Intro to Topological Data Analysis

Network and Homology

Chunyin Siu, Feb 21, 2025

Intro to Topological $\boxed{\text{Data Analysis}}$

Network and Homology

Chunyin Siu, Feb 21, 2025

What we learnt

- Beliefs
- Randomness
- A little dimension reduction

Question:
Which paradigm of statistics does dimension reduction belong to?

Dimension Reduction

Dimension Reduction

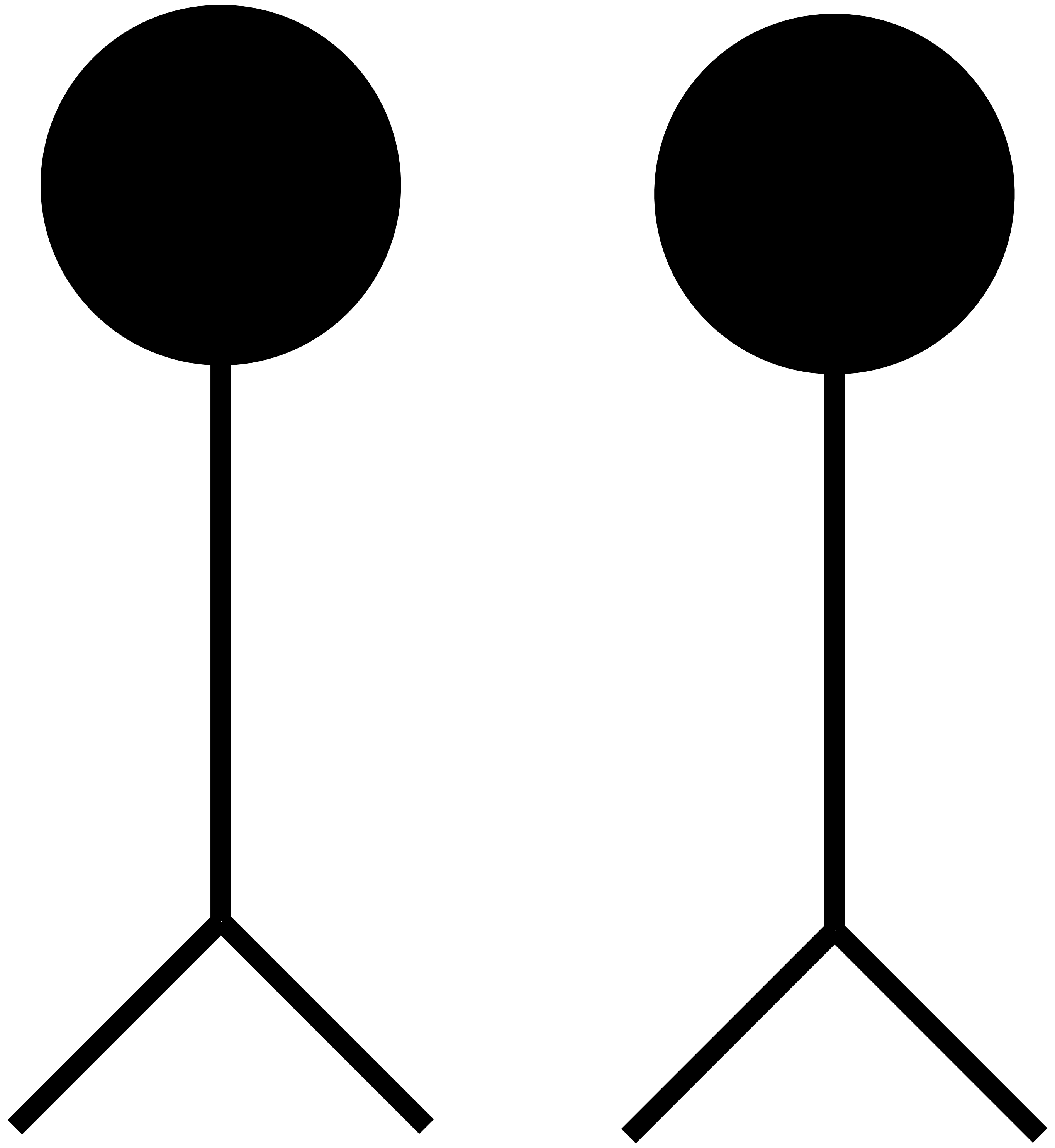
Enchanted, 2007

Dimension Reduction

Enchanted, 2007

Input, Output

Enchanted, 2007



Different Methods

Principal Component Analysis

Multi-Dimensional Scaling

UMAP

Mapper

Different Methods

Principal Component Analysis	Preserve linear information
Multi-Dimensional Scaling	
UMAP	
Mapper	

Different Methods

Principal Component Analysis	Preserve linear information
Multi-Dimensional Scaling	Preserve distances as far as possible
UMAP	
Mapper	

Different Methods

Principal Component Analysis	Preserve linear information
Multi-Dimensional Scaling	Preserve distances as far as possible
UMAP	Preserve local local distances as far as possible
Mapper	

Different Methods

Principal Component Analysis		Preserve linear information
Multi-Dimensional Scaling	Geometric Methods	Preserve distances as far as possible
UMAP		Preserve local local distances as far as possible
Mapper		

Different Methods

Principal Component Analysis		Preserve linear information
Multi-Dimensional Scaling	Geometric Methods	Preserve distances as far as possible
UMAP		Preserve local local distances as far as possible
Mapper	Topological Methods	

Different Methods

Principal Component Analysis		Preserve linear information
Multi-Dimensional Scaling	Geometric Methods	Preserve distances as far as possible
UMAP		Preserve local local distances as far as possible
Mapper	Topological Methods	Preserve connectivity information in a specific sense

As good as possible…

As good as possible…

- Nearby points in high-dim remains nearby in low-dim.
- Reverse?

As good as possible…

- Nearby points in high-dim remains nearby in low-dim.
- Reverse?

Cinderella III: A Twist in Time (2017)

As good as possible…

- Nearby points in high-dim remains nearby in low-dim.
- Reverse needs cheating

Alice in the Wonderland (1951)

Mapper

- Nearby points in high-dim remains nearby in low-dim.
- Reverse guaranteed by using graph topology to transcend Euclidean geometry

Alice in the Wonderland (1951)

Mapper

- Nearby points in high-dim remains nearby in low-dim.
- Reverse guaranteed by using graph topology to transcend Euclidean geometry
- Price: loss of metric control

Alice in the Wonderland (1951)

Mapper

- Nearby points in high-dim remains nearby in low-dim.
- Reverse guaranteed by using graph topology to transcend Euclidean geometry
- Price: loss of metric control
- Get to keep: connectivity

Alice in the Wonderland (1951)

What is Mapper?

What is Mapper?

Input		
Data points		

What is Mapper?

Input		
Data points Metric		

What is Mapper?

Input		
Data points Metric Coloring scheme (optional)		

What is Mapper?

Input		Output
Data points Metric Coloring scheme (optional)		A (colored) graph, where

What is Mapper?

Input		Output
Data points Metric Coloring scheme (optional)		A (colored) graph, where nodes are clusters of data points

What is Mapper?

Input		Output
Data points Metric Coloring scheme (optional)		A (colored) graph, where nodes are clusters of data points edges are nonempty intersection of clusters

What is Mapper?

Input	Parameters	Output
Data points Metric Coloring scheme (optional)	Discretization parameters	A (colored) graph, where nodes are clusters of data points edges are nonempty intersection of clusters

What is Mapper?

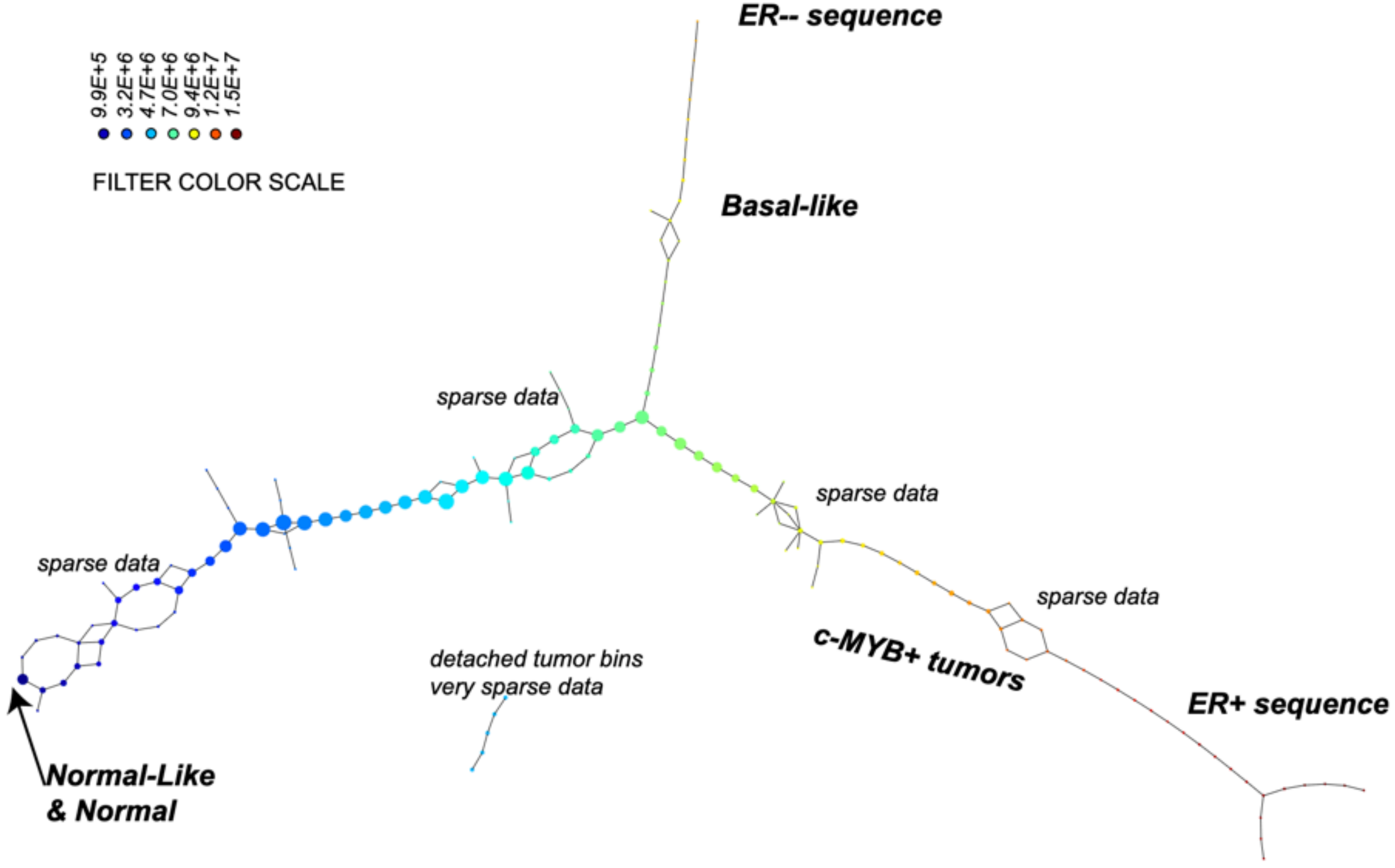
Input	Parameters	Output
Data points Metric Coloring scheme (optional)	Discretization parameters Lens (if lens-based)	A (colored) graph, where nodes are clusters of data points edges are nonempty intersection of clusters

What is Mapper?

Input	Parameters	Output
Data points Metric Coloring scheme (optional)	Discretization parameters Lens (if lens-based) Subroutine parameters	A (colored) graph, where nodes are clusters of data points edges are nonempty intersection of clusters

Nicolau et al, 2011

- data: Transcriptional microarray data
- sample: 295 tumors
- features: 262 genes
- lens: Normal component to the linear subspace of healthy tissues



What is Mapper Good For?

proximity guarantee, and…?

What is Mapper Good For?

proximity guarantee, and…?

- Distinguish points that are indistinguishable through the lens

What is Mapper Good For?

proximity guarantee, and…?

- Distinguish points that are indistinguishable through the lens
- Good for sub-typing

What is Mapper Good For?

proximity guarantee, and…?

- Distinguish points that are indistinguishable through the lens
- Good for sub-typing
- Finding out relationship between them

Limitations of Mapper

Limitations of Mapper

- cannot retain metric information

Limitations of Mapper

- cannot retain metric information
- choices of lens

Limitations of Mapper

- cannot retain metric information
- choices of lens
- parameter tuning

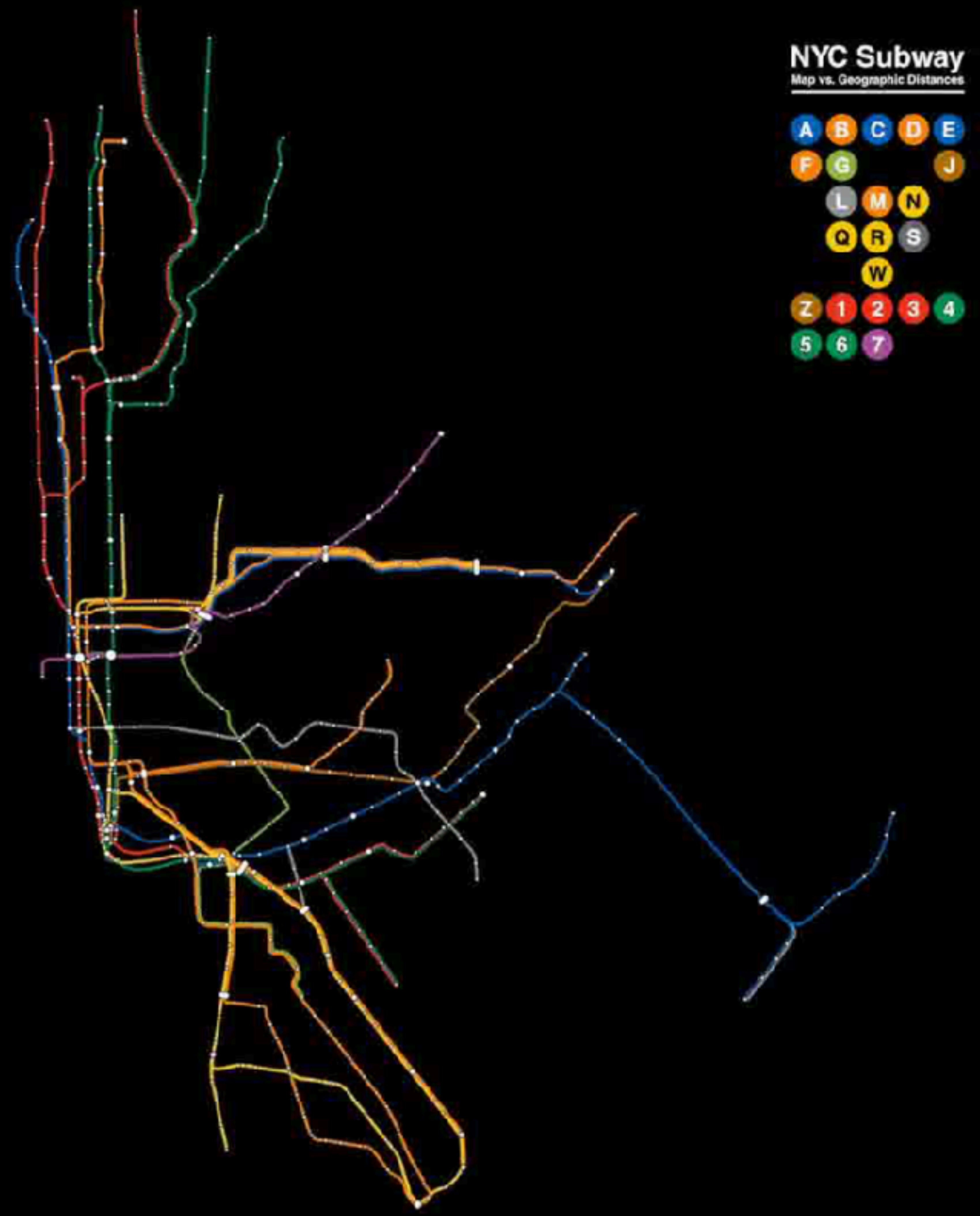
Limitations of Mapper

- cannot retain metric information
- choices of lens
- parameter tuning
- difficulty of clustering

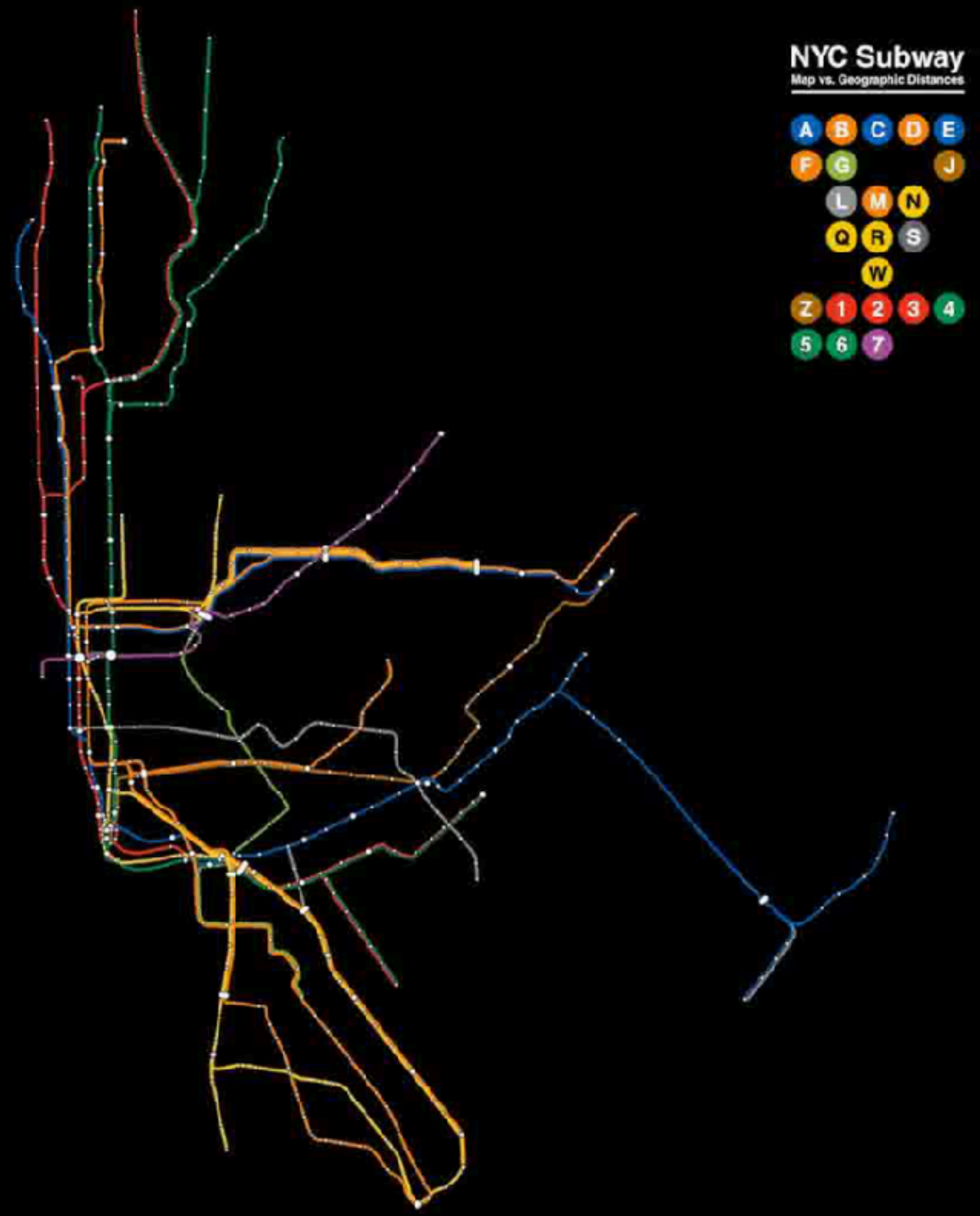
Intro to $\boxed{\text{Topological}}$ Data Analysis

Network and Homology

Chunyin Siu, Feb 21, 2025



Playhouse_animation
https://www.reddit.com/r/dataisbeautiful/comments/6c51re/nyc_subway_map_distances_vs_geographic_distances/



Playhouse_animation
https://www.reddit.com/r/dataisbeautiful/comments/6c51re/nyc_subway_map_distances_vs_geographic_distances/

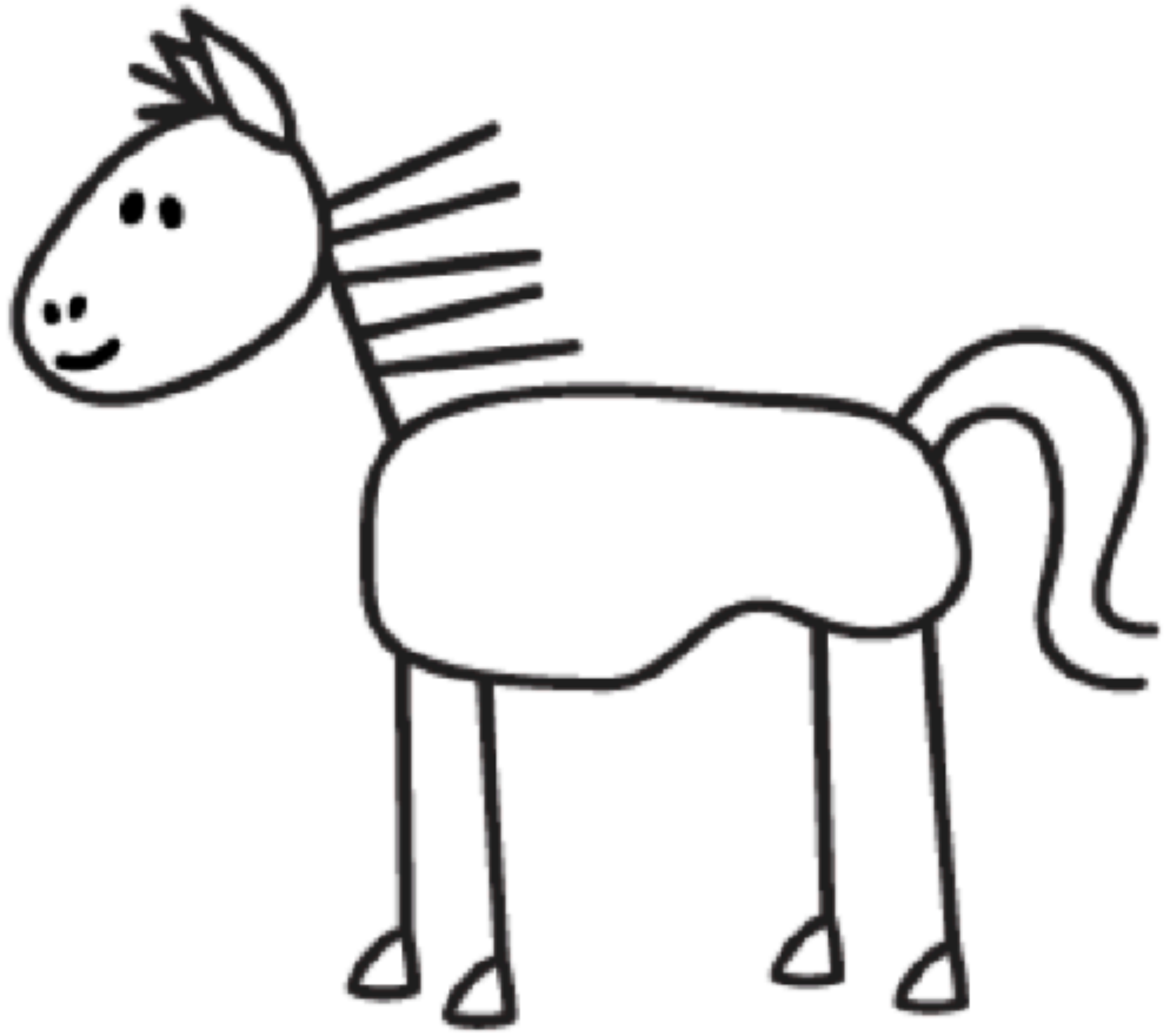
Raw Data Good

Combinatorial Representation Better

Photo by David Dibert from Pexels: https://www.pexels.com/photo/brown-horse-on-grass-field-635499/

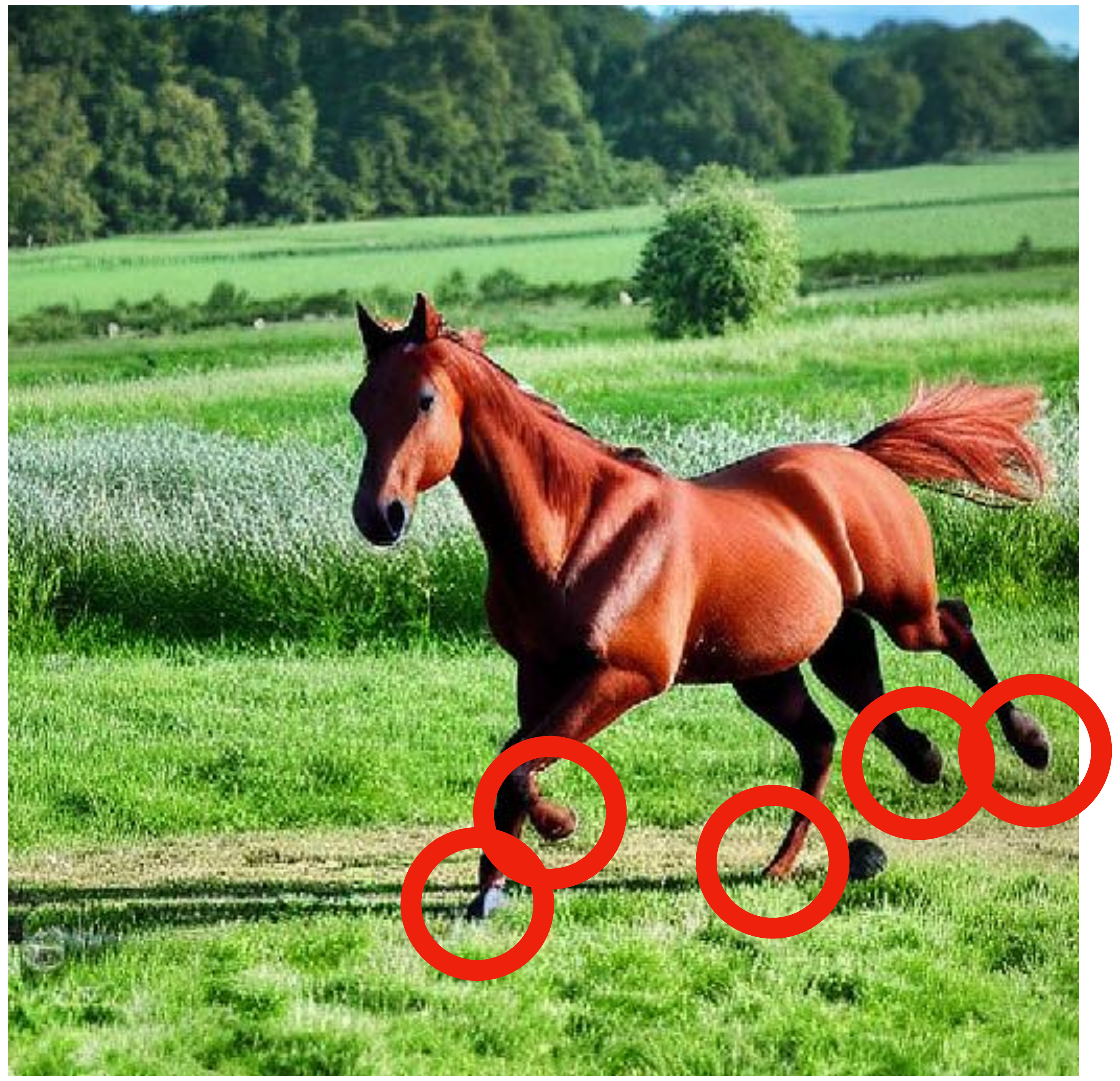
Photo by Pixabay from Pexels: https://www.pexels.com/photo/white-horse-461717/

Mea Culpa



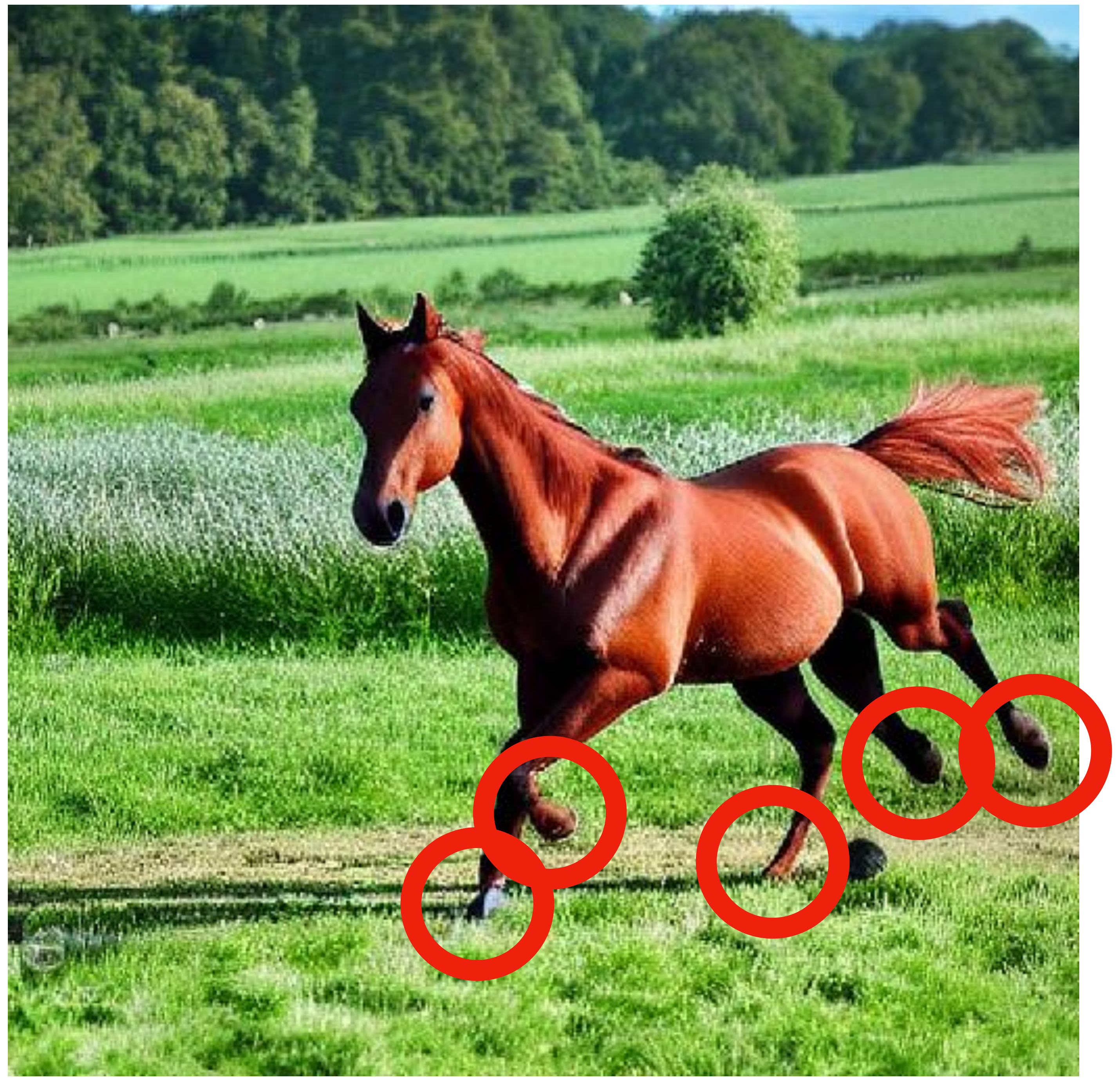
from Cliparts: https://cliparts.co/clipart/2503420

from Francois Chollet's x account: https://twitter.com/fchollet/status/1573836241875120128



from Francois Chollet's x account: https://twitter.com/fchollet/status/1573836241875120128

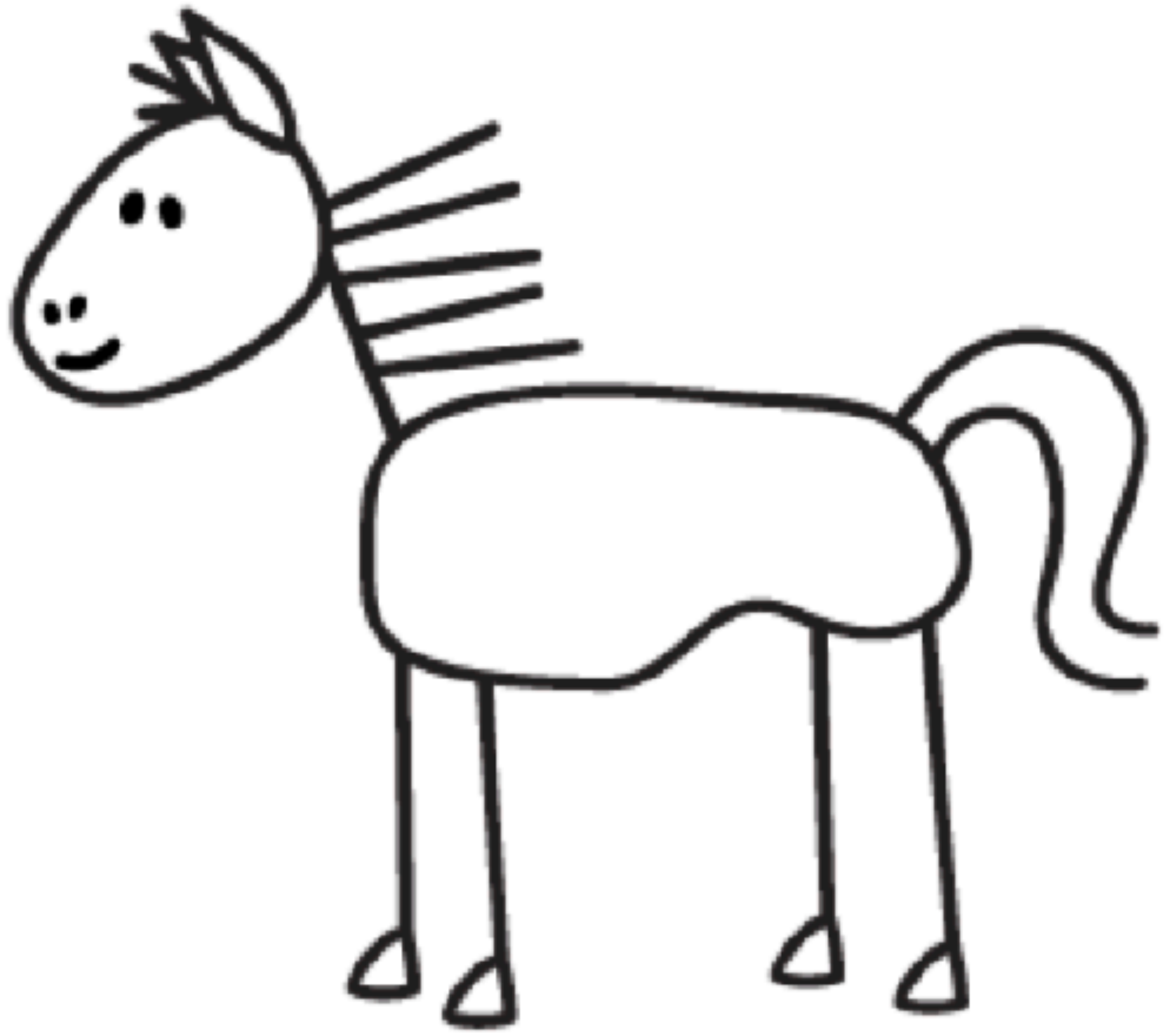
Mea Maxima Culpa



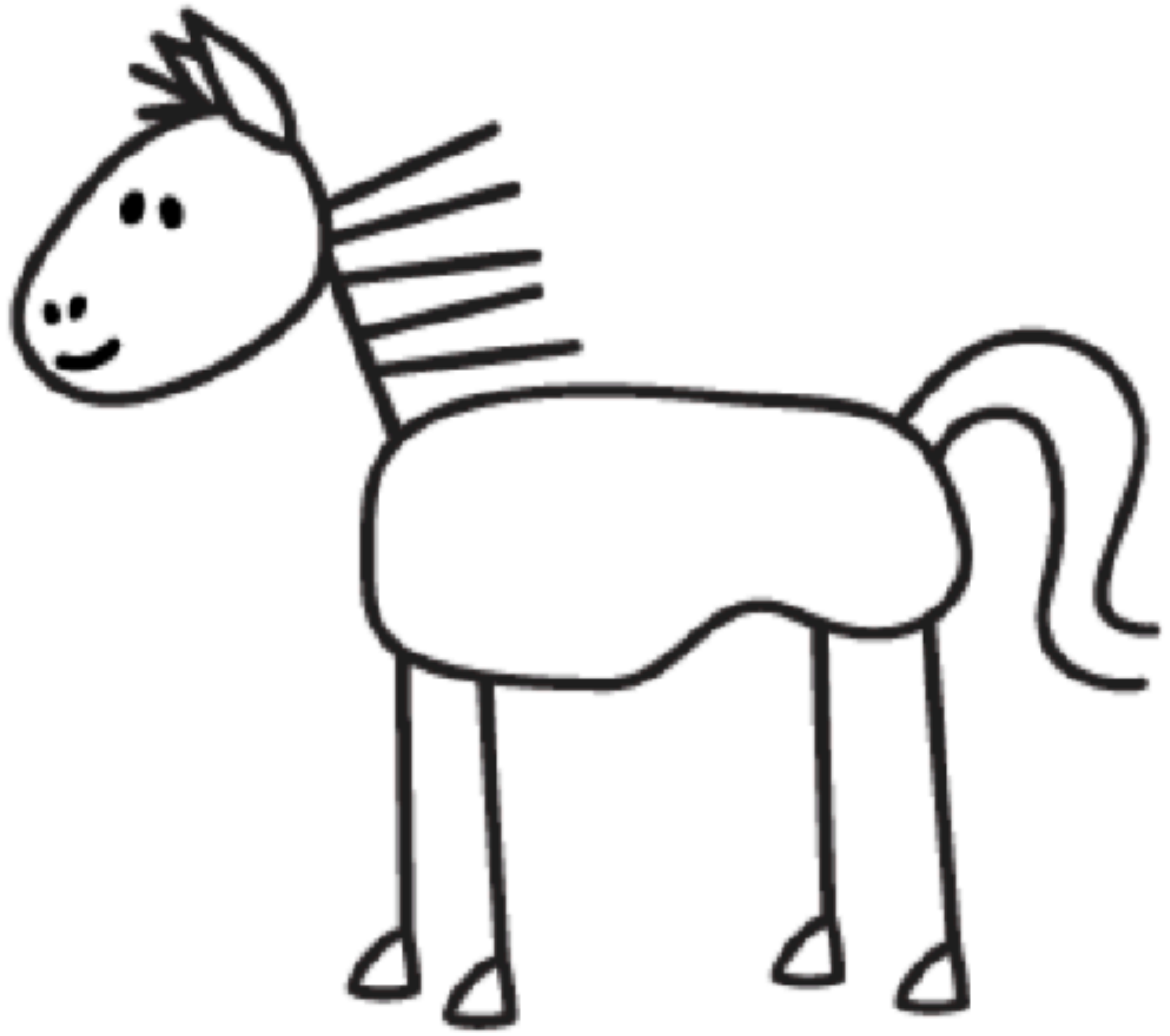
from Francois Chollet's x account: https://twitter.com/fchollet/status/1573836241875120128

Forgive me Father
for my combinatorial representation is wrong


NYC Subway
Map vs. Geographic Distances
A B C D E
F G J
L M N
Q R S
W
Z 1 2 3 4
5 6 7



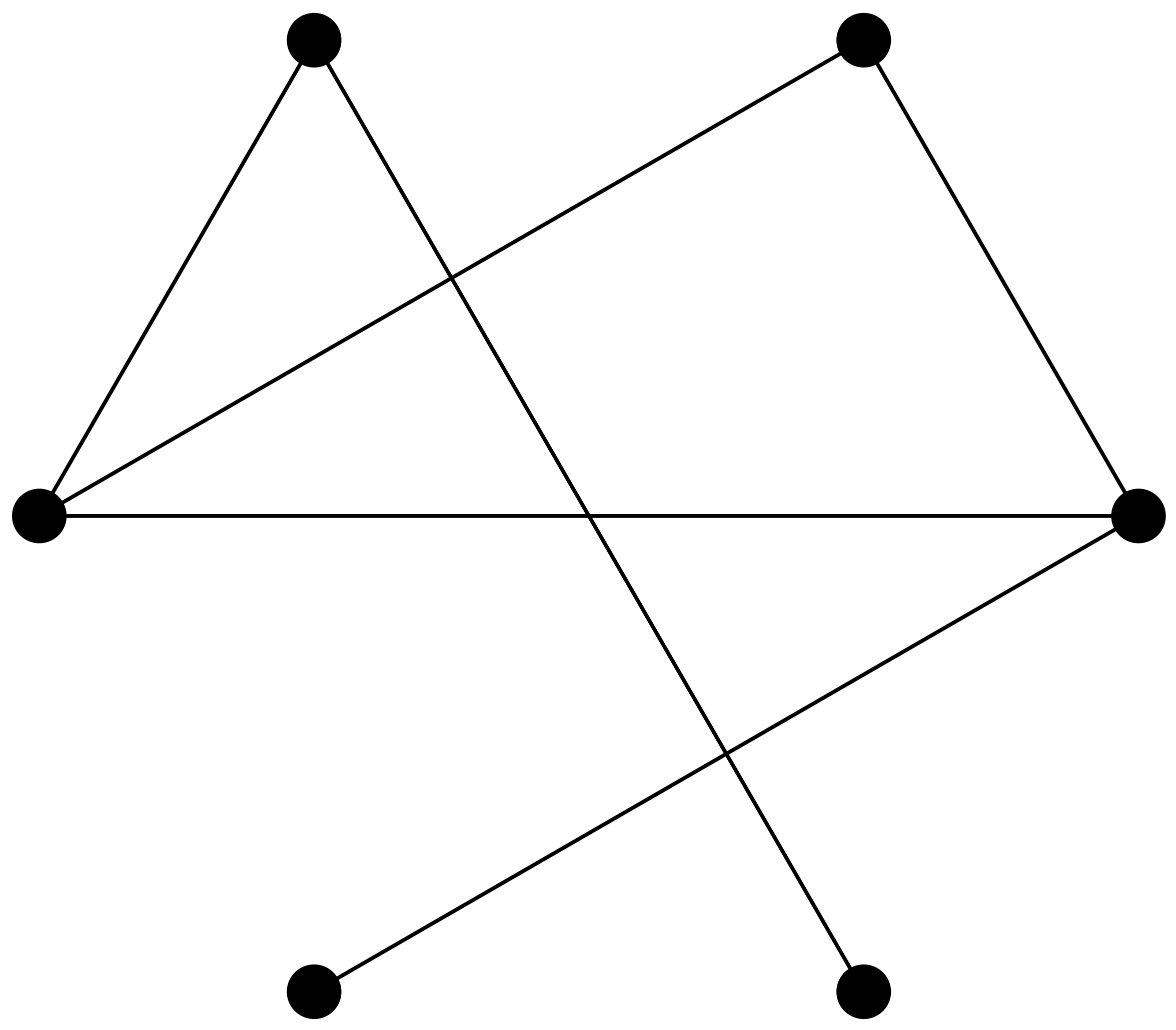
But how do we compare them?



Topology

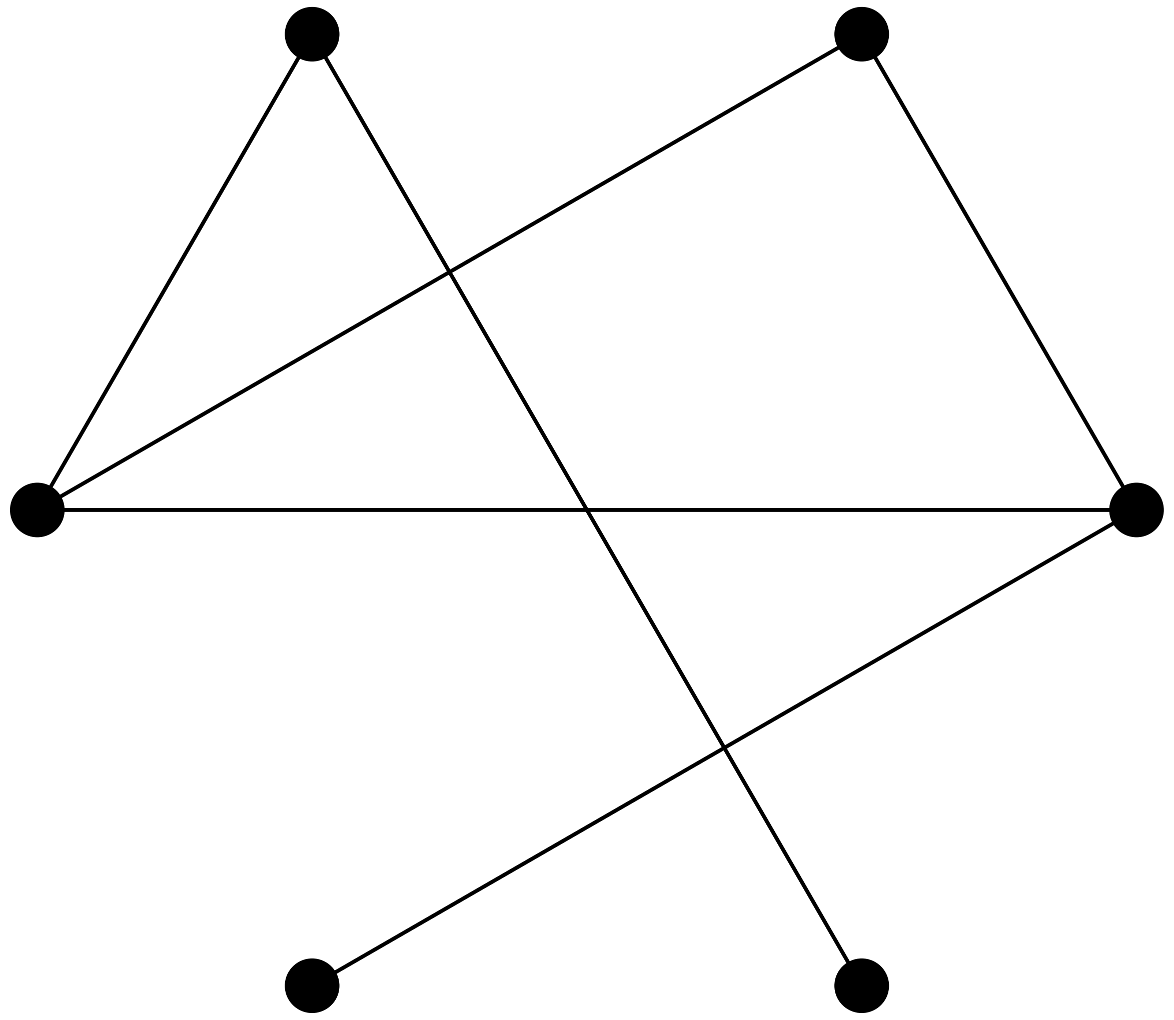
Disambiguation

Disambiguation



Graph theory

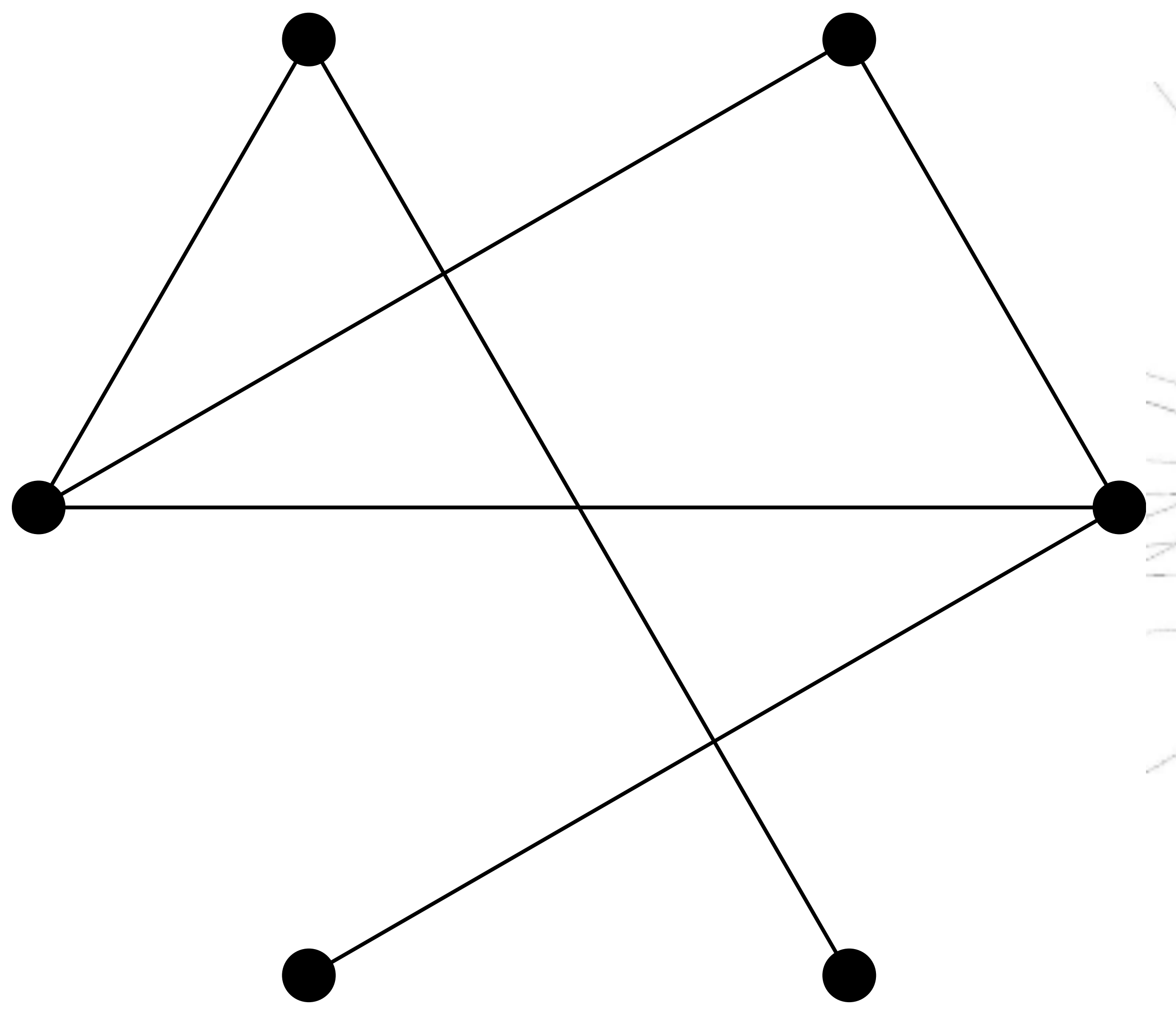
Disambiguation



mathematical study of
objects with vertices and
edges

Graph theory

Disambiguation



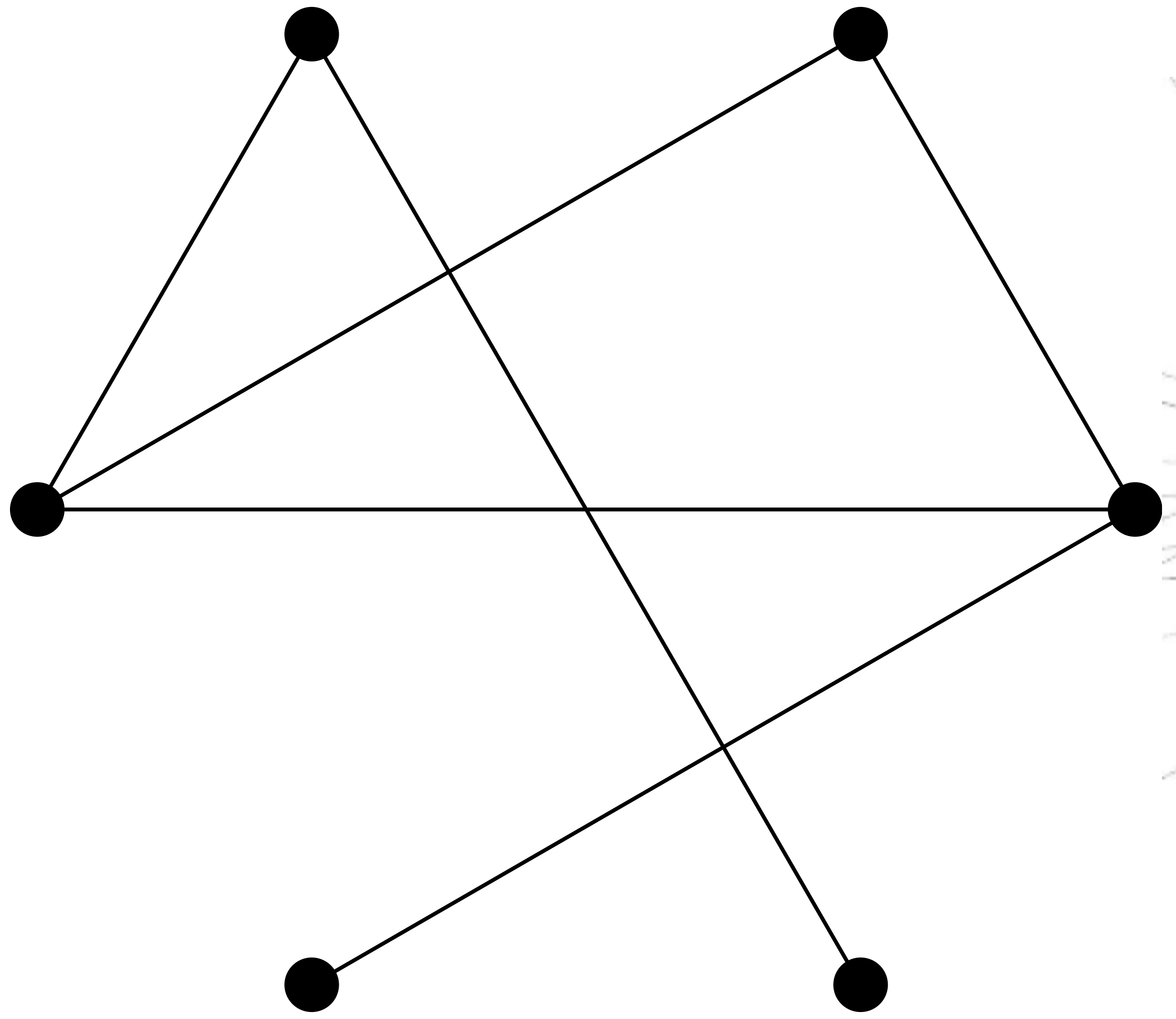
applied graph theory

Stephen Coast
https://www.fractalus.com/steve/stuff/ipmap/

Graph theory

Network science

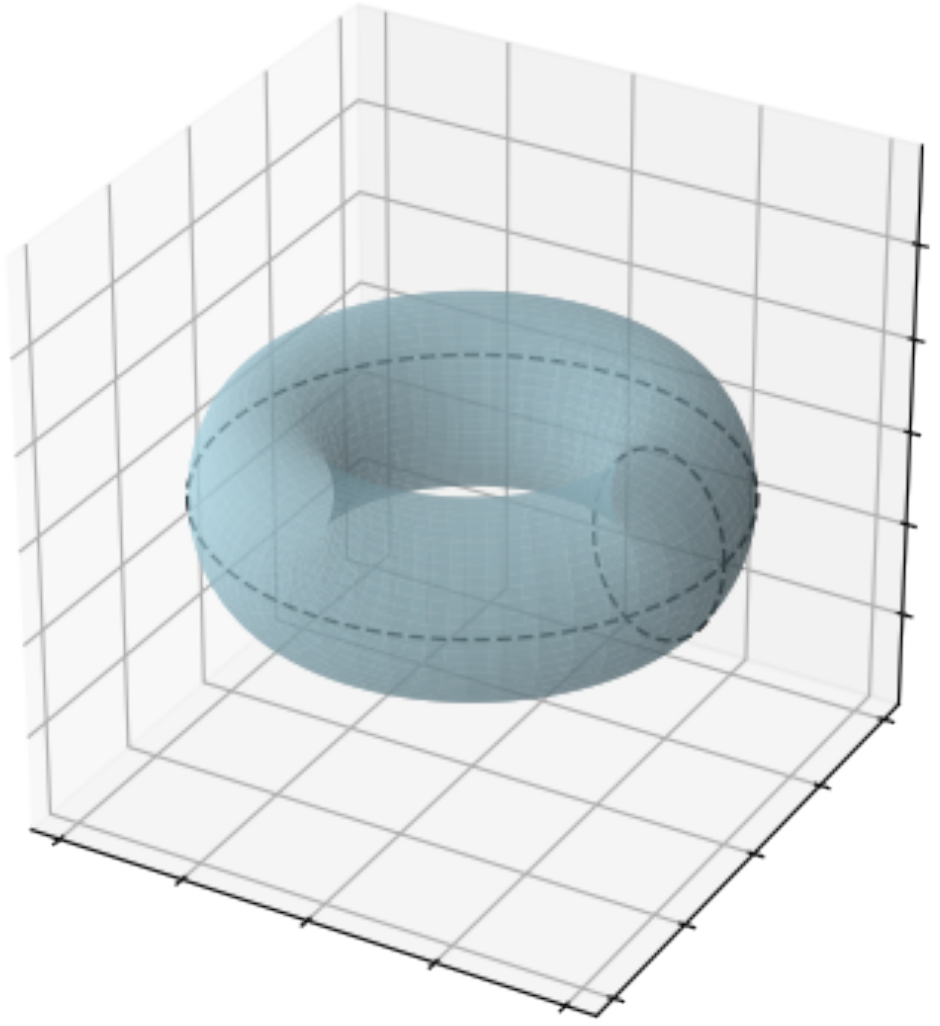
Disambiguation



Graph theory

Stephen Coast
https://www.fractalus.com/steve/stuff/ipmap/

Network science



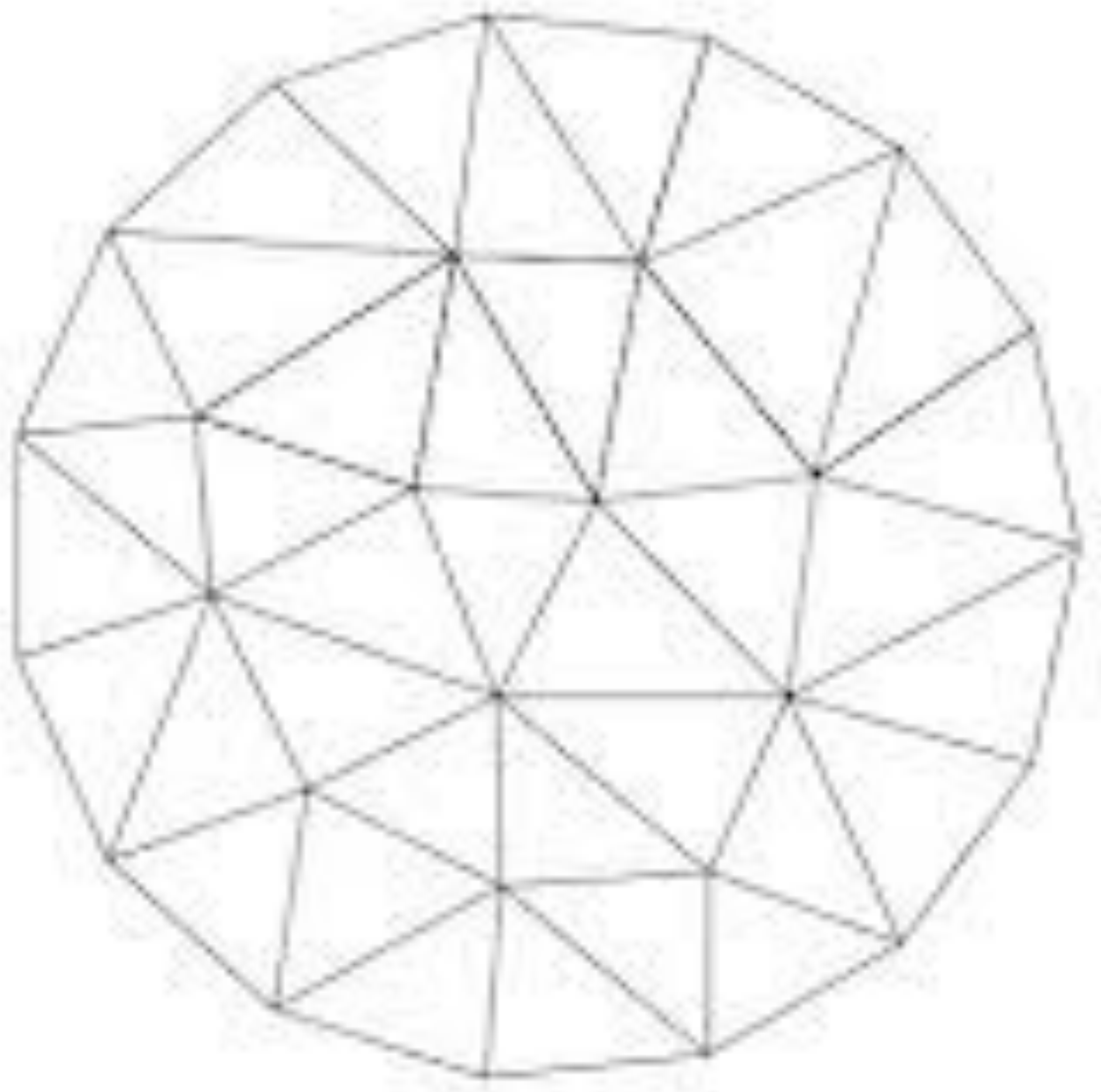
Andrey Yao

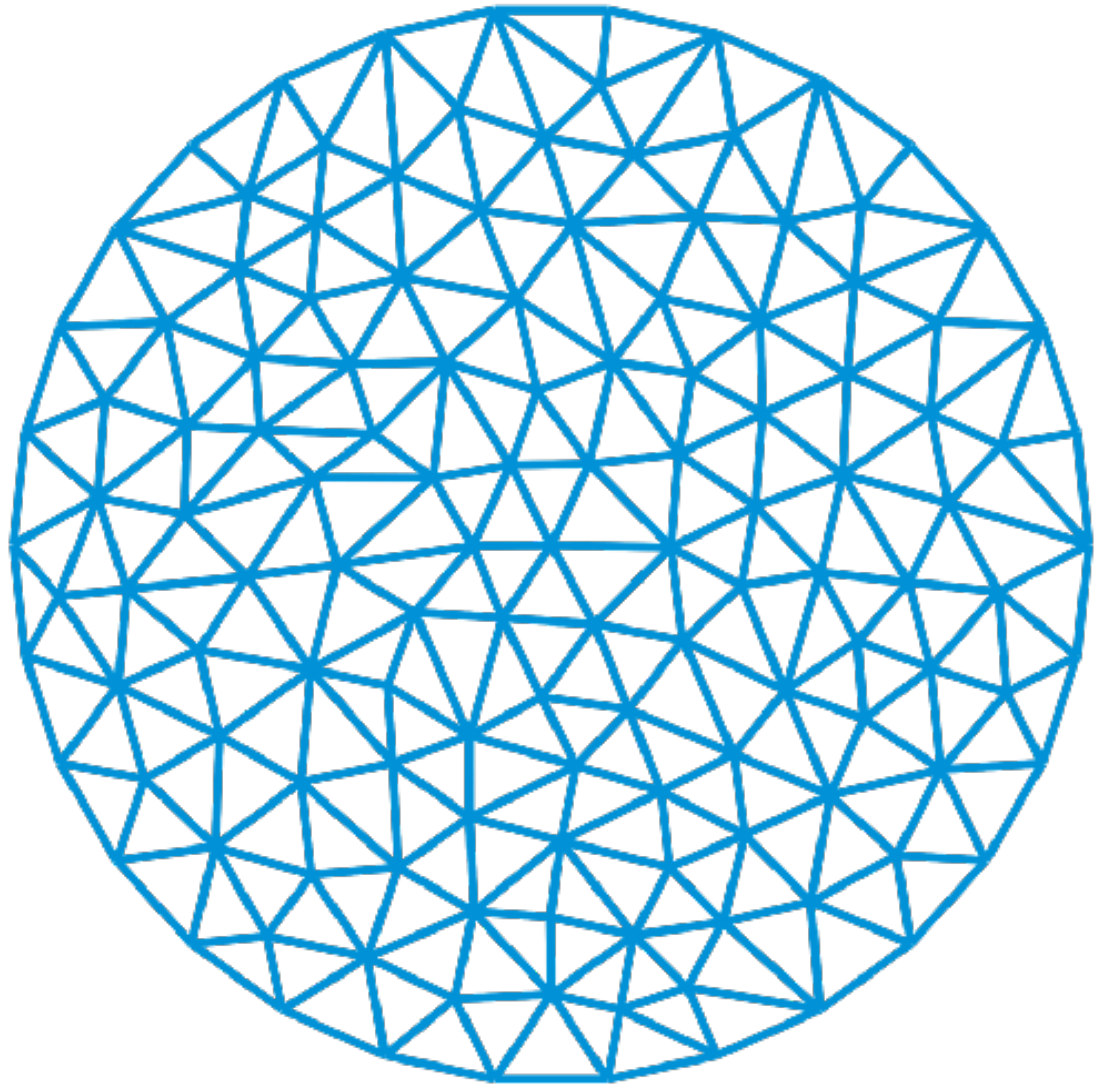
Algebraic topology

mathematical study of combinatorial objects made up of vertices, edges, triangles, etc

Why triangles and higher dimensions?

Why triangles and higher dimensions?



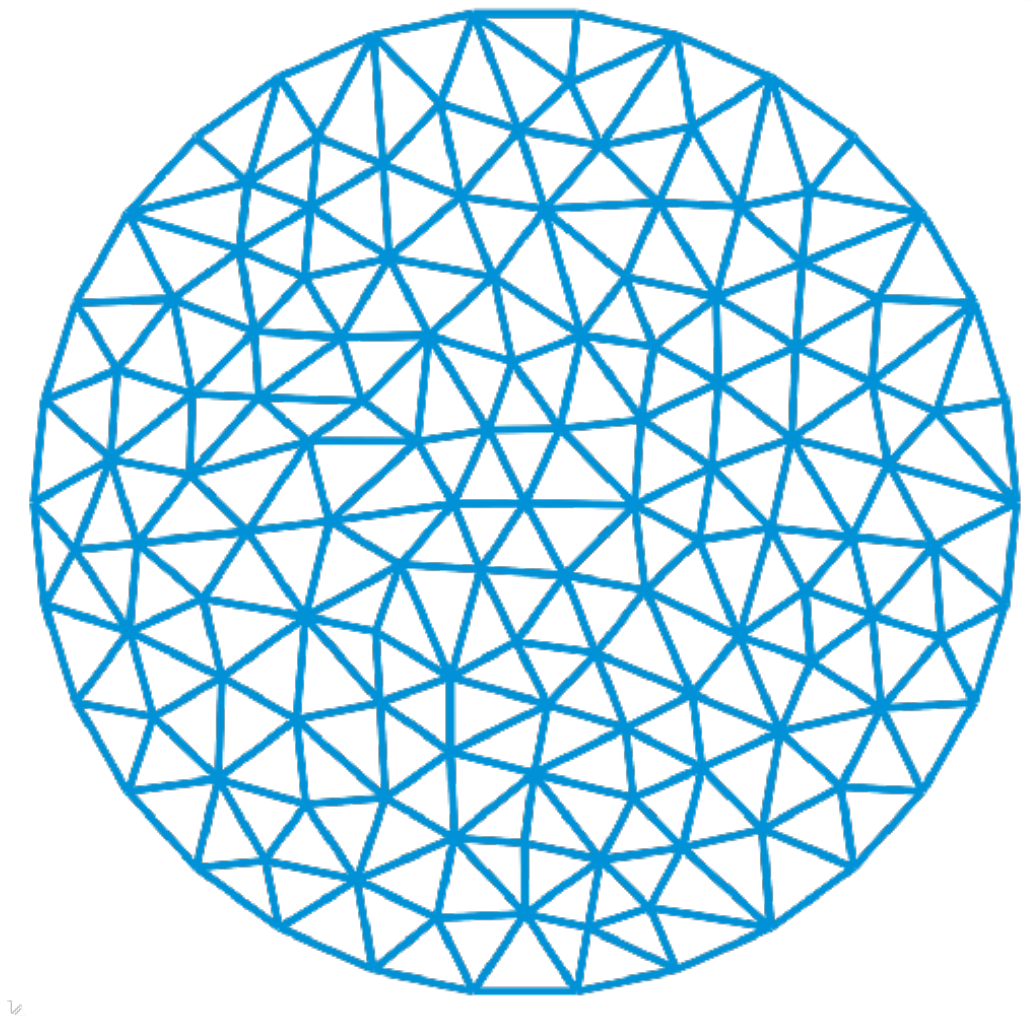


Graphs v.s. Higher-Dimensional Objects

- when nodes and edges are well defined

Graphs v.s. Higher-Dimensional Objects

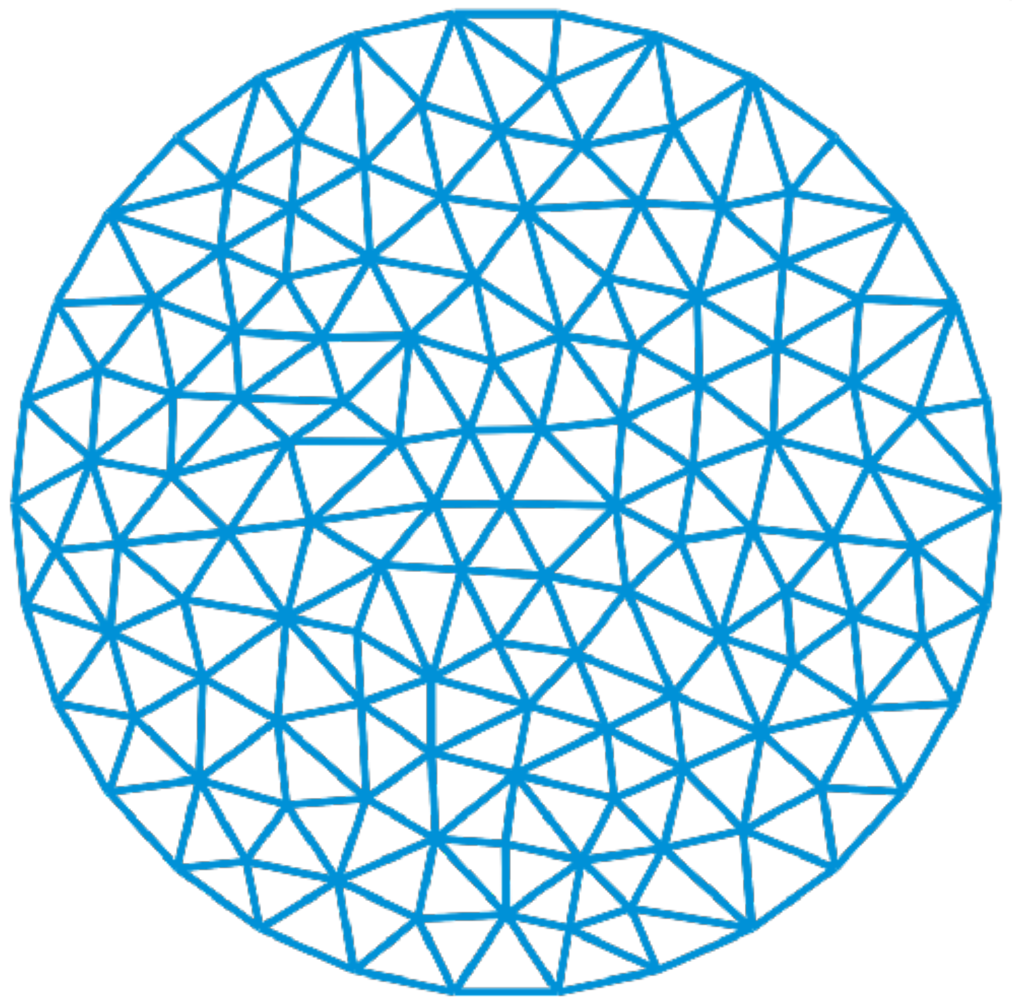
- when nodes and edges are well defined



- when nodes and edges are noisy

Graphs v.s. Higher-Dimensional Objects

- when nodes and edges are well defined



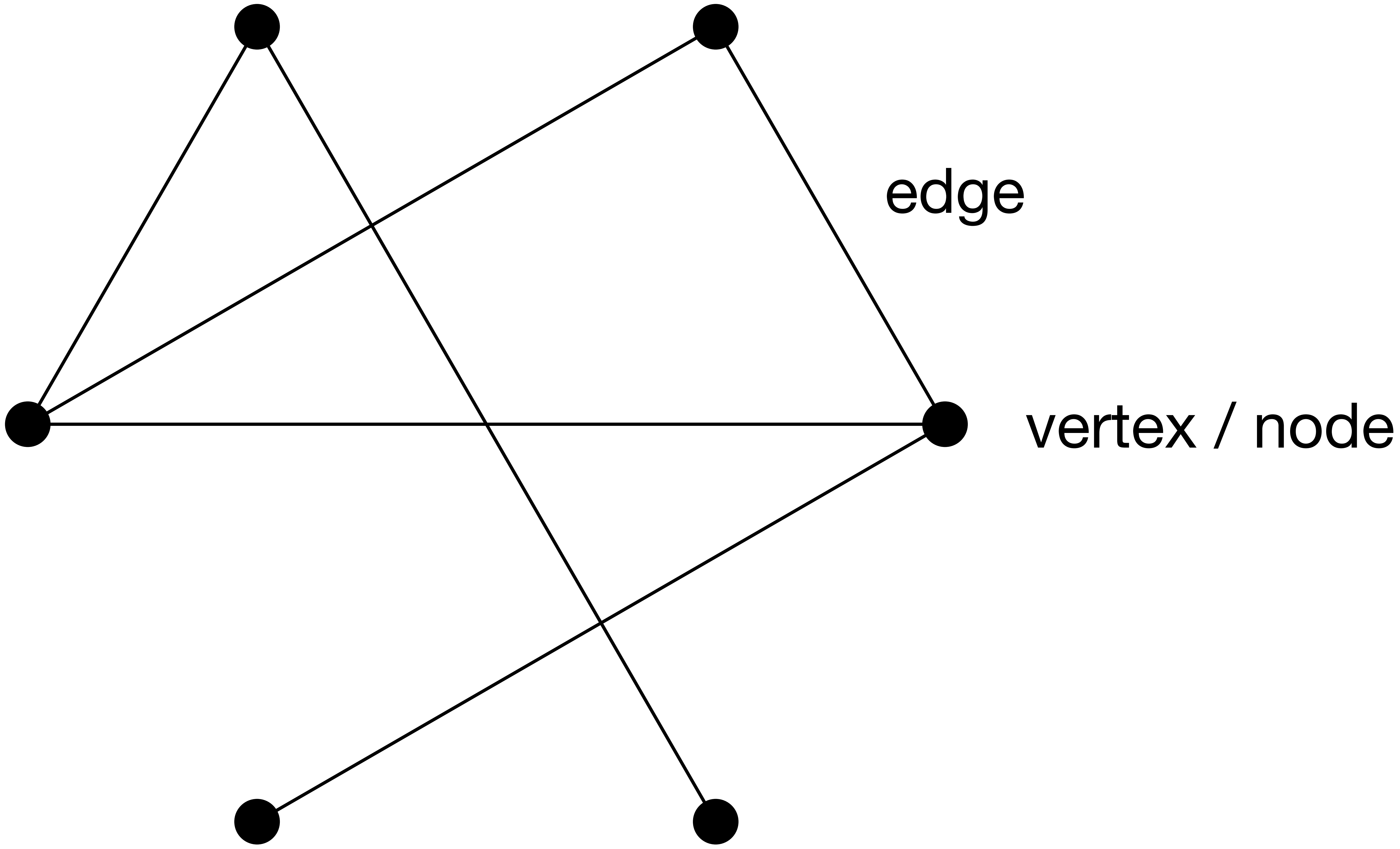
- when nodes and edges are noisy
- when there is a continuous underlying objects

Take-Home Message

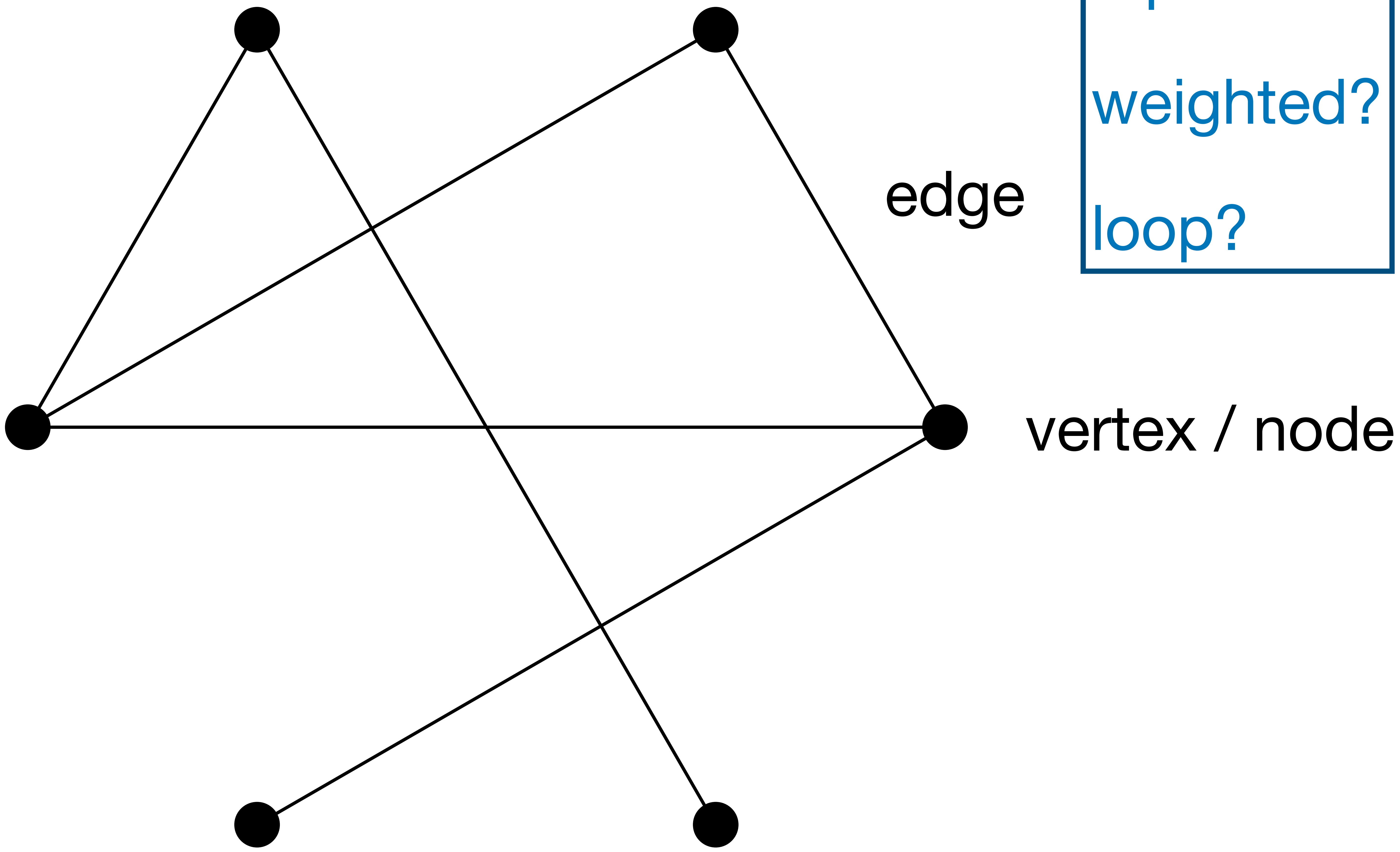
- Combinatorial representations highlight “relevant” features.
- Networks are good when vertices and edges are meaningful.
- Algebraic topology is the study of comparison of combinatorial objects.

Network Science

What is a graph?



What is a graph?



directed?

repeated?

weighted?

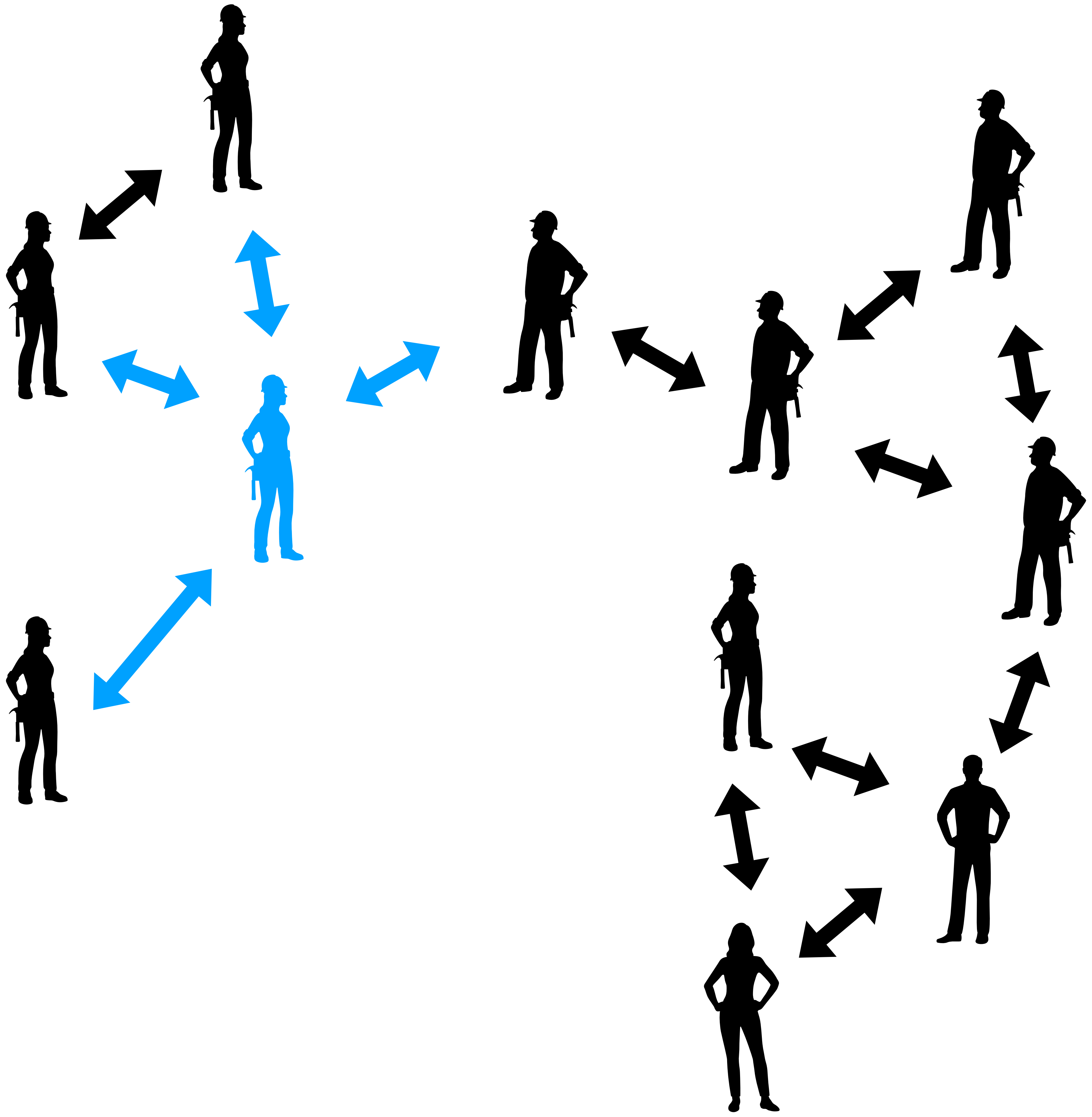
loop?

Storing a graph

- list of edges
- adjacency matrix

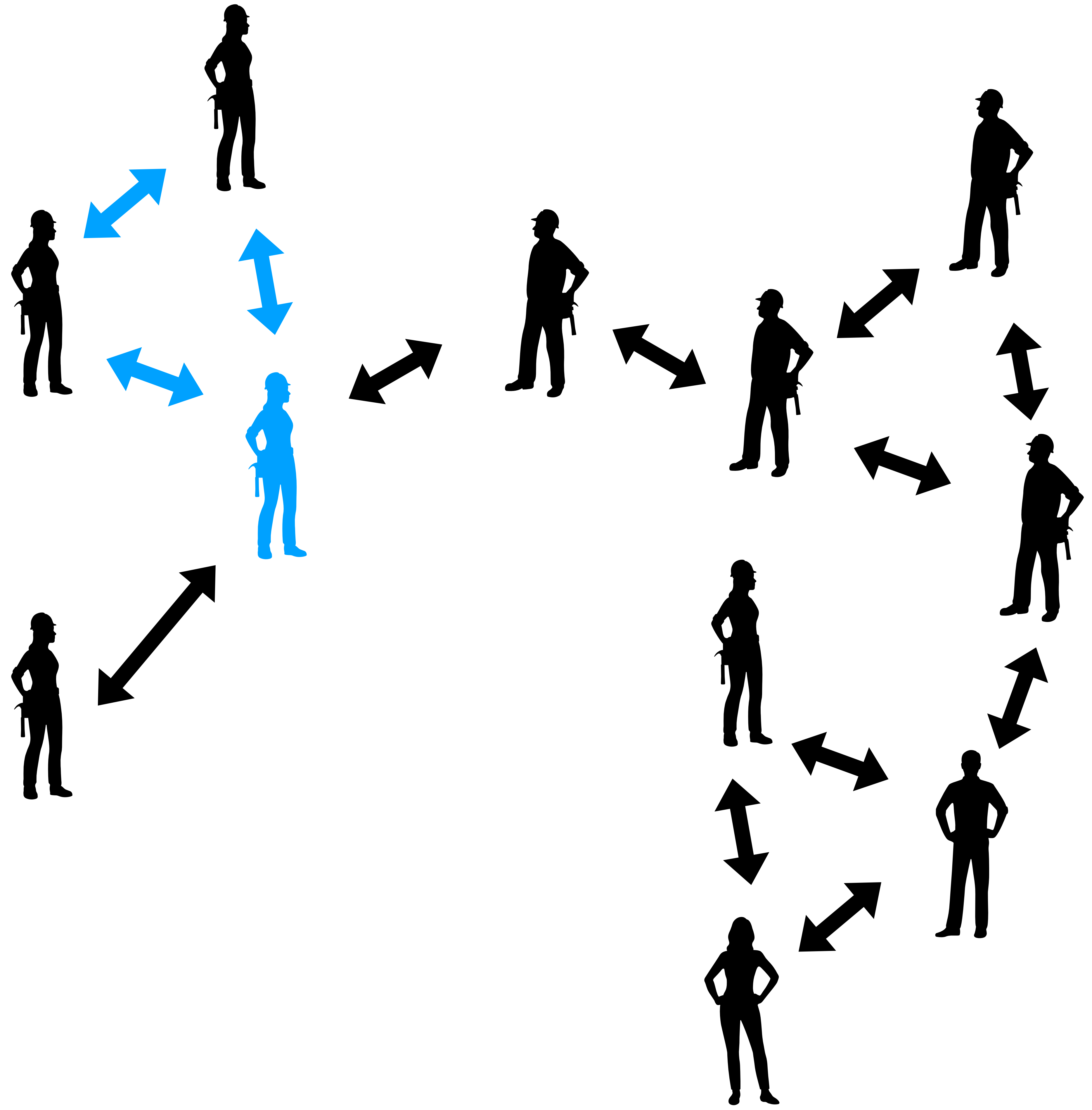
Counting

- Degree

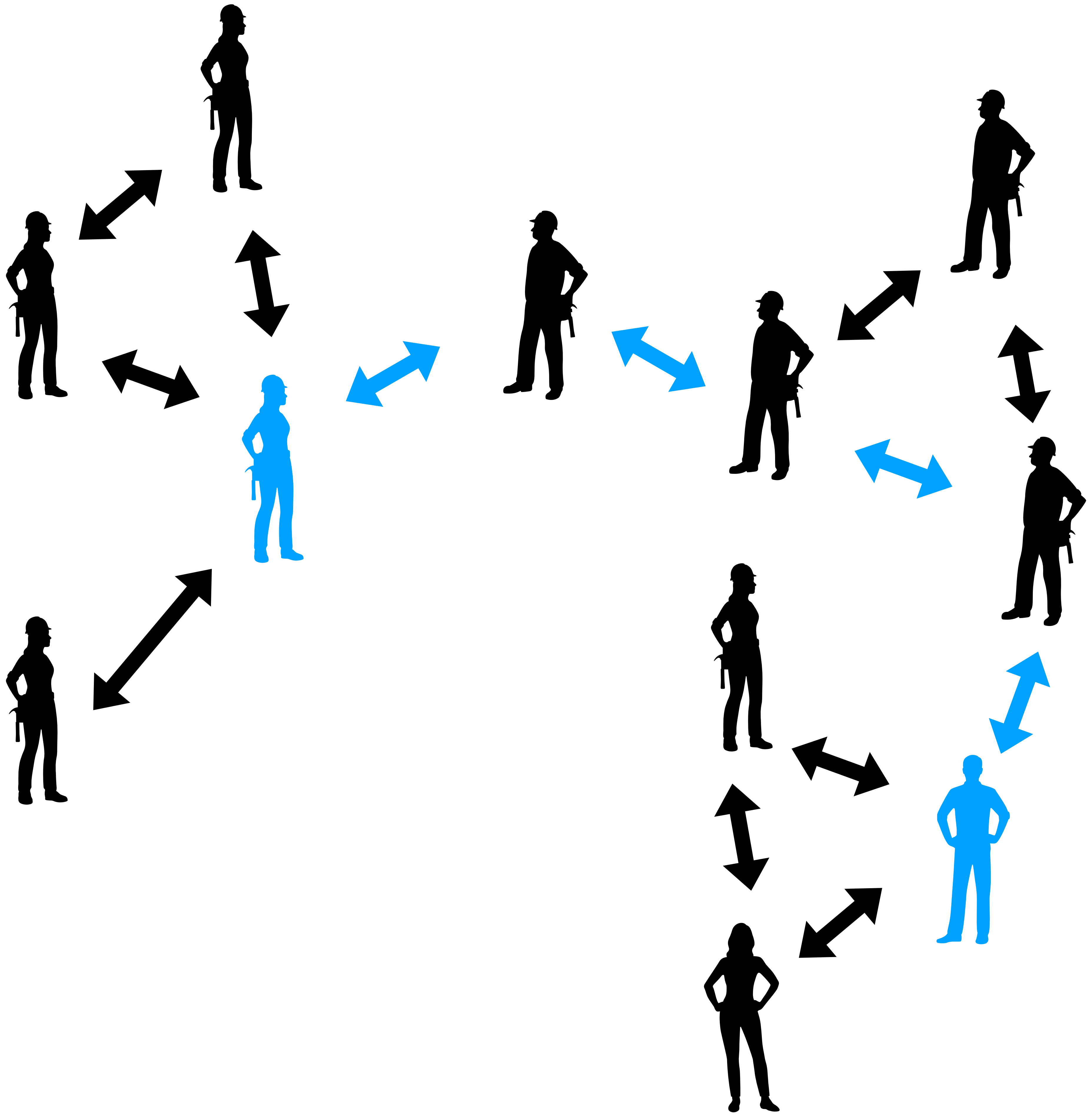


Counting

- Degree
- Clustering coefficients

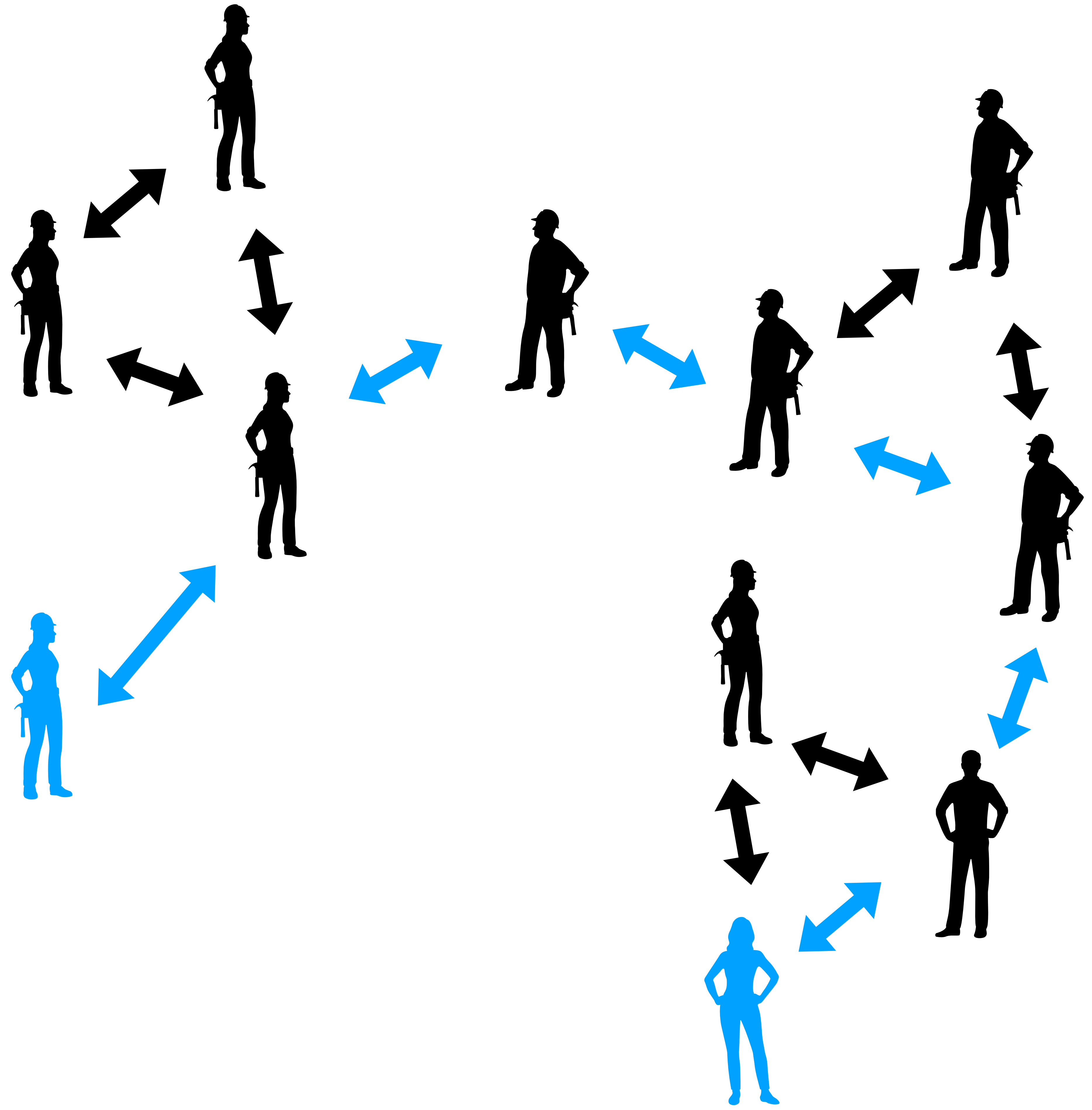


Geometry



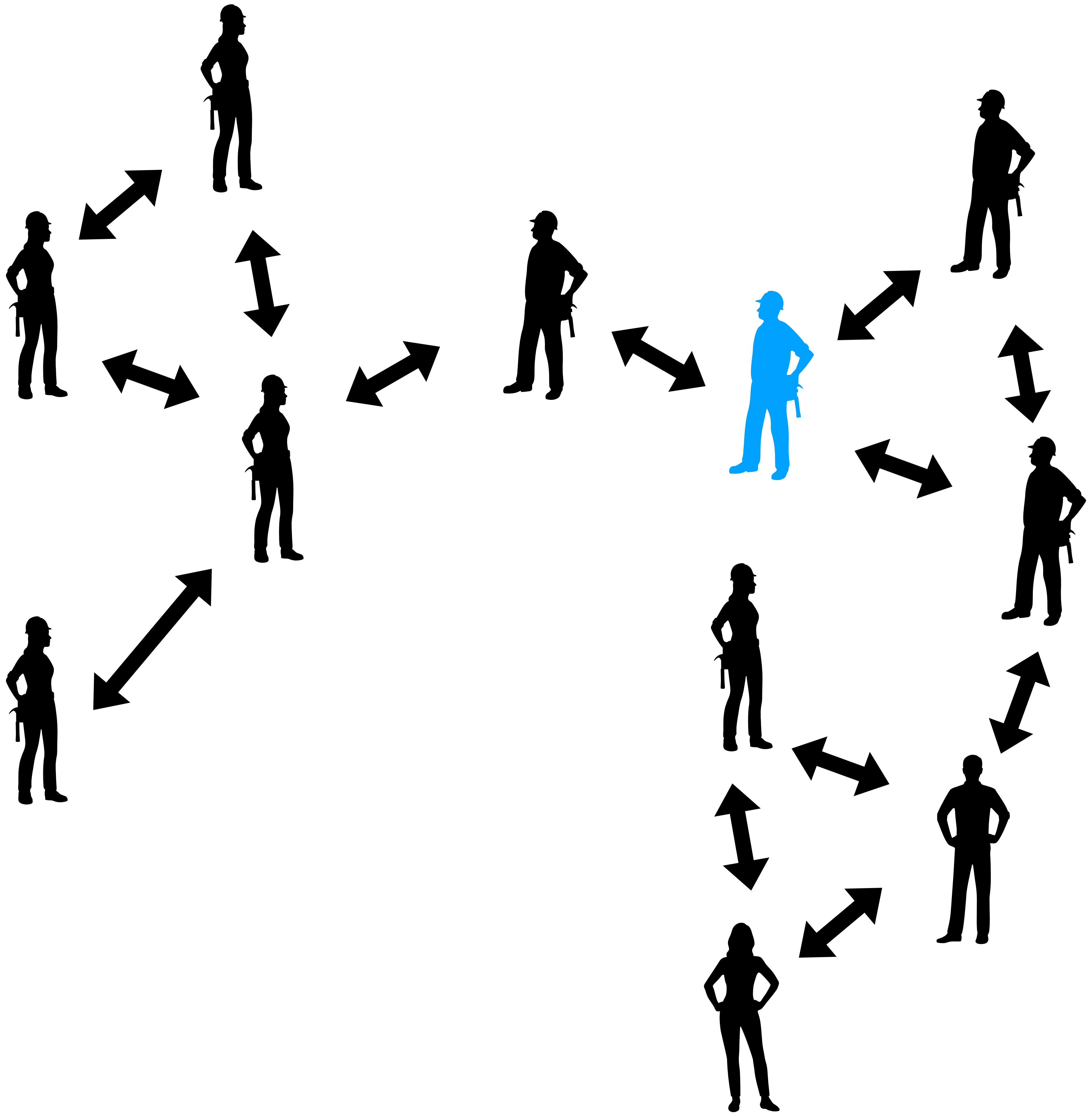
Geometry

- Diameter



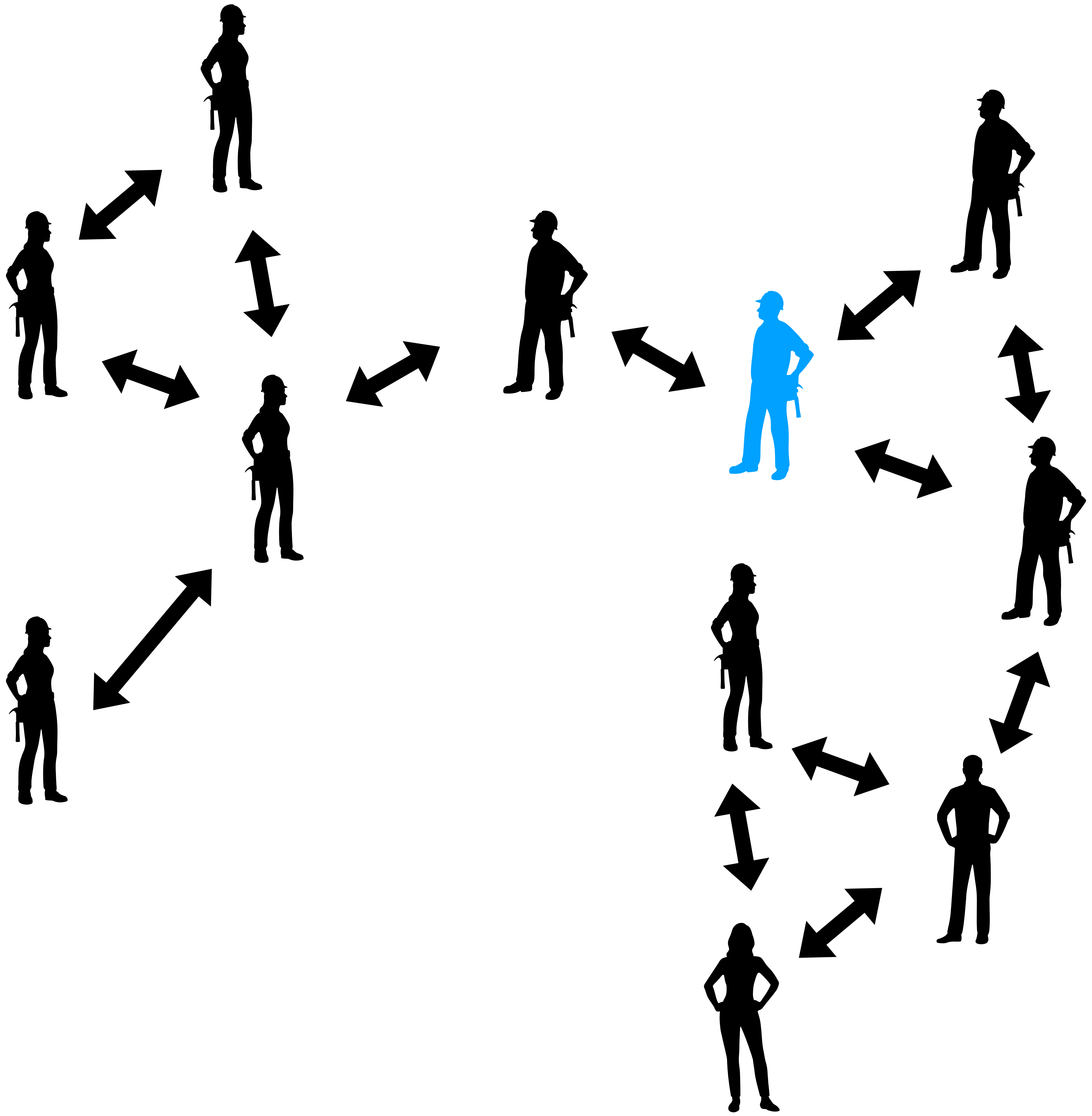
Geometry

- Diameter
- centrality

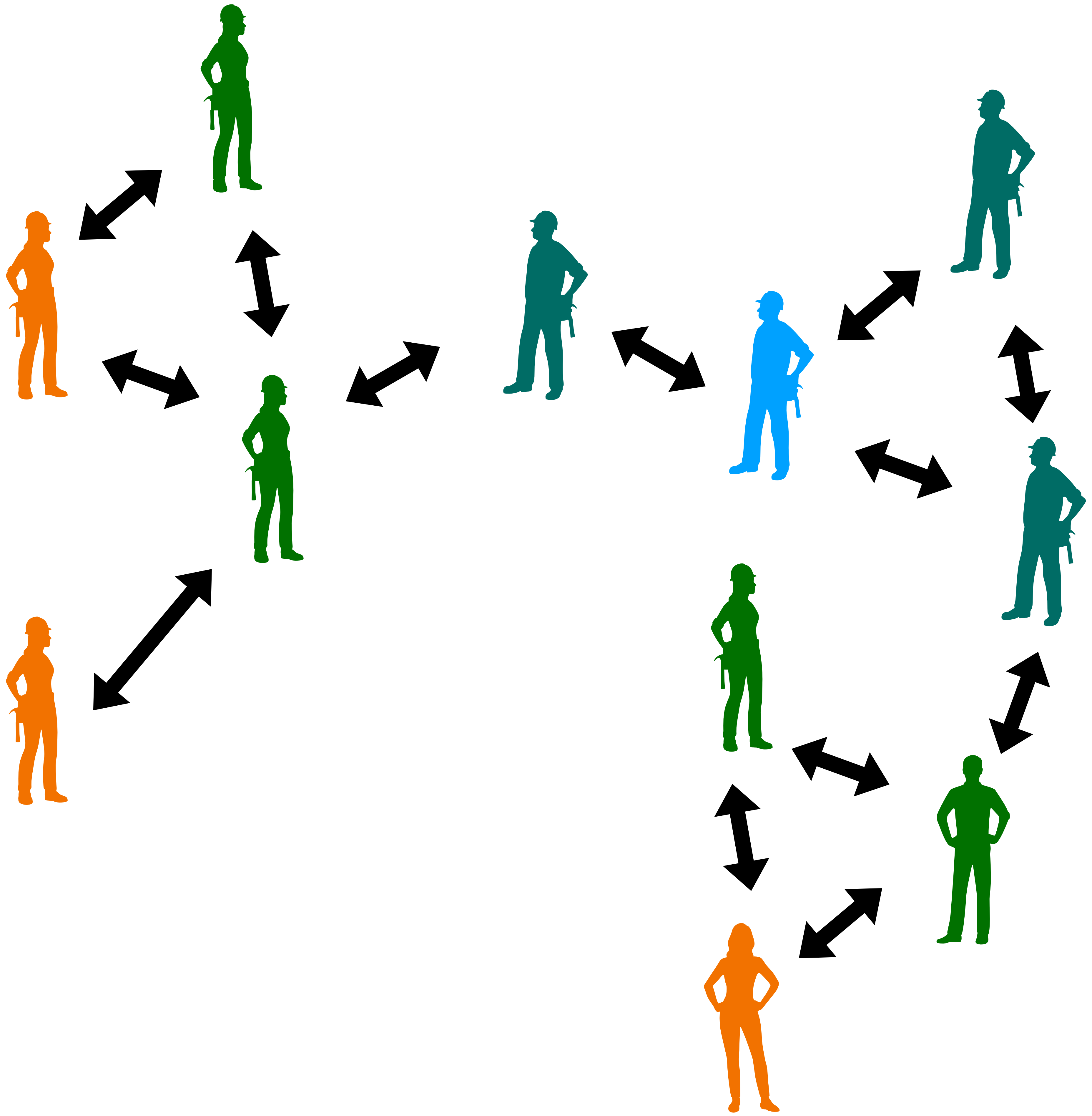


Geometry

- Diameter
- centrality
    - closeness centrality of a node (1/average distance from other points)
    - betweenness centrality of a node (fraction of shortest paths through a node)

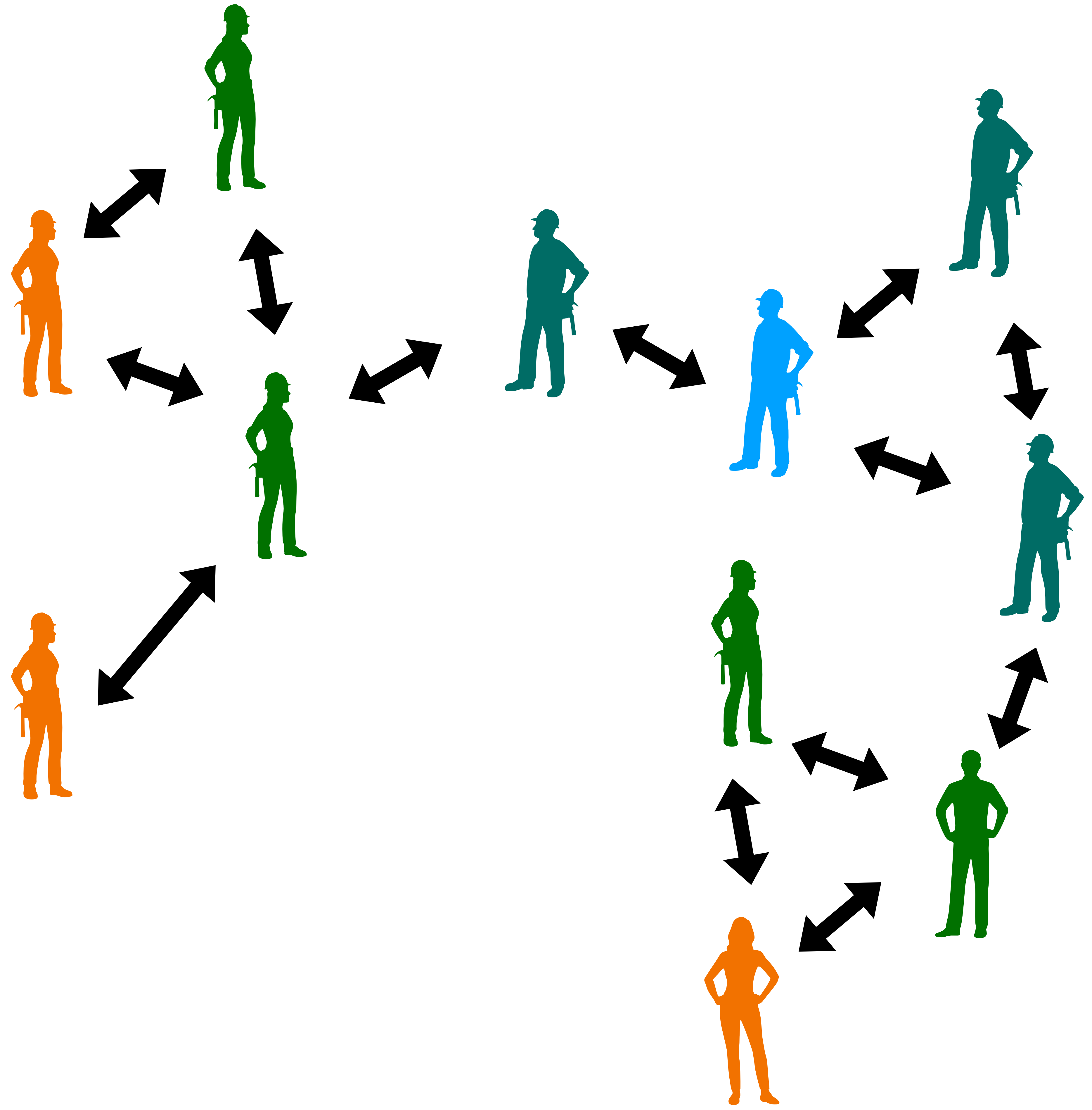


Spectral Geometry



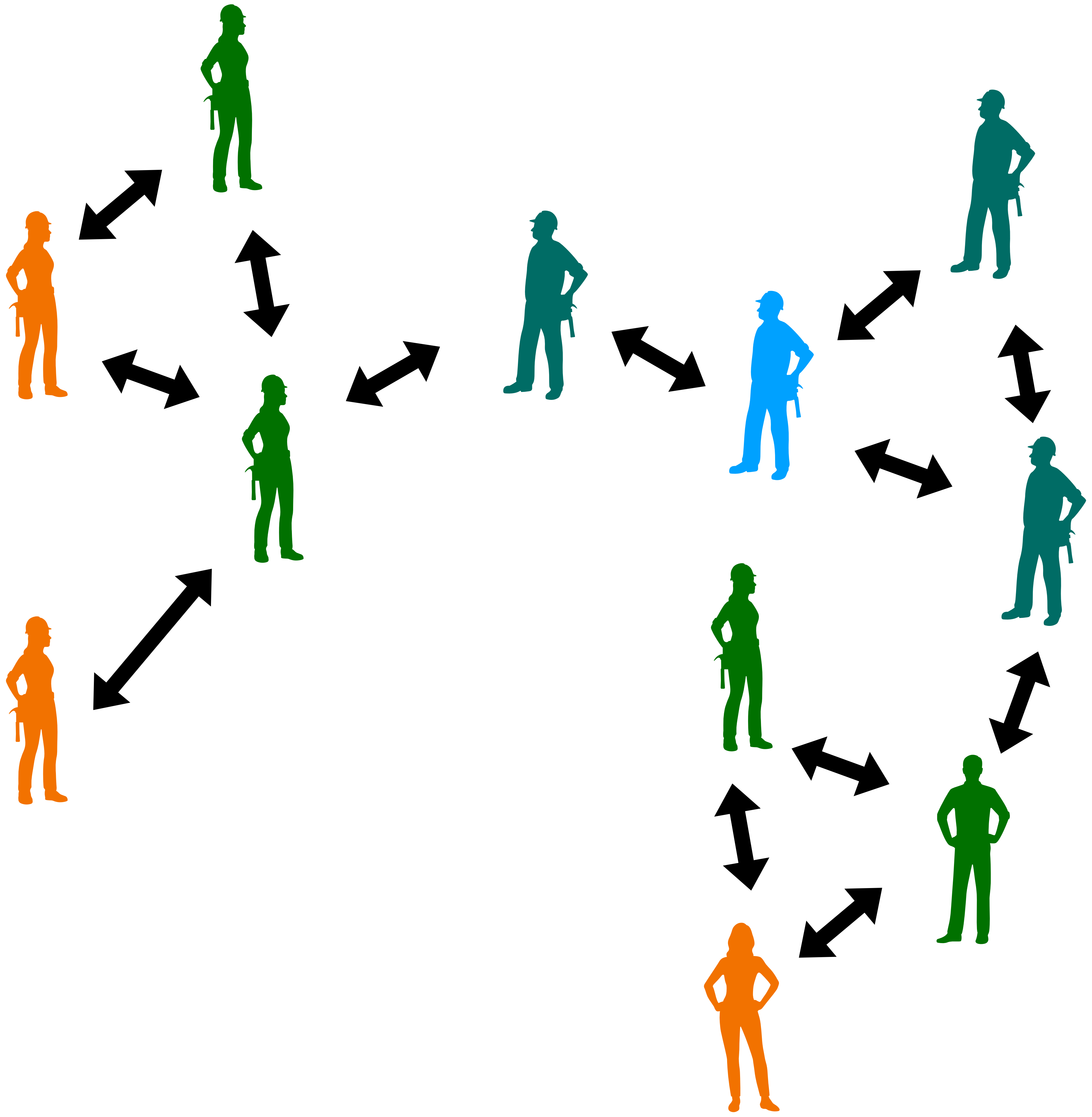
Spectral Geometry

- Graph Laplacian



Spectral Geometry

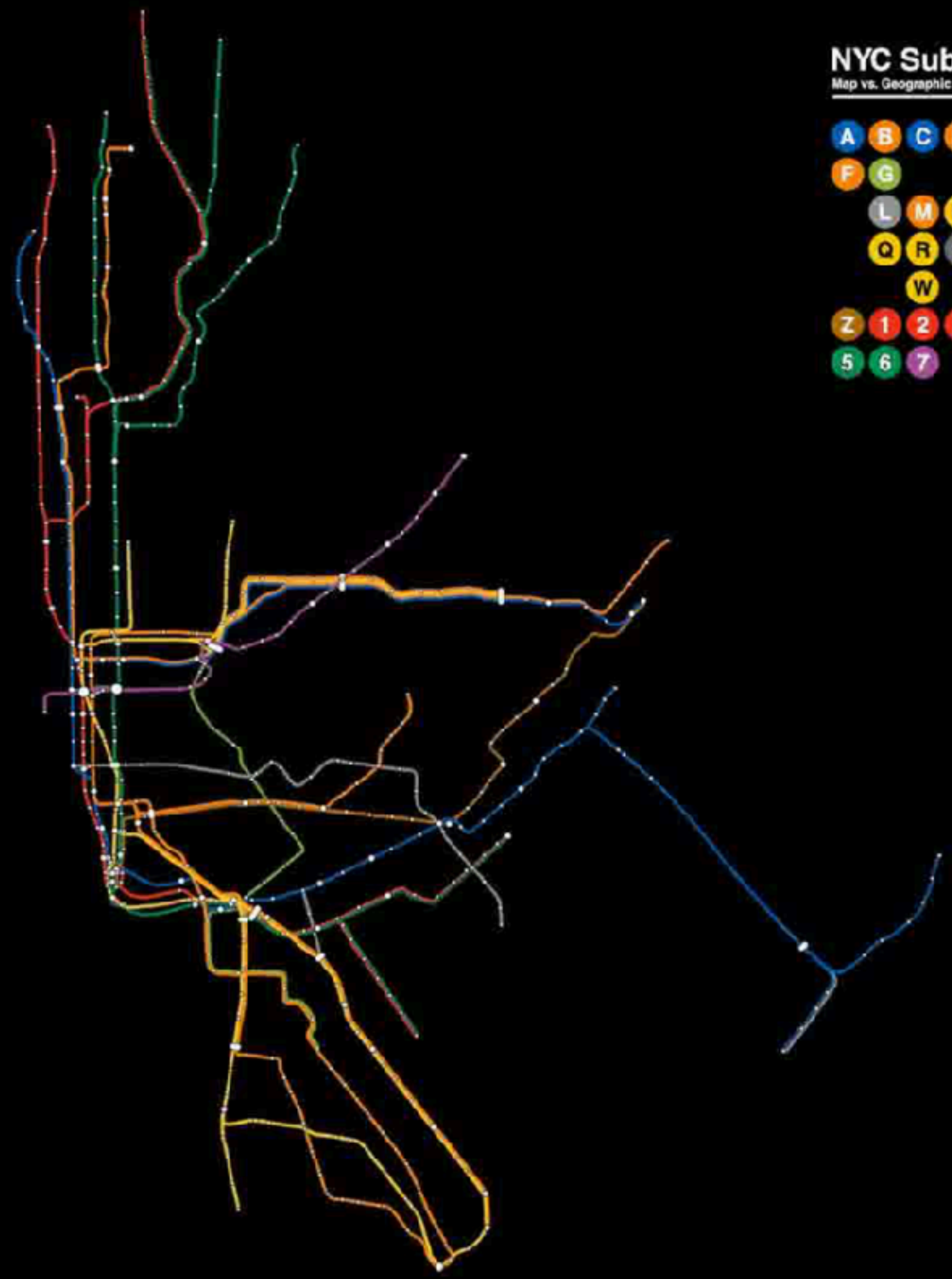
- Graph Laplacian
- Eigenvector centrality



Network

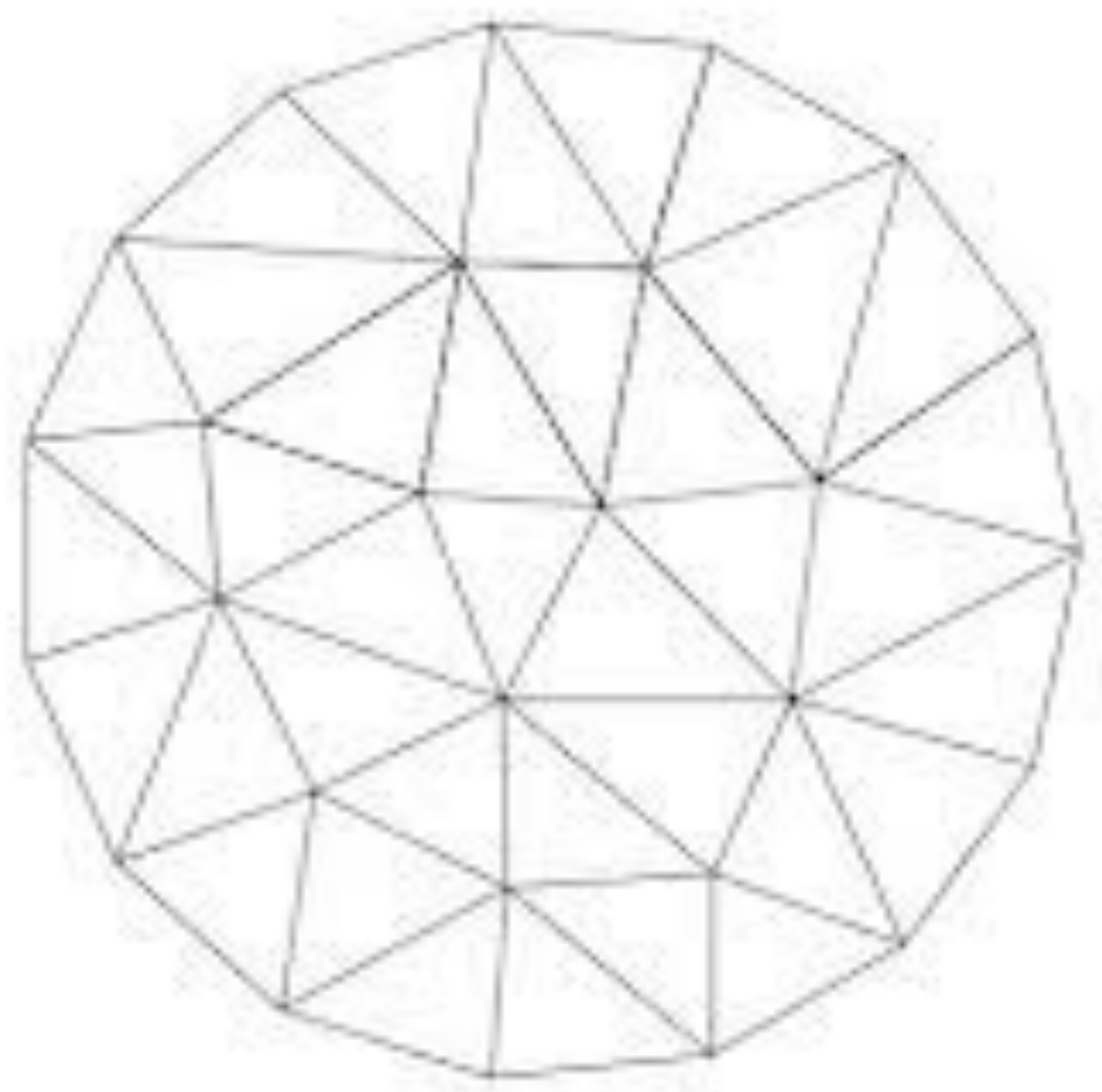
Algebraic Topology

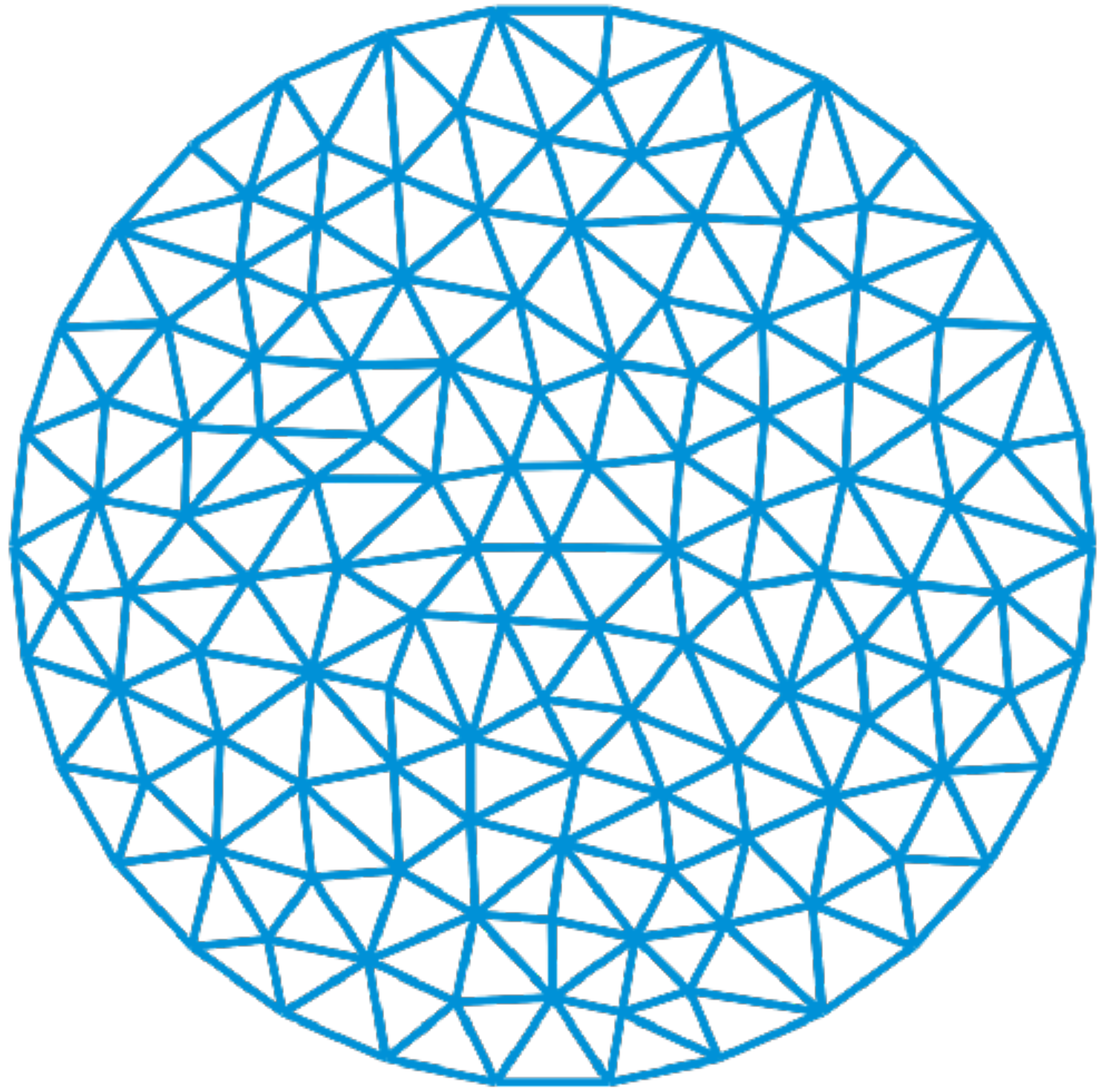
How to compare combinatorial objects



Playhouse_animation
https://www.reddit.com/r/dataisbeautiful/comments/
6c51re/nyc_subway_map_distances_vs_geographic_distances/

Triangles and beyond…



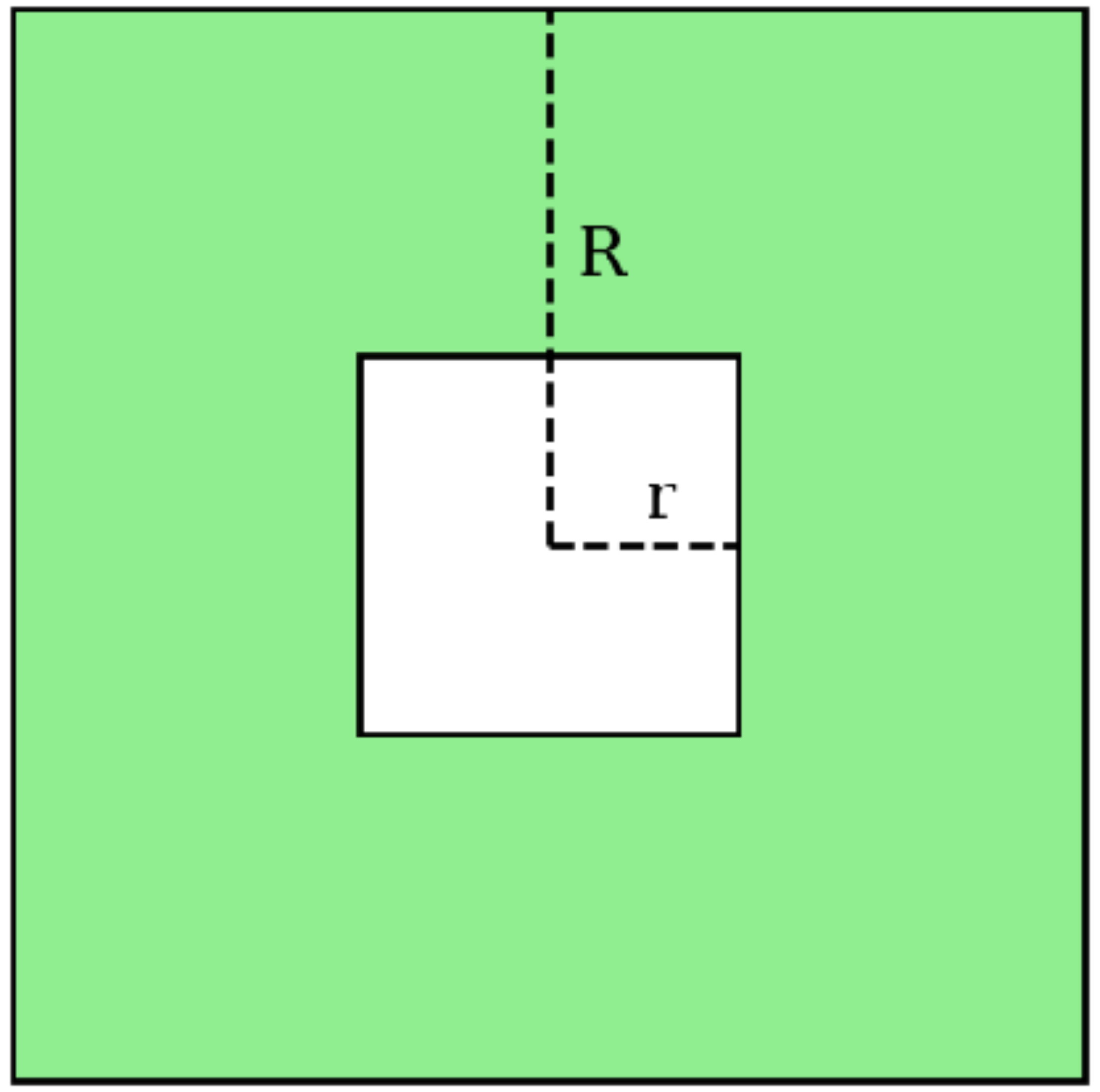


(Henry Segerman and Keenan Crane
https://www.youtube.com/watch?v=9NlqYr6-TpA)

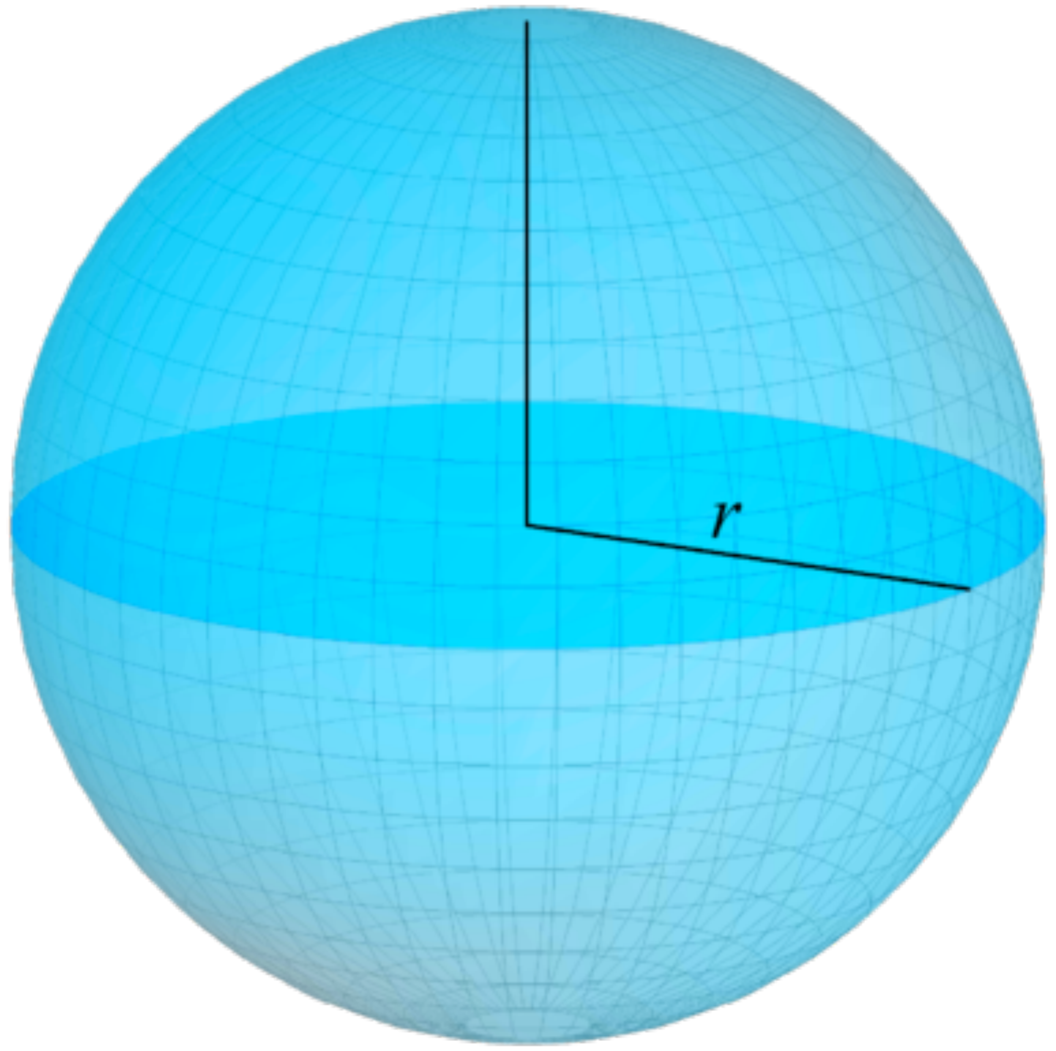
Connectivity

Homology H_q and Betti numbers $\beta_q = \dim H_q$

Count of Holes



$\beta_1 = 1 :$ 1 loop

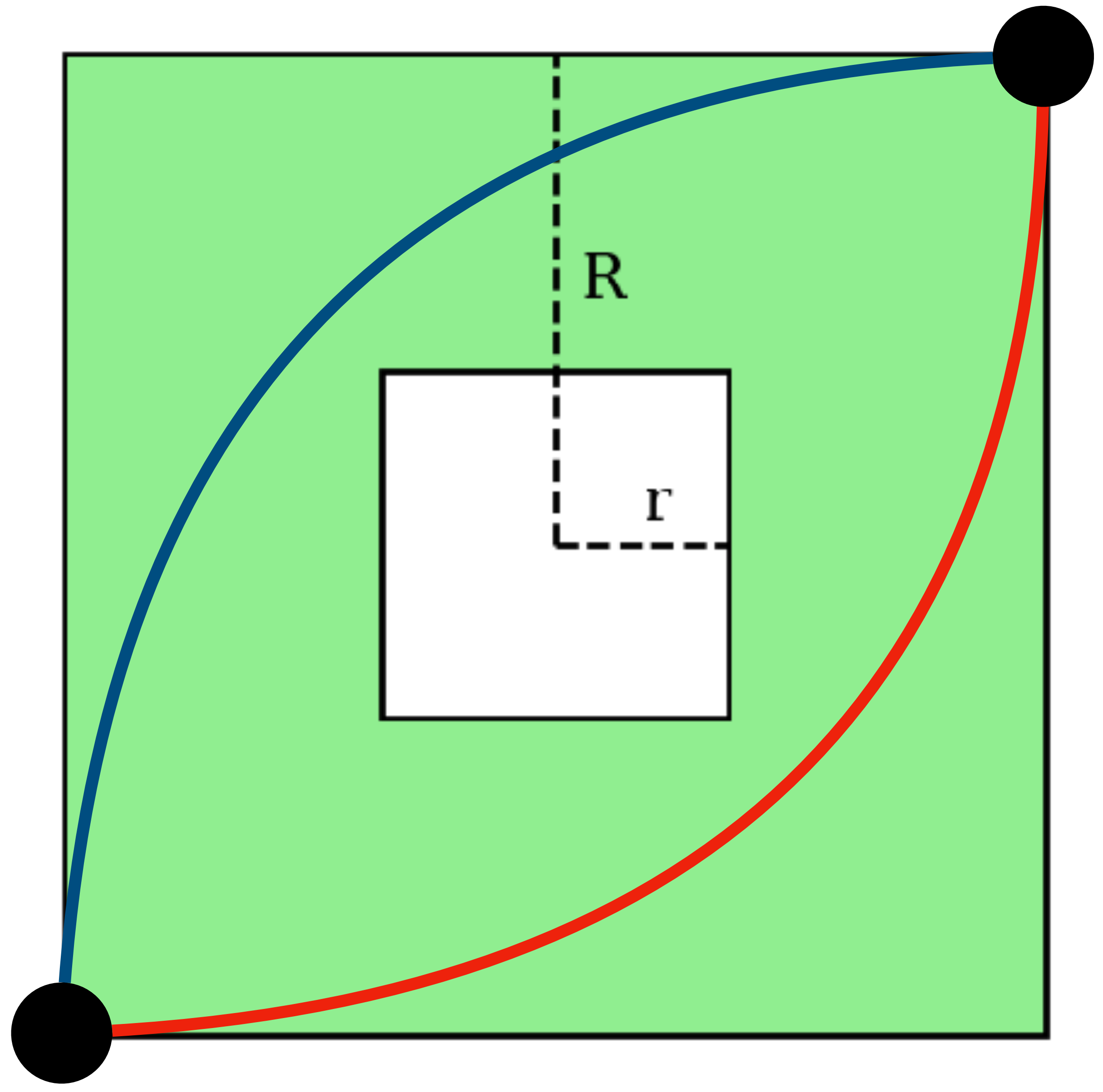


$\beta_1 = 0 :$ 0 loop

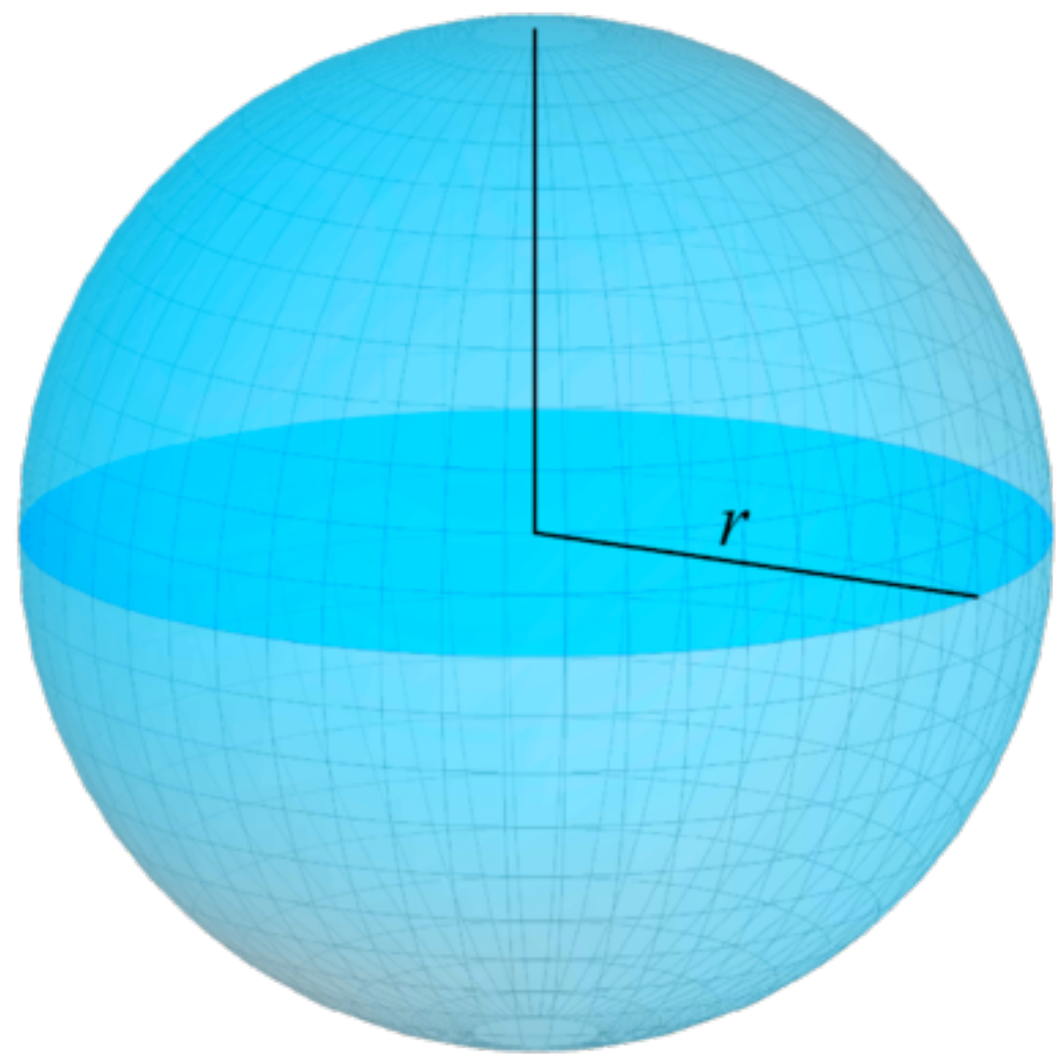
$\beta_2 = 1 :$ 1 cavity

Homology H_q and Betti numbers $\beta_q = \dim H_q$

Count of Repeated Connections



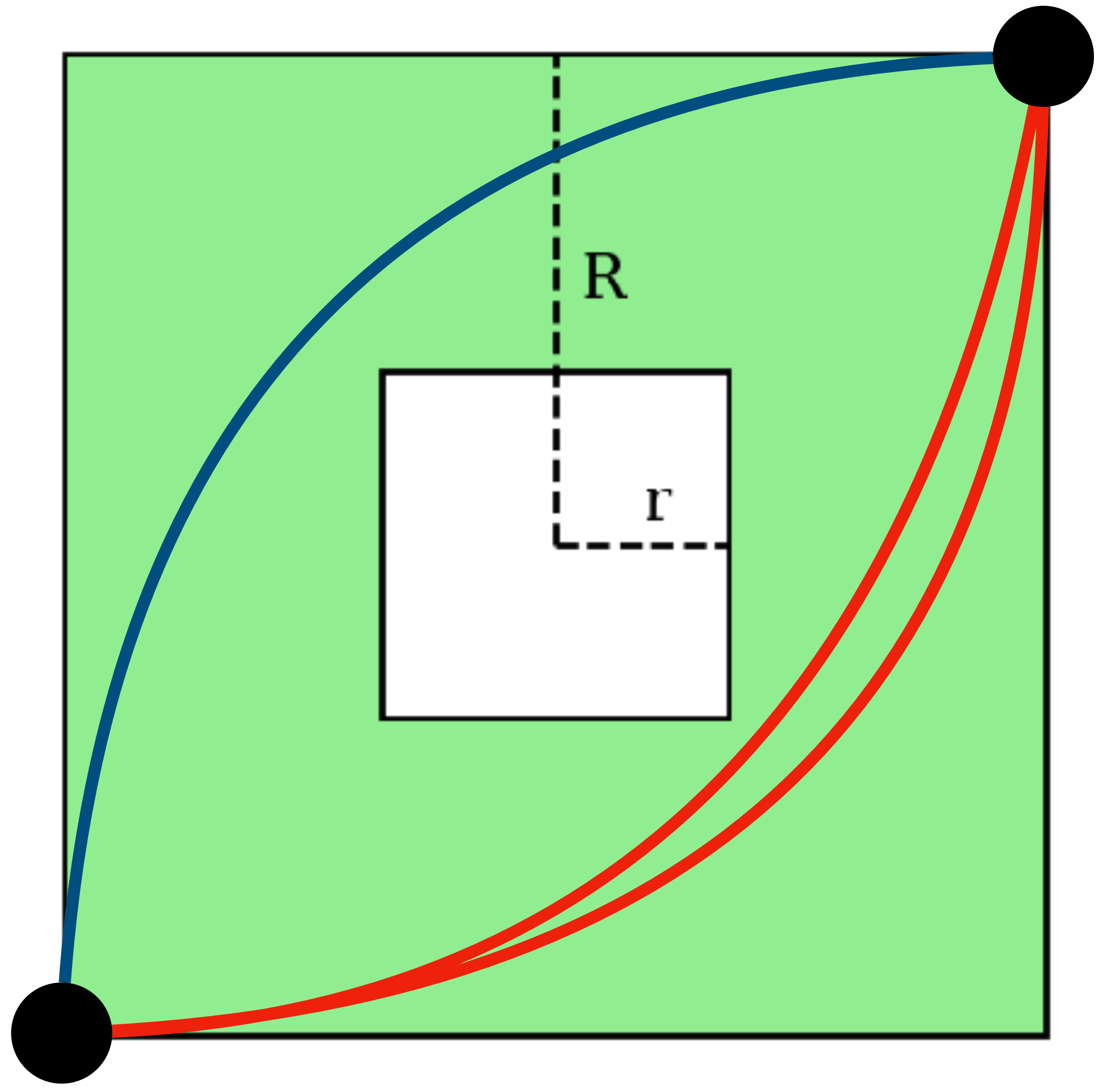
1 alternative path



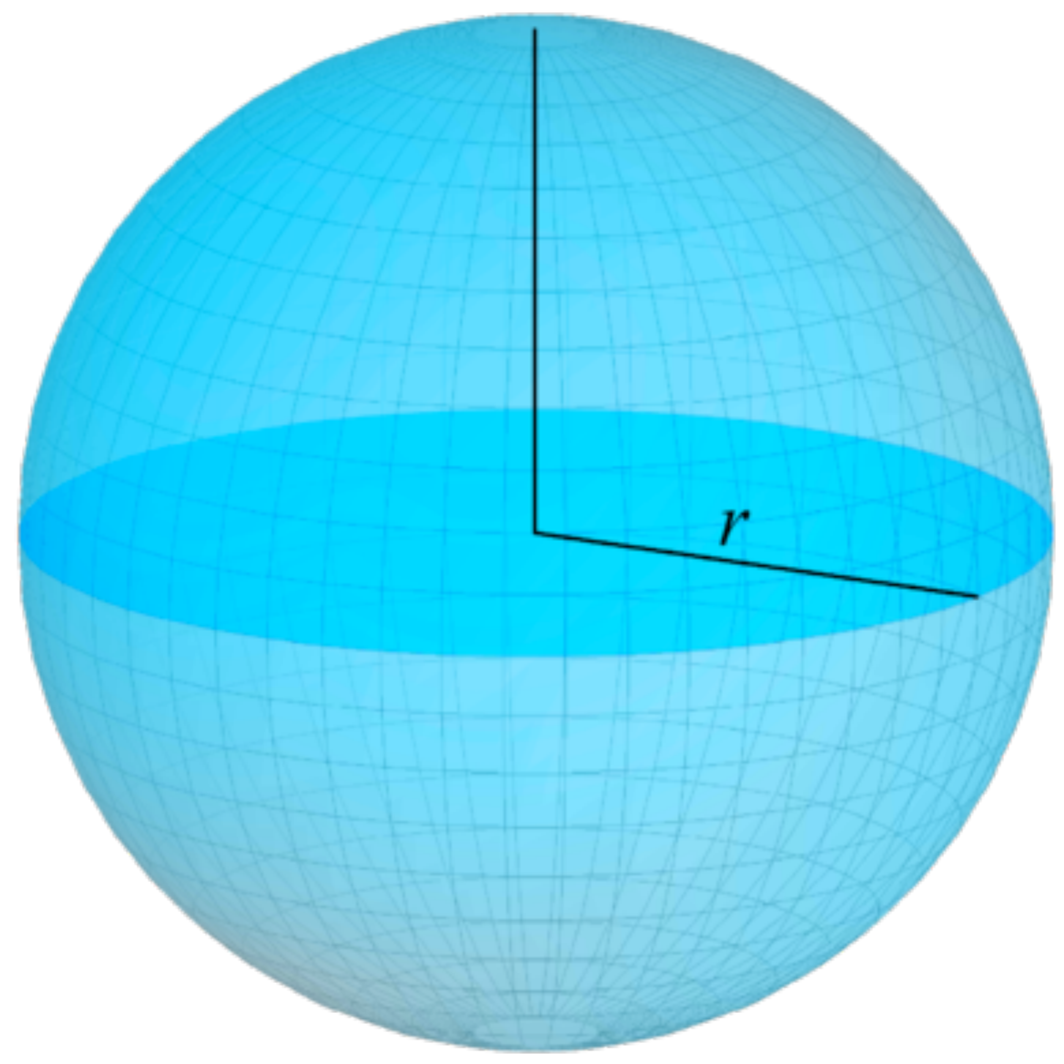
0 loop
1 cavity

Homology H_q and Betti numbers $\beta_q = \dim H_q$

Count of (Independent) Repeated Connections



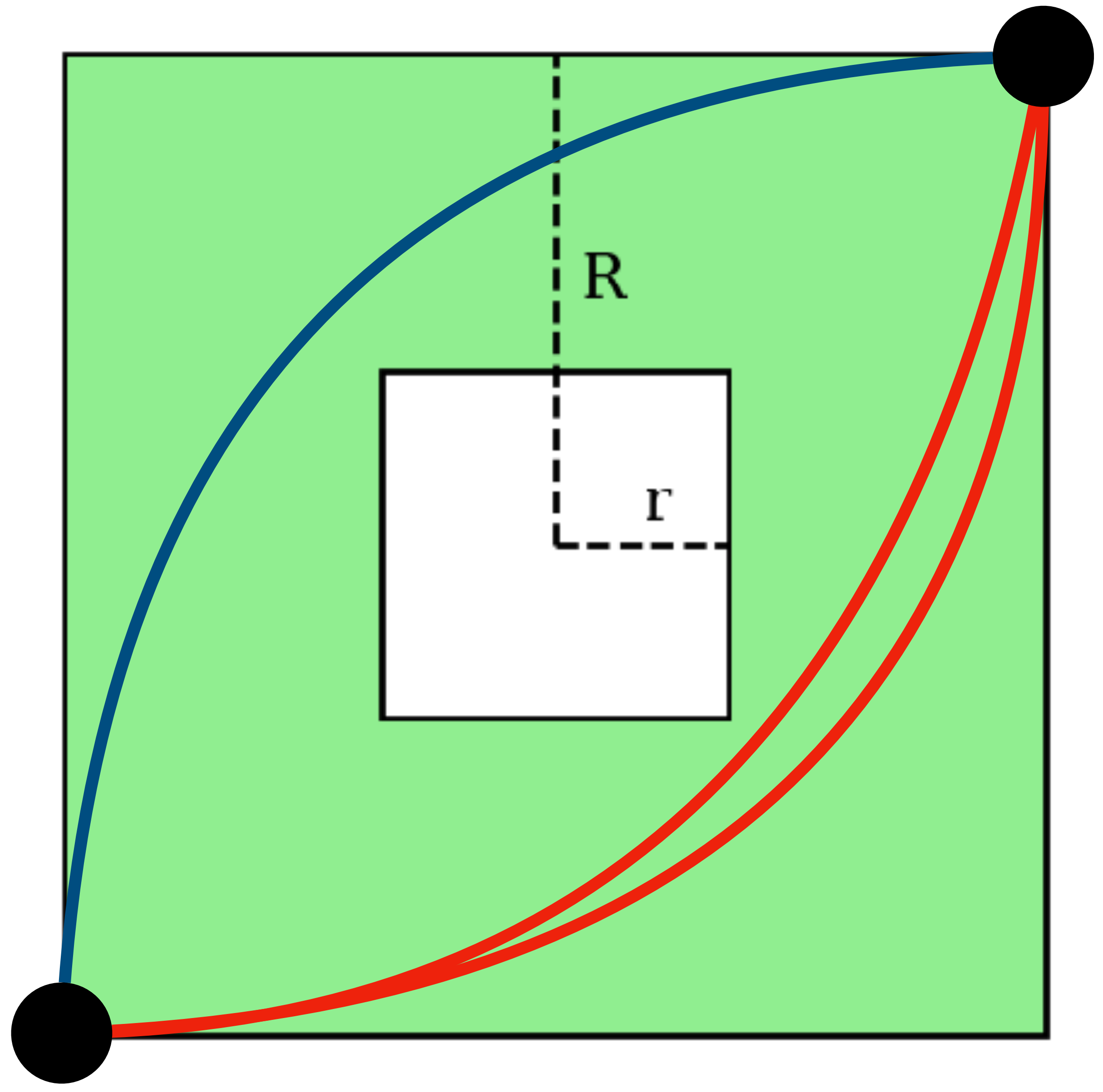
1 alternative path



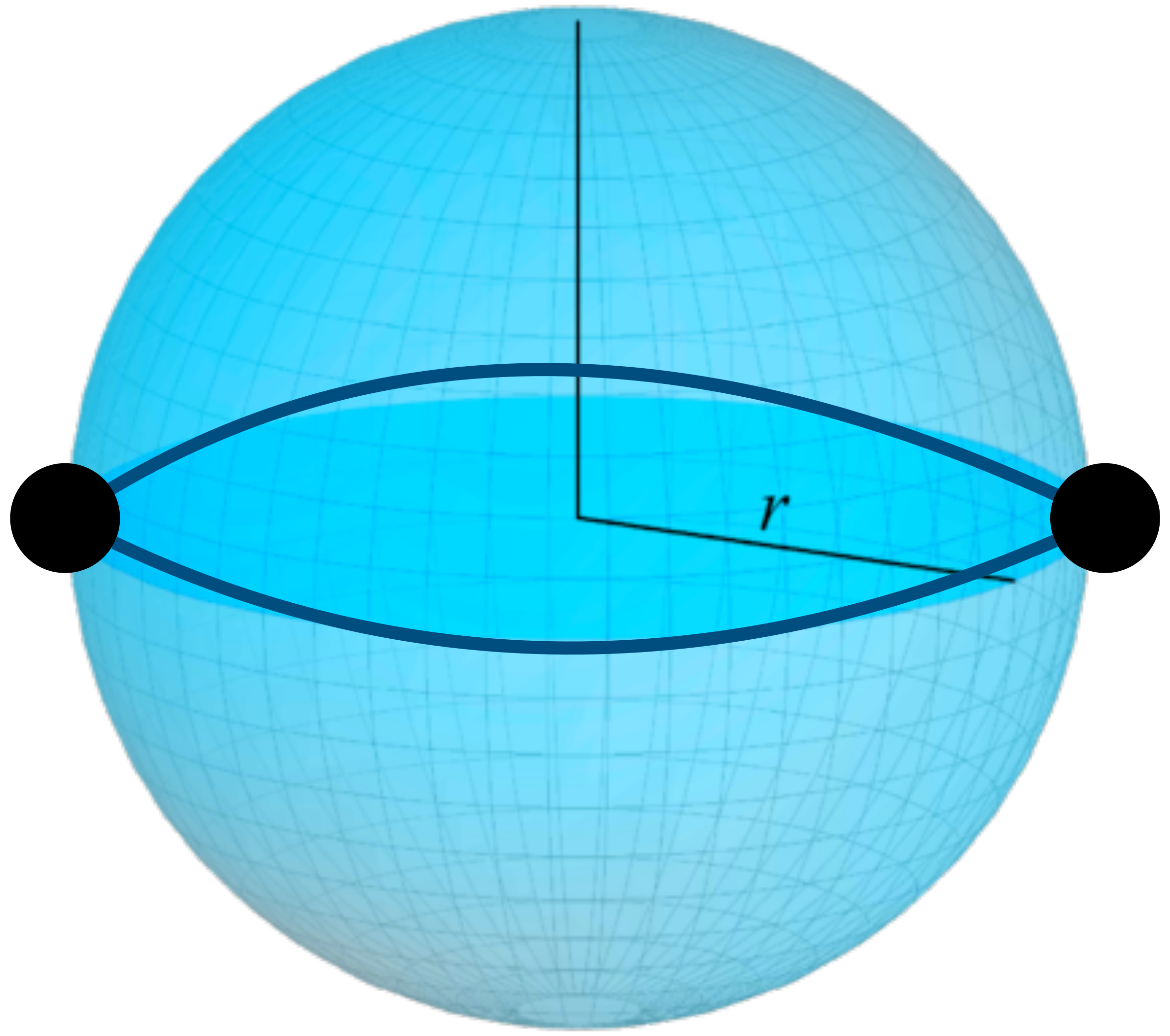
0 loop
1 cavity

Homology H_q and Betti numbers $\beta_q = \dim H_q$

Count of (Independent) Repeated Connections



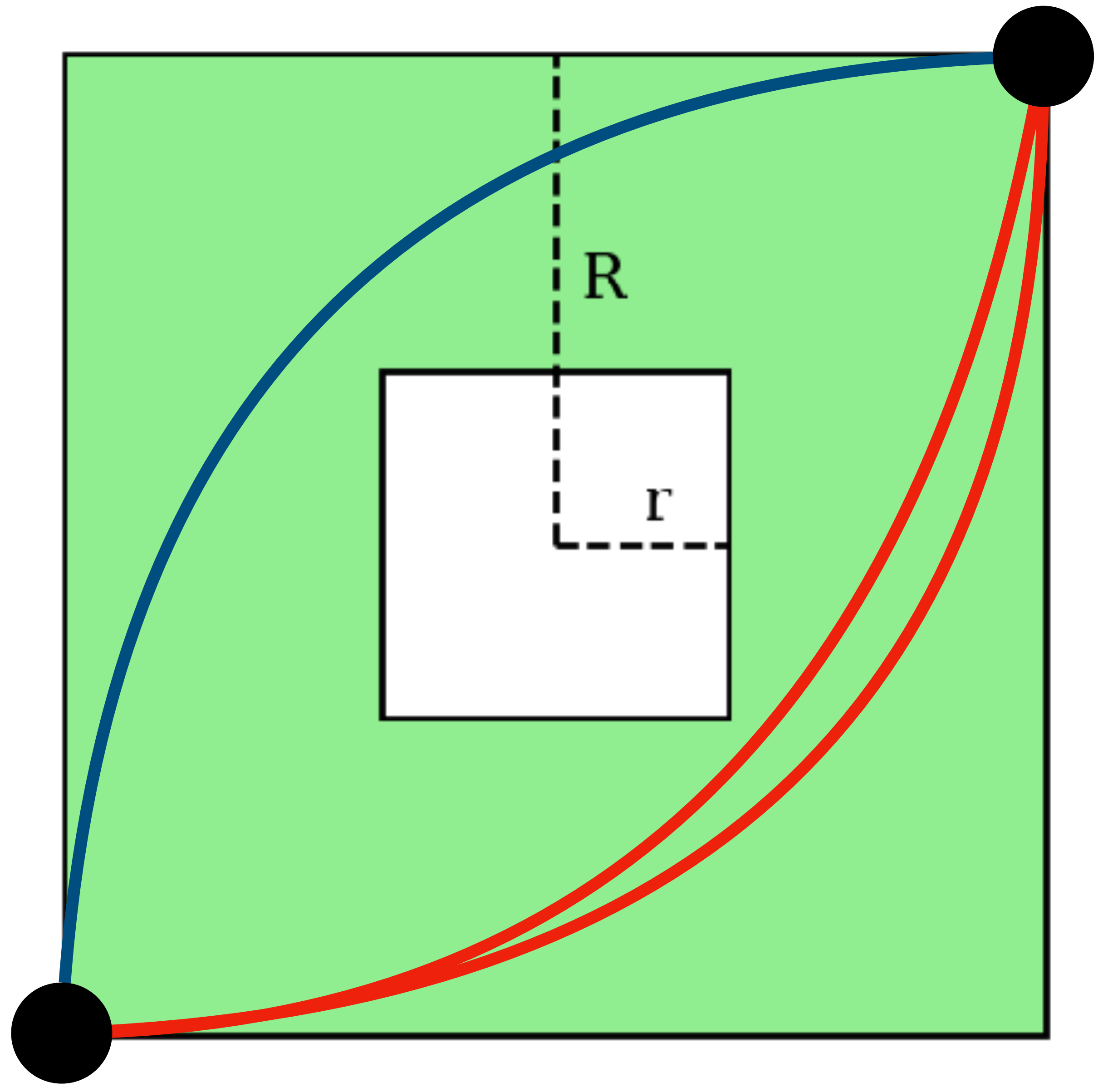
1 alternative path



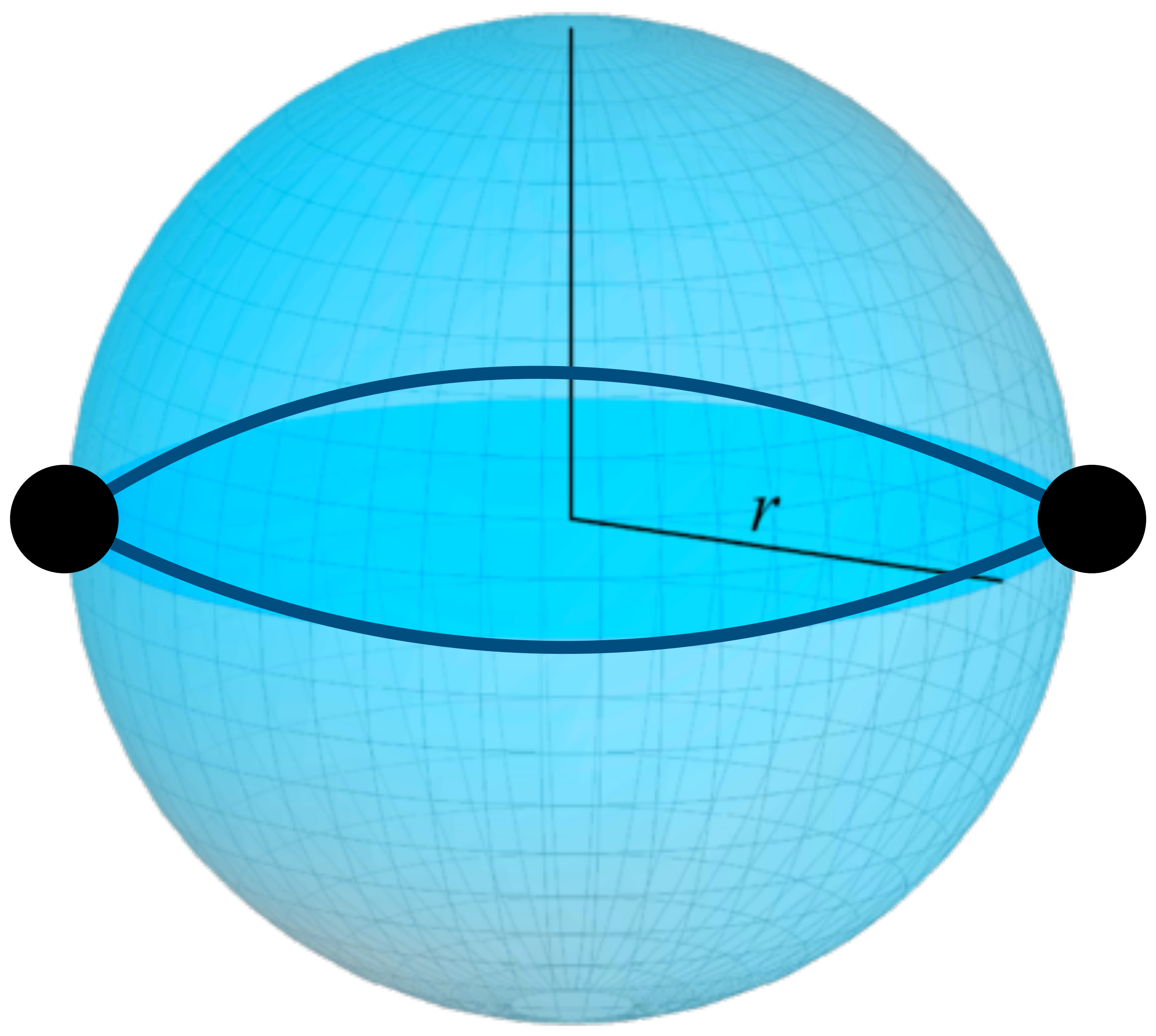
0 alternative path (slide through upper hemisphere)
1 cavity

Homology H_q and Betti numbers $\beta_q = \dim H_q$

Count of (Independent) Repeated Connections



1 alternative path



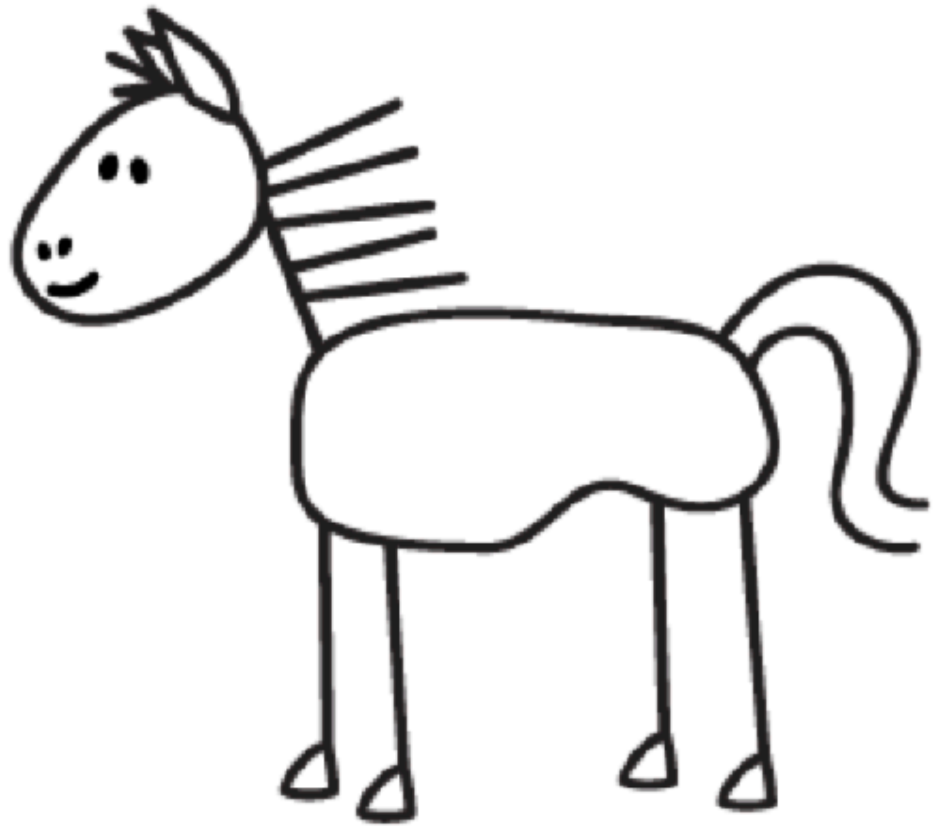
0 alternative path (slide through upper hemisphere)
1 alternative way to slide a path

Betti numbers count repeated connections “in all dimensions”.

Intermission

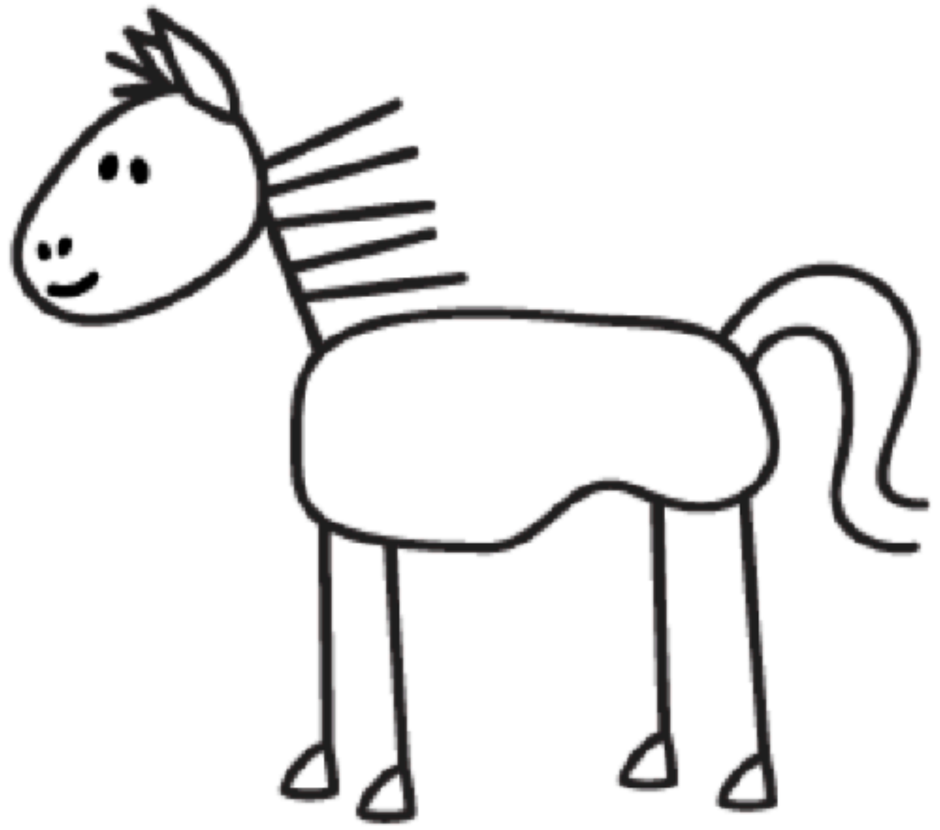
— because everybody is confused by now

Recap

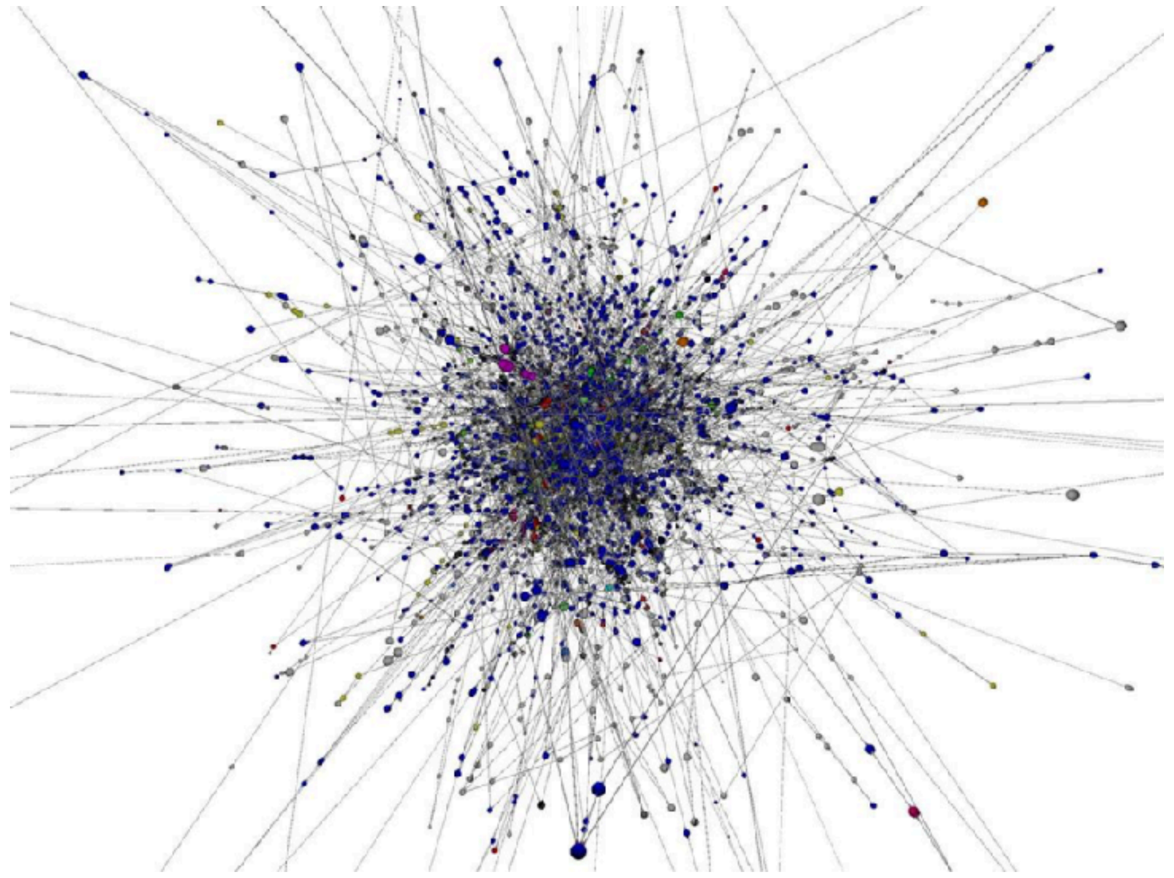


Combinatorial
representations are good

Recap

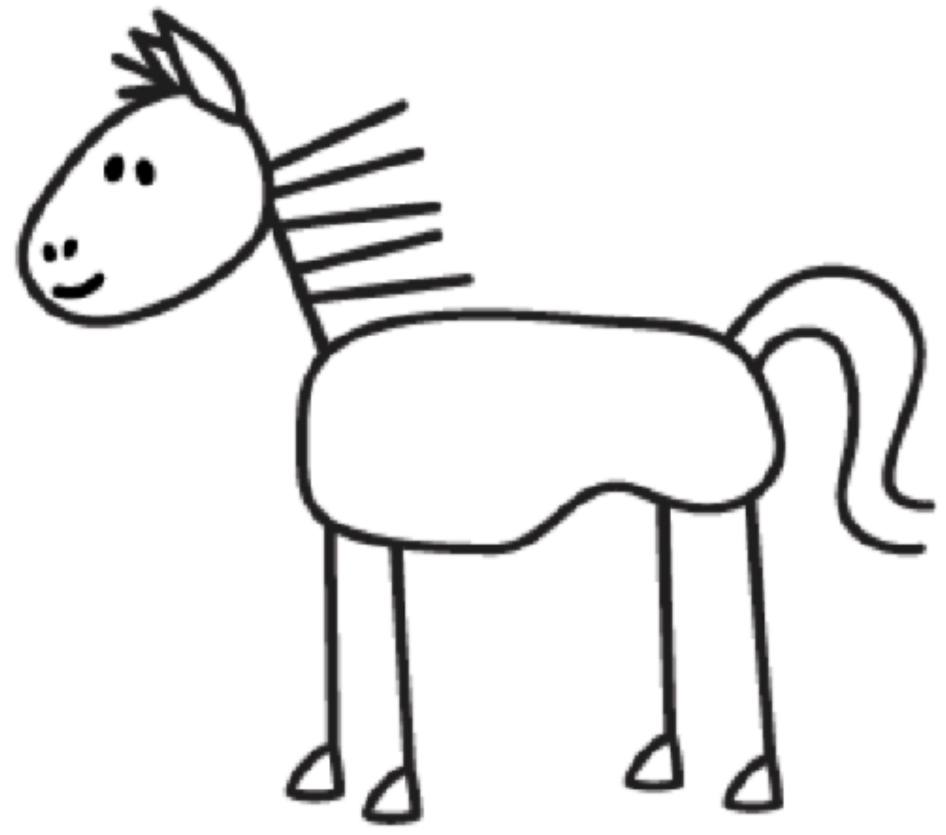


Combinatorial representations are good

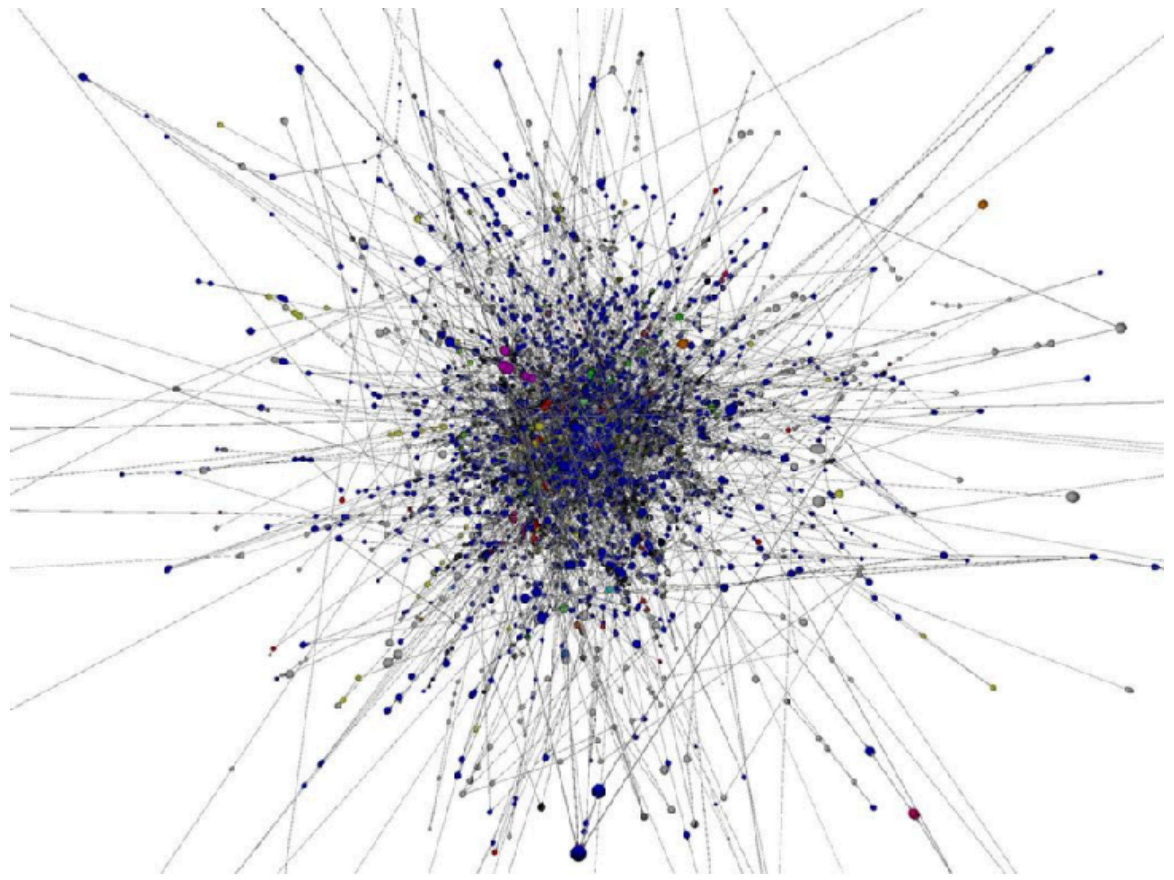


We have ways to describe networks.

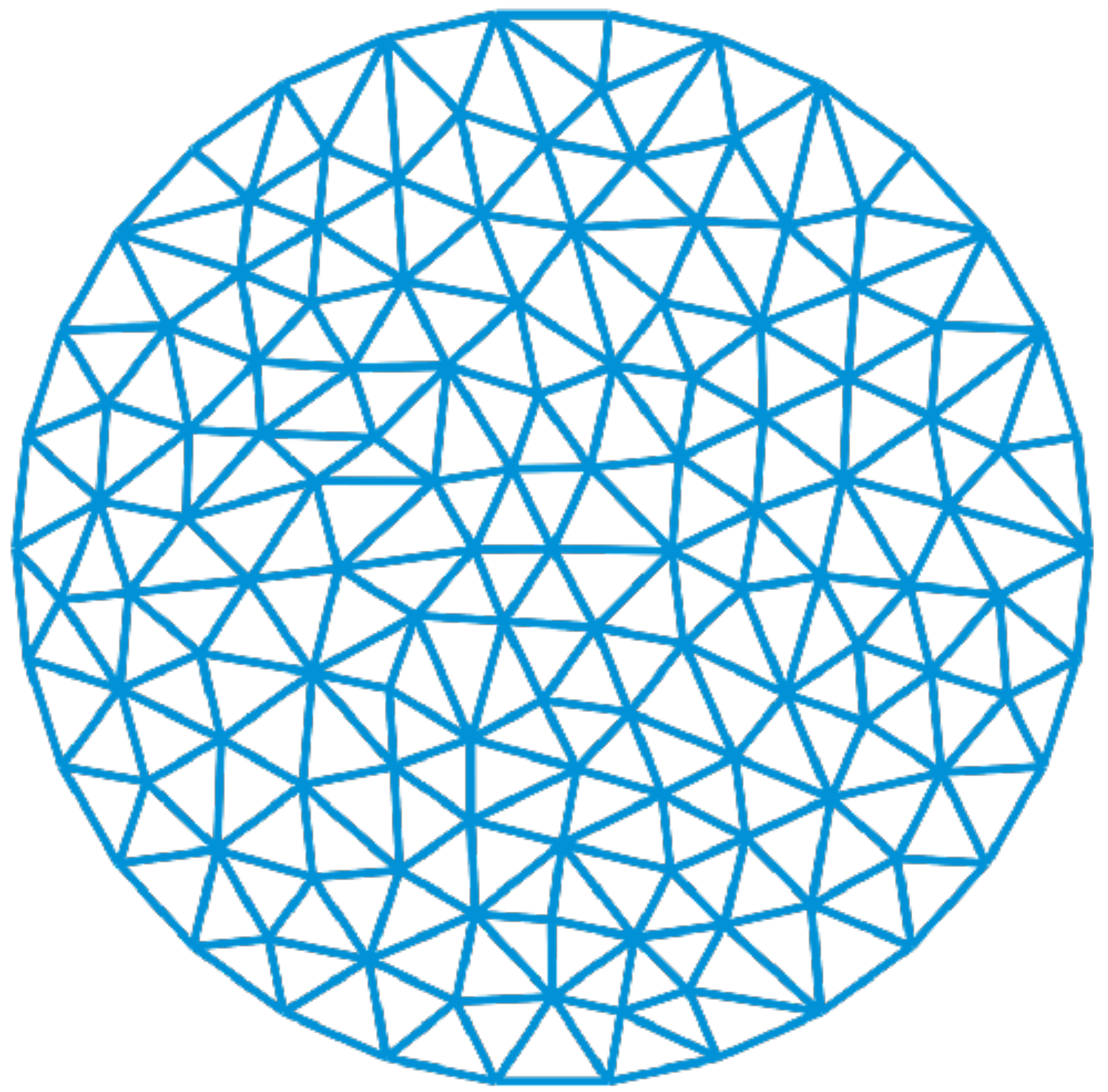
Recap



Combinatorial
representations are good

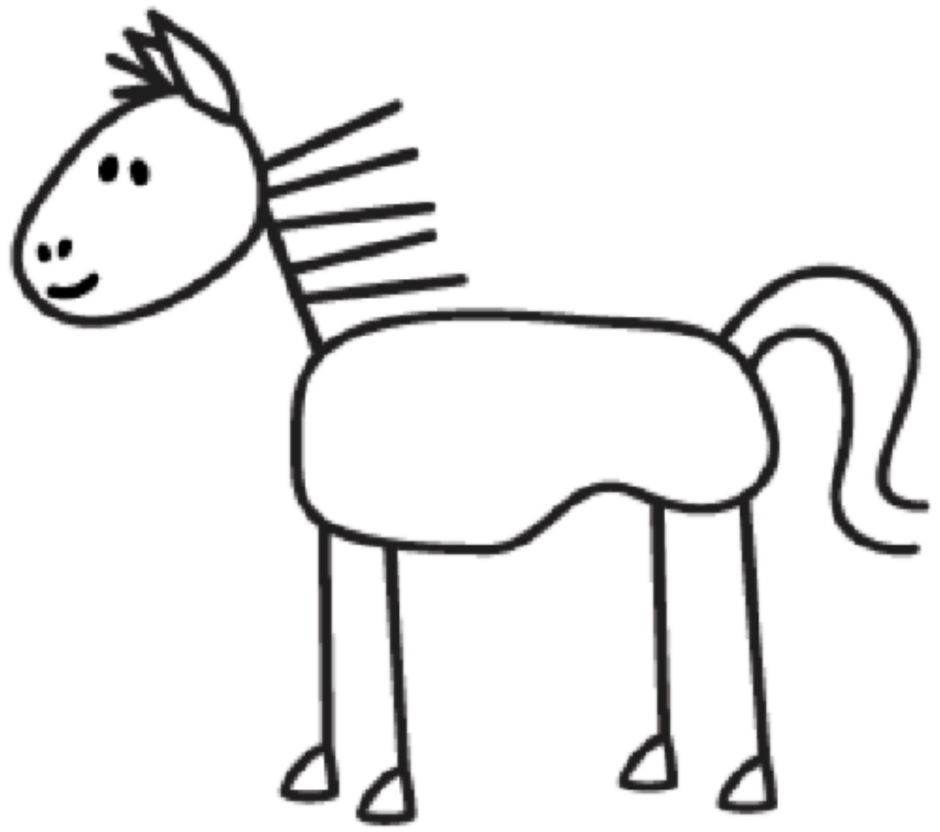


We have ways to describe
networks.

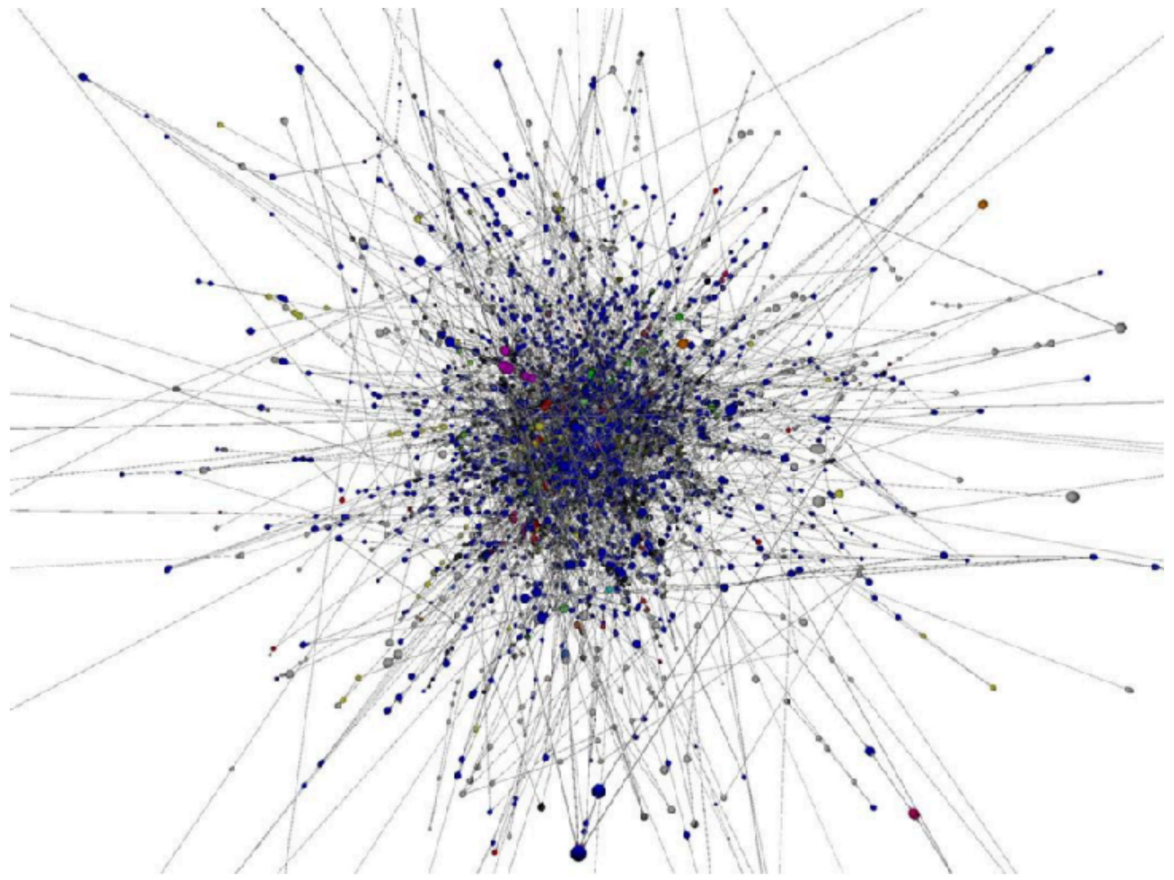


Algebraic topology is helpful
for comparing different
combinatorial representations

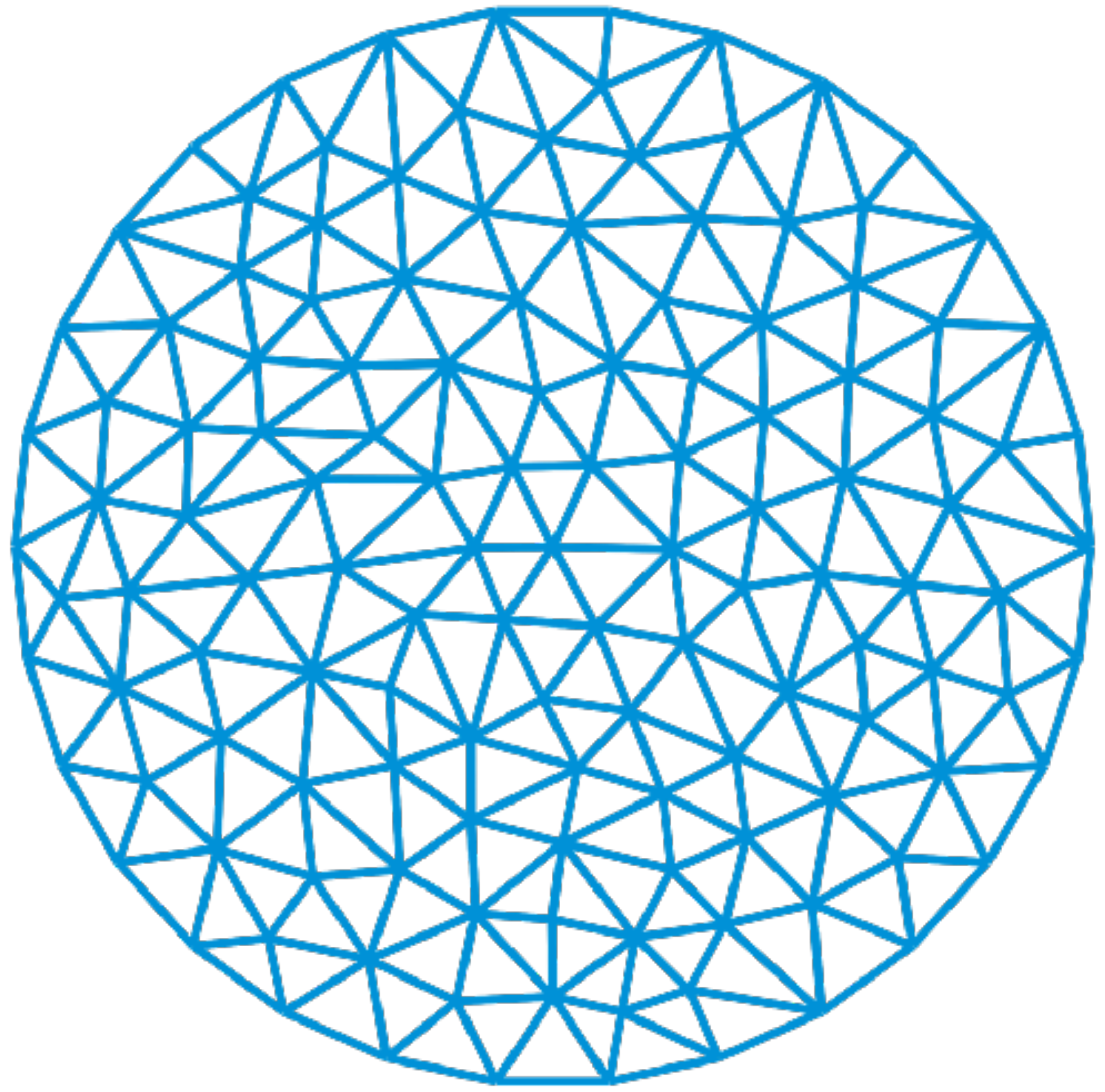
Recap



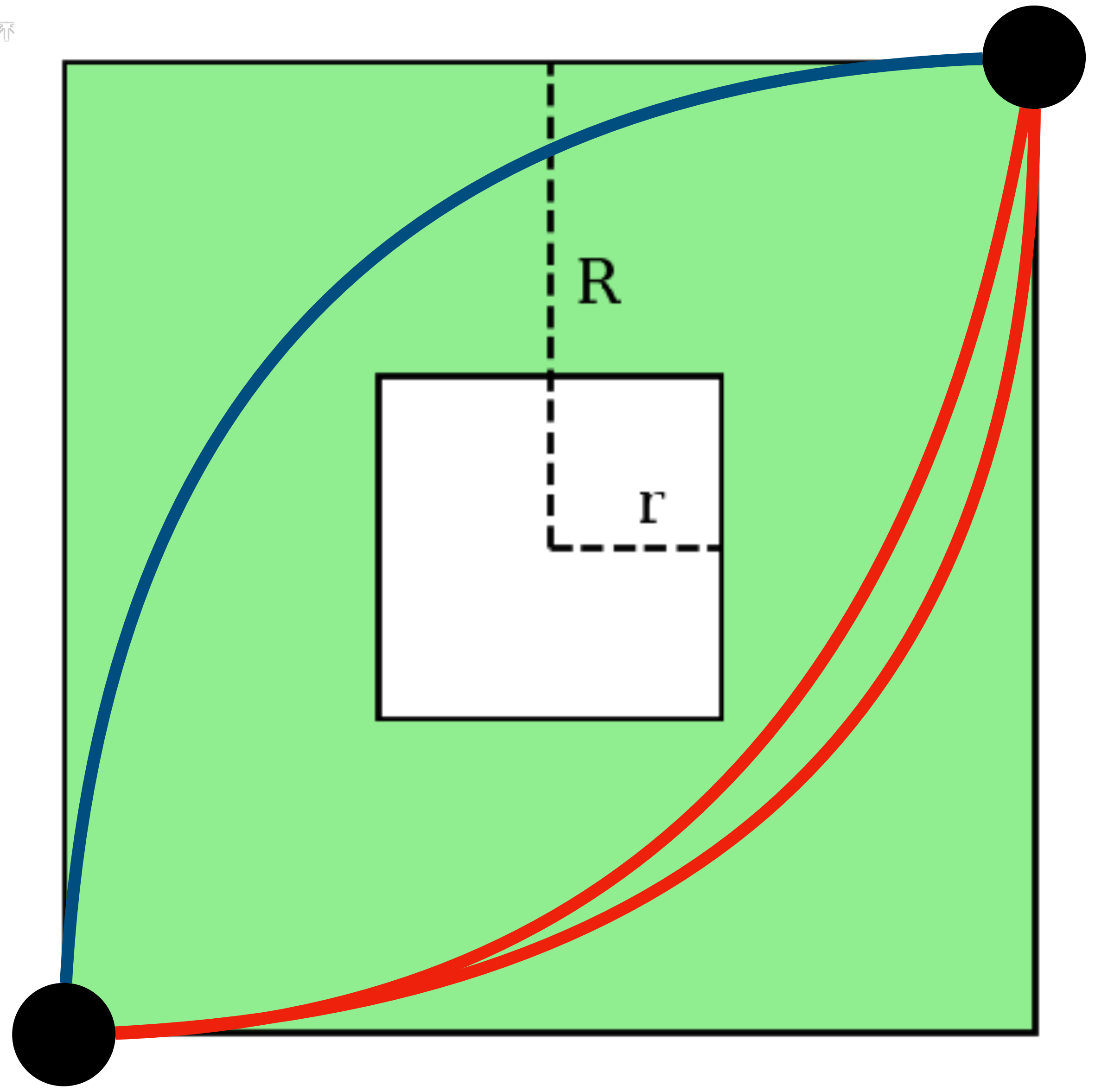
Combinatorial representations are good



We have ways to describe networks.

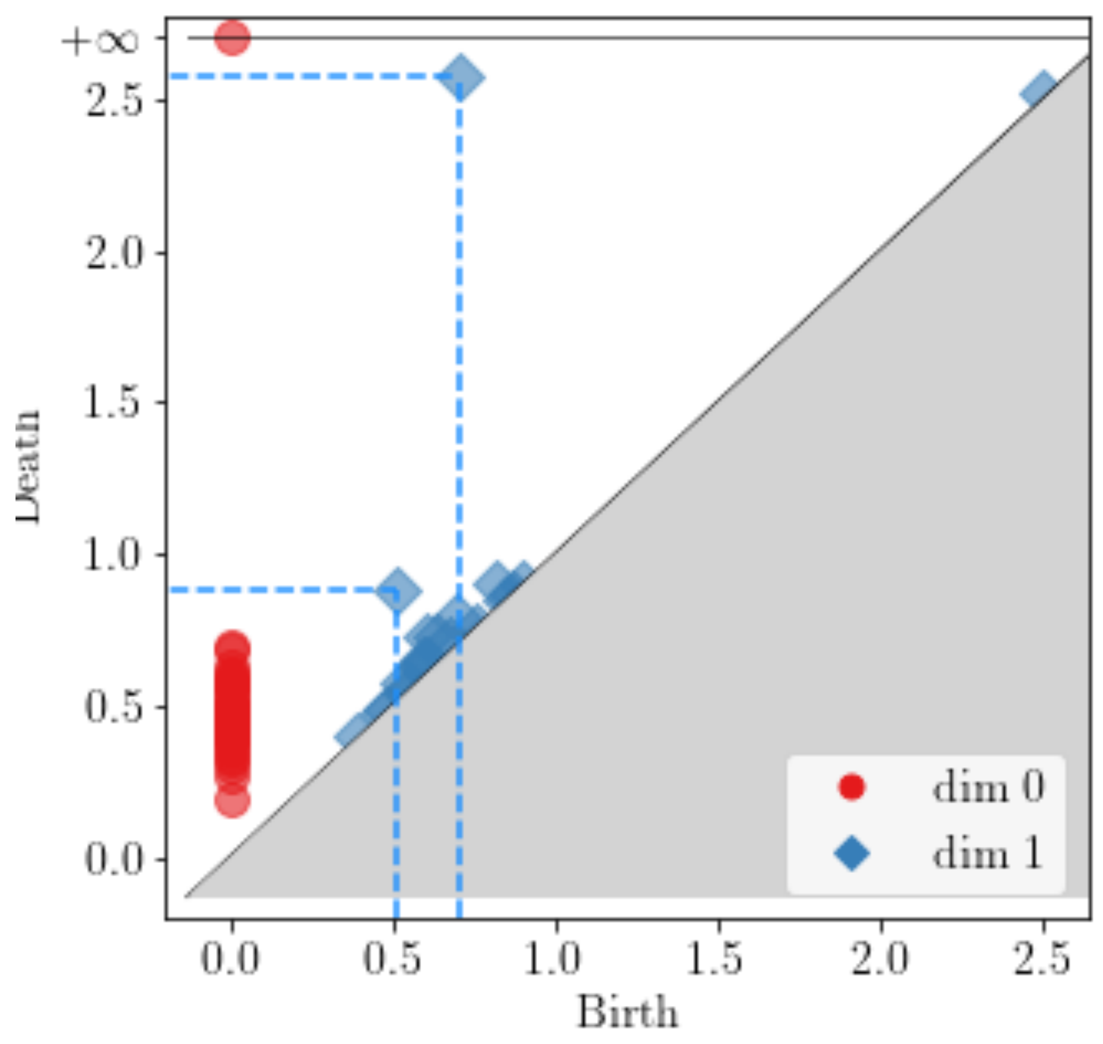


Algebraic topology is helpful for comparing different combinatorial representations

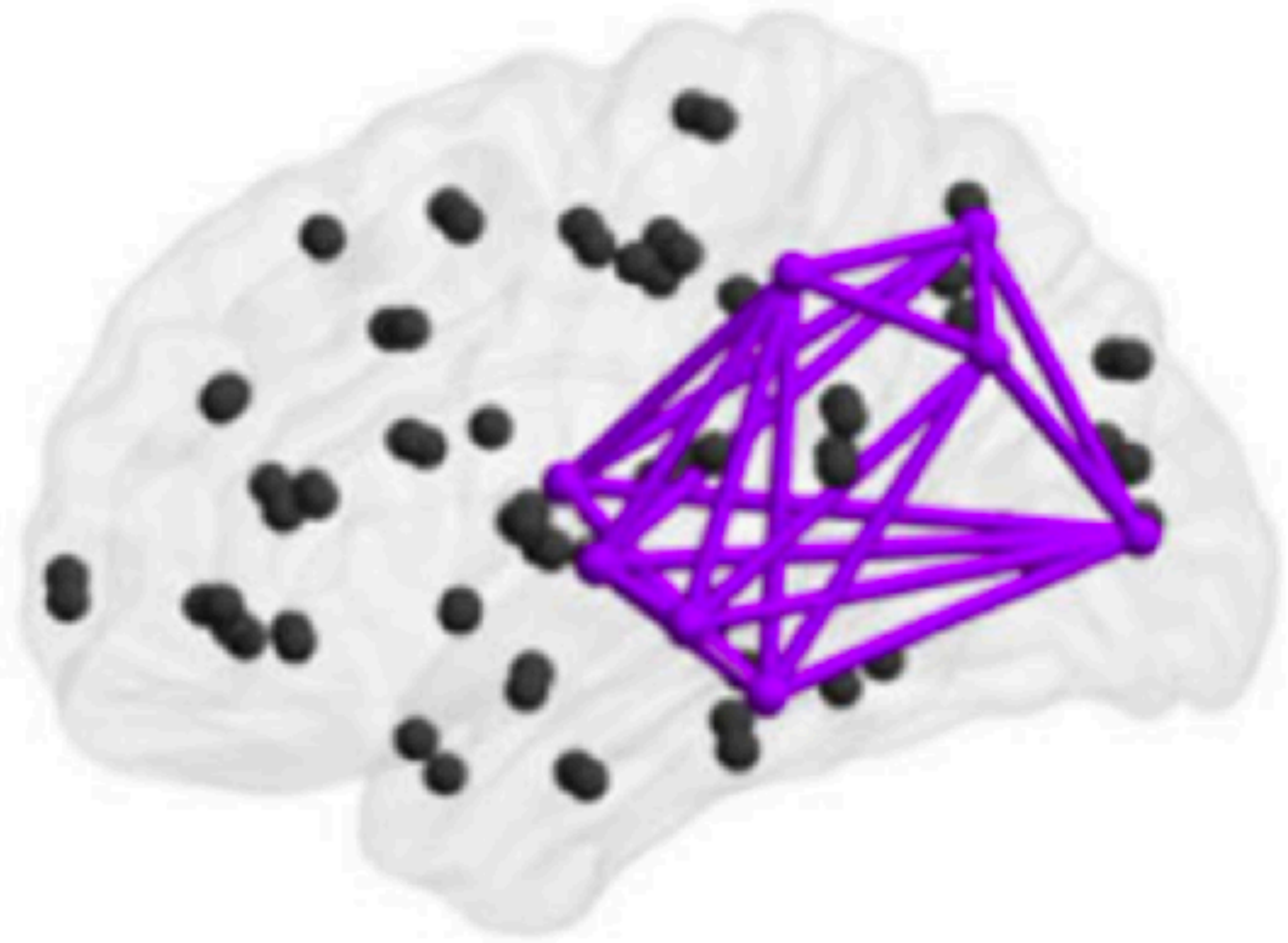


Betti numbers count repeated connections

What's next



Persistent Homology

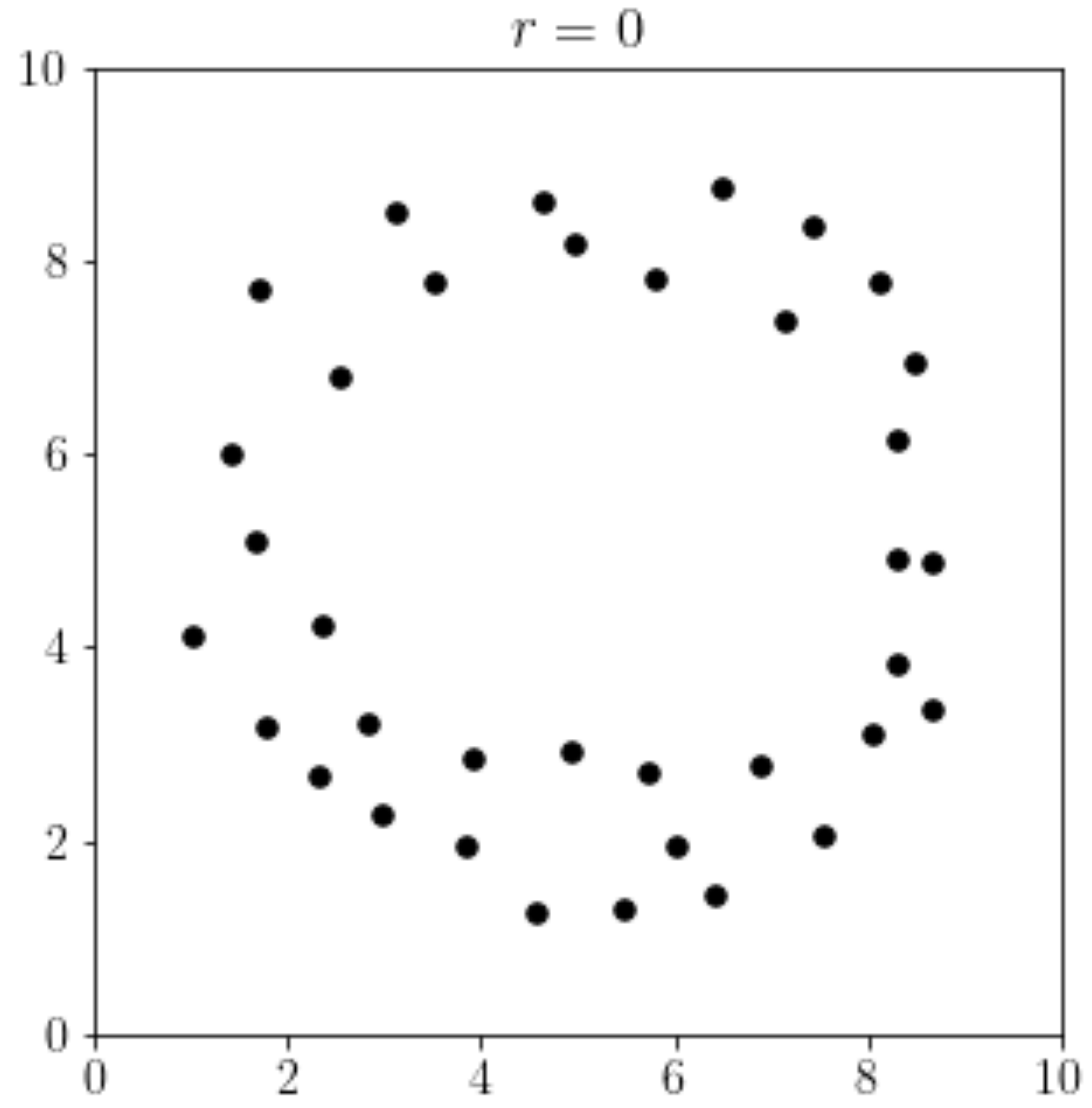


Applications

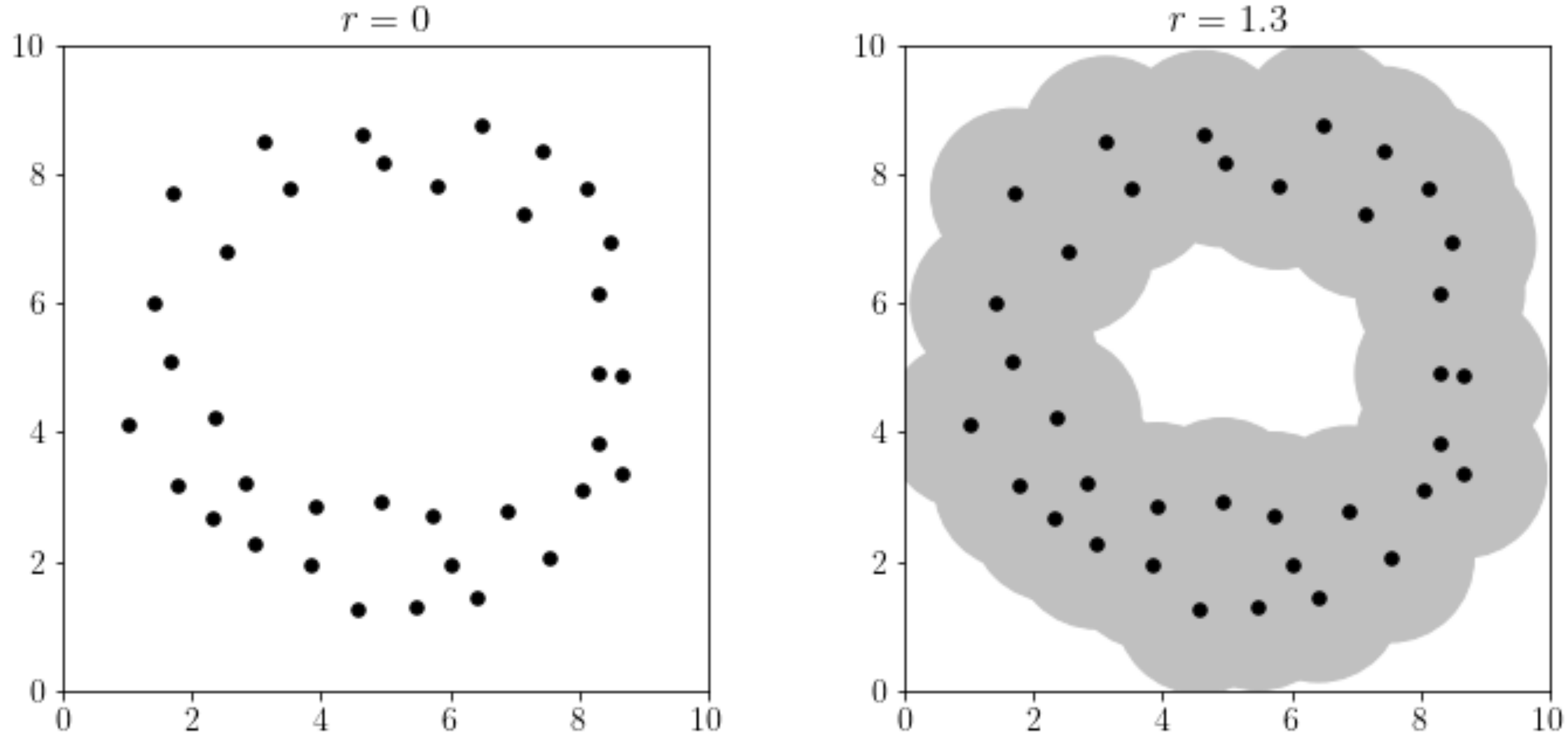
Persistent Homology

A very confusing buzz word

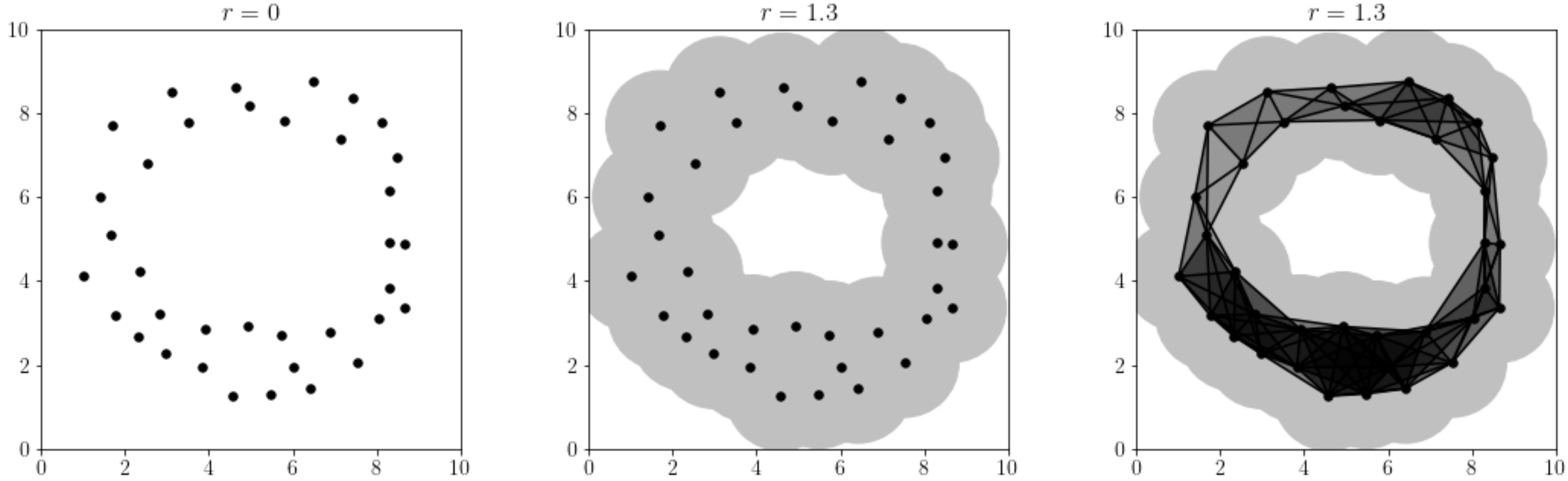
How to Build a “Higher” Network



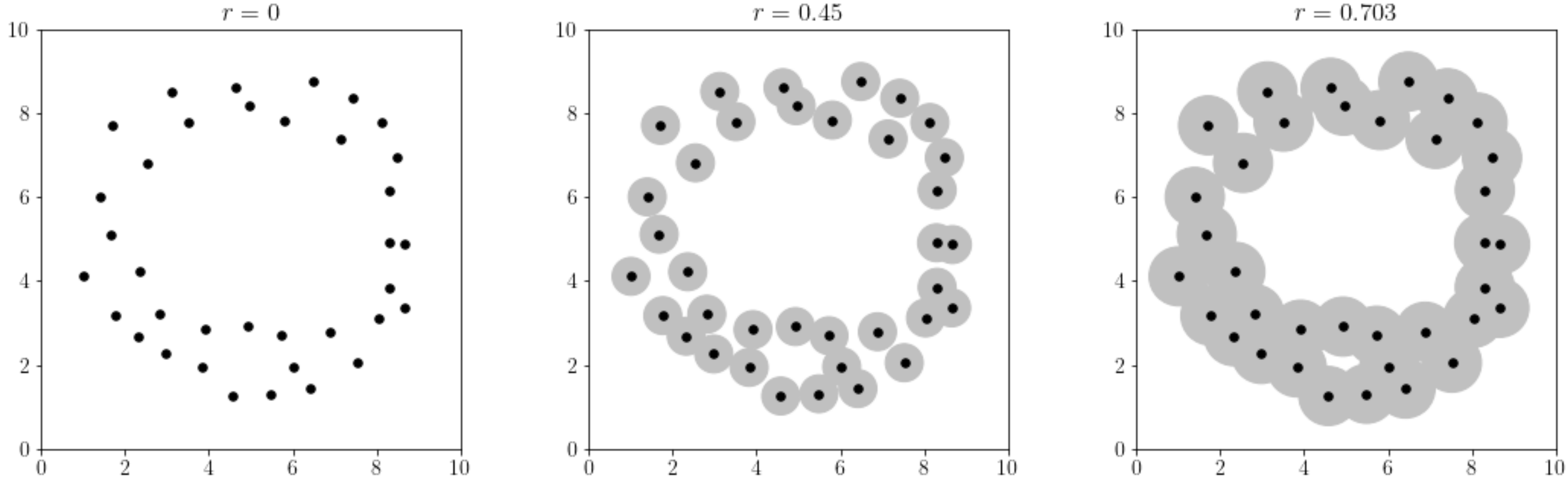
How to Build a “Higher” Network



How to Build a “Higher” Network



How to Choose the threshold?



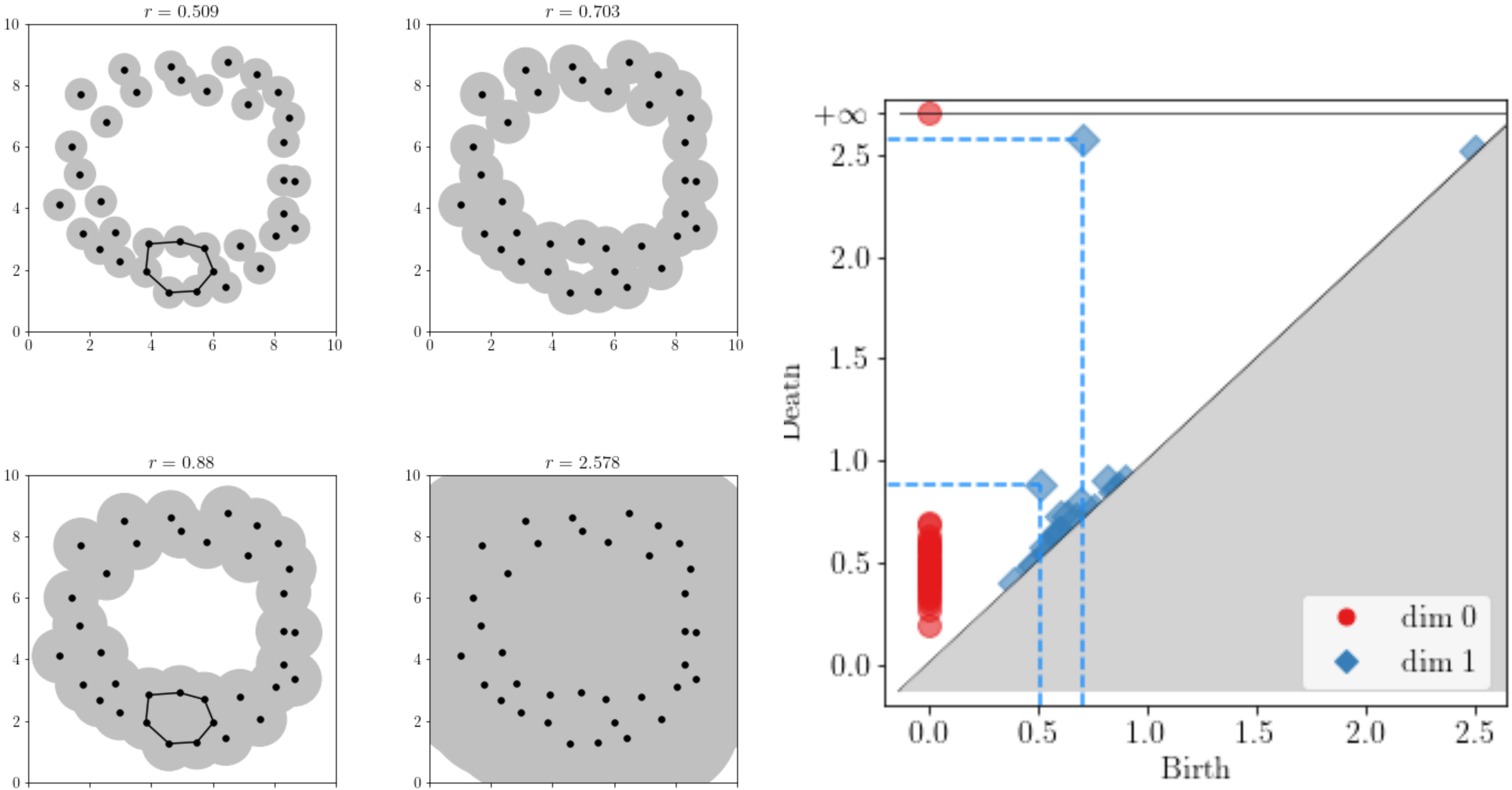


diagram credit: Andrey Yao

Application

Sizemore, Giusti, Kahn, Vettel, Betzel, Bassett, 2017

Setup

- DSI data from 8 volunteers

Setup

- DSI data from 8 volunteers
- 83 nodes (Lausanne atlas)

Setup

- DSI data from 8 volunteers
- 83 nodes (Lausanne atlas)
- Edges weighted by white matter tract density

Setup

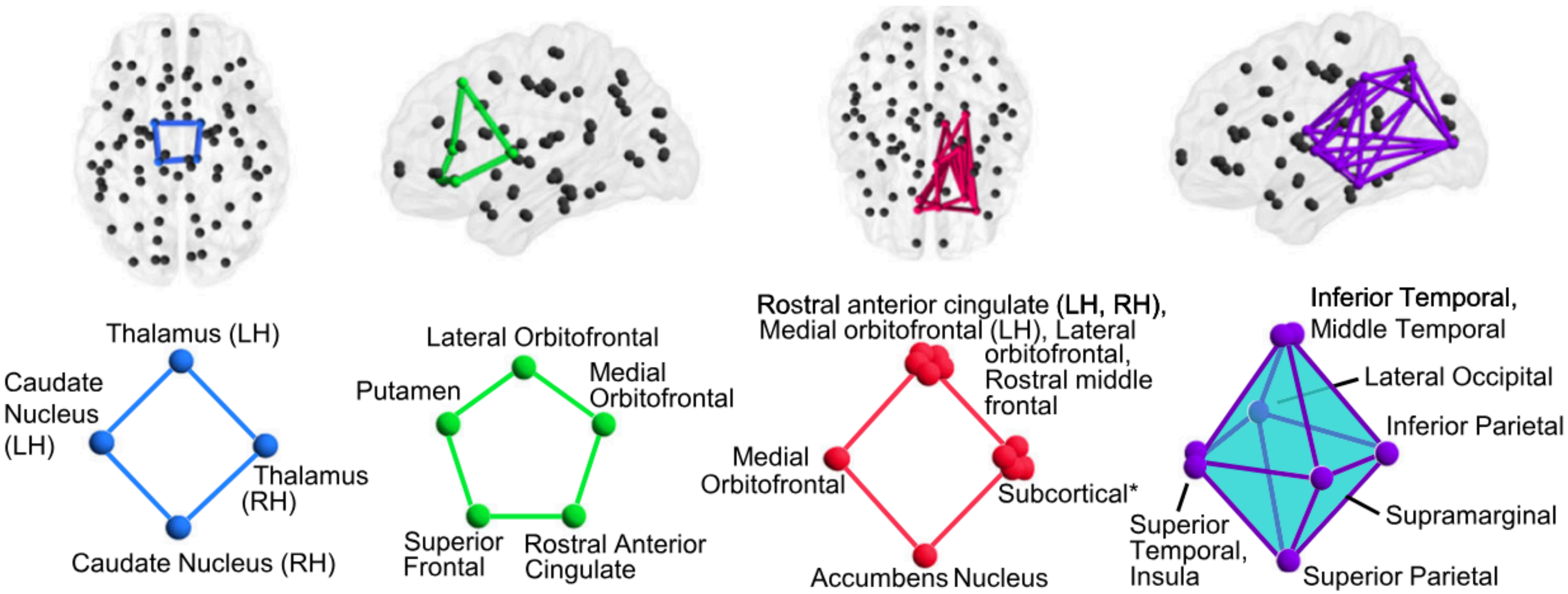
- DSI data from 8 volunteers
- 83 nodes (Lausanne atlas)
- Edges weighted by white matter tract density
- Add edges one by one starting from the mostly heavily weighted edge

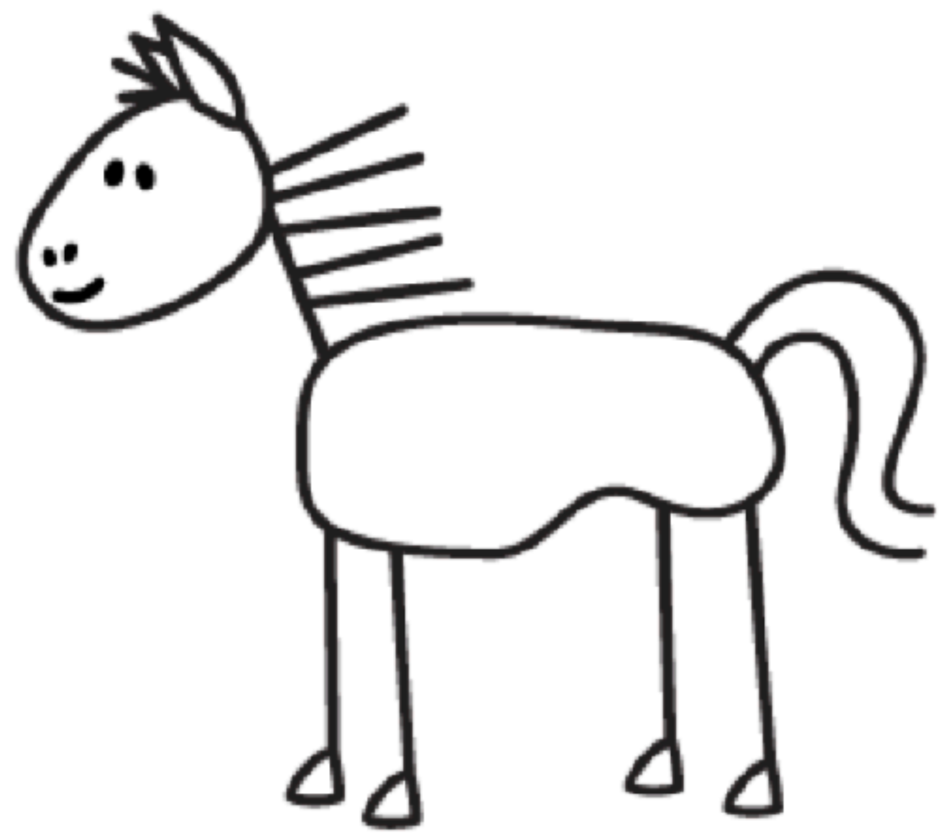
Setup

- DSI data from 8 volunteers
- 83 nodes (Lausanne atlas)
- Edges weighted by white matter tract density
- Add edges one by one starting from the mostly heavily weighted edge
- Add a triangle whenever three nodes are pairwise connected, likewise for tetrahedra and so on

Persistent Loops and Cavities

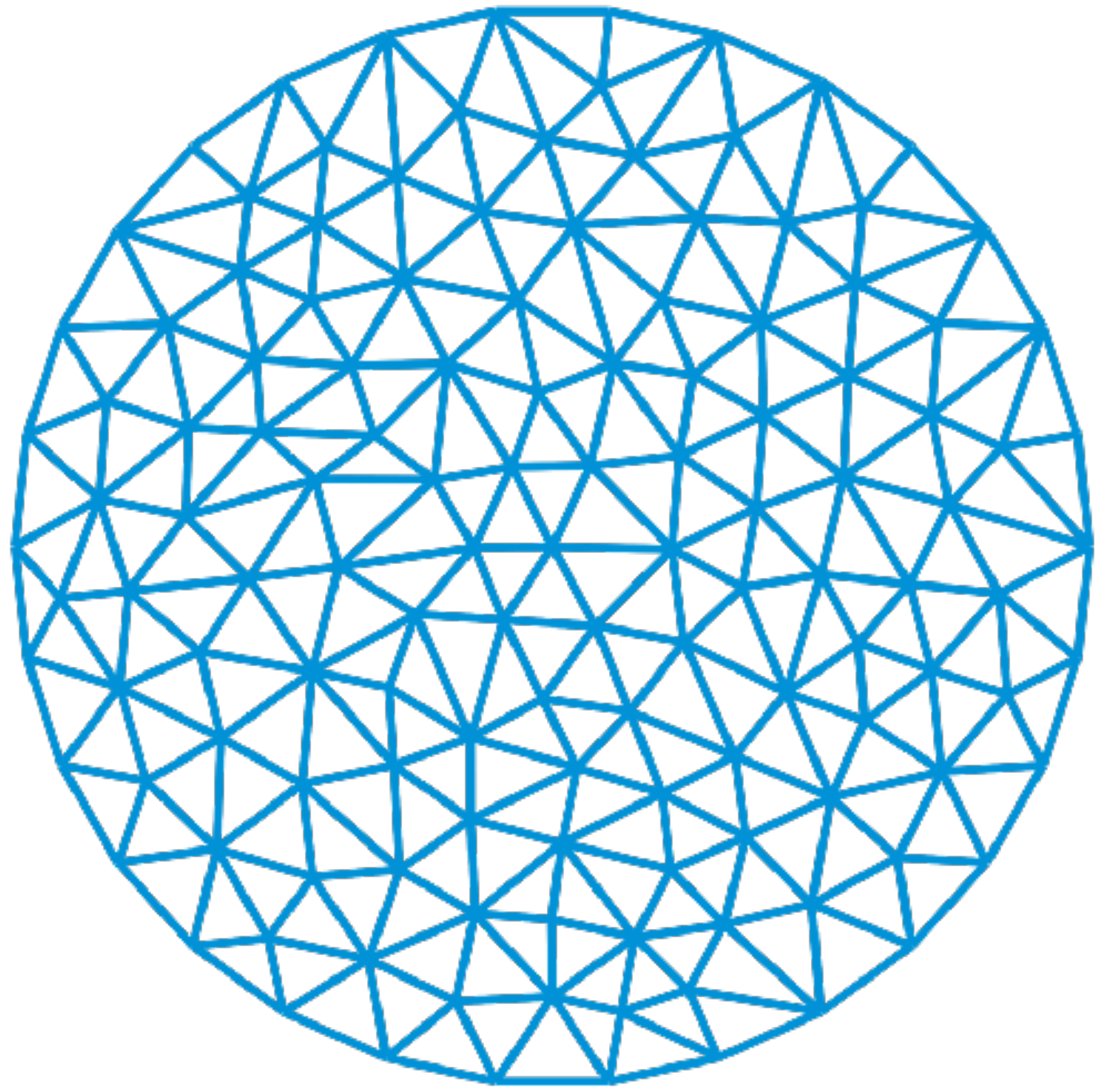
connection between evolutionarily old structures and more recently-developed neo-cortical regions



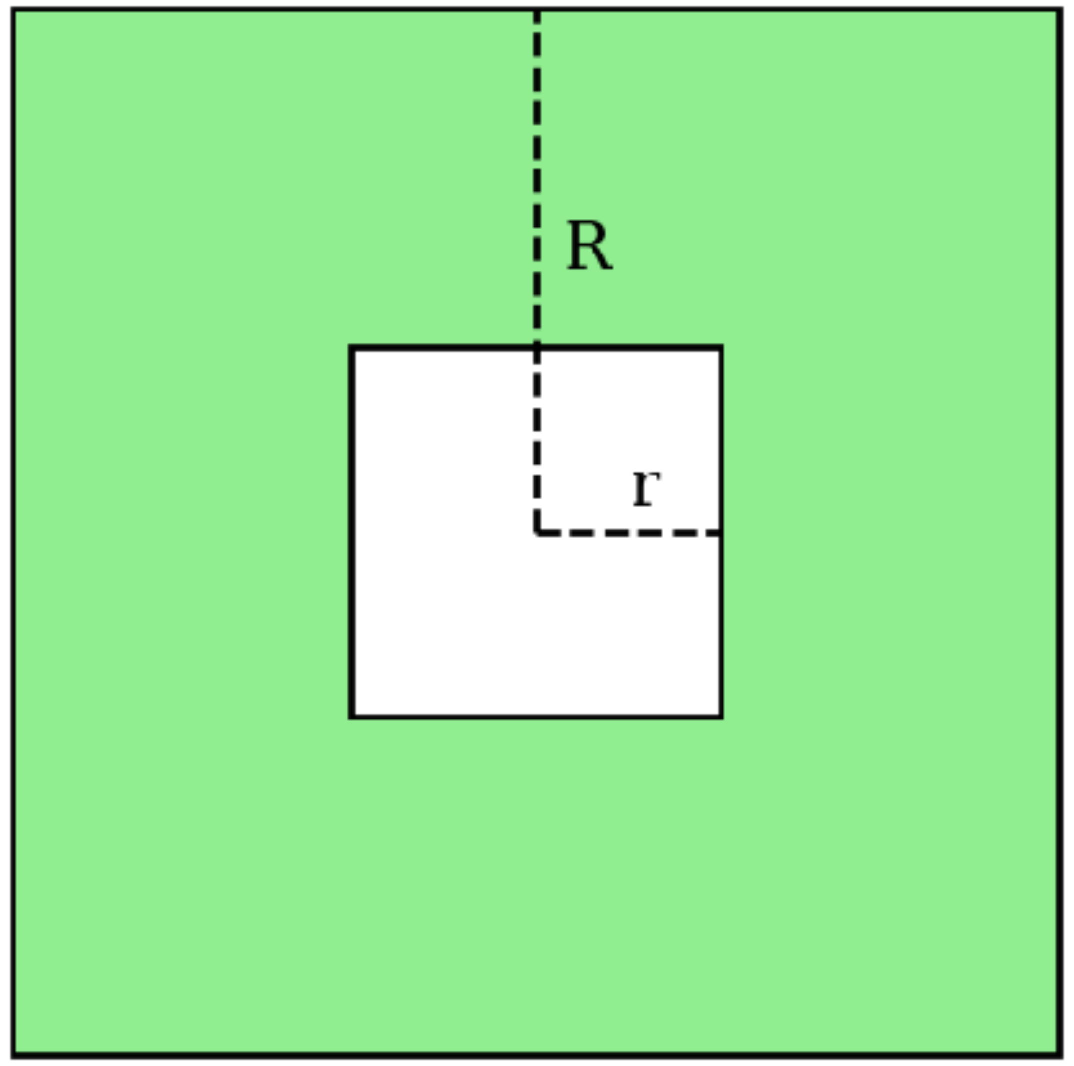


Combinatorial representations are good

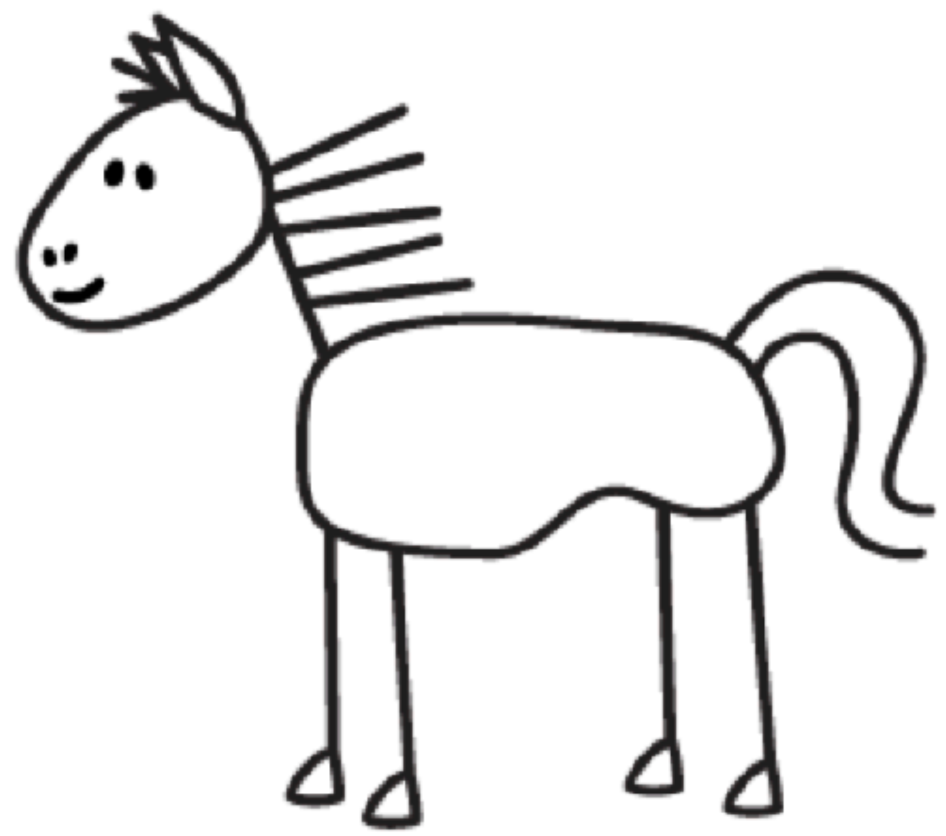
We have ways to describe networks.



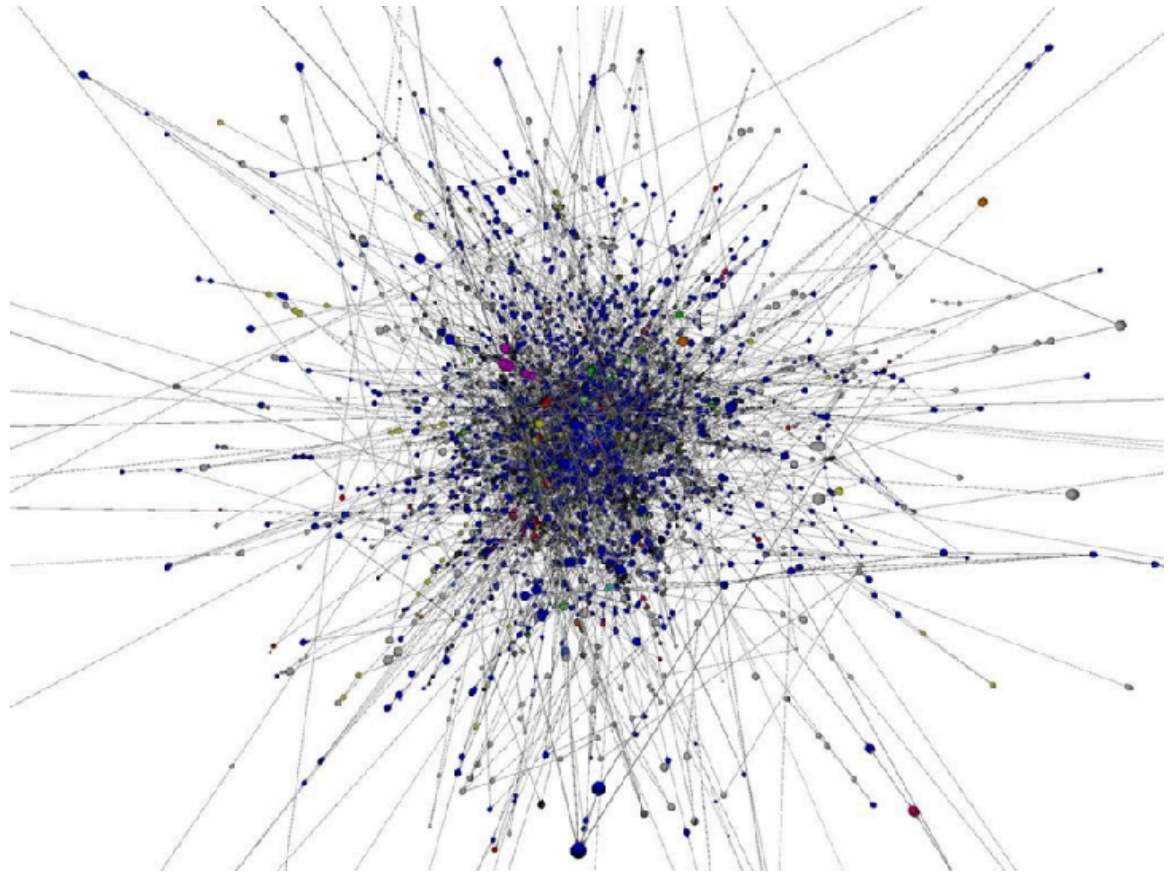
Algebraic topology is helpful for comparing different combinatorial representations



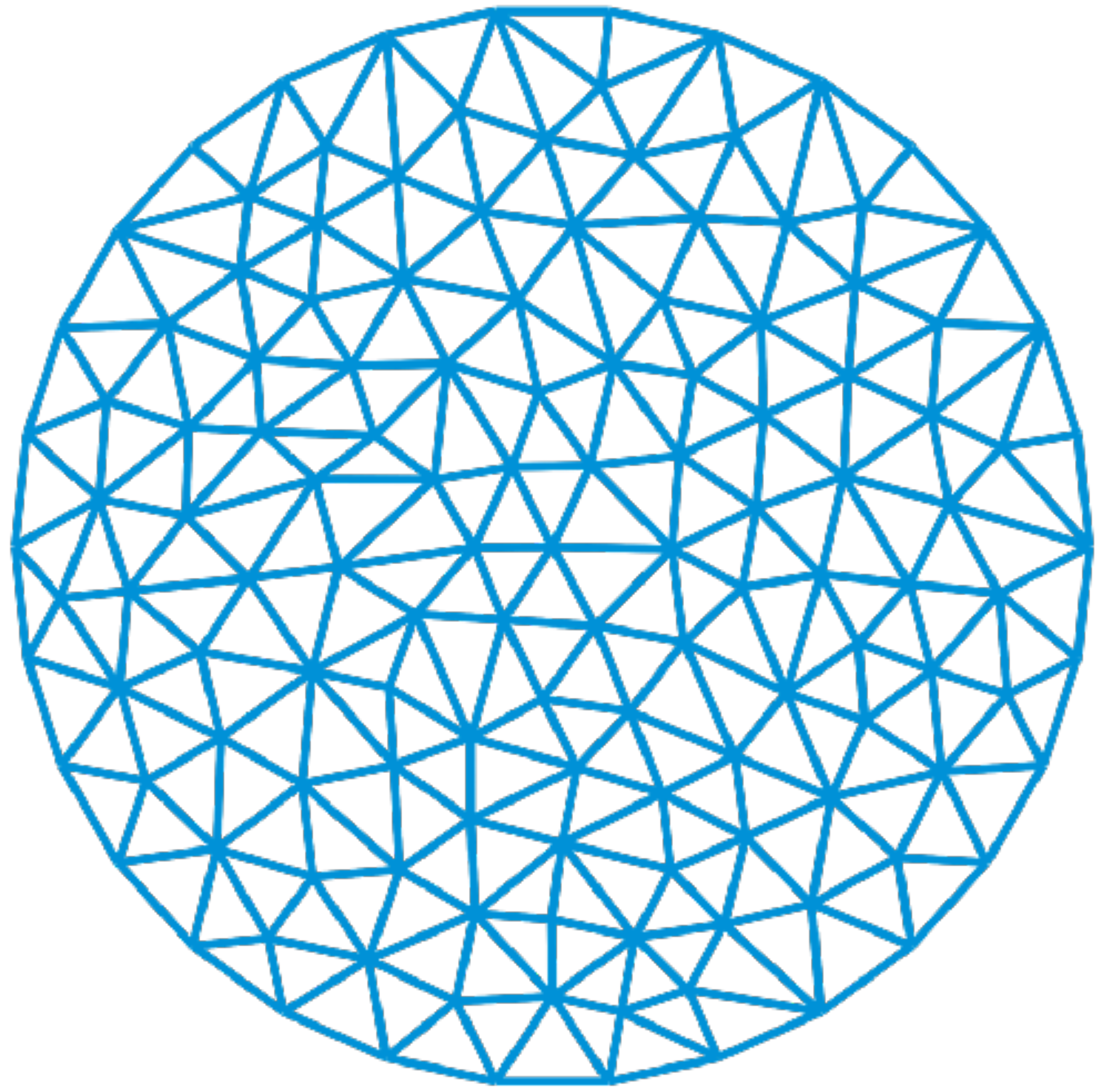
Betti numbers count repeated connections



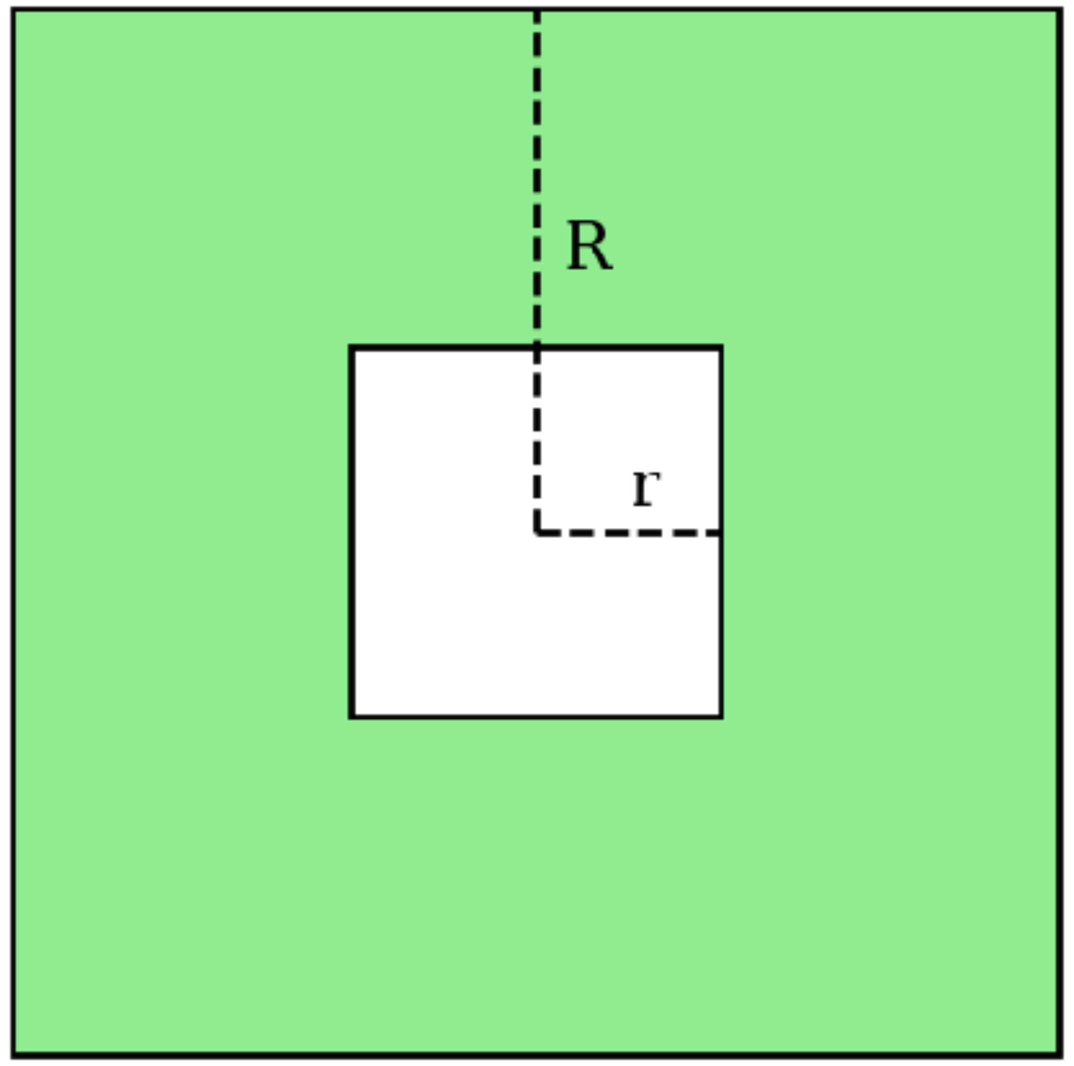
Combinatorial representations are good



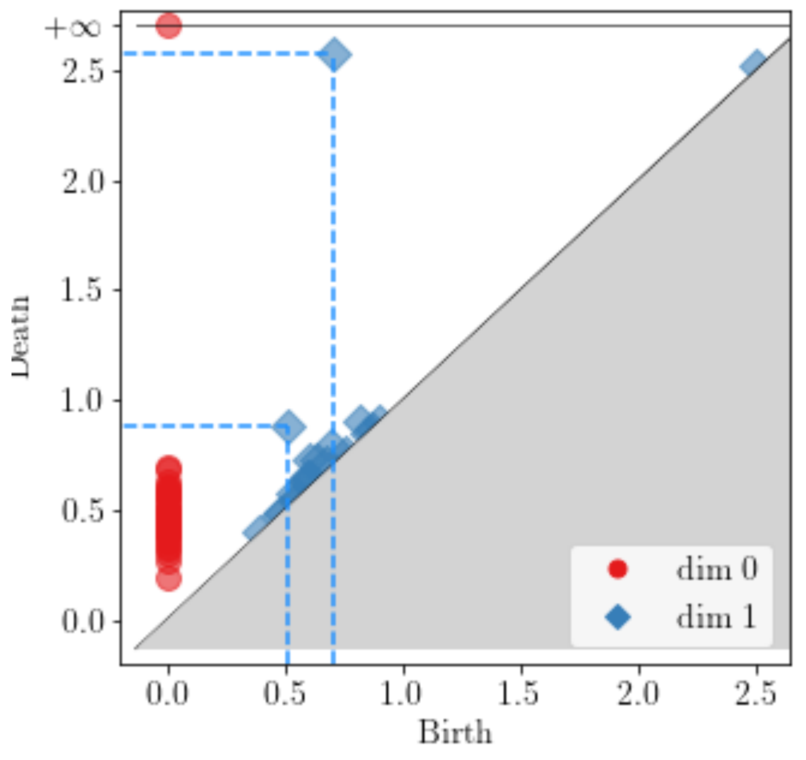
We have ways to describe networks.



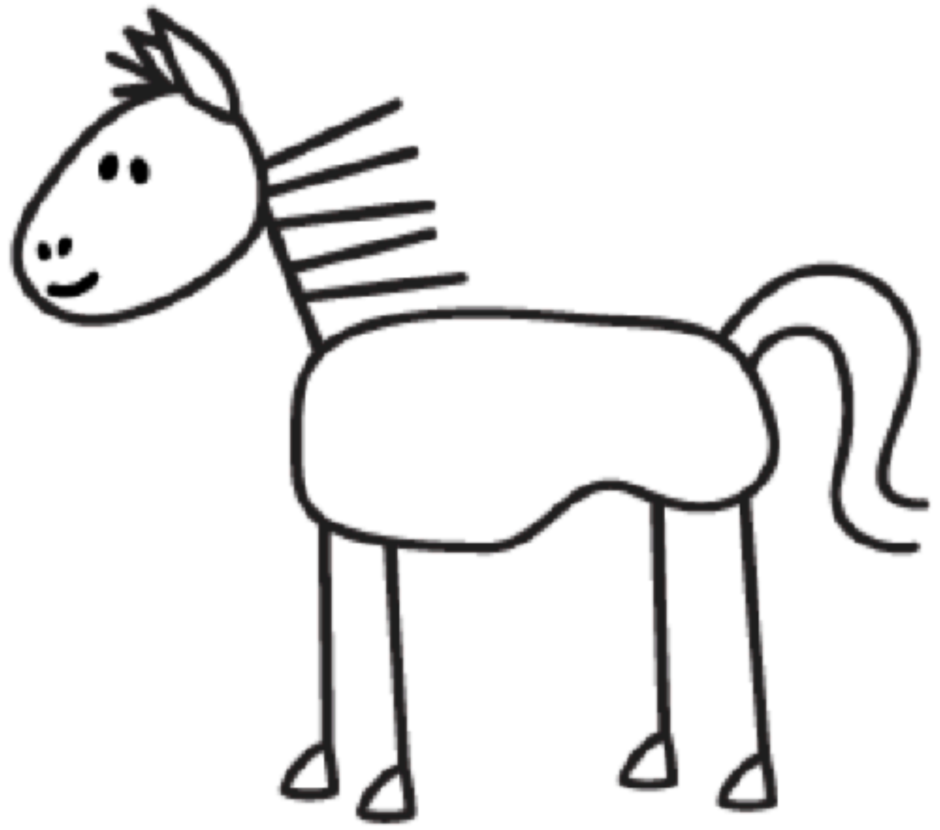
Algebraic topology is helpful for comparing different combinatorial representations



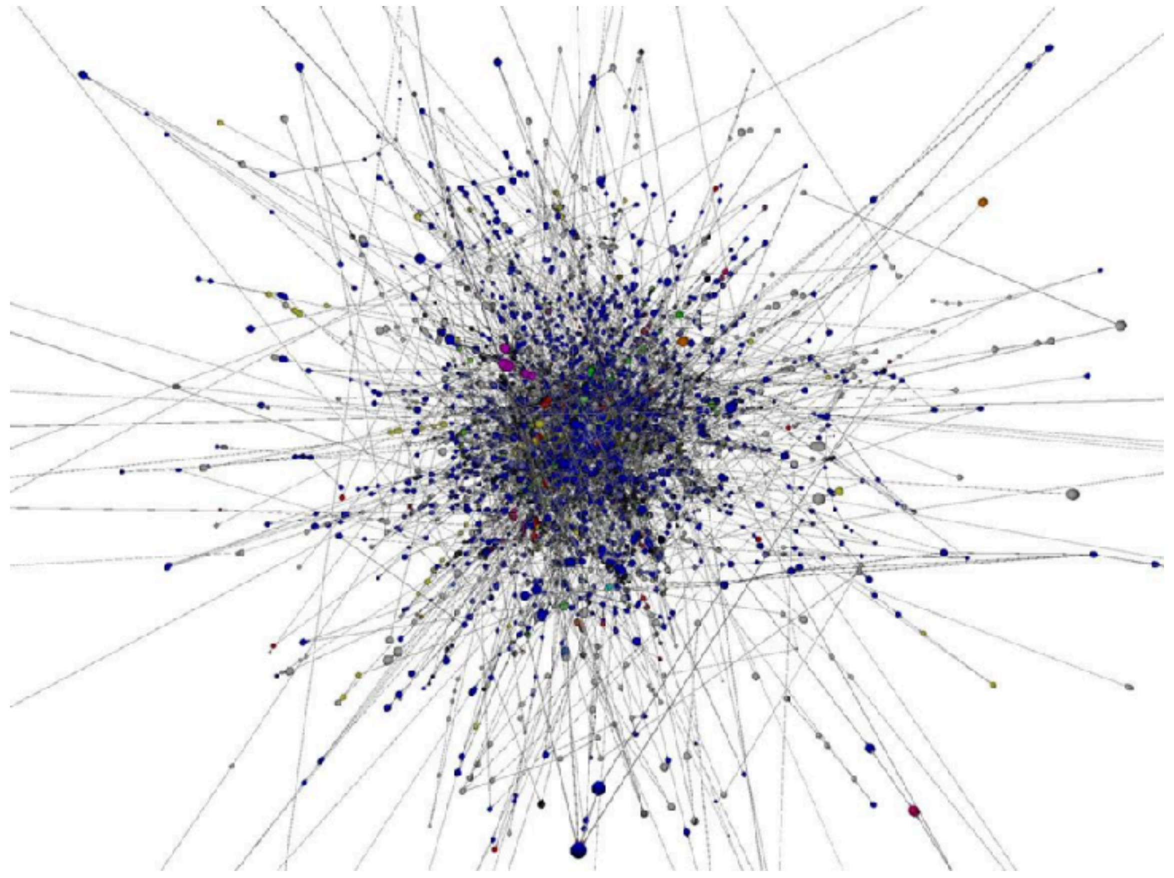
Betti numbers count repeated connections



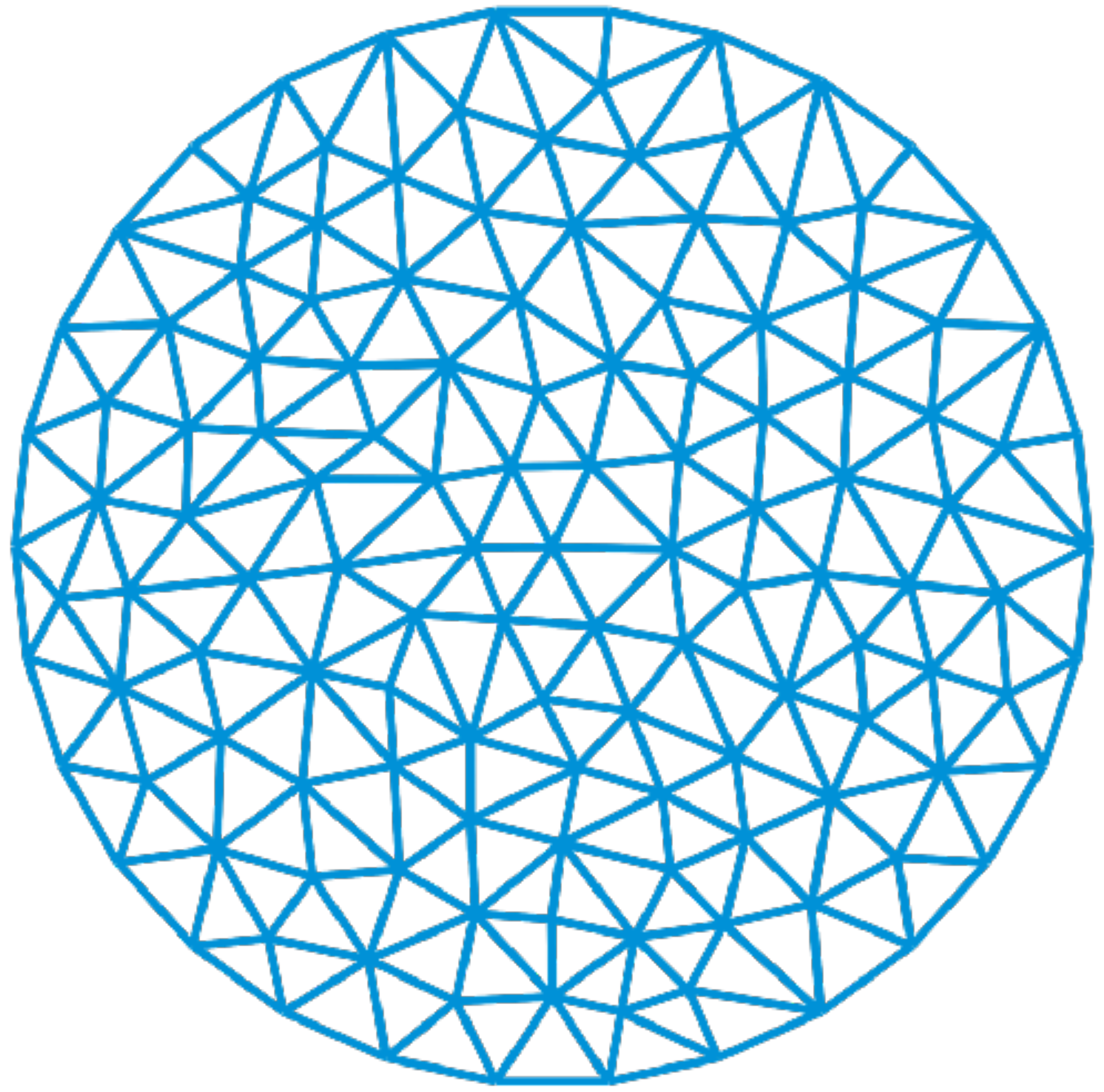
Persistent homology is not that scary.



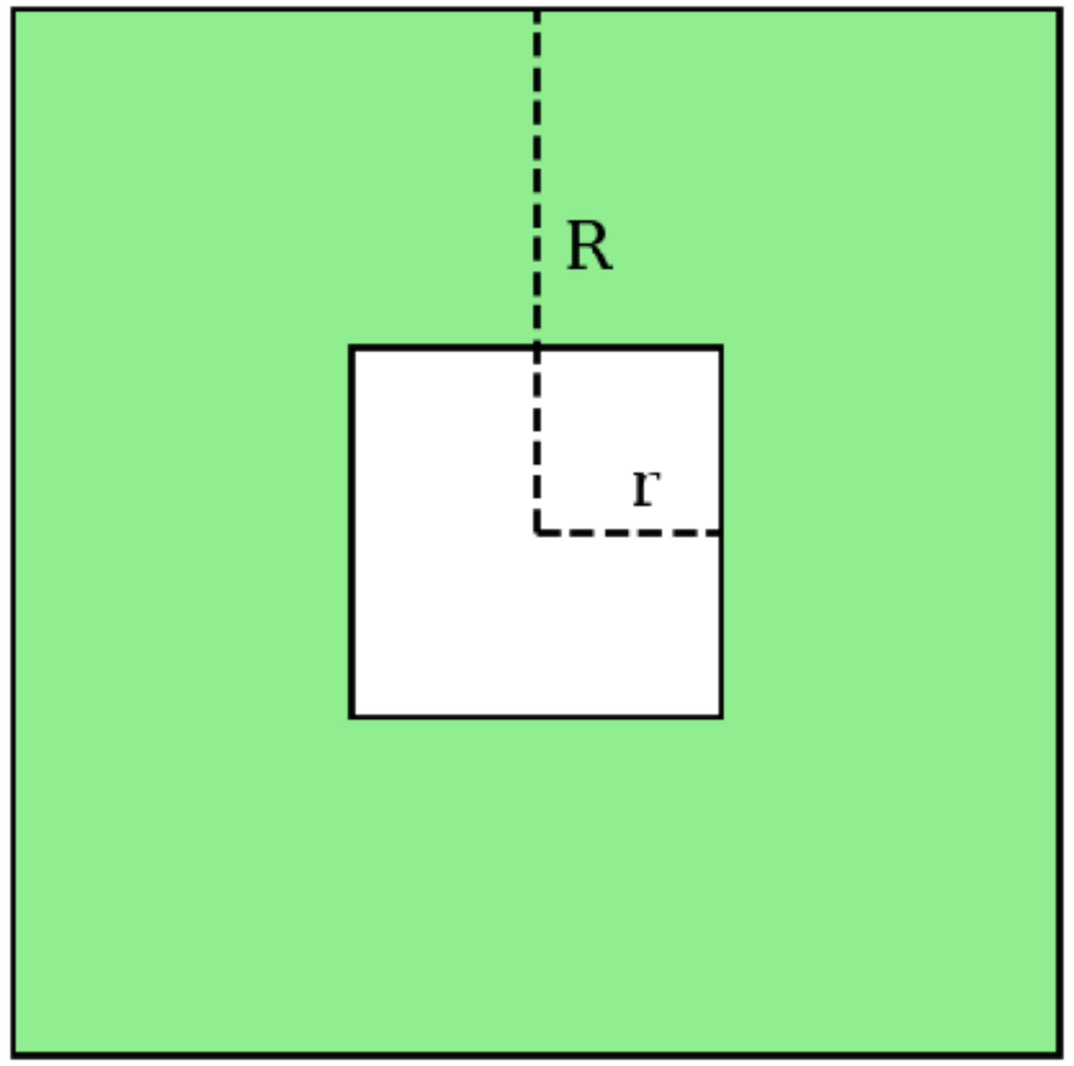
Combinatorial representations are good



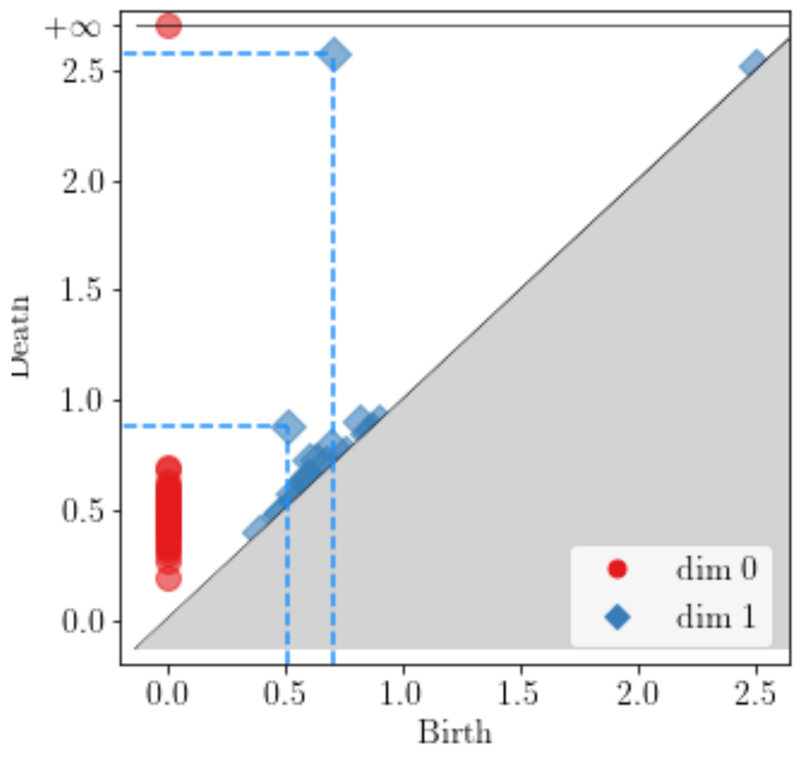
We have ways to describe networks.



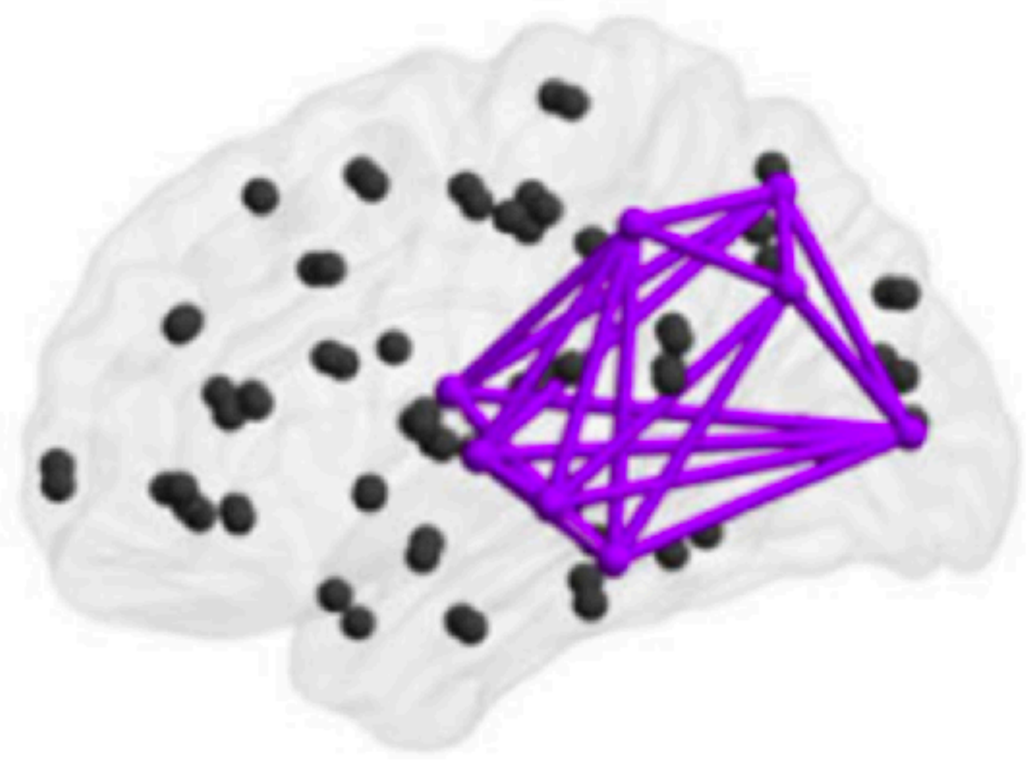
Algebraic topology is helpful for comparing different combinatorial representations



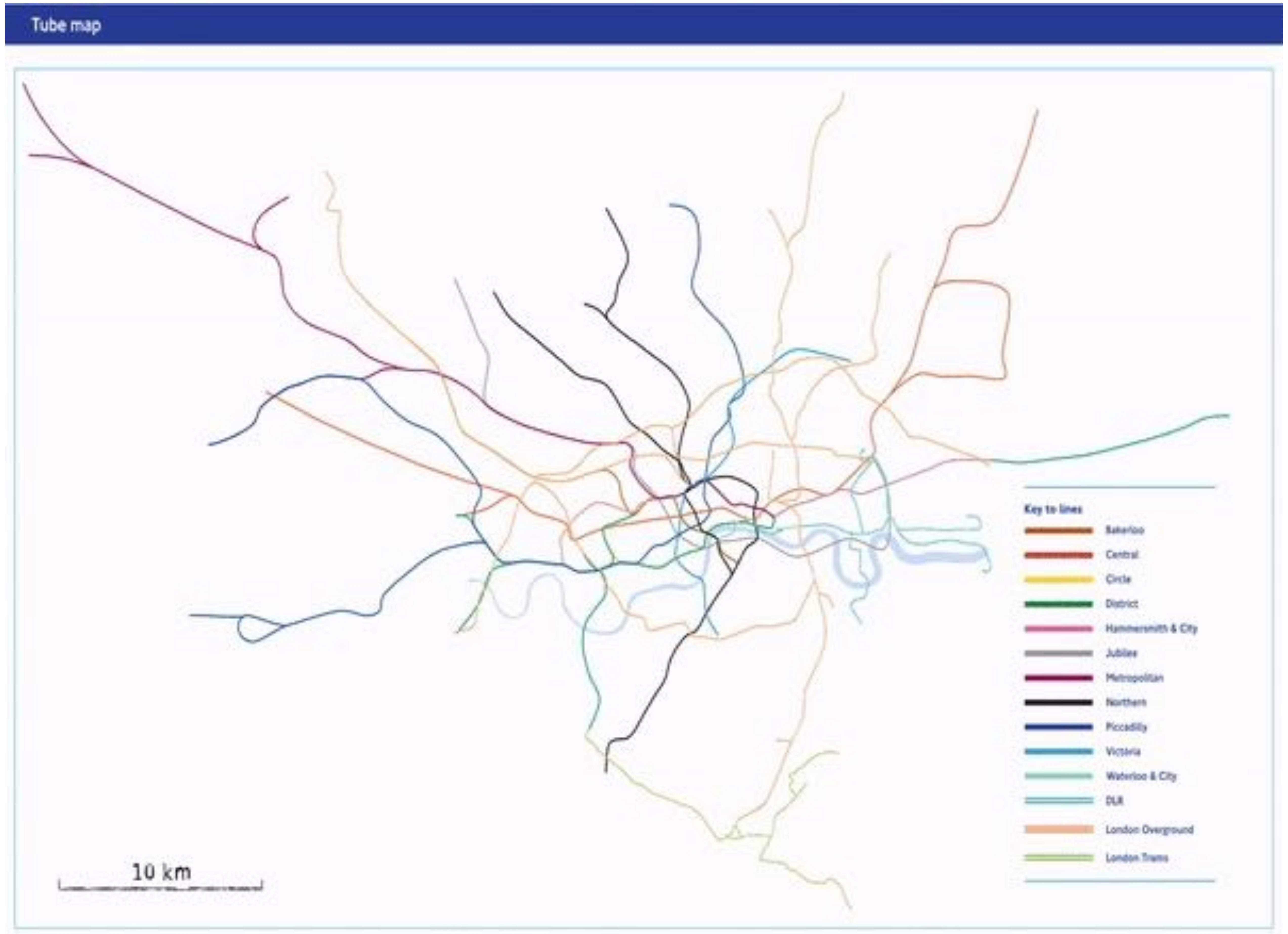
Betti numbers count repeated connections



Persistent homology is not that scary.



Repeated pathways in the brain may have biological significance



A_wandering_chemist
https://www.reddit.com/r/dataisbeautiful/comments/b8ihhr/comparison_between_the_london_tube_map_and_its/

A_wandering_chemist
https://www.reddit.com/r/dataisbeautiful/comments/b8ihhr/comparison_between_the_london_tube_map_and_its/

(Dis-) Connectivity

- Connected components

Partition of graph

- Modularity (the “difference” between proportion of in-group edges and proportion of inter-group edges)
- Participation score of a node with respect to a partition (1 - mean squared distributions of connections to different modules in the partition)