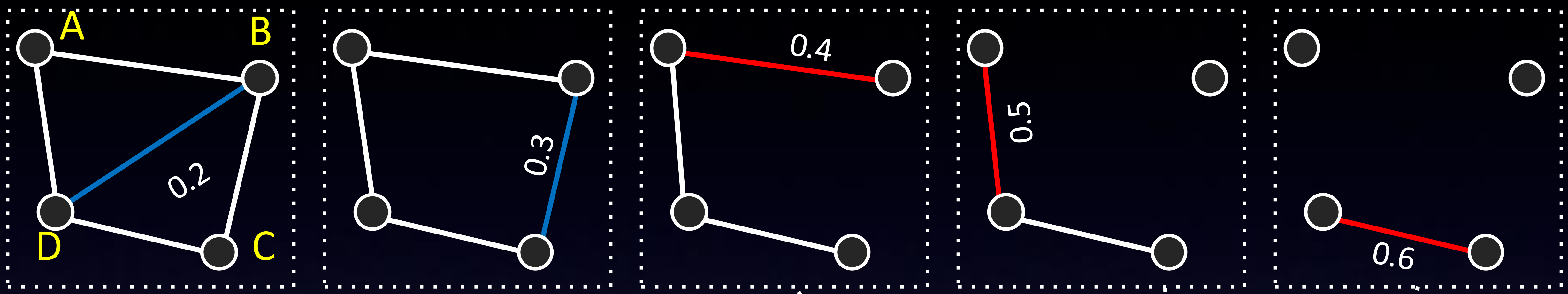
Persistence = Life time (death – birth) of a feature



Theorem *Birth & death sets partition* the edge set

E_1 Edges destroy cycles

E_0 Edges create components

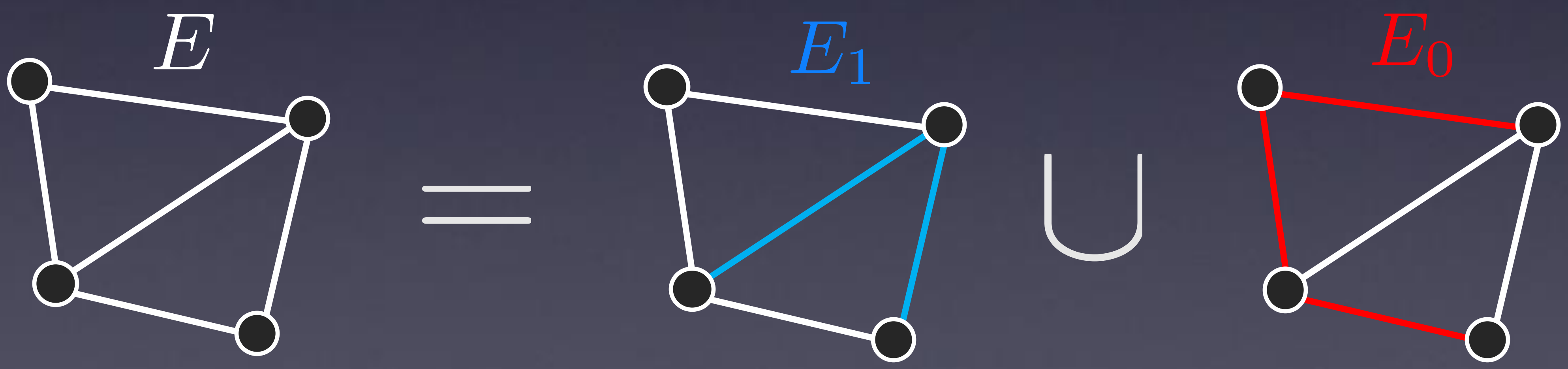


$$\#(E_1) = 1 + \frac{|V|(|V|-3)}{2}$$

$$\#(E_0) = |V| - 1$$

$$\#(E) = \frac{|V|(|V|-1)}{2}$$

Maximum spanning tree



$O(|E| \log |V|)$

Songdechakraiwut & Chung 2020 arXiv: 2012.00675

codes

Betti plots
http://brainimaging.waisman.wisc.edu/%7Echung/barcodes

Exact topological inference
http://www.stat.wisc.edu/~mchung/TDA

Paired image (twin, longitudinal) related TDA
http://pages.stat.wisc.edu/~mchung/twins

Morse Filtration

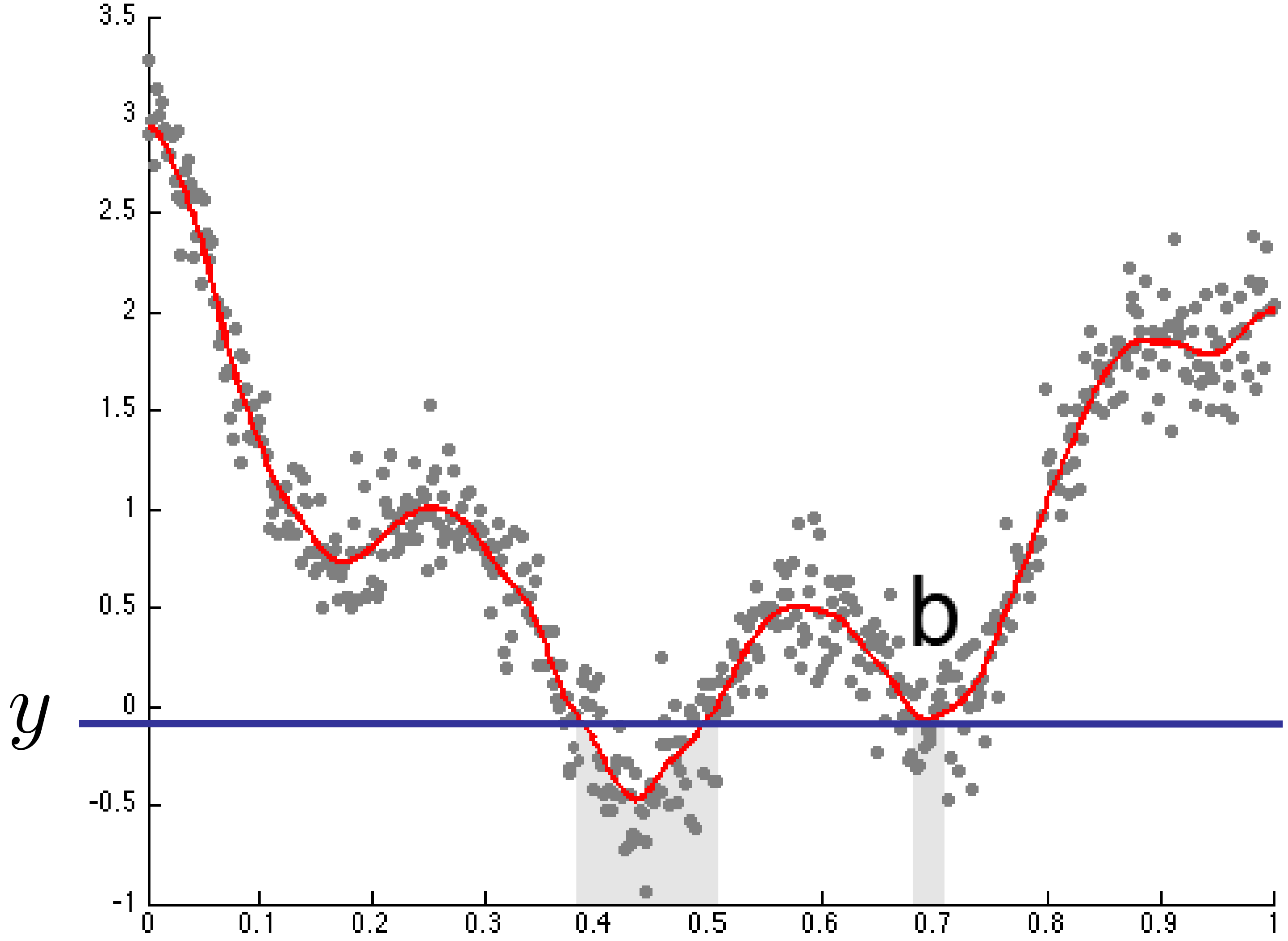
Most useful in functional and time series data

Morse theory for functional data

$$Y = \mu + \epsilon$$

Chung et al., 2009
Information Processing in Medical Imaging (IPMI) 5636:386-397.

Unknown signal μ is assumed to be a Morse function: all critical values are unique.

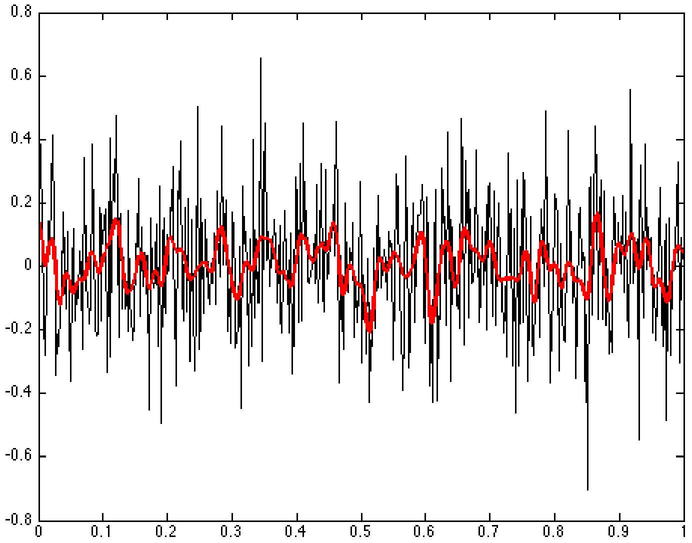


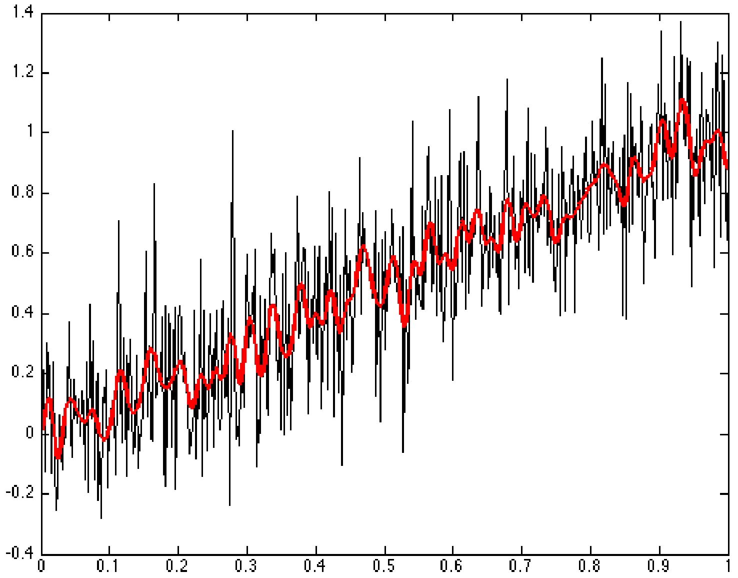
Sublevel set

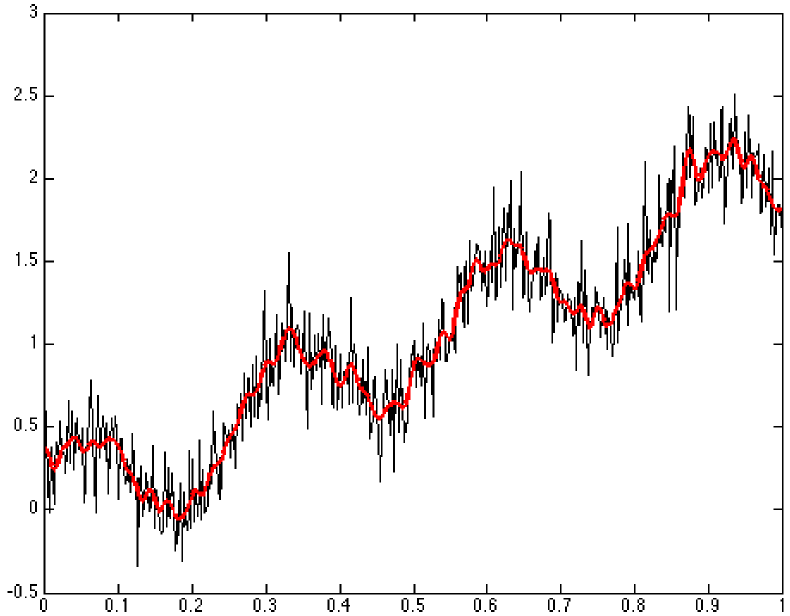
$$R(y) = \mu^{-1}(-\infty, y]$$

Number of connected components $\#R(y)$

Critical values capture the pattern of signal changes

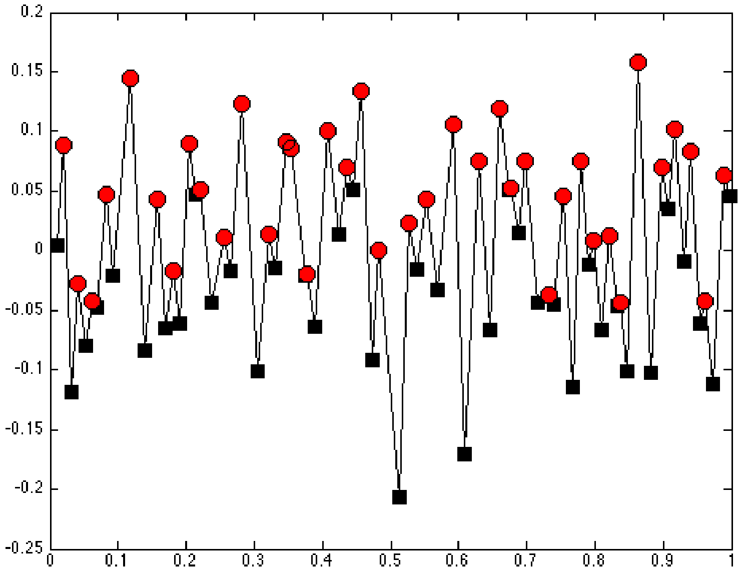


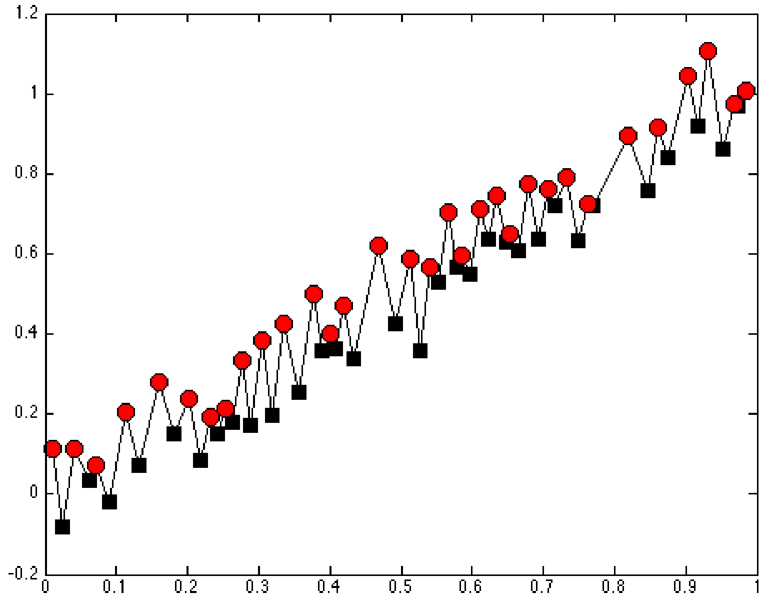


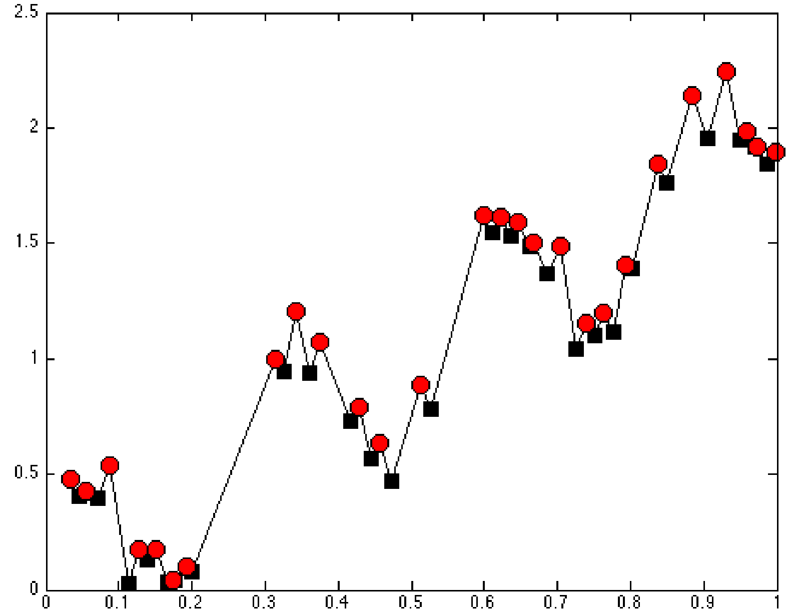


$$f(t) = e(t)$$

$$f(t) = t + e(t)$$



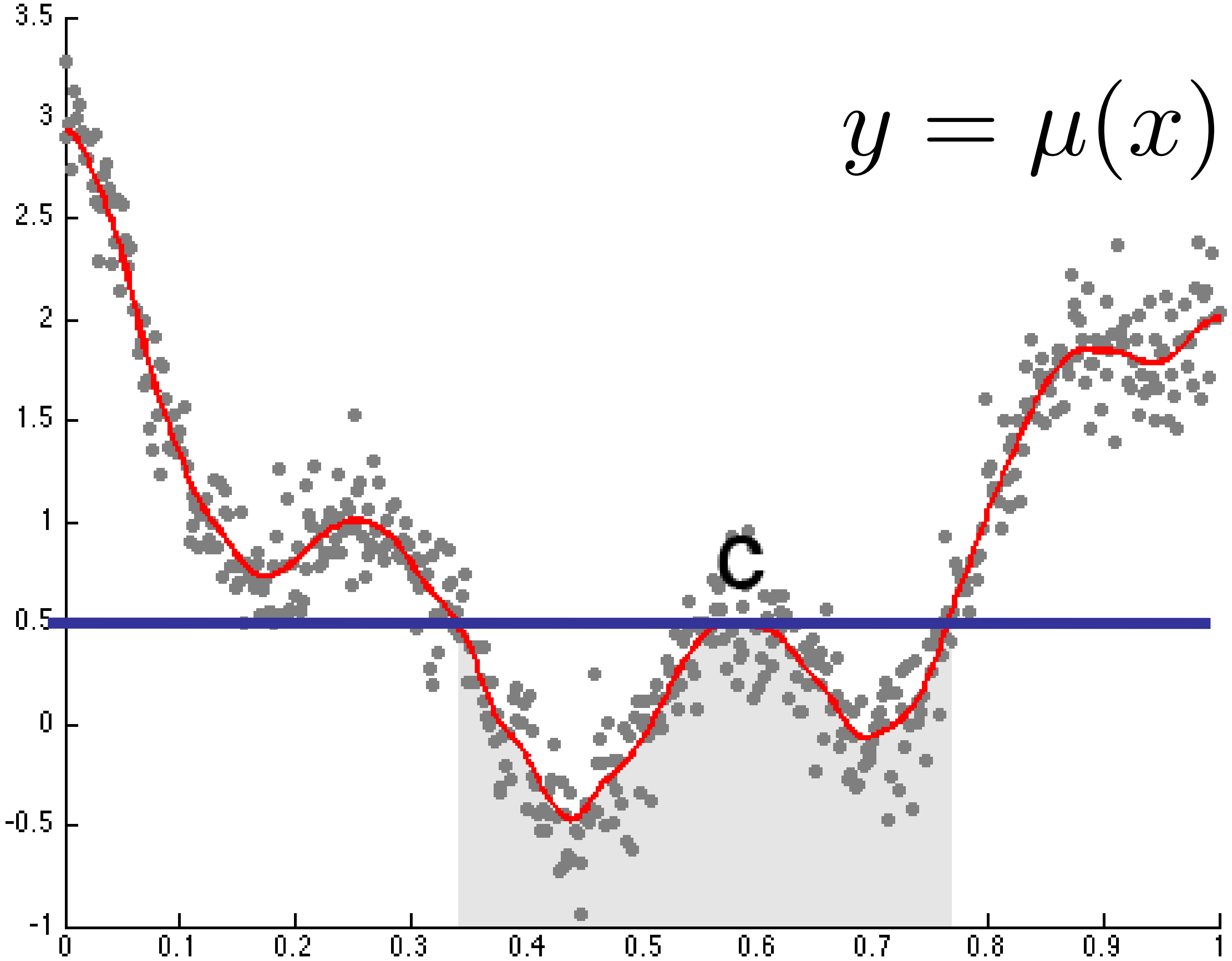




Morse filtration

Consider a sublevel set

$$R(y) = \mu^{-1}(-\infty, y]$$



For critical values

$$b < c$$

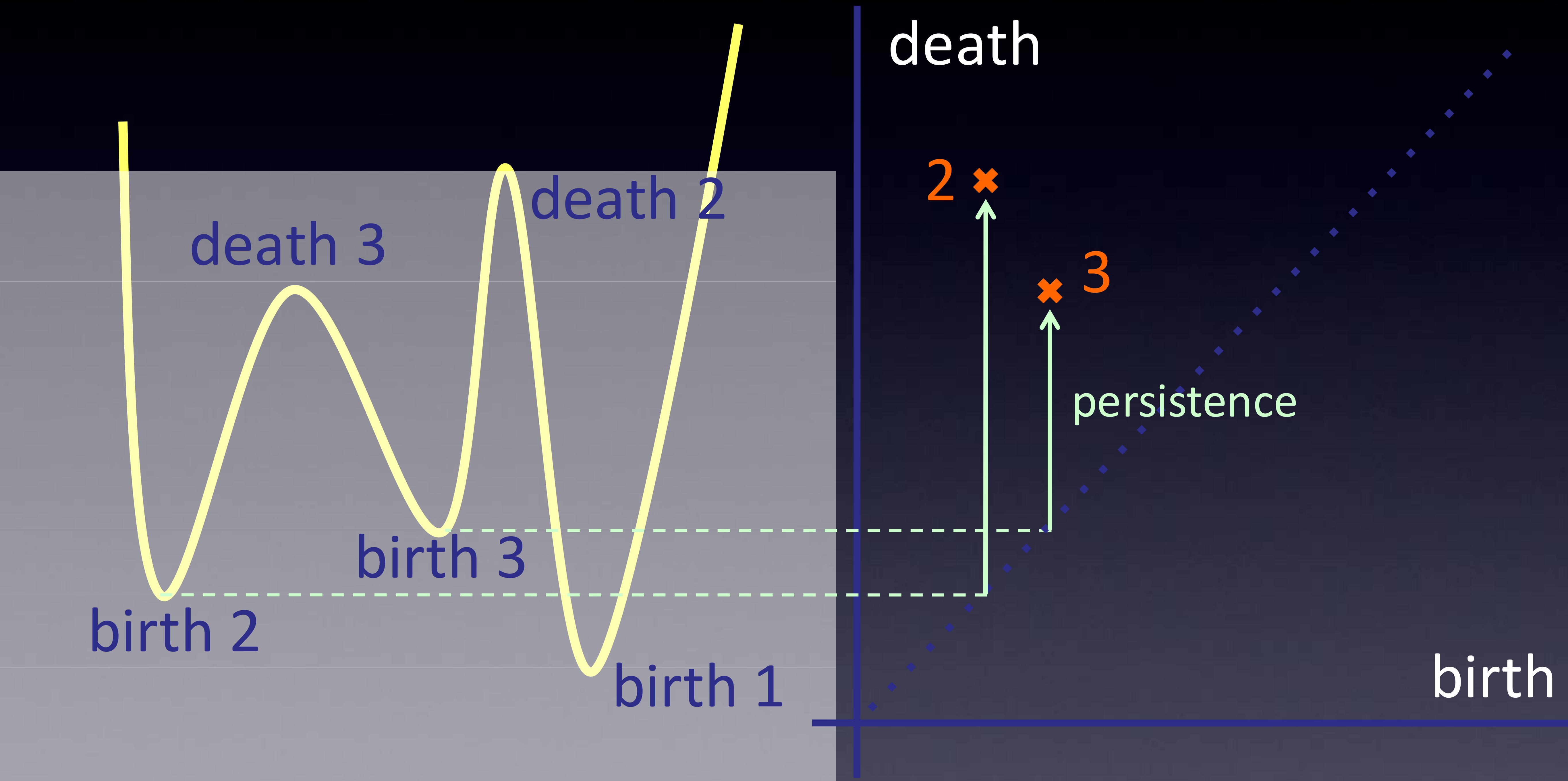
$$R(b) \subset R(c)$$

# of connected components

$$\#R(b) = \#R(c) - 1$$

Persistence Diagram (PD)

O(*n* log *n*)



Pair the time of death with the time of the closest earlier birth.
Birth 1 is paired to infinity or ignored.

Pairing Brackets

(((())(()(() (() ())) ()()()()((((())))

(((())(()(() (() ())) ()()()()((((())))

(((())(()(() (() ())) ()()()()((((())))

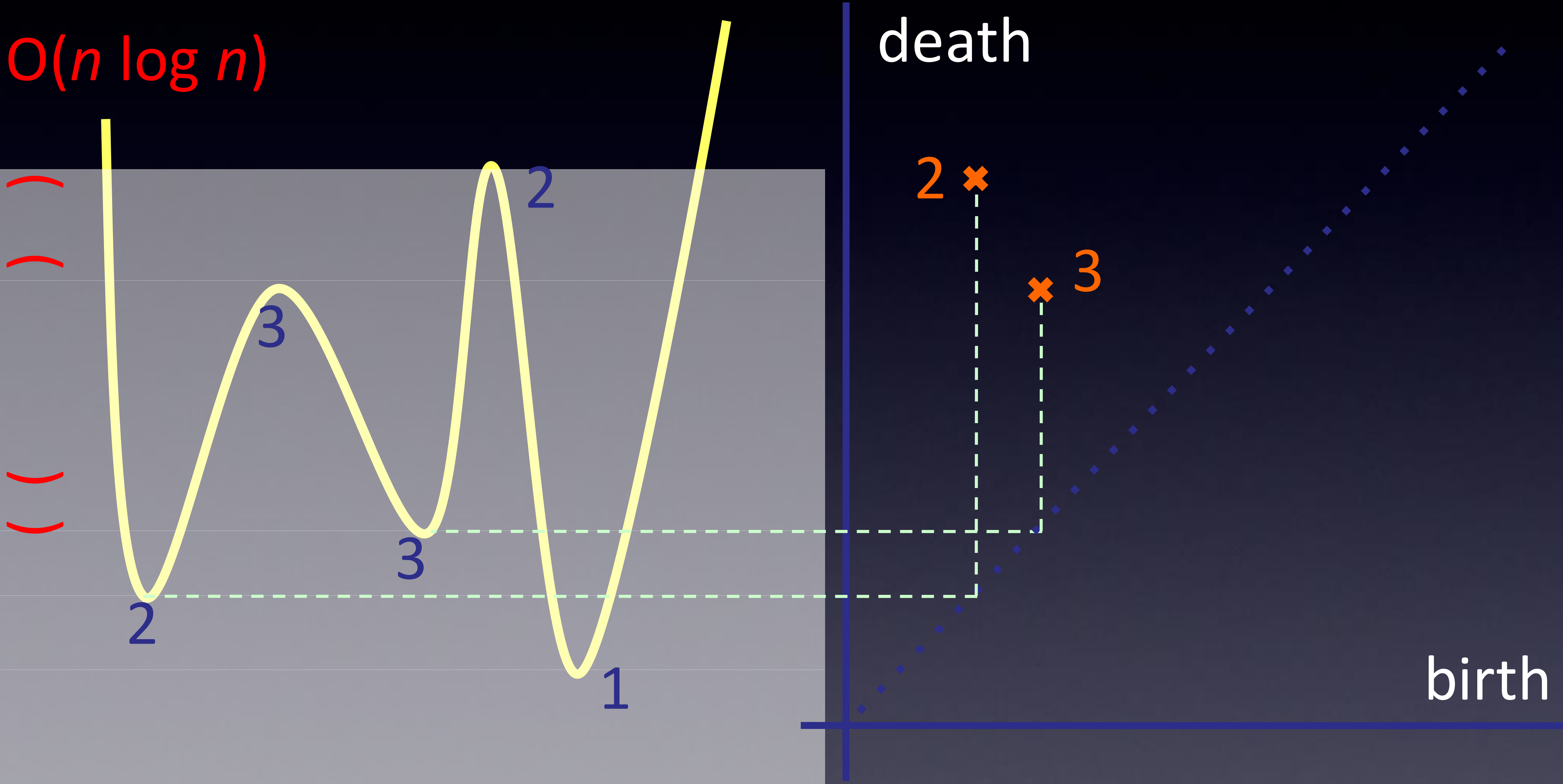
(((   )(()(() (() ())) ()()()()((((())))

(((   )(()(() (() ())) ()()()()((((())))

(((()(() (() ())) ()()()()((((())))

(((()(() (() ())) ()()()()((((())))

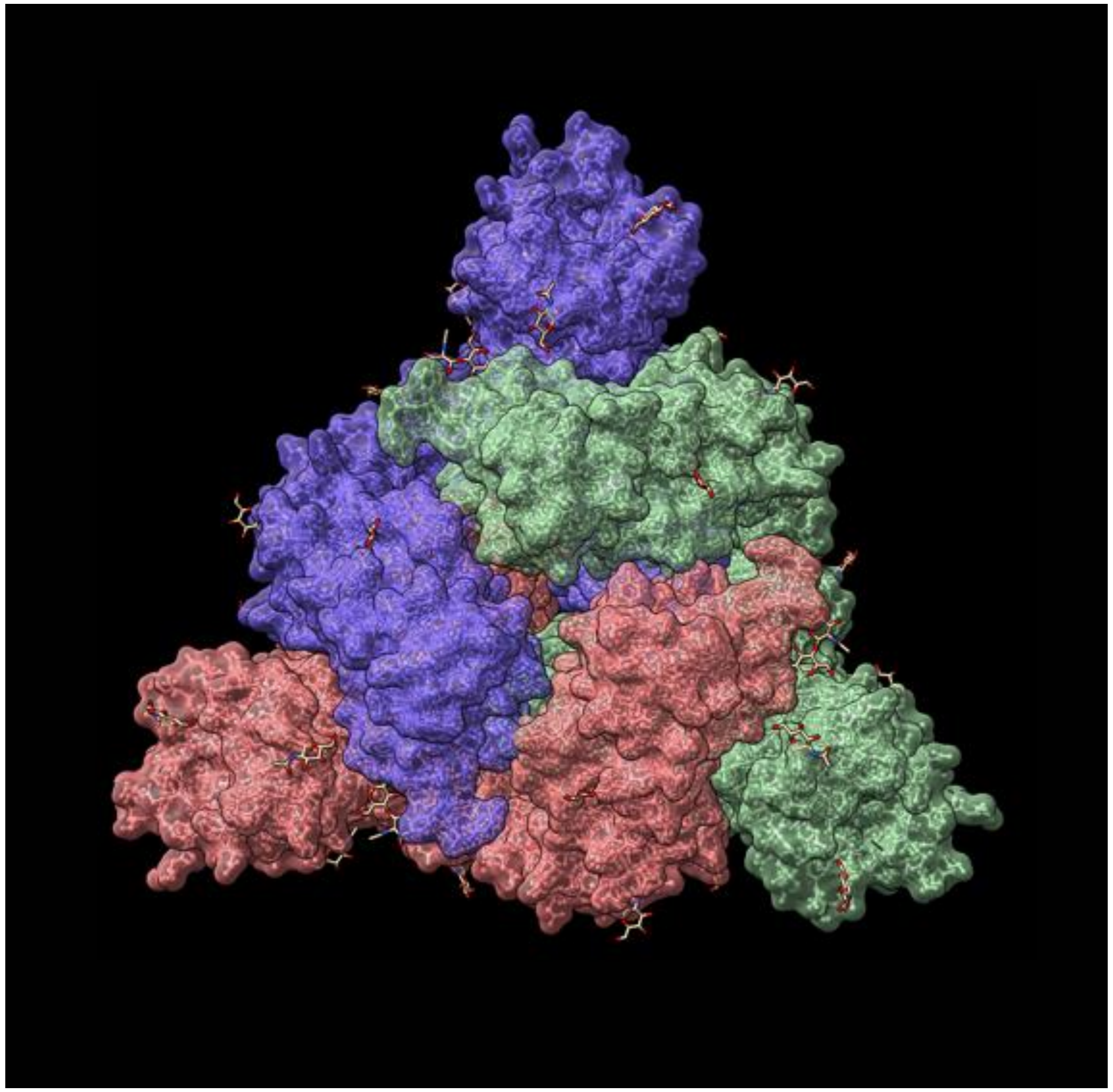
Persistence Diagram (PD)

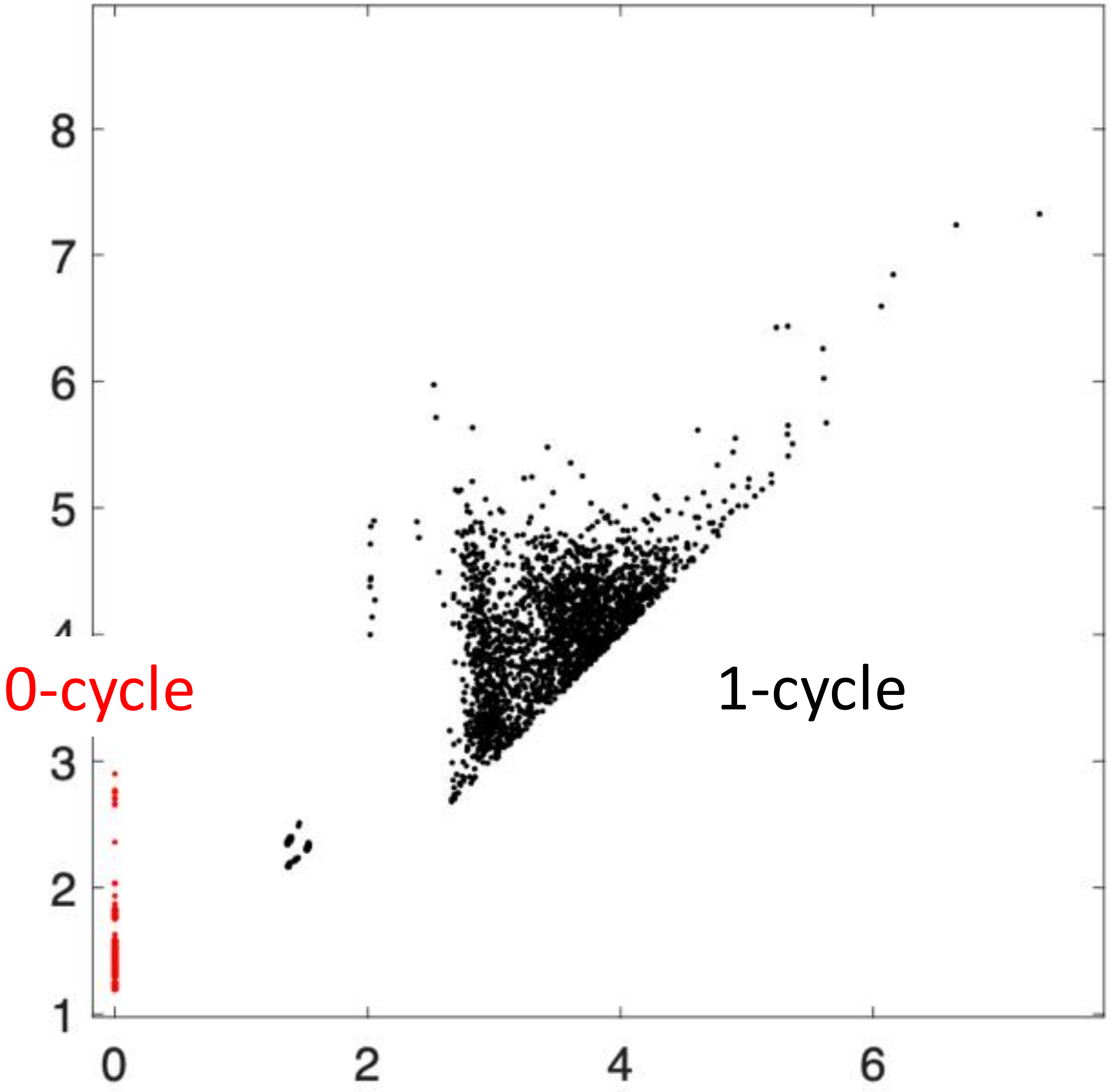


Pair the time of death with the time of the closest earlier birth

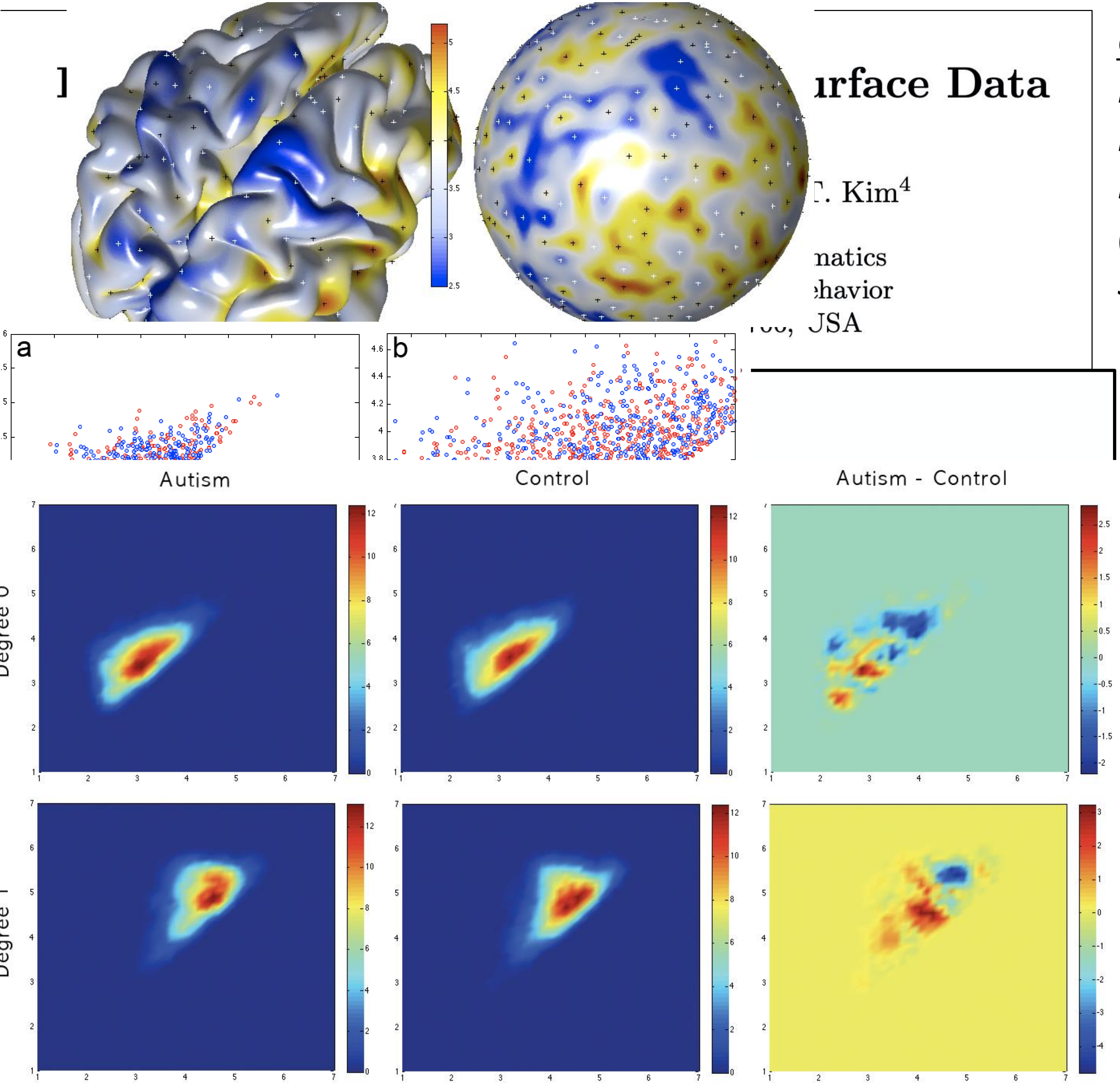
Persistence Diagram (PD) of a protein r

Rips filtration on distance between 8000 atoms





Extremely slow computation → Simply use graph filtration



]

urface Data

T. Kim^4

natics

ehavior

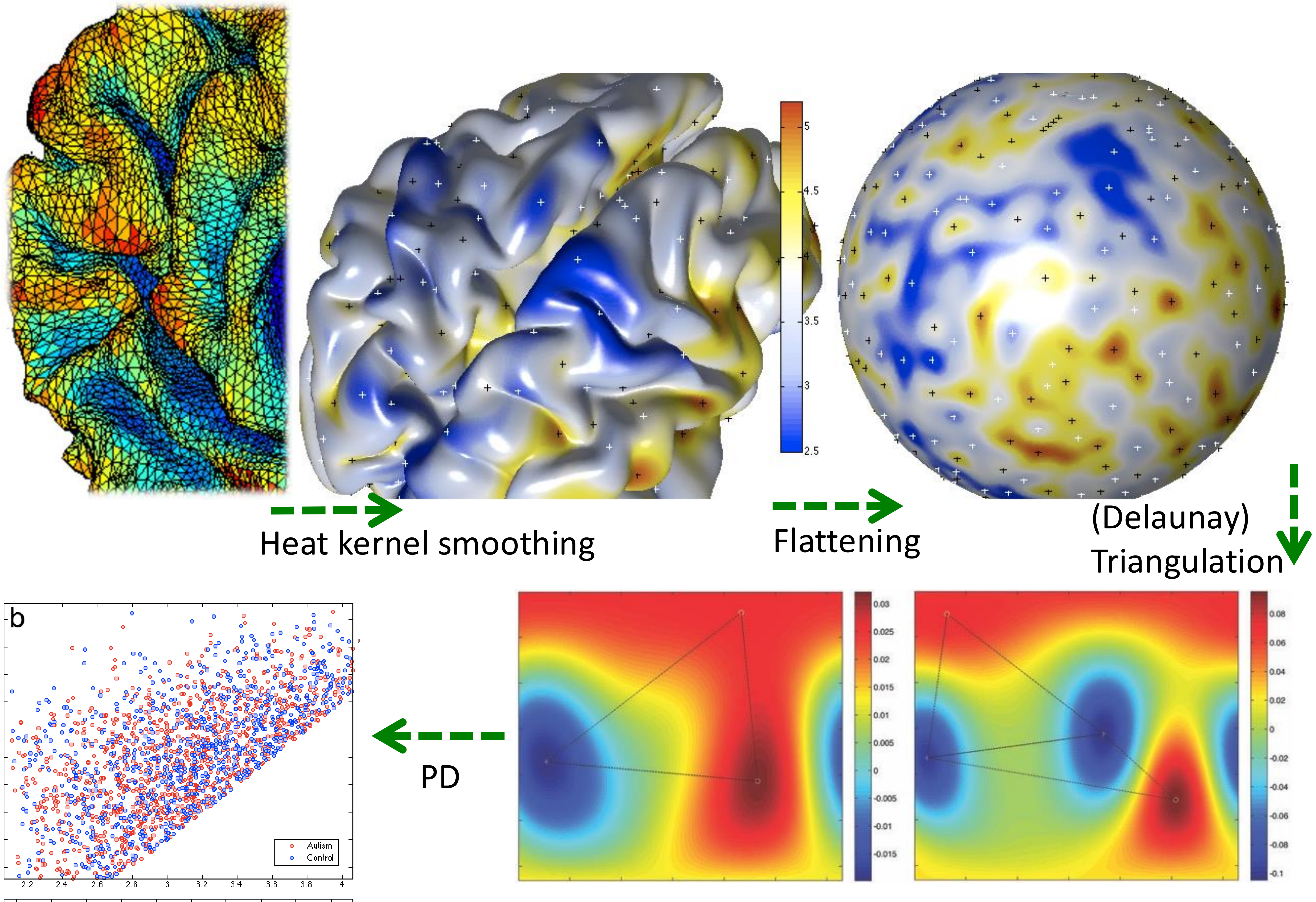
USA

Chung et al., 2009

Information Processing in Medical Imaging (IPMI) 5636:386-397.

First persistent homology study applied to medical imaging

Persistent homology on cortical manifolds



Heat kernel smoothing

Flattening

(Delaunay) Triangulation

PD