An Introduction to Topological Data Analysis

Yuan Yao

Department of Mathematics
HKUST

Outline

Methods for Imposing a Geometry

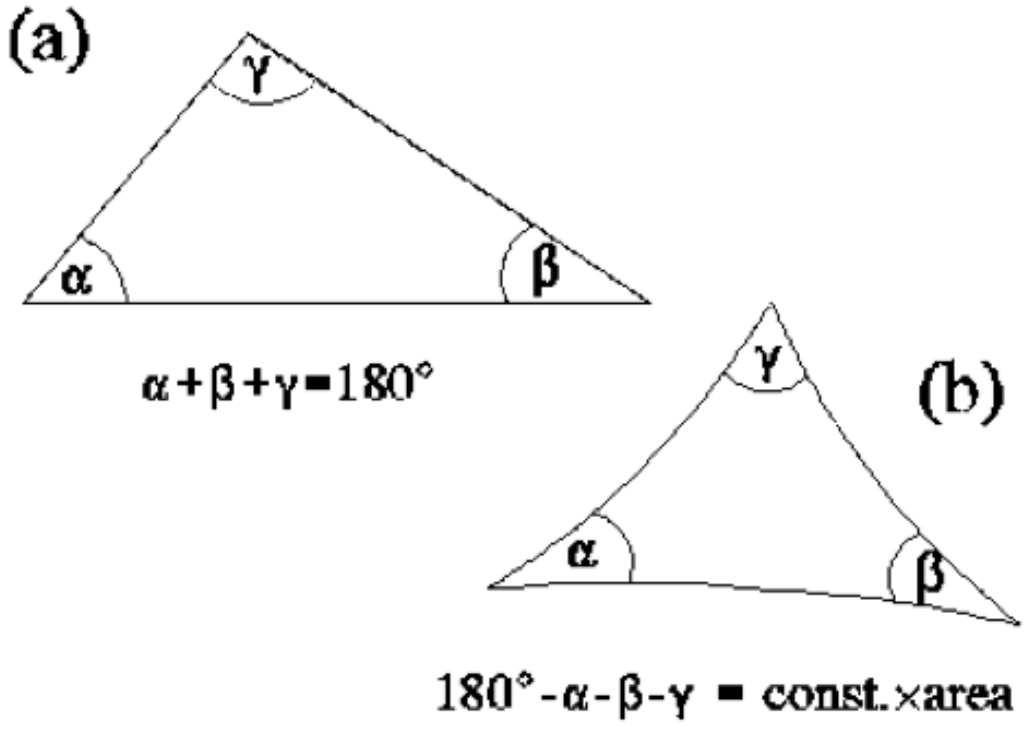


Figure: Define a metric

Methods for Summarizing or Visualizing a Geometry

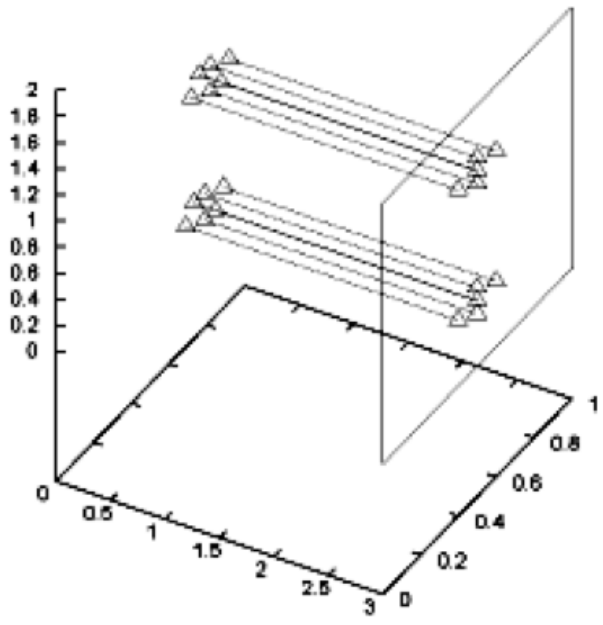


Figure: Linear projection (PCA, MDS, etc. Euclidean Metric)

Methods for Summarizing or Visualizing a Geometry

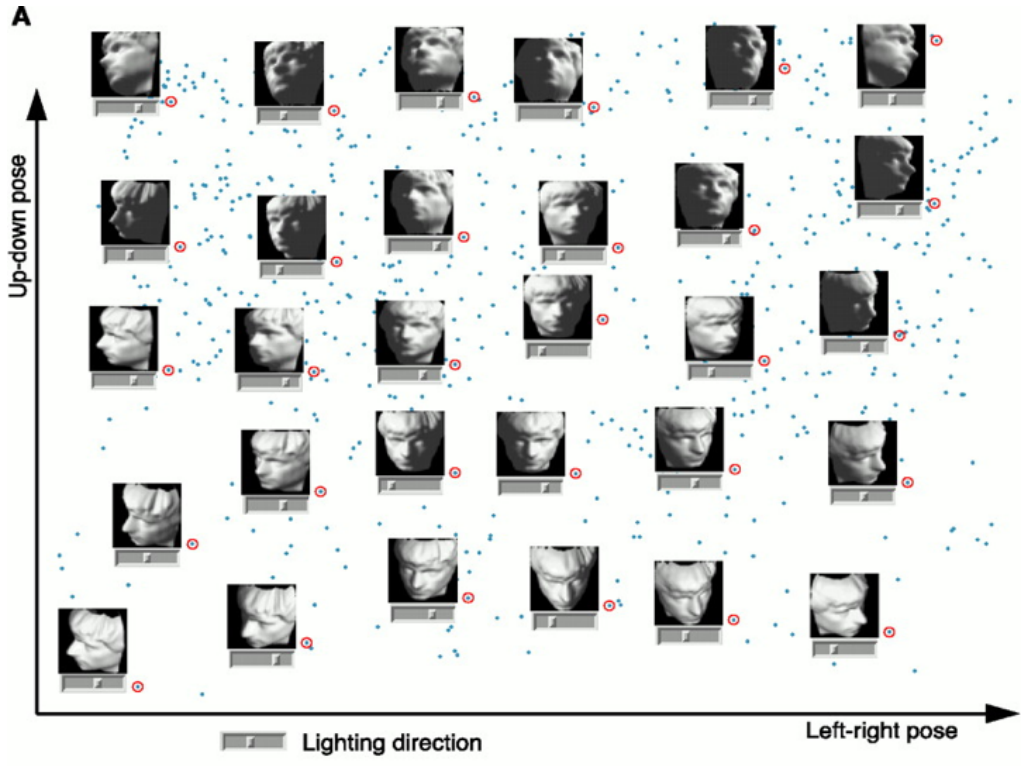


Figure: Nonlinear Dimensionality Reduction (ISOMAP, LLE etc. Riemannian Metric)

Geometric Data Reduction

- General method of manifold learning takes the following Spectral Kernal Embedding approach
  - construct a neighborhood graph of data, G
  - construct a positive semi-definite kernel on graphs, K
  - find global embedding coordinates of data by eigen-decomposition of $K = YY^T$
- Geometric reconstruction can be relaxed via Semi-definite Programming (SDP)
- Sometimes 'distance metric' is just a similarity measure (nonmetric MDS, ordinal embedding)
- Sometimes coordinates are not a good way to organize/visualize the data (e.g. $d > 3$)
- Sometimes all that is required is a qualitative view

Methods for Summarizing or Visualizing a Geometry

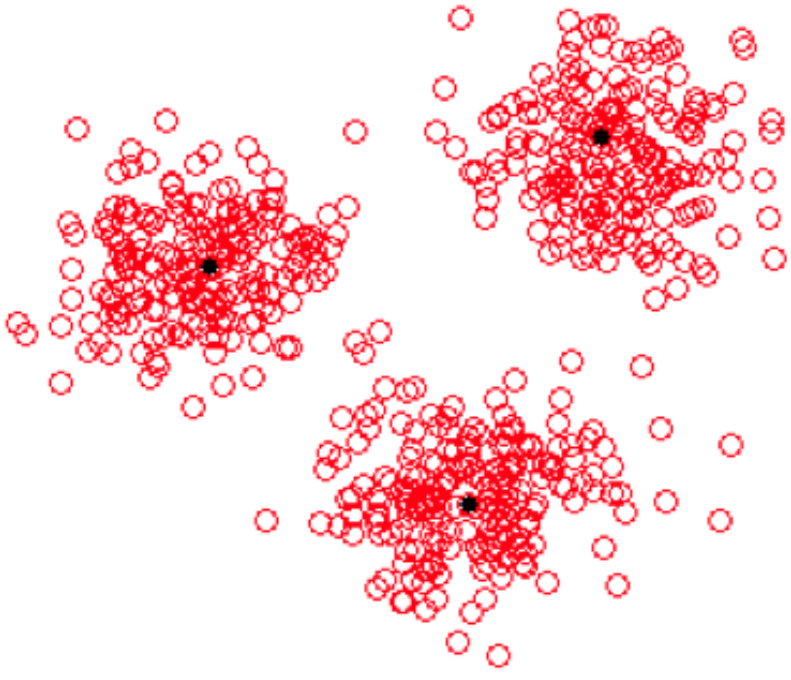


Figure: Clustering the data

Methods for Summarizing or Visualizing a Geometry

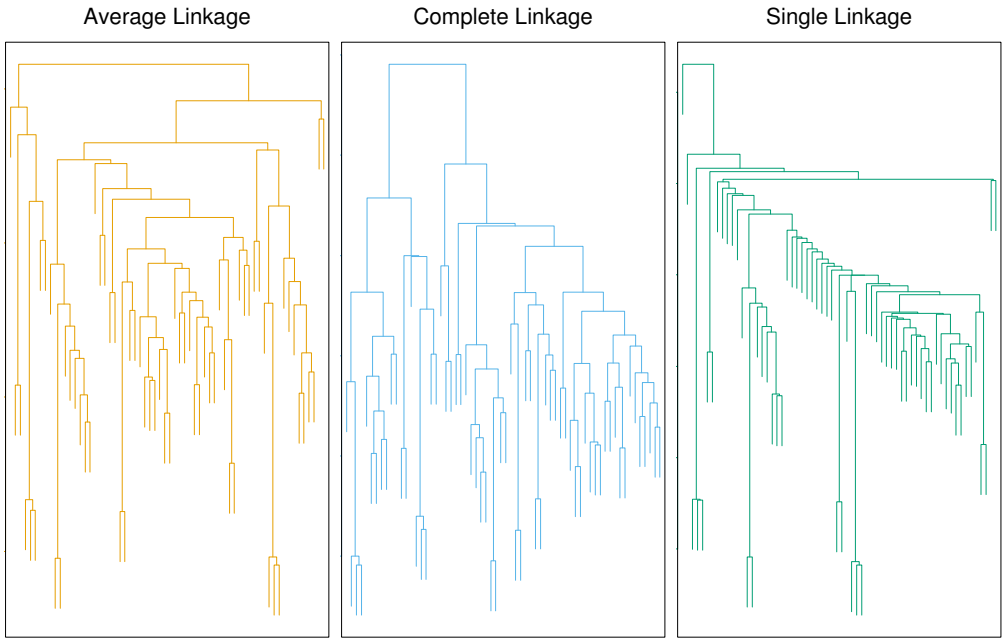


Figure: Cluster trees: Average, complete, and single linkage. From *Introduction to Statistical Learning with Applications in R*.

Hierarchical Cluster Trees

1. Start with each data point as its own cluster;
2. Repeatedly merge two "closest" clusters, where notions of "distance" between two clusters are given by:
   - Single linkage: closest pair of points
   - Complete linkage: furthest pair of points
   - Average linkage (several variants):
     (i) distance between centroids
     (ii) average pairwise distance
     (iii) Ward's method: increase in k-means cost due to merger

Methods for Summarizing or Visualizing a Geometry

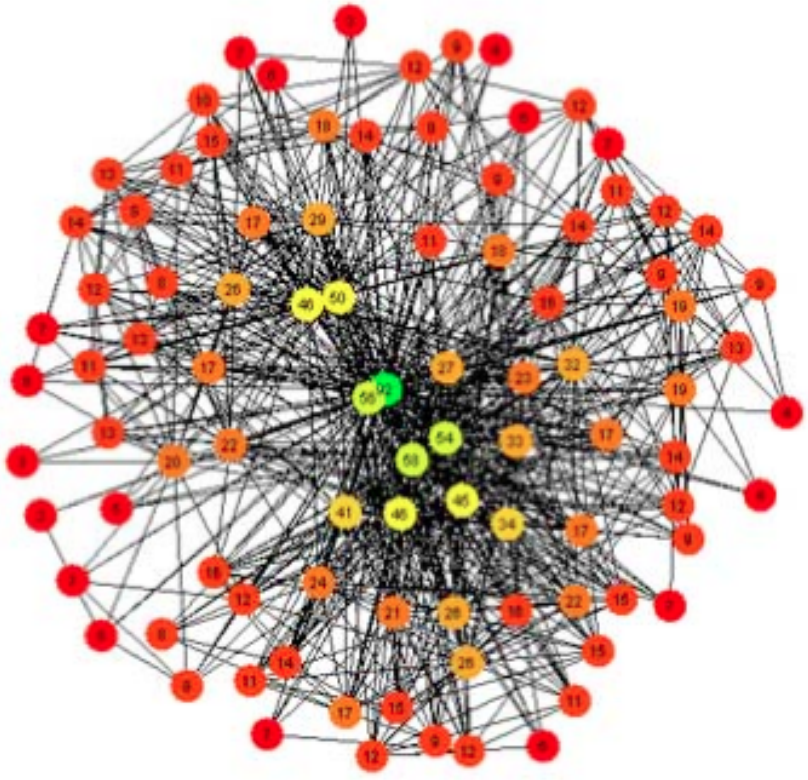


Figure: Define a graph or network structure

Topology

- Origins of Topology in Math
  - Leonhard Euler 1736, Seven Bridges of Königsberg
  - Johann Benedict Listing 1847, Vorstudien zur Topologie
  - J.B. Listing (orbituary) Nature 27:316-317, 1883. "qualitative geometry from the ordinary geometry in which quantitative relations chiefly are treated."

RNA hairpin folding pathways

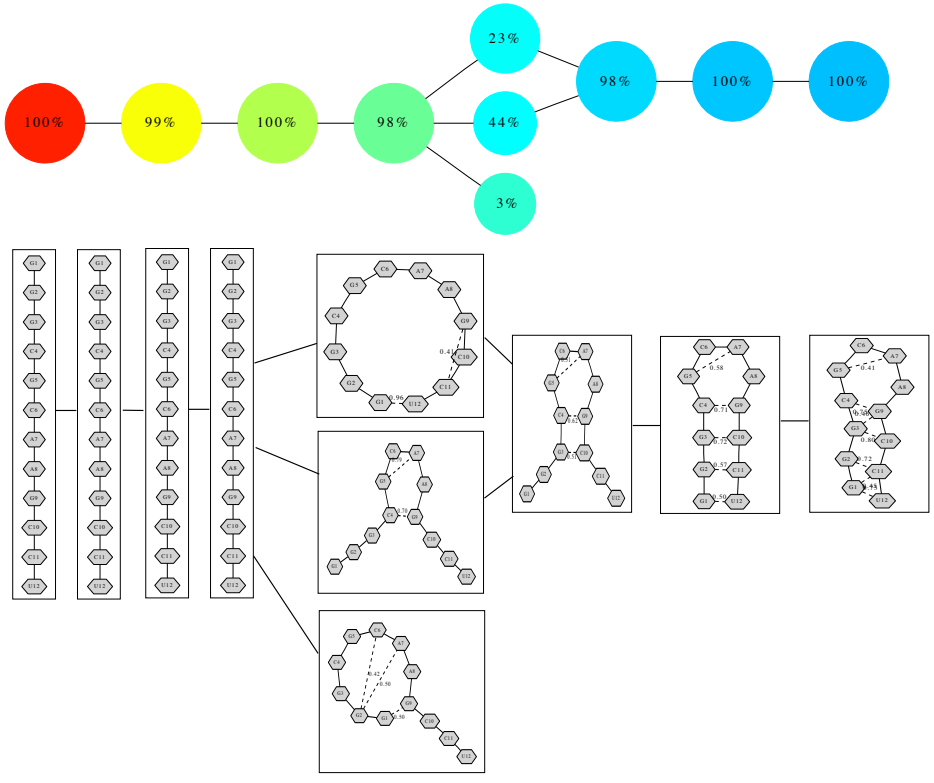


Figure: Jointly with Xuhui Huang, Jian Sun, Greg Bowman, Gunnar Carlsson, Leo Guibas, and Vijay Pande, *JACS'08*, *JCP'09*

Differentiation process from murine embryonic stem cells to motor neurons

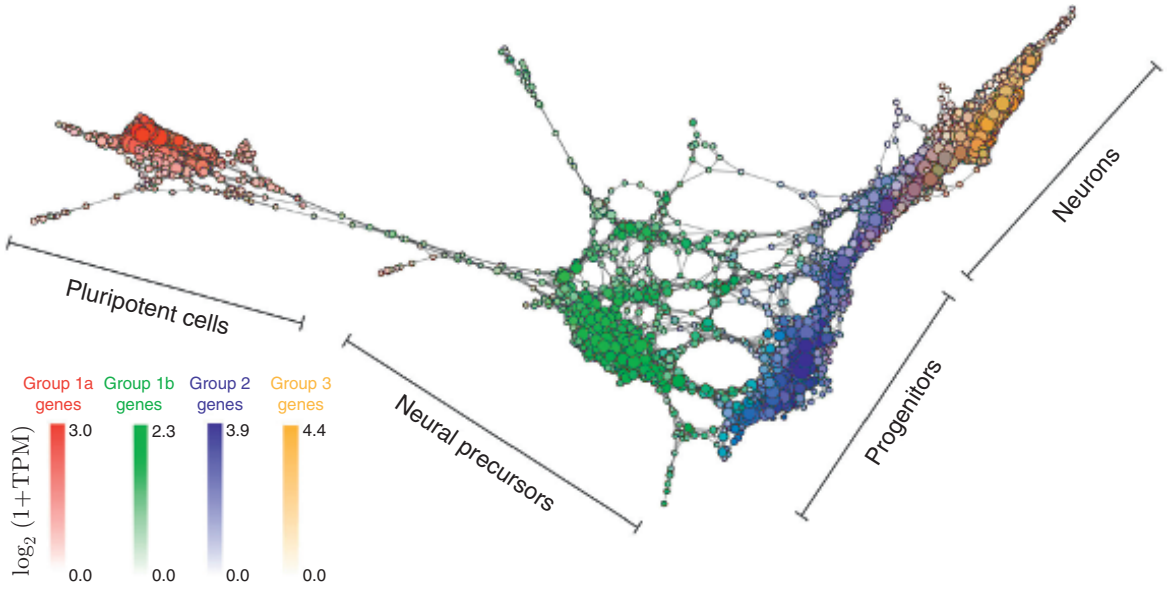


Figure: Mapper graph of single cell data, where the different regions in the Mapper graph nicely line up with different points along the differentiation timeline. Rizvi et al. *Nature Biotechnol.* 35.6 (2017), 551-560.

Key elements

- Coordinate free representation
- Invariance under deformations
- Compressed qualitative representation

Topology in continuous spaces

- To see points in neighborhood the *same* requires distortion of distances, i.e. stretching and shrinking
- We do not permit *tearing*, i.e. distorting distances in a discontinuous way

Continous Topology

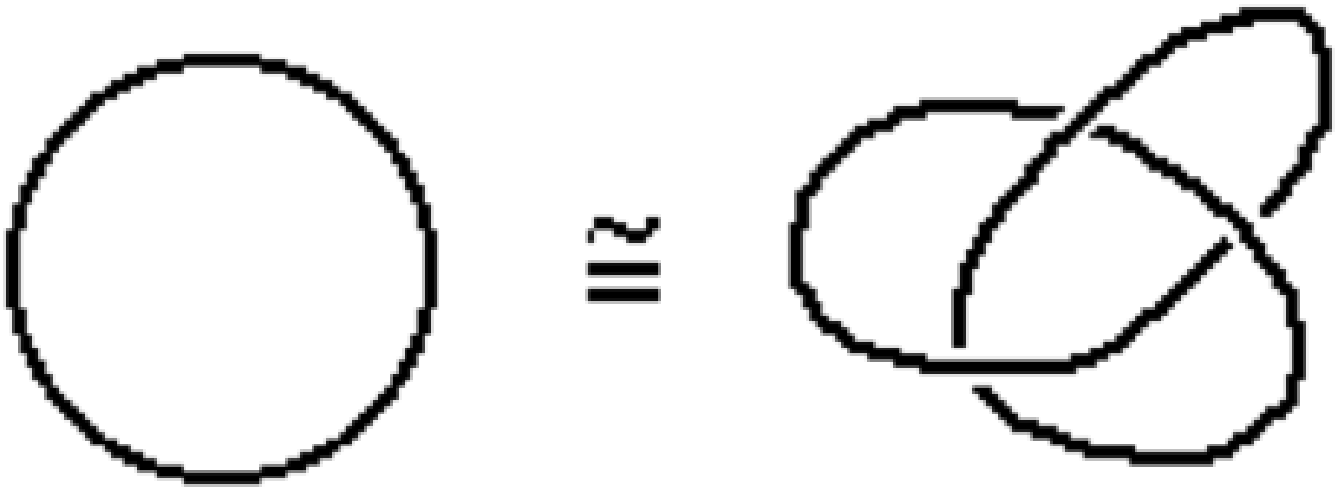


Figure: Homeomorphic

Continuous Topology

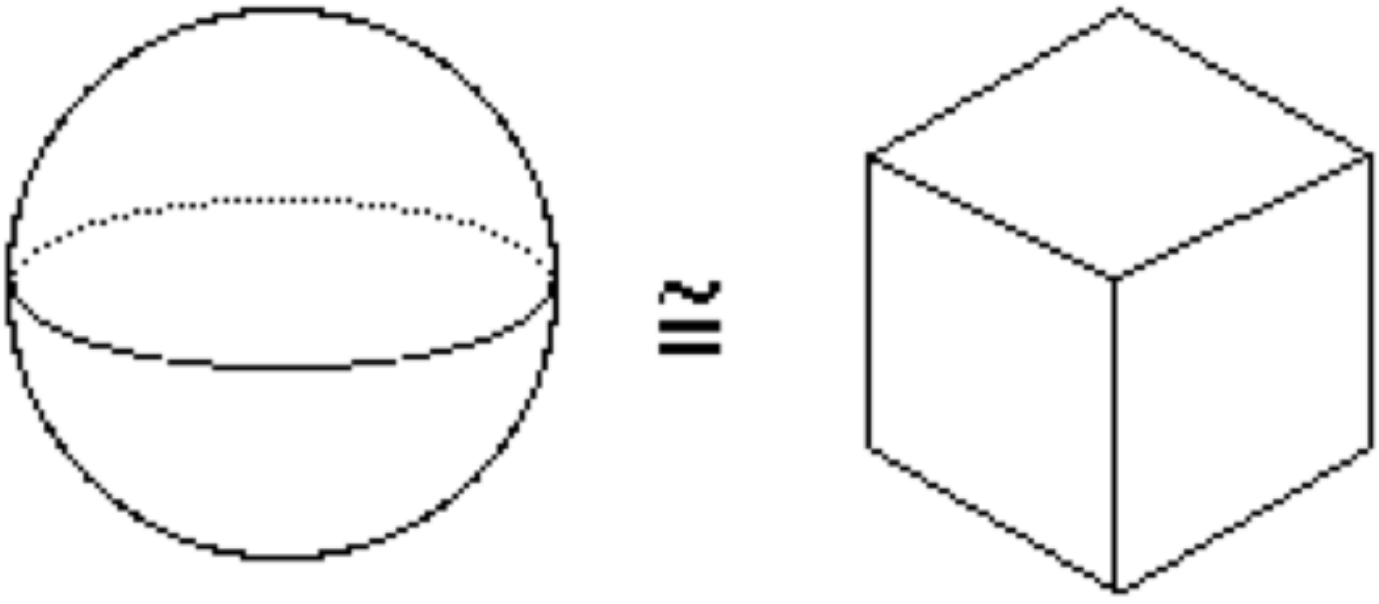


Figure: Homeomorphic

Discrete case?

How does topology make sense, in discrete and noisy setting?

Properties of Data Geometry

Fact

We Don't Trust Large Distances!

- In life or social sciences, distance (metric) are constructed using a notion of similarity (proximity), but have no theoretical backing (e.g. distance between faces, gene expression profiles, Jukes-Cantor distance between sequences)
- Small distances still represent similarity (proximity), but long distance comparisons hardly make sense

Properties of Data Geometry

Fact

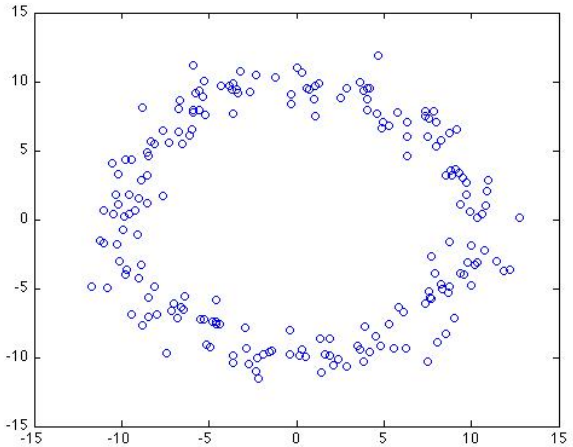
We Only Trust Small Distances a Bit!

- Both pairs are regarded as similar, but the strength of the similarity as encoded by the distance may not be so significant
- Similar objects lie in neighborhood of each other, which suffices to define topology

Properties of Data Geometry

Fact

Even Local Connections are Noisy, depending on observer's scale!



- Is it a circle, dots, or circle of circles?
- To see the circle, we ignore variations in small distance (tolerance for proximity)

So we need robust topology against metric distortions

- Distance measurements are noisy
- Physical device like human eyes may ignore differences in proximity (or as an average effect)
- Topology is the crudest way to capture invariants under distortions of distances
- At the presence of noise, one need topology varied with scales

What kind of topology?

- Topology studies (global) mappings between spaces
- Point-set topology: continuous mappings on open sets
- Differential topology: differentiable mappings on smooth manifolds
  - Morse theory tells us topology of continuous space can be learned by discrete information on critical points
- Algebraic topology: homomorphisms on algebraic structures, the most concise encoder for topology
- Combinatorial topology: mappings on simplicial (cell) complexes
  - Simplicial complex may be constructed from data
  - Algebraic, differential structures can be defined here

Topological Data Analysis

- What kind of topological information often useful
  - 0-homology: clustering or connected components
  - 1-homology: coverage of sensor networks; paths in robotic planning
  - 1-homology as obstructions: inconsistency in statistical ranking; harmonic flow games
  - high-order homology: high-order connectivity?
- How to compute homology in a stable way?
  - *simplicial complexes* for data representation
  - *filtration* on simplicial complexes
  - *persistent homology*

Outline

Simplicial Complexes for Data Representation

Definition (Simplicial Complex)

An abstract simplicial complex is a collection Σ of subsets of V which is closed under inclusion (or deletion), i.e. $\tau \in \Sigma$ and $\sigma \subseteq \tau$, then $\sigma \in \Sigma$.

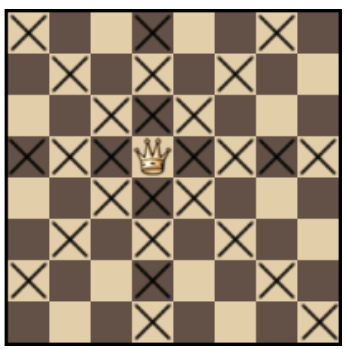
- Chess-board Complex
- Term-document cooccurance complex
- Nerve complex
- Point cloud data in metric spaces:
  - Čech, Rips, Witness complex
  - Mayer-Vietoris Blowup
- Clique complex in pairwise comparison graphs
- Strategic complex in game theory

Chess-board Complex

Definition (Chess-board Complex)

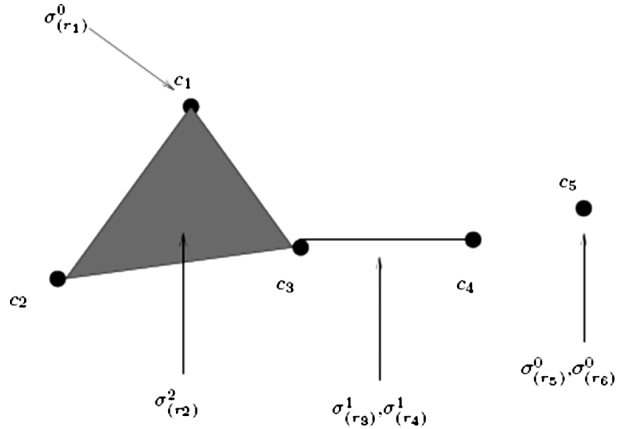
Let V be the positions on a Chess board. Σ collects position subsets of V where one can place queens (rooks) without capturing each other.

- Closedness under deletion: if $\sigma \in \Sigma$ is a set of "safe" positions, then any subset $\tau \subseteq \sigma$ is also a set of "safe" positions



Term-Document Co-occurrence Complex

	c_1	c_2	c_3	c_4	c_5
r_1	1	0	0	0	0
r_2	1	1	1	0	0
r_3	0	0	1	1	0
r_4	0	0	1	1	0
r_5	0	0	0	0	1
r_6	0	0	0	0	1



- Left is a term-document co-occurrence matrix
- Right is a simplicial complex representation of terms
- Connectivity analysis captures more information than Latent Semantic Index (Li & Kwong 2009)

Nerve complex

Definition (Nerve Complex)

Define a cover of X, $X = \cup_\alpha U_\alpha$. $V = \{U_\alpha\}$ and define $\Sigma = \{U_I : \cap_{\alpha \in I} U_I \neq \emptyset\}$.

- Closedness under deletion
- Can be applied to any topological space X

Nerve Theorem

Theorem (Nerve Theorem)

Consider the nerve complex of X,

$$\Sigma = \{U_I : \cap_{\alpha \in I} U_I \neq \emptyset, X = \cup_\alpha U_\alpha\}.$$

If every U_I is contractible, then X has the same homotopy type as Σ.

Nerve complex example

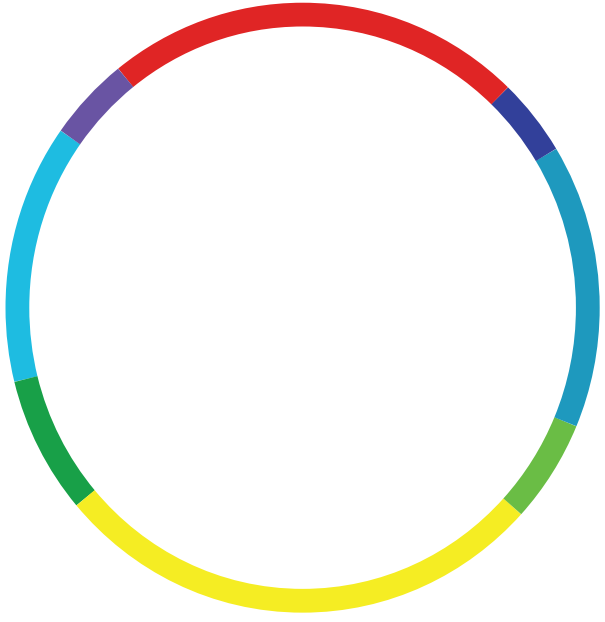


Figure: Covering of circle

Nerve complex example

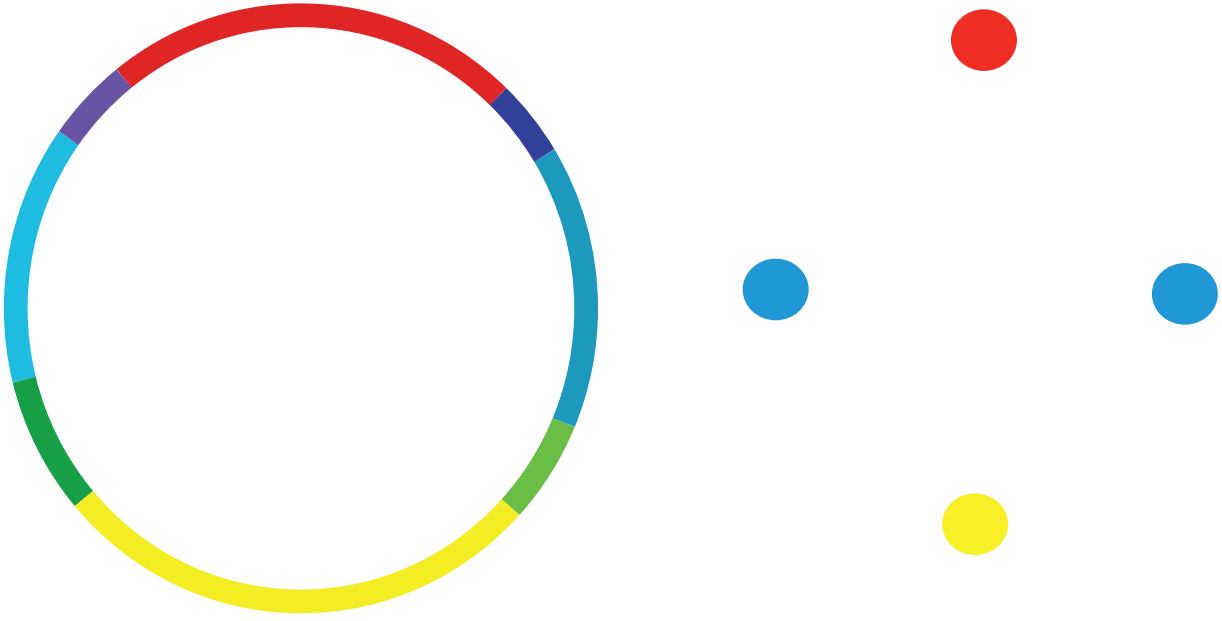


Figure: Create nodes

Nerve complex example

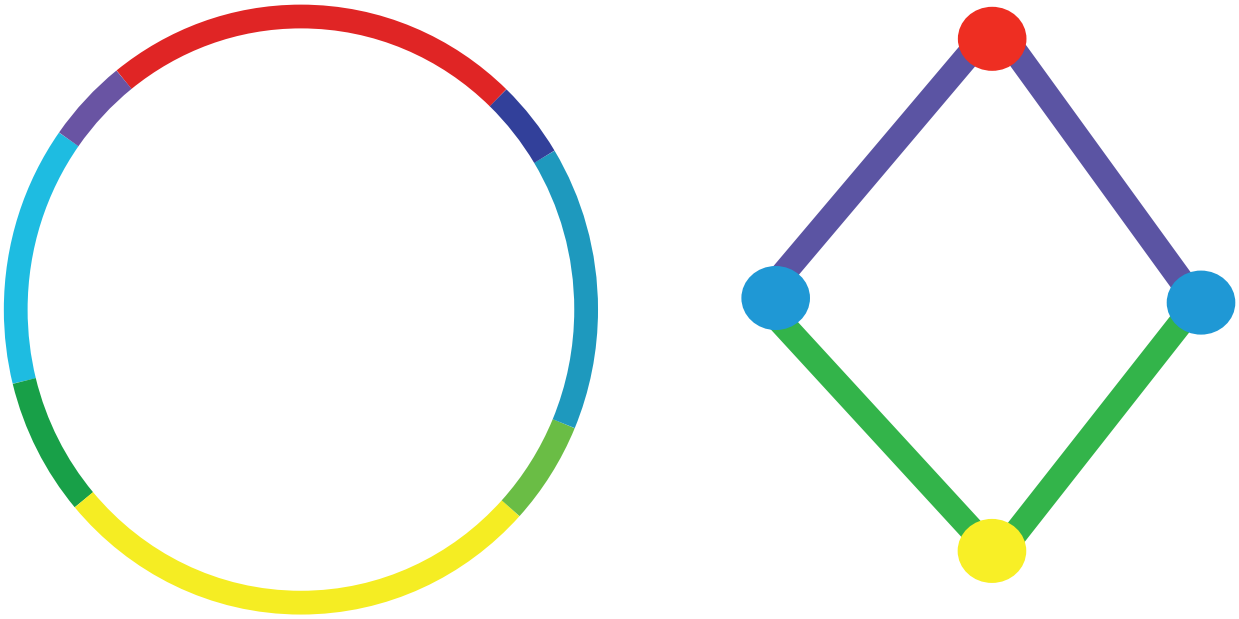


Figure: Create edges, that gives a Nerve complex (graph)

Nerve of Seven Bridges of Königsberg

Figure: Nerve graph of Seven Bridges of Könisberg

Point cloud data

- Now given point cloud data $\mathcal{X} = \{x_1, \ldots, x_n\}$, and a covering $V = \{U_\alpha\}$, where each U_α is a cluster of data
- Build a simplicial complex (Nerve) in the same way, but components replaced by clusters

Mapping

- How to choose coverings?
- Create a reference map (or filter) $h : \mathcal{X} \to \mathcal{Z}$, where $\mathcal{Z}$ is a topological space often with interesting metrics (e.g. $\mathbb{R}$, $\mathbb{R}^2$, S^1 etc.), and a covering $\mathcal{U}$ of $\mathcal{Z}$, then construct the covering of $\mathcal{X}$ using inverse map $\{h^{-1}U_\alpha\}$.

Example: Morse Theory and Reeb graph

- a nice (Morse) function: $h : \mathcal{X} \to \mathbb{R}$, on a smooth manifold $\mathcal{X}$
- topology of $\mathcal{X}$ reconstructed from level sets $h^{-1}(t)$
- topological of $h^{-1}(t)$ only changes at 'critical values'
- Reeb graph: a simplified version, contracting into points the connected components in $h^{-1}(t)$

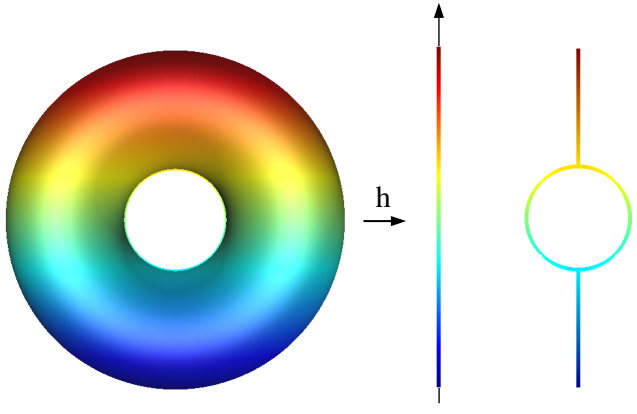


Figure: Construction of Reeb graph; h maps each point on torus to its height.

Mapper: from Continuous to Discrete...

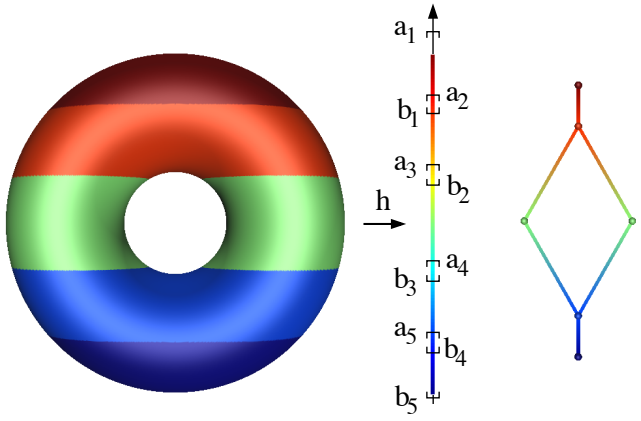


Figure: An illustration of Mapper.

Note:

- degree-one nodes contain local minima/maxima;
- degree-three nodes contain saddle points (critical points);
- degree-two nodes consist of regular points

Mapper algorithm

[*Singh-Memoli-Carlsson. Eurograph-PBG, 2007*] Given a data set $\mathcal{X}$,

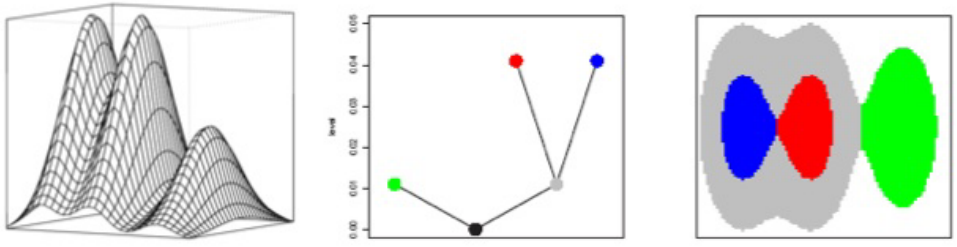
- choose a filter map $h : \mathcal{X} \to \mathcal{Z}$, where $\mathcal{Z}$ is a topological space such as $\mathbb{R}$, S^1, $\mathbb{R}^d$, etc.
- choose a cover $\mathcal{Z} \subseteq \cup_\alpha U_\alpha$
- cluster/partite level sets $h^{-1}(U_\alpha)$ into $V_{\alpha,\beta}$
- graph representation: a node for each $V_{\alpha,\beta}$, an edge between $(V_{\alpha_1,\beta_1}, V_{\alpha_2,\beta_2})$ iff $U_{\alpha_1} \cap U_{\alpha_2} \neq \emptyset$ and $V_{\alpha_1,\beta_1} \cap V_{\alpha_2,\beta_2} \neq \emptyset$.
- extendable to simplicial complex representation.

Note: it extends Reeb Graph from $\mathbb{R}$ to general topological space $\mathcal{Z}$; may lead to a particular implementation of Nerve theorem through filter map h.

In applications.

Reeb graph has found various applications in computational geometry, statistics under different names.

- computer science: contour trees, Reeb graphs
- statistics: density cluster trees (Hartigan)

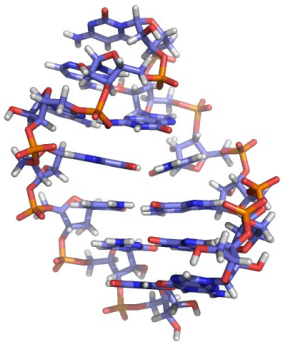


Reference Mapping

Typical one dimensional filters/mappings:

- Density estimators
- Measures of data (ec-)centrality: e.g. $\sum_{x' \in \mathcal{X}} d(x, x')^p$
- Geometric embeddings: PCA/MDS, Manifold learning, Diffusion Maps etc.
- Response variable in statistics: progression stage of disease etc.

Example: RNA Tetraloop



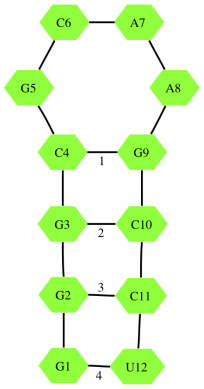


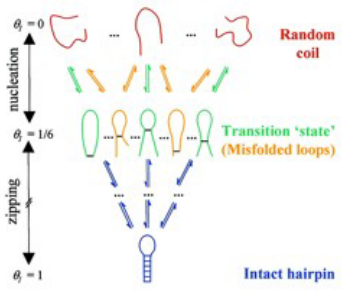
Figure: RNA GCAA-Tetraloop

Biological relevance:

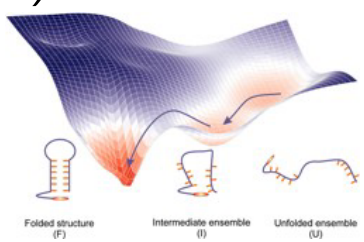
- serve as nucleation site for RNA folding
- form sequence specific tertiary interactions
- protein recognition sites
- certain Tetraloops can pause RNA transcription

Note: simple, but, biological debates over intermediate states on folding pathways

Debates: Two-state vs. Multi-state Models



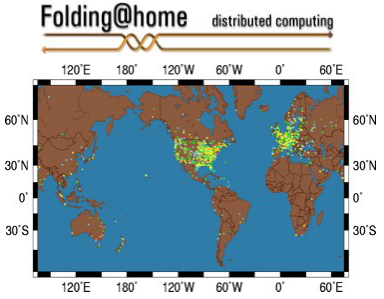
(a) 2-state model



(b) multi-state model

- 2-state: transition state with any one stem base pair, from thermodynamic experiments [*Ansari A, et al. PNAS, 2001, 98: 7771-7776*]
- multi-state: there is a stable intermediate state, which contains collapsed structures, from kinetic measurements [*Ma H, et al. PNAS, 2007, 104:712-6*]
- experiments: no structural information
- computer simulations at full-atom resolution:
  - exisitence of intermediate states
  - if yes, what's the structure?

MD Simulation by Folding@Home



Simulation Box.

[*Bowman, Huang, Y., Sun, ... Vijay. JACS, 2008*]

- 2800 SREMD (Serial Replica Exchange Molecular Dynamics) simulations with RNA hairpin (5'-GGGCGCAAGCCU-3')
- 389 RNA atoms, ∼4000 water and 11 Na^+
- SREMD random walks in temperature space (56 ladders from 285K to 646K) with molecular dynamic trajectories
- 210,000 ns simulations with ∼105,000,000 configurations
- Unfortunately, sampling still not converged!

Dimensionality Reduction using Contact Map

- Massive volume and high dimensionality: $100M$ samples in $12K$ Cartesian coordinates $\Rightarrow$ contact maps as 55-bit string
- Samples are not in equilibrium distribution
- Looking for a needle in a haystack:
  - intermediates/transition states of interests are of low-density
  - folded/unfolded states are dominant

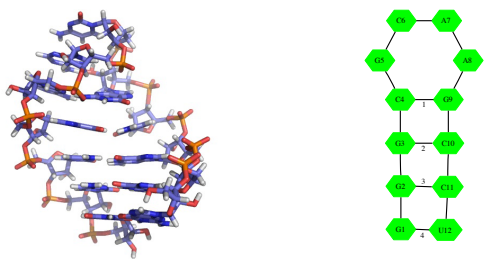


Figure: Left: NMR structure of the GCAA tetraloop. Right: Contact map.

Mapper with density filters in biomolecular folding

Reference: Bowman-Huang-Yao et al. J. Am. Chem. Soc. 2008; Yao, Sun, Huang, et al. J. Chem. Phys. 2009.

- densest regions (energy basins) may correspond to metastates (e.g. folded, extended)
- intermediate/transition states on pathways connecting them are relatively sparse

Therefore with Mapper

- clustering on density level sets helps separate and identify metastates and intermediate/transition states
- graph representation reflects kinetic connectivity between states

A vanilla version

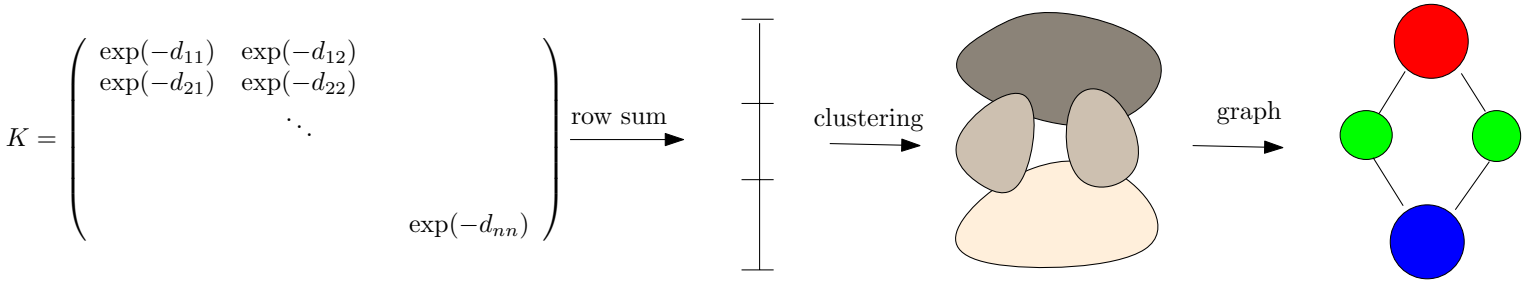


Figure: Mapper Flow Chart

1. Kernel density estimation $h(x) = \sum_i K(x, x_i)$ with Hamming distance for contact maps
2. Rank the data by h and divide the data into n overlapped sets
3. Single-linkage clustering on each level sets
4. Graphical representation

Mapper output for Unfolding Pathways

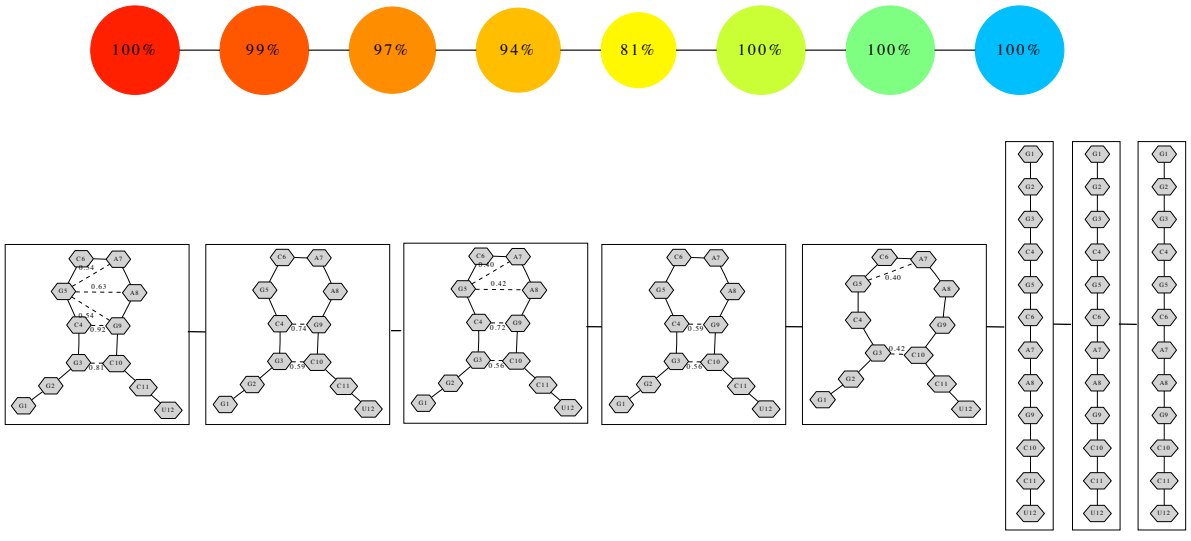


Figure: Unfolding pathway

Mapper output for Refolding Pathways

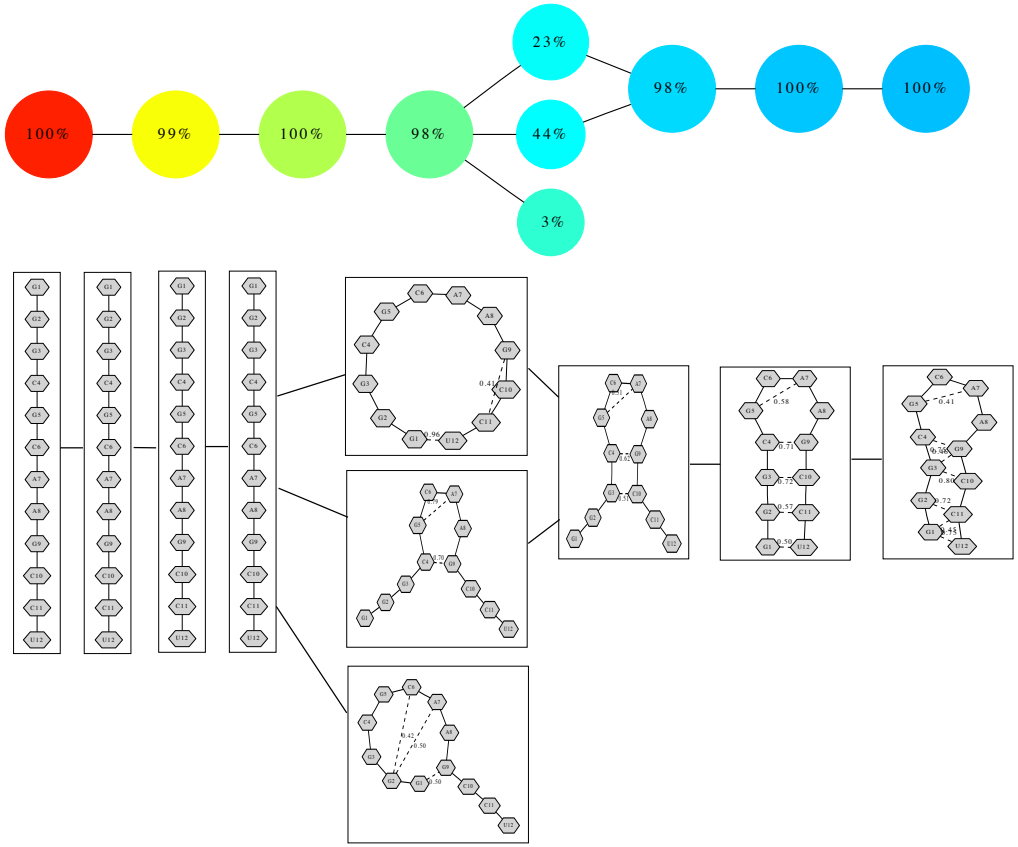


Figure: Refolding pathway

Example: Progression of Breast Cancer

- We study samples of expression data in $\mathbb{R}^n$ $(n = 262)$ from 295 breast cancers as well as additional samples from normal breast tissue.
  - The distance metric was given by the correlation between (projected) expression vectors.
  - The filter function used was a measure taking values in $\mathbb{R}$ of the deviation of the expression of the tumor samples relative to normal controls (l_2-eccentrality).
  - The cover was overlapping intervals in $\mathbb{R}$.
- Two branches of breast cancer progression are discovered.

Progression of Breast Cancer: l_2-eccentrality

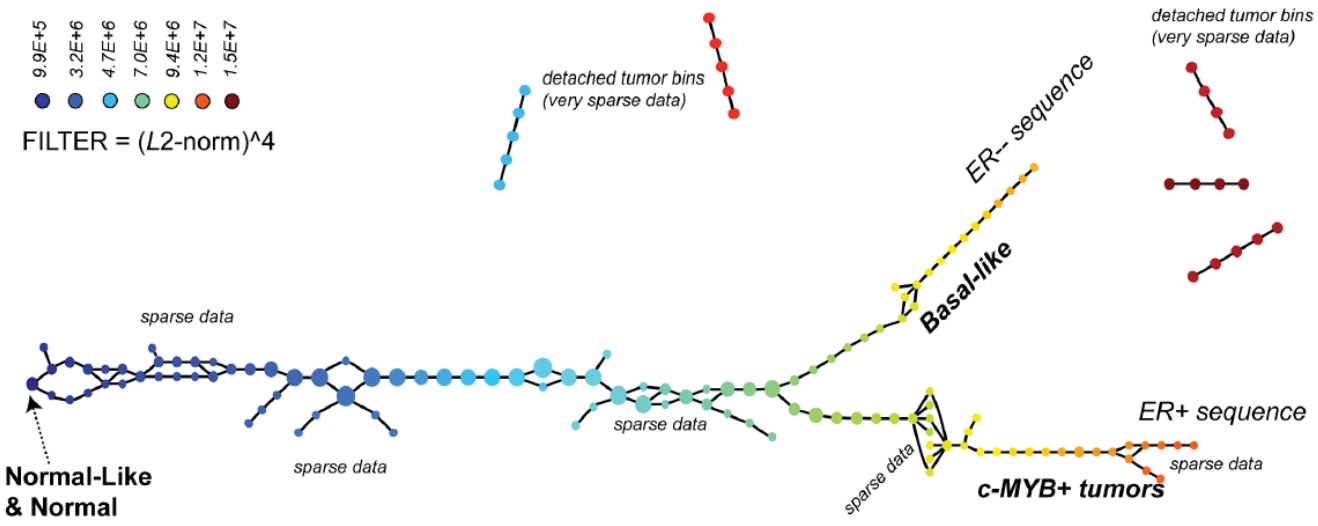


Figure: Monica Nicolau, A. Levine, and Gunnar Carlsson, *PNAS'10*

Note: Progression of Breast Cancer

- The lower right branch itself has a subbranch (referred to as c-MYB+ tumors), which are some of the most distinct from normal and are characterized by high expression of genes including c-MYB, ER, DNALI1 and C9ORF116. Interestingly, all patients with c-MYB+ tumors had very good survival and no metastasis.
- These tumors do not correspond to any previously known breast cancer subtype; the grouping seems to be invisible to classical hierarchical clustering methods.

Example: differentiation process using single cell data

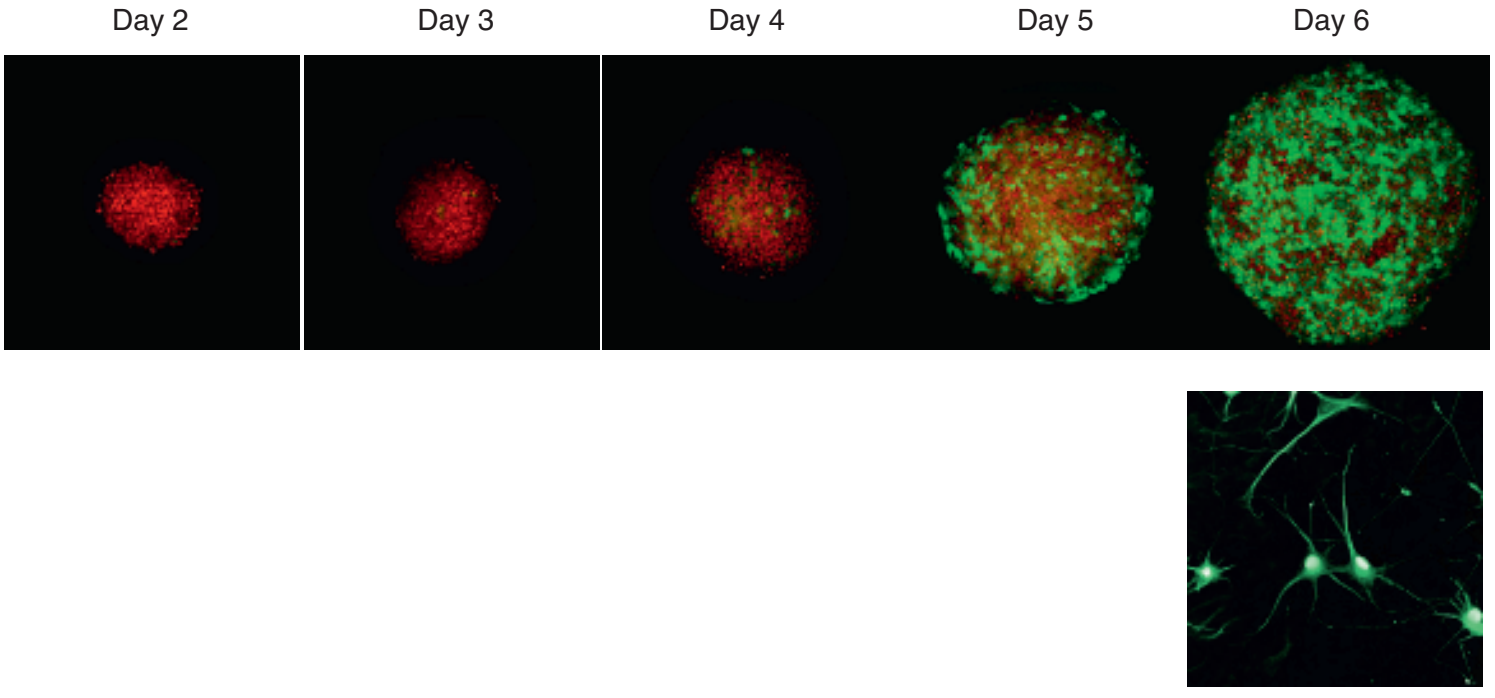


Figure 2.31 Over time, embryonic stem cells differentiate into distinct cell types. These pictures capture the in vitro differentiation of mouse embryonic stem cells into motor neurons over the course of a week. Embryonic stem cells are marked in red, and fully differentiated neurons in green. Figure from experiment performed by Elena Kandror, Abbas Rizvi and Tom Maniatis at Columbia University.

Differentiation process visualization by Mapper

- Over time, undifferentiated embryonic cells become differentiated motor neurons when retinoic acid and sonic hedgehog (a differentiation-promoting protein) are applied.
- Mapper graph of differentiation process from murine embryonic stem cells to motor neurons:
  - The data generated corresponds to RNA expression profiles from roughly 2000 single cells.
  - The distance metric was provided by correlation between expression vectors.
  - The filter function used was multidimensional scaling (MDS) projection into $\mathbb{R}^2$.
  - The cover was overlapping rectangles in $\mathbb{R}^2$.

Mapper Graph of Differentiation Process

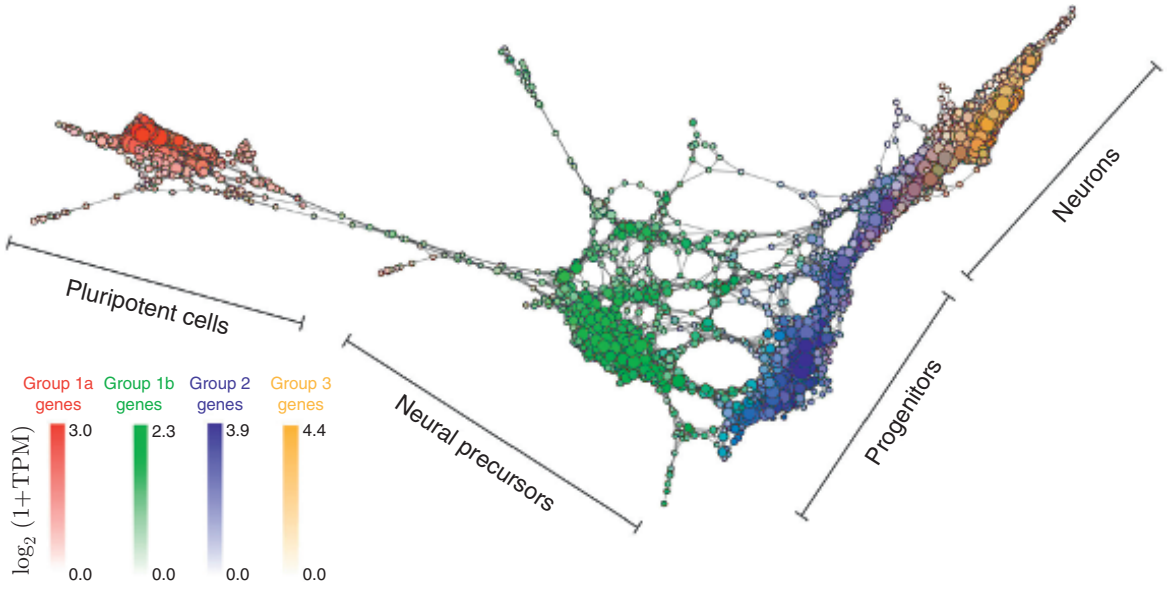


Figure: The different regions in the Mapper graph nicely line up with different points along the differentiation timeline. Rizvi et al. *Nature Biotechnol.* 35.6 (2017), 551-560.

Example: Brain Tumor

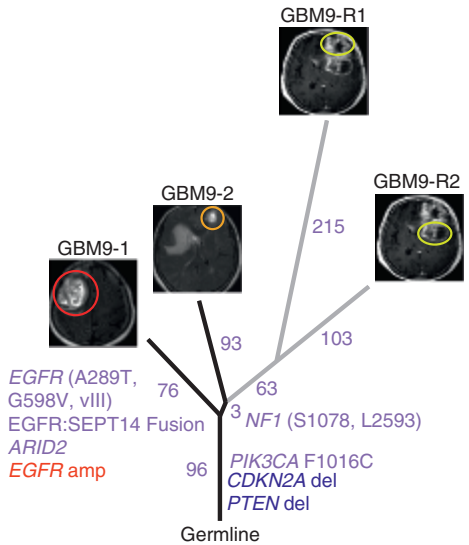
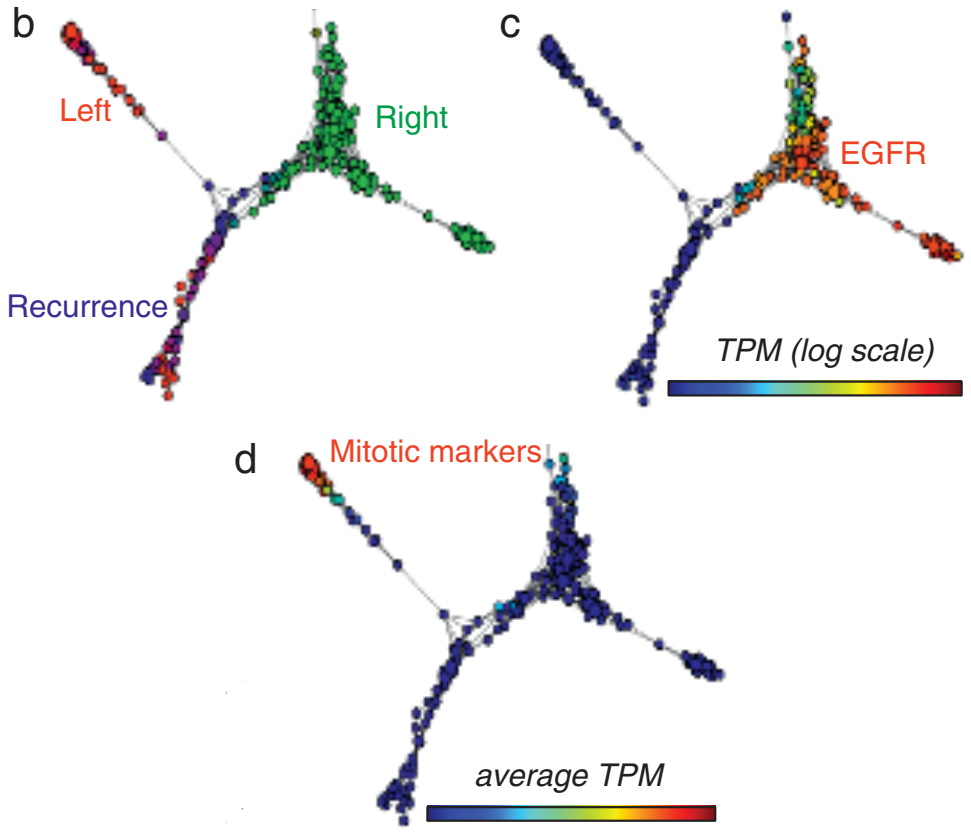


Figure: A patient with two focal glioblastomas, on the left and right hemispheres. After surgery and standard treatment, the tumor reappeared on the left side. Genomic analysis shows that the initial tumors were seeded by two independent, but related clones. The recurrent tumor was genetically similar to the left one. Jin-Ku Lee et al. *Nature Genetics* 49.4 (2017): 594-599.

Mapper Graph of Single Cell Seq.



Note: Mapper Graph

- Using Mapper, one can appreciate a more continuous structure that recapitulates the clonal and genetic history.
  - The tumor on the right appears to be transcriptionally distinct from the left tumor and the recurrence tumor.
  - Expression profiles from cells in the recurrence tumor resembled the originating initial tumor.
  - This is an important finding, as it shows a continued progression at the expression level, with a few cells at diagnosis having a similar pattern as cells at relapse.
  - It also shows that EGFR mutation is a subclonal event, occurring only in the tumor at diagnosis that is not responsible for the relapse. So tumors with heterogeneous populations of cells are less sensitive specific therapies which target a subpopulation..

Čech complex

Definition (Čech Complex C_ϵ)

In a metric space (X, d), define a cover of X, $X = \cup_\alpha U_\alpha$ where $U_\alpha = B_\epsilon(t_\alpha) := \{x \in X : d(x - t_\alpha) \leq \epsilon\}$. $V = \{U_\alpha\}$ and define $\Sigma = \{U_I : \cap_{\alpha \in I} U_I \neq \emptyset\}$.

- Closedness under deletion
- Can be applied to any metric space X
- Nerve Theorem: if every U_I is contractible, then X has the same homotopy type as Σ.

Example: Čech Complex

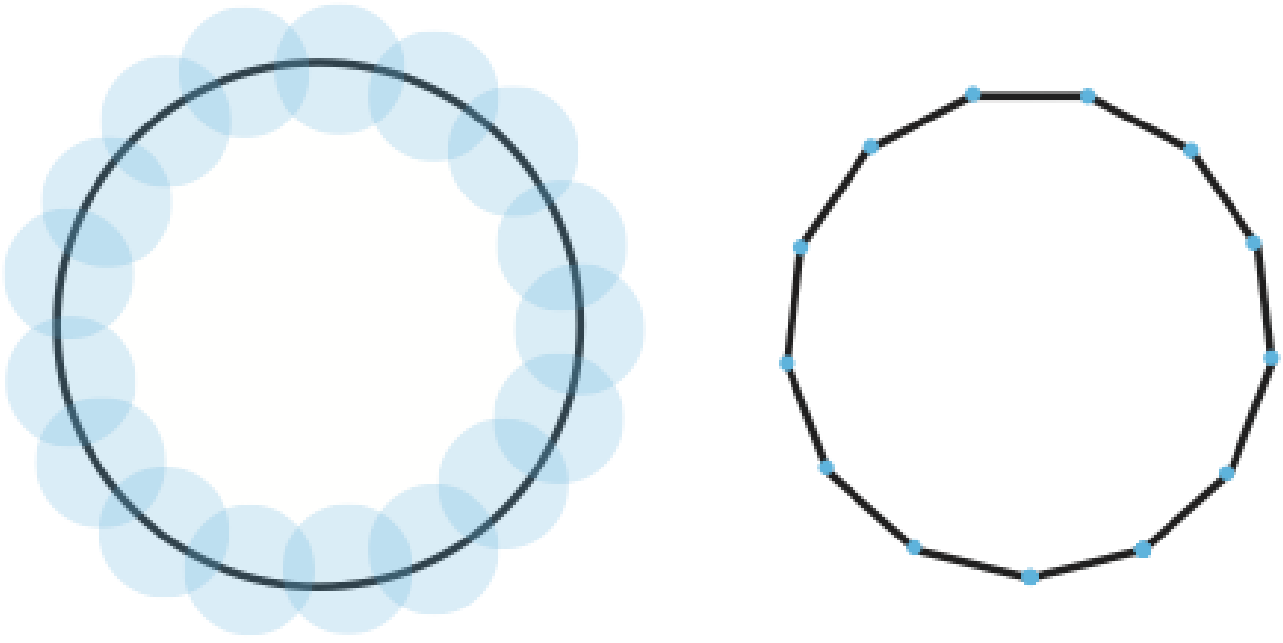


Figure: Čech complex of a circle, C_ϵ, covered by a set of balls.

Vietoris-Rips complex

- Čech complex is hard to compute, even in Euclidean space
- One can easily compute an upper bound for Čech complex
  - Construct a Čech subcomplex of 1-dimension, i.e. a graph with edges connecting point pairs whose distance is no more than ϵ.
  - Find the clique complex, i.e. maximal complex whose 1-skeleton is the graph above, where every k-clique is regarded as a $k-1$ simplex

Definition (Vietoris-Rips Complex)

Let $V = \{x_\alpha \in X\}$. Define $VR_\epsilon = \{U_I \subseteq V : d(x_\alpha, x_\beta) \leq \epsilon, \alpha, \beta \in I\}$.

Example: Rips Complex

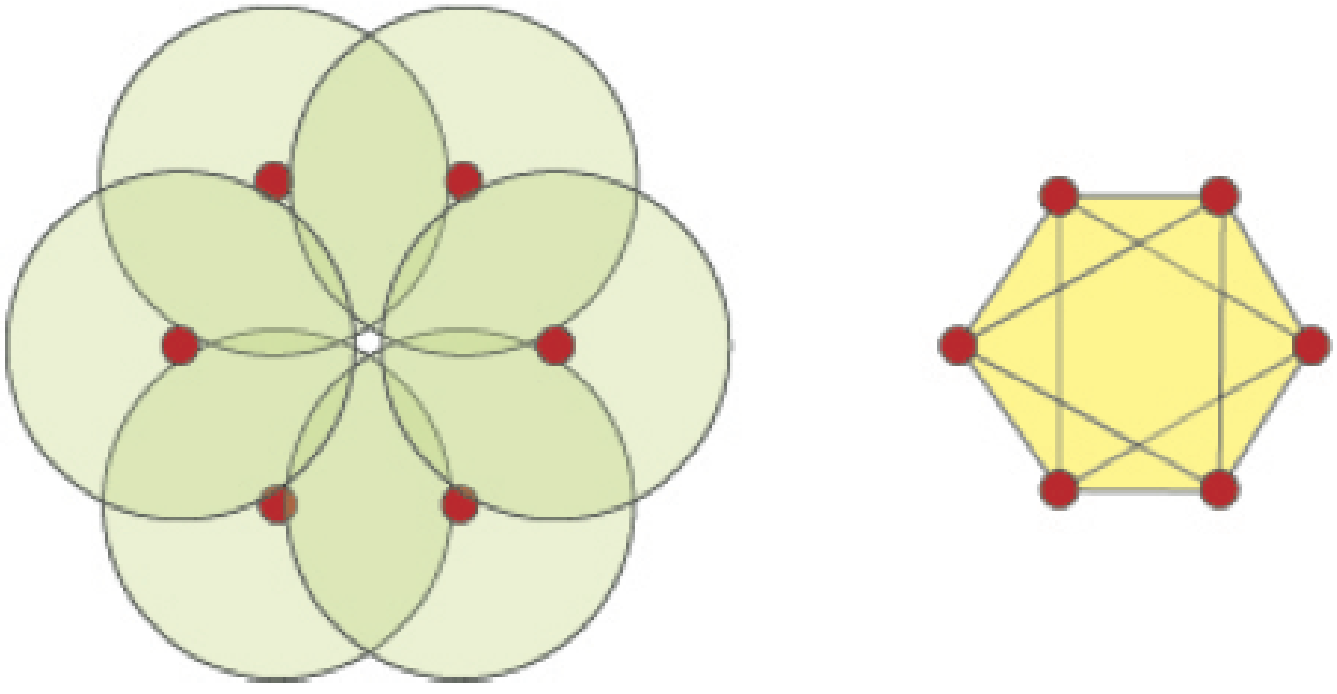


Figure: Left: Čech complex gives a circle; Right: Rips complex gives a sphere S^2.

Generalized Vietoris-Rips for Symmetric Relations

Definition (Symmetric Relation Complex)

Let V be a set and a symmetric relation $R = \{(u, v)\} \subseteq V^2$ such that $(u, v) \in R \Rightarrow (v, u) \in R$. Σ collects subsets of V which are in pairwise relations.

- Closedness under deletion: if $\sigma \in \Sigma$ is a set of related items, then any subset $\tau \subseteq \sigma$ is a set of related items
- Generalized Vietoris-Rips complex beyond metric spaces
- E.g. Zeeman's tolerance space
- C.H. Dowker defines simplicial complex for unsymmetric relations

Sandwich Theorems

- Rips is easier to compute than Cech
  - even so, Rips is exponential to dimension generally
- However Vietoris-Rips CAN NOT preserve the homotopy type as Cech
- But there is still a hope to find a lower bound on homology –

Theorem ("Sandwich")

$$VR_\epsilon \subseteq C_\epsilon \subseteq VR_{2\epsilon}$$

- If a homology group "persists" through $R_\epsilon \to R_{2\epsilon}$, then it must exists in C_ϵ; but not the vice versa.

A further simplification: Witness complex

Definition (Strong Witness Complex)

Let $V = \{t_\alpha \in X\}$. Define
$W_\epsilon^s = \{U_I \subseteq V : \exists x \in X, \forall \alpha \in I, d(x, t_\alpha) \leq d(x, V) + \epsilon\}$.

Definition (Week Witness Complex)

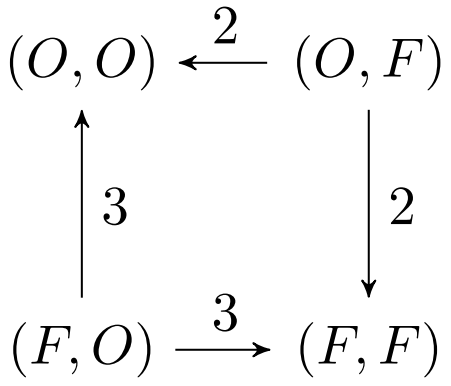
Let $V = \{t_\alpha \in X\}$. Define
$W_\epsilon^w = \{U_I \subseteq V : \exists x \in X, \forall \alpha \in I, d(x, t_\alpha) \leq d(x, V_{-I}) + \epsilon\}$.

- V can be a set of landmarks, much smaller than X
- Monotonicity: $W_\epsilon^* \subseteq W_{\epsilon'}^*$ if $\epsilon \leq \epsilon'$
- But not easy to control homotopy types between W^* and X

Strategic Simplicial Complex for Flow Games

	O	F
O	3, 2	0, 0
F	0, 0	2, 3

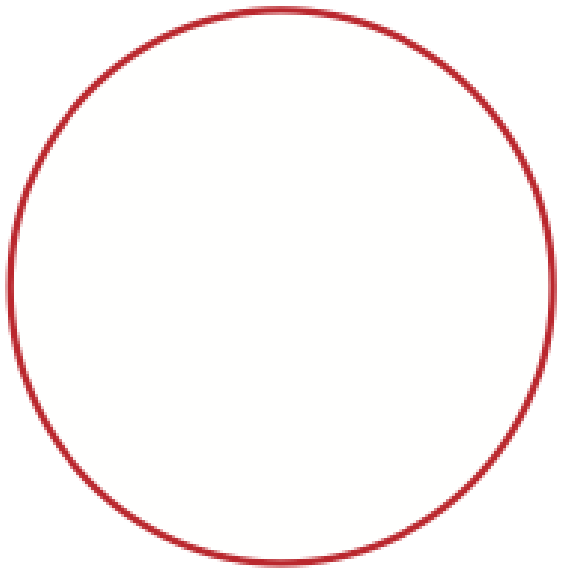
(a) Battle of the sexes



- Strategic simplicial complex is the clique complex of pairwise comparison graph above, inspired by ranking
- Every game can be decomposed as the direct sum of potential games and zero-sum games (harmonic games) (Candogan, Menache, Ozdaglar and Parrilo 2010)

Outline

Betti Numbers: the number of i-dim holes



$\beta_0 = 1$, $\beta_1 = 1$, and $\beta_i = 0$ for $i \geq 2$

Betti Numbers: the number of i-dim holes

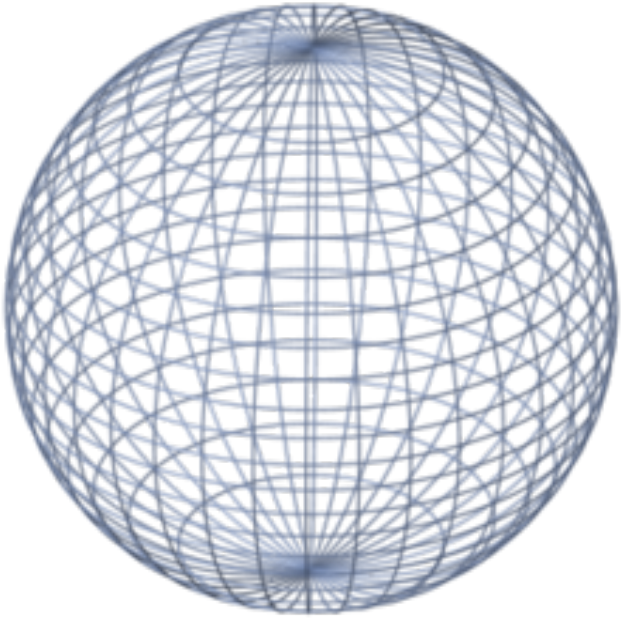
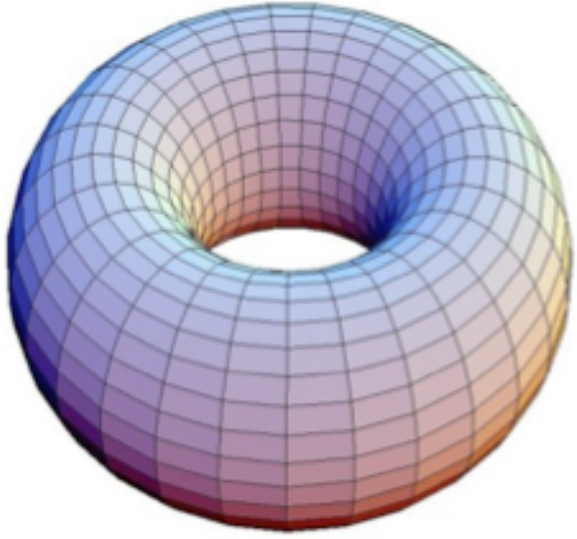


Figure: Sphere: $\beta_0 = 1$, $\beta_1 = 0$, $\beta_2 = 1$, and $\beta_k = 0$ for $k \geq 3$

Betti Numbers: the number of i-dim holes

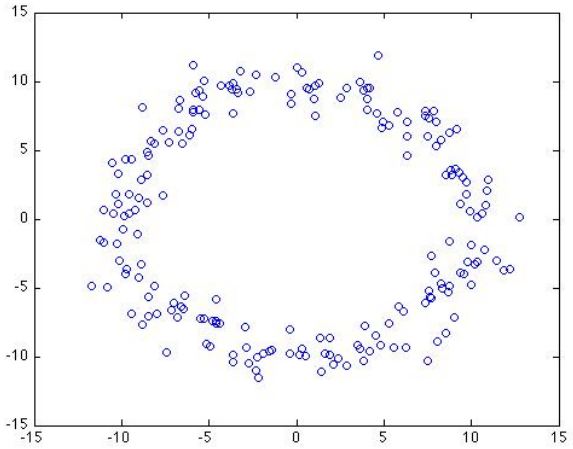


$\beta_0 = 1$, $\beta_1 = 2$, $\beta_2 = 1$, and $\beta_k = 0$ for $k \geq 3$

Betti Numbers and Homology Groups

- Betti numbers are computed as dimensions of Boolean vector spaces (E. Noether, $\mathbb{Z}_2$-homology group)
- $\beta_i(X) = \dim H_i(X, \mathbb{Z}_2)$, $\mathbb{Z}_2$-homology or more general Homology group associated with any fields or integral domain (e.g. $\mathbb{Z}$, $\mathbb{Q}$, and $\mathbb{R}$)
- $H_i(X)$ is *functorial*, i.e. continuous mapping $f : X \to Y$ induces linear transformation $H_i(f) : H_i(X) \to H_i(Y)$, structure preserving
- computation is simple linear algebra over fields or integers
- data representation by *simplicial complexes*

Topology at Different Scales



- Is it a circle, dots, or circle of circles?
- How to find robust topology at different scales?

Example I: Persistent Homology of Čech Complexes

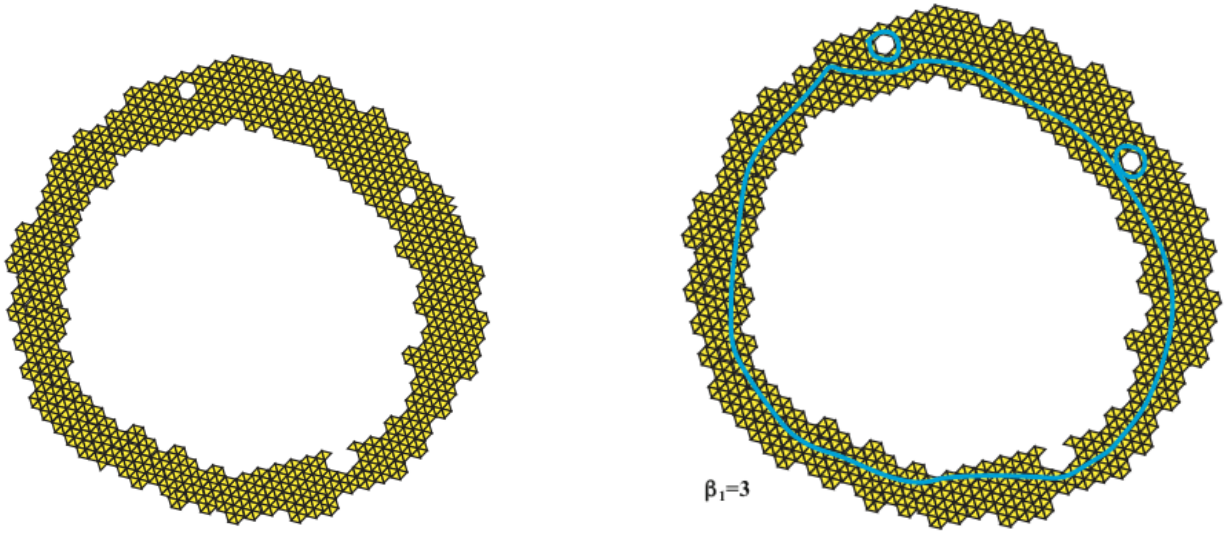


Figure: Scale ϵ_1: $\beta_0 = 1$, $\beta_1 = 3$

Example I: Persistent Homology of Čech Complexes

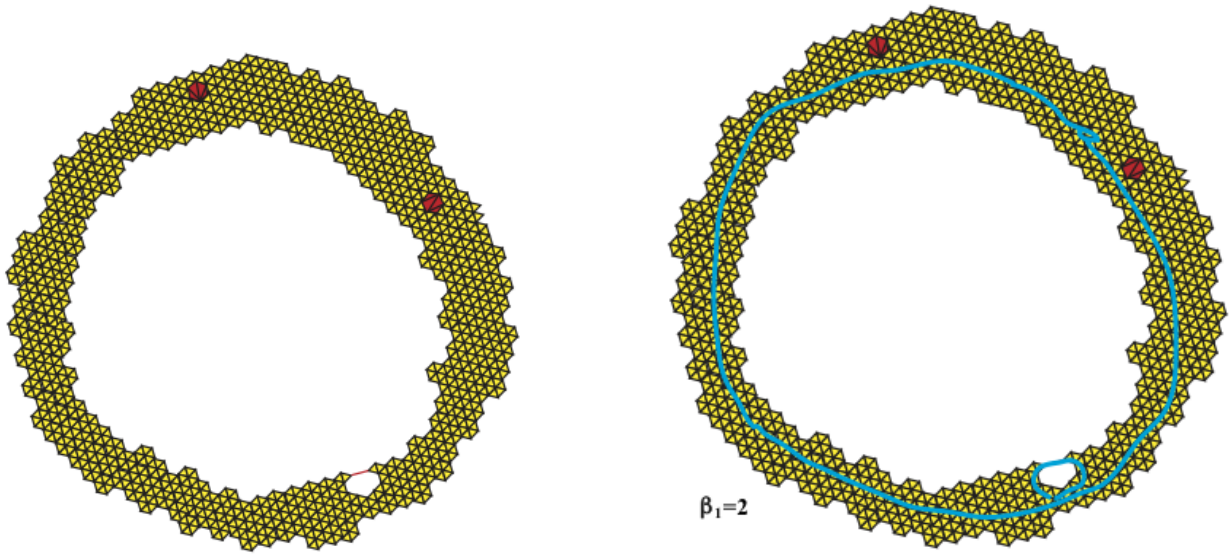


Figure: Scale $\epsilon_2 > \epsilon_1$: $\beta_0 = 1$, $\beta_1 = 2$. Persistent $\beta_0 = 1$ and $\beta_1 = 1$ from ϵ_1 to ϵ_2 suggest that a connected component and a loop are stable topological features here.

Example II: Persistence 0-Homology induced by Height Function

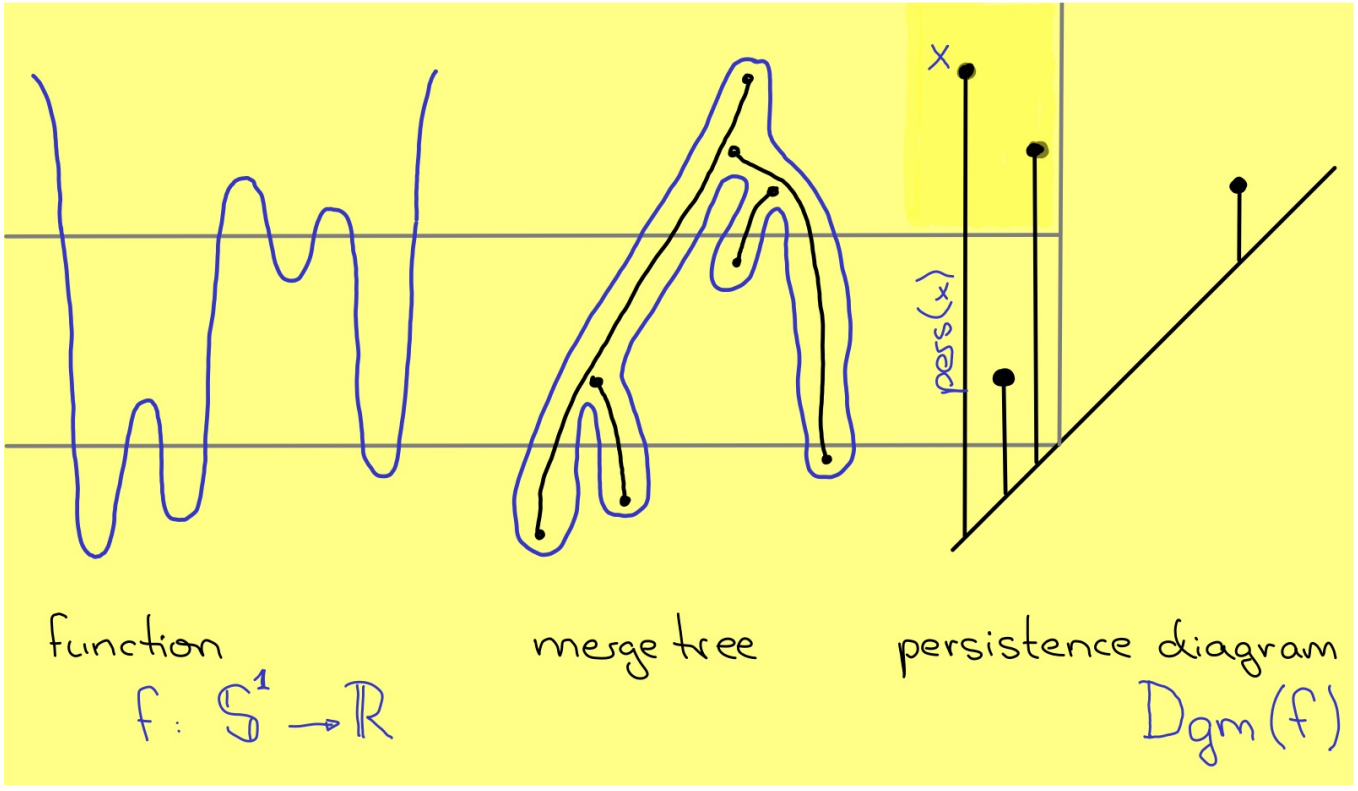


Figure: The birth and death of connected components.

Example III: Persistent Homology as Online Algorithm to Track Topology Changements

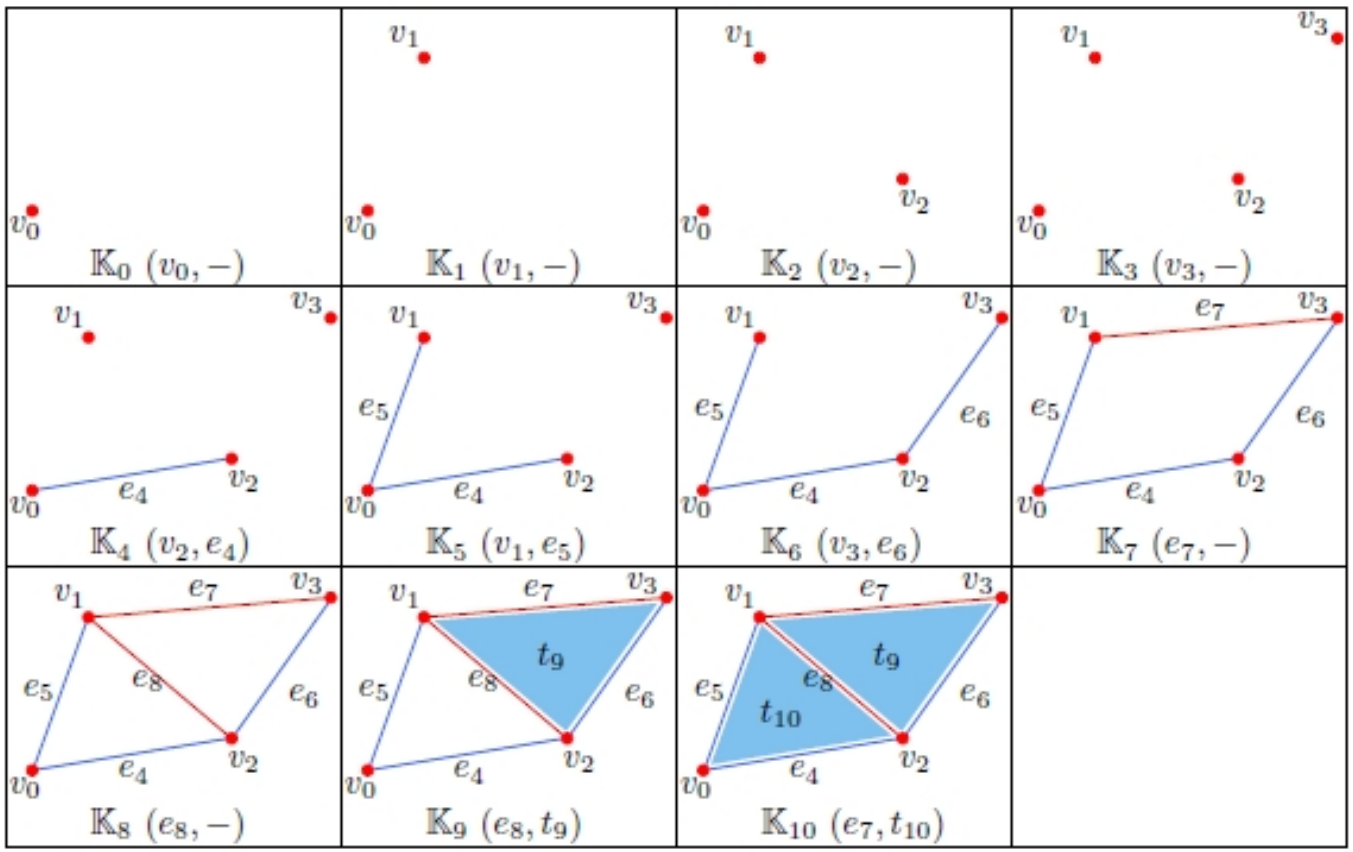
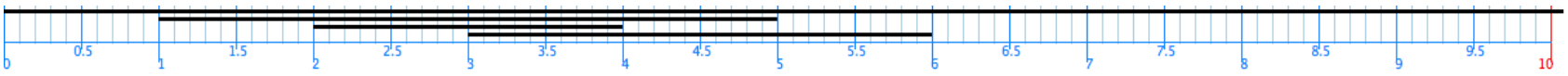
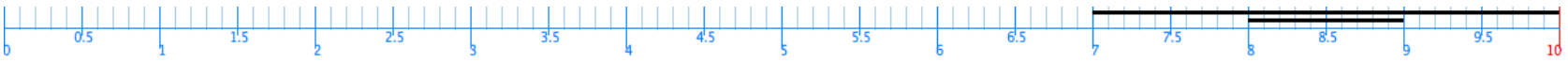


Figure: The birth and death of simplices.

Persistent Betti Numbers: Barcodes



Barcodes: Dimension 0



Barcodes: Dimension 1

- Toolbox: JavaPlex (`https://github.com/appliedtopology/javaplex/wiki/Tutorial`)
  - Java version of Plex, work with matlab
  - Rips, Witness complex, Persistence Homology
- Other Choices: Plex 2.5 for Matlab (not maintained any more), Dionysus (Dimitry Morozov)

Persistent Homology: Algebraic Characterization

- All above gives rise to a filtration of simplicial complex

$$\emptyset = \Sigma_0 \subseteq \Sigma_1 \subseteq \Sigma_2 \subseteq \dots$$

- Functoriality of inclusion: there are homomorphisms between homology groups

$$0 \to H_1 \to H_2 \to \dots$$

- A persistent homology is the image of H_i in H_j with $j > i$.

Persistent 0-Homology of Rips Complex

- Equivalent to single-linkage clustering or minimal spanning tree
- Barcode is the single linkage dendrogram (tree) without labels
- Kleinberg's Impossibility Theorem for clustering: no clustering algorithm satisfies scale invariance, richness, and consistency
- Memoli & Carlsson 2009: single-linkage is the unique persistent clustering (functorial) with scale invariance
- Open Question: but, is persistence the necessity for clustering?

- Notes: try matlab command `linkage` or R `hclust` for single-linkage clustering.

Application: Evolutionary Trees

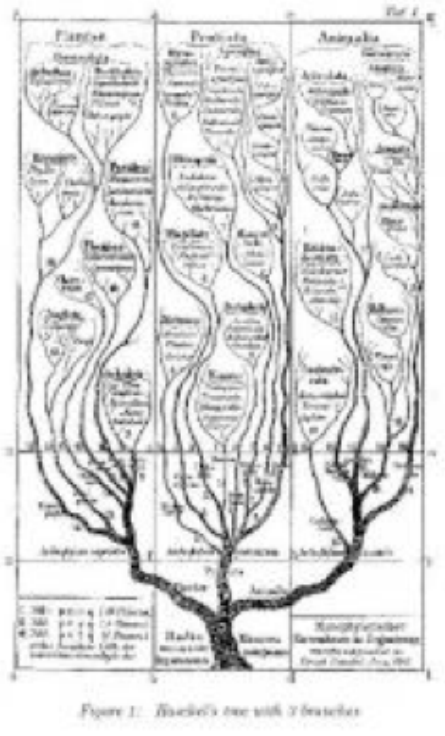


Figure: Are phylogenetic trees good representations for evolution?

Virus gene reassortment may introduce loops

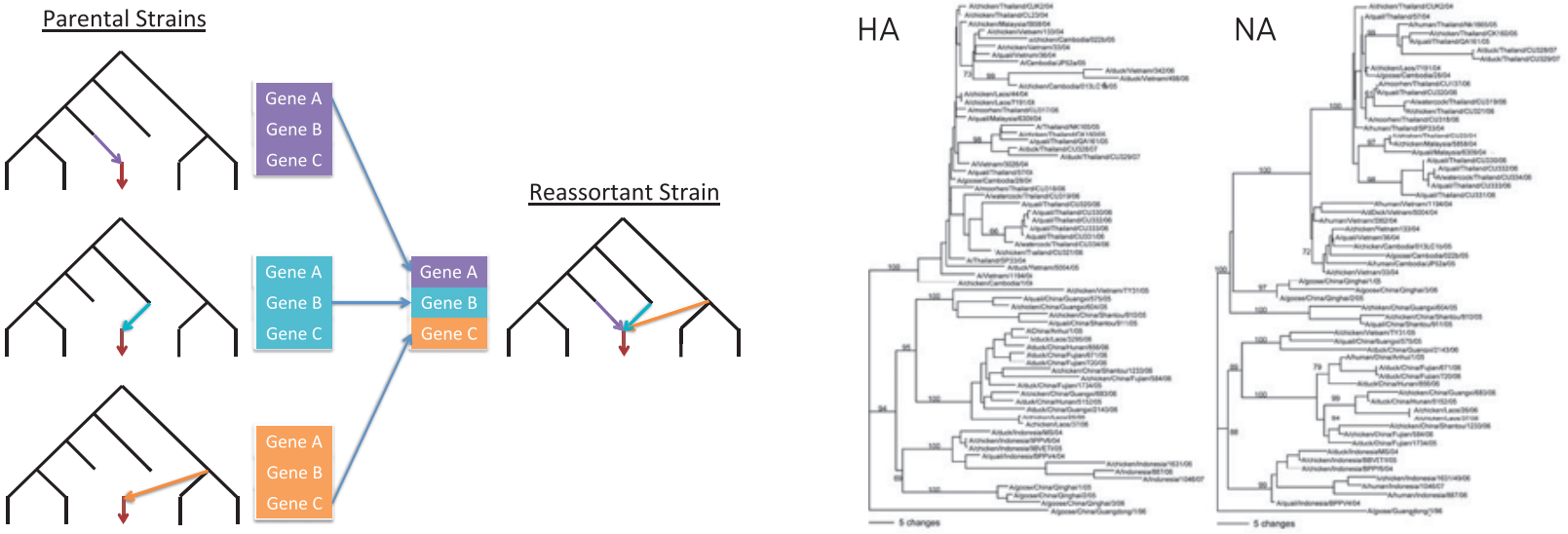


Figure 5.16 Left: Reassortments in viruses lead to incompatibility between trees. Reticulate network representing the reassortment of three parental strains. The reticulate network results from merging the three parental phylogenetic trees. Source: [100]. Right: Indeed, incompatibility between tree topologies inferred from different genes is a criterion used for the identification of events of genomic material exchange. Here we represent two genes of influenza A virus with different topologies using phylogenetic networks. From Joseph Minhow Chan, Gunnar Carlsson, and Raúl Rabadán, 'Topology of viral evolution', *Proceedings of the National Academy of Sciences* 110.46 (2013): 18566–18571. Reprinted with Permission from Proceedings of the National Academy of Sciences.

Influenza

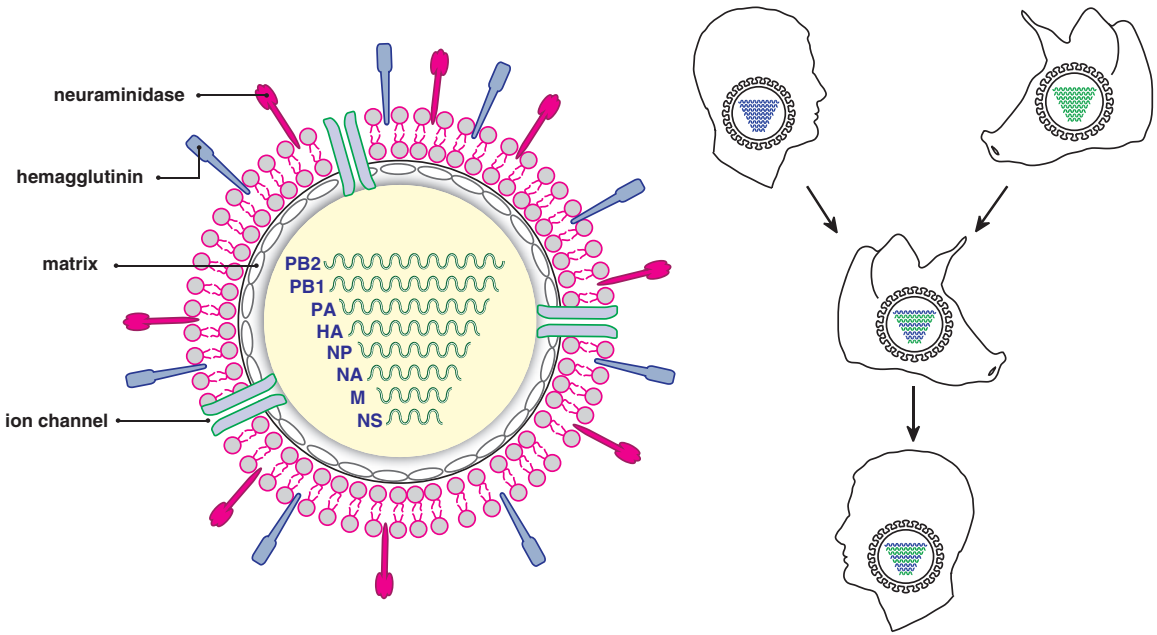


Figure 5.14 Influenza A is an antisense single-stranded RNA virus whose genome is composed of eight different segments containing one or two genes per segment. This virus contains an envelope borrowed from the infected cell that expressed two viral proteins, hemagglutinin and neuraminidase. When circulating viruses co-infect the same cell, new viruses can be created that contain segments from both parents. This phenomenon, called reassortment, can lead to dramatic adaptations to novel environments, and it is thought to be one of the contributing factors to human influenza pandemics.

Origins of H1N1-2009

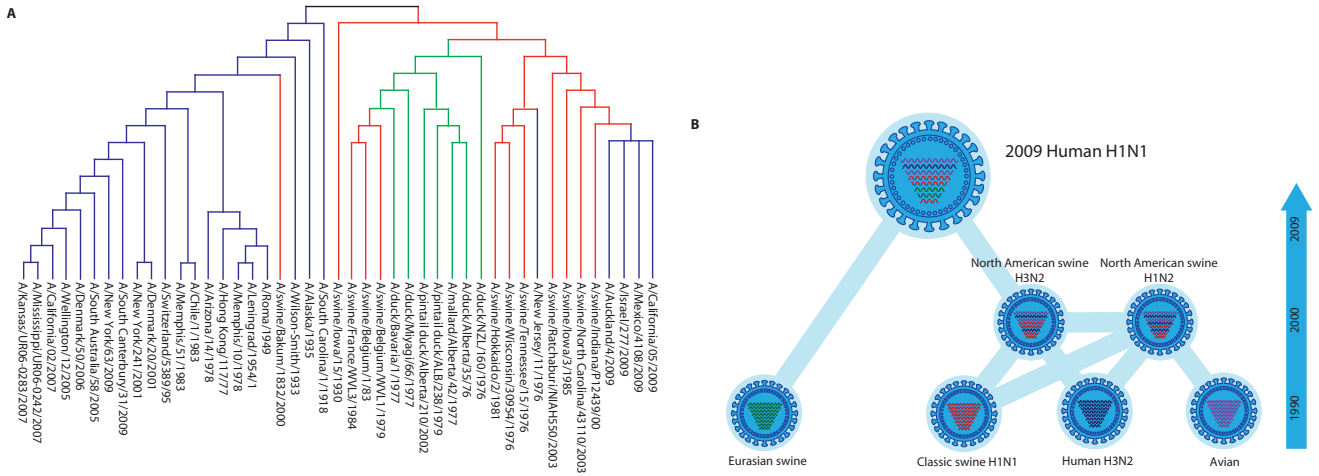


Figure: Origins of H1N1 2009 pandemic virus. Using phylogenetic trees, the history of the HA gene of the 2009 H1N1 pandemic virus was reconstructed. It was related to viruses that circulated in pigs potentially since the 1918 H1N1 pandemic. These viruses had diverged since that date into various independent strains, infecting humans and swine. Major reassortments between strains led to new sets of segments from different sources. In 1998, triple reassortant viruses were found infecting pigs in North America. These triple reassortant viruses contained segments that were circulating in swine, humans and birds. Further reassortment of these viruses with other swine viruses created the ancestors of this pandemic. Until this day, it is unclear how, where or when these reassortments happened. Source: [506]. From New England Journal of Medicine, Vladimir Trifonov, Hossein Khiabanian, and Raúl Rabadán, Geographic dependence, surveillance, and origins of the 2009 influenza A (H1N1) virus, 361.2, 115–119.

When Persistent Betti-0 meets Pylogenetic Trees

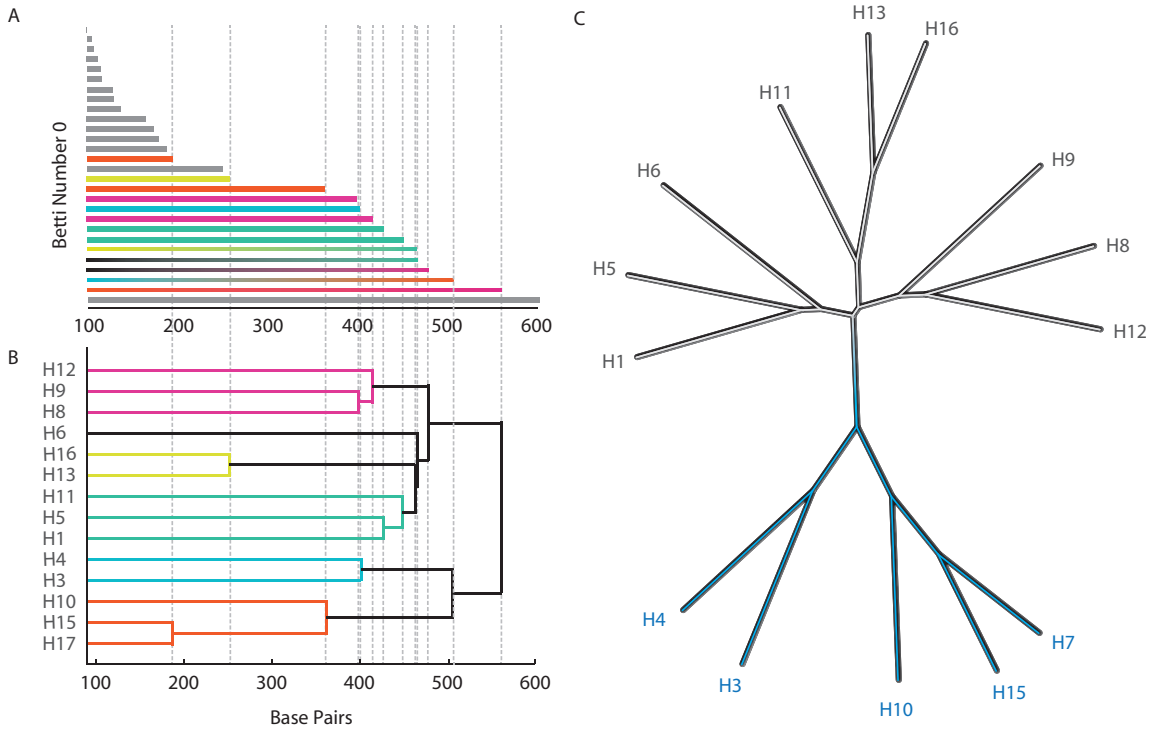


Figure: In case of vanishing higher dimensional homology, zero dimen- sional homology generates trees. When applied to only one gene of influenza A, in this case hemagglutinin, the only significant homology occurs in dimen- sion zero (panel A). The barcode represents a summary of a clustering procedure (panel B), that recapitulates the known phylogenetic relation between different hemagglutinin types (panel C). Source: [100]. From Joseph Minhow Chan, Gunnar Carlsson, and Raúl Rabadán, 'Topology of viral evolution', Proceedings of the National Academy of Sciences 110.46 (2013): 18566–18571.

Whole Genomic Persistent Betti Numbers

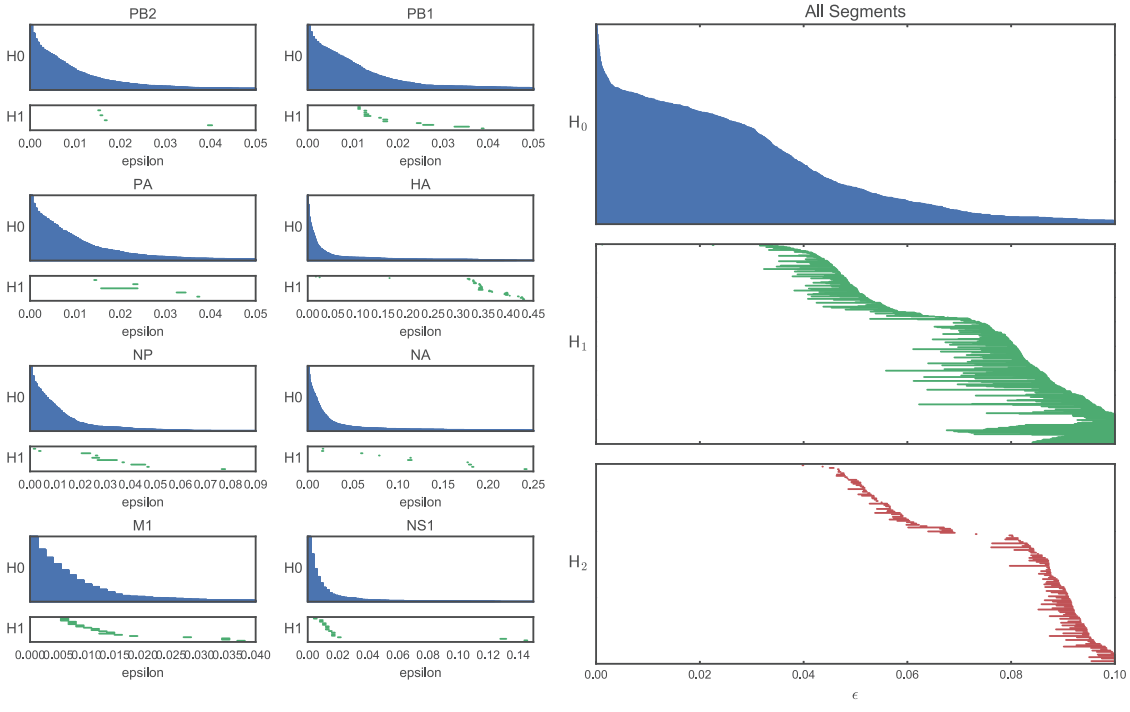


Figure 5.18 Influenza evolves through mutations and reassortment. When the persistent homology approach is applied to finite metric spaces derived from only one segment, up to small noise, the homology is zero dimensional suggesting a tree-like process (left). However, when different segments are put together, the structure is more complex revealing non-trivial homology at different dimensions (right). 3105 influenza whole genomes were analyzed. Data from isolates collected between 1956 to 2012; all influenza A subtypes.

Two modes in persistent β_1 distributions suggest intra- and inter-subtypes

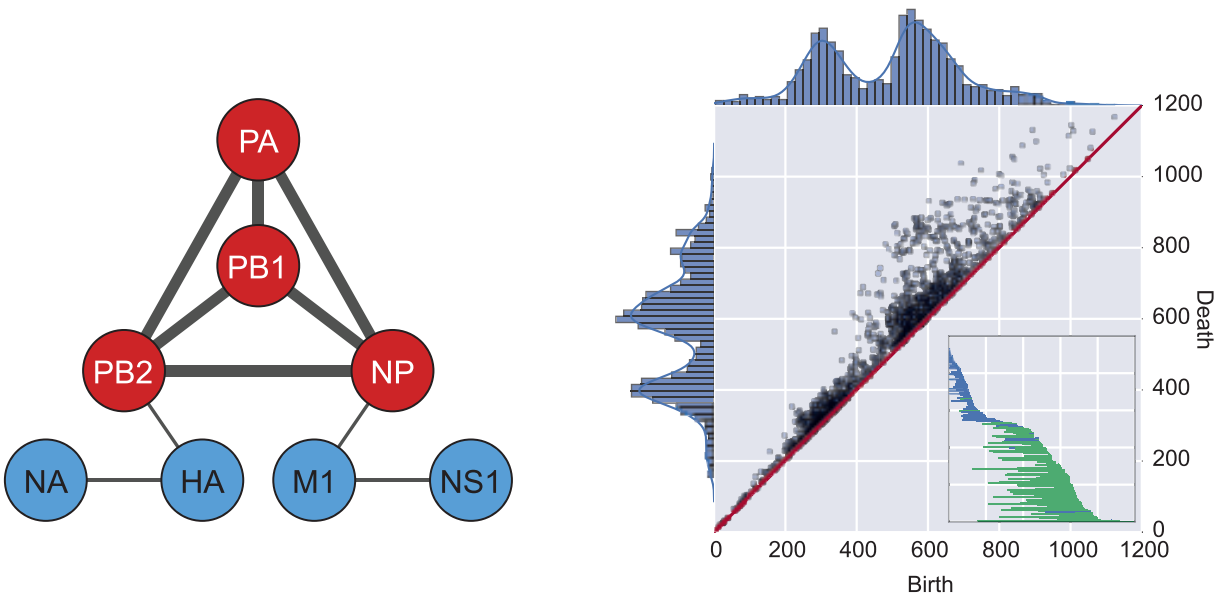


Figure: Co-reassortment of viral segments as structure in persistent homol- ogy diagrams. Left: The non-random cosegregation of influenza segments was measured by testing a null model of equal reassortment. Significant cosegregation was identified within PA, PB1, PB2, NP, consistent with the cooperative func- tion of the polymerase complex. Source: [100]. Right: The persistence diagram for whole-genome avian flu sequences revealed bimodal topological structure. Annotating each interval as intra- or inter-subtype clarified a genetic barrier to reassortment at intermediate scales. From Joseph Minhow Chan, Gunnar Carlsson, and Raúl Rabadán, 'Topology of viral evolution', Proceedings of the National Academy of Sciences 110.46 (2013): 18566–18571.

Application: Sensor Network Coverage by Persistent Homology

- V. de Silva and R. Ghrist (2005) Coverage in sensor networks via persistent homology.
- Ideally sensor communication can be modeled by Rips complex
  - two sensors has distance within a short range, then two sensors receive strong signals;
  - two sensors has distance within a middle range, then two sensors receive weak signals;
  - otherwise no signals

Sandwich Theorem

Theorem (de Silva-Ghrist 2005)

Let X be a set of points in R^d and $C_\epsilon(X)$ the Čech complex of the cover of X by balls of radius $\epsilon/2$. Then there is chain of inclusions

$$R_{\epsilon'}(X) \subset C_\epsilon(X) \subset R_\epsilon(X) \quad \textit{whenever} \quad \frac{\epsilon}{\epsilon'} \geq \sqrt{\frac{2d}{d+1}}.$$

Moreover, this ratio is the smallest for which the inclusions hold in general.

Note: this gives a sufficient condition to detect holes in sensor network coverage

- Čech complex is hard to compute while Rips is easy;
- If a hole persists from $R_{\epsilon'}$ to R_ϵ, then it must exists in C_ϵ.

Persistent 1-Homology in Rips Complexes

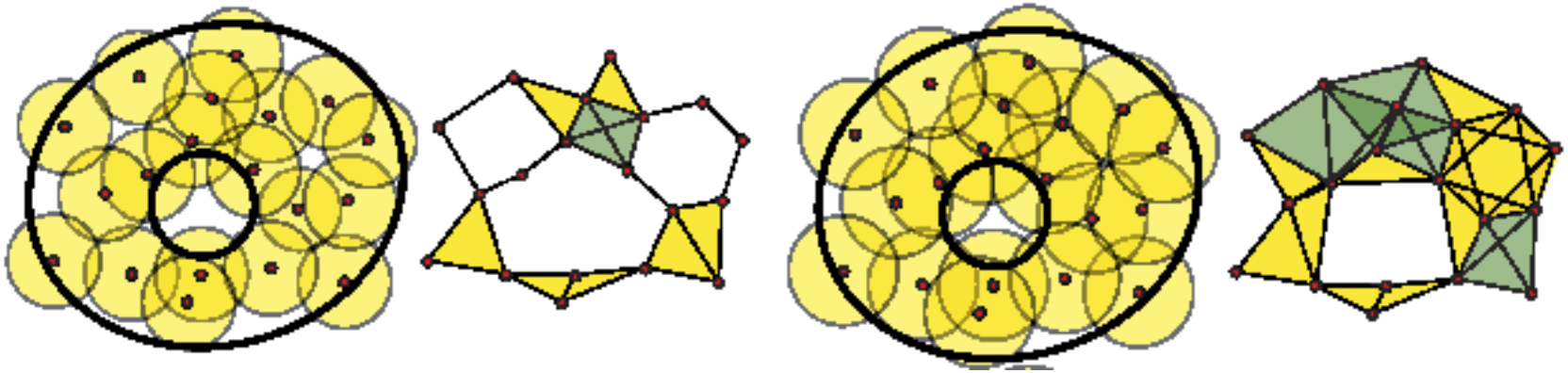


Figure: Left: $R_{\epsilon'}$; Right: R_{ϵ}. The middle hole persists from $R_{\epsilon'}$ to R_{ϵ}.

Application: Natural Image Statistics

- G. Carlsson, V. de Silva, T. Ishkanov, A. Zomorodian (2008) On the local behavior of spaces of natural images, *International Journal of Computer Vision*, 76(1):1-12.
- An image taken by black and white digital camera can be viewed as a vector, with one coordinate for each pixel
- Each pixel has a "gray scale" value, can be thought of as a real number (in reality, takes one of 255 values)
- Typical camera uses tens of thousands of pixels, so images lie in a very high dimensional space, call it pixel space, $\mathcal{P}$

Natural Image Statistics

- **D. Mumford**: What can be said about the set of images $\mathcal{I} \subseteq \mathcal{P}$ one obtains when one takes many images with a digital camera?
- **Lee, Mumford, Pedersen**: Useful to study local structure of images statistically

Natural Image Statistics

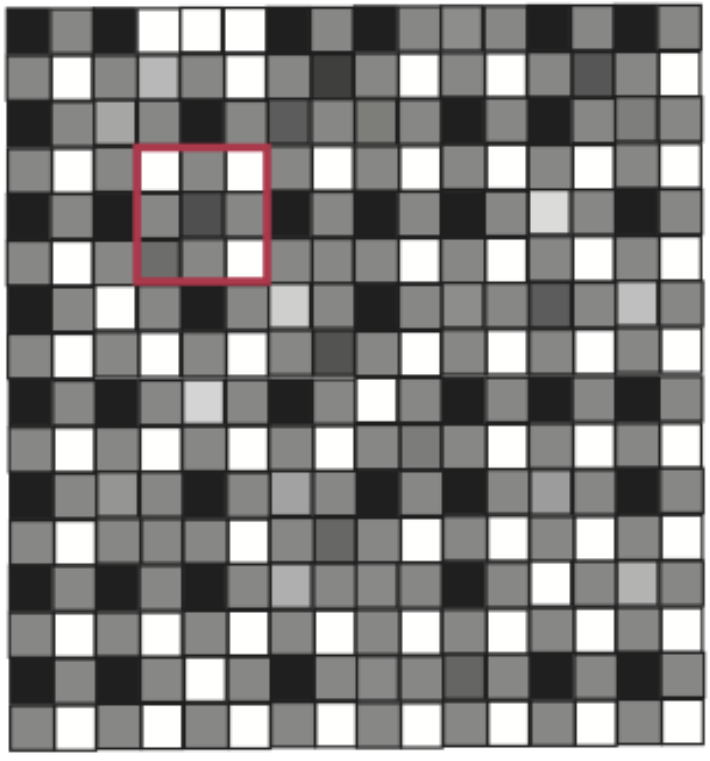


Figure: 3×3 patches in images

Natural Image Statistics

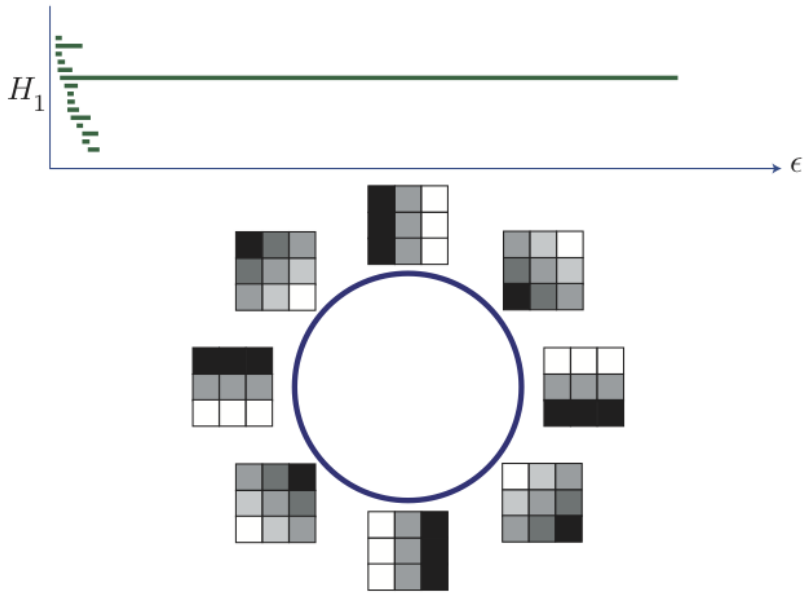
Lee-Mumford-Pedersen [LMP] study only high contrast patches.

- Collect: $4.5M$ high contrast patches from a collection of images obtained by van Hateren and van der Schaaf
- Normalize mean intensity by subtracting mean from each pixel value to obtain patches with mean intensity $= 0$
- Puts data on an 8-D hyperplane, $\approx R^8$
- Furthermore, normalize contrast by dividing by the norm, so obtain patches with norm $= 1$, whence data lies on a 7-D ellipsoid, $\approx S^7$

Natural Image Statistics: Primary Circle

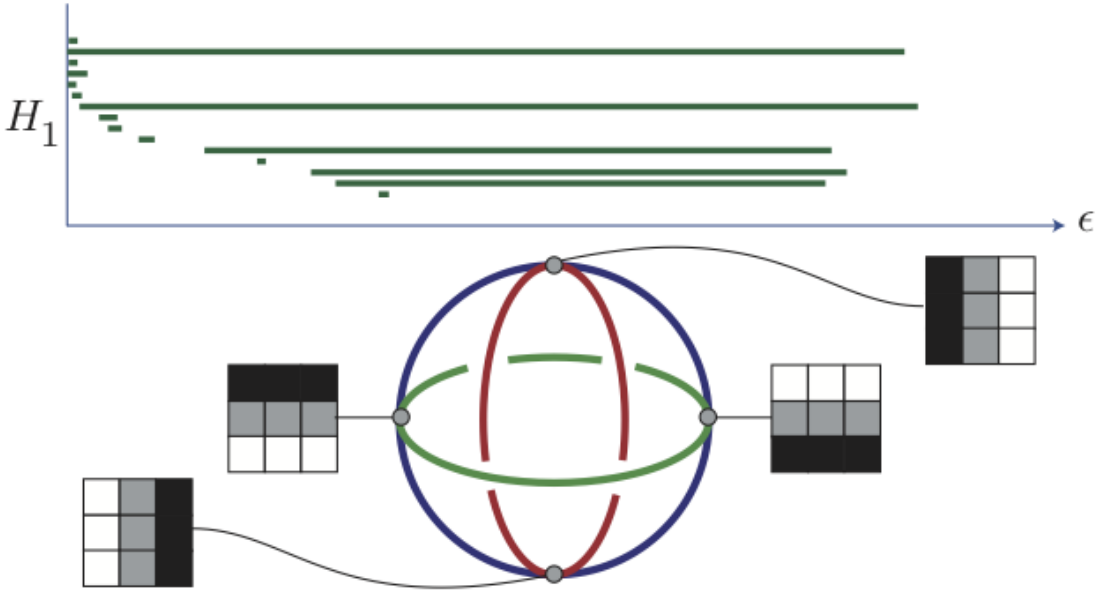
High density subsets $\mathcal{M}(k = 300, t = 0.25)$:

- Codensity filter: $d_k(x)$ be the distance from x to its k-th nearest neighbor
  - the lower $d_k(x)$, the higher density of x
- Take $k = 300$, the extract $5,000$ top $t = 25\%$ densest points, which concentrate on a primary circle



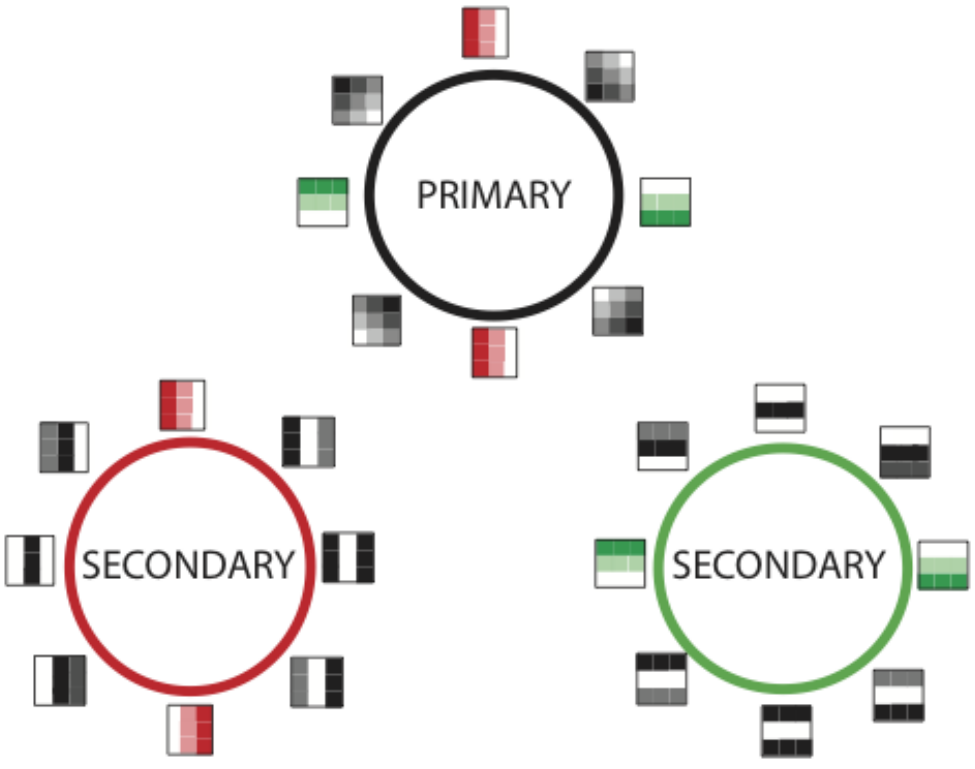
Natural Image Statistics: Three Circles

- Take $k = 15$, the extract $5,000$ top 25% densest points, which shows persistent $\beta_1 = 5$, 3-circle model

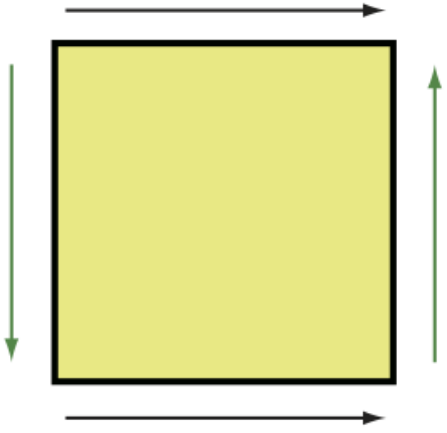


Natural Image Statistics: Three Circles

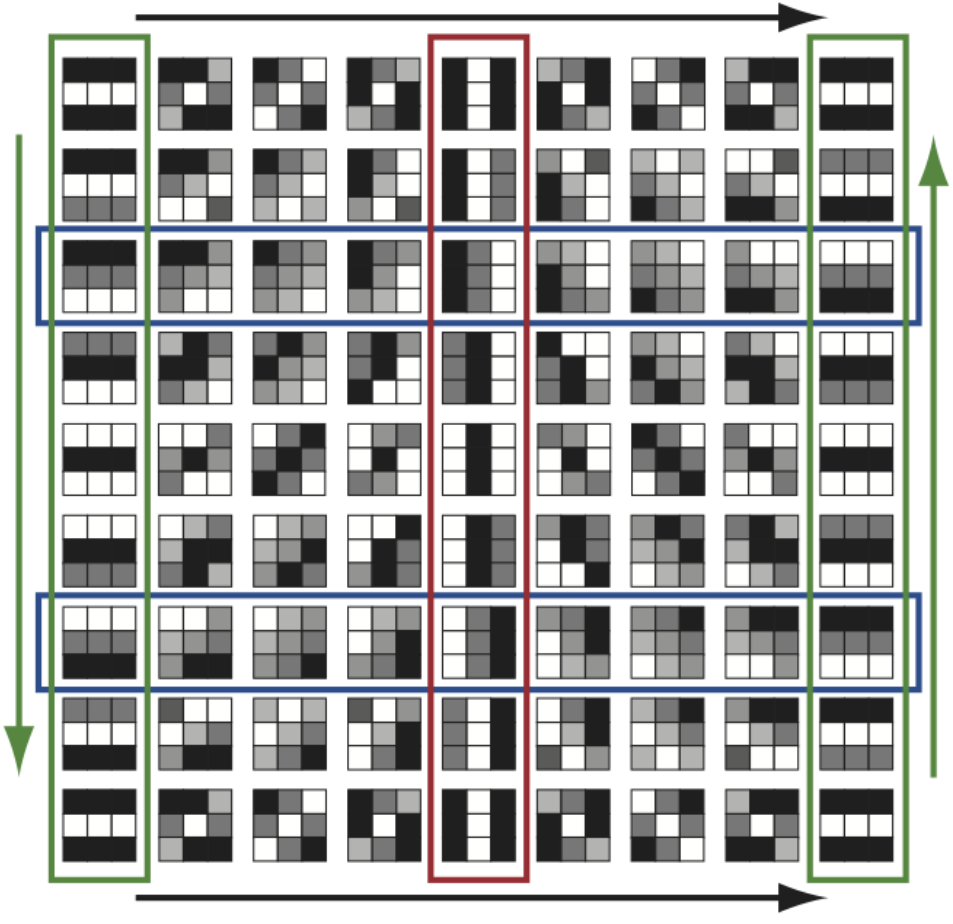
Generators for 3 circles



Natural Image Statistics: Klein Bottle



Natural Image Statistics: Klein Bottle Model



Reference

- Edelsbrunner, Letscher, and Zomorodian (2002) Topological Persistence and Simplification.
- Ghrist, R. (2007) Barcdes: the Persistent Topology of Data. *Bulletin of AMS*, 45(1):61-75.
- Edelsbrunner, Harer (2008) Persistent Homology - a survey. *Contemporary Mathematics*.
- Carlsson, G. (2009) Topology and Data. *Bulletin of AMS*, 46(2):255-308.
- Camara et al. (2016) Topological Data Analysis Generates High-Resolution, Genome-wide Maps of Human Recombination, *Cell Systems*, 3(1): 83–94.
- Wei, Guowei, (2017) Persistent Homology Analysis of Biomolecular Data, *SIAM News*.
- Raul Rabadan and Andrew J. Blumberg (2020). Topological Data Analysis for Genomics and Evolution. *Cambridge University Press*.