The Waisman Laboratory for Brain Imaging and Behavior

University of Wisconsin
SCHOOL OF MEDICINE AND PUBLIC HEALTH

Lecture 1 Simplicial homology and persistent homology

Moo K. Chung
Department of Biostatistics and Medical Informatics
University of Wisconsin-Madison

https://github.com/laplcebeltrami/ISBI2023TDA

Acknowledgement

Zijian Chen, Vijay Anand D, Sixtus Dakurah, Soumya Das, Tananun Songdechakraiwut, Univ. of Wisconsin-Madison, US

Shih-Gu Huang, Anqi Qiu National University of Singapore, Singapore

Anass El Yaagoubi Bourakna, Hernando Ombao KAUST, Saudi Arabia

Hyekyung Lee, Dong Soo Lee Seoul National University, Korea

Grants: *NIH R01 EB022856, R01 EB028753, NSF DMS-2010778*

References

Gunnar Carlsson 2009, A User's Guide to Topological Data Analysis

Herbert Edelsbrunner and John L. Harer
Computational Topology: An Introduction 2010, American Mathematical Society

Chung et al. 2020 Review: Toplogical distance and losses in brain networks arXiv:2102.08623

Matlab toolbox

PH-STAT

Statistical Inference on Persistent Homology

https://github.com/laplcebeltrami/PH-STAT

Topological Data Analysis

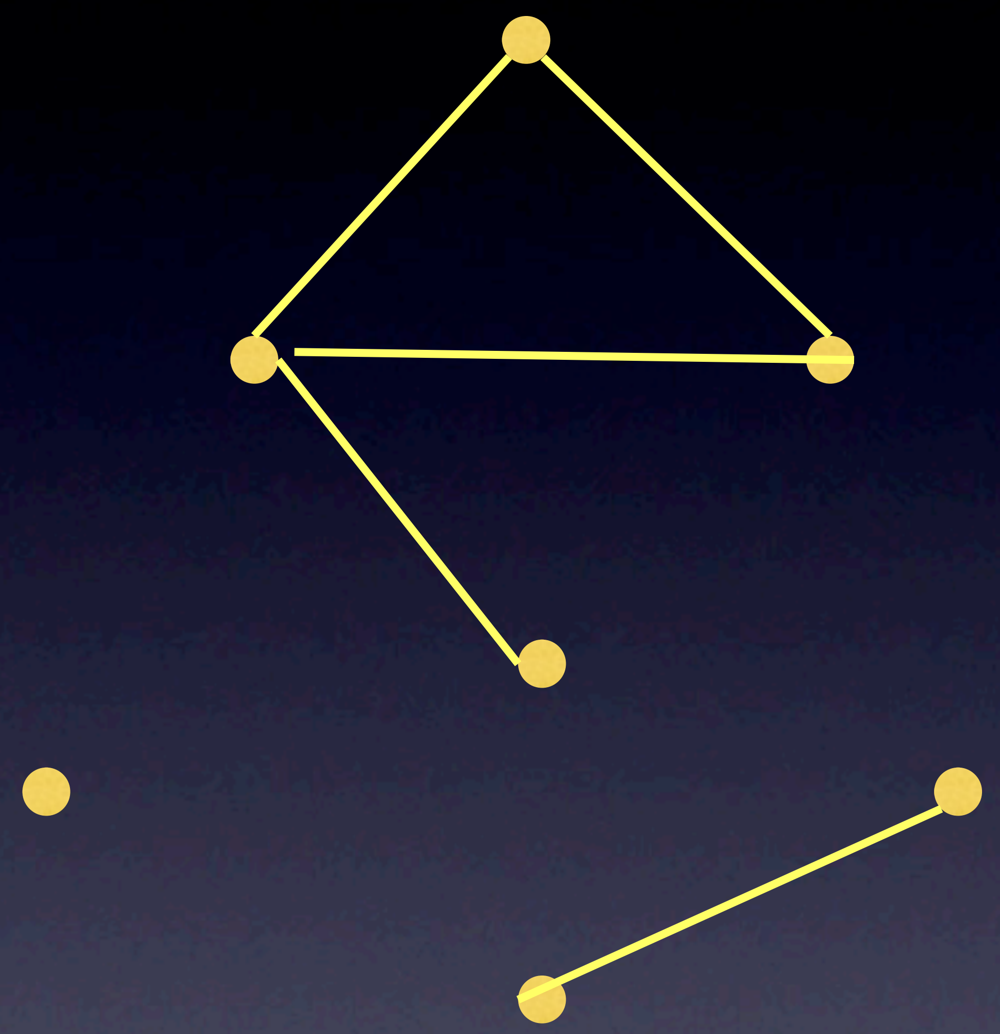
- Branch of data science that uses topology
- Study properties of data that remain <u>invariant</u> under continuous transformations
- Identify underlying patterns using topology

Persistent Homology

- Captures the topological features of data across different scales
- More persistent topological features → signal
- Less persistent topological features → noise
- Multiscale approach

Betti numbers β_i

of i-dimensional holes/loops

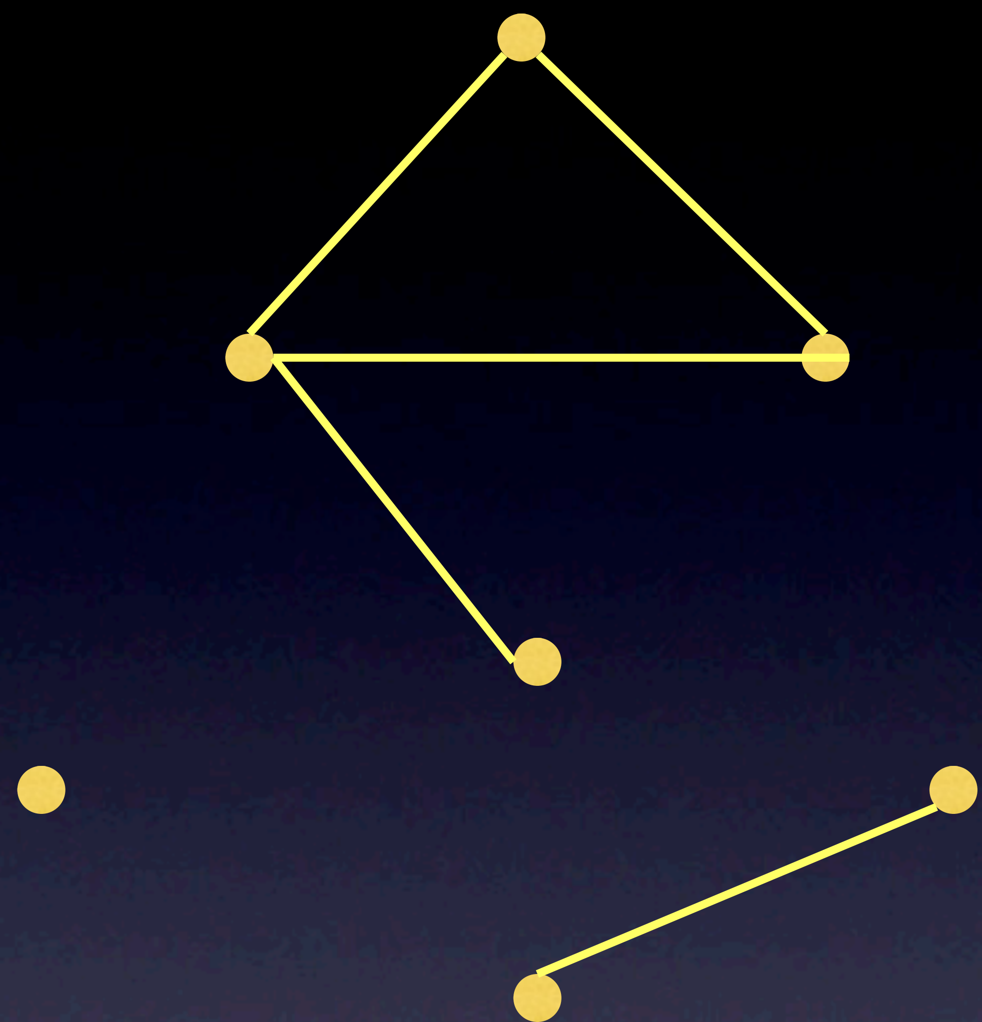


β_0 = # of connected components = 3

β_1 = # of cycles = 1

Euler characteristic: $\chi = 3 - 1 = 2$

Betti numbers β_i # of i-dimensional holes/loops



β_0 = # of connected components = 3

β_1 = # of 1D holes = 1

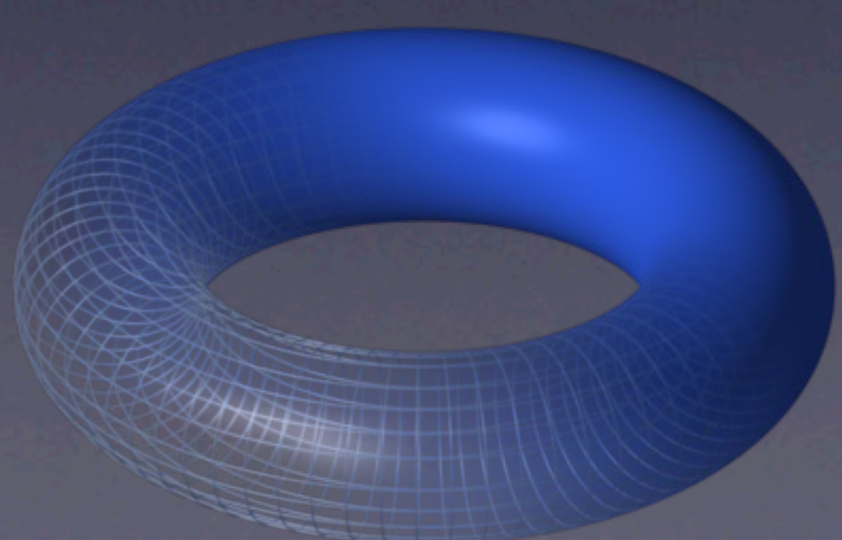
β_2 = # of 2D cavities =0

Betti-number representation:

(3,1,0,0,...)

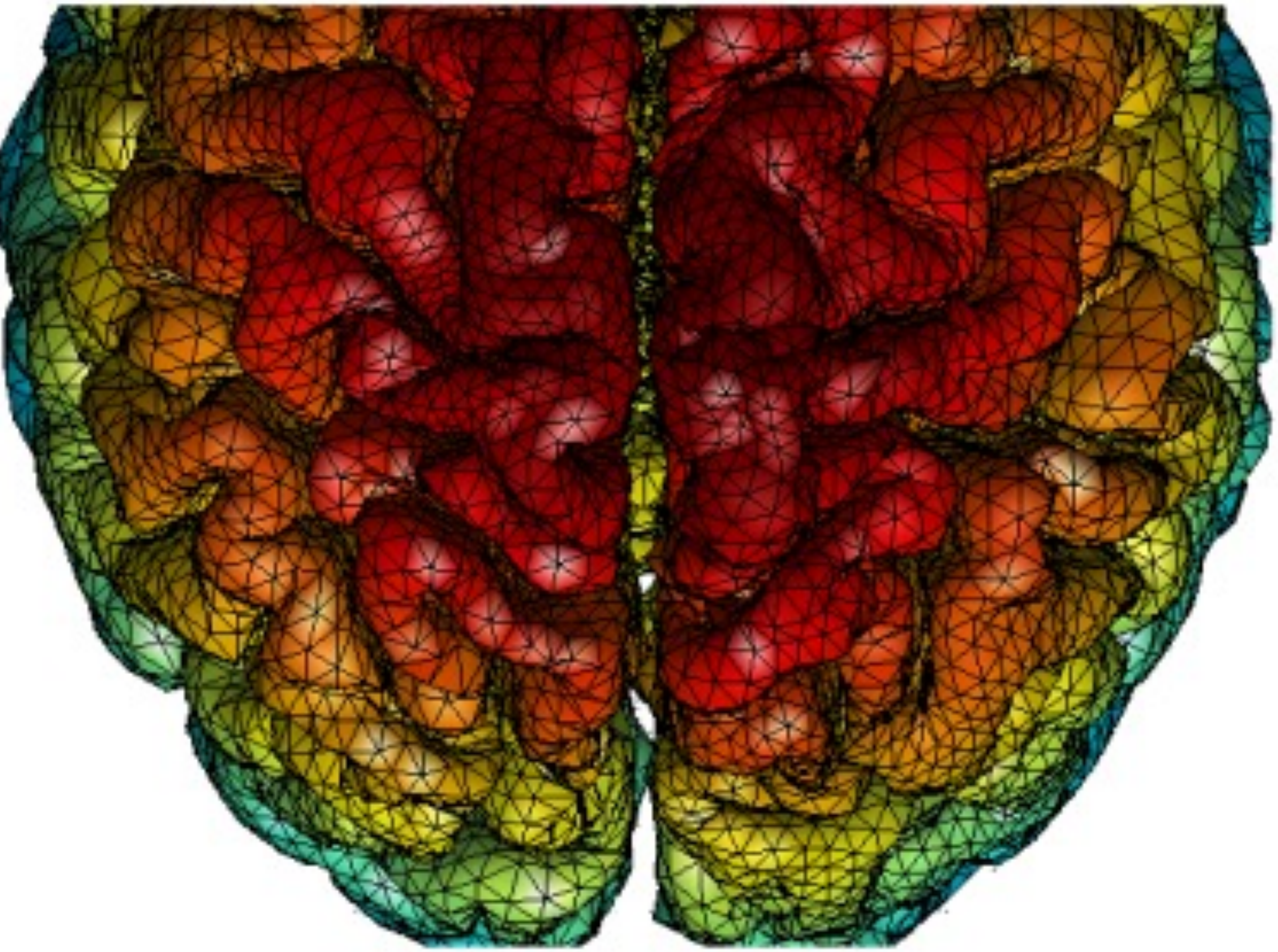
Euler characteristic:

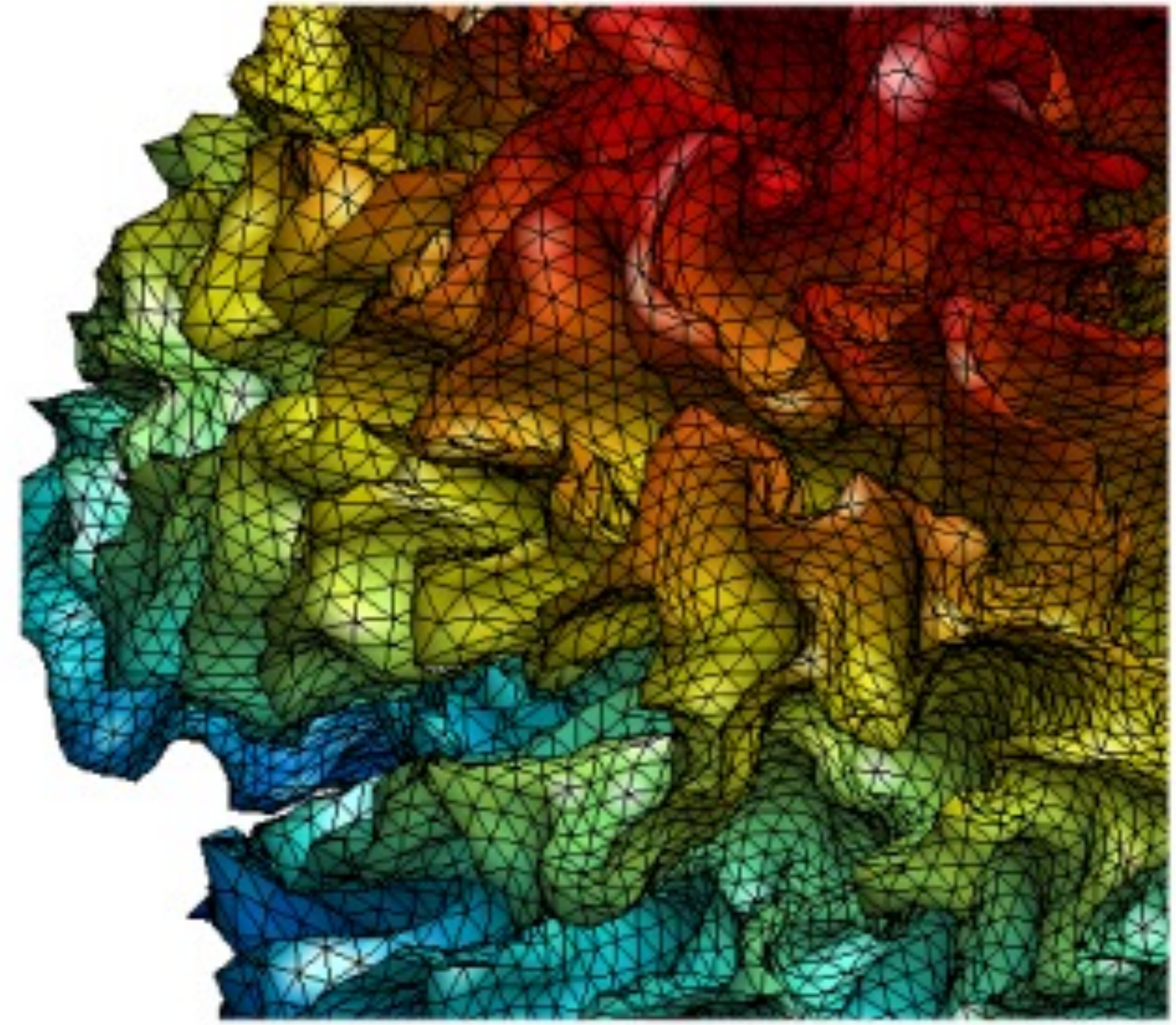
$$\chi = \beta_0 - \beta_1 = 2$$



$$\beta_0 = 1, \beta_1 = 2, \beta_2 = 1$$

(1,2,1,0,0,...)

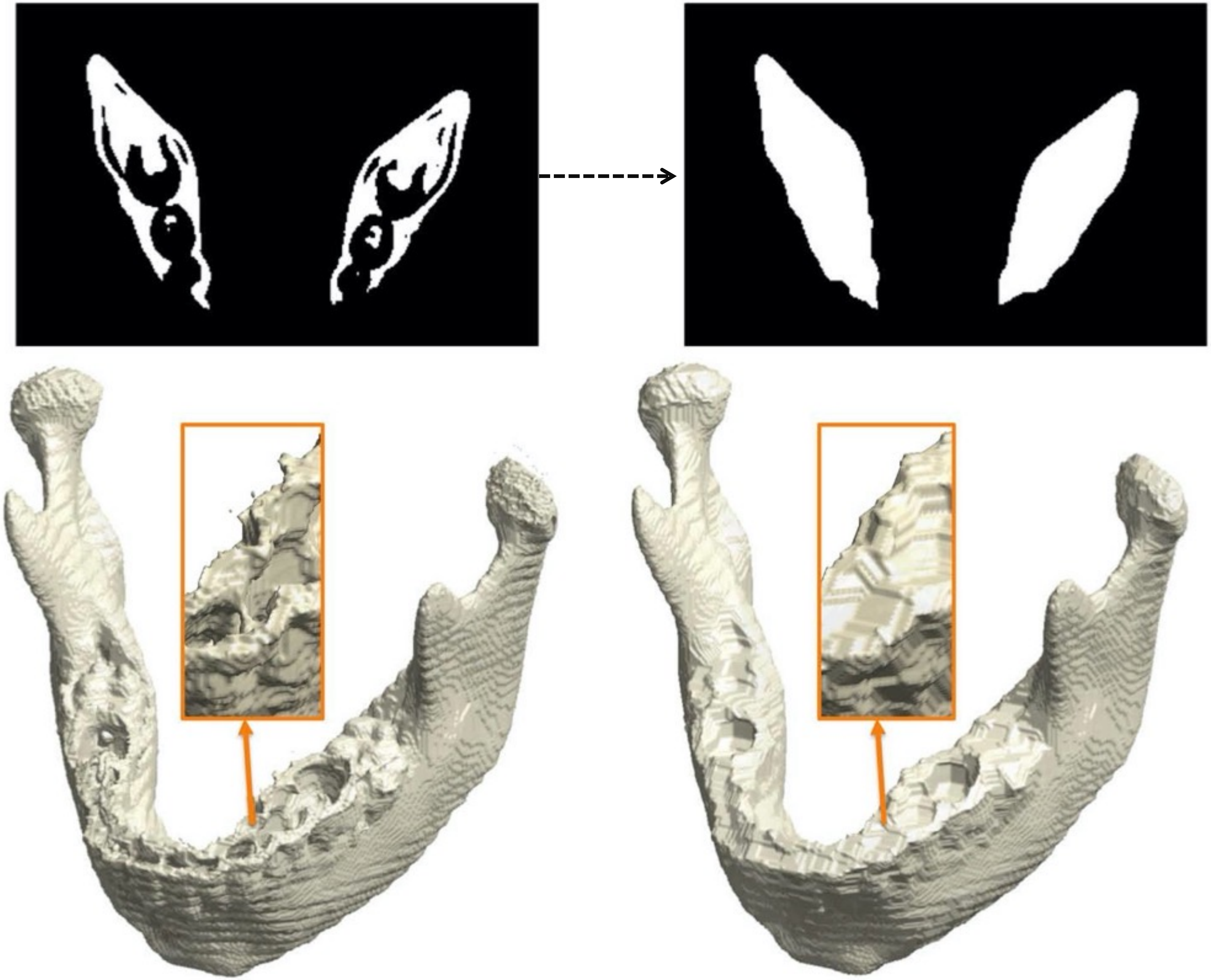




Euler characteristic of a surface mesh from SurfStat

N - E +F = 2 for a surface topologically equivalent to a sphere. For each triangle, there are three edges. Since two adjacent triangles share the same edge, the total number of edges is E = 3F/2. Hence, we have F=2N – 4 for a closed surface.

Topology correction in CT segmentation



Hole & handles corrected using Euler characteristic

Chung et al. 2015 Medical Image Analysis. 22:63-76

Keith Worsley's random field theory

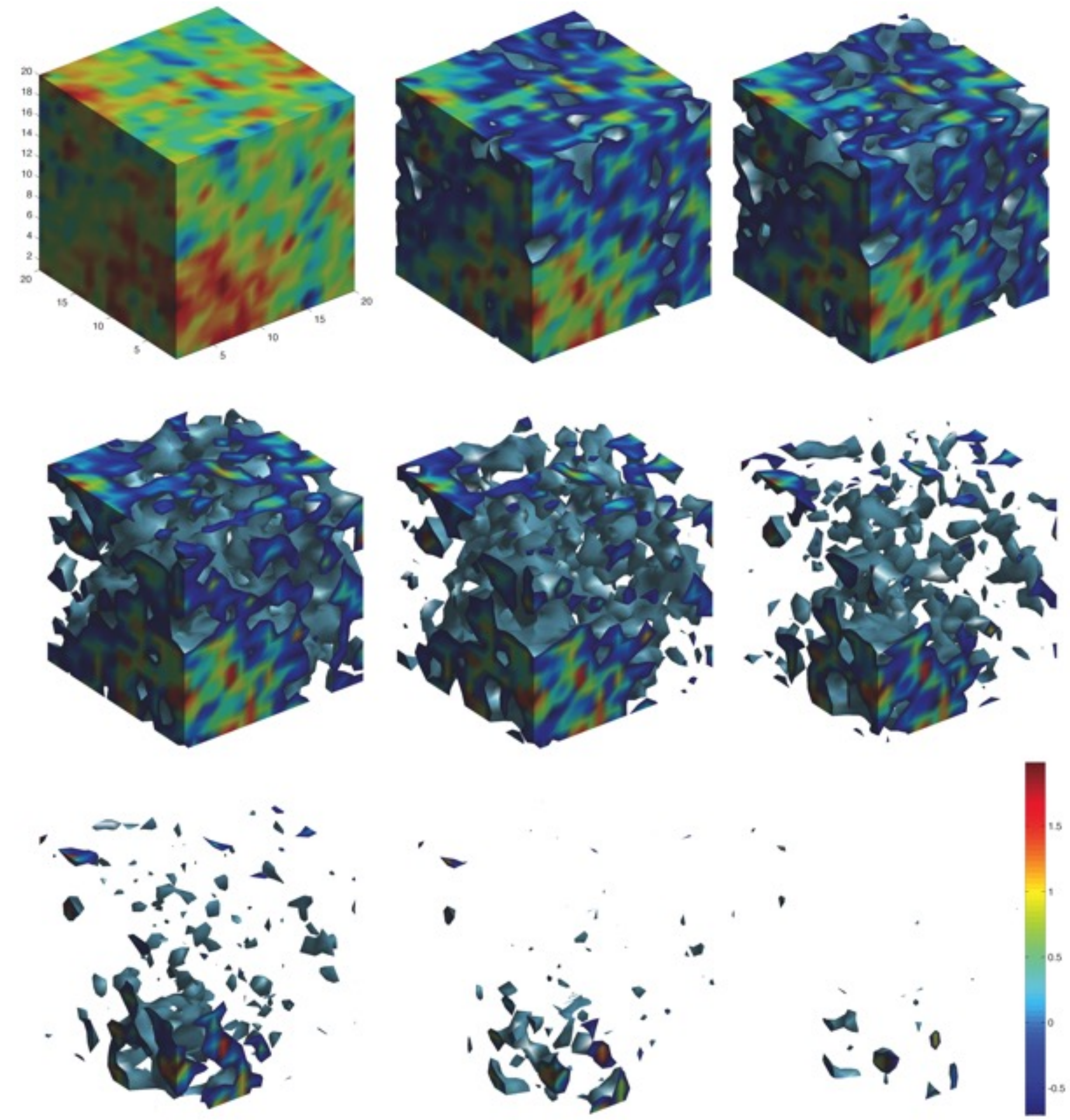
Random field

$$P\Big(\sup_{x\in\mathbb{M}} T(x) > h\Big)$$

$$A_h = \{x \in \mathbb{M} : T(x) > h\}$$

$$P\Big(\sup_{x\in\mathbb{M}} T(x) > h\Big) = \mathbb{E}\chi(A_h)$$

$$\chi(A_h) = \sum_j (-1)^j \beta_j(A_h)$$



Milor 1963 Morse theory
Adler, 1994 The geometry of random fields
Worsley et al., 1996 Human Brain Mapping

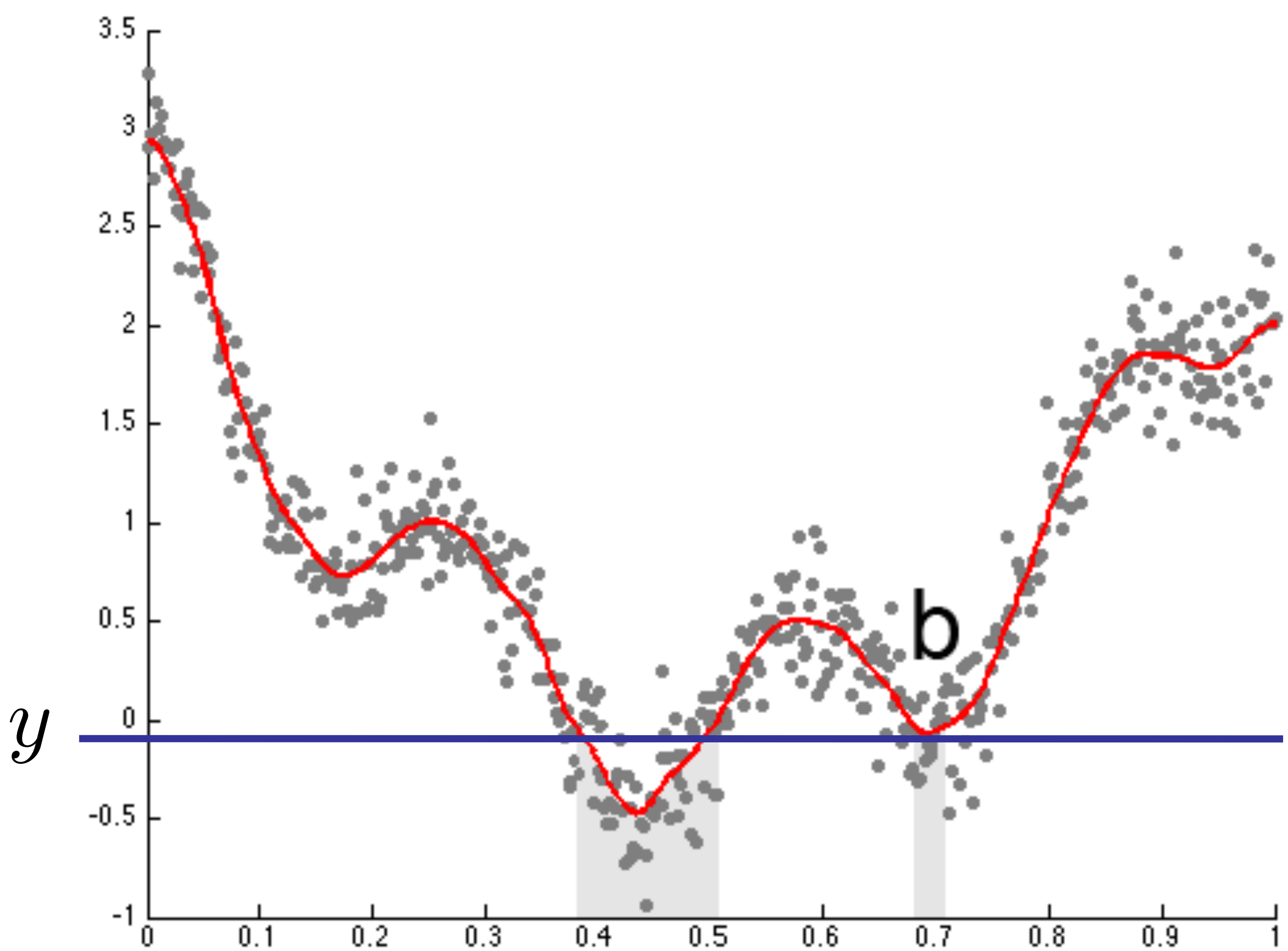
Morse Filtration

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

Morse theory for functional data

$$Y(t) = \mu(t) + \epsilon(t)$$

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



Sublevel set

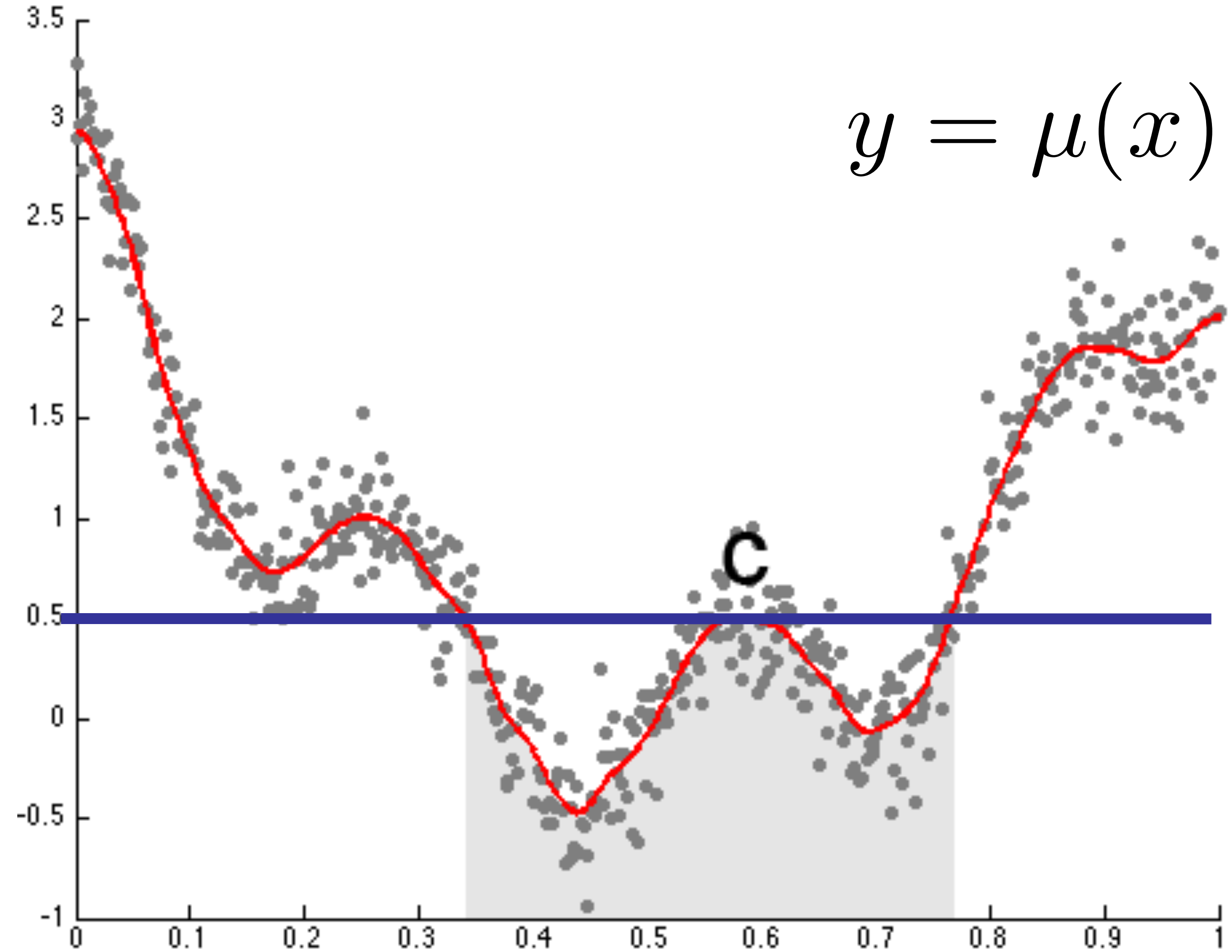
$$R_y = \{t : \mu(t) \leq y\}$$

The topology of sublevel set is characterized by Betti-0 number only

Morse filtration

Sublevel set

$$R_y = \{t : \mu(t) \le y\}$$



Morse filtration

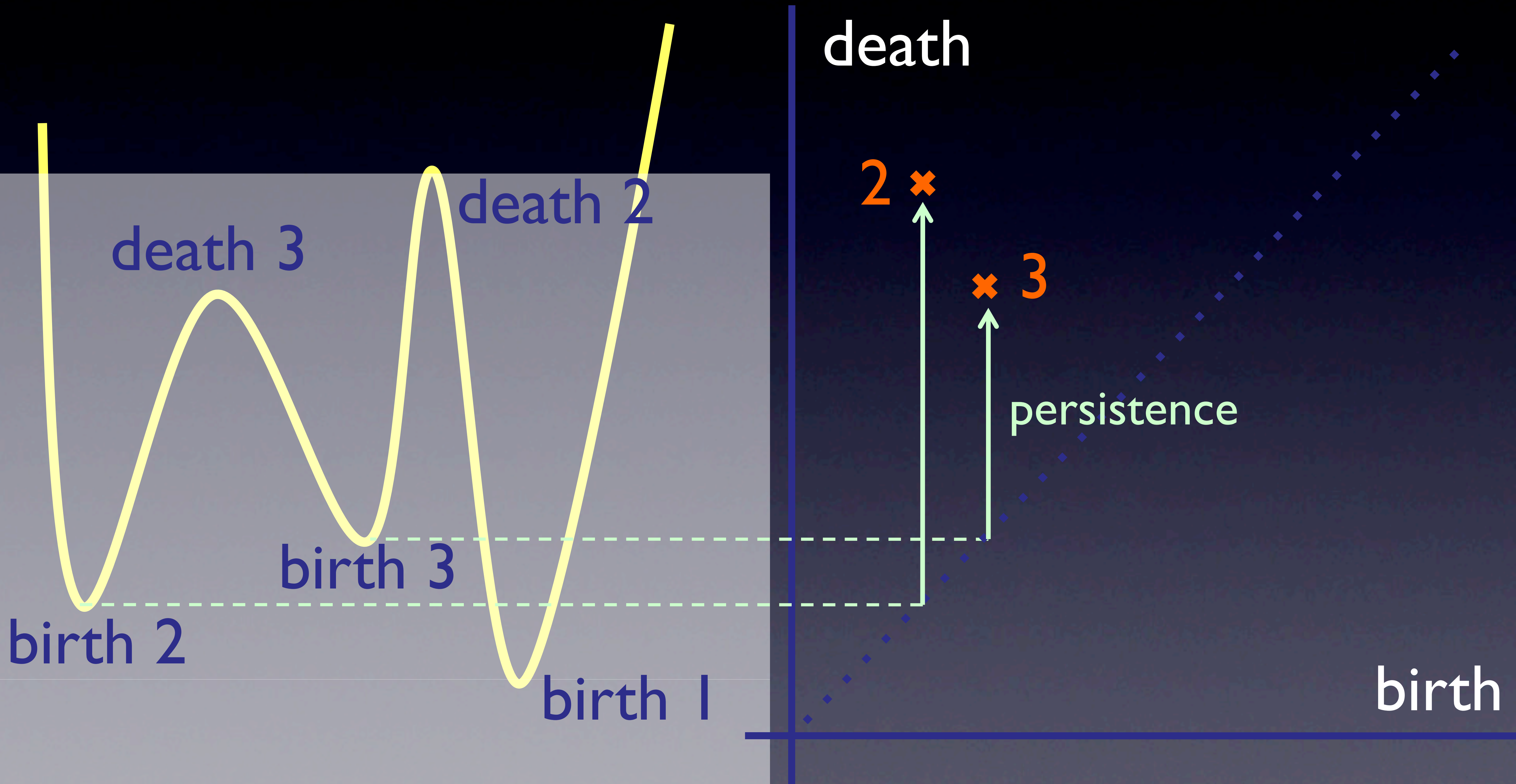
$$R_b \subset R_c$$

Component dies at c

$$\beta_0(R_c) = \beta_0(R_b) - 1$$

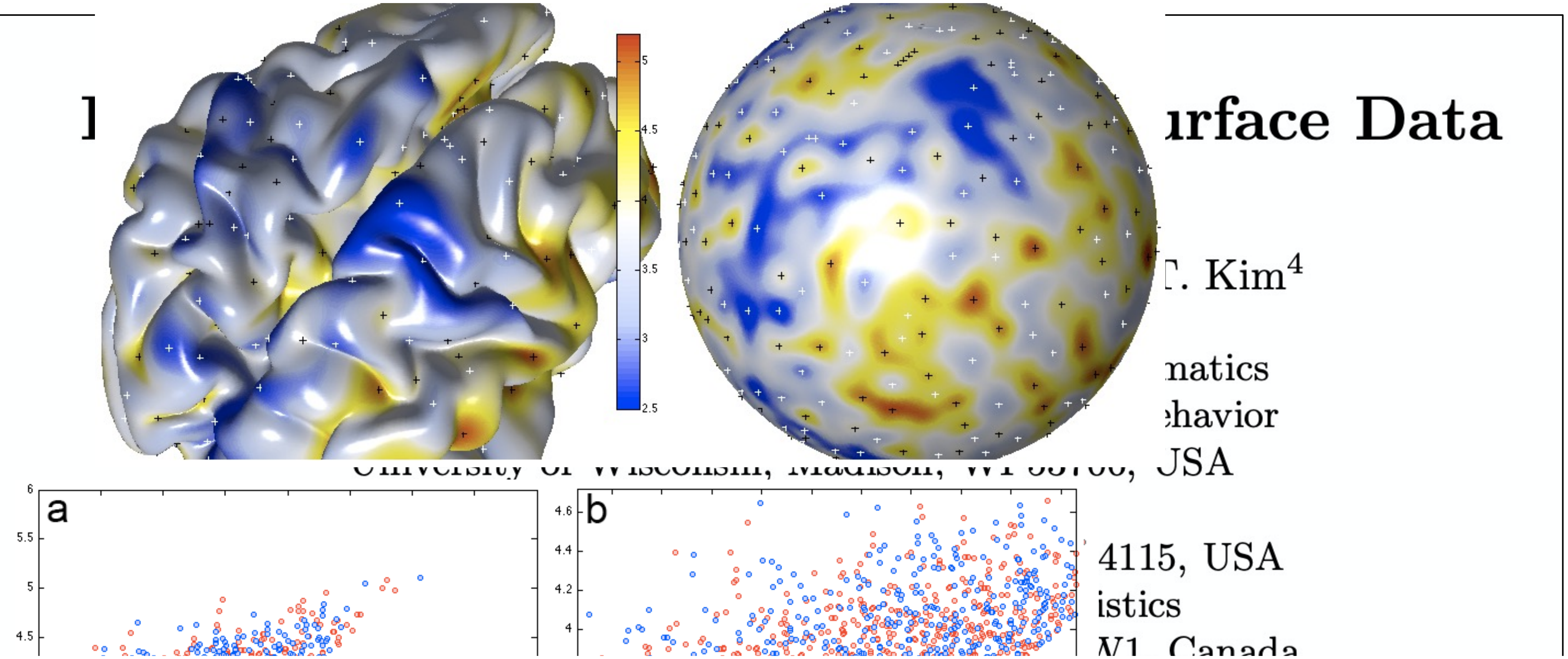
Persistence Diagram (PD)

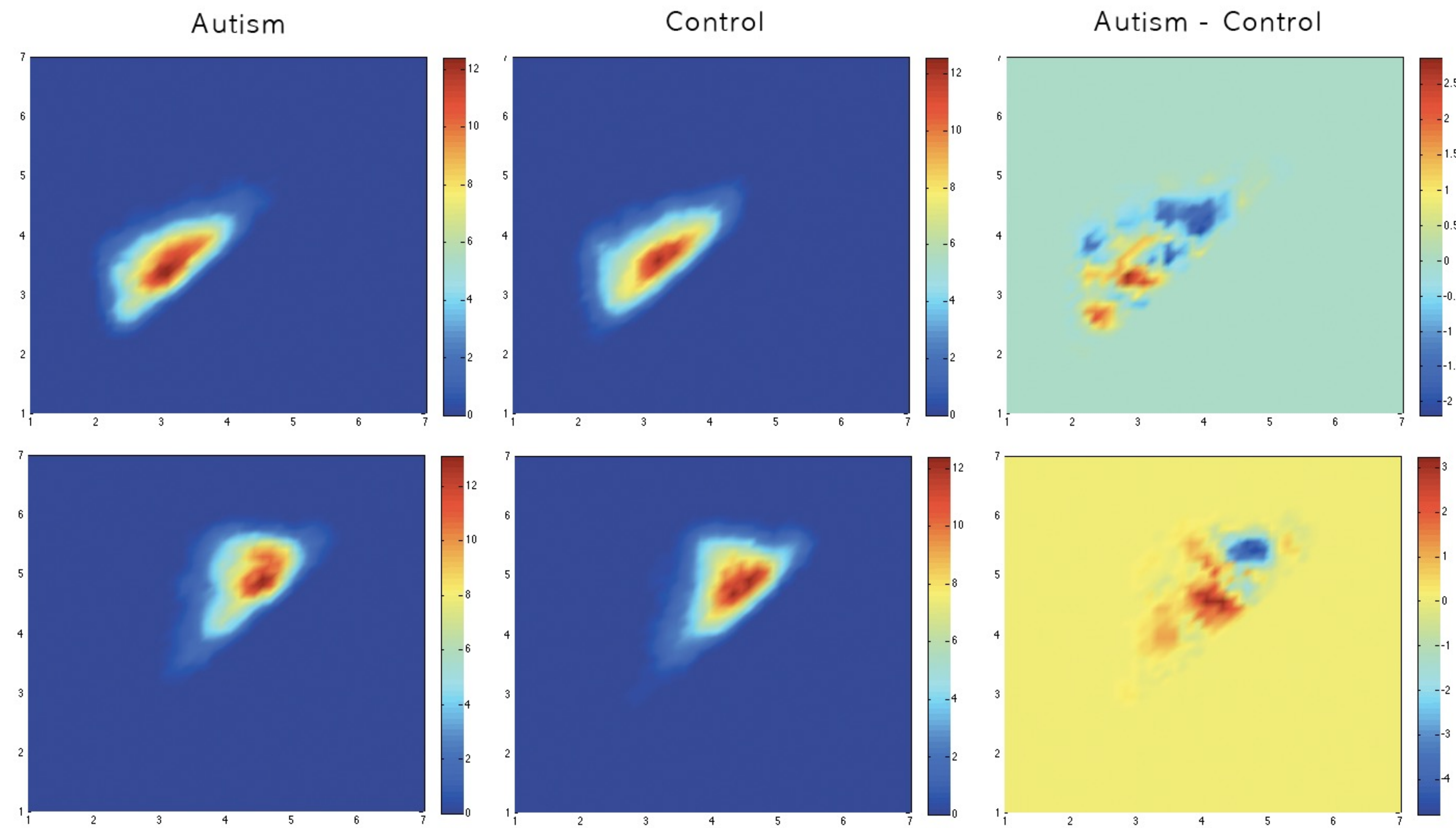
O(n log n)



Elder's rule:

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





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

First TDA paper in medical imaging

Rips filtrations

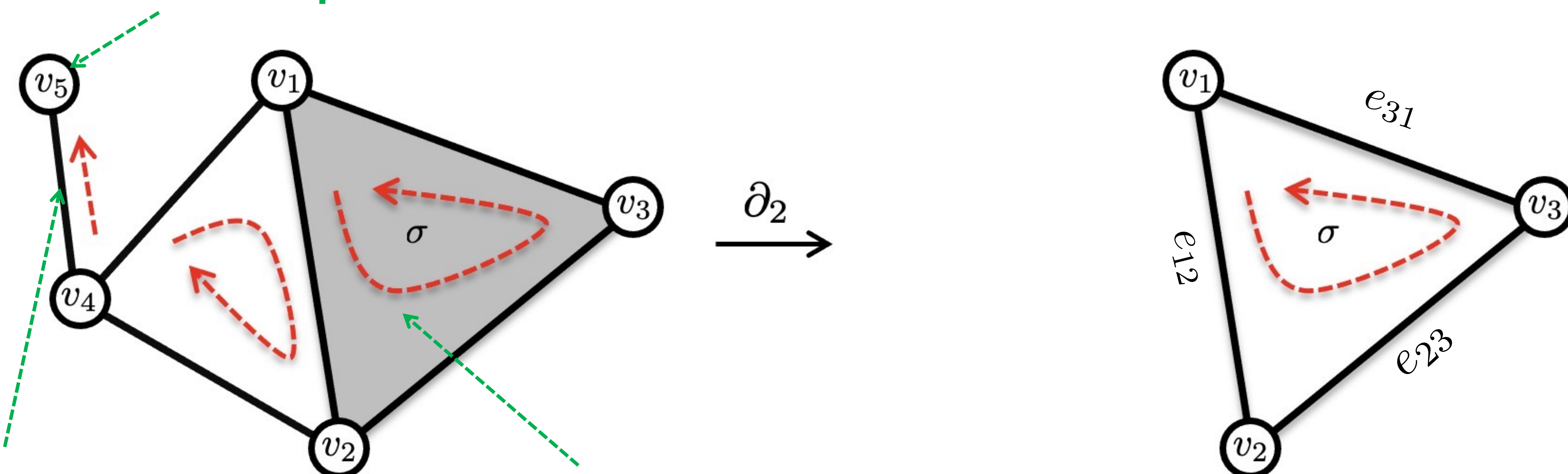
n-simplex

The basic building block of persistent homology

The smallest convex set containing *n+1* points

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

0-simplex $[v_5]$



1-simplex

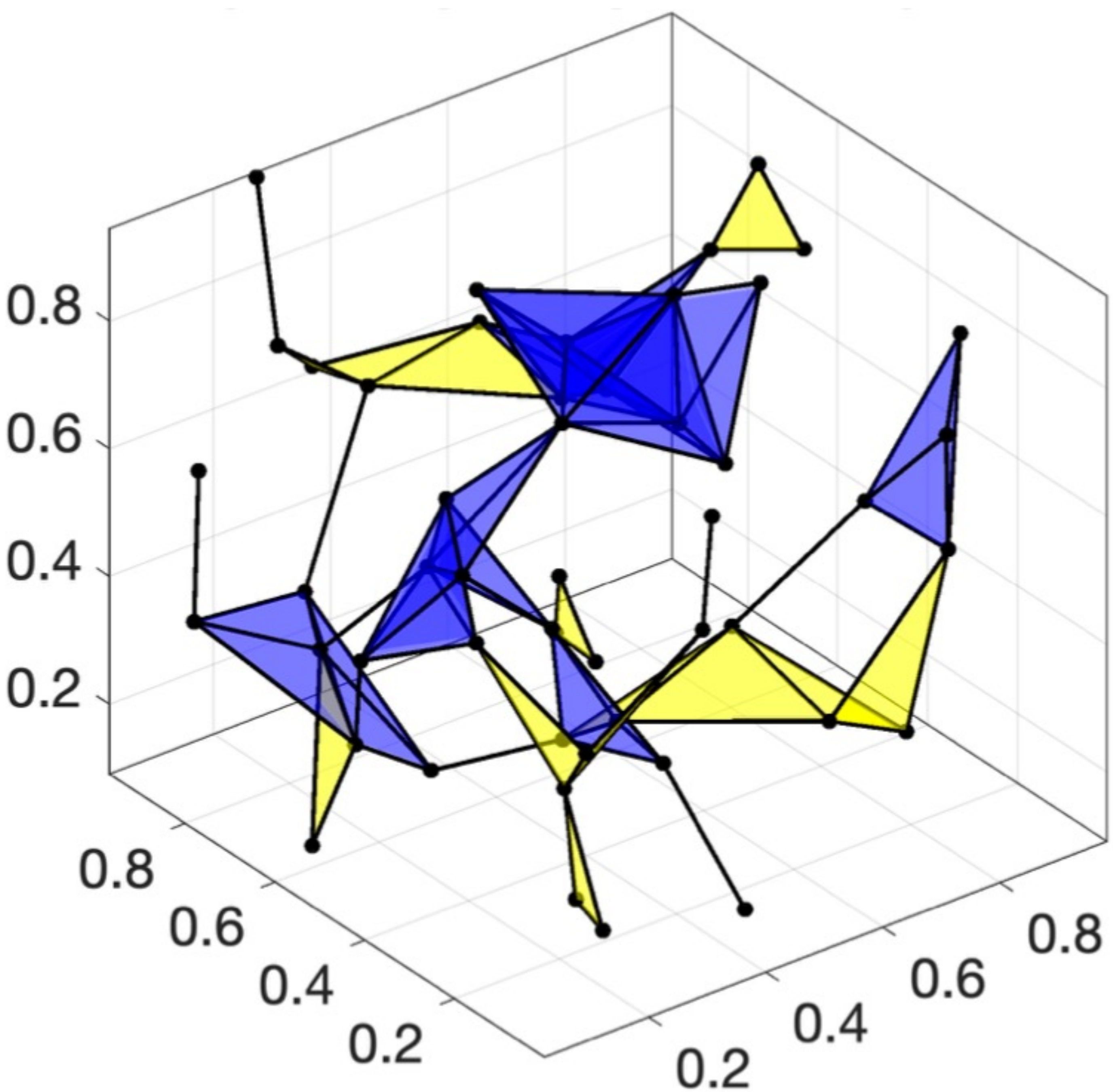
$[v_4, v_5] = -[v_5, v_4]$

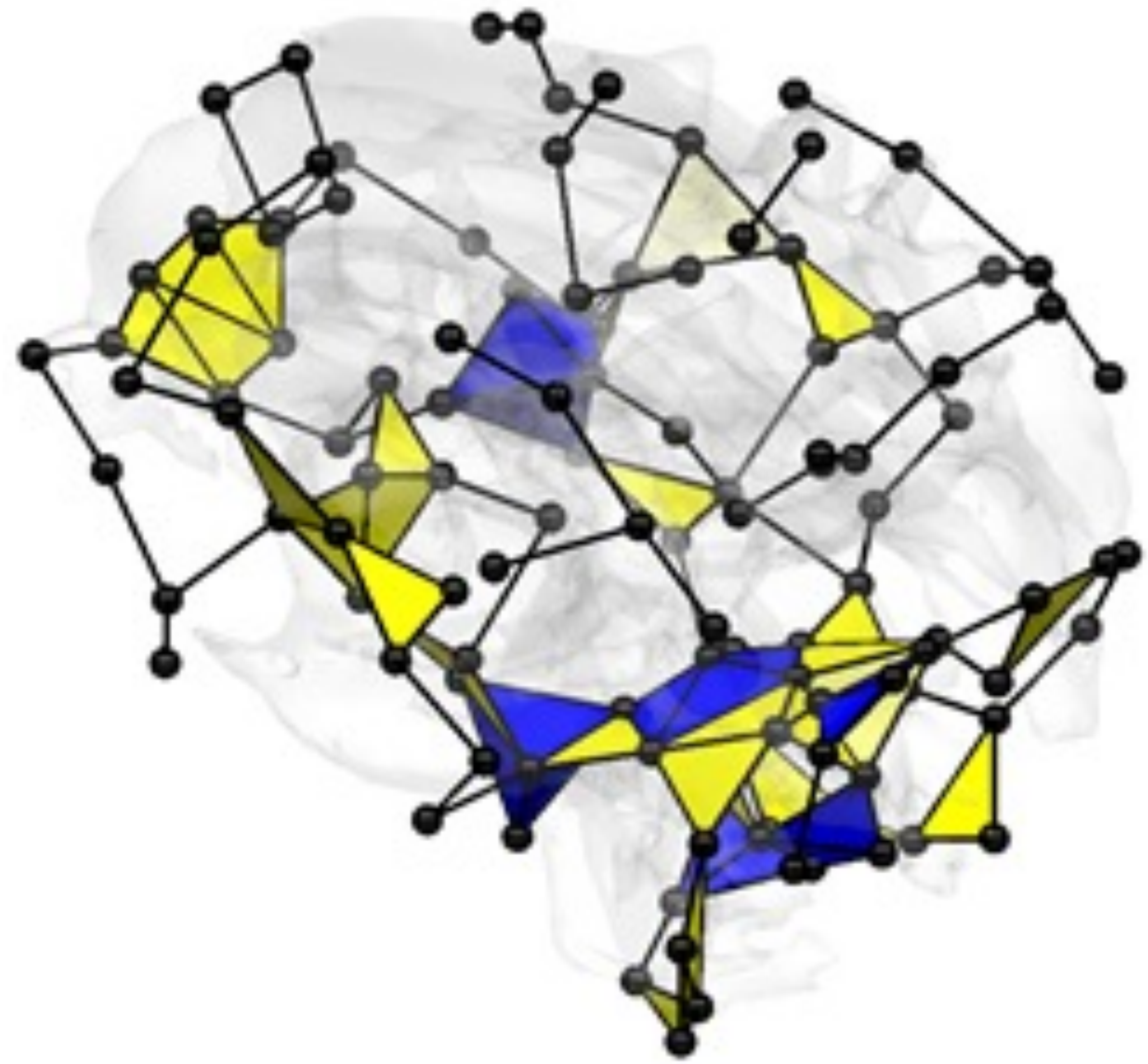
2-simplex

$\sigma = [v_1, v_2, v_3]$

Simplicial complex

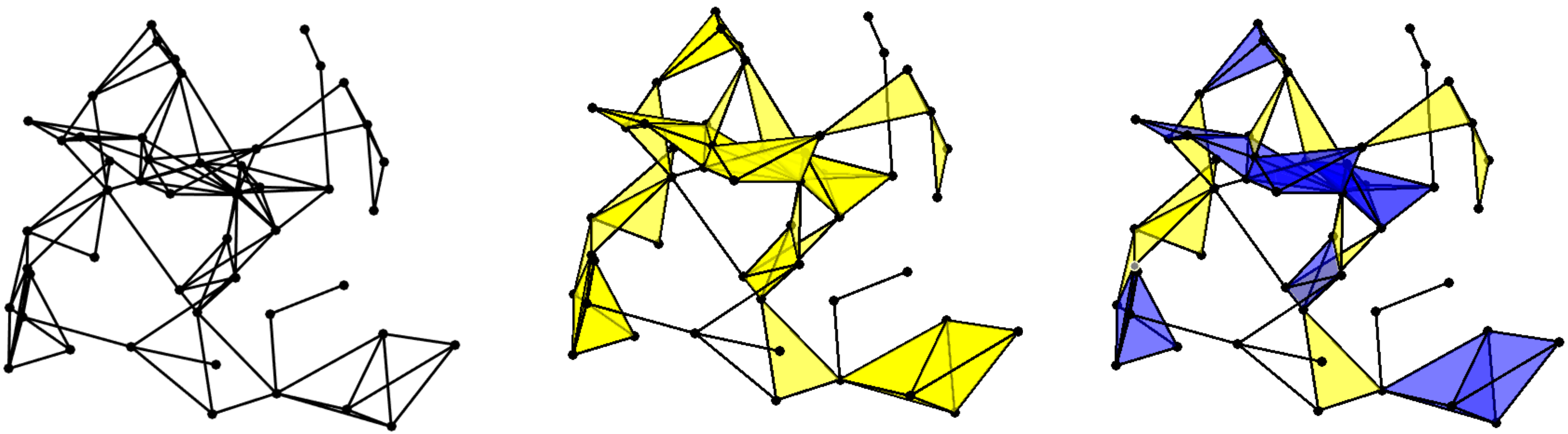
A simplicial complex is a set composed of points, line segments, triangles, and their n-dimensional counterparts.





k-skeleton

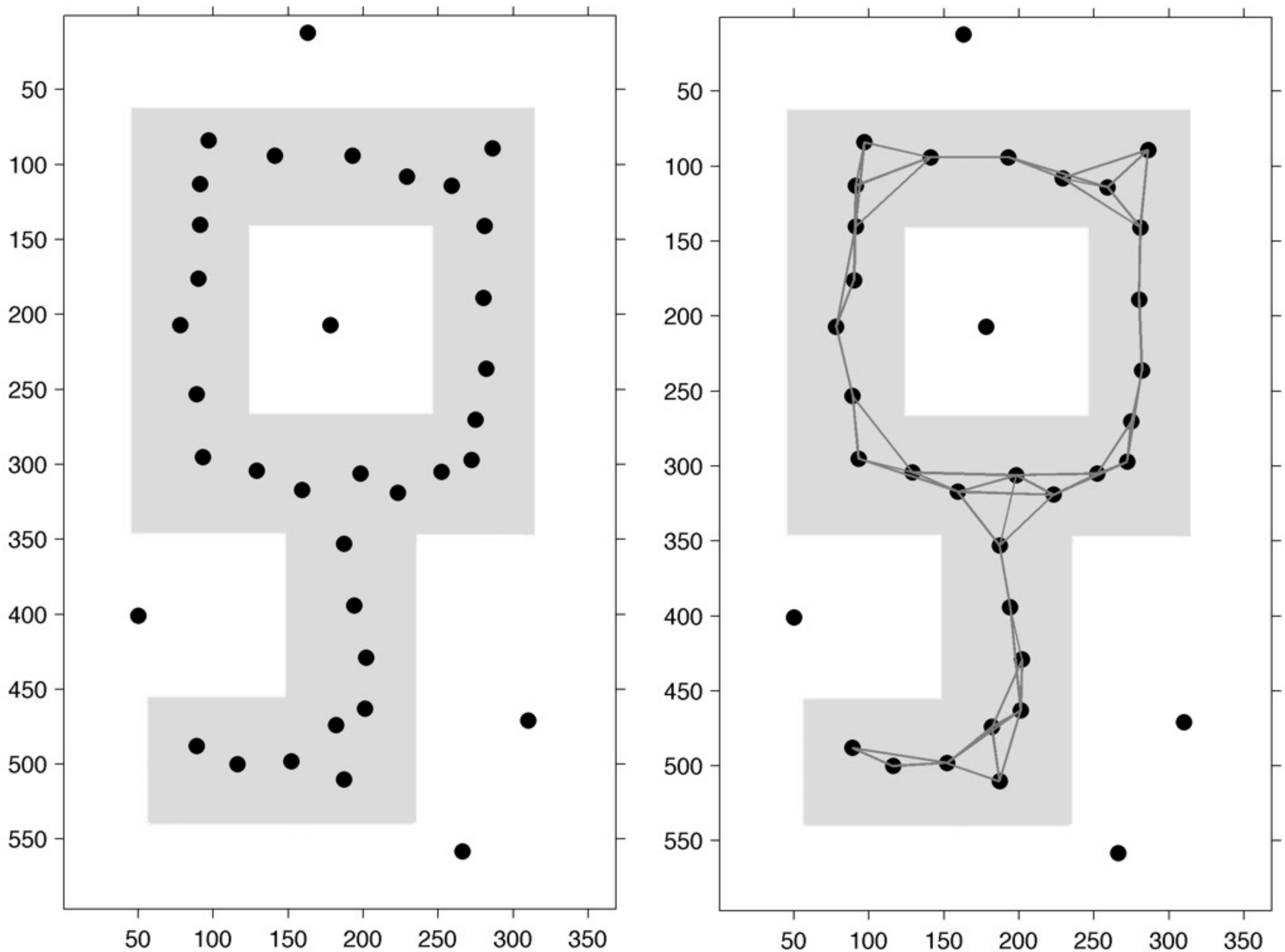
A simplicial complex consisting of up to k-simplices



1-skeleton 2-skeleton 3-skeleton

1-skeleton of point cloud data

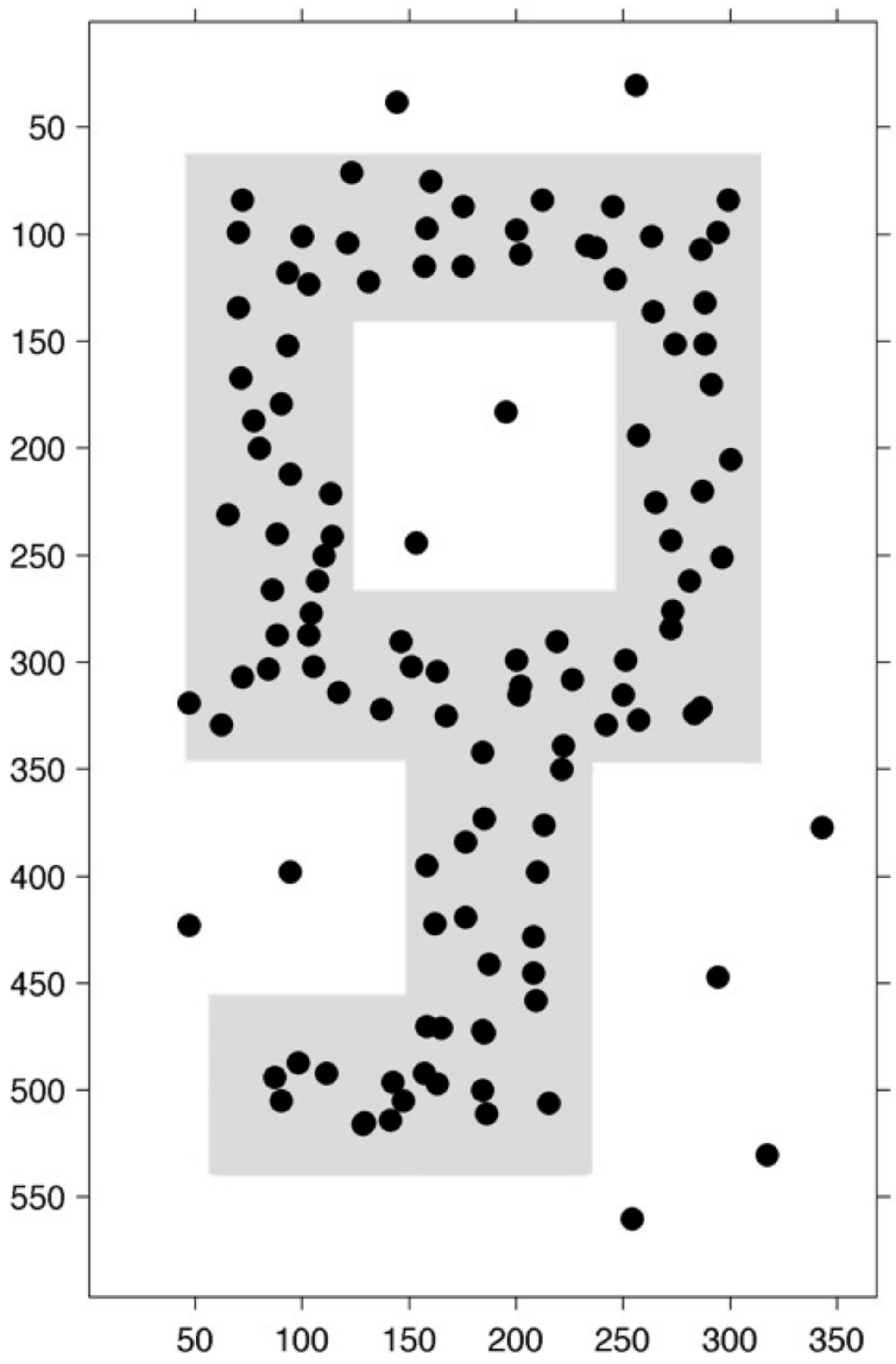
$\varepsilon = 70\text{mm}$

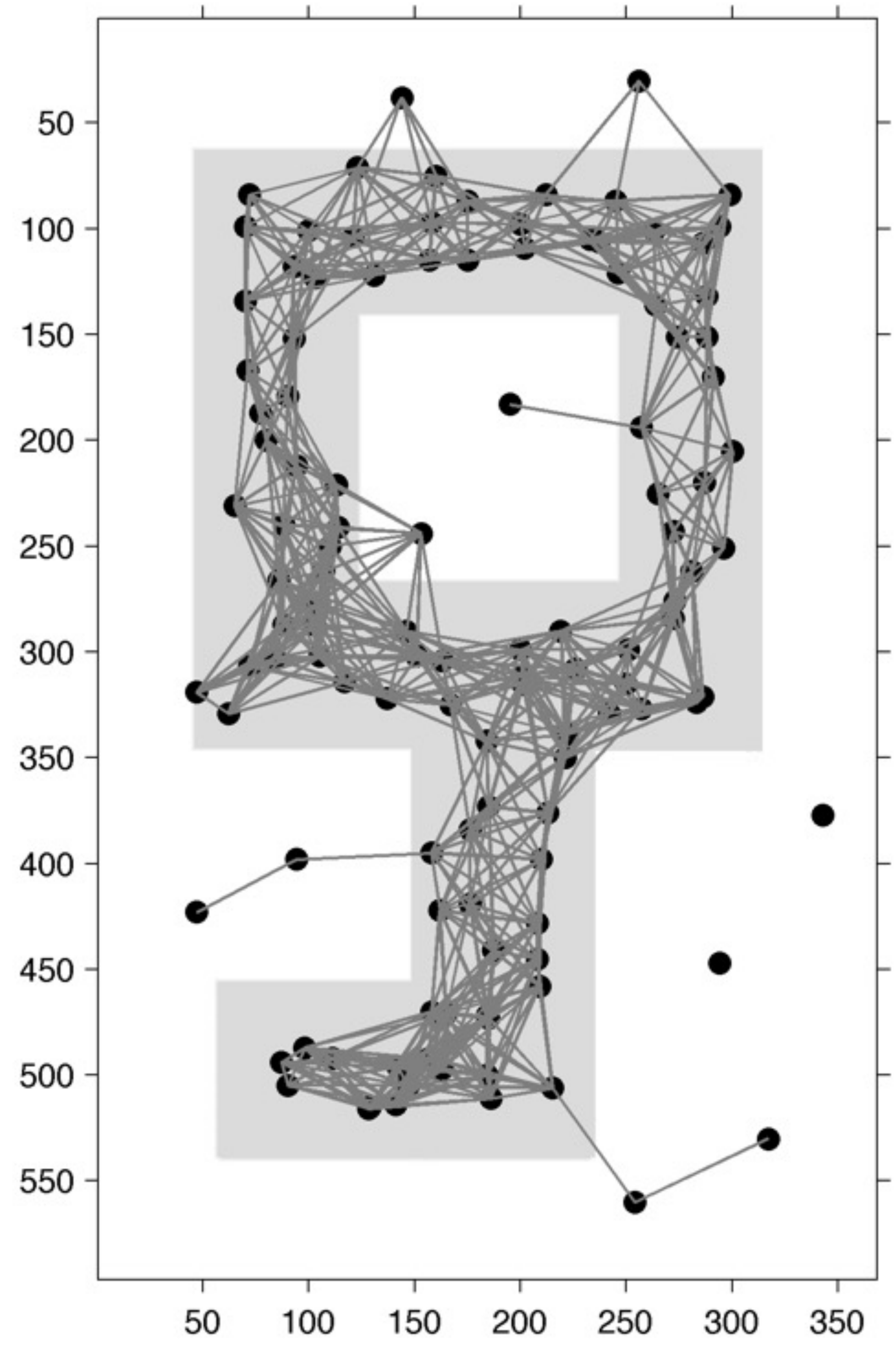


Recovering underlying topology

1-skeleton of point cloud data

ε = 70mm

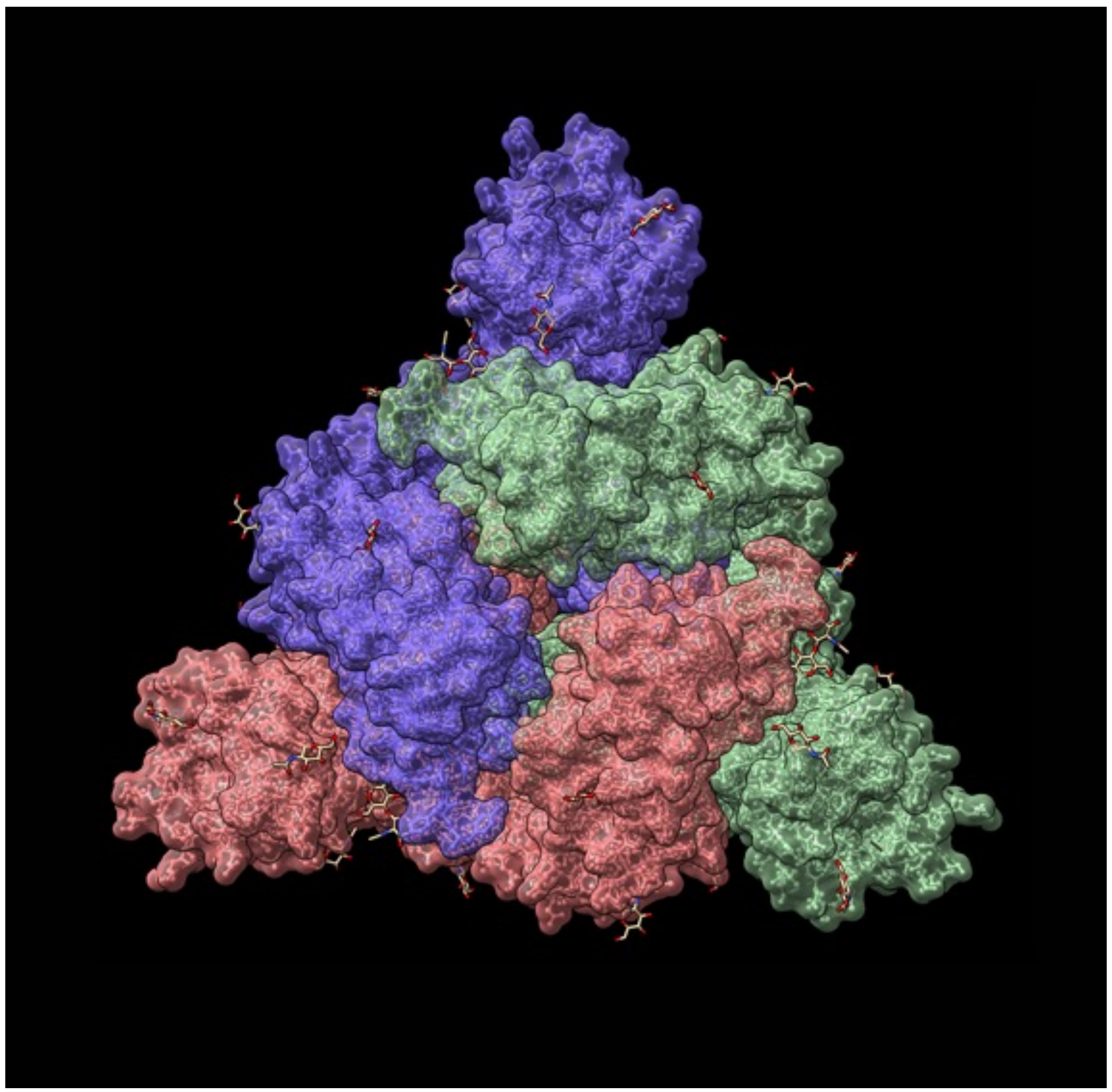


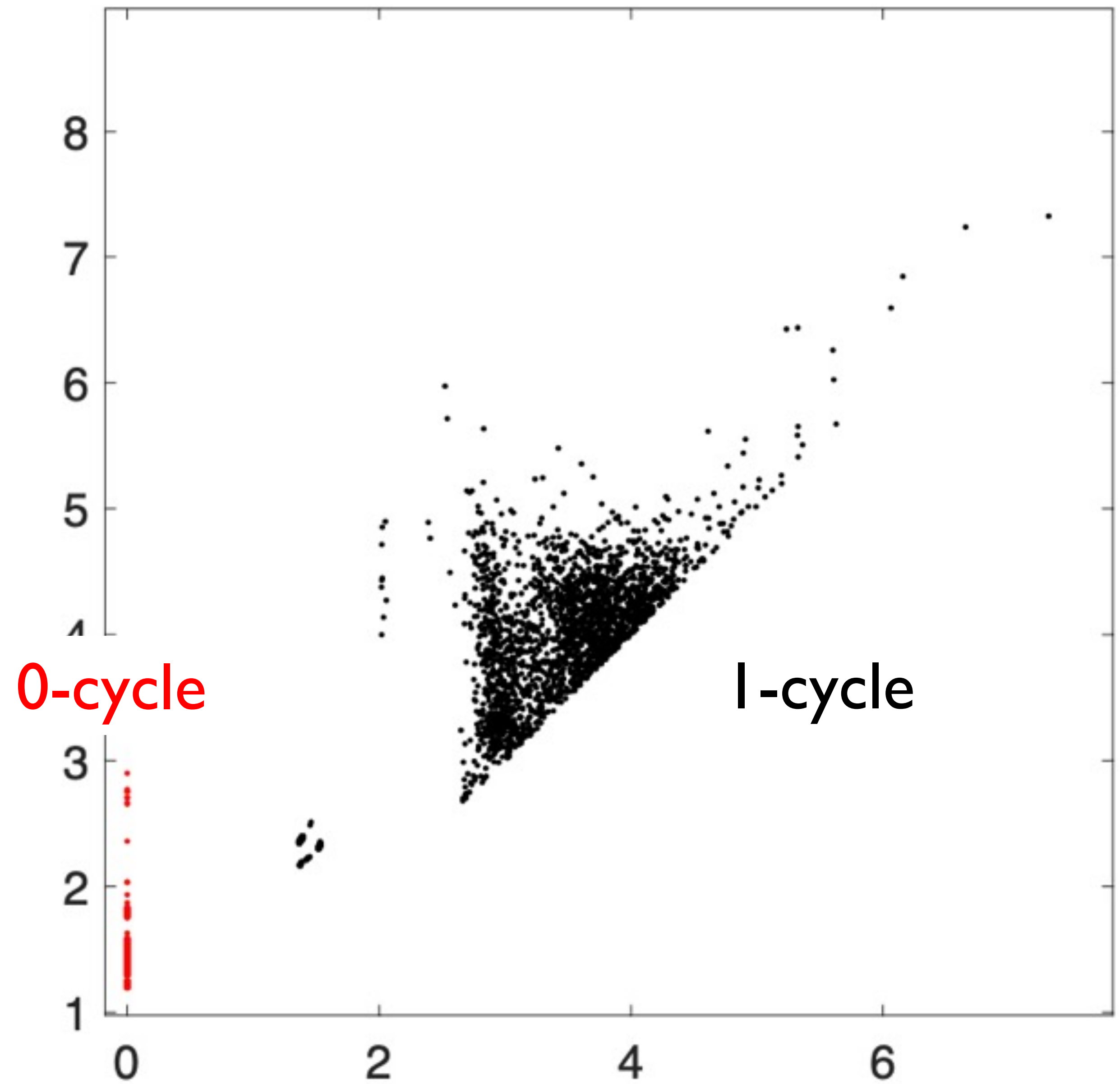


Better approach: perform kernel smoothing and then Morse filtration

Persistence Diagram (PD) of a protein molecule

Rips filtration on distance between 8000 atoms





Extremely slow computation → Simply use graph filtration

Graph filtrations

Baseline filtration for brain networks first introduced in

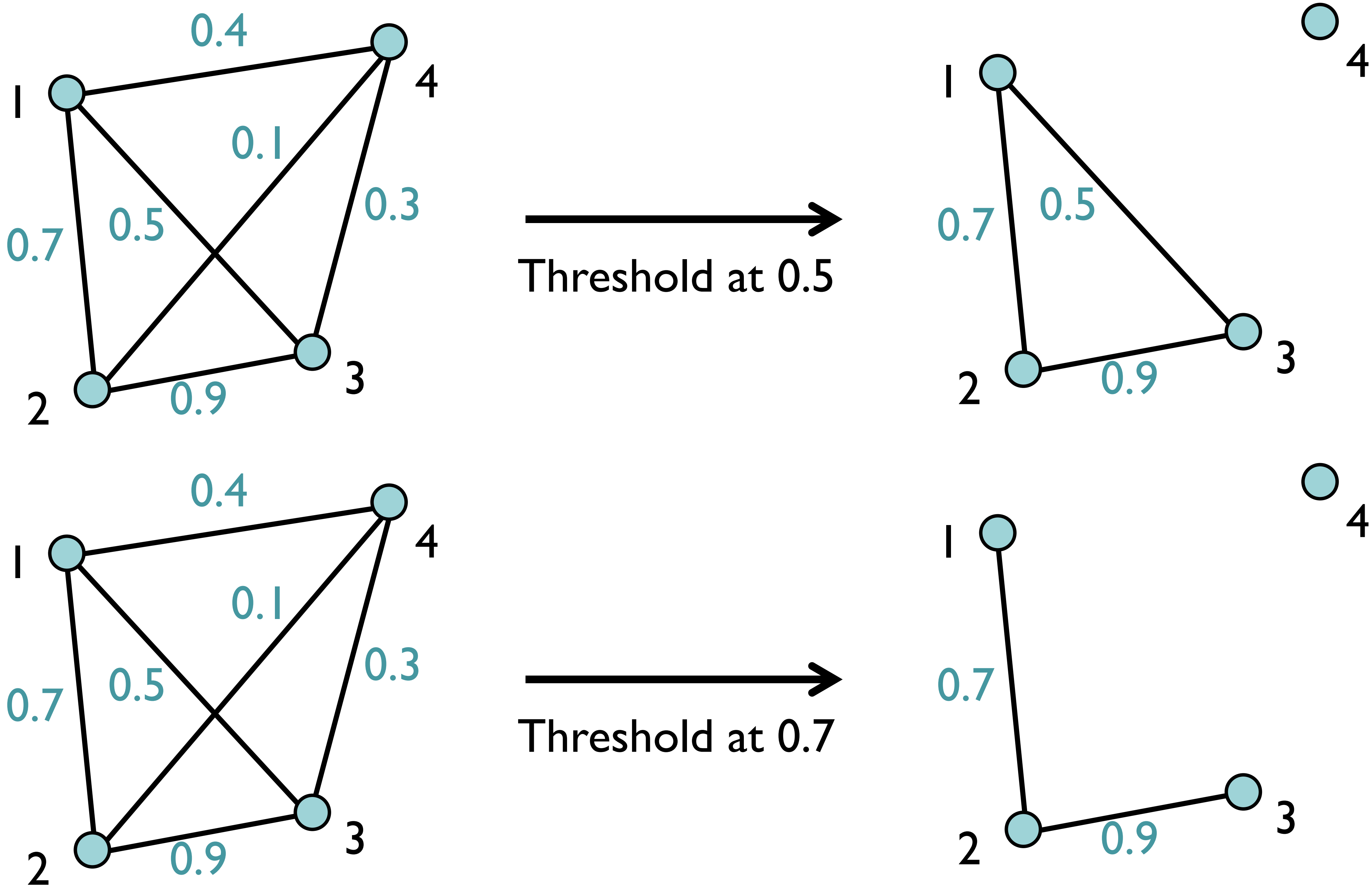
Lee et al. 2011 MICCAI 302-309

Lee et al. 2012 IEEE Transactions on Medical Imaging 31:2267-2277

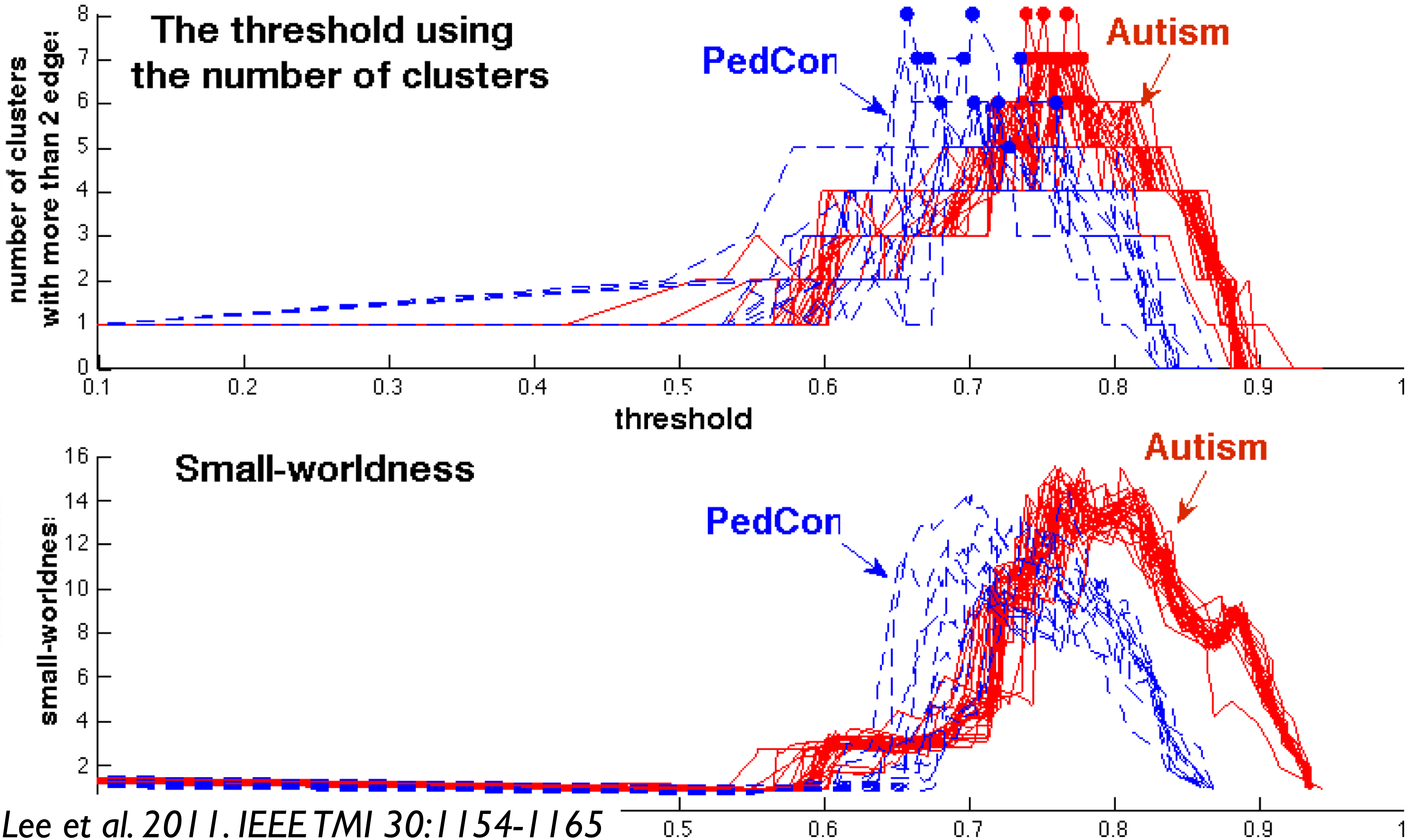
What is wrong with the arbitrary thresholding?

Edge weight ρ_{ij} between node *i* and *j*

→ Connectivity matrix $\rho = (\rho_{ij})$



Single threshold often suboptimal → TDA-based network analysis



Lee et al. 2011. IEEE TMI 30:1154-1165

Rips filtration vs. graph filtration

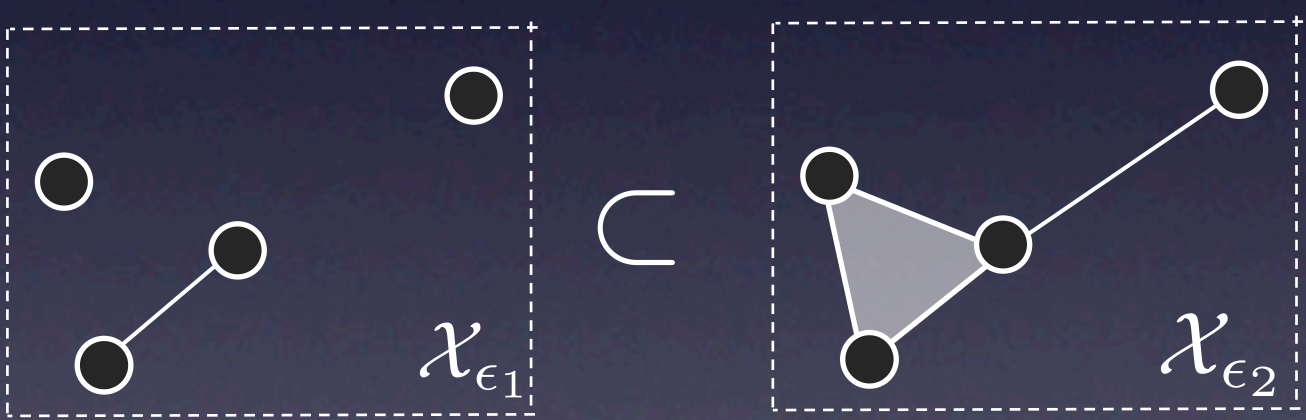
Metric space

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

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

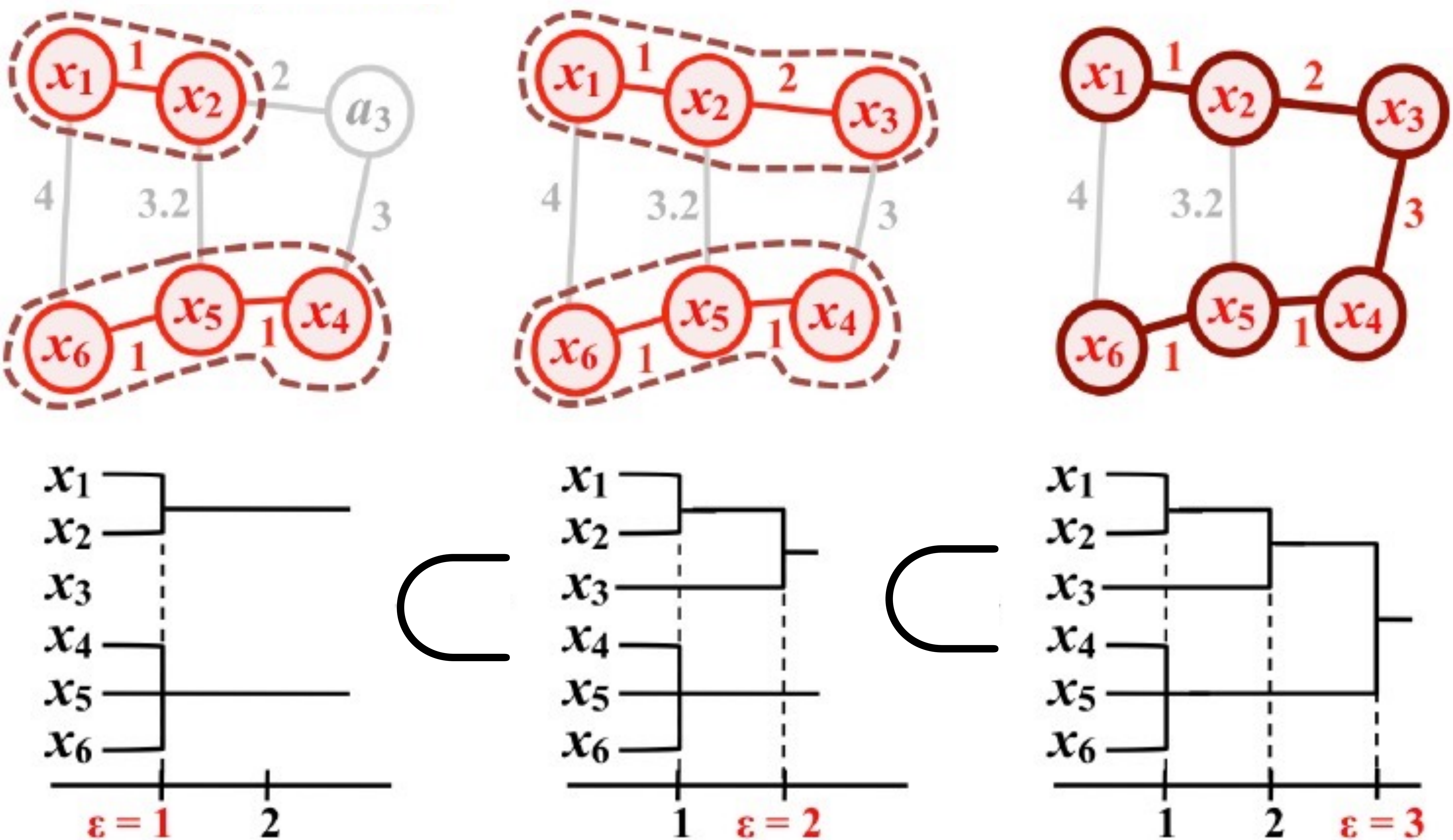
Graph filtration

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

for increased edge weights

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

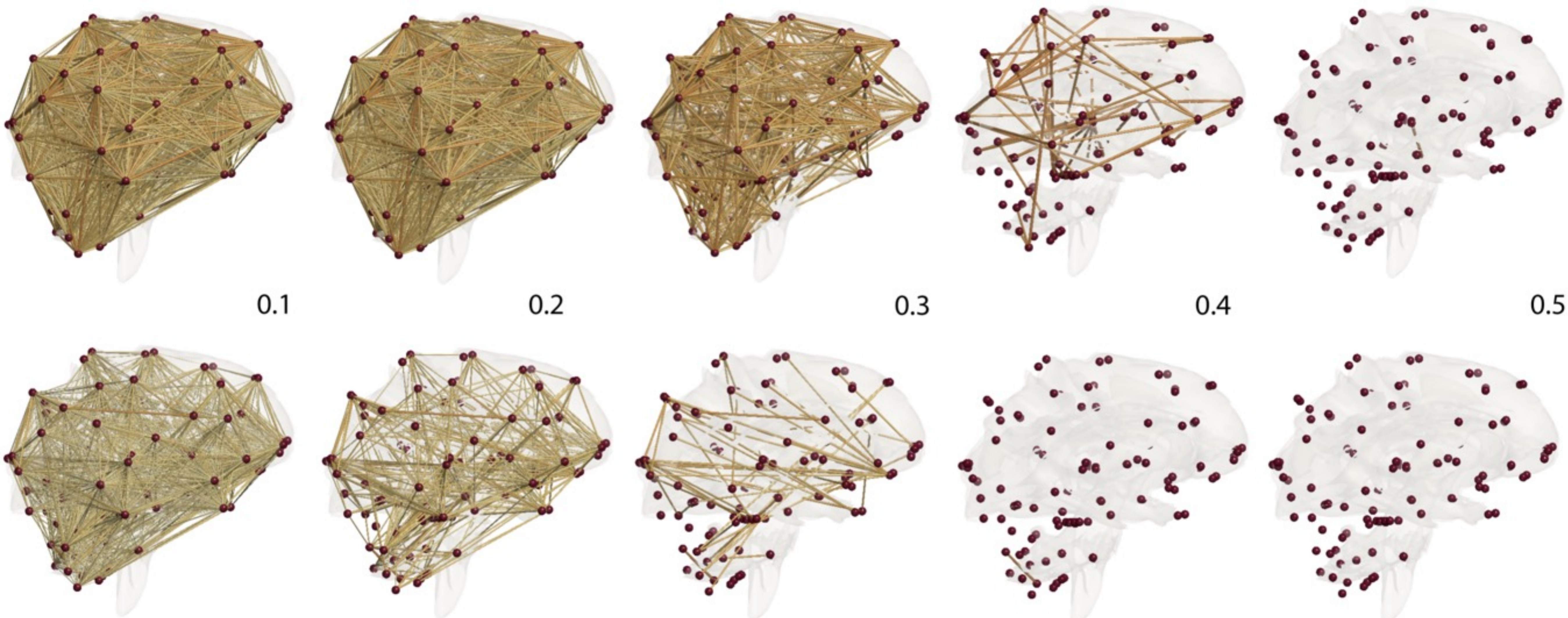
Graph filtration=single linkage clustering



Lee et al. 2011 MICCAI 302-309

Graph filtrations on resting-state fMRI

MZ-twins



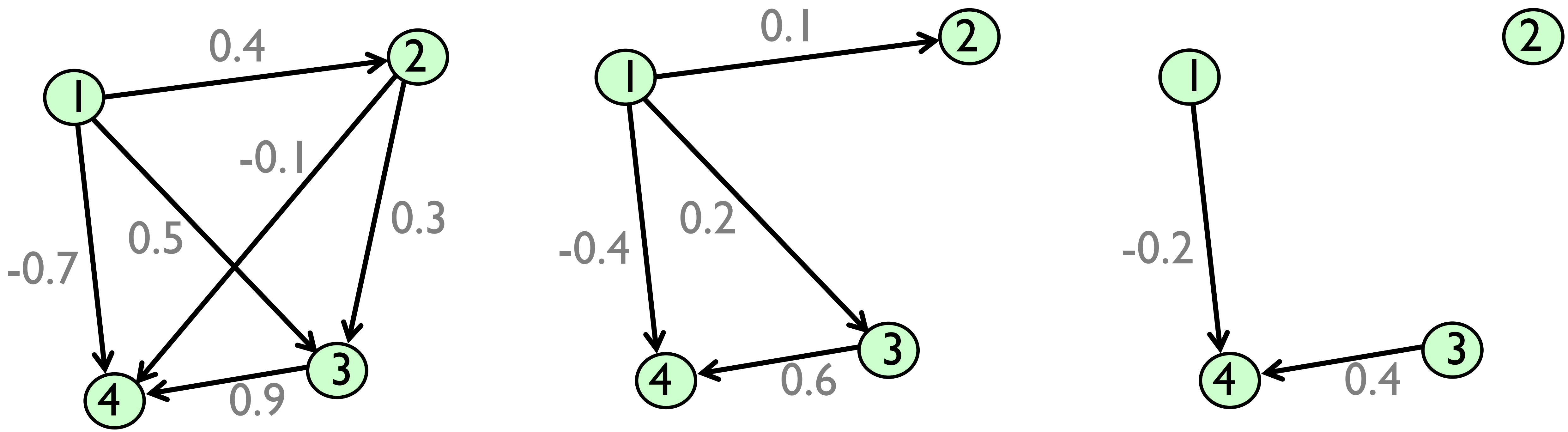
DZ-twins

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.

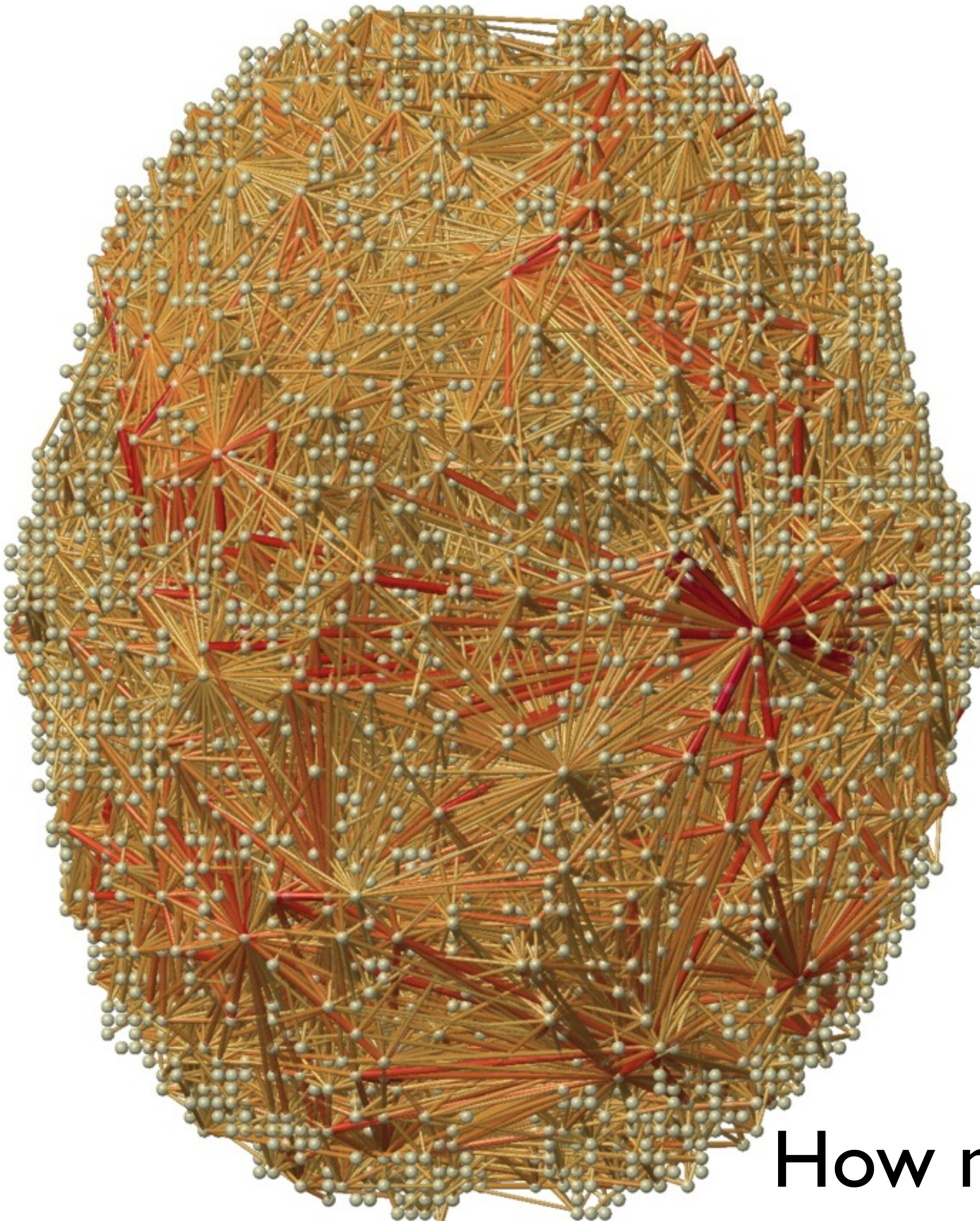
Graph filtration on directed graphs

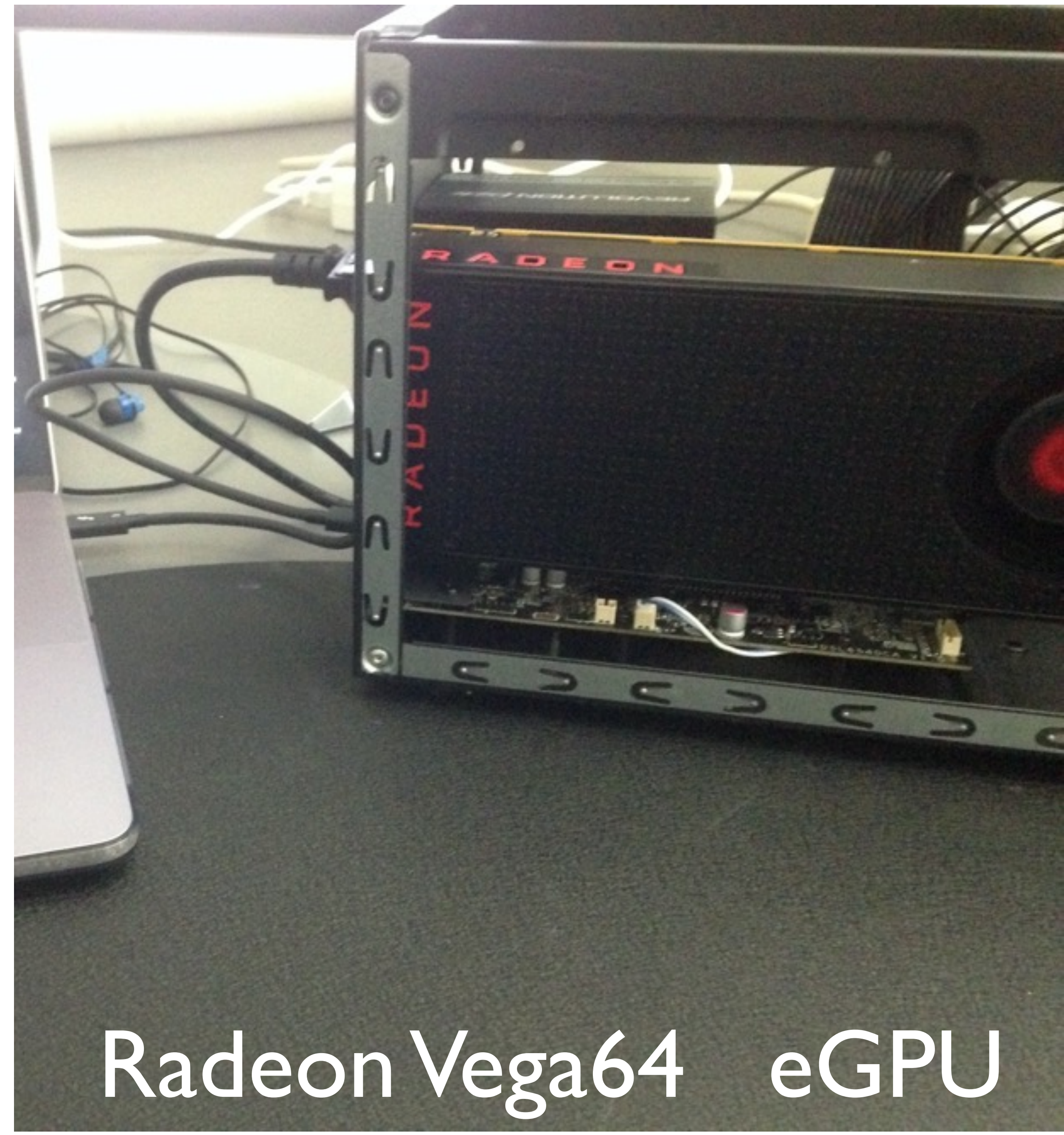


$$G(0) \supset G(0.3) \supset G(0.5)$$

Question: How to build graph filtration coherently on directed graphs?

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





Radeon Vega64 eGPU

How many cycles in the network?

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)
```

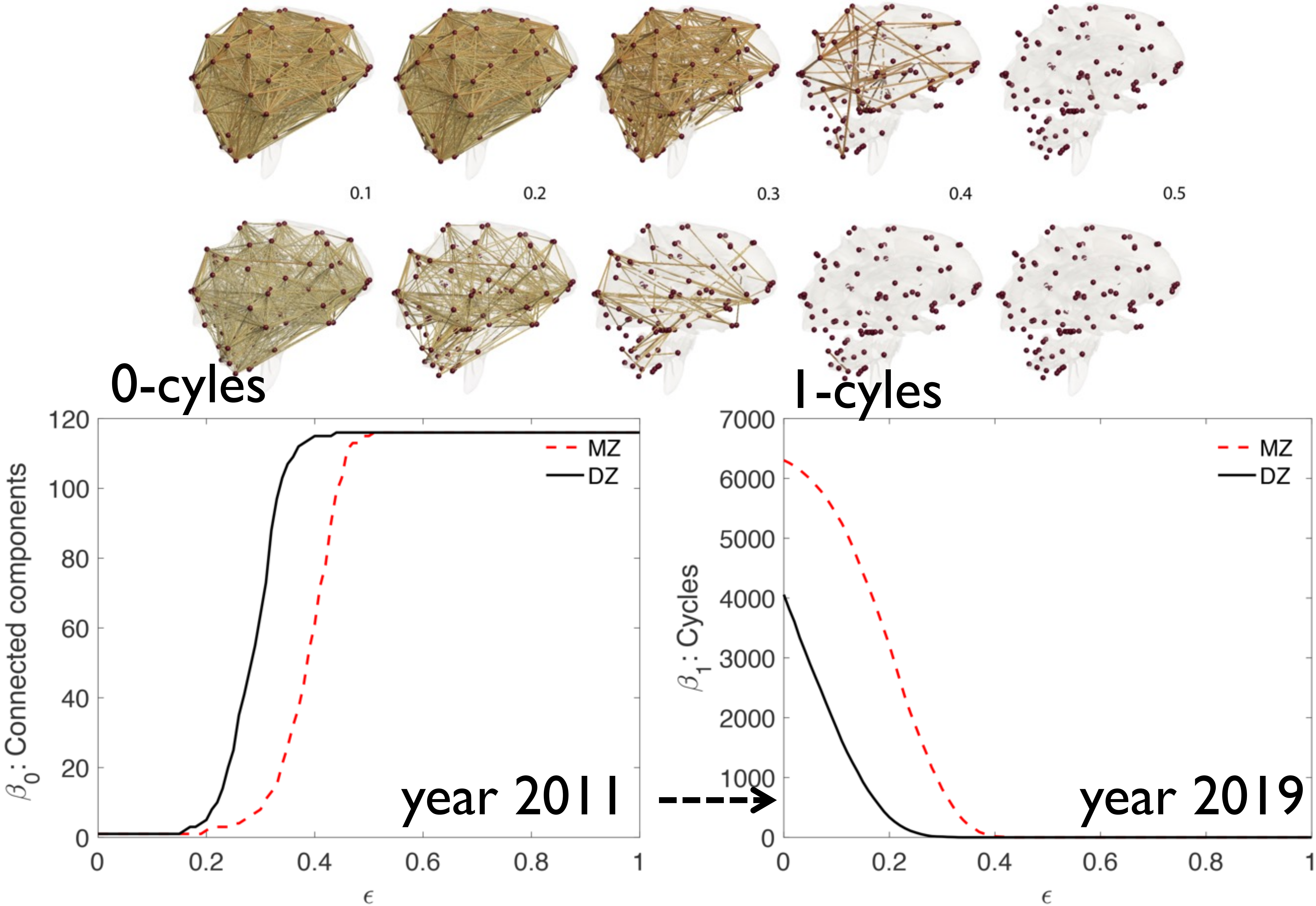
Computation of β_1: As a function of β_0

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

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

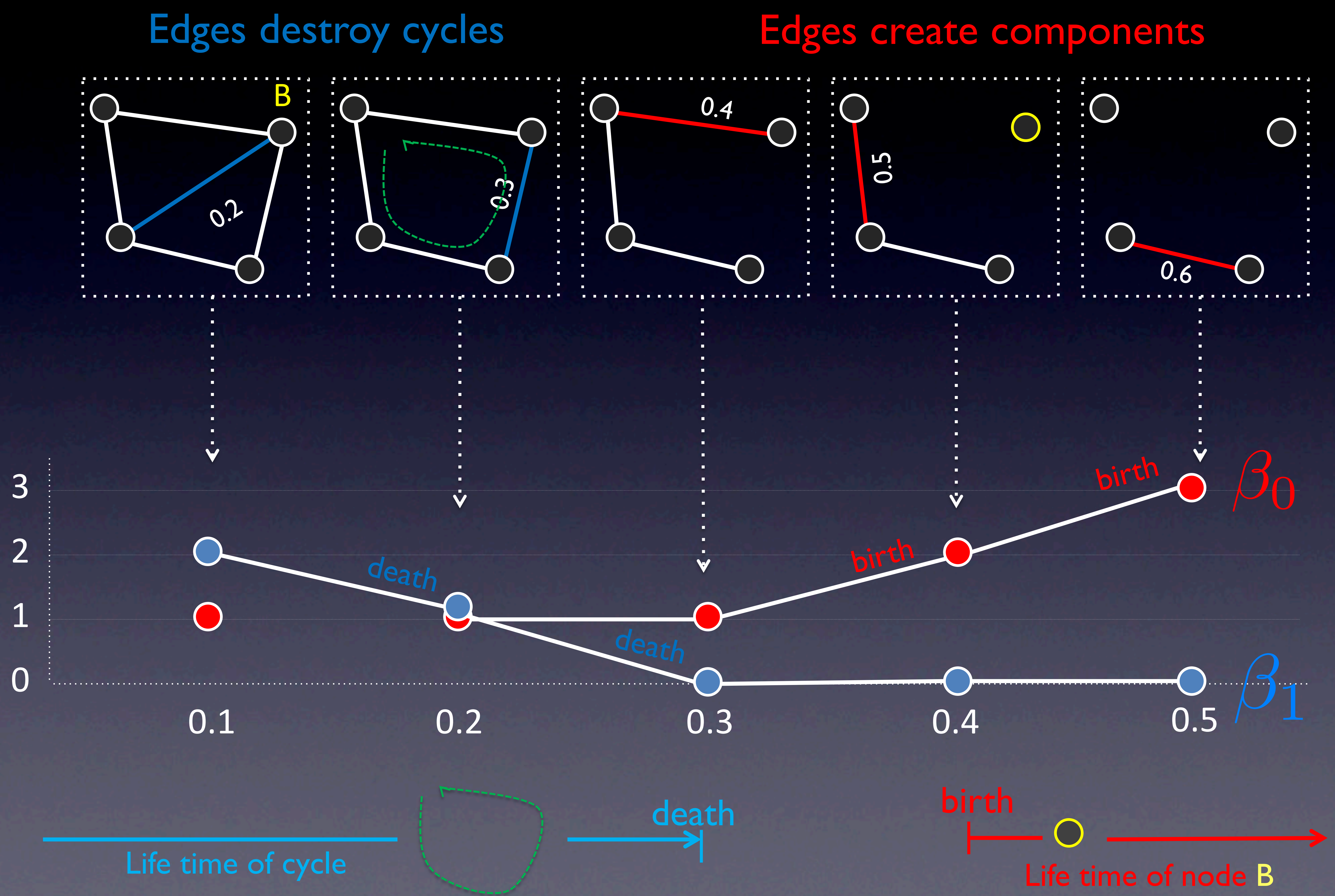
<u>This is not efficient</u>. Need an incremental algorithm that updates as we delete one edge at a time.

Betti-plots in 116 nodes network



Birth and death decomposition

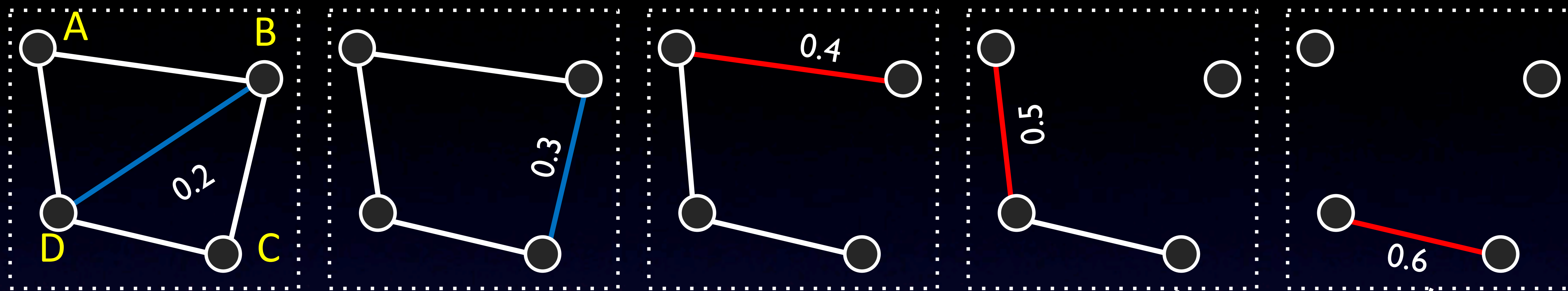
Persistence = Life time (death – birth) of a feature



Theorem *Birth & death sets partition* the edge set

E_1 Edges destroy cycles

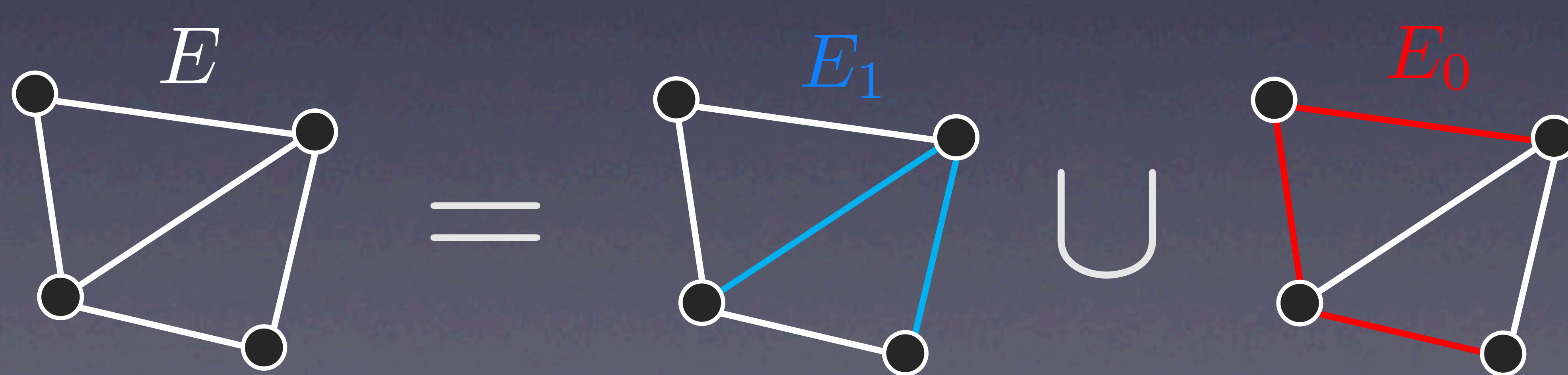
E_0 Edges create components



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

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

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

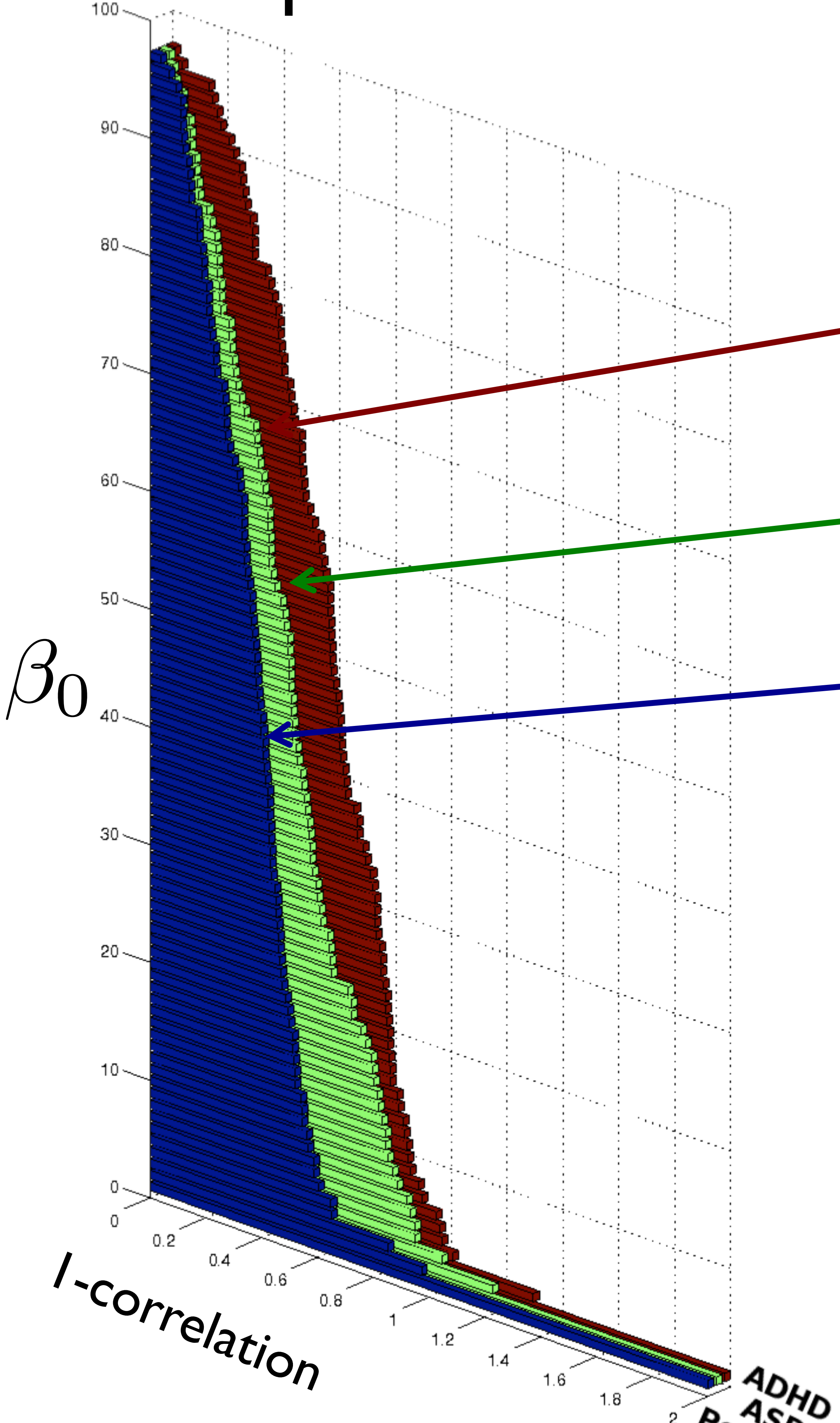


Maximum spanning tree

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

Songdechakraiwut & Chung 2020 arXiv: 2012.00675

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

Hodge Laplacian

Multiplicity of eigenvalues

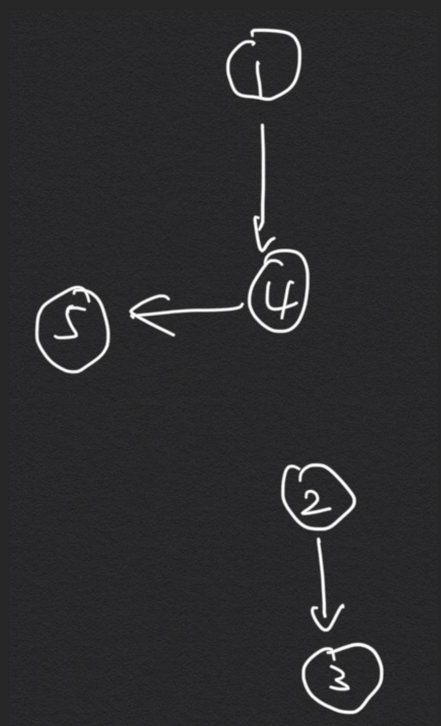
Multiplicity m of eigenvalue λ :

$$L\mathbf{v}_1 = \lambda\mathbf{v}_1, \cdots, L\mathbf{v}_m = \lambda\mathbf{v}_m$$

Theorem: The multiplicity of eigenvalue $\lambda_0 = 0$ is the number of connected components: 0-th Betti number β_0.

Equivalently $\beta_0 = p - rank(L)$

Multiplicity of eigenvalues



$$\begin{array}{c|ccccc} & 1 & 4 & 5 & 2 & 3 \\ \hline 1 & & & & & \\ 4 & & & & & \\ 5 & & & & & \\ 2 & & & & & \\ 3 & & & & & \end{array} \quad \begin{bmatrix} 1 & -1 & 0 & 0 & 0 \\ -1 & 2 & -1 & 0 & 0 \\ 0 & -1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & -1 \\ 0 & 0 & 0 & -1 & 1 \end{bmatrix}$$

```
>> [F, D] = eig(L)
```

F =

$$\begin{matrix} 0 & -0.5774 & -0.7071 & 0 & 0.4082 \\ 0 & -0.5774 & 0.0000 & 0 & -0.8165 \\ 0 & -0.5774 & 0.7071 & 0 & 0.4082 \\ -0.7071 & 0 & 0 & -0.7071 & 0 \\ -0.7071 & 0 & 0 & 0.7071 & 0 \end{matrix}$$

D =

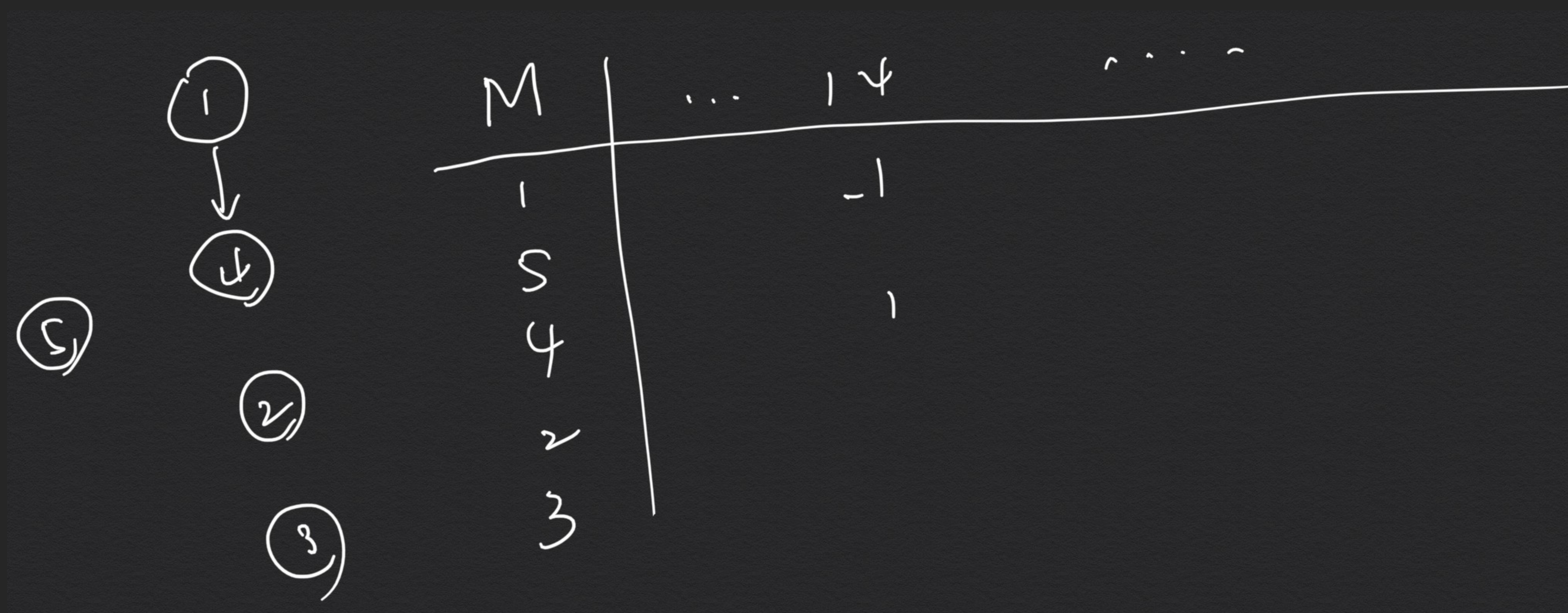
$$\begin{matrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0.0000 & 0 & 0 & 0 \\ 0 & 0 & 1.0000 & 0 & 0 \\ 0 & 0 & 0 & 2.0000 & 0 \\ 0 & 0 & 0 & 0 & 3.0000 \end{matrix}$$

$rank(L) = 3$ $\qquad$ $\beta_0 = 2$

Theorem: $\beta_0 = p - rank(L)$

Proof. $$rank(L) = rank(MM^\top) = rank(M^\top)$$

$$kerL = \{\sigma : L\sigma = 0\} = \{\sigma : M^\top\sigma = 0\}$$



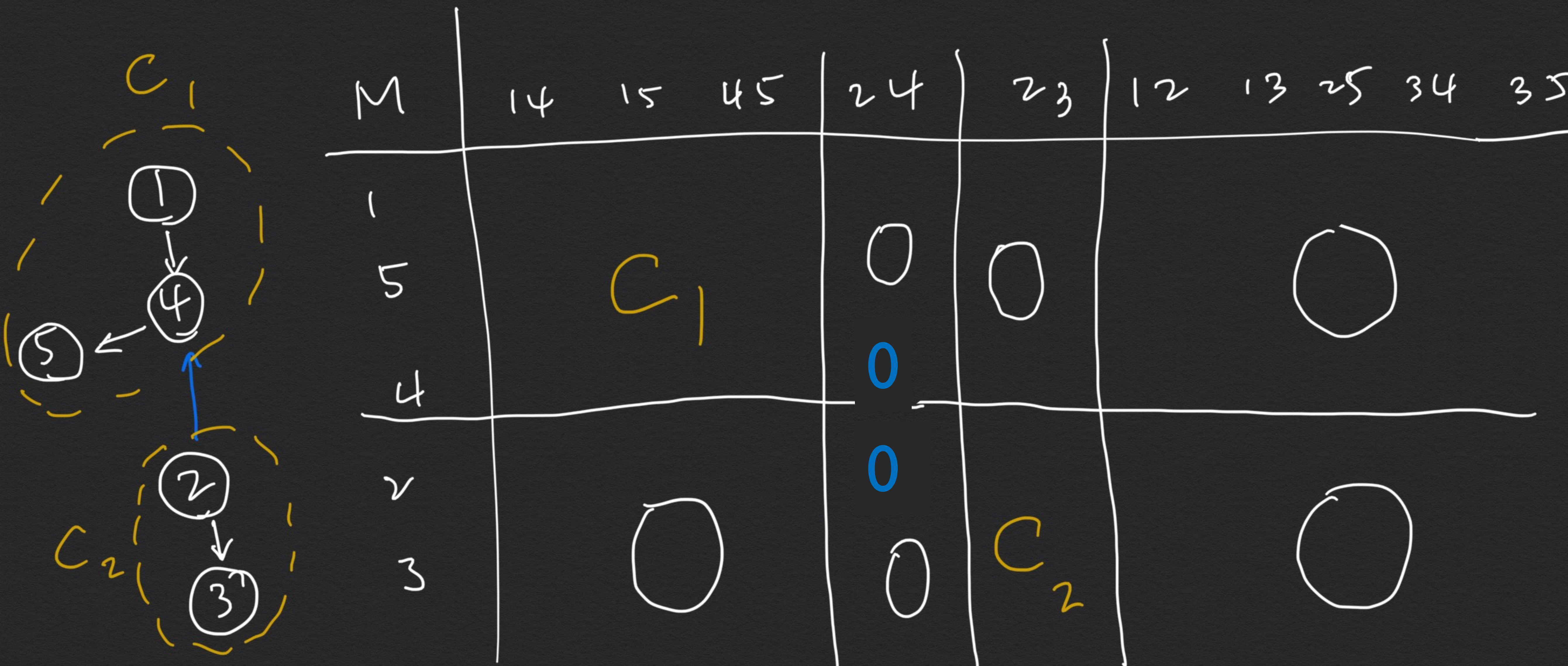
$$rank(kerL) = 4$$

$$\sigma = [a, a, b, c, d]^\top$$

$$rank(L) = 1$$

Graph with 2 disconnected components

$$ker L = \{\sigma : L\sigma = 0\} = \{\sigma : M^\top \sigma = 0\}$$



$$\sigma = [a, a, a, b, b]^\top$$

$$rank(ker L) = \beta_0 = 2$$

$$rank(L) = p - 2$$

Rank-nullity theorem for graph Laplacian

$$kerL = \{\sigma : L\sigma = 0\} = \{\sigma : M^\top \sigma = 0\}$$

$$imgL = \{L\sigma : \sigma \in \mathbb{R}^p\}$$

$$rank(imgL) + rank(kerL) = p$$

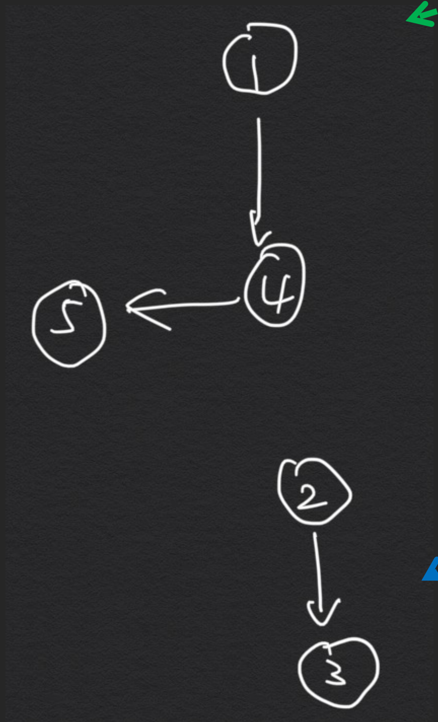
$rank(L)$ β_0

How to find the connected components?

$$kerL = \{\sigma : L\sigma = 0\} = \{\sigma : M^\top \sigma = 0\}$$

The kernel space is spanned by the eigenvectors corresponding to zero eigenvalue

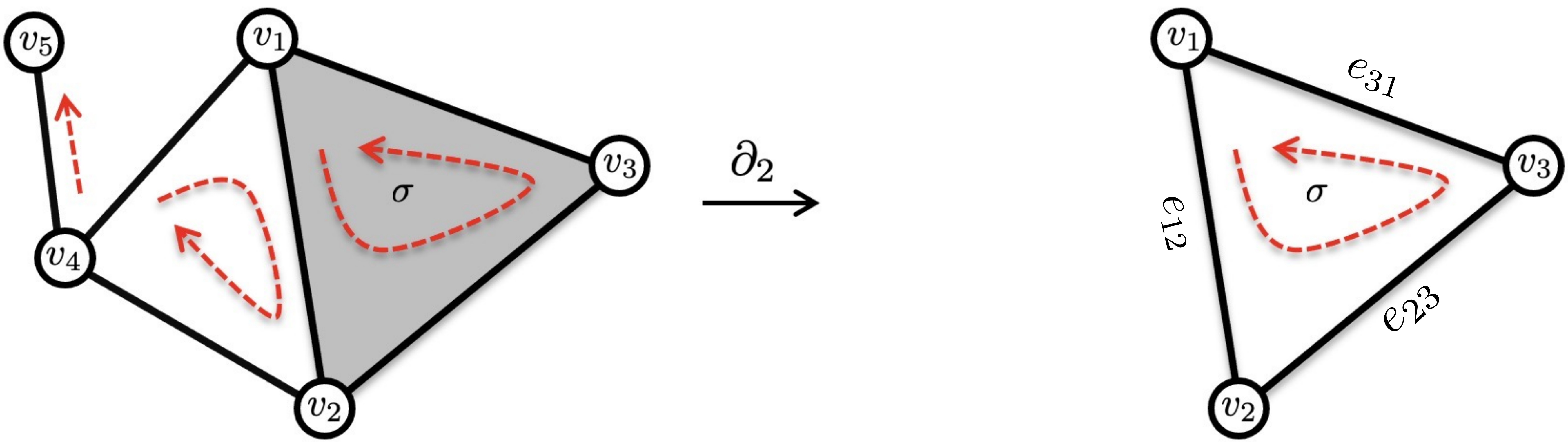
```
>>  [F, D] = eig(L)

F =
         0   -0.5774   -0.7071         0    0.4082
         0   -0.5774    0.0000         0   -0.8165
         0   -0.5774    0.7071         0    0.4082
   -0.7071         0         0   -0.7071         0
   -0.7071         0         0    0.7071         0

D =
         0         0         0         0         0
         0    0.0000         0         0         0
         0         0    1.0000         0         0
         0         0         0    2.0000         0
         0         0         0         0    3.0000
```



Boundary operators ∂_k

∂_k Removes the filled-in interior of *k*-simplexes

$\partial_k : C_k \to C_{k-1}$



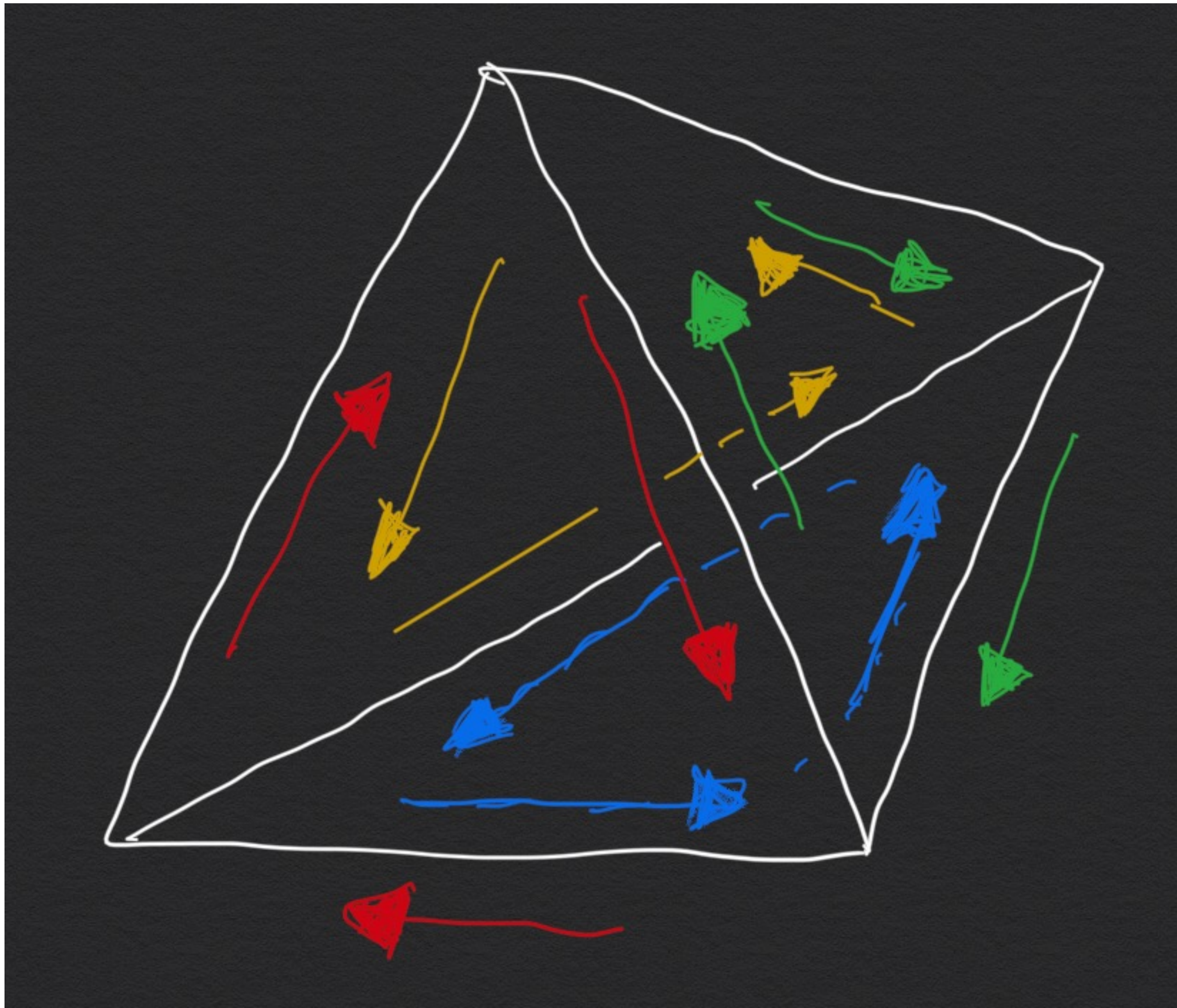
$$\partial_2 \sigma = [v_1, v_2] + [v_2, v_3] + [v_3, v_1]$$

$$\partial_1 \partial_2 \sigma = 0$$

$$\partial_1 \partial_2 \sigma = v_2 - v_1 + v_3 - v_2 + v_1 - v_3 = 0$$

Theorem. $\partial_{k-1} \partial_k \sigma = 0$

Example. Boundary of boundary of a filled-in tetrahedron = edges



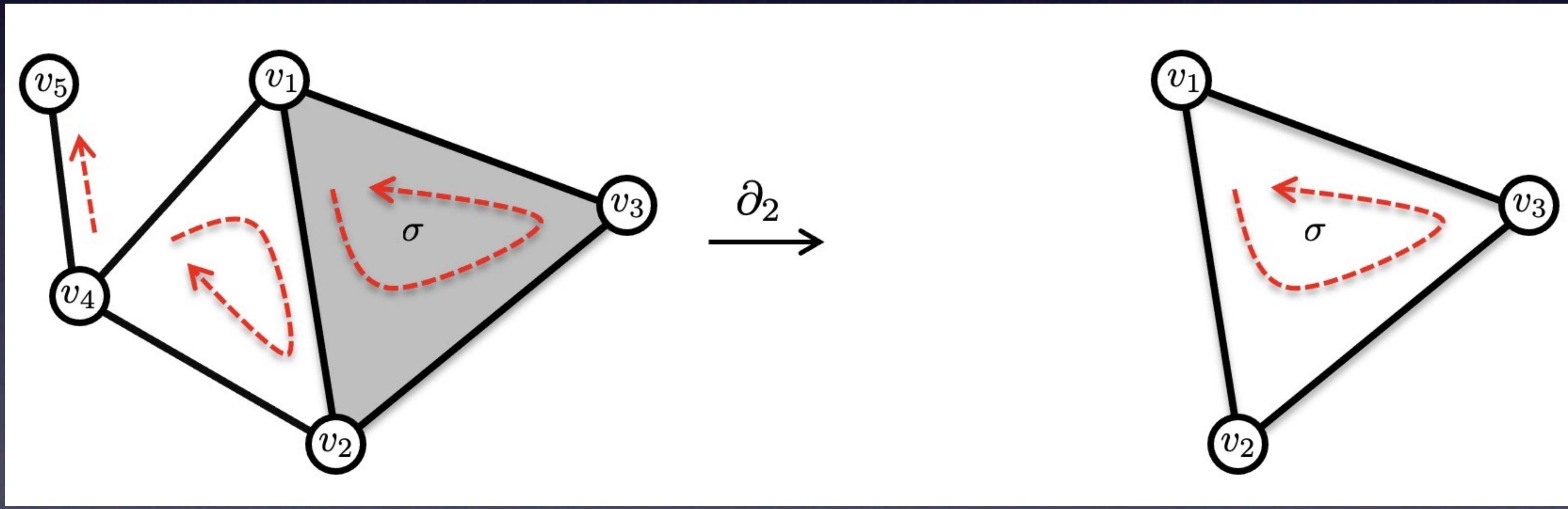
→ Does the sum of edges vanish?

Image of boundary operator

$$img\ \partial_{k+1} = \{\partial_{k+1}\sigma : \sigma \in C_{k+1}\}$$

collection of *k*-boundaries

$\partial_2\sigma = [v_1, v_2] + [v_2, v_3] + [v_3, v_1]$ is 1-boundary



Kernel of boundary operator

$$ker\ \partial_k = \{\sigma \in C_k : \partial_k \sigma = 0\}$$

collection of *k*-cycles

$$\partial_1 \partial_2 \sigma = v_2 - v_1 + v_3 - v_2 + v_1 - v_3 = 0$$

$$\partial_2 \sigma = [v_1, v_2] + [v_2, v_3] + [v_3, v_1]$$ is 1-cycle

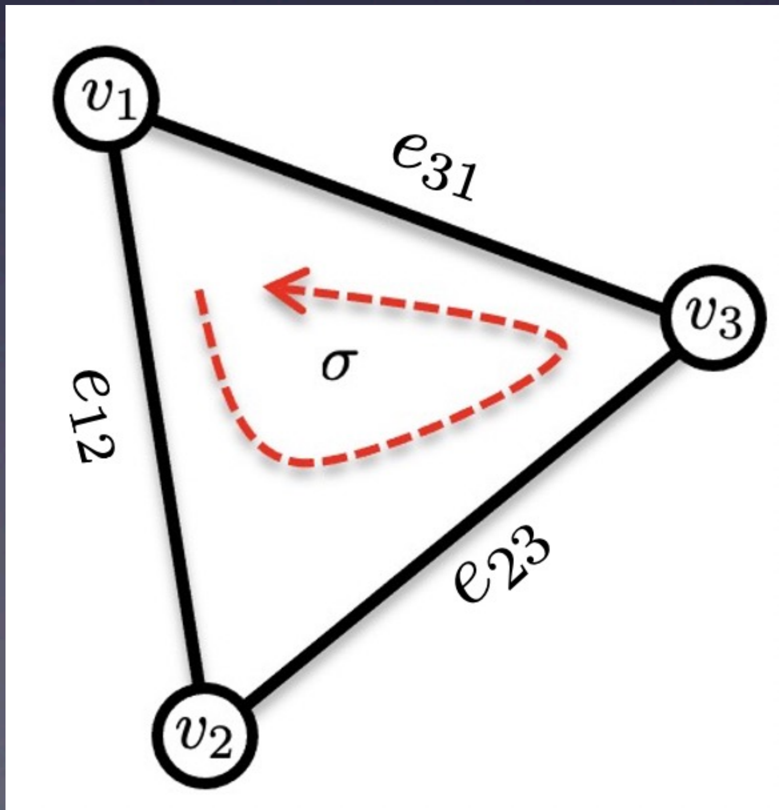


Image of boundary operator

From theorem $\partial_k \partial_{k+1} \sigma = 0$
boundary $\partial_{k+1}\sigma$ is always a cycle.

$$ker\partial_k \supset img\partial_{k+1}$$

Set of cycles Set of boundaries

Quotient space:

$$H_k = ker\partial_k / img\partial_{k+1}$$

Total number of algebraically independent cycles (# of basis).

$$\beta_k = rank(ker\partial_k) - rank(img\partial_{k+1})$$

Lee et al. 2014 (MICCAI 8675:297-304)

Boundary matrix = incidence matrix ∂_k

of *k*-dimensional simplices σ_j

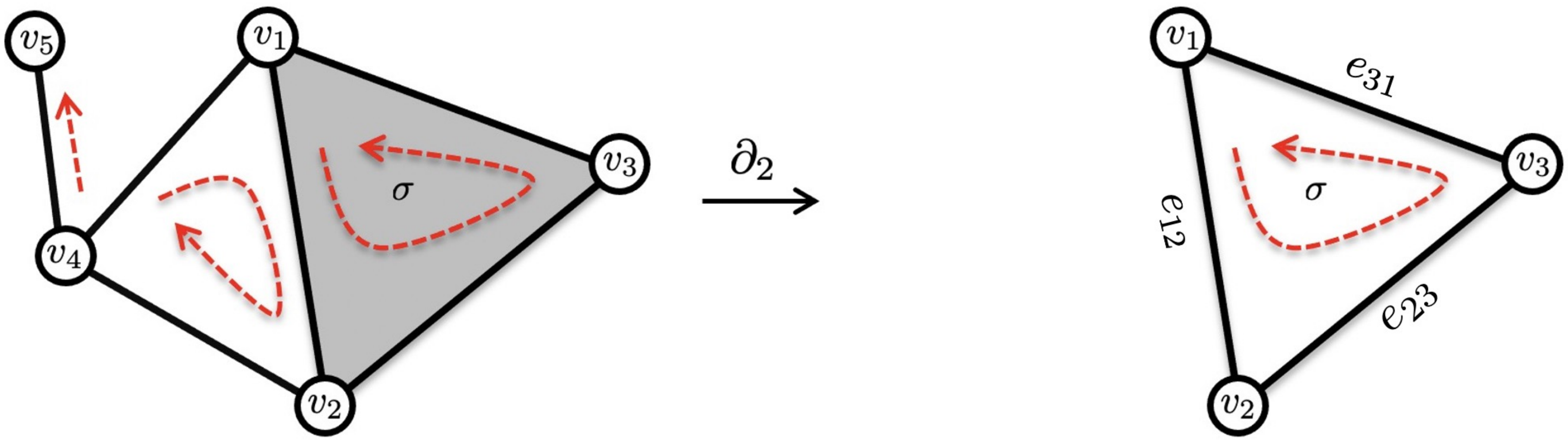
of (*k-1*)-dimensional simplices τ_i

$$
\partial_k = \begin{array}{|ccccccc|}
\hline
1 & 0 & 1 & \dots & 1 & 0 & 1 \\
\cdot & & & & \cdot & & \\
\cdot & & & & \cdot & & \\
\cdot & & & & \cdot & & \\
1 & 1 & & & & 0 & 0 \\
0 & 1 & 0 & \dots & 0 & 1 & -1 \\
\hline
\end{array}
$$

(*i,j*)-th entry = 1 if $\tau_i \subset \sigma_j$

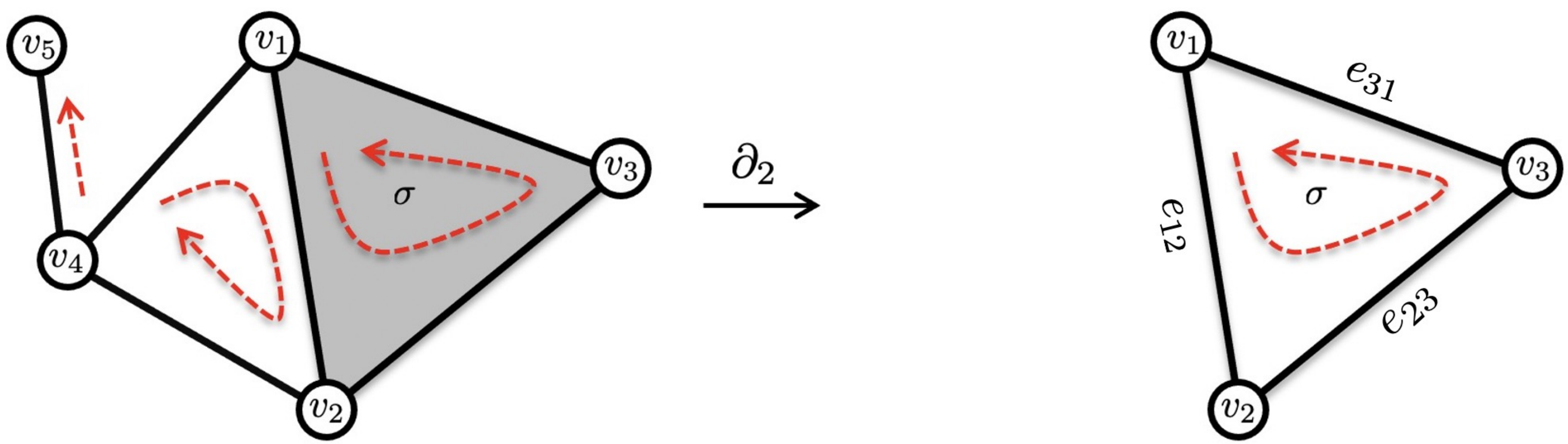
Sign depends on the orientation of τ_i

Boundary matrix ∂_0



$$\boldsymbol{\partial}_0 = \begin{array}{cc} & \begin{array}{ccccc} v_1 & v_2 & v_3 & v_4 & v_5 \end{array} \\ 0 & \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \end{pmatrix} \end{array}$$

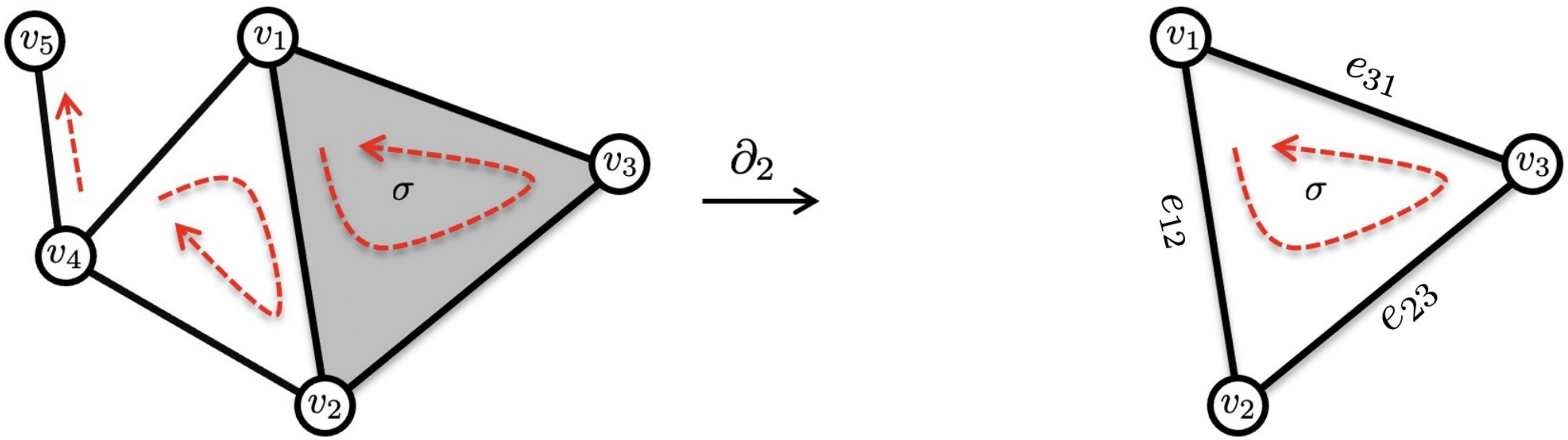
Boundary matrix ∂_1



$$\partial_1 \partial_2 \sigma = v_2 - v_1 + v_3 - v_2 + v_1 - v_3 = 0$$

$$\boldsymbol{\partial}_1 = \begin{array}{c} \\ v_1 \\ v_2 \\ v_3 \\ v_4 \\ v_5 \end{array} \begin{array}{c} \begin{array}{cccccc} \overbrace{e_{12} \quad e_{23} \quad e_{31}}^{\sigma} & e_{24} & e_{41} & e_{45} \end{array} \\ \begin{pmatrix} -1 & 0 & 1 & 0 & 1 & 0 \\ 1 & -1 & 0 & -1 & 0 & 0 \\ 0 & 1 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & -1 & -1 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \end{array}$$

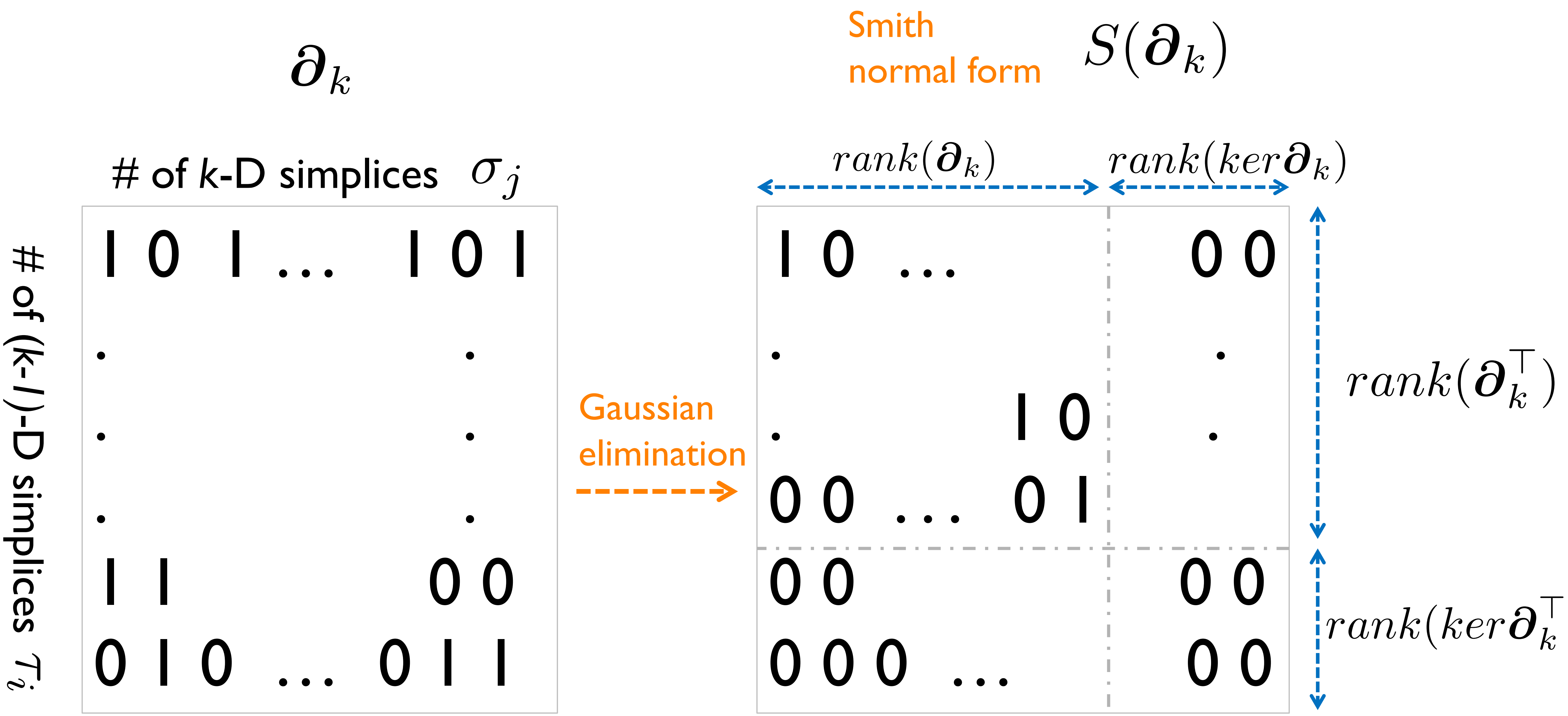
Boundary matrix ∂_2



$$\partial_2 \sigma = [v_1, v_2] + [v_2, v_3] + [v_3, v_1]$$

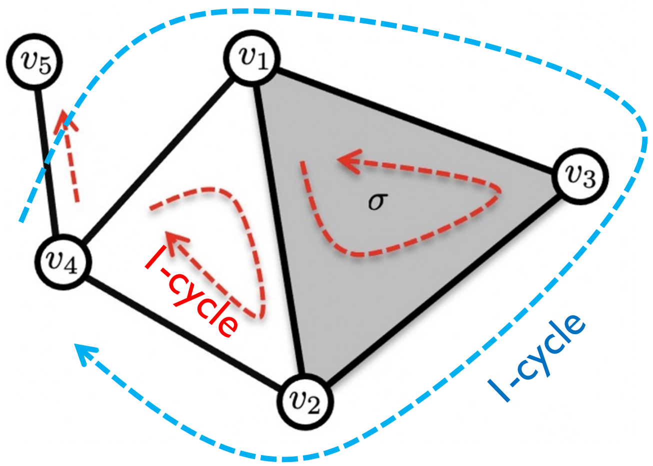
$$\boldsymbol{\partial}_2 = \begin{array}{c c} & \sigma \\ \begin{array}{c} e_{12} \\ e_{23} \\ e_{31} \\ e_{24} \\ e_{41} \\ e_{45} \end{array} & \begin{pmatrix} 1 \\ 1 \\ 1 \\ 0 \\ 0 \\ 0 \end{pmatrix} \end{array}$$

Rank nullity theorem for boundary matrix



$$\beta_k = rank(ker\boldsymbol{\partial}_k) - rank(\boldsymbol{\partial}_{k+1})$$

Computing Betti numbers



$$\mathcal{S}(\boldsymbol{\partial}_1) = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}, \quad \mathcal{S}(\boldsymbol{\partial}_2) = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

$$\beta_0 = rank(ker\boldsymbol{\partial}_0) - rank(\boldsymbol{\partial}_1) \;\; = 5\text{-}4$$

$$\beta_1 = rank(ker\boldsymbol{\partial}_1) - rank(\boldsymbol{\partial}_2) \;\; = 2\text{-}1$$

Far simpler approach given in [*Chung et al. 2019 Network Neuroscience 3:674-694*]

k-th Hodge Laplacian

$$\Delta_k = \partial_{k+1}\partial_{k+1}^\top + \partial_k^\top \partial_k$$

0th Hodge Laplacian
Graph Laplacian

$$\Delta_0 = \underbrace{\boxed{\partial_1\partial_1^\top}}_{}$$

# of nodes × # of nodes

~~1st Hodge Laplacian~~ $\cancel{\Delta_1 = \partial_2\partial_2^\top + \partial_1^\top\partial_1}$

1st Hodge Laplacian for graphs: $\Delta_1 = \boxed{\partial_1^\top\partial_1}$

# of edges × # of edges

k-th Hodge Laplacian

$$\Delta_k = \partial_{k+1}\partial_{k+1}^\top + \partial_k^\top \partial_k$$

0th Hodge Laplacian
Graph Laplacian

# of nodes

$$\Delta_0 = \boxed{\partial_1 \partial_1^\top}$$

# of nodes

1st Hodge Laplacian

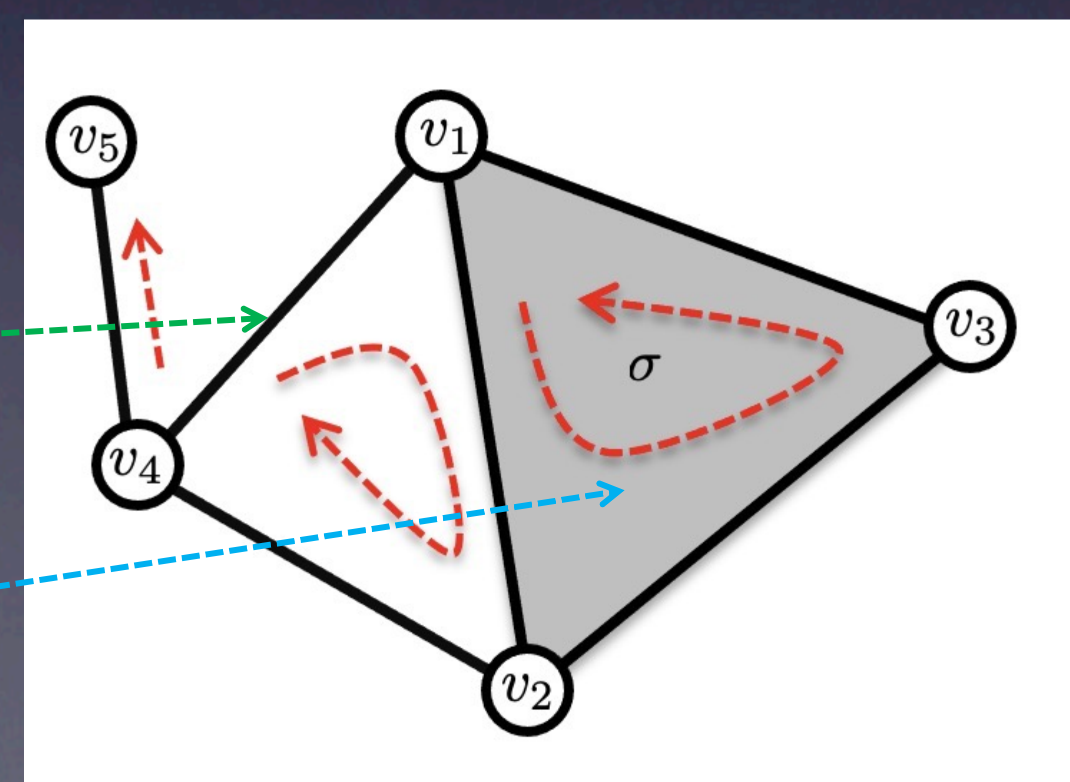
# of edges

# of edges

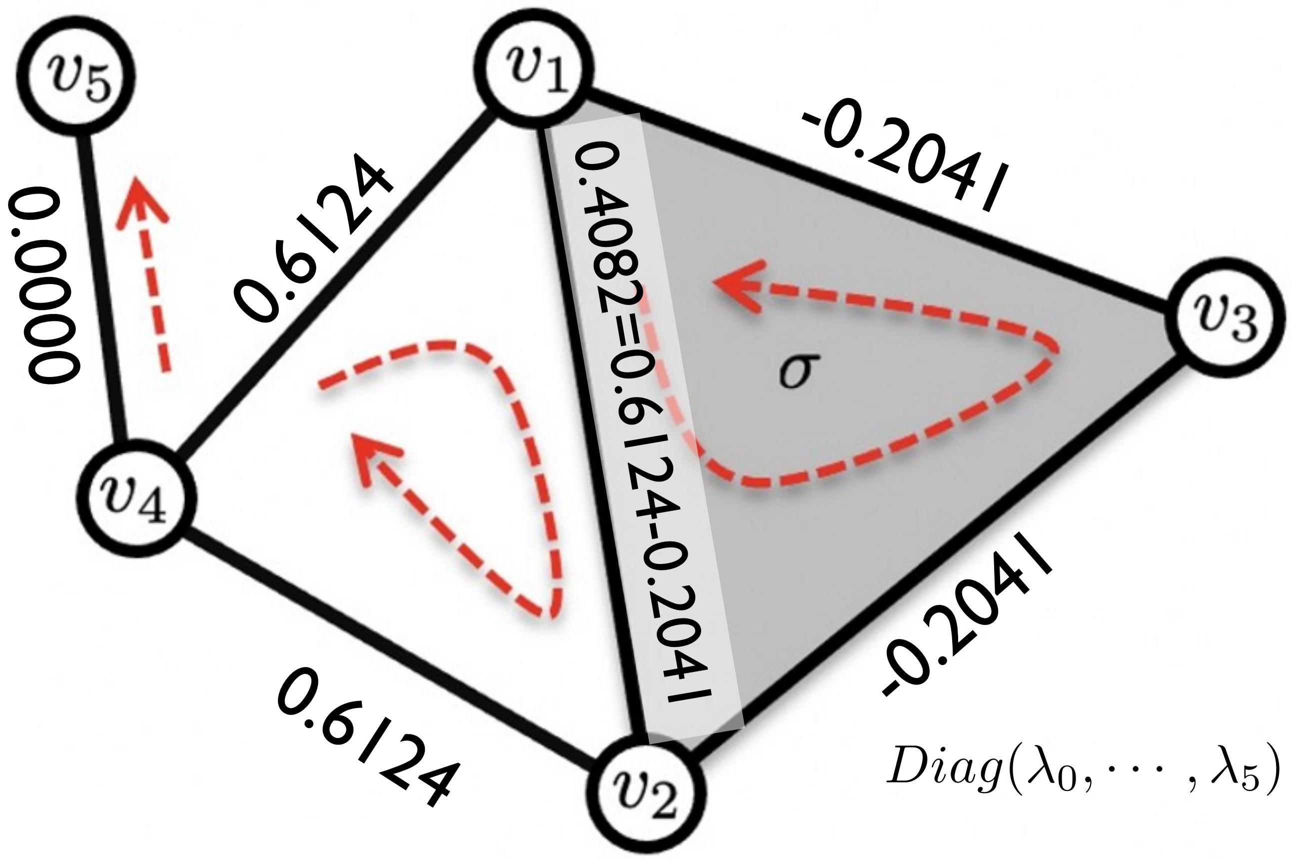
$$\Delta_1 = \boxed{\partial_2 \partial_2^\top} + \boxed{\partial_1^\top \partial_1}$$

# of edges

# of edges



Eigenvectors of Hodge Laplacian of zero eigenvalue



$$Diag(\lambda_0, \cdots, \lambda_5) = \begin{pmatrix} 0.0000 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0.8299 & 0 & 0 & 0 & 0 \\ 0 & 0 & 2.6889 & 0 & 0 & 0 \\ 0 & 0 & 0 & 3.0000 & 0 & 0 \\ 0 & 0 & 0 & 0 & 4.0000 & 0 \\ 0 & 0 & 0 & 0 & 0 & 4.4812 \end{pmatrix}$$

$$(\mathbf{f}_0, \cdots, \mathbf{f}_5) = \begin{pmatrix} 0.4082 & -0.0000 & 0.0000 & 0.5774 & 0.7071 & -0.0000 \\ -0.2041 & 0.1993 & -0.5765 & 0.5774 & -0.3536 & 0.3578 \\ -0.2041 & -0.1993 & 0.5765 & 0.5774 & -0.3536 & -0.3578 \\ 0.6124 & -0.4325 & 0.1793 & -0.0000 & -0.3536 & 0.5299 \\ 0.6124 & 0.4325 & -0.1793 & -0.0000 & -0.3536 & -0.5299 \\ 0.0000 & -0.7392 & -0.5207 & 0.0000 & 0.0000 & -0.4271 \end{pmatrix}$$

1st Hodge Laplacian for graphs

1st Hodge Laplacian for 2-simplex

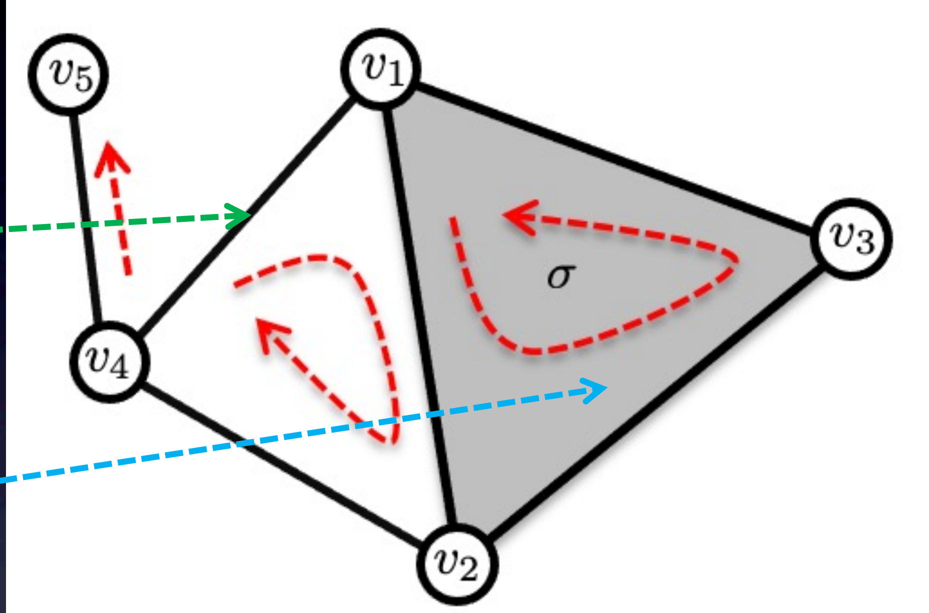
of edges

# of edges

$$\Delta_1 = \boxed{\partial_2 \partial_2^\top} + \boxed{\partial_1^\top \partial_1}$$

# of edges

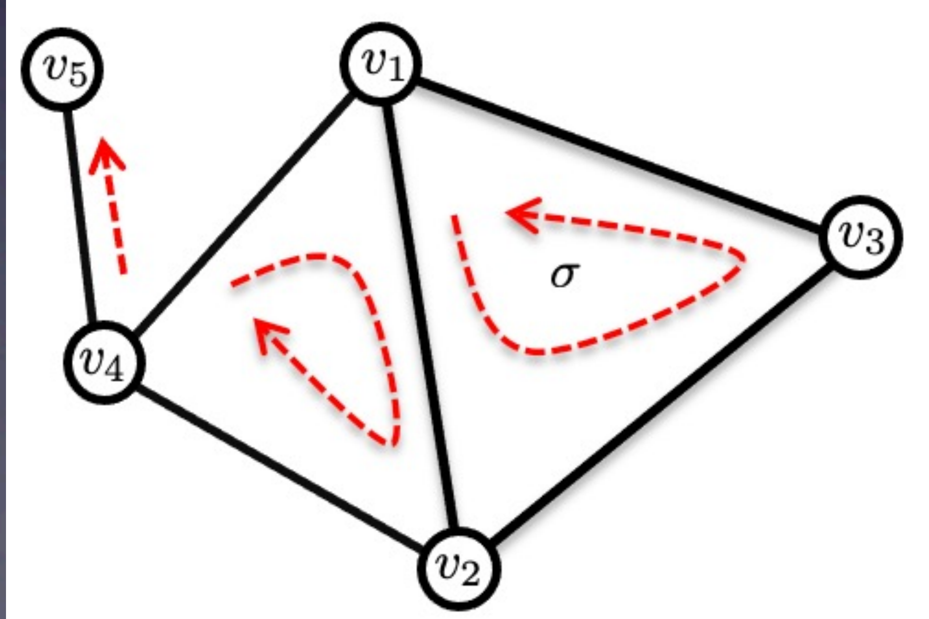
# of edges



# of edges

$$\Delta_1 = \boxed{\partial_1^\top \partial_1}$$

# of edges



vector representation of 1-cycles

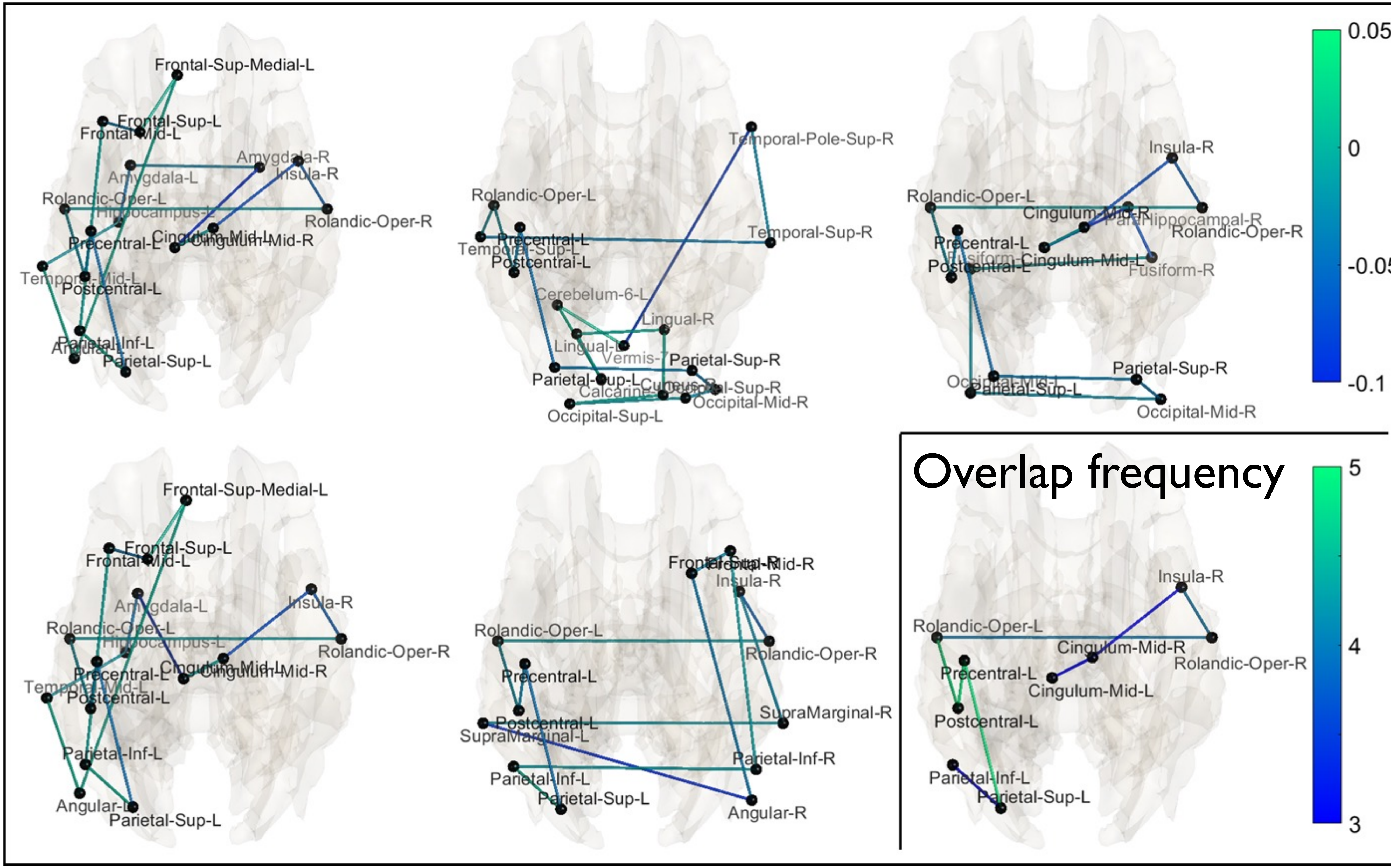
$$cycle = \sum_{i<j} a_{ij} e_{ij}$$

$$\begin{array}{cc} & \beta_1 \\ \text{\# of edges} & \begin{bmatrix} 2.3 & \ldots & 1.39 \\ . & & . \\ . & & . \\ . & & . \\ -1.4 & \ldots & 0 \end{bmatrix} \end{array} \xrightarrow{\text{Gaussian elimination}} \begin{array}{cc} & \beta_1 \\ \text{\# of edges} & \begin{bmatrix} 1 & 0\ldots & & 0 \\ 1 & 0 & . & 0 \\ 0 & 1 & . & 0 \\ 0 & 1 & . & 1 \\ 1. & \ldots & 0 & 1 \end{bmatrix} \end{array}$$

Linearly independent columns

More meaningful representation

Five biggest cycle differences (male – female) in HCP

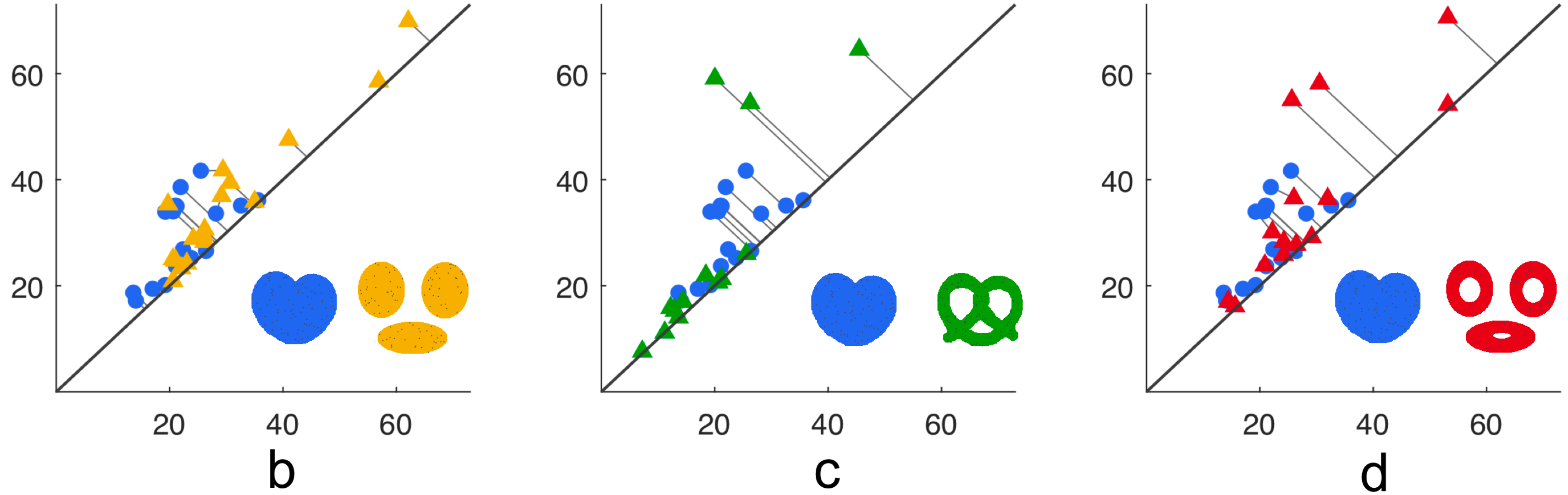


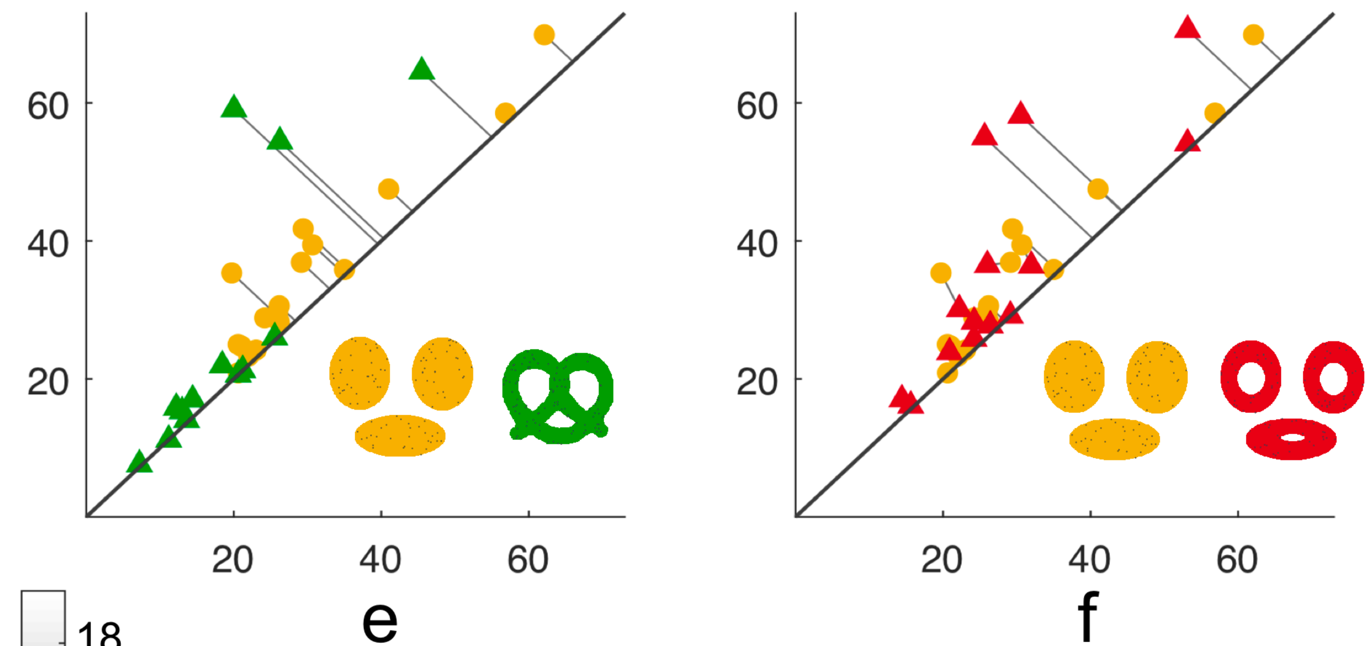
Anand et al. 2021, IEEE TMI arXiv:2110.14599

p-value = 0.03

Topological distances

Bottleneck distance





For definition, simply
read the assigned reading PDF