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

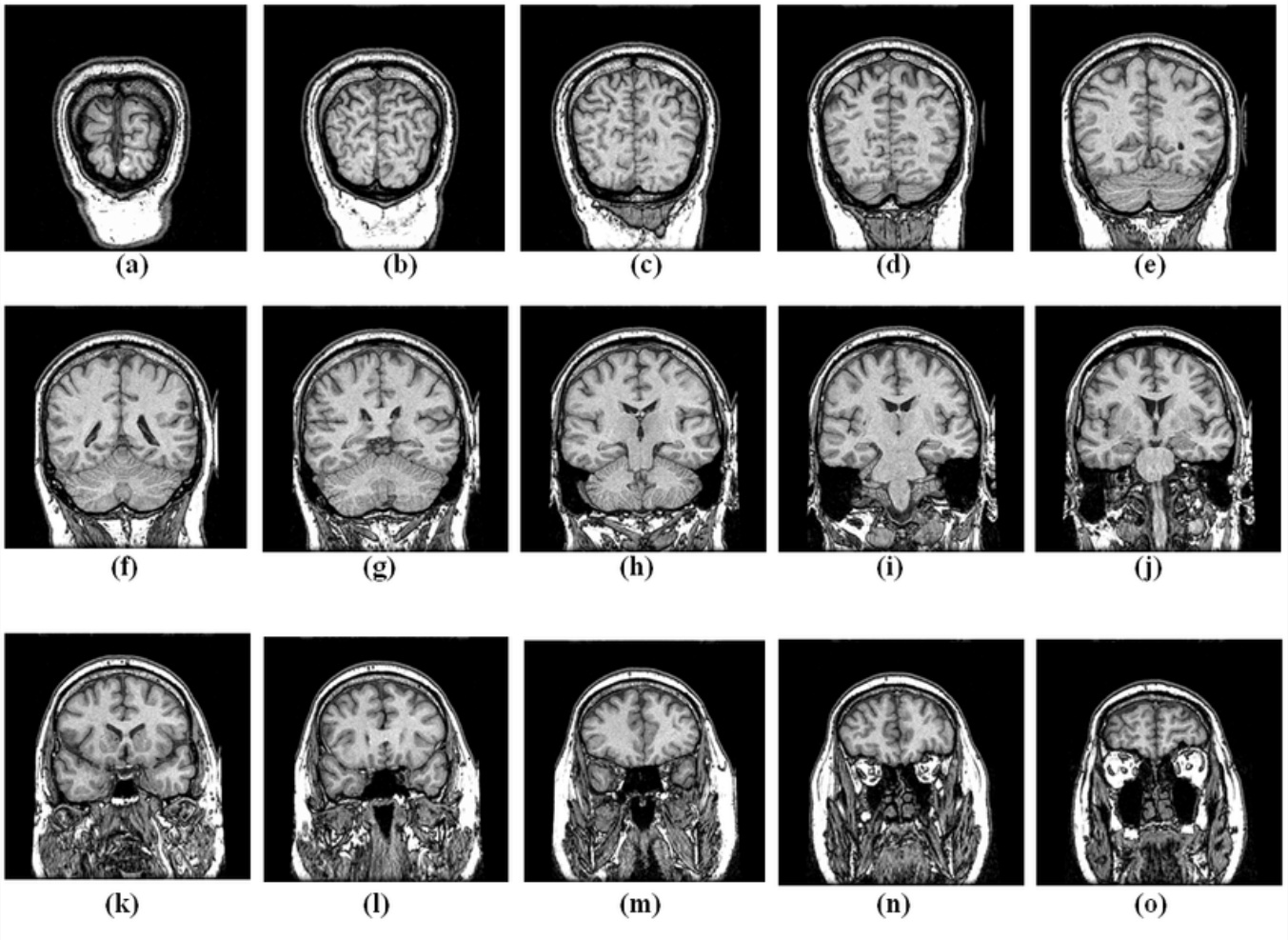
Aptamers

Mapping out Future Directions

Section Map (!)

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

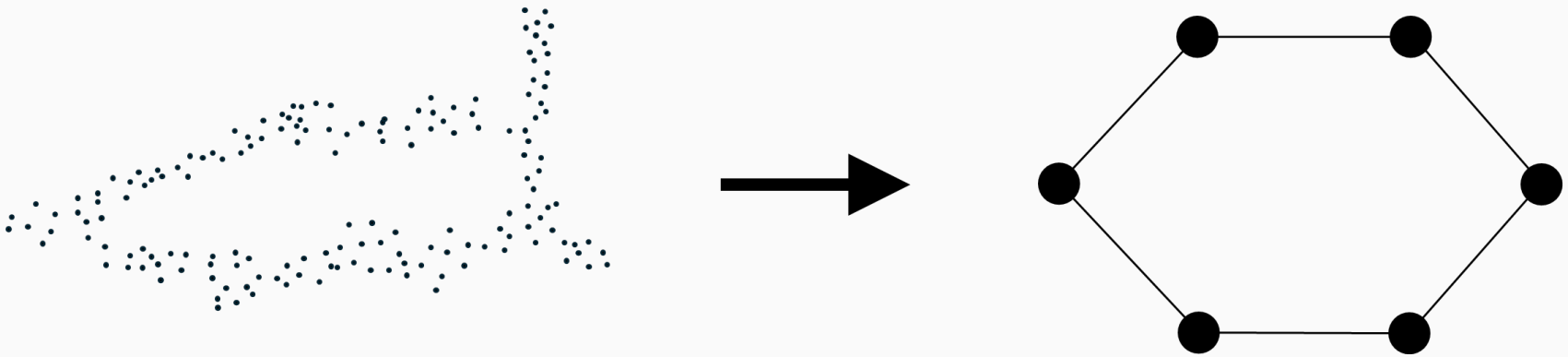
- High-dimensional space is extremely unintuitive
- If $V_n(r)$ is the volume of the n-dimensional ball with radius r, then for any $\varepsilon > 0$,

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

  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 $\Longrightarrow$ 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)



Section Map (!)

Why TDA?

Mapper and Its Flavors

Cytoscape to the Rescue

Aptamers

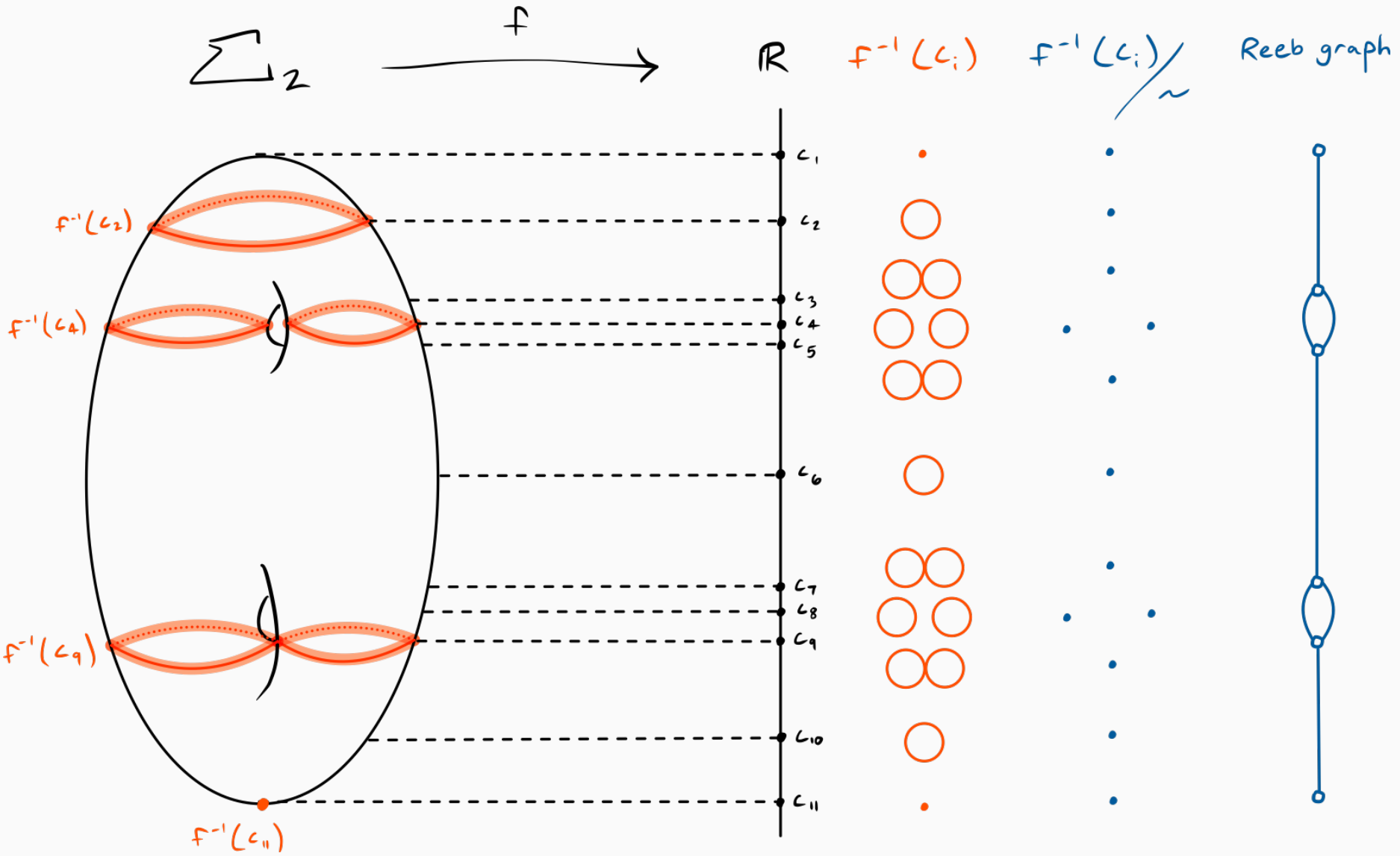
Mapping out Future Directions

Motivation: Reeb Graph

- Idea: construct graph reflecting level sets of a "filter" function
- Formally, given a topological space X and a continuous function $f : X \to \mathbb{R}$, 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.

[1] Despite names this is not always a graph

Motivation: Reeb Graph



Mapper: Original Flavor

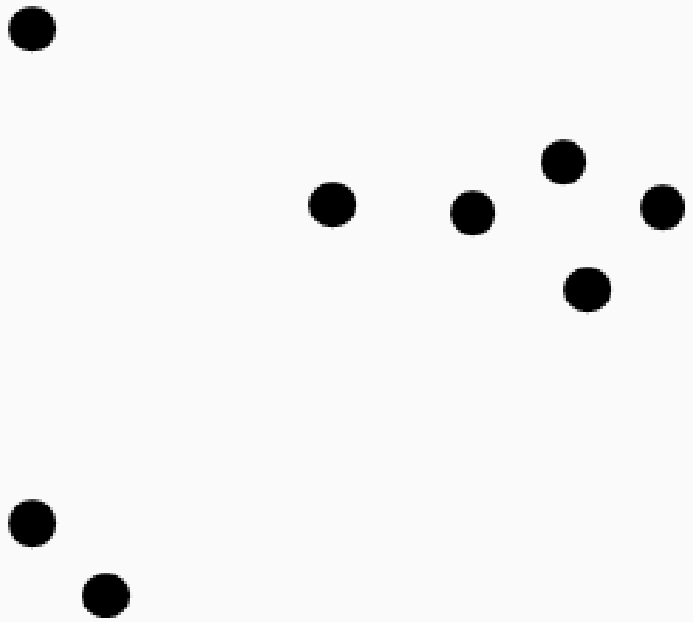
- How can we apply this to the discrete setting?
- Topological space $X \implies$ point cloud P (a discrete set of points in a space)
- Filter function: $f : P \implies \mathbb{R}$
- Level sets of points $\implies$ level sets of overlapping intervals
- Connected components $\implies$ clusters
- Quotient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

How to Cluster, Hierarchically

- Two ingredients:
  - A discrete set X of n equipped with a distance function $d : X \times X \to \mathbb{R}$
  - A **linking criterion**, which is a function $\ell : \mathscr{P}(X) \times \mathscr{P}(X) \to \mathbb{R}$ which measures distance between 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 can record the resulting hierarchy of clusters using a **dendrogram**, which records information about the merging process.

Example: Single Linkage Clustering

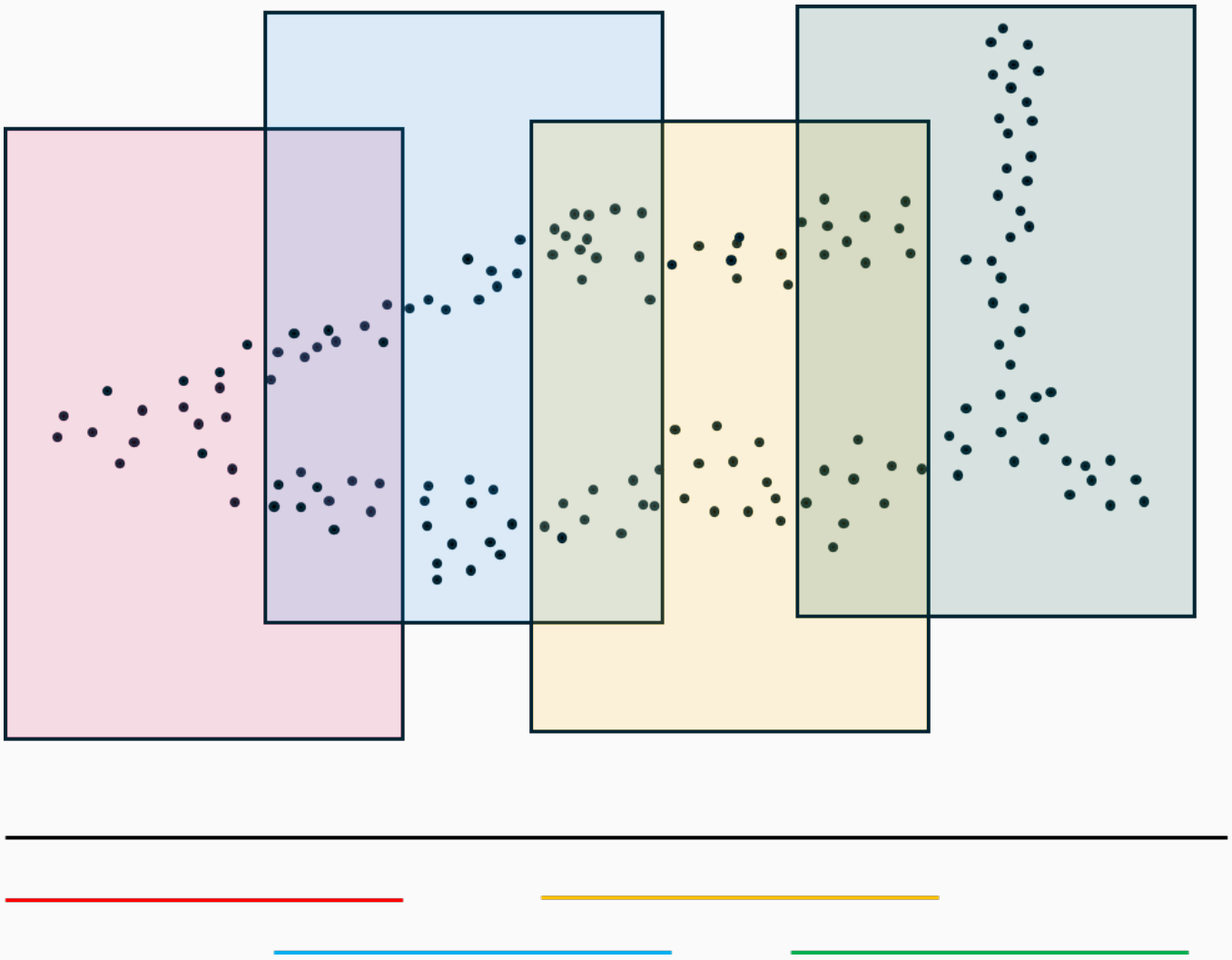
- Data: 8 points in $\mathbb{R}^2$, with the usual metric



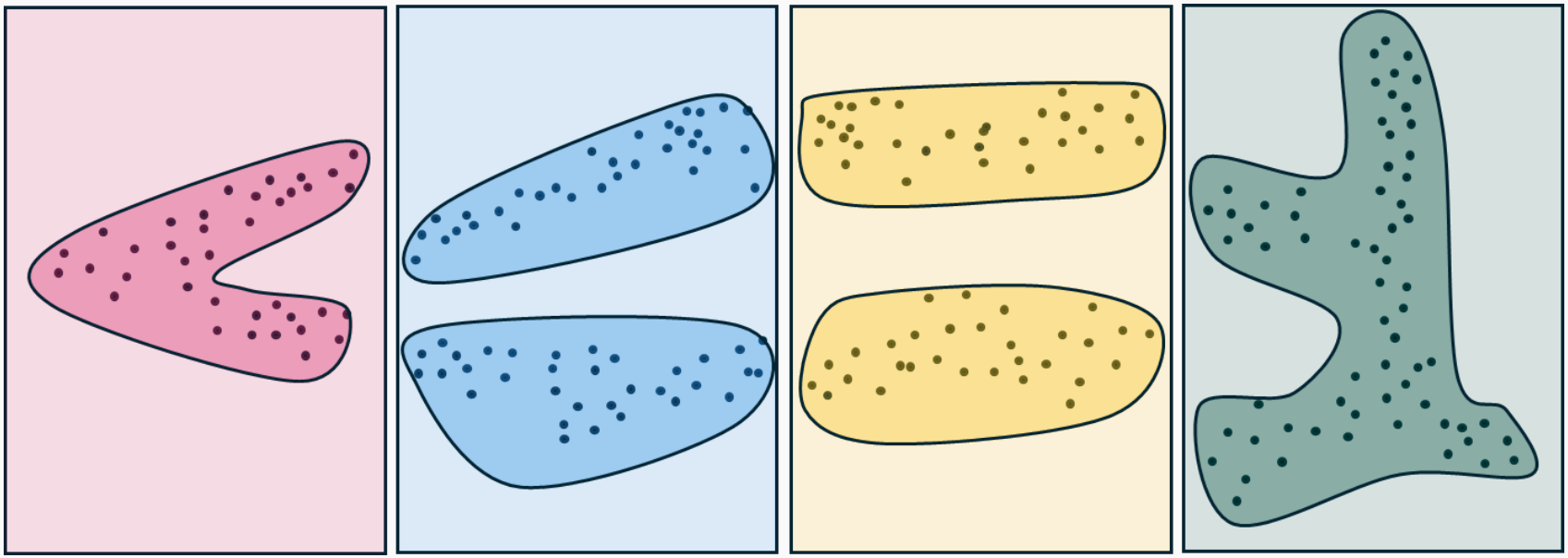
- Linking criterion:

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

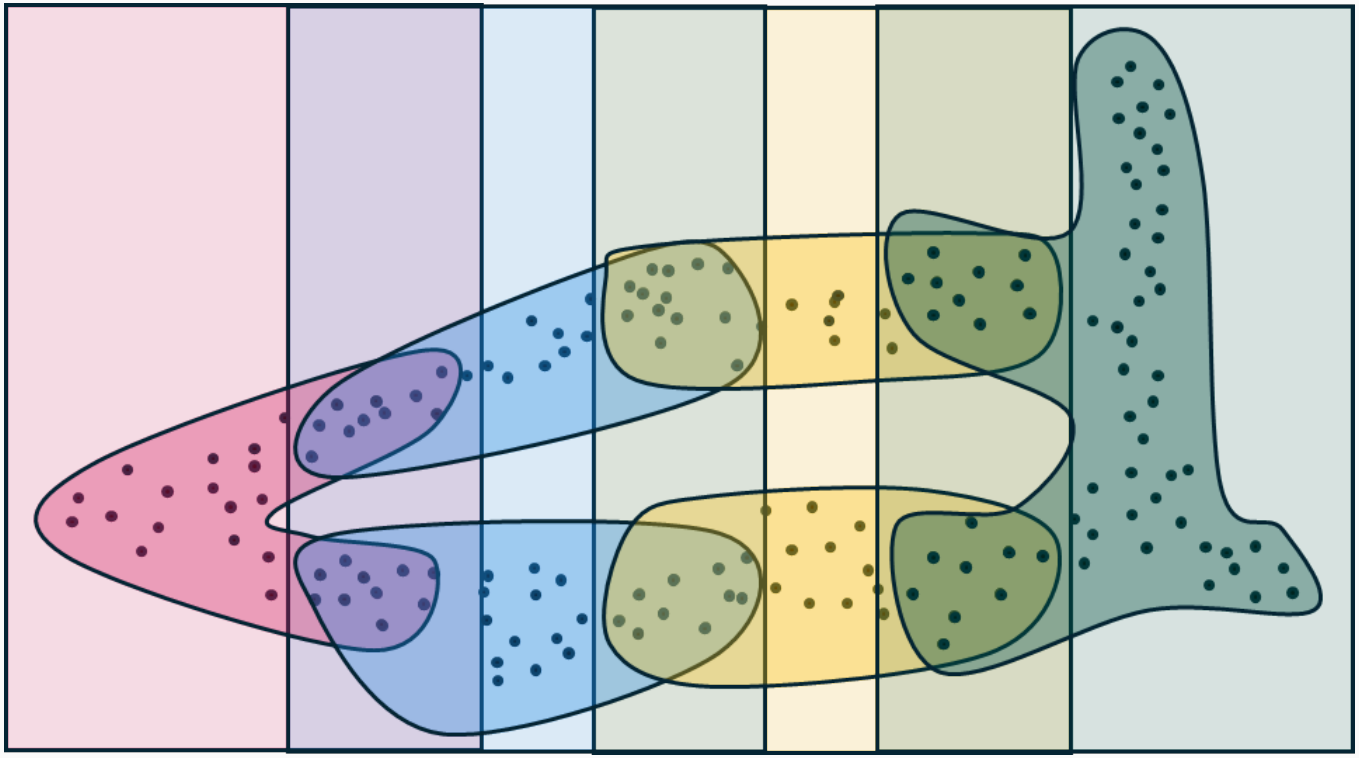
Filtering



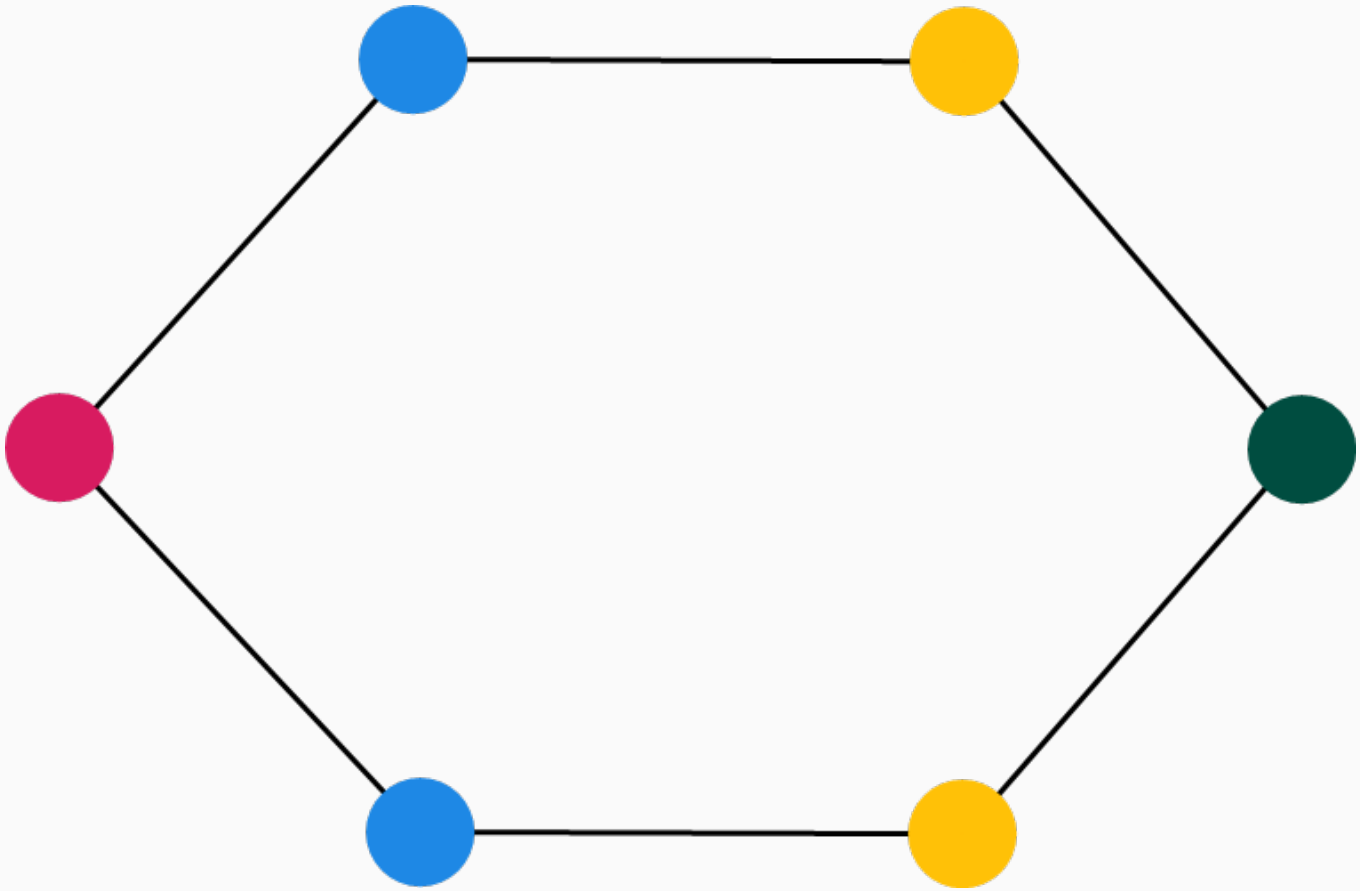
Clustering



Nerve Construction



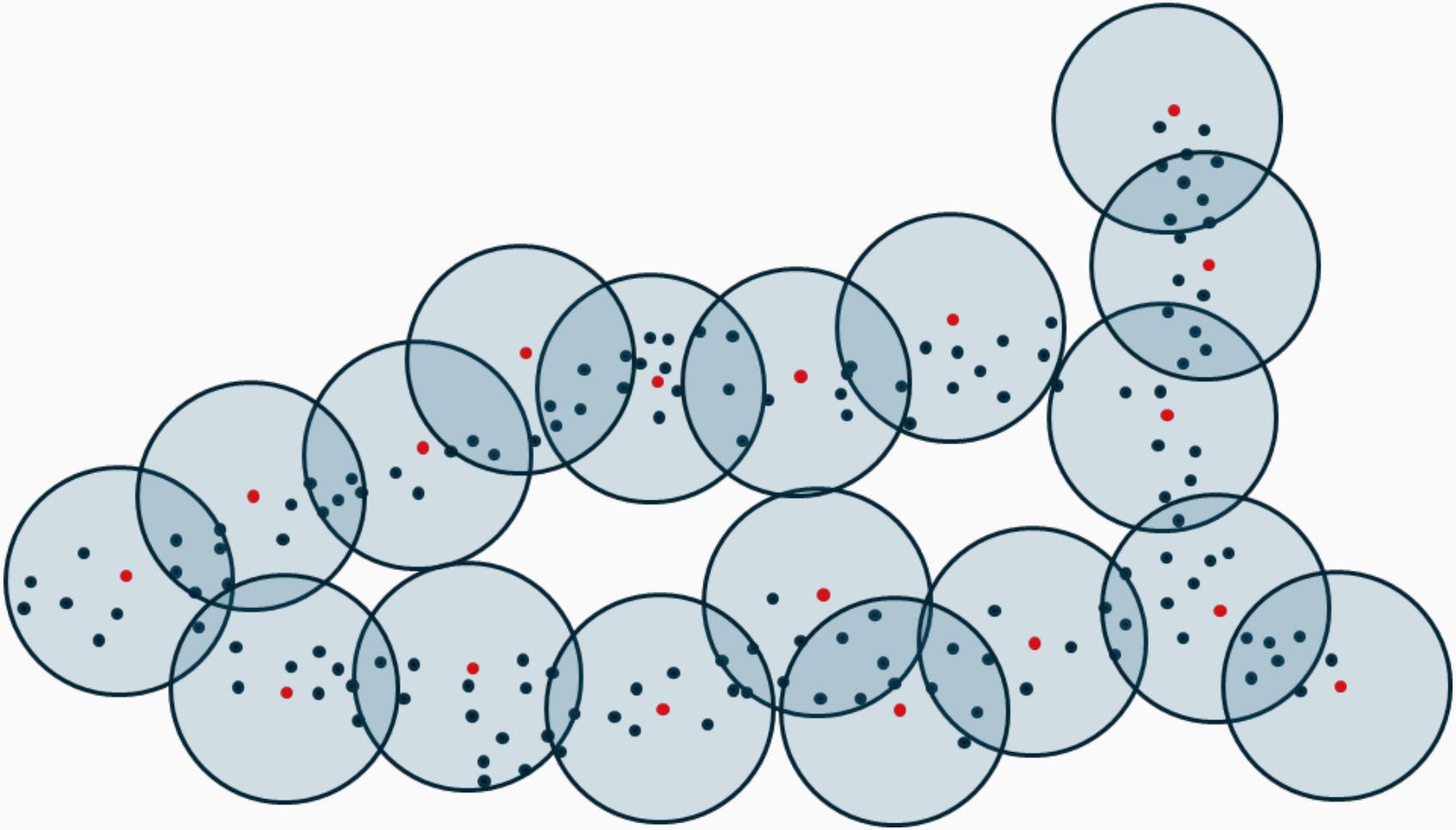
Output



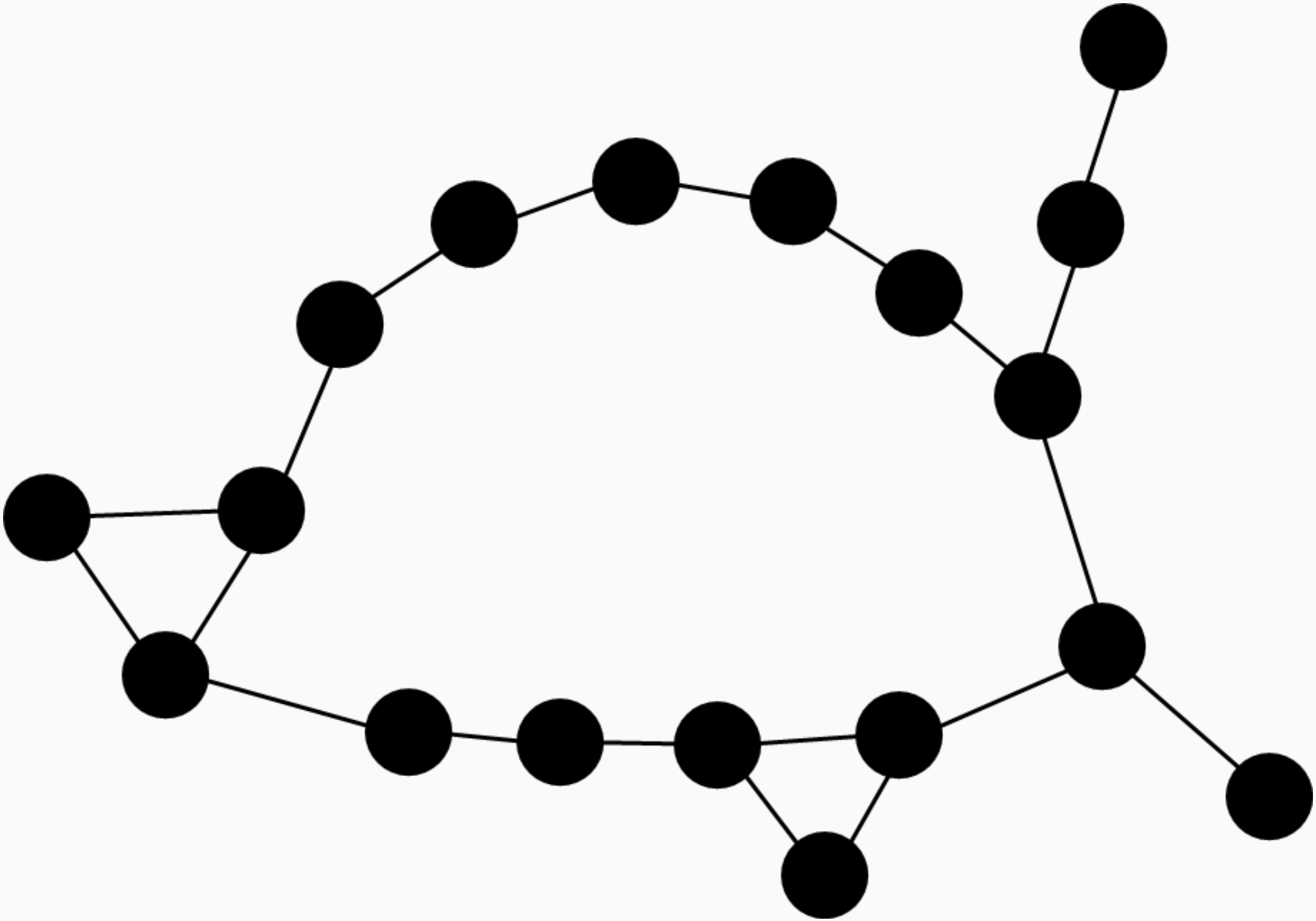
Ballmapper

- Original flavor has a lot of choices to make
- Idea: come up with a one-parameter Mapper
- Ballmapper: 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 some $B(x_i, \varepsilon)$ for some x_i
  - If x_j is a ball center, then the only ball containing it is $B(x_j, \varepsilon)$
- Nerve construction is the same

Ballmapper: Cover

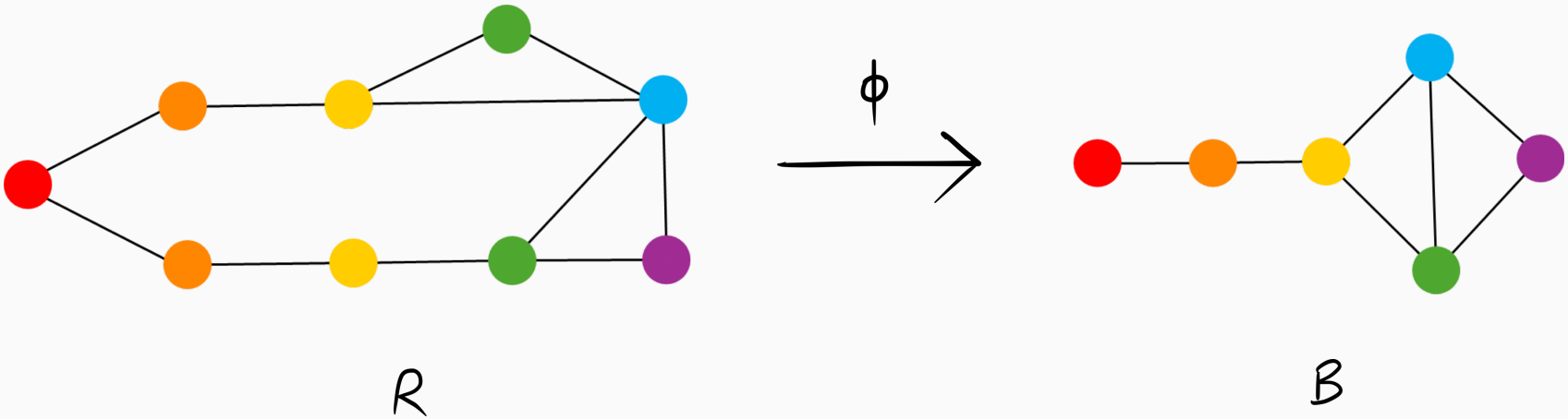


Ballmapper: Nerve



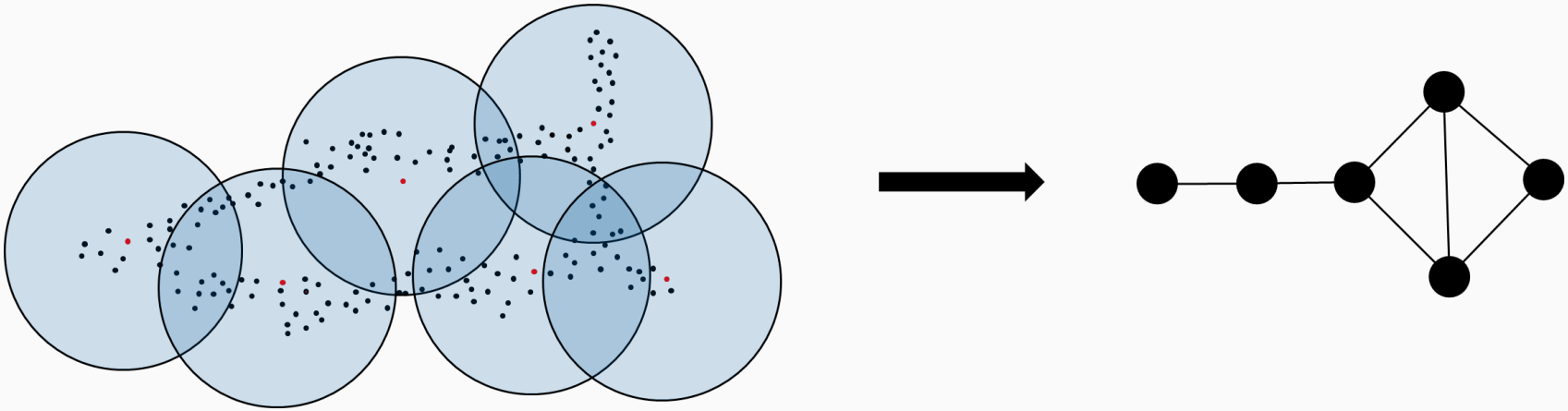
Refined Ballmapper

- Idea: combine Ball filtering and Original clustering
- Bin by balling, then cluster within balls as in Original
- Allows for comparison of two different metrics[2] at the same time on the same data

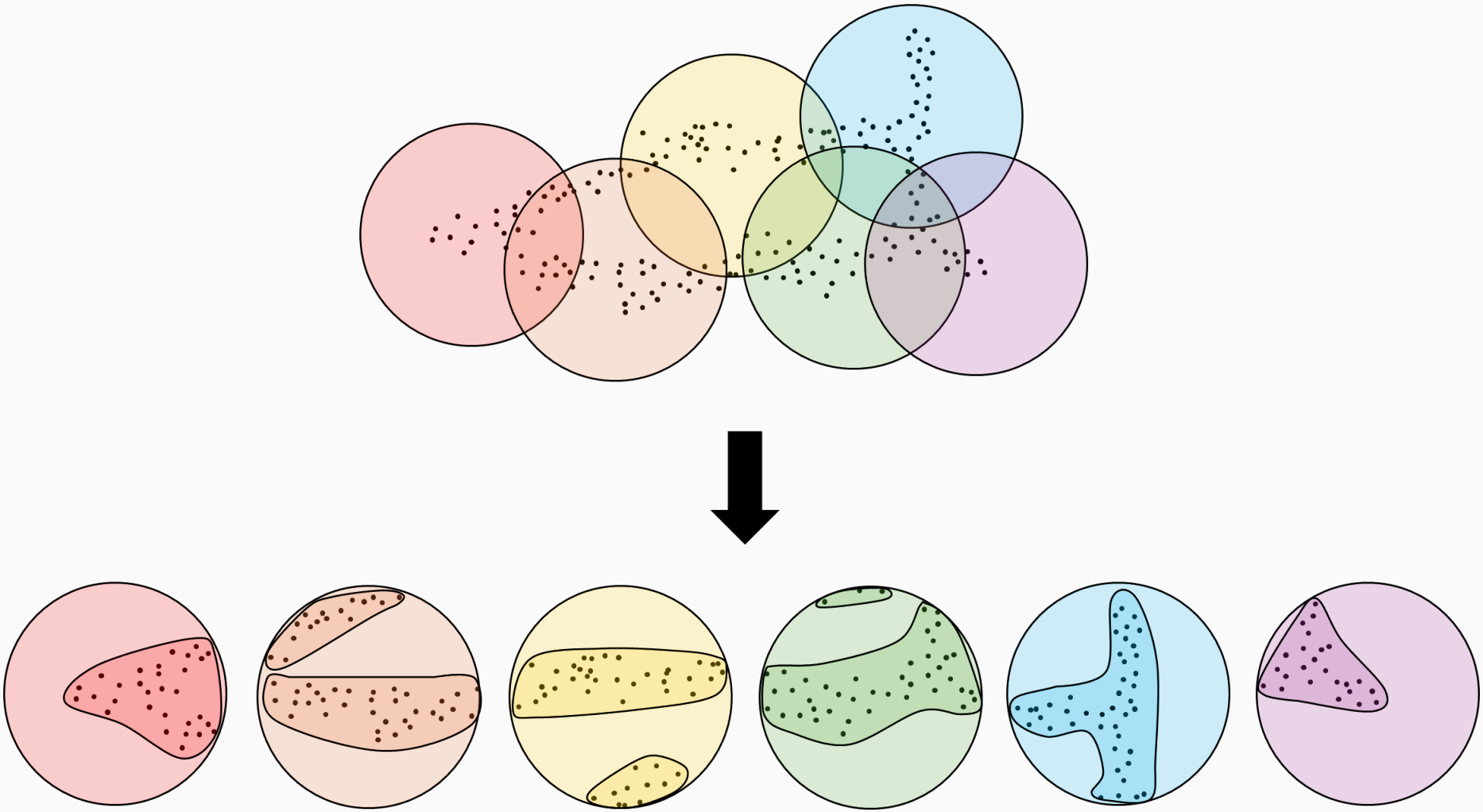


[2] Could be not strictly metrics!

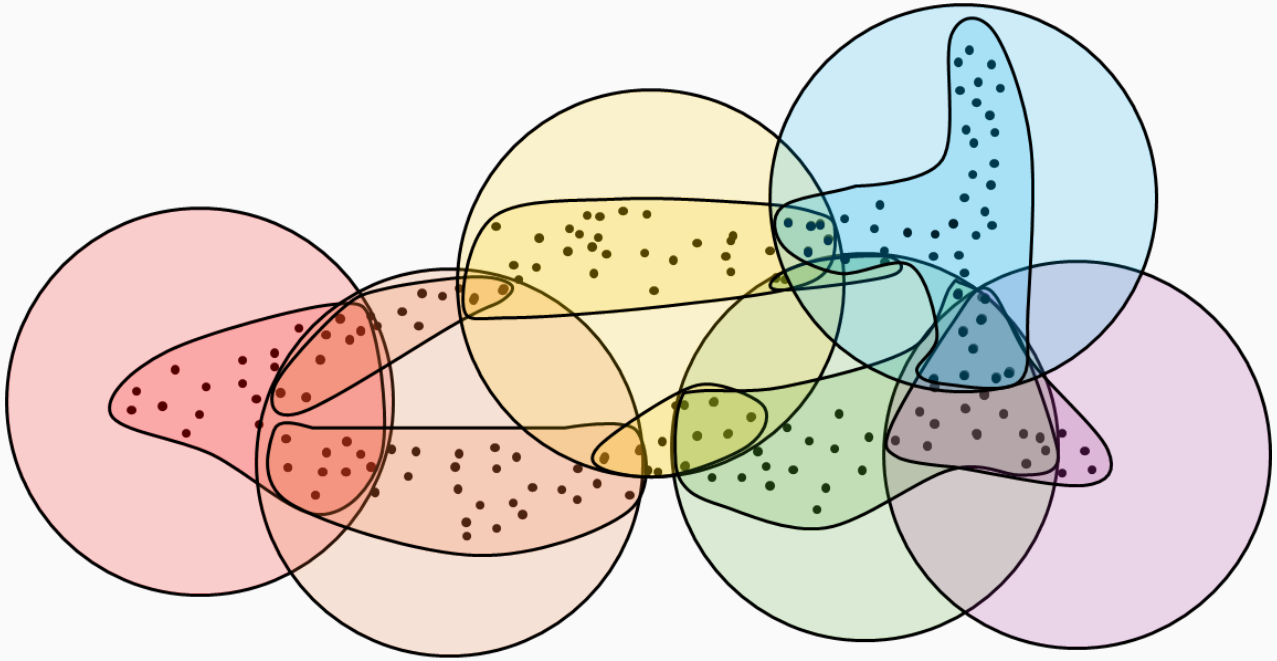
Standard Ballmapper

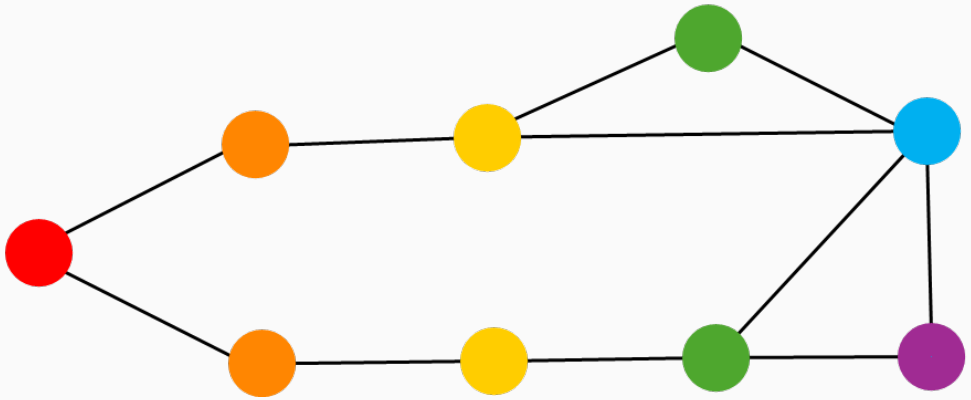


Balls as Bins



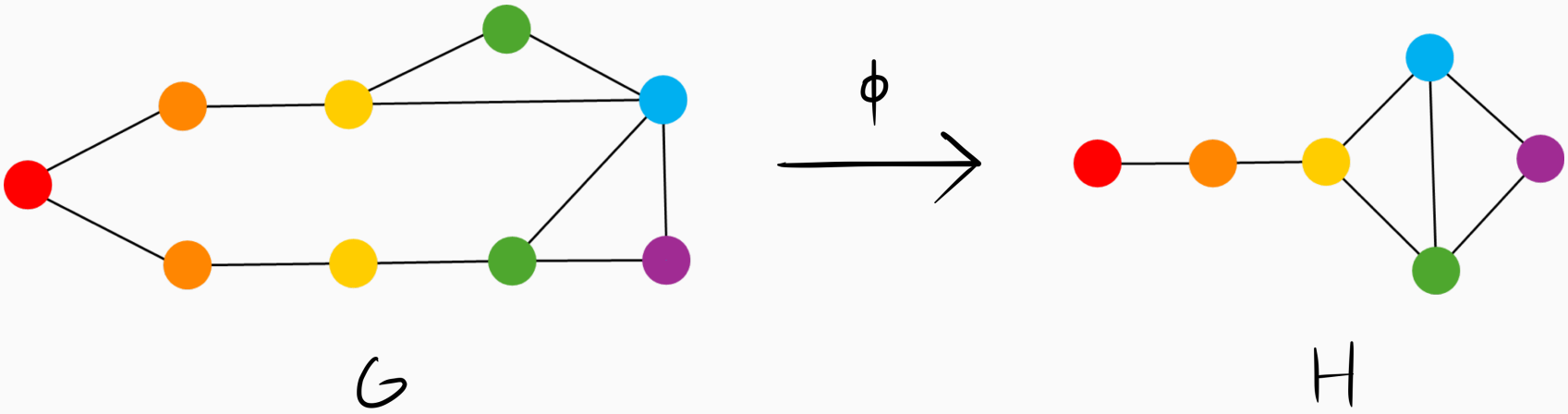
Refined Ballmapper Graph





Graph Theoretic Relation

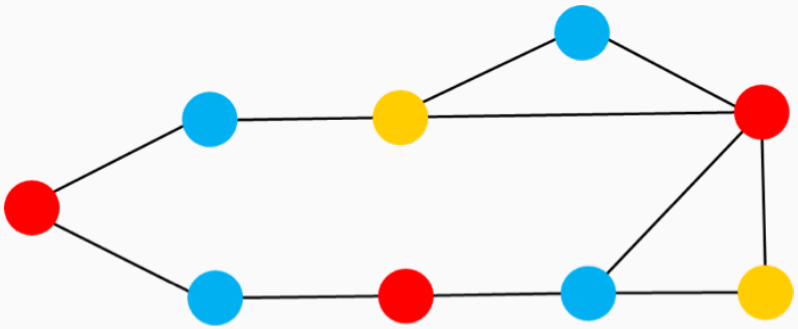
- A function ϕ between the vertices of two graphs G and H is called a **graph homomorphism** if $uv \in E(G)$ implies $\phi(uv) \in E(H)$.
- G and H are called **homomorphically equivalent** (hom-equivalent) if there exist graph homomorphisms $f : G \to H$ and $g : H \to G$.



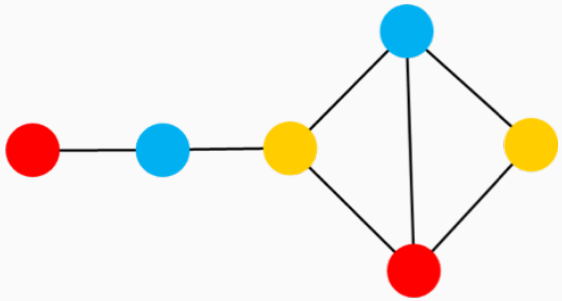
Why Might We Care? Possibility: Cores

- 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



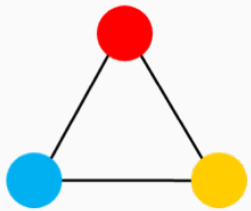
G



H

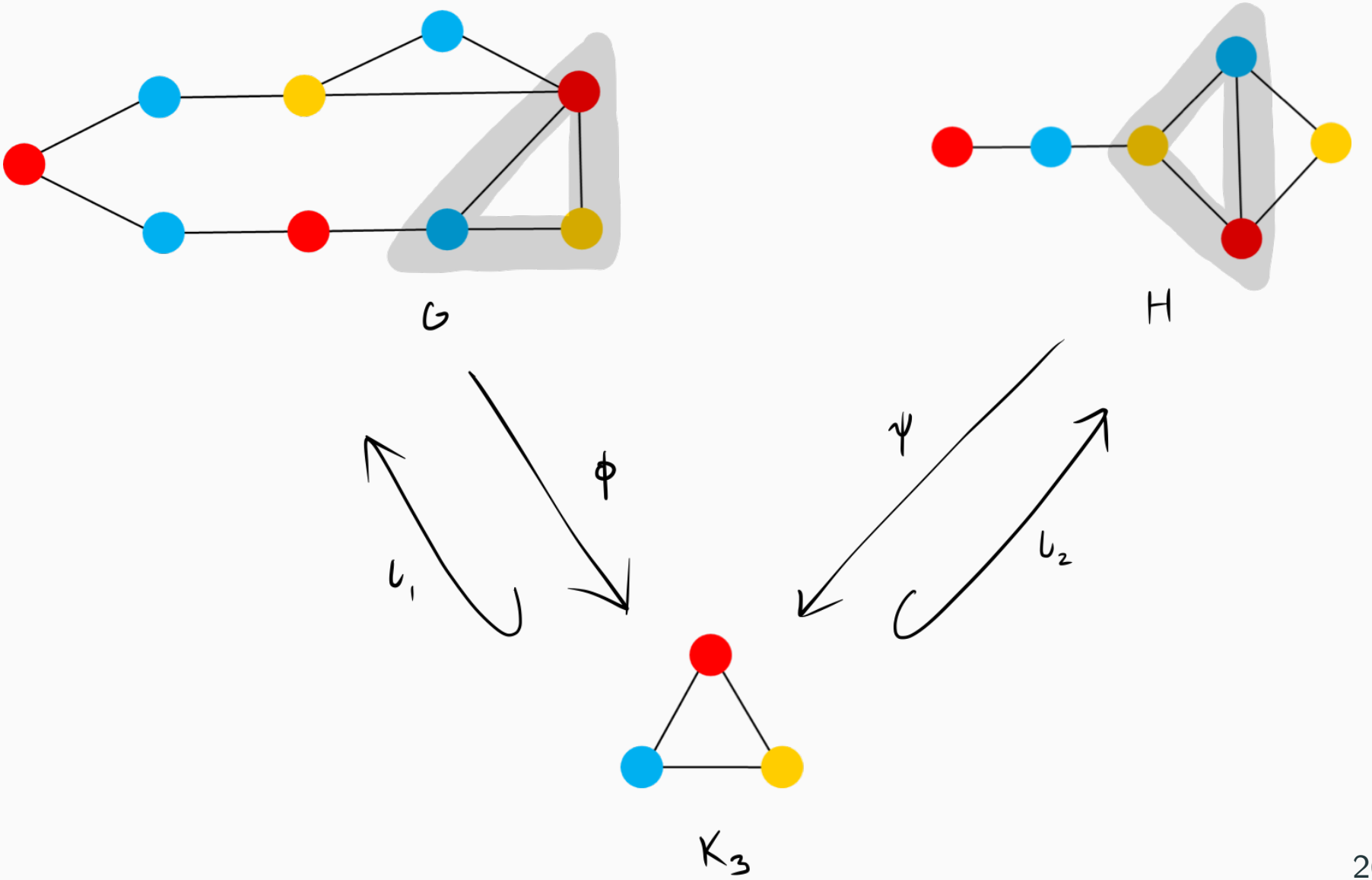
ϕ

ψ

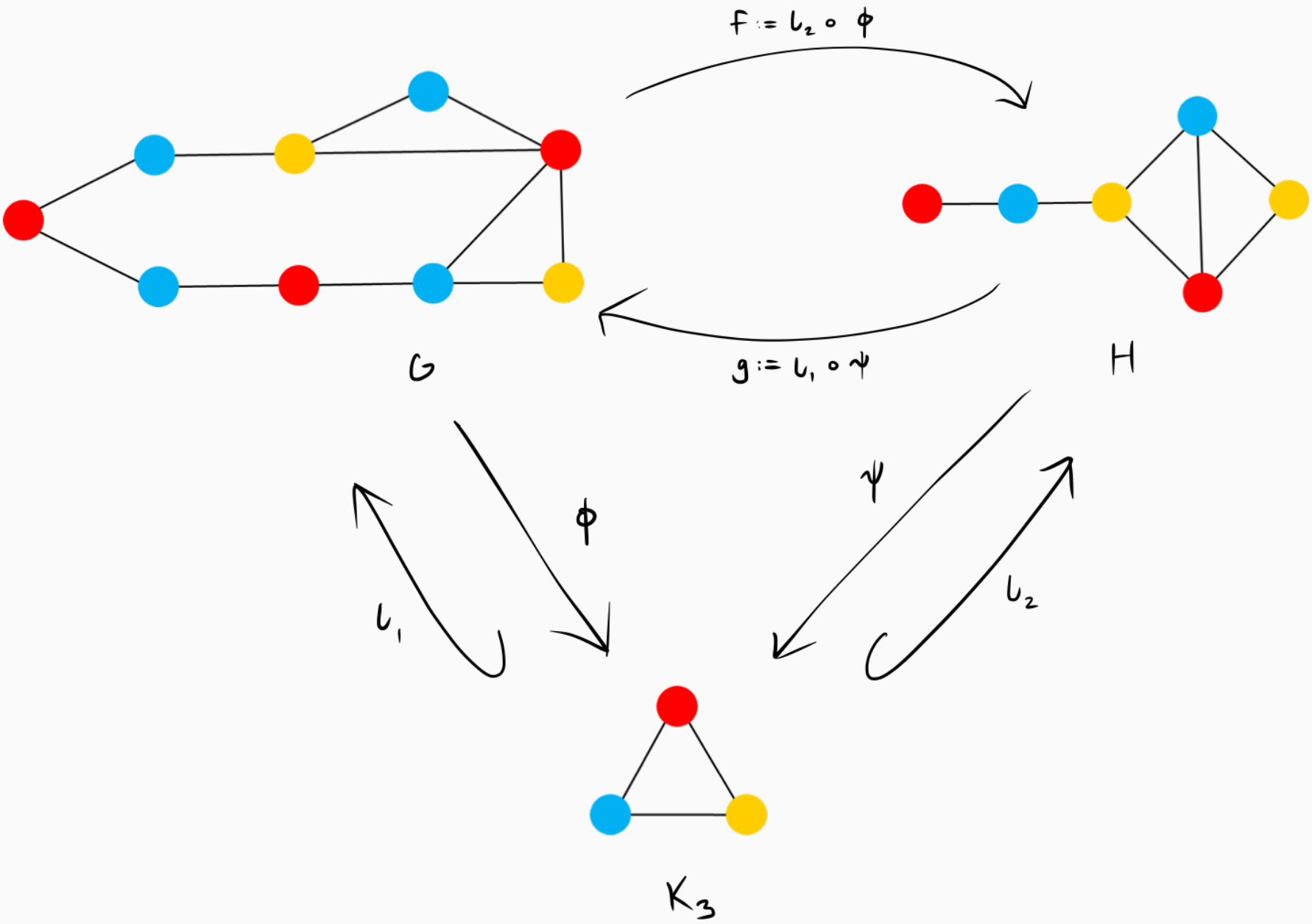


K_3

Example Equivalence via Core



Example Equivalence via Core



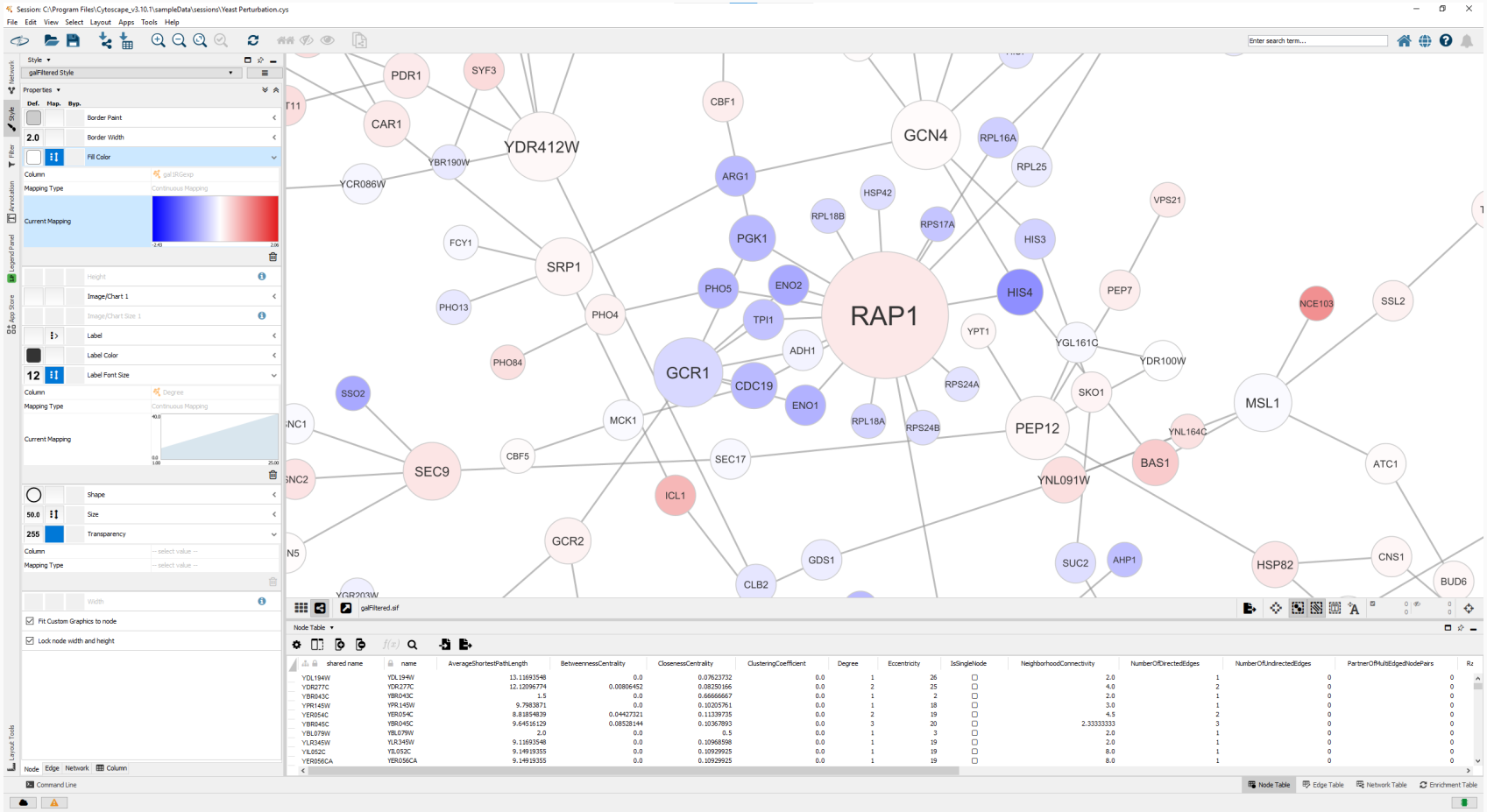
Problems

- We lost so much stuff! That clustering took work...
- Graphs are abstract combinatorial structures; they convey no geometrical information
- Potentially interesting features can “bypass” the filter
- Output heavily depends on your choices of parameters and clustering method

Section Map (!)

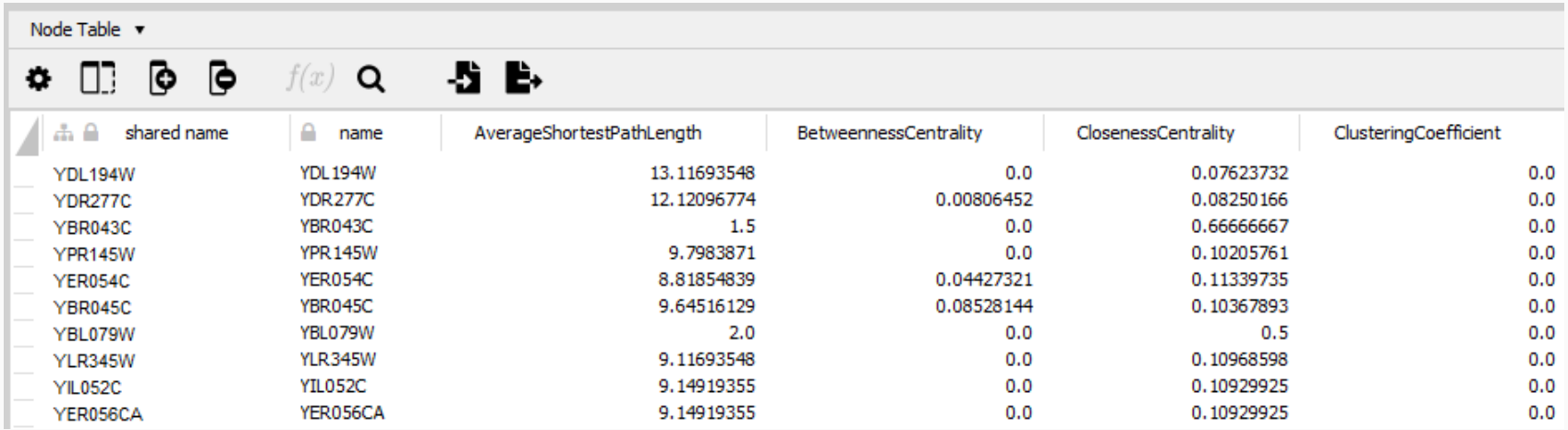
What is Cytoscape?

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



Features

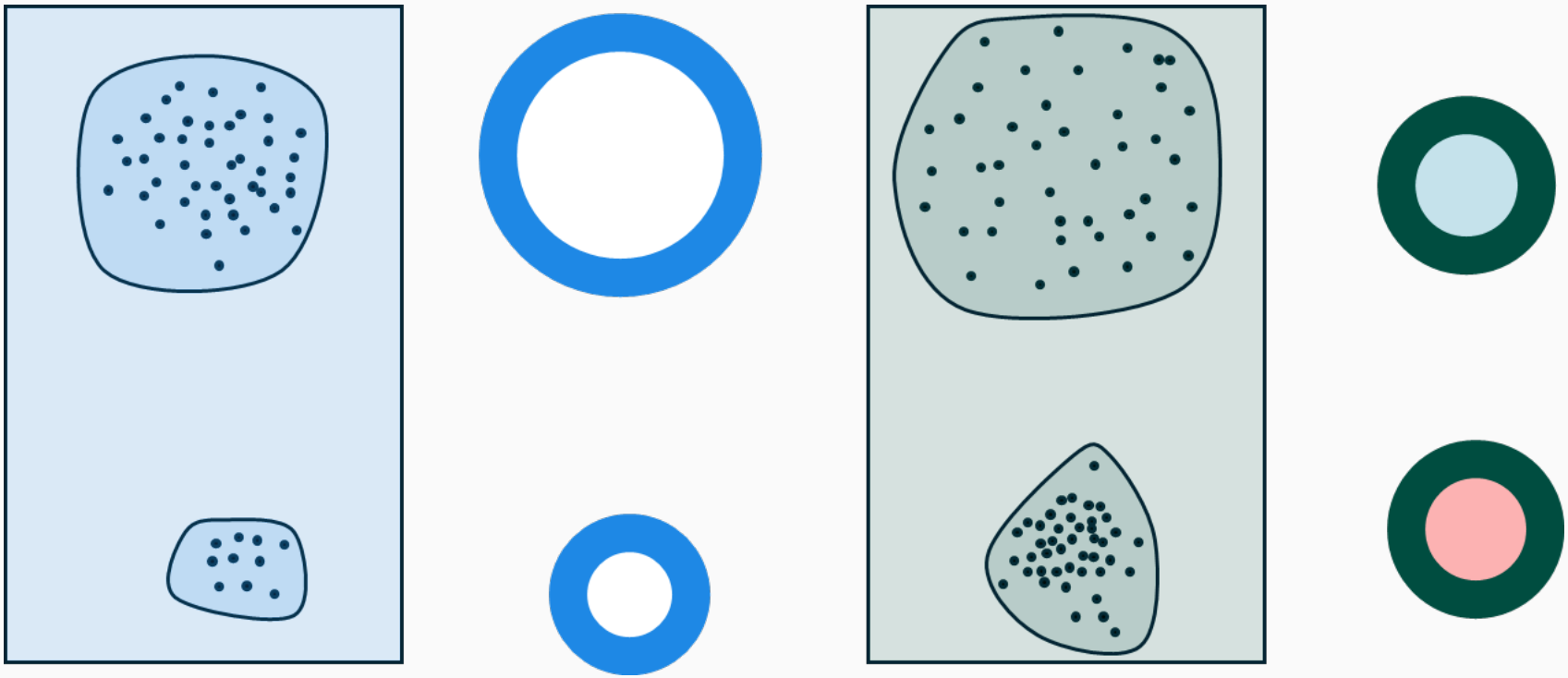
- Networks are stored as separate tables of nodes and edges
- Tables can be augmented with any number or type of columns
- Any column can then be associated with a visual characteristic of the network



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

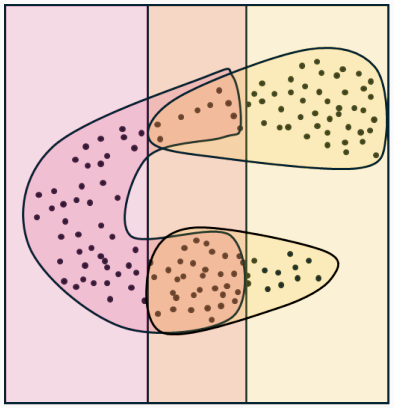
Styling Mapper: Vertices

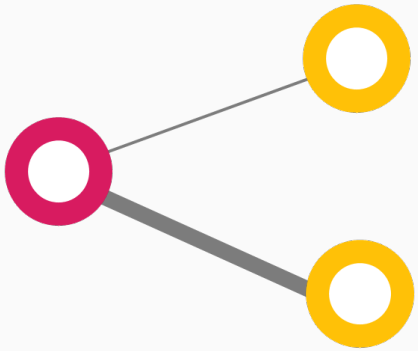
- Node size/label $\rightarrow$ cluster size
- Node border color $\rightarrow$ associated level set/filter value/ball
- Node fill color $\rightarrow$ cluster dispersion

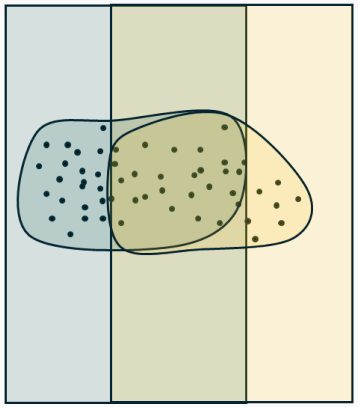


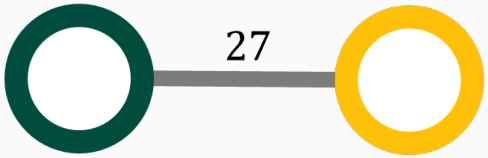
Styling Mapper: Edges

- Edge thickness/opacity $\rightarrow$ cluster intersection strength
- Edge label $\rightarrow$ cluster intersection size



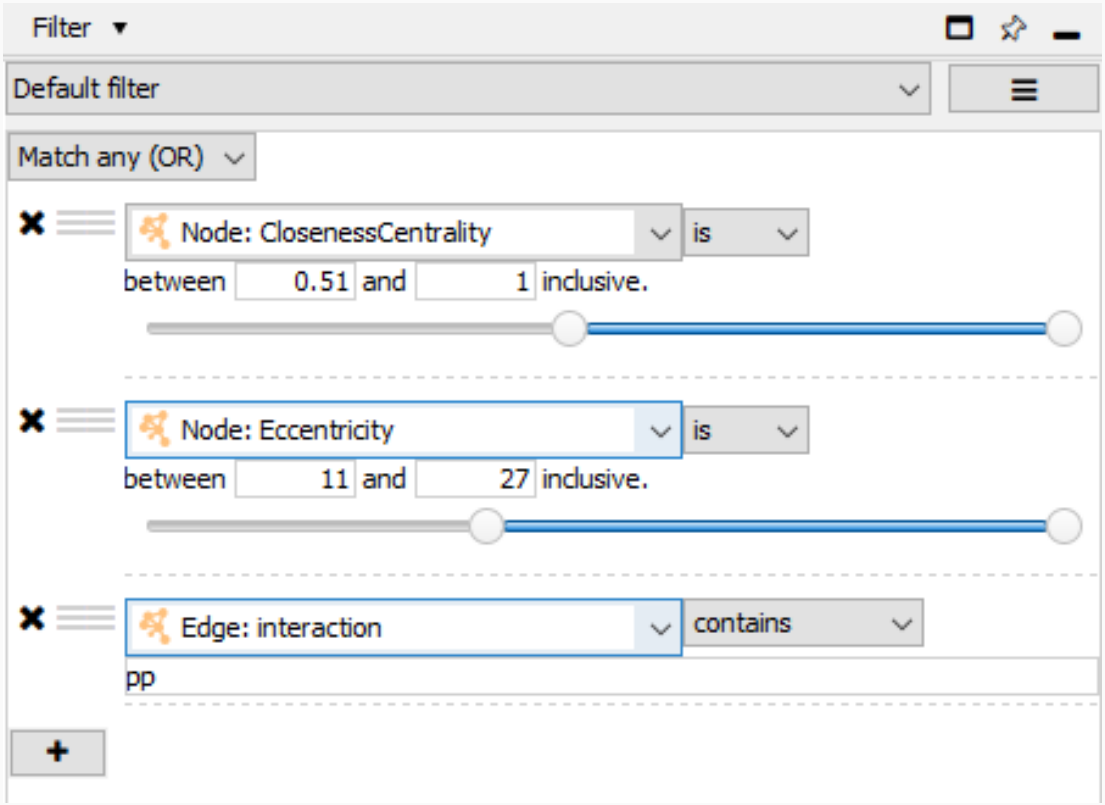






Exploring the Graph

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



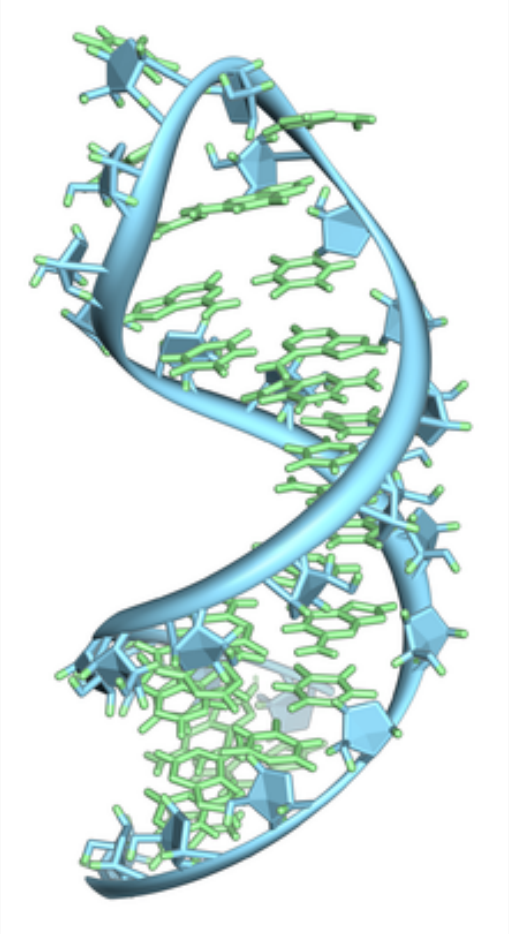
Possible Capabilities

- 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)

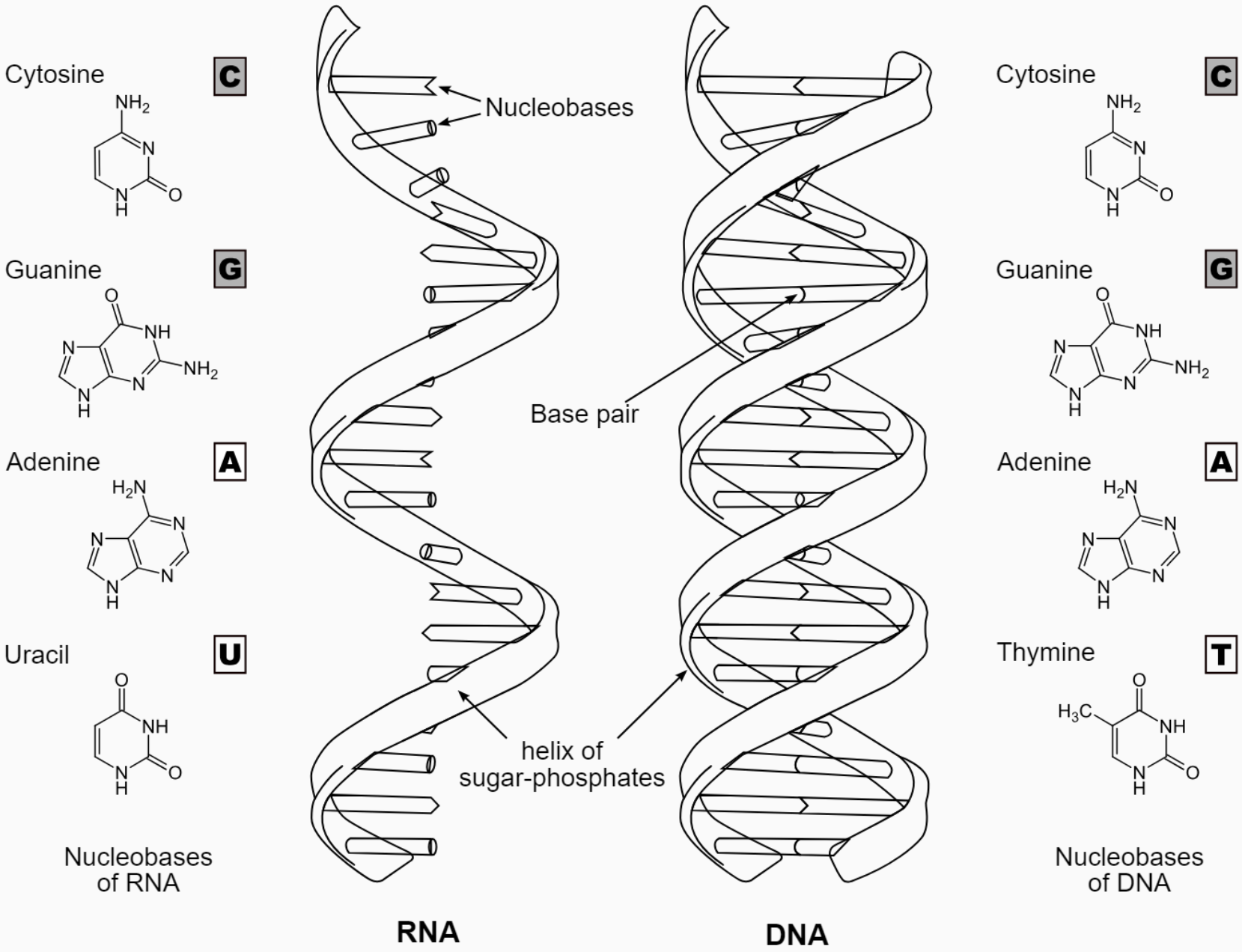
Section Map (!)

Biology CliffNotes CliffNotes

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

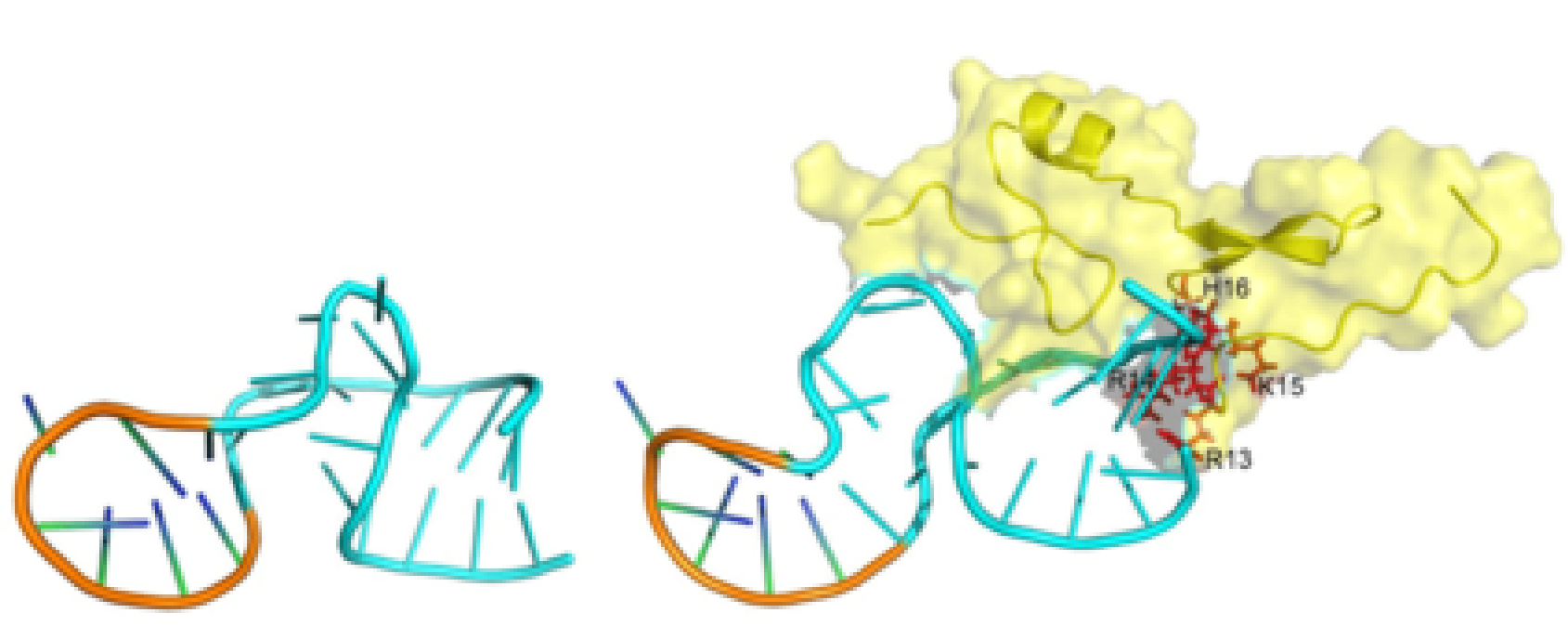


RNA and DNA


Cytosine
C
Guanine
G
Adenine
A
Uracil
U
Nucleobases of RNA
Nucleobases
Base pair
helix of sugar-phosphates
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**



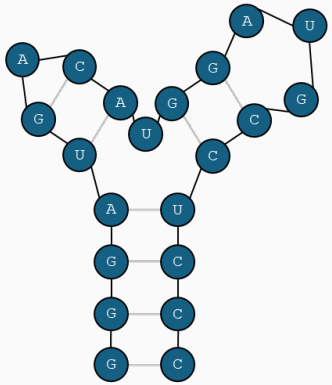
Comparing Aptamers

- For TDA to work we need a metric[3] by which to compare aptamers
- Aptamers have two characteristics: their genetic code and their structure
- Distance between sequences: Levenshtein distance
- Distance between structures: tree distance

[3] ehhhh...

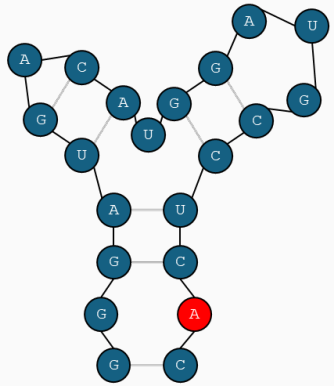
Aptamer Metrics

Aptamer X



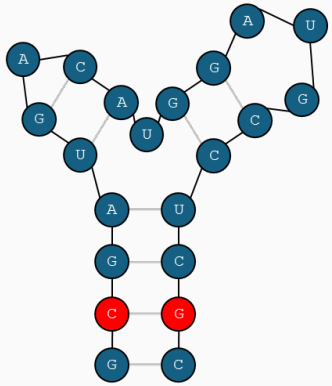
GGGAUGACAUGGAUGCCUCCC
((((((.)).((...))))))

Aptamer Y



GGGAUGACAUGGAUGCCUCAC
(.((((.)).((...)))).)

Aptamer Z



GGGCUGACAUGGAUGCCGCCC
((((((.)).((...))))))

δ_{Lev}: Levenshtein distance

δ_{Tree}: Tree distance

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

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

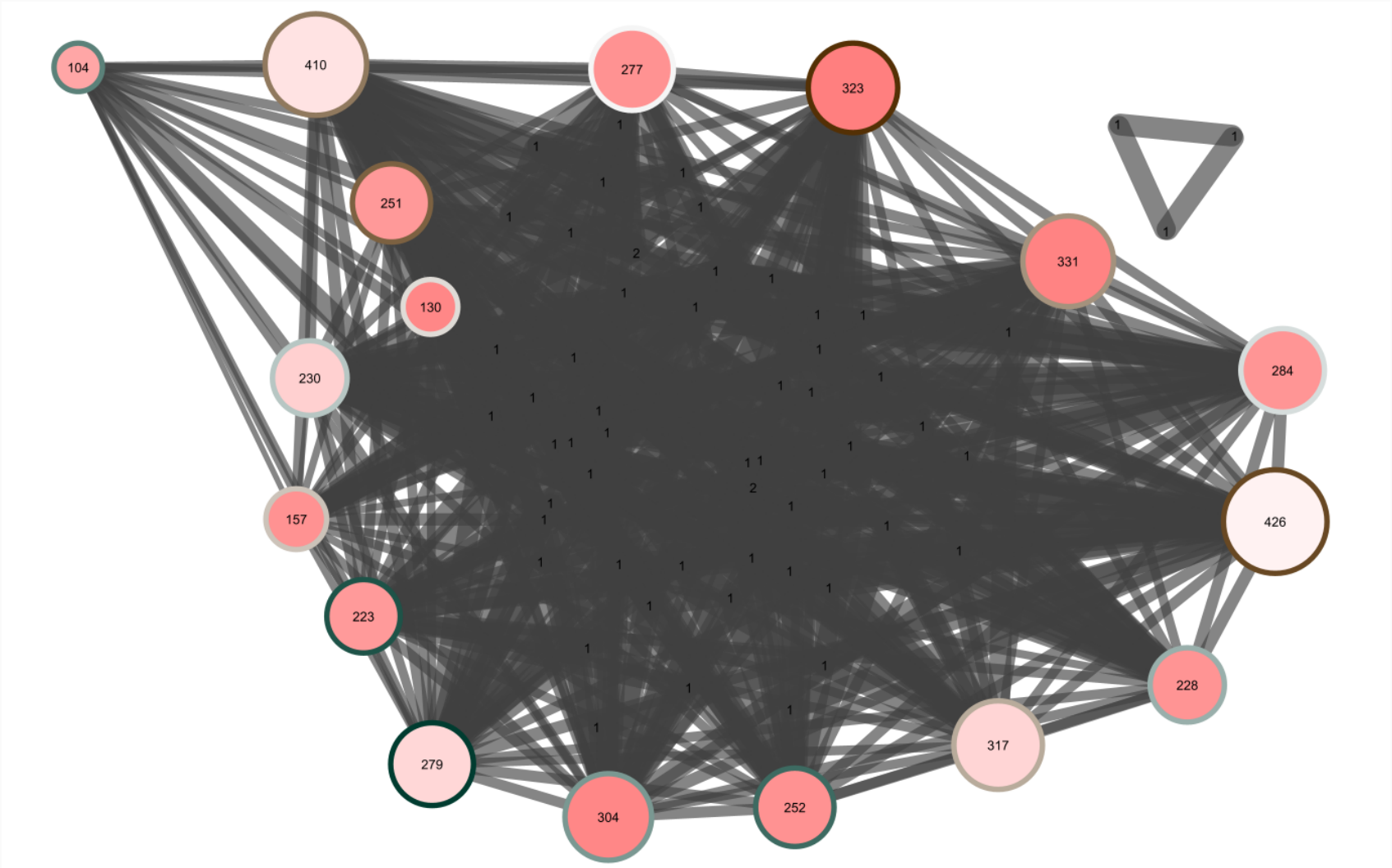
$\delta_{Lev}(X, Z) = 2$

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

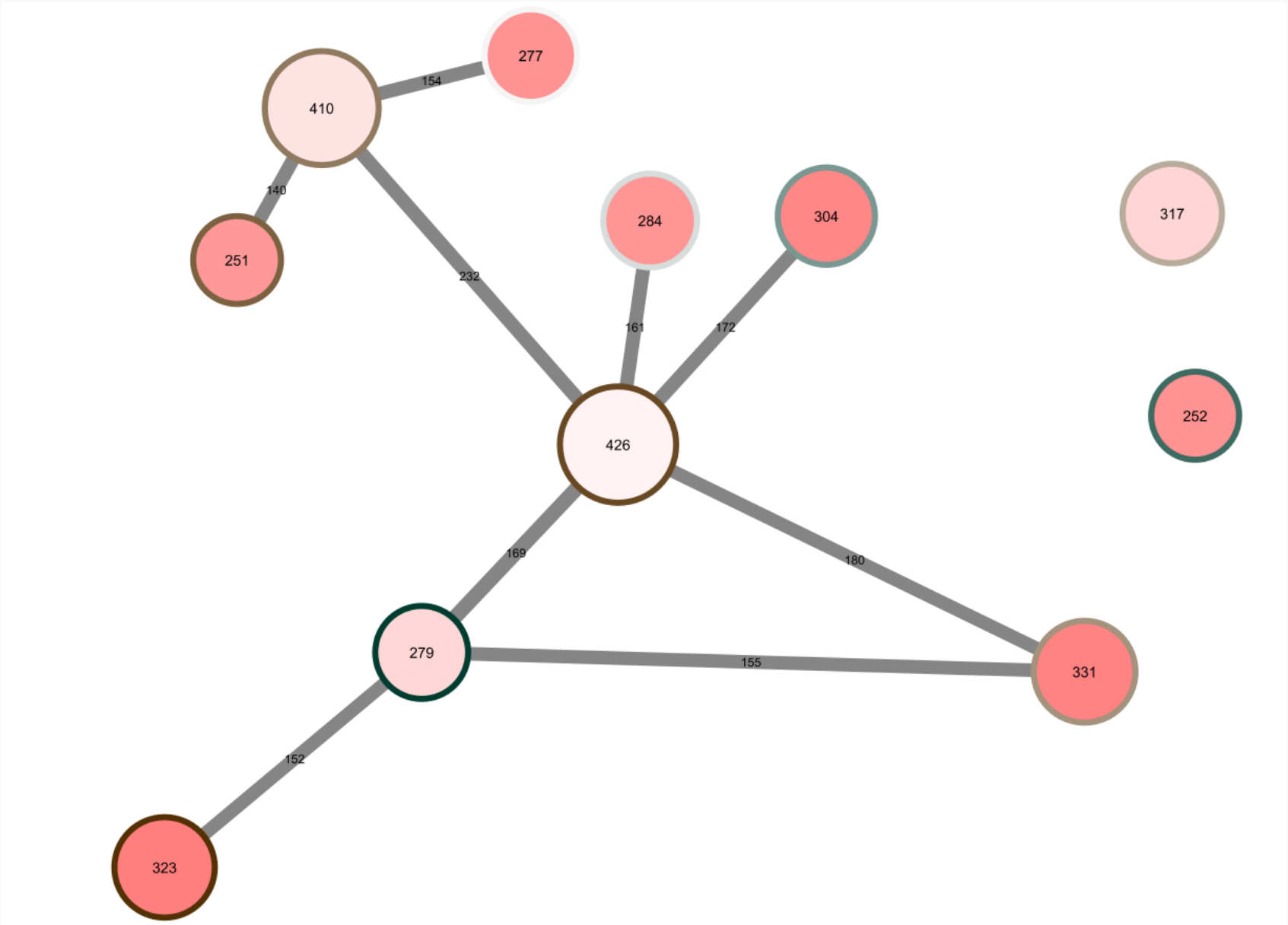
Aptamer Clustering With Mapper

- 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 Mapptamer Graph



Pruned Mapptamer Graph



Section Map (!)

Clustering

Stability

Persistence

Return of Reeb

More Aptamers

Neuroscience