

Topological Data Analysis with Mapper: an Implementation in Cytoscape and an Application to Aptamers

George Clare Kennedy

August 29, 2024

University of Iowa

Outline

Why TDA?

Mapper and Its Flavors

Cytoscape to the Rescue

Aptamers

Mapping out Future Directions

Section Map (!)

Why TDA?

Mapper and Its Flavors

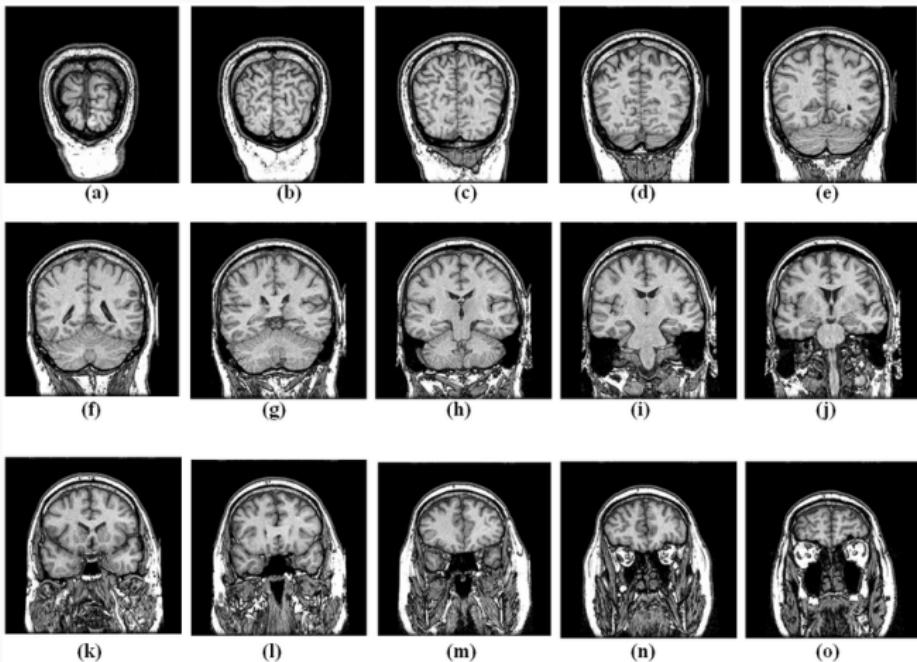
Cytoscape to the Rescue

Aptamers

Mapping out Future Directions

Data is Big

- Modern techniques allow for rich data collection and storage
- Size of datasets can be enormous in both observations (rows) and variables (columns)



Data is Big

- Modern techniques allow for rich data collection and storage
- Size of datasets can be enormous in both observations (rows) and variables (columns)



Geometry is Hard

- High-dimensional space is extremely unintuitive
- If $V_n(r)$ is the volume of the n -dimensional ball with radius r , then for any $\varepsilon > 0$,

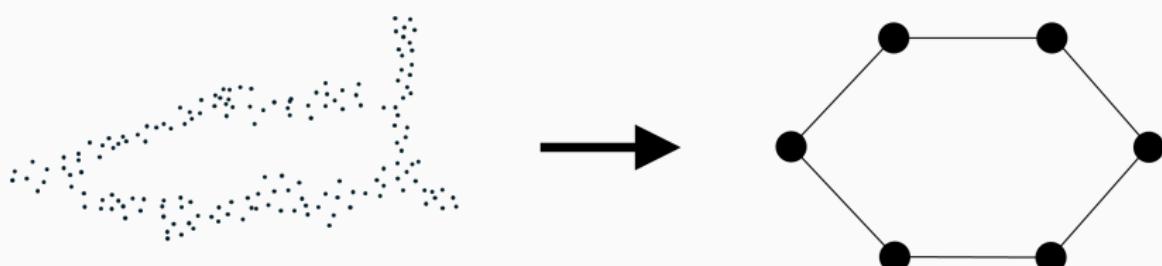
$$\lim_{n \rightarrow \infty} \frac{V_n(1 - \varepsilon)}{V_n(1)} = 0$$

i.e., the volume of balls lives almost entirely at the boundary

- Trying to analyze many characteristics creates combinatorial problems ($n!$ is big!)

Toning It Down

- Broad idea: high dimensions \implies low dimensions
- More specific idea: build a simplicial complex
- Simpler idea: build a 1-dimensional simplicial complex (that is, a graph)
- Enter: the Mapper algorithm (Singh et al, 2007) [3]



Section Map (!)

Why TDA?

Mapper and Its Flavors

Cytoscape to the Rescue

Aptamers

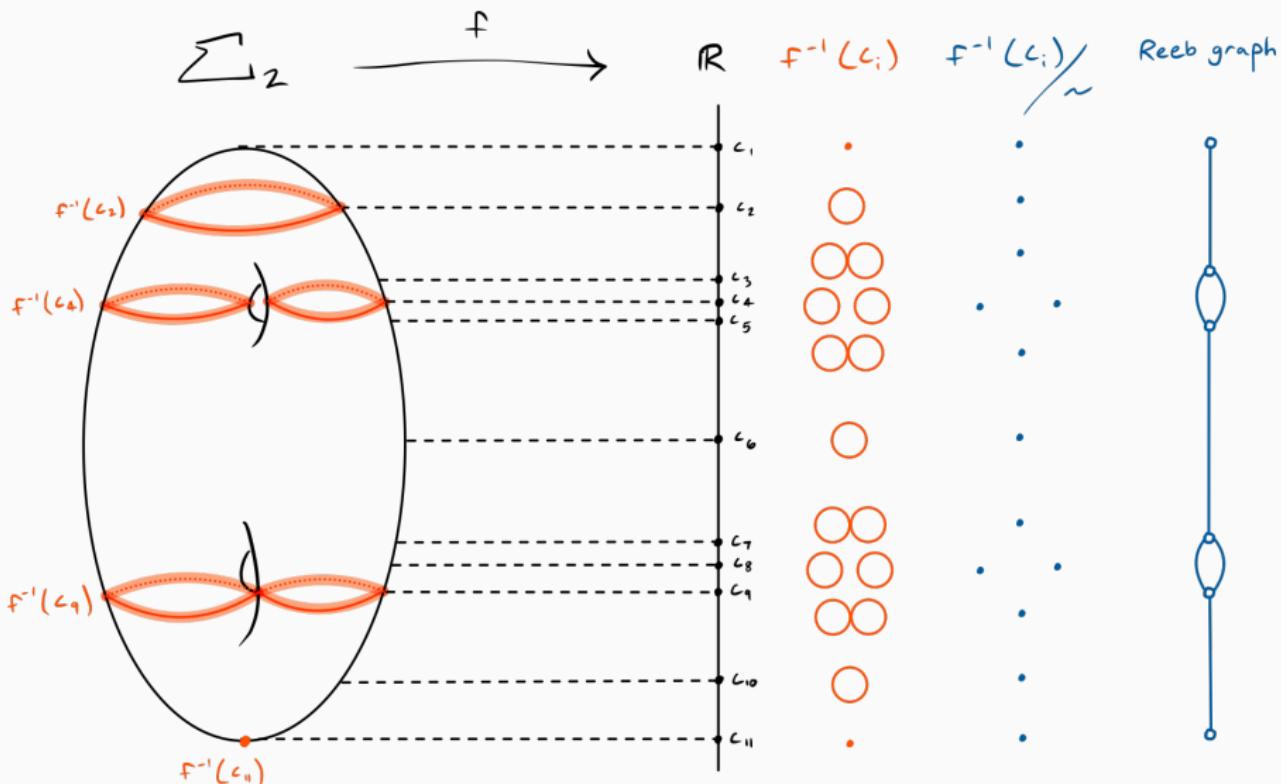
Mapping out Future Directions

Motivation: Reeb Graph

- Idea: construct graph reflecting level sets of a “filter” function
- Formally, given a topological space X and a continuous function $f : X \rightarrow \mathbb{R}$, define an equivalence relation \sim on X where $x \sim y$ if x and y live in the same connected component of a level set $f^{-1}(c)$ for some $c \in \mathbb{R}$.
- The **Reeb graph**¹ is X / \sim , taken with the quotient topology.

¹Despite names this is not always a graph

Motivation: Reeb Graph



Mapper: Original Flavor

- How can we apply this to the discrete setting?
- Topological space $X \implies$ point cloud P (a discrete set of points in a space)
- Filter function: $f : P \rightarrow \mathbb{R}$
- Level sets of points \implies level sets of overlapping intervals
- Connected components \implies clusters
- Quotient space \implies intersection graph

Original Mapper Algorithm: Overview

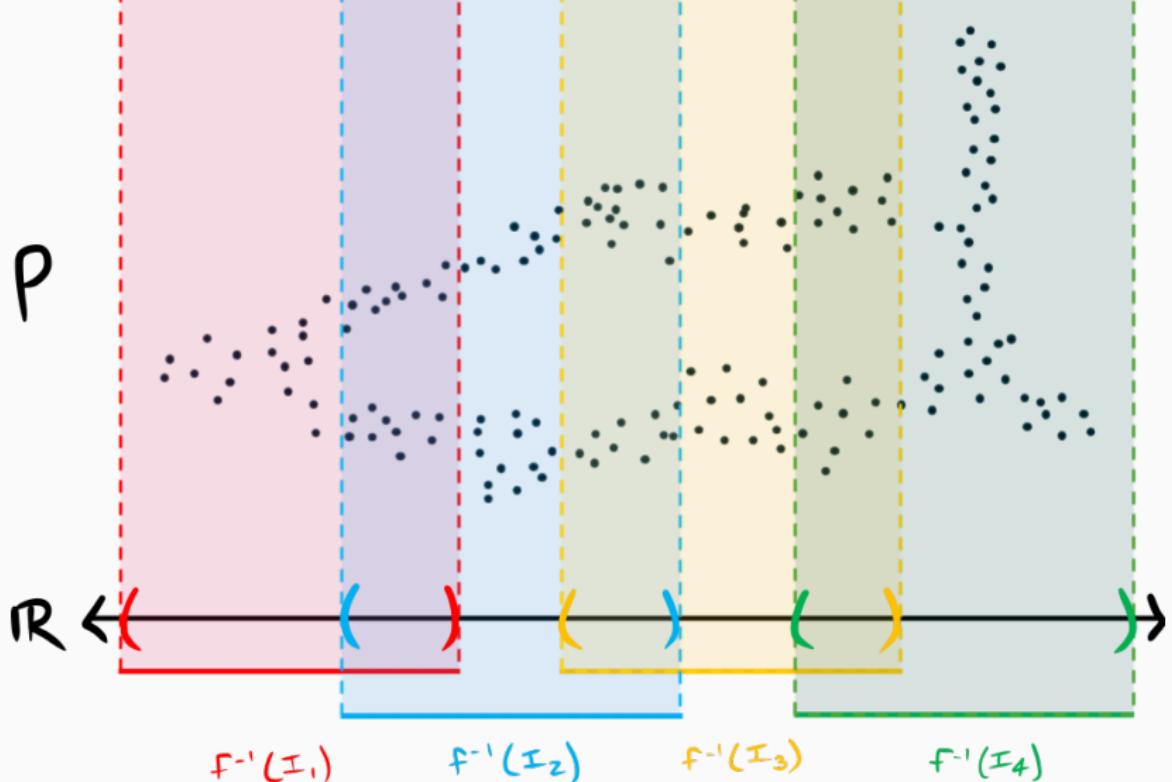
- Ingredients:
 - Point cloud P
 - Filter function $f : P \rightarrow \mathbb{R}$
 - Collection of overlapping intervals $\{I_1, \dots, I_k\}$
 - Clustering algorithm
- Output:
 - Finite graph M

Point Cloud

Our example point cloud will live in \mathbb{R}^2 :

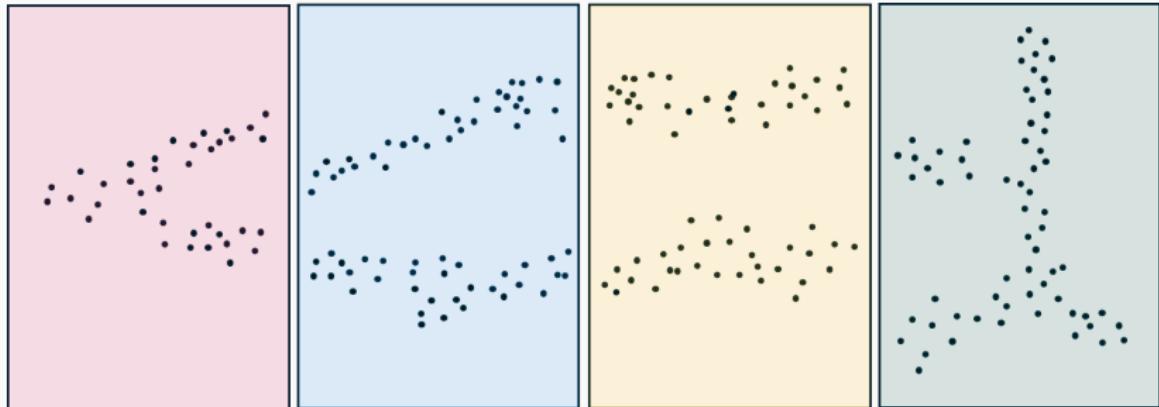


Filtering

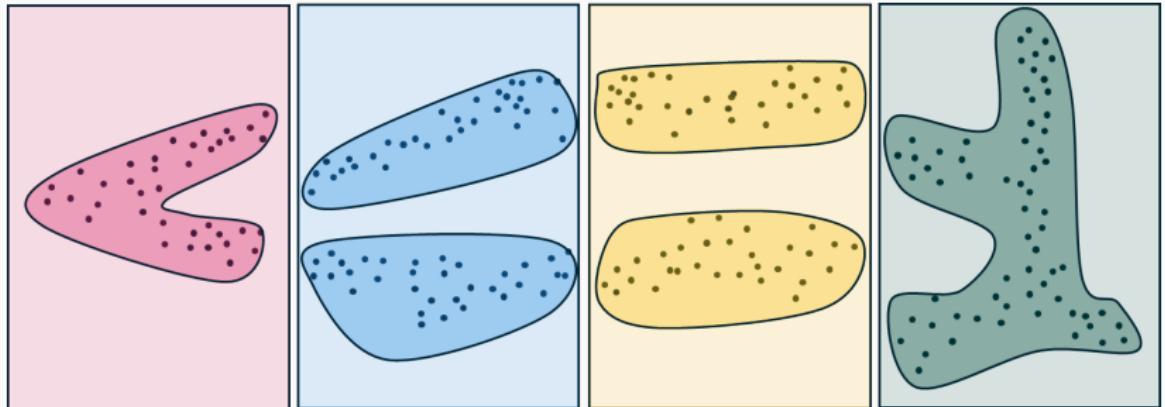


Here, $f : P \rightarrow \mathbb{R}$ is projection to the x -coordinate.

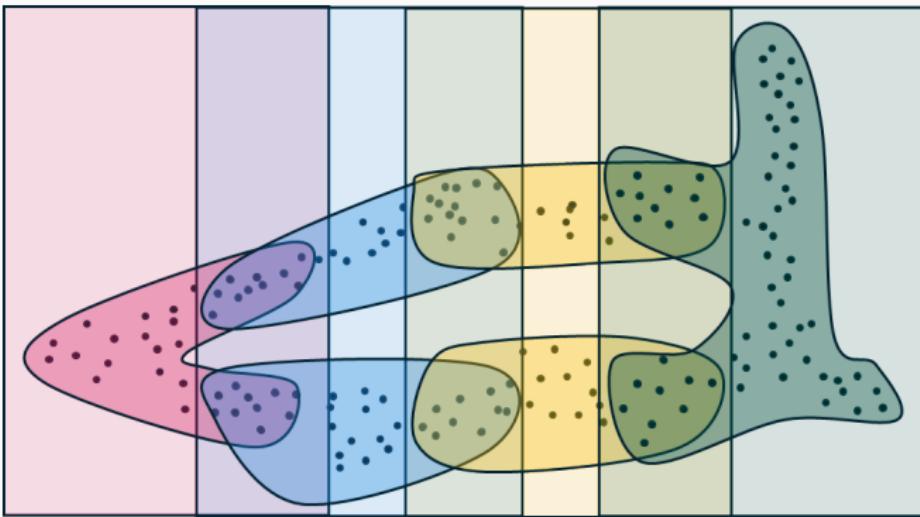
Filtering



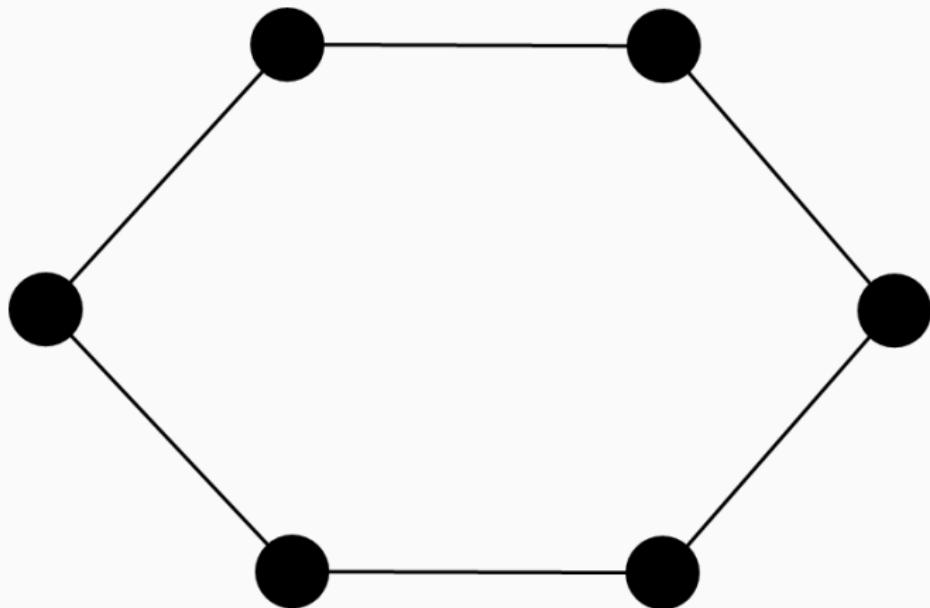
Clustering



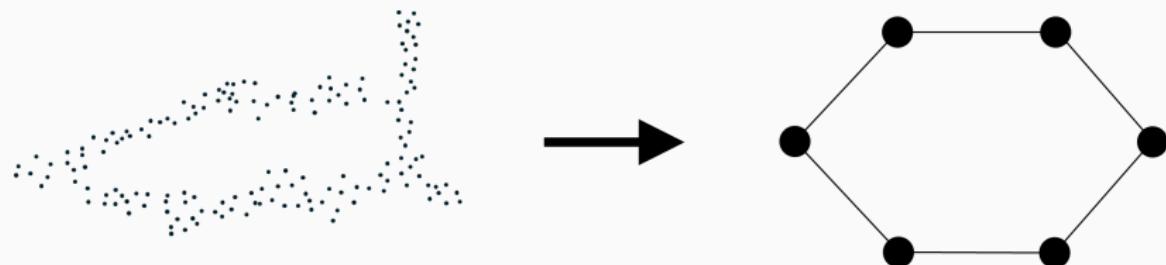
Overlapping Clusters



Output

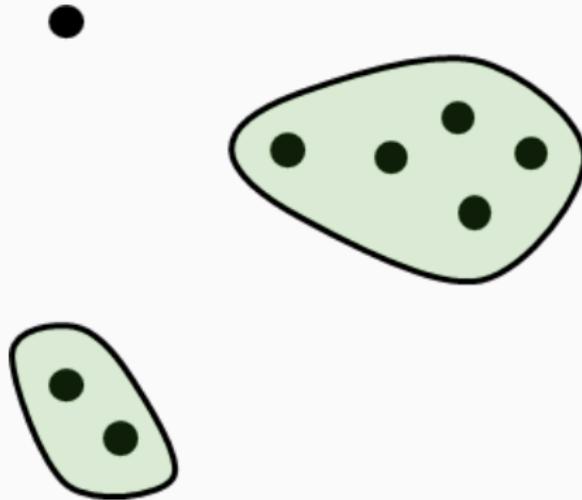


Original Mapper: Problems



- We lost so much stuff! Clustering takes work.
- Graphs are abstract combinatorial structures; they convey no other information
- Potentially interesting features can “bypass” the filter
- Large number of parameters complicates effectiveness
- Output heavily depends on choice of clustering method

Speaking of Clustering Methods...



There are a large variety of clustering algorithms, including:

- Hierarchical clustering
- k -means clustering (centroid based)
- DBSCAN (density based)
- Topological clustering

How to Cluster, Hierarchically

- Two ingredients:
 - A discrete set of points X , equipped with a distance function $d : X \times X \rightarrow \mathbb{R}$
 - A **linking criterion**, a (partial) function $\ell : \mathcal{P}(X) \times \mathcal{P}(X) \rightarrow \mathbb{R}$ which measures the distance between disjoint sets of points
- Process:
 1. Begin by considering each point as an individual cluster as part of a collection \mathcal{C} .
 2. Find the two $A, B \in \mathcal{C}$ that minimize ℓ . Merge these clusters.
 3. Repeat step 2 until \mathcal{C} consists of up to a single cluster.
- We record the resulting hierarchy of clusters using a **dendrogram**, which records information about the merging process.

Example: Single Linkage Clustering

- Data: 8 points in \mathbb{R}^2 , with the usual metric



- Linking criterion:

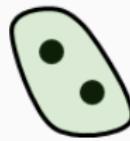
$$\ell(A, B) = \min_{a \in A, b \in B} \{d(a, b)\}$$

- Starting dendrogram:



Single Linkage Clustering In Action

$$\ell(A, B) = \min_{a \in A, b \in B} \{d(a, b)\}$$

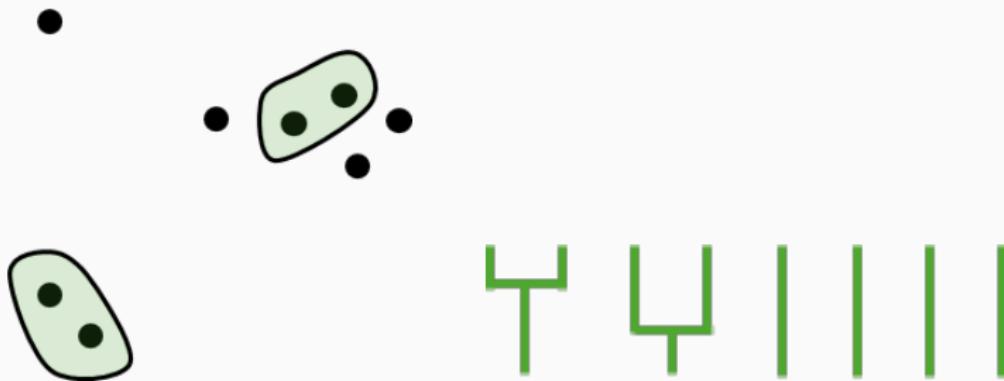


We merge edges at a height proportional to the corresponding value of ℓ (the merge height):



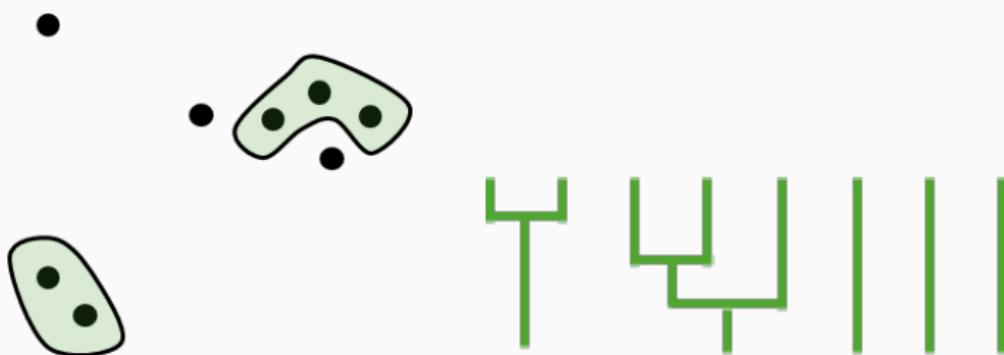
Single Linkage Clustering In Action

$$\ell(A, B) = \min_{a \in A, b \in B} \{d(a, b)\}$$



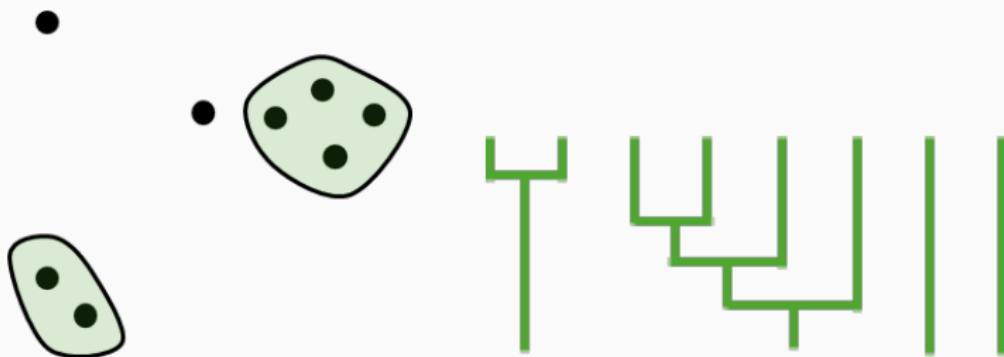
Single Linkage Clustering In Action

$$\ell(A, B) = \min_{a \in A, b \in B} \{d(a, b)\}$$



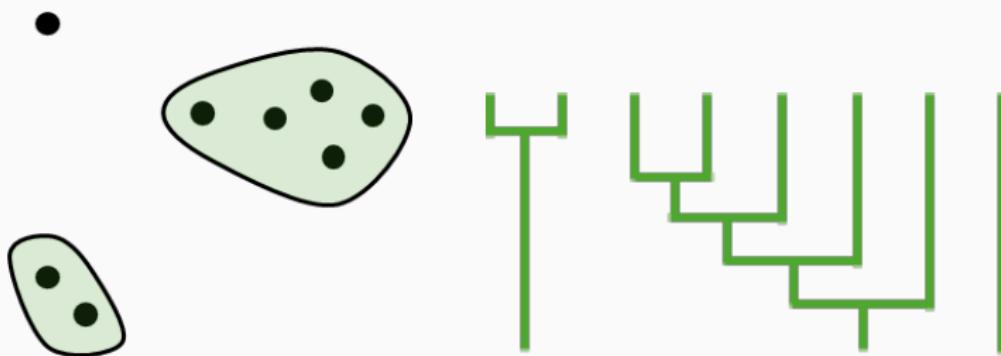
Single Linkage Clustering In Action

$$\ell(A, B) = \min_{a \in A, b \in B} \{d(a, b)\}$$



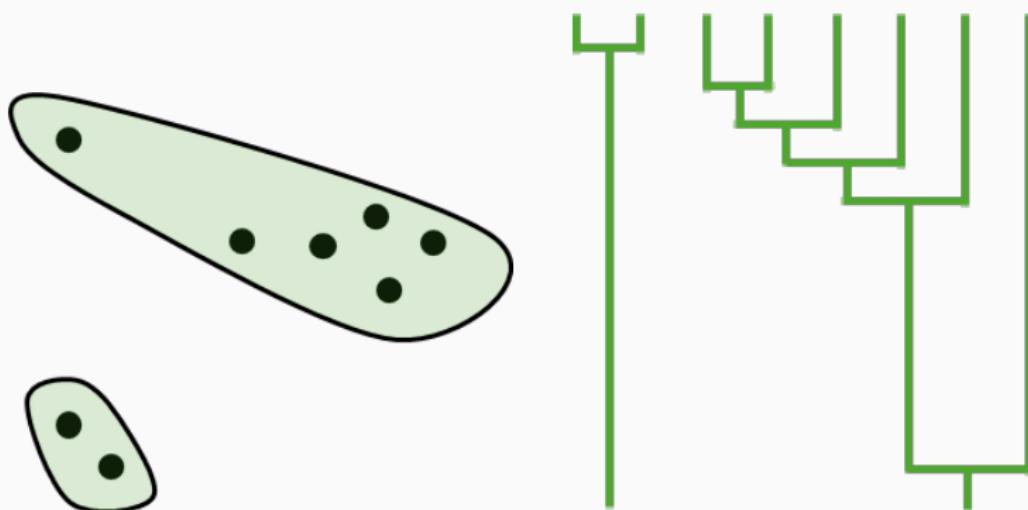
Single Linkage Clustering In Action

$$\ell(A, B) = \min_{a \in A, b \in B} \{d(a, b)\}$$



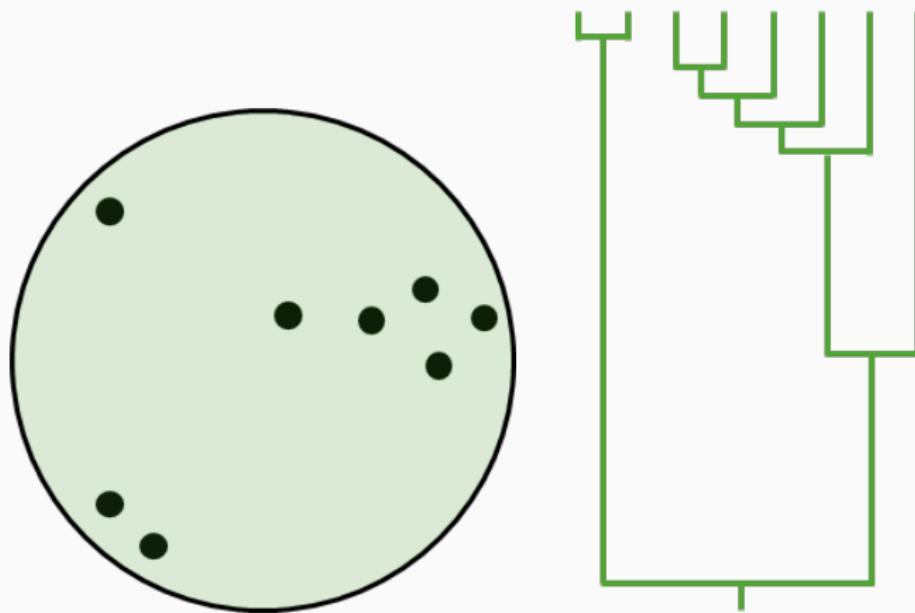
Single Linkage Clustering In Action

$$\ell(A, B) = \min_{a \in A, b \in B} \{d(a, b)\}$$



Single Linkage Clustering In Action

$$\ell(A, B) = \min_{a \in A, b \in B} \{d(a, b)\}$$



Ballmapper (Dłotko, 2019)

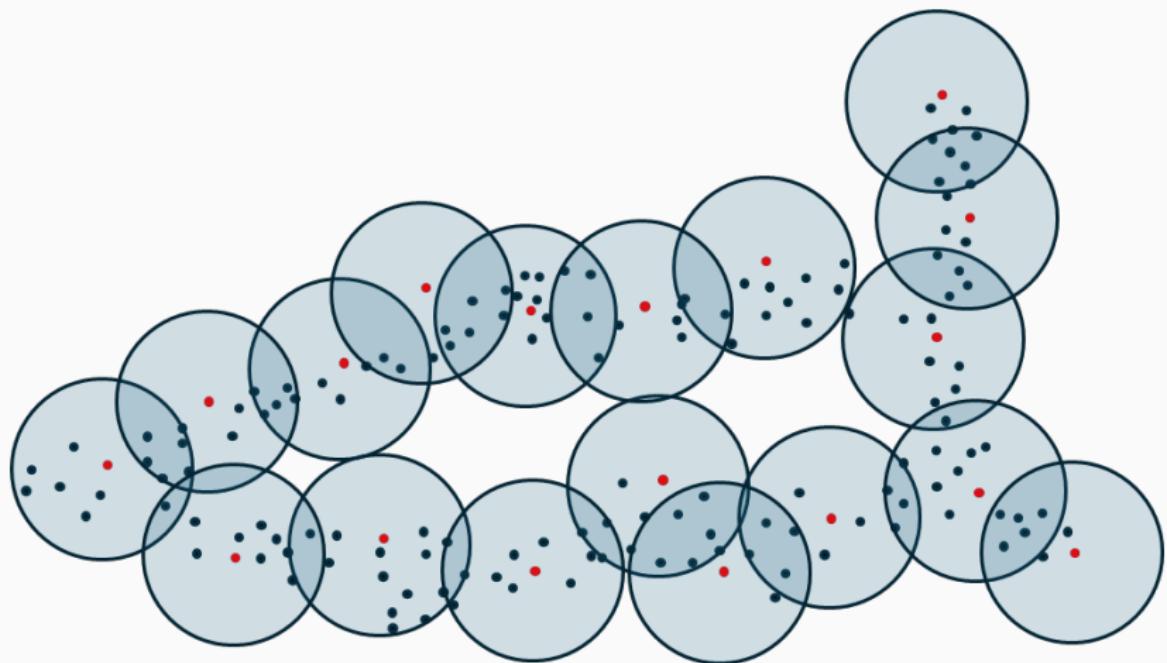
- Original flavor has a lot of choices to make
- Idea: come up with a one (ish)-parameter Mapper
- **Ballmapper** [2]: in place of a conventional filter, cover the dataset with overlapping ε -balls
- Specifically, we want a cover $C = \bigcup_i B(x_i, \varepsilon)$ such that:
 - Every datapoint x is contained in $B(x_i, \varepsilon)$ for some x_i
 - If x_j is a ball center, then the only ball containing it is $B(x_j, \varepsilon)$
- Graph construction unchanged

Ballmapper Algorithm: Overview

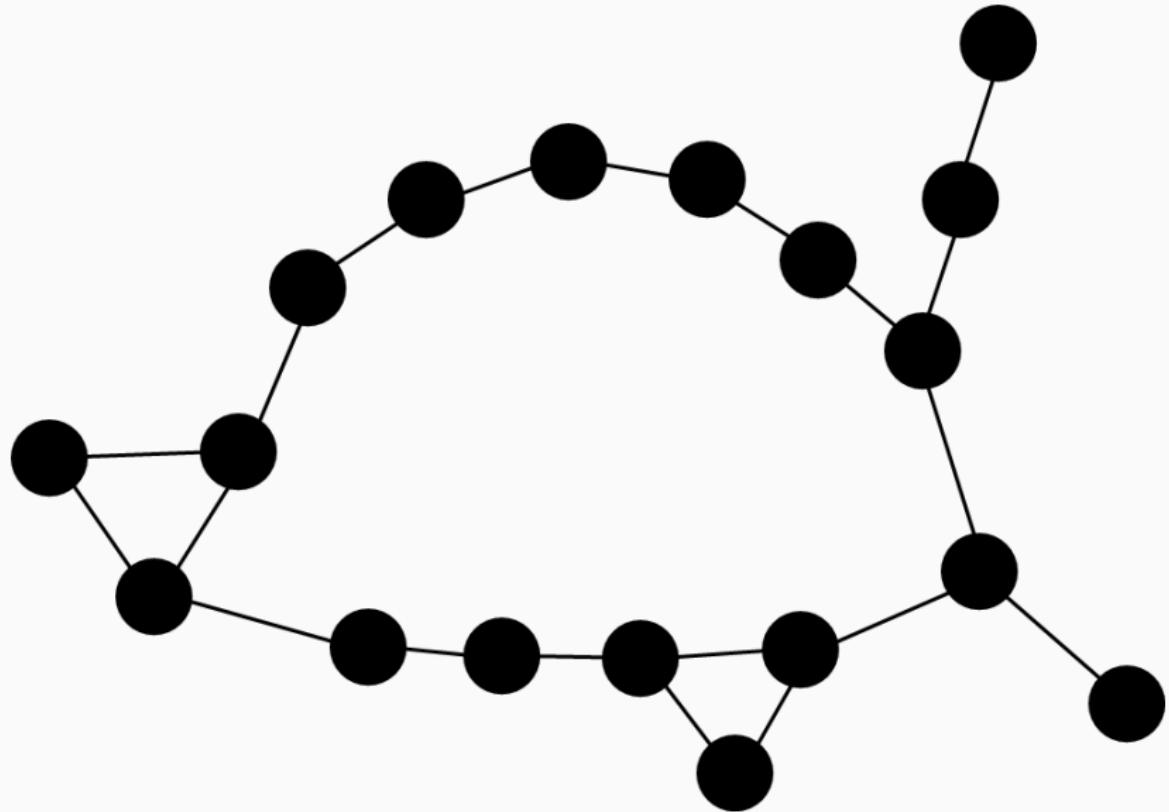
- Ingredients:
 - Point cloud P with distance function d
 - Ball radius $\varepsilon > 0$
 - Suitable cover of data $\bigcup B_\varepsilon(x_i)$
- Output:
 - Finite graph BM

Balling the Data

- Can be done quickly with a greedy method
- May also use k -means clustering, etc.



Output



Ballmapper: Problems



- Output is still just a graph!
- Balls become black boxes
- Can be quite noisy if ε is too small, and meaningless for ε too large

Refined Ballmapper

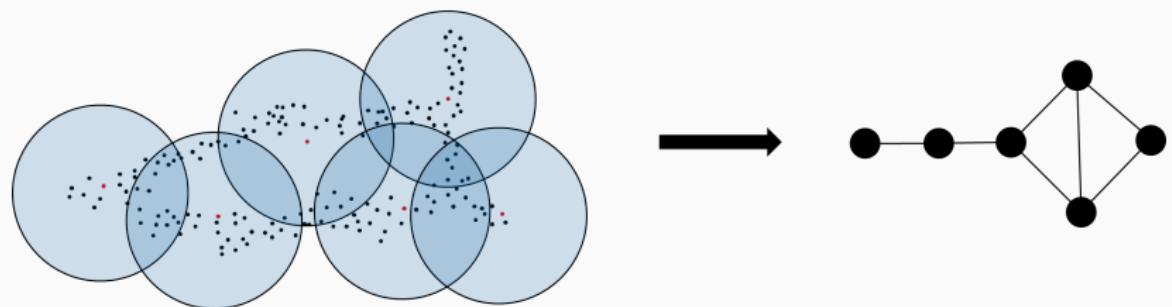
- Idea: combine Ball filtering and Original clustering
- Bin by balling, then cluster within balls as in Original
- Allows for comparison of two different metrics² at the same time on the same data

²Could be not strictly metrics!

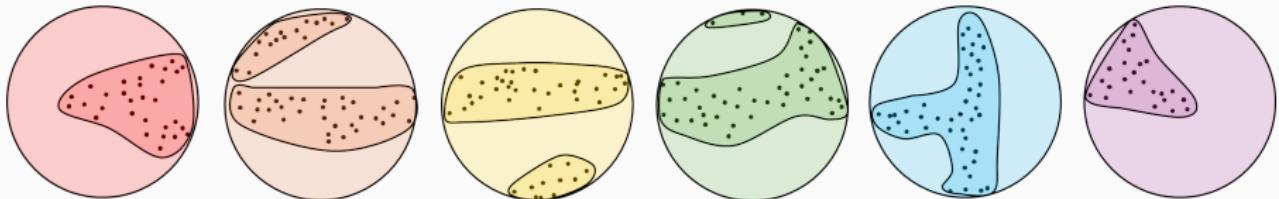
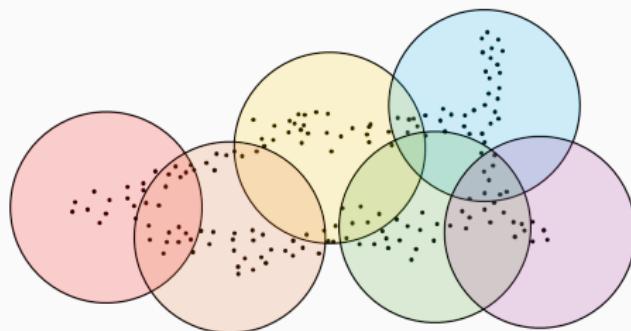
Refined Ballmapper: Overview

- Ingredients:
 - Point cloud P with distance function d
 - Ball radius $\varepsilon > 0$
 - Suitable cover of data $\bigcup B_\varepsilon(x_i)$
 - Clustering algorithm
- Output:
 - Finite graph RBM

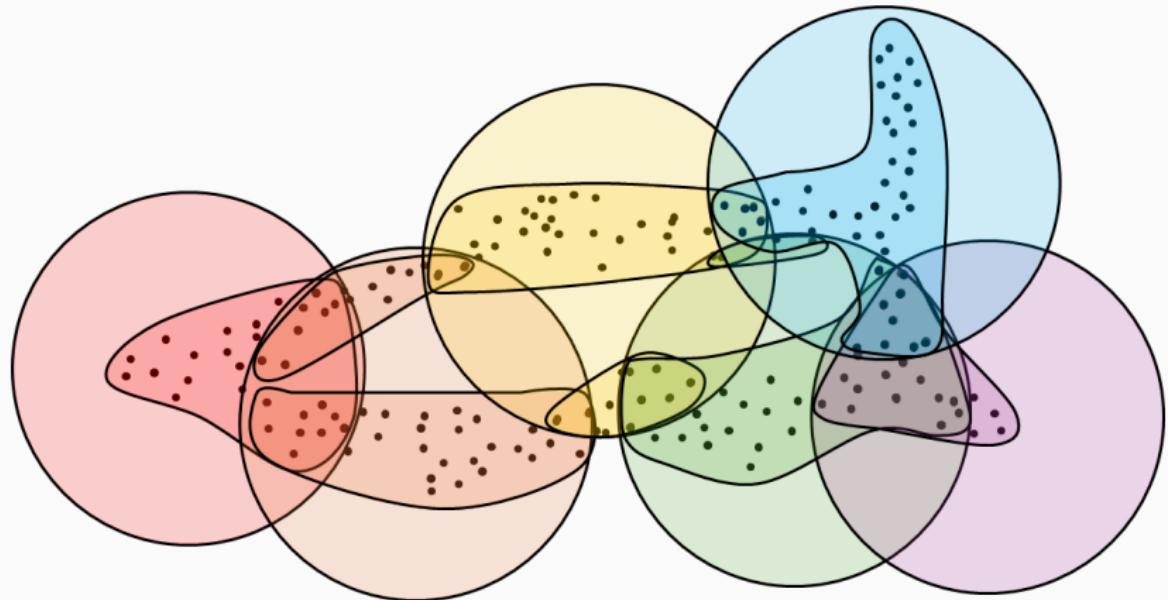
Standard Ballmapper



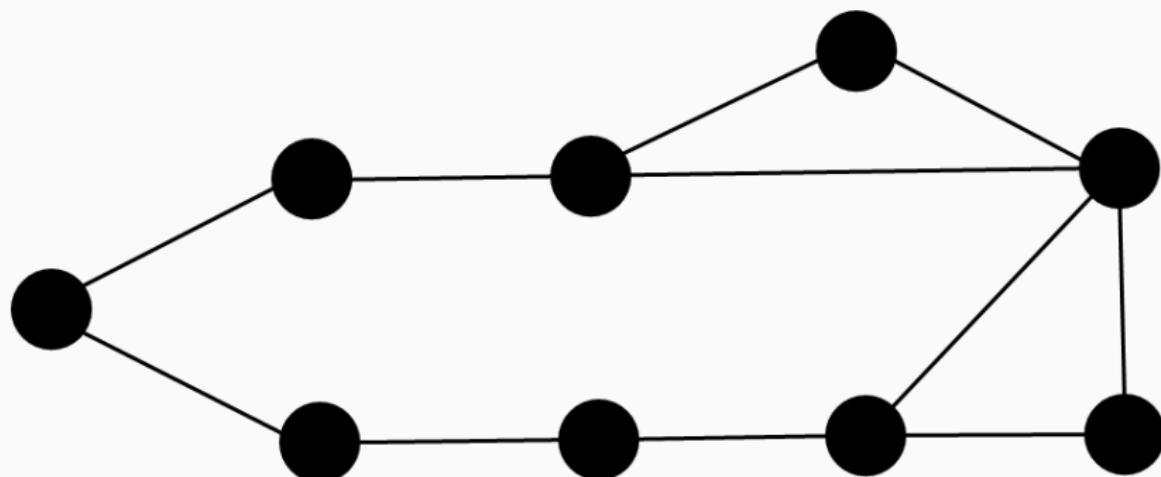
Balls as Bins



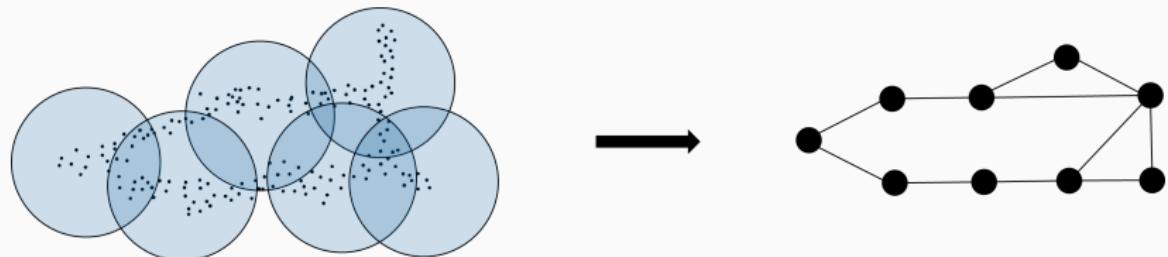
Balls Together



Output



Refined Ballmapper: Problems



- Choice of cover now strongly influences meaningfulness of clusters
- Choice of clustering algorithm is still an issue
- Still have that graph problem!

Section Map (!)

Why TDA?

Mapper and Its Flavors

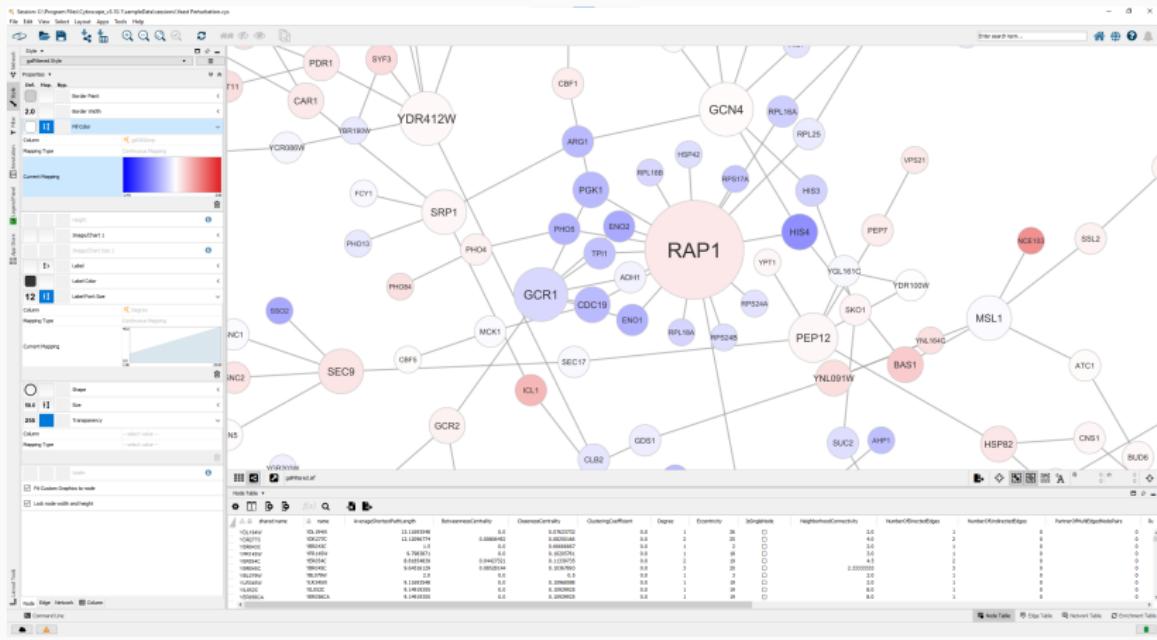
Cytoscape to the Rescue

Aptamers

Mapping out Future Directions

What is Cytoscape?

- Powerful network analysis software (written in Java)
 - Used primarily by bioinformaticists but is a general use program



Features

- Networks are stored as separate tables of nodes and edges
- Tables can be augmented with any number or type of columns
- Any column can then be associated with a visual characteristic of the network

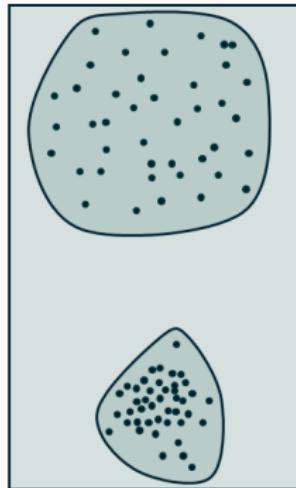
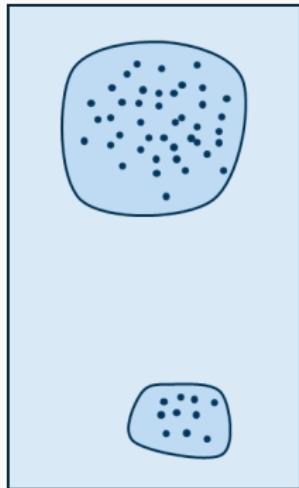
Node Table ▾

The screenshot shows a software interface titled "Node Table". At the top, there is a toolbar with icons for settings, refresh, search, and file operations. Below the toolbar is a header row containing the column names: "shared name", "name", "AverageShortestPathLength", "BetweennessCentrality", "ClosenessCentrality", and "ClusteringCoefficient". The main body of the table lists 12 rows of data, each representing a node. The data is as follows:

shared name	name	AverageShortestPathLength	BetweennessCentrality	ClosenessCentrality	ClusteringCoefficient
YDL194W	YDL194W	13.11693548	0.0	0.07623732	0.0
YDR277C	YDR277C	12.12096774	0.00806452	0.08250166	0.0
YBR043C	YBR043C	1.5	0.0	0.66666667	0.0
YPR145W	YPR145W	9.7983871	0.0	0.10205761	0.0
YER054C	YER054C	8.81854839	0.04427321	0.11339735	0.0
YBR045C	YBR045C	9.64516129	0.08528144	0.10367893	0.0
YBL079W	YBL079W	2.0	0.0	0.5	0.0
YLR345W	YLR345W	9.11693548	0.0	0.10968598	0.0
YIL052C	YIL052C	9.14919355	0.0	0.10929925	0.0
YER056CA	YER056CA	9.14919355	0.0	0.10929925	0.0

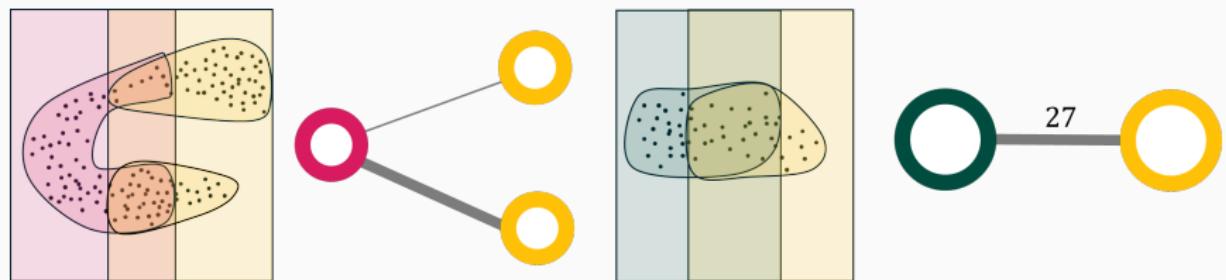
Styling Mapper: Vertices

- Node size/label → cluster size
- Node border color → associated level set/filter value/ball
- Node fill color → cluster dispersion



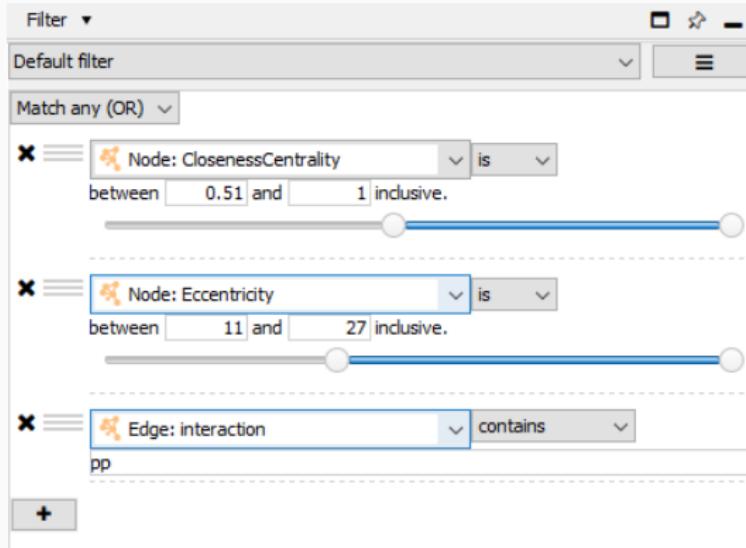
Styling Mapper: Edges

- Edge thickness/opacity → cluster intersection strength
- Edge label → cluster intersection size



Exploring the Graph

- Cytoscape can calculate classical network statistics
 - Centrality measures
 - Clustering coefficients
 - Modularity classes
- We can filter out nodes/edges by characteristics



Possible Capabilities

- Cytoscape is open source and was designed to be modified
- Possible projects here include:
 - Assign energy function to edges and apply layout algorithm
 - Animation between networks (say, from RBM to BM or reverse)
 - More graph algorithms (finding cores, clique detection, etc)

Section Map (!)

Why TDA?

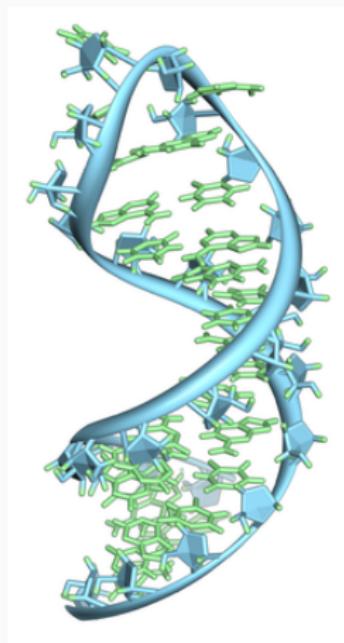
Mapper and Its Flavors

Cytoscape to the Rescue

Aptamers

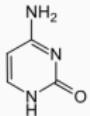
Mapping out Future Directions

- RNA (ribonucleic acid) and DNA (deoxyribonucleic acid) are polymers which carry genetic sequences and have additional structure



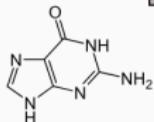
RNA and DNA

Cytosine



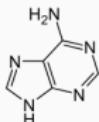
C

Guanine



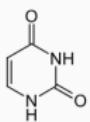
G

Adenine



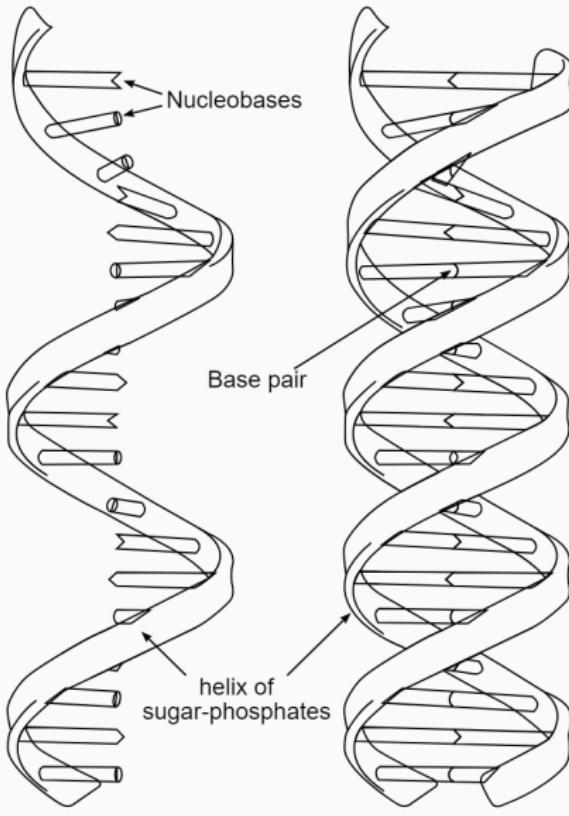
A

Uracil



U

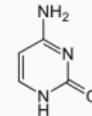
Nucleobases
of RNA



RNA

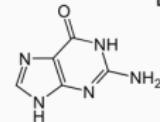
DNA

Cytosine



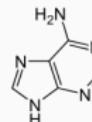
C

Guanine



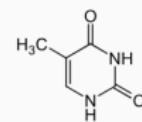
G

Adenine



A

Thymine

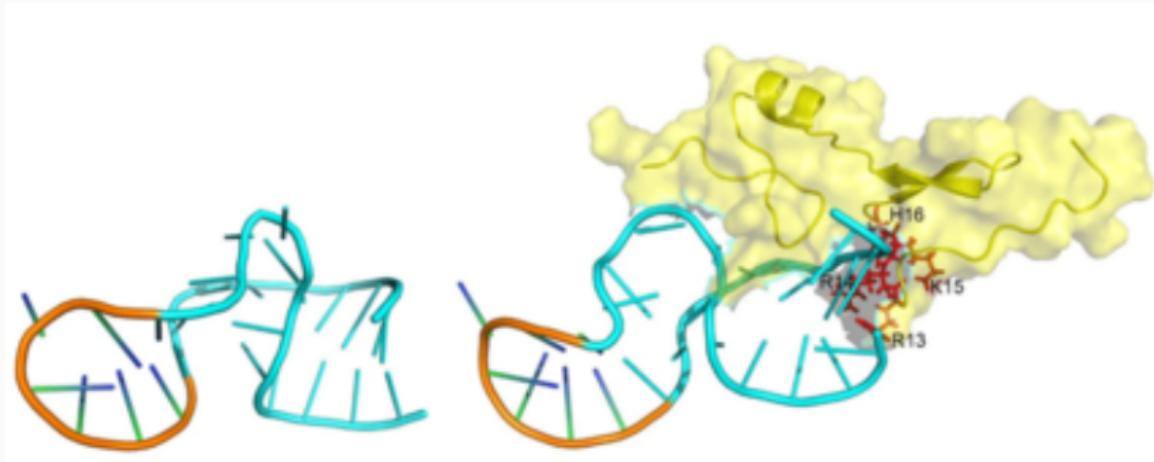


T

Nucleobases
of DNA

What Is an Aptamer?

- Aptamers are synthetic RNA molecules that bind to a specific target
- Similar function to antibodies, but much smaller
- **Genetic code not expressed**



Comparing Aptamers

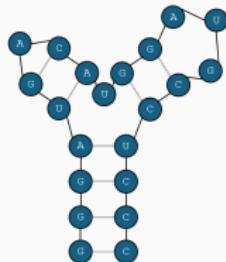
- For TDA to work we need a notion of distance among aptamers
- Aptamers have two characteristics: their genetic code and their structure
- Distance between sequences: Levenshtein distance
- Distance between structures: tree distance

Distance Functions

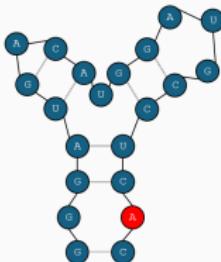
- The **Levenshtein distance** between two strings a and b is the minimum number of deletions, insertions, or substitutions needed to change a into b .
 - $\text{lev}(\text{house}, \text{mouse}) = 1$ (single substitution)
 - $\text{lev}(\text{cat}, \text{tarp}) = 3$ (2 substitutions, 1 insertion)
 - $\text{lev}(\text{mister}, \text{mister}) = 0$ (!)
- The **graph edit distance** between two graphs G and H is the minimum number of vertex/edge insertions or deletions needed to make G isomorphic to H .
 - $\text{dist}(K_n, K_{n-1}) = n$
 - $\text{dist}(C_n, C_k) = 2(|n - k| + 1)$

Aptamer Metrics

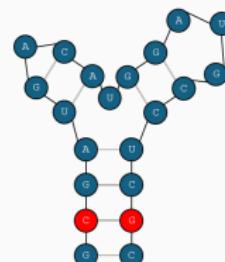
Aptamer X



Aptamer Y



Aptamer Z



GGGAUGACAUGGAUGGCCUCCC
((((((.)).((...))))))

δ_{Lev} : Levenshtein distance

δ_{Tree} : Tree distance

GGGAUGACAUGGAUGGCCUCAC
(.((((.)).((...)))).)

$$\delta_{\text{Lev}}(X, Y) = 1$$

$$\delta_{\text{Tree}}(X, Y) = 1$$

GGGCUGACAUGGAUGGCCGCC
((((((.)).((...))))))

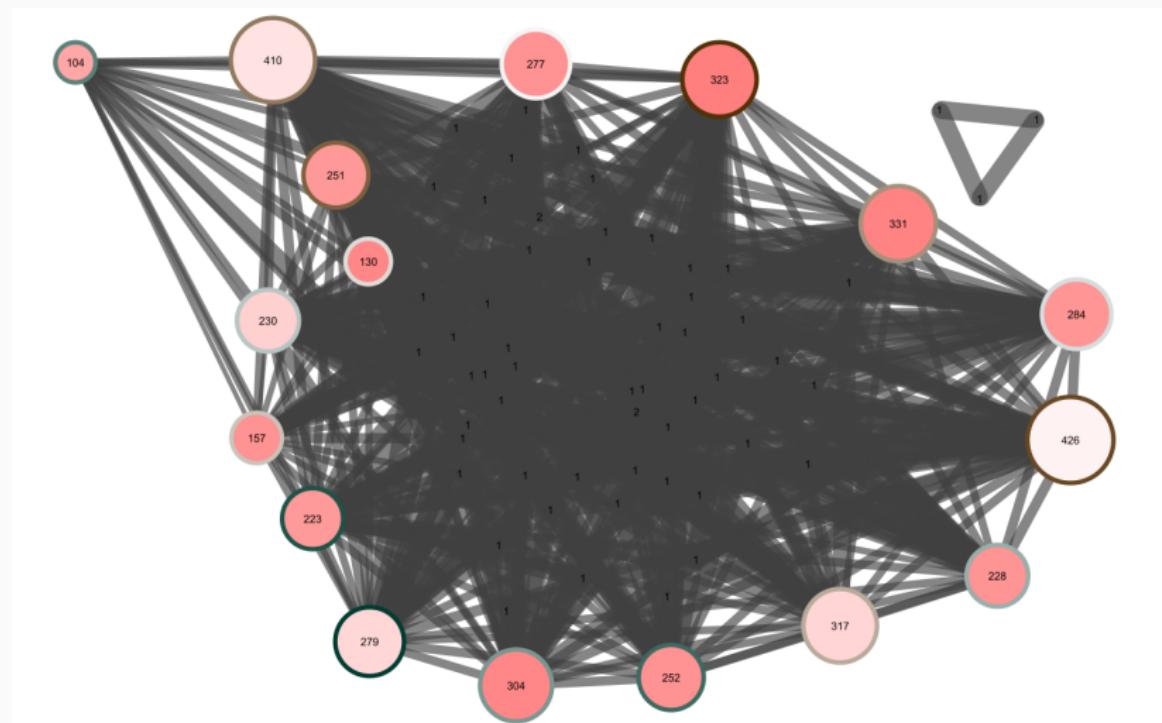
$$\delta_{\text{Lev}}(X, Z) = 2$$

$$\delta_{\text{Tree}}(X, Z) = 0$$

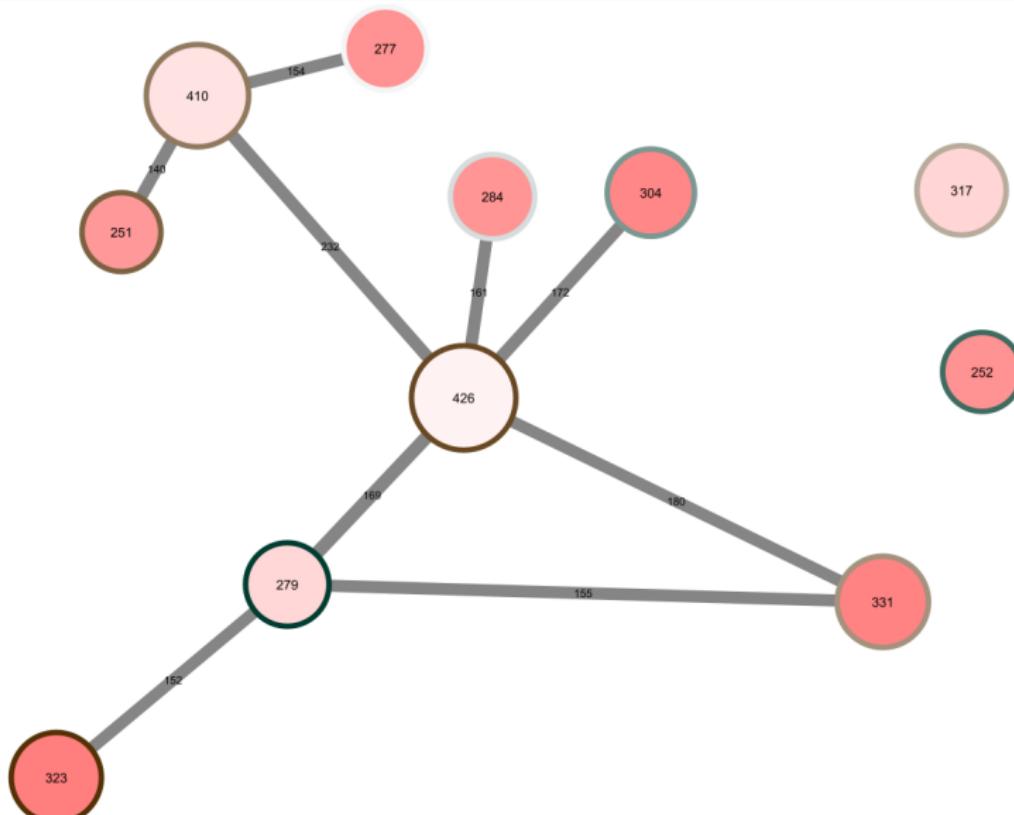
Aptamer Clustering With Mapper

- Flavor: Refined Ballmapper
- Idea: Ball using tree distance, cluster using Levenshtein distance
- Clustering method: single linkage hierarchical
- Vertices of the graph are clusters of aptamers related in both sequence and structure
- Graph structure may highlight families of aptamers or reveal other insights

Big Maptamer Graph



Pruned Maptamer Graph



Section Map (!)

Why TDA?

Mapper and Its Flavors

Cytoscape to the Rescue

Aptamers

Mapping out Future Directions

Other Clustering Algorithms

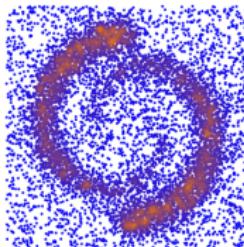
- Hierarchical:
 - Complete linkage:

$$\ell(A, B) = \max_{a \in A, b \in B} \{d(a, b)\}$$

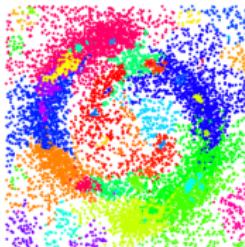
- (Unweighted) average linkage:

$$\ell(A, B) = \frac{1}{|A||B|} \sum_{a \in A} \sum_{b \in B} d(a, b)$$

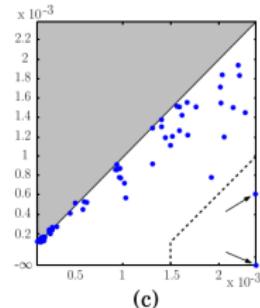
- k -means
- Topological Mode Analysis Tool (ToMATo) [1]



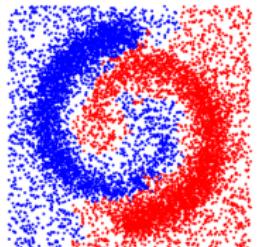
(a)



(b)



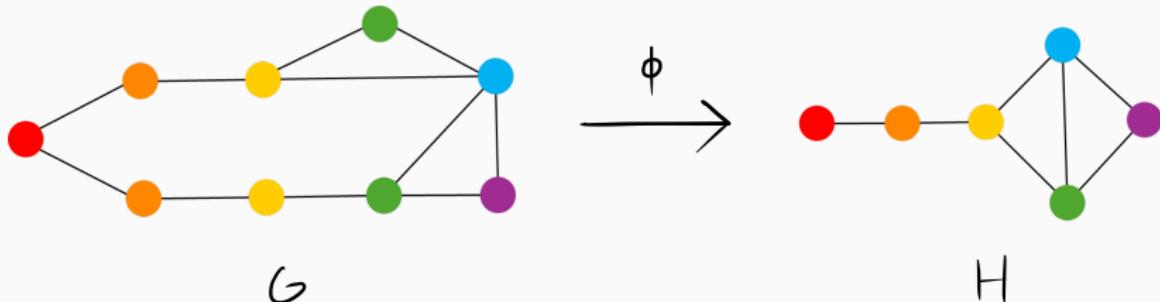
(c)



(d)

Refined Ballmapper and Graph Theory

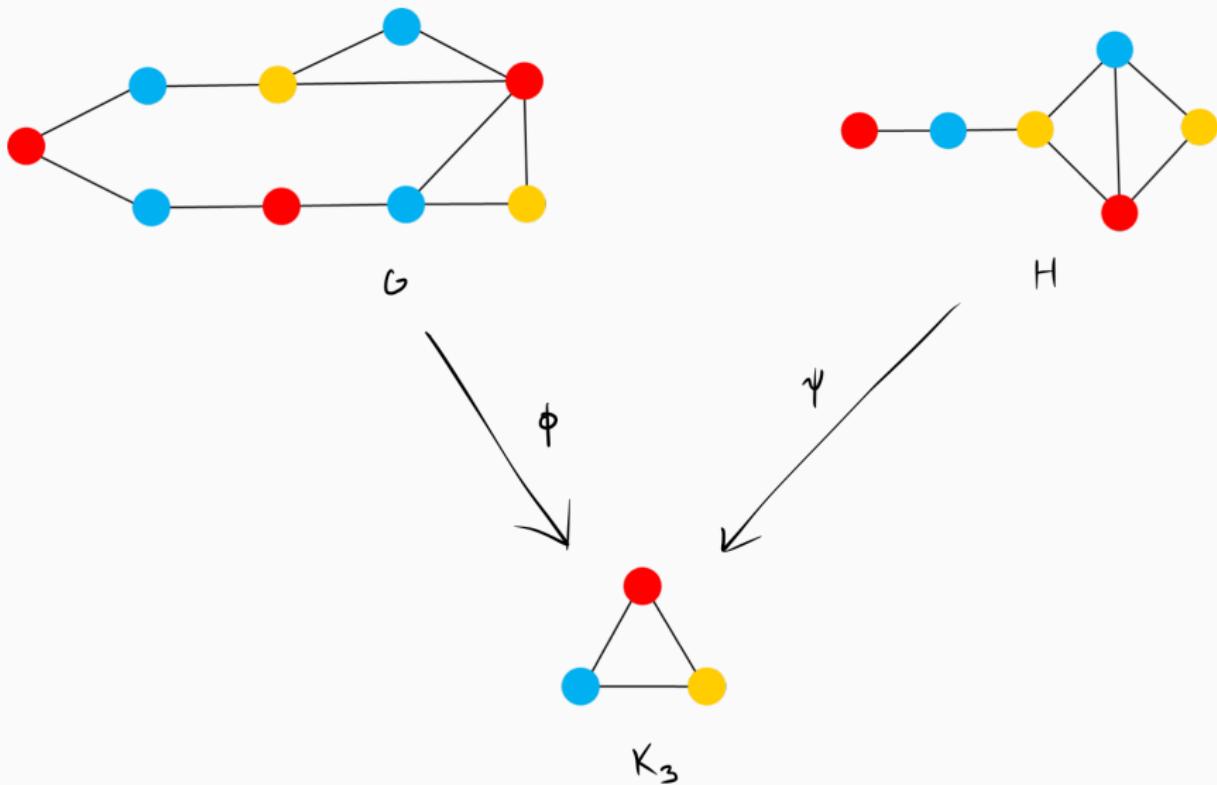
- A function ϕ between the vertices of two graphs G and H is called a **graph homomorphism** if $uv \in E(G)$ implies $\phi(uv) \in E(H)$.
- G and H are called **homomorphically equivalent** (hom-equivalent) if there exist graph homomorphisms $f : G \rightarrow H$ and $g : H \rightarrow G$.



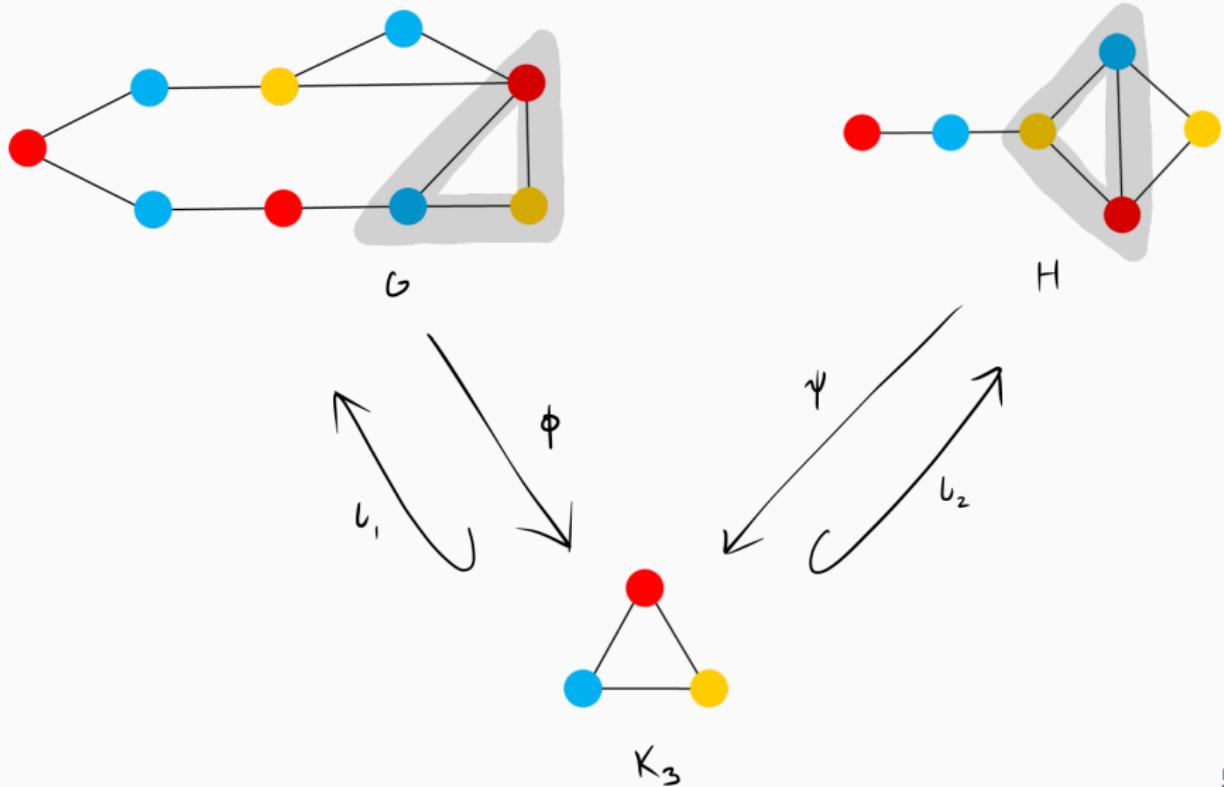
Why Might We Care? Possibility: Cores

- A **core** C of a graph G is a graph such that G and C are hom-equivalent, and C is the smallest such graph.
 - Complete graphs, odd cycles, etc
- Every finite graph has a core, and it is unique (up to isomorphism).
- Graphs with the same cores are necessarily hom-equivalent, and vice versa.
- Core-finding complexity: NP-complete :(
- Applications to relational algebra

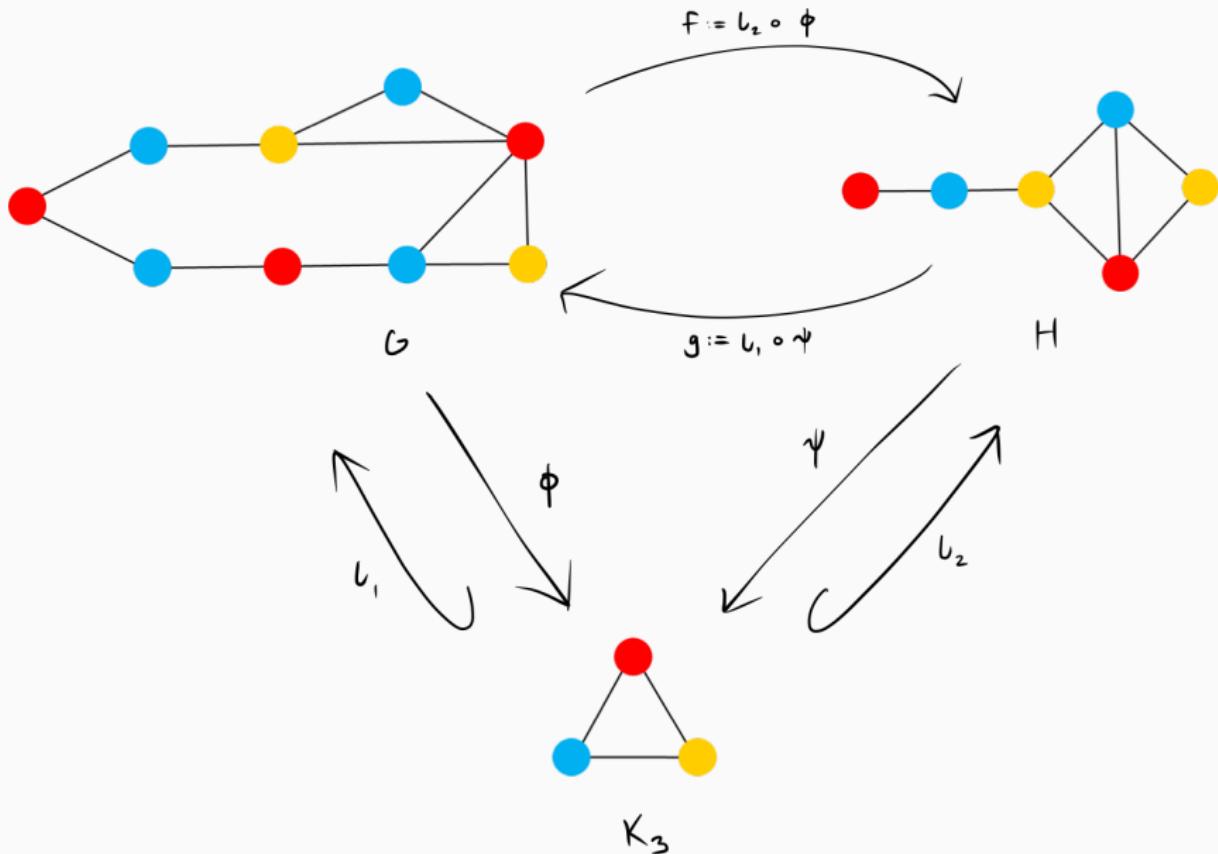
Example Equivalence via Core



Example Equivalence via Core



Example Equivalence via Core



Stability

Persistence

Return of Reeb

More Aptamers

- [1] F. Chazal et al. “**Persistence-based clustering in Riemannian manifolds.**”. In: (2013).
- [2] Paweł Dłotko. **Ball mapper: a shape summary for topological data analysis.** 2019. arXiv: 1901.07410 [math.AT]. URL: <https://arxiv.org/abs/1901.07410>.
- [3] Gurjeet Singh, Facundo Memoli, and Gunnar Carlsson. “**Topological Methods for the Analysis of High Dimensional Data Sets and 3D Object Recognition**”. In: *Eurographics Symposium on Point-Based Graphics*. Ed. by M. Botsch et al. The Eurographics Association, 2007. ISBN: 978-3-905673-51-7. DOI: /10.2312/SPBG/SPBG07/091-100.