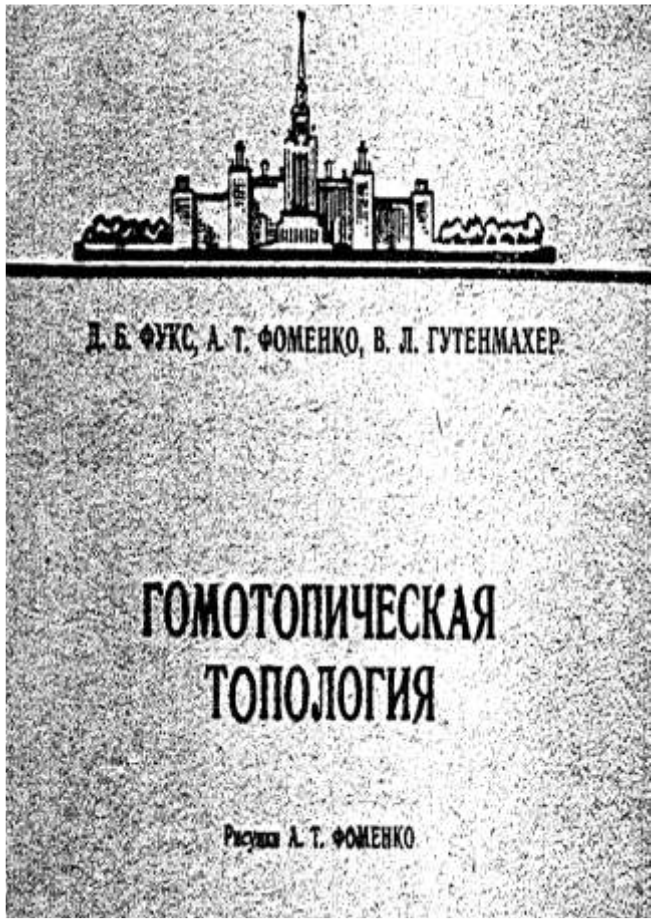
Topological approaches to data analysis

Václav Snášel

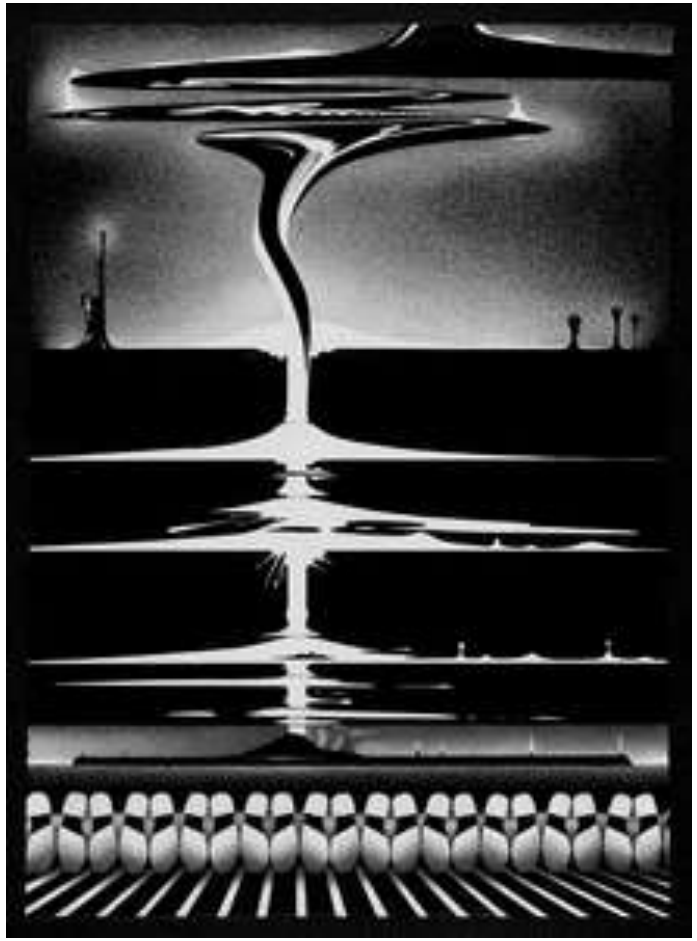
2018

Images



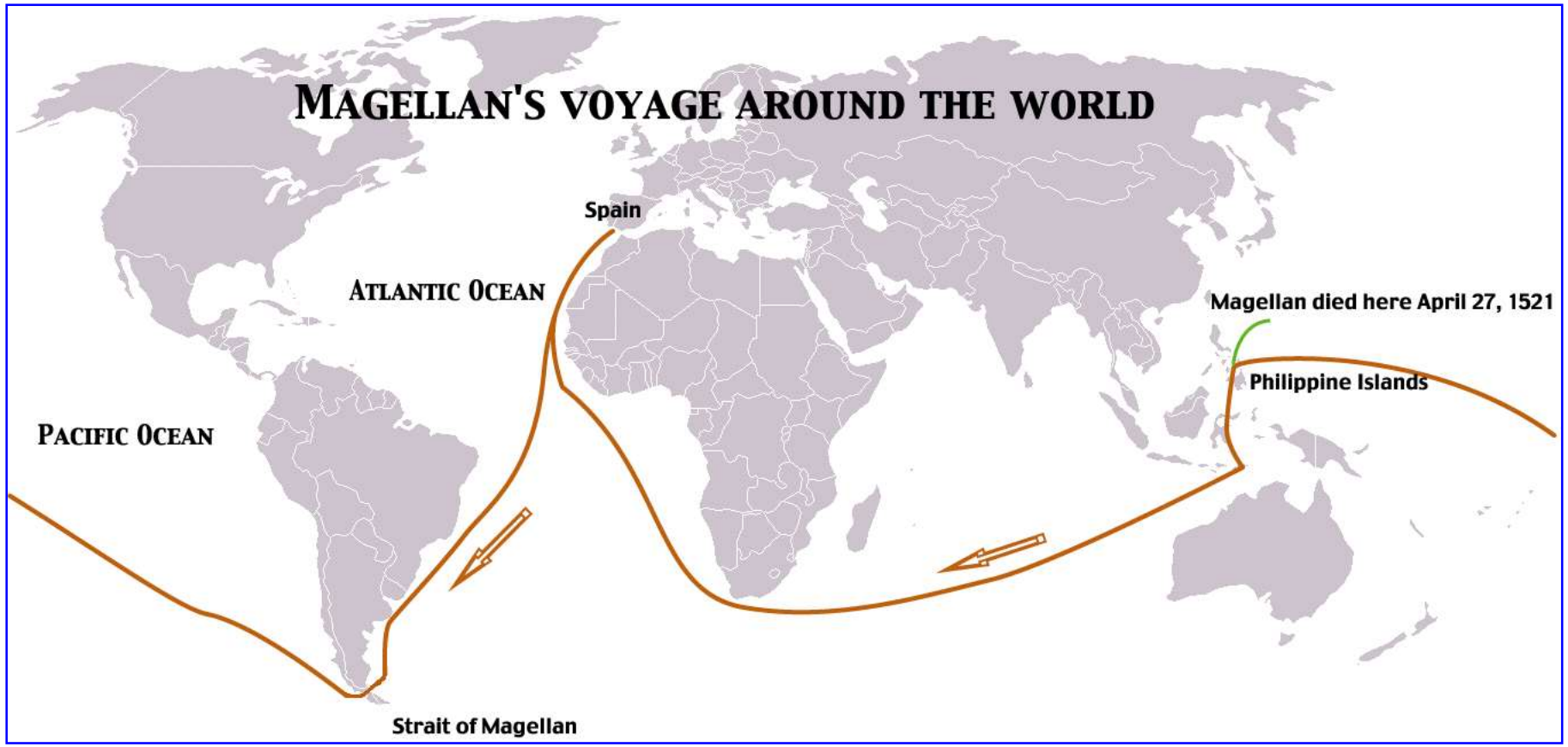
А.Т.Фоменко, Математика и Миф Сквозь Призму Геометрии,
http://dfgm.math.msu.su/files/fomenko/myth-sec6.php

More Images



Magellan's Journey

- August 10, 1519 — September 6, 1522; Start: about 250 men
- Return: about 20 men



Introduction - historical overview

The Erdös Number Project

This is the website for the Erdös Number Project, which studies research collaboration among mathematicians.

*The site is maintained by **Jerry Grossman** at **Oakland University**. **Patrick Ion**, a retired editor at **Mathematical Reviews**, and **Rodrigo De Castro** at the **Universidad Nacional de Colombia, Bogota** provided assistance in the past. Please address all comments, additions, and corrections to Jerry at **grossman@oakland.edu**.*

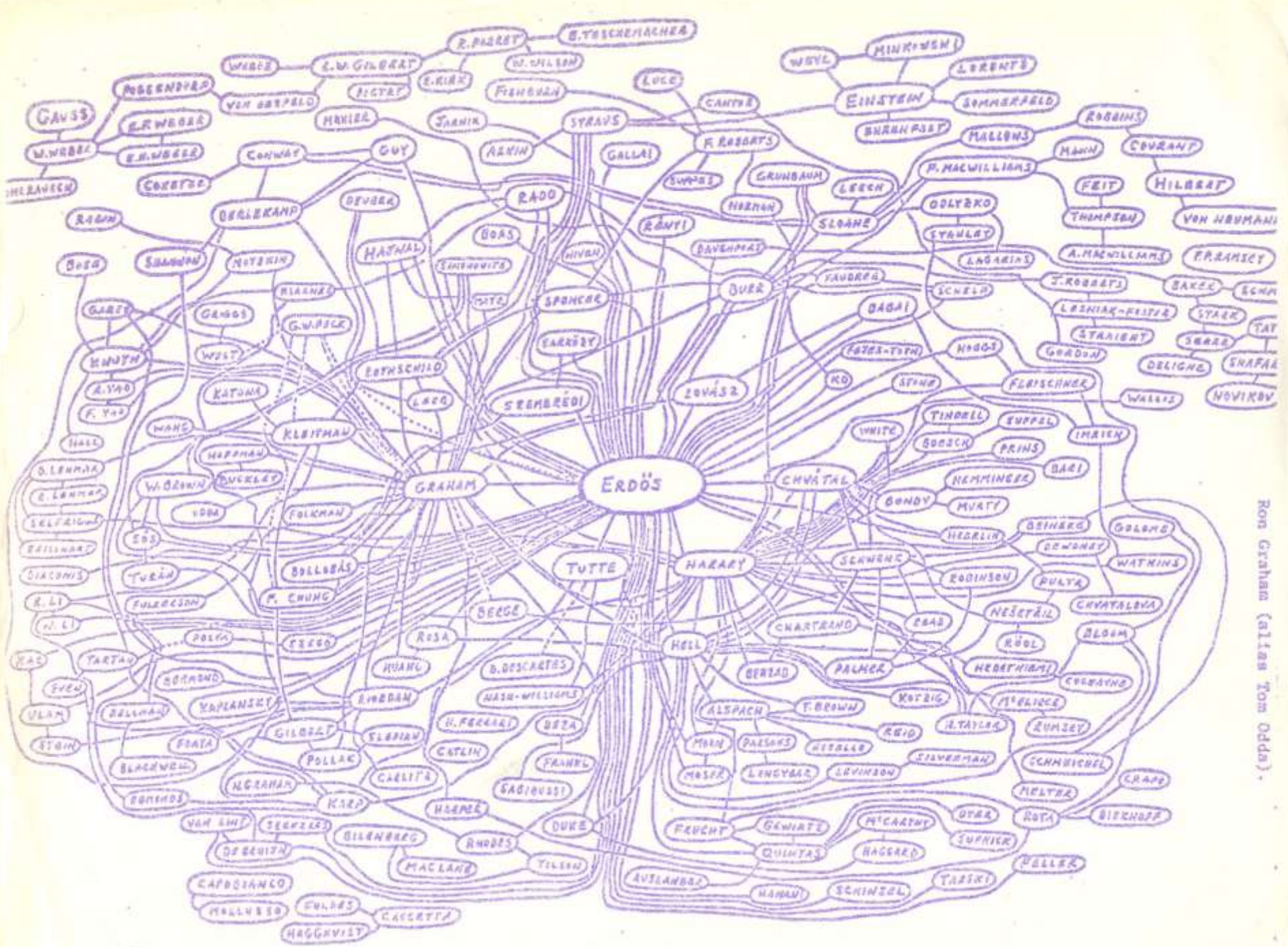
Erdös numbers have been a part of the **folklore of mathematicians** throughout the world for many years. For an introduction to our project, a description of what Erdös numbers are, what they can be used for, who cares, and so on, choose the "What's It All About?" link below. To find out who **Paul Erdös** is, look at this **biography** at the MacTutor History of Mathematics Archive, or choose the "Information about Paul Erdös" link below. Some useful information can also be found in **this Wikipedia article**, which may or may not be totally accurate.

http://www.oakland.edu/enp/

Erdös number **(1913-1996)**

1 475 papers

0	---	1 person
1	---	504 people
2	---	6593 people
3	---	33605 people
4	---	83642 people
5	---	87760 people
6	---	40014 people
7	---	11591 people
8	---	3146 people
9	---	819 people
10	---	244 people
11	---	68 people
12	---	23 people
13	---	5 people

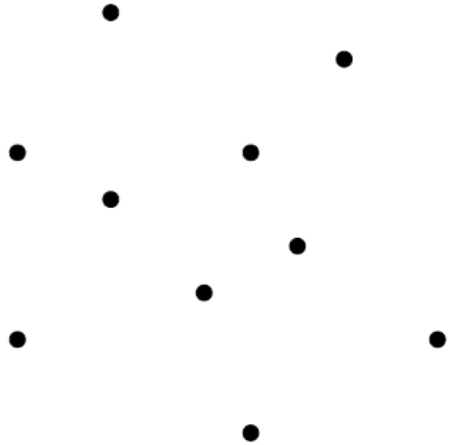


Ron Graham (alias Tom Odda).

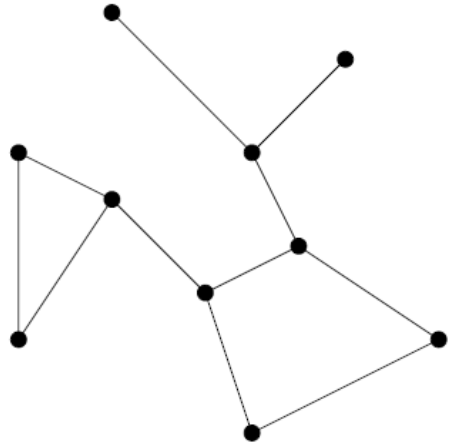
Figure 1

To appear in Topics in Graph Theory (F. Harary, ed.) New York Academy of Sciences (1979).

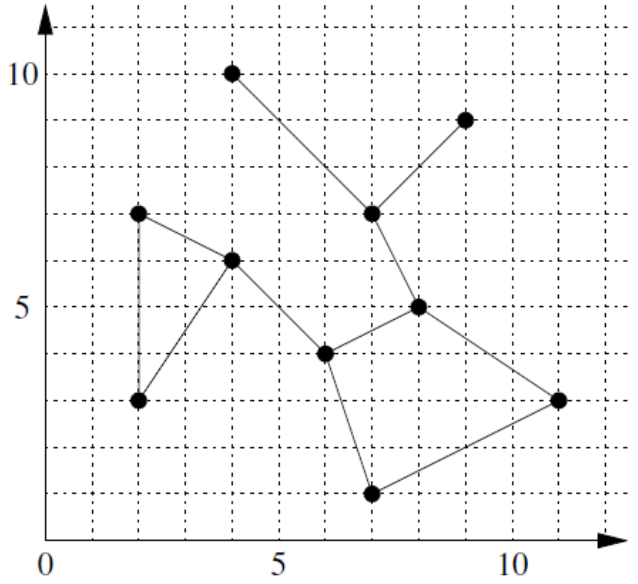
Topology



(a) A space



(b) A topological space



(c) A metric space

Motivation

Anatoly Fomenko and Dmitry Fuchs, Homotopical Topology, Springer, (Graduate Texts in Mathematics), 2016.

Dimitry Kozlov, Combinatorial Algebraic Topology, Springer, (Algorithms and Computation in Mathematics), 2008.

Allen Hatcher, Algebraic Topology, Cambridge University Press, 2001.

Tomasz Kaczynski, Konstantin Mischaikow, Marian Mrozek, Computational Homology, (Applied Mathematical Sciences), Springer, 2004.

Motivation

Afra J. Zomorodian, Topology for Computing, (Cambridge Monographs on Applied and Computational Mathematics), American Mathematical Society, 2009.

Steve Y. Oudot, Persistence Theory: From Quiver Representations to Data Analysis, (Mathematical Surveys and Monographs), American Mathematical Society, 2017.

Afra J. Zomorodian, Advances in Applied and Computational Topology (Proceedings of Symposia in Applied Mathematics), 2012.

Motivation

Herbert Edelsbrunner and John L. Harer, Computational Topology: An Introduction, American Mathematical Society, 2009.

Robert Ghrist, Elementary Applied Topology, 2014.

Motivation

Julien Tierny, Topological Data Analysis for Scientific Visualization (Mathematics and Visualization), Springer, 2018.

Julien Tierny, Topological Data Analysis for Scientific Visualization, (Mathematics and Visualization), Springer, 2017.

Valerio Pascucci, Xavier Tricoche, Hans Hagen, Julien Tierny, Topological Methods in Data Analysis and Visualization: Theory, Algorithms, and Applications, (Mathematics and Visualization), Springer, 2011.

Motivation

Gunnar Carlsson, Topology and data, Bull. Amer. Math. Soc. 46 (2009), 255-308.

Gunnar Carlsson, Topological pattern recognition for point cloud data, Acta Numerica, Volume 23, May 2014, 289 – 368.

Topological Space

A topological space is a set X together with a collection τ
of subsets of X (i.e., τ is a subset of the power set of X) satisfying the following axioms:

1. The empty set $\emptyset$ and X are in τ.
2. The union of any collection of sets in τ is also in τ.
3. The intersection of any finite collection of sets in τ is also in τ.

The set τ is called a topology on X. The sets in τ are referred to as open sets, and their complements in X are called closed sets.

A topology specifies "nearness"; an open set is "near" each of its points.

A function between topological spaces is said to be continuous if the inverse image of every open set is open.

Metric Spaces

A metric is a „distance“ function, defined as follows:

If X is a set, then a metric on X is a function d

$$d: X \times X \to \mathbb{R}_+$$

which satisfied the following properties:

- $d(x, x) \geq 0$
- $d(x, y) = d(y, x)$
- $d(x, y) + d(y, z) \geq d(x, z)$ (Triangle inequality)

(X, d) is called metric space.

From Metric Space to Topological Space

In any metric space M we can define the r-neighborhoods as the sets of the form $B(x,r) = \{y \in M : d(x,y) < r\}$.

A point x is an interior point of a set E if there exists an r-neighborhood of x that is a subset of E.

A point x is a limit point of a set E, if every r-neighborhood of x contains a point $y \neq x$ in E.

A set E is open if all points of E are interior points of E.
A set E is closed of all limit points of E belong to E.

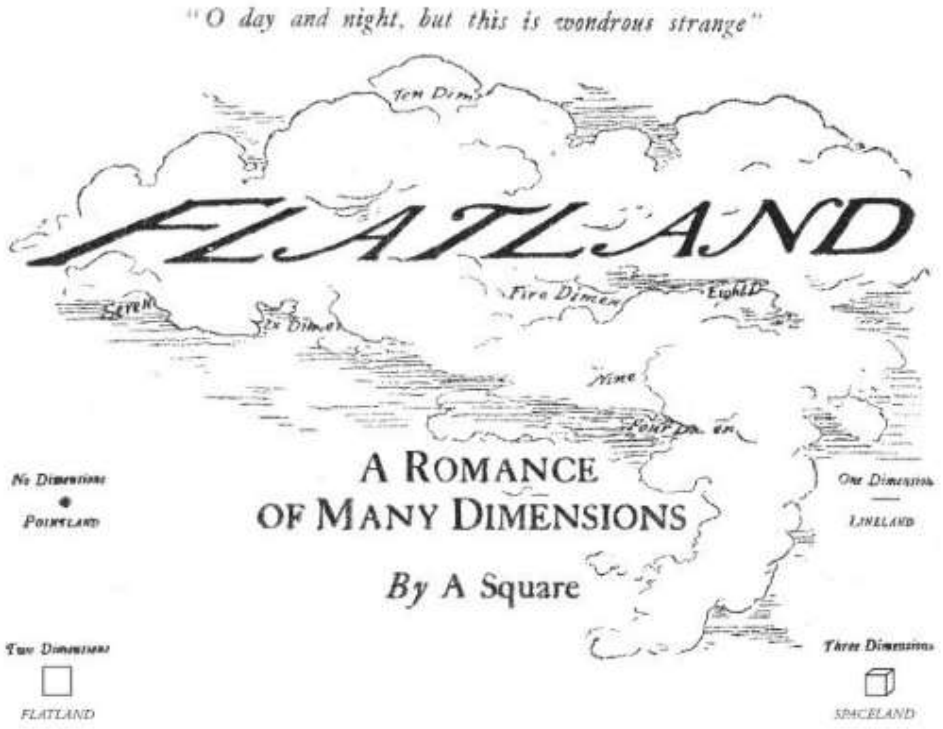
Theorem: A set is open if and only if its complement is closed.

General Topology Overview

Branches

- Point-Set Topology
  - Based on sets and subsets
  - Connectedness
  - Compactness
- Algebraic Topology
  - Derived from Combinatorial Topology
  - Models topological entities and relationships as algebraic structures such as groups or a rings
- Smooth Manifold
  - Morse theory
  - Field theory

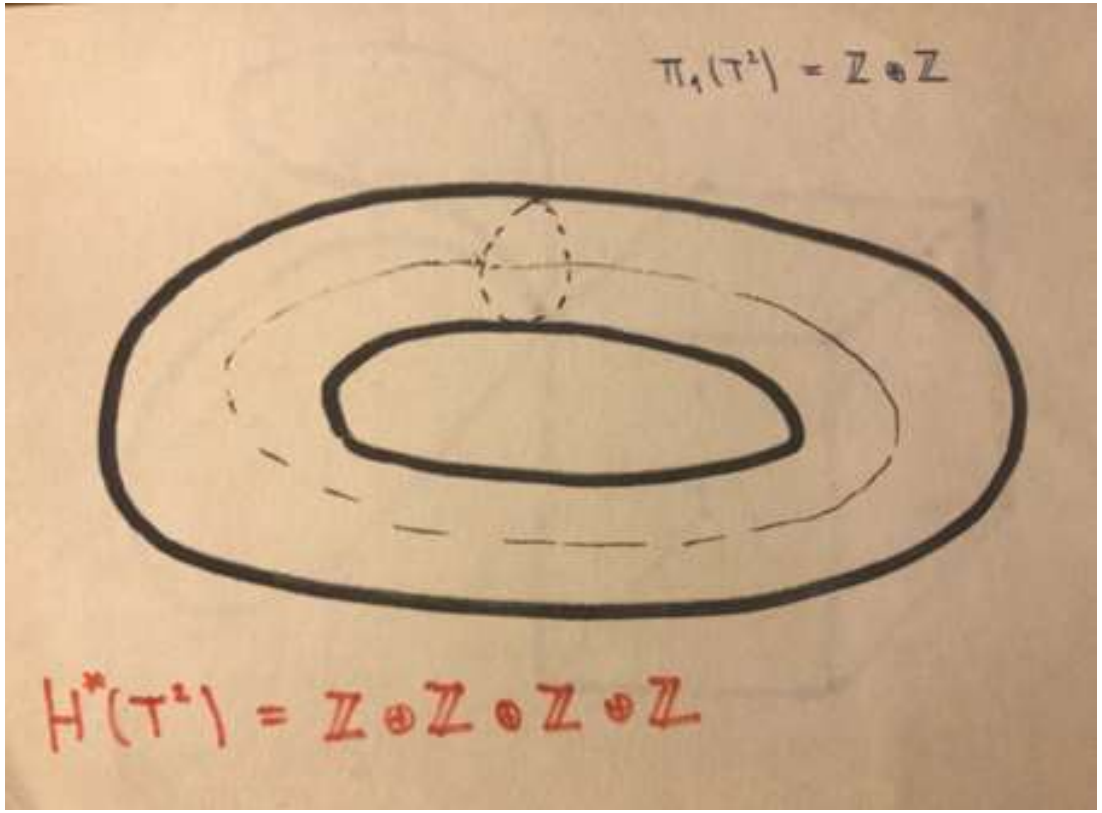
FlatLand
A Romance of Many Dimensions EDWIN ABBOTT

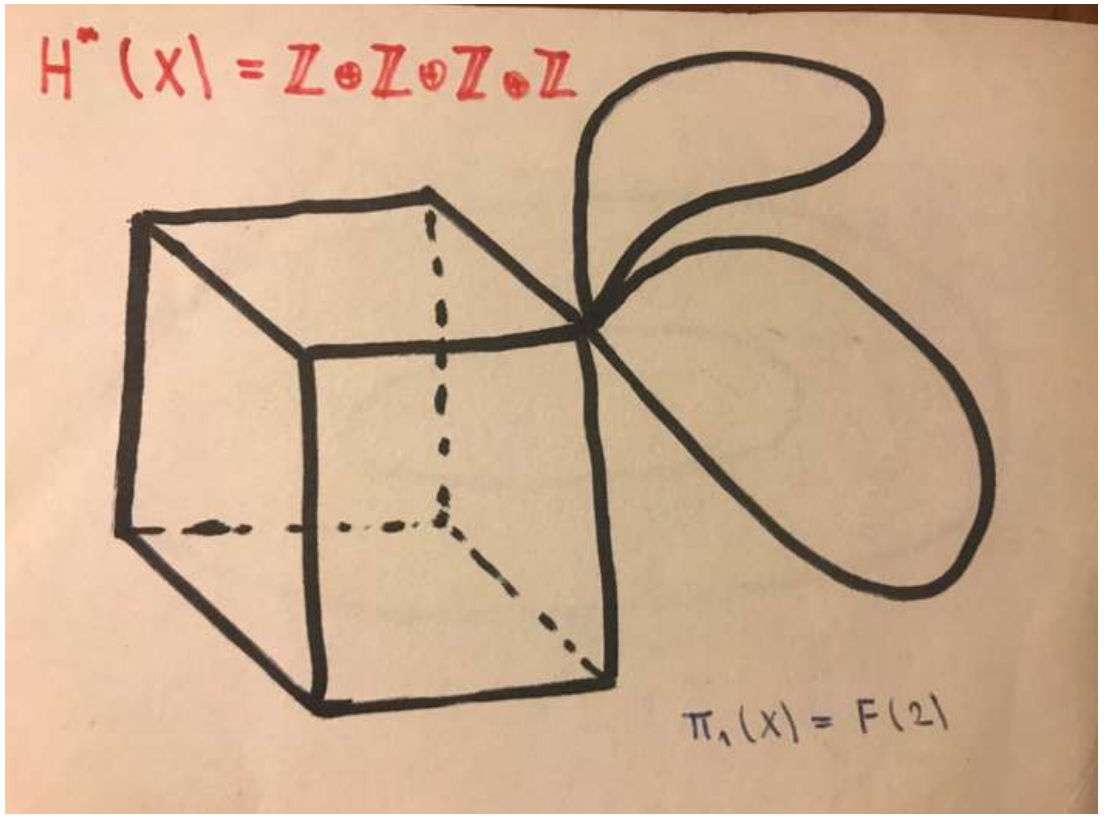


"Fie, fie, how franticly I square my talk!"

PRINCETON UNIVERSITY PRESS
PRINCETON AND OXFORD
1926

Cycle in topology





Albrecht Dold, Lectures on Algebraic Topology, Springer, 1992.
Edward H. Spanier, Algebraic Topology, McGraw-Hill Inc., 1966.

Perspectives - Topology

Gunnar Carlsson: Topology and Data

Bulletin of The American Mathematical Society, Volume 46, Number 2, April 2009, Pages 255–308

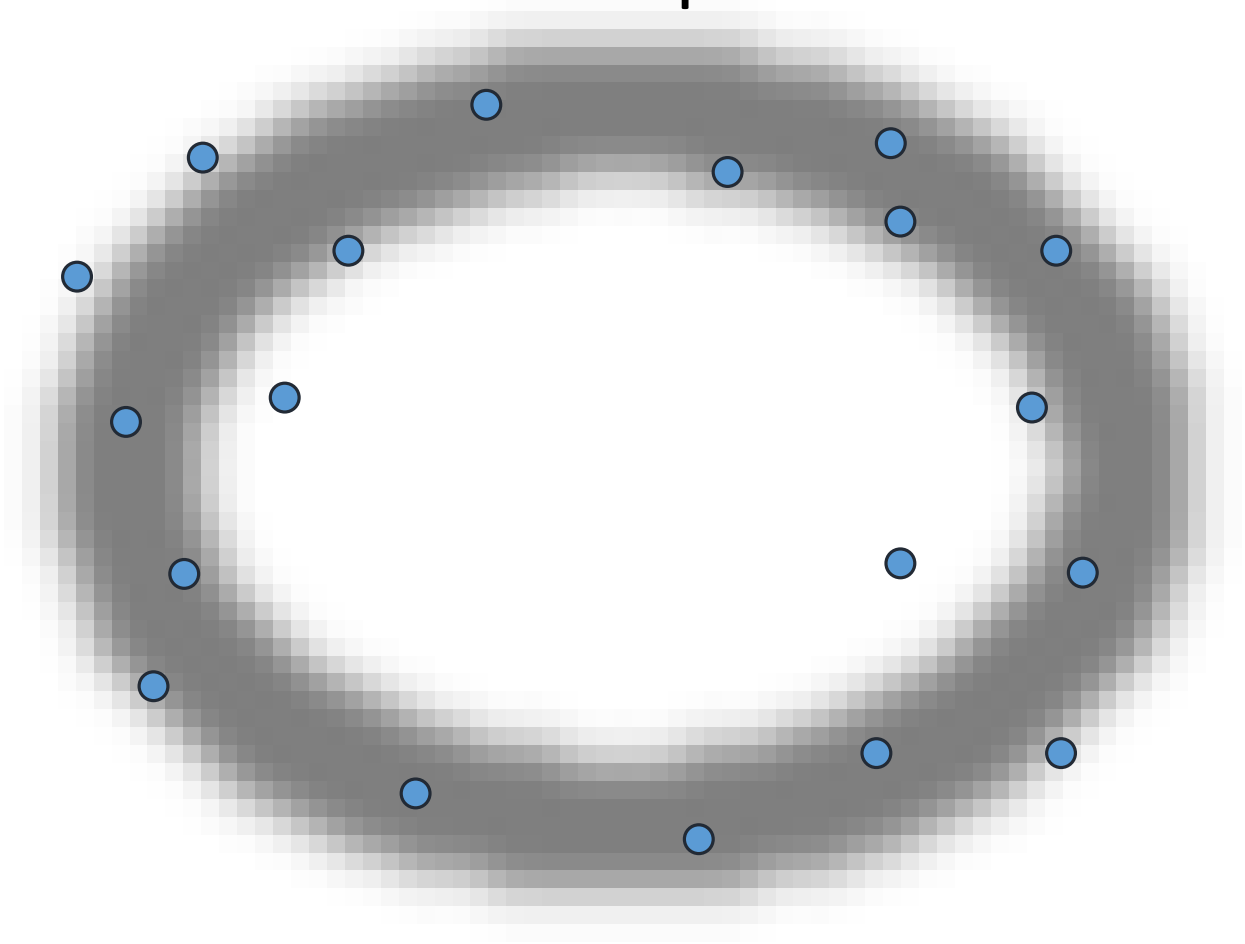
- **Qualitative information is needed**: One important goal of data analysis is to allow the user to obtain knowledge about the data, i.e. to understand how it is organized on a large scale.
- **Metrics are not theoretically justified**: In physics, the phenomena studied often support clean explanatory theories which tell one exactly hat metric to use. In biological problems, on the other hand, this is much less clear. In the biological context, notions of distance are constructed using some intuitively attractive measures of similarity
- **Coordinates are not natural**: Although we often receive data in the form of vectors of real numbers, it is frequently the case that the coordinates, like the metrics mentioned above.

Topological approaches to data analysis

Topological approaches to data analysis are based around the notion that there is an idea of proximity between these data points.

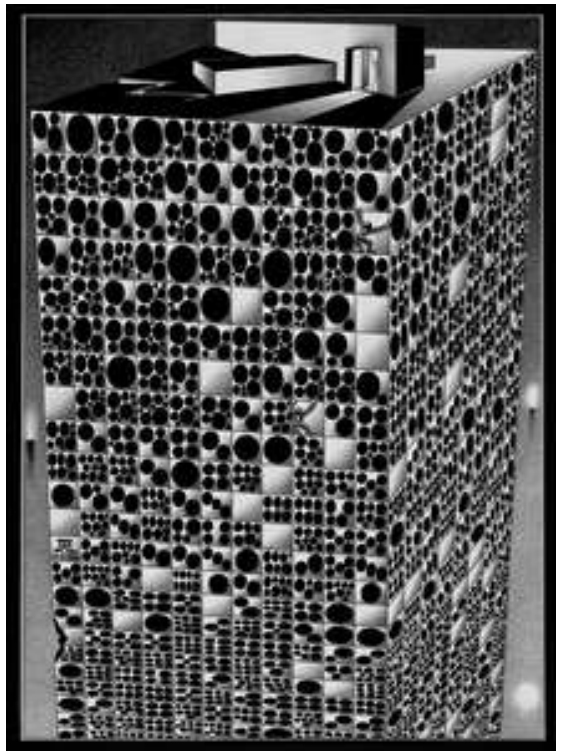
For each data point $\boldsymbol{x} = (x_1, \ldots, x_n)$ consists of n numerical values, we have a natural definition of proximity that comes from the standard Euclidean distance: this is the generalization of the standard distance in the plane $d(\boldsymbol{x}, \boldsymbol{y}) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$

Example: What is the shape of the data?



Problem: Discrete points have trivial topology.

Data Has Shape
And Shape Has Meaning



Basic Concepts of Graph

- Graphs $\mathrm{G} = (V, E)$
- V: the set of nodes
- E: the set of edges
- v_i: a node from V
- $e(v_i, v_j)$: an edge between node v_i and v_j
- A: the adjacency matrix; $A_{ij} = 1$ if exists edge between node v_i and v_j else $A_{ij} = 0$
- d_i: the degree of node v_i
- D: degree matrix; $D_{ii} = d_i$ else $D_{ij} = 0$
- geodesic: a shortest path between two nodes
  - geodesic distance

Graphs

Many data sets can be transformed to a graph representation by simple means: → similarity graphs

Given:

- data „points“ $x_1, \ldots, x_n$ in R^m
- similarity values $s(x_i, x_j)$ or distance values $d(x_i, x_j)$

Construct graph:

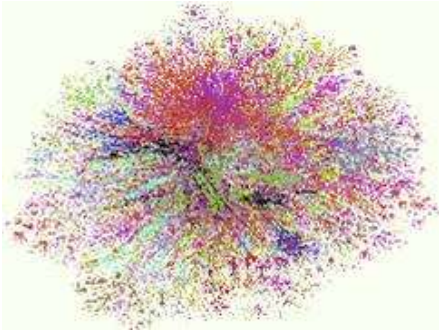
- Data point are vertices of the graph
- Connect points which are „close“

Intuition: graph captures local neighborhoods

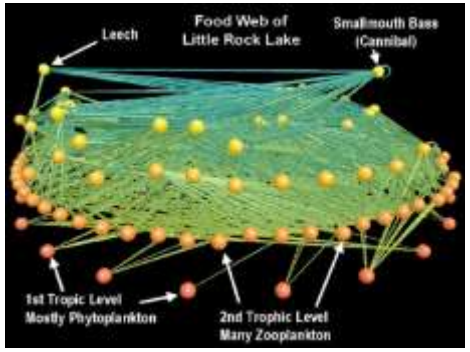
Constructing graph

- data „points“ $x_1, \ldots, x_n$ in R^m
- Nodes x_i and x_j are connected by edge if $\| x_i - x_j \|^2 < \varepsilon$
- Nodes x_i and x_j are connected by edge if x_i is among k nearest neighbors of x_j or if x_j is among k nearest neighbors of x_i

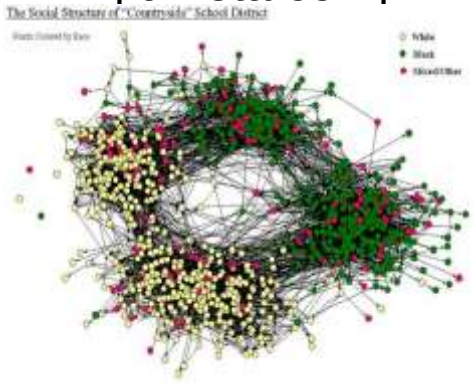
Graphs - why should we care?



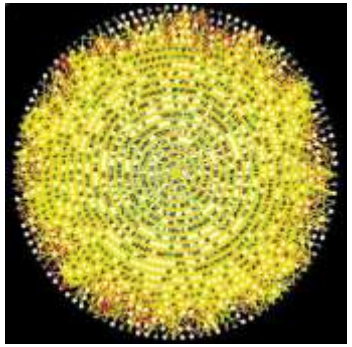
Internet Map
[lumeta.com]



Food Web
[Martinez ’91]



Friendship Network
[Moody ’01]



Protein Interactions
[genomebiology.com]

Datasets in the form of matrices - graphs

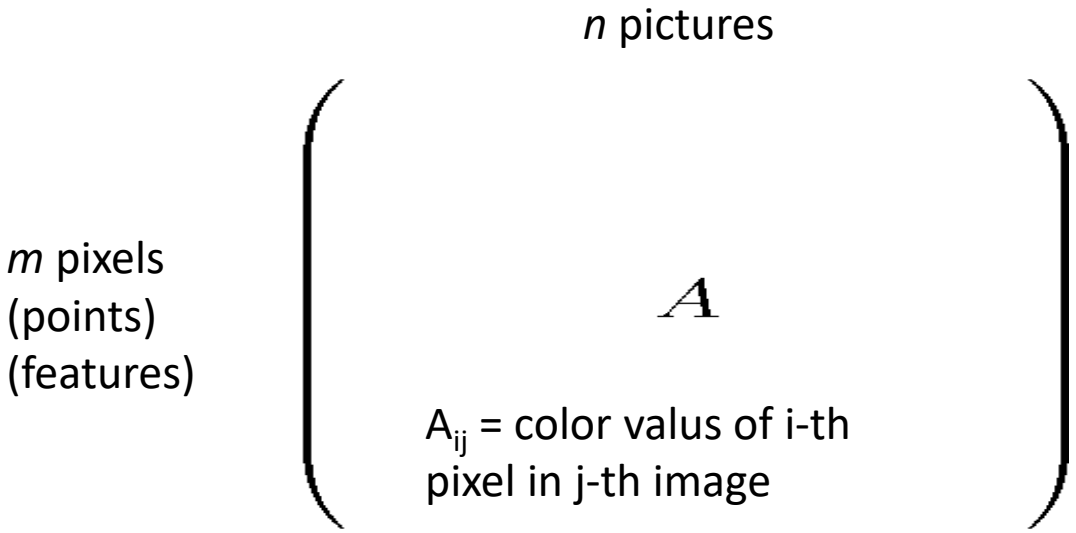
We are given m objects and n features describing the objects.
(Each object has n numeric values describing it.)

Dataset
An m-by-n matrix A, A_{ij} shows the "importance" of feature j for object i.
Every row of A represents an object.

Goal
We seek to understand the structure of the data, e.g., the underlying process generating the data.

Images matrices

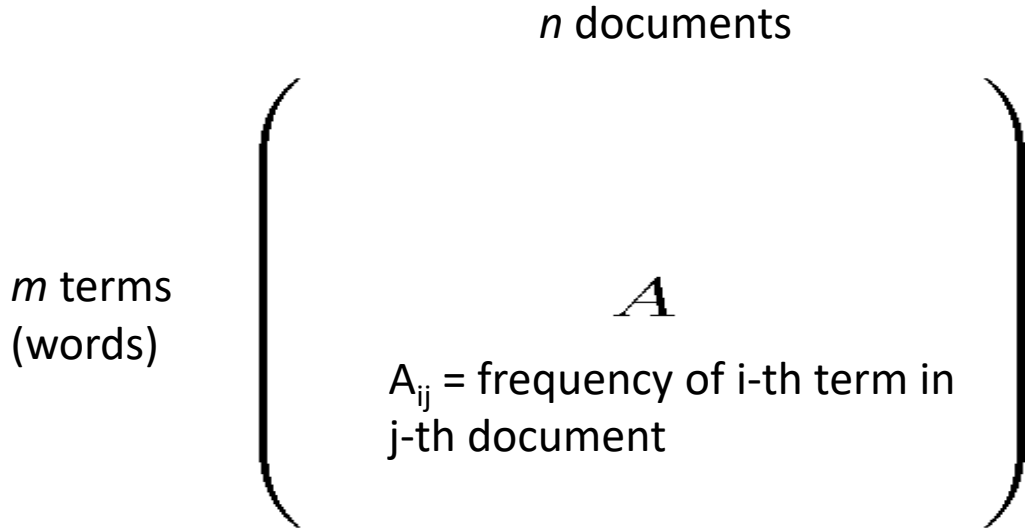
A collection of images is represented by an m-by-n matrix



Data mining tasks
- Cluster or classify images
- Find “nearest neighbors”
- Feature selection: find a subset of features that (accurately) clusters or classifies images.

Document-term matrices

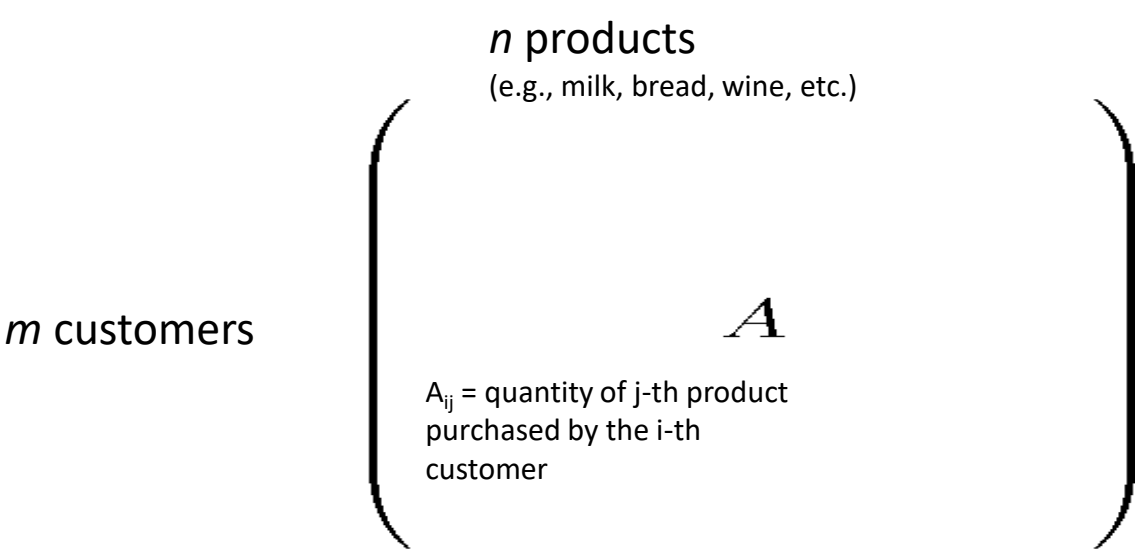
A collection of documents is represented by an m-by-n matrix



Data mining tasks
- Cluster or classify documents
- Find “nearest neighbors”
- Feature selection: find a subset of terms that (accurately) clusters or classifies documents.

Market basket matrices

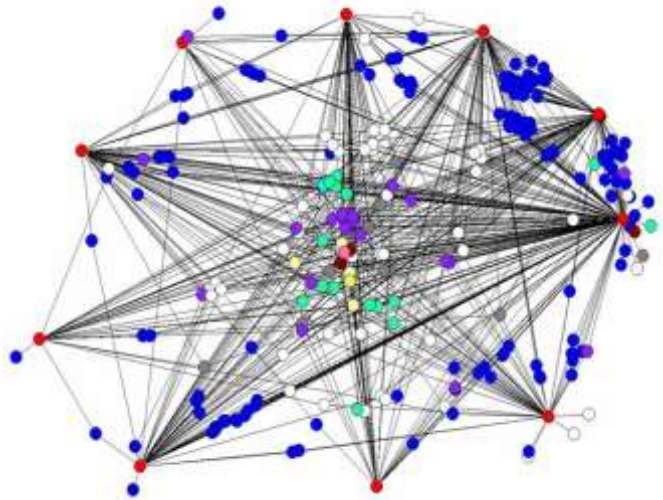
Common representation for association rule mining.



Data mining tasks
- Find association rules
E.g., customers who buy product x buy product y with probility 89%.
- Such rules are used to make item display decisions, advertising decisions, etc.

Social networks (e-mail graph, FaceBook, MySpace, etc.)

Represents the email communications (relationships) between groups of users.



n users

m users $\begin{pmatrix} & A & \end{pmatrix}$

A_{ij} = number of emails exchanged between users i and j during a certain time period

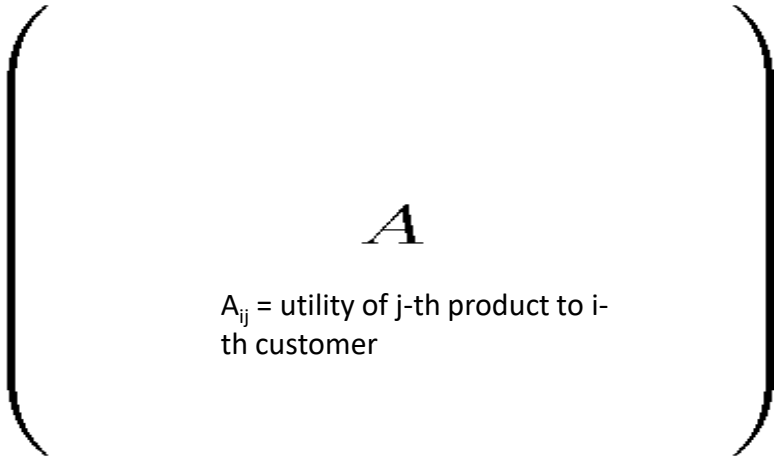
Data mining tasks
- cluster the users
- identify “dense” networks of users (dense subgraphs)

Recommendation systems

The m-by-n matrix A represents m customers and n products.

products

customers



A

A_{ij} = utility of j-th product to i-th customer

Data mining task
Given a few samples from A, recommend high utility products to customers.

Intrusion detection

The m-by-n matrix A represents m records and n attributes. The data for our experiments was prepared by the 1998 DARPA intrusion detection evaluation program by MIT Lincoln Labs

attributes

records

$$A$$

A_{ij} = utility of j-th attribute to i-th record

Data mining task

Reduce noise in the data.

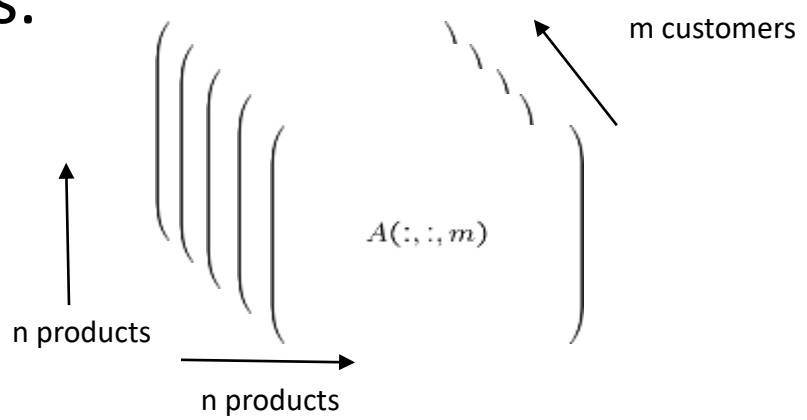
Tensors: recommendation systems

Economics:

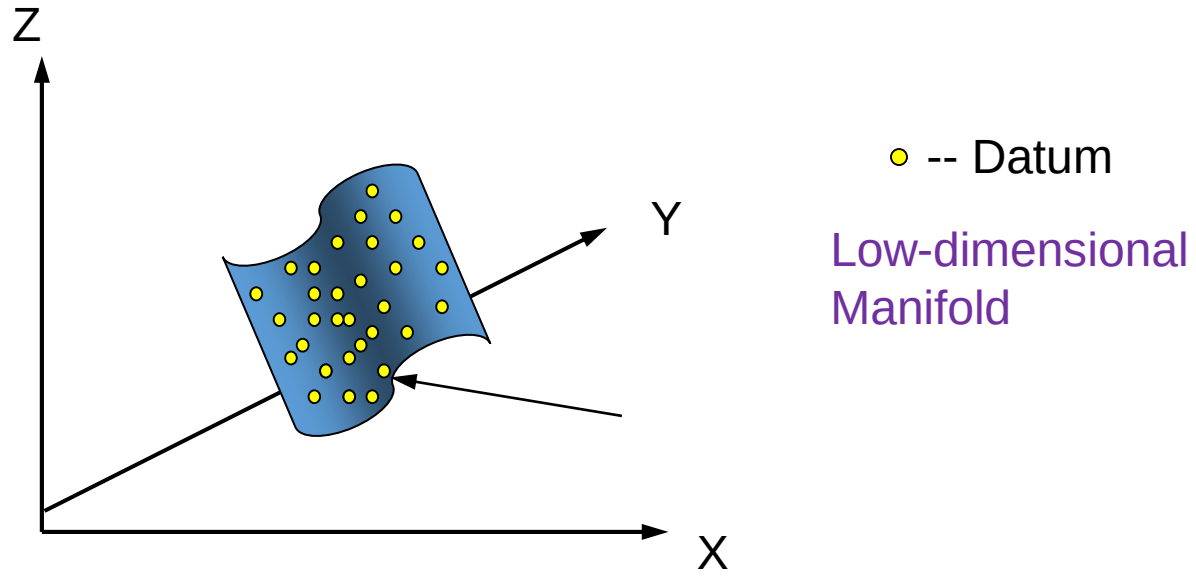
- Utility is ordinal and not cardinal concept.
- Compare products; don't assign utility values.

Recommendation Model Revisited:

- Every customer has an n-by-n matrix (whose entries are +1,-1) and represent pair-wise product comparisons.
- There are m such matrices, forming an n-by-n-by-m 3-mode tensor A.



Data as manifolds



- Data lie on a low-dimensional manifold. The shape of the manifold is not known *a priori*.

Reeb graphs

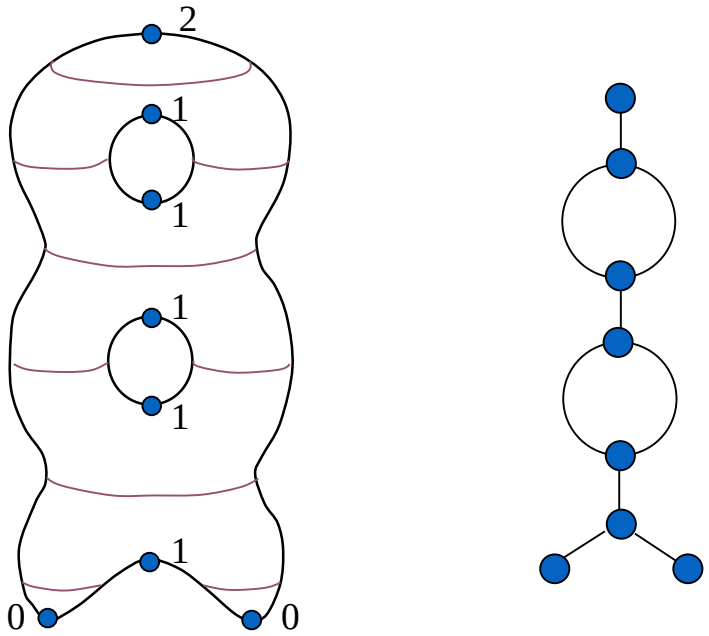
A Reeb graph (named after Georges Reeb by René Thom) is a mathematical object reflecting the evolution of the level sets of a real-valued function on a manifold. Reeb graph is based on Morse theory.

Similar concept was introduced by G.M. Adelson-Velskii and A.S. Kronrod and applied to analysis of Hilbert's thirteenth problem.

Reeb graphs found a wide variety of applications in computational geometry and computer graphics, including computer aided geometric design, topology-based shape matching, topological data analysis, topological simplification and cleaning, surface segmentation and parametrization, efficient computation of level sets, and geometrical thermodynamics

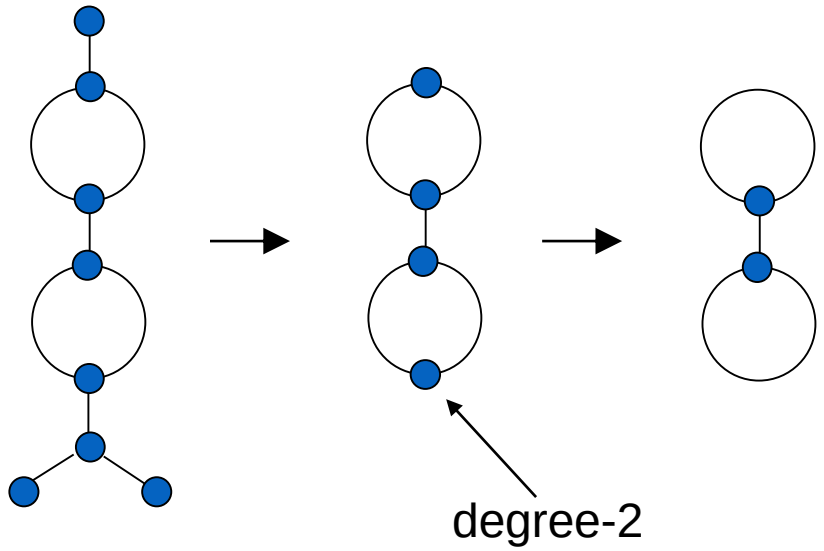
Reeb graphs

- Schematic way to present a Morse function
- **Vertices** of the graph are **critical points**
- **Arcs** of the graph are connected components of the level sets of f, **contracted** to points

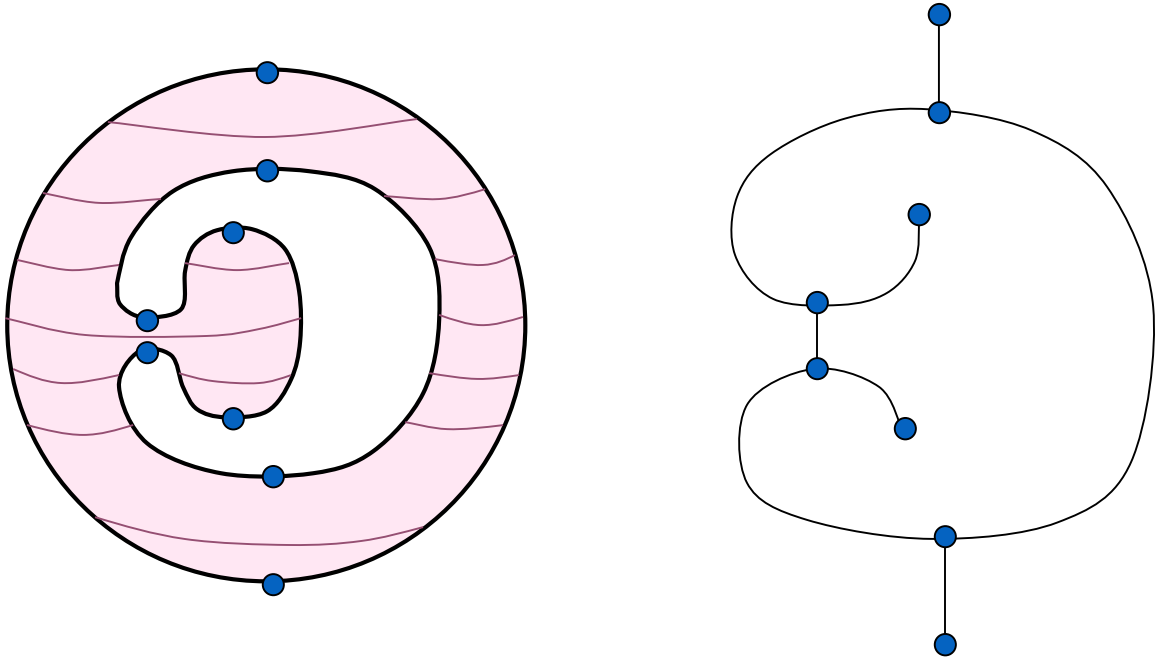


Reeb graphs and genus

- The number of loops in the Reeb graph is equal to the surface genus
- To count the loops, simplify the graph by contracting degree-1 vertices and removing degree-2 vertices

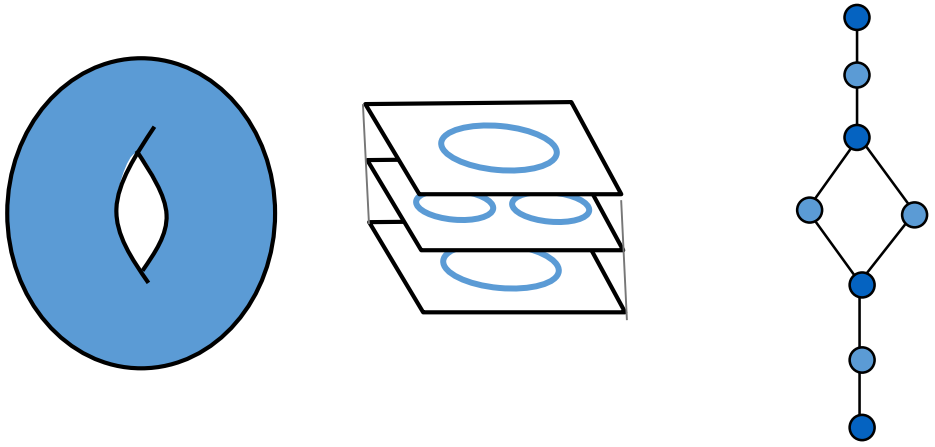


Another Reeb graph example



Discretized Reeb graph

- Take the critical points and “samples” in between
- Robust because we know that nothing happens between consecutive critical points

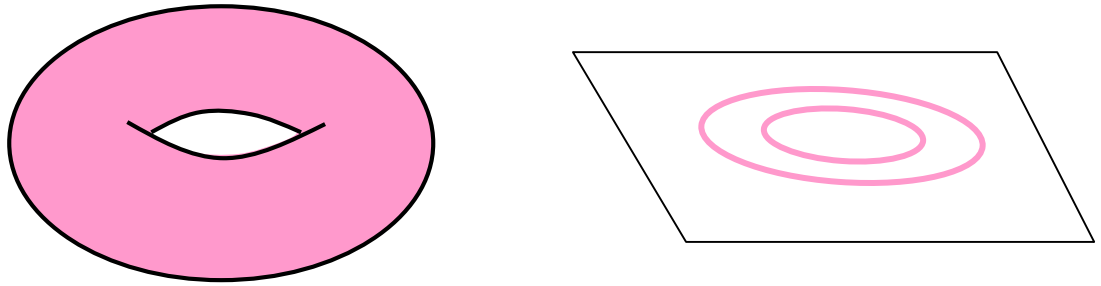


Reeb graphs for Shape Matching

- Reeb graph encodes the behavior of a Morse function on the shape
- Also tells us about the topology of the shape
- Take a meaningful function and use its Reeb graph to compare between shapes!

Choose the right Morse function

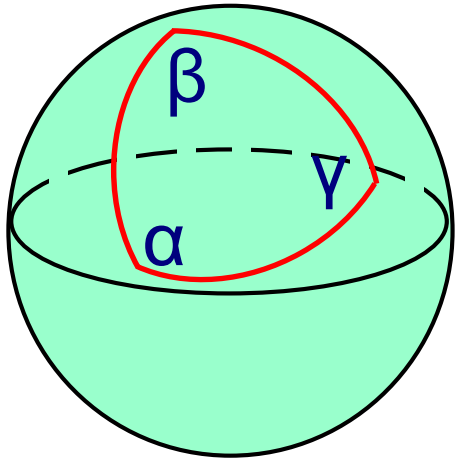
- The height function $f(\mathbf{p}) = z$ is not good enough – not rotational invariant
- Not always a Morse function



Constant curvature K

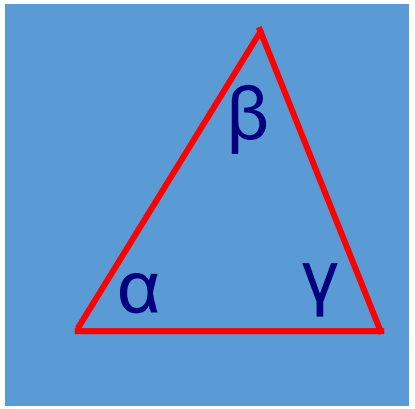
Sphere K>0

($K = 1/R^2$)



$\alpha + \beta + \gamma > 180$

Plane K =0

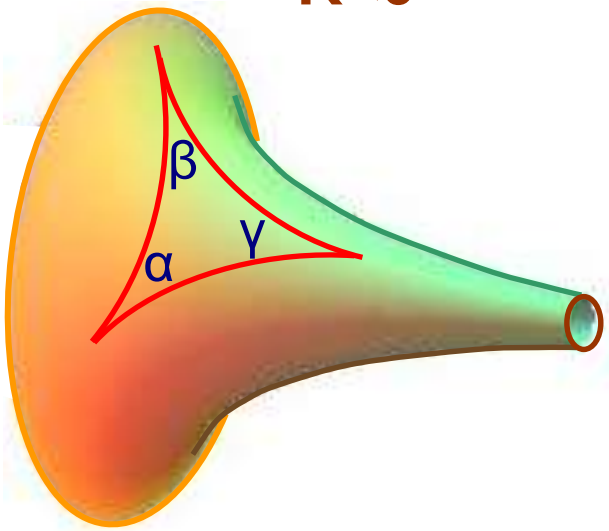


$\alpha + \beta + \gamma = 180$

Pseudosphere

(part of Hyperbolic plane)

K<0



$\alpha + \beta + \gamma < 180$

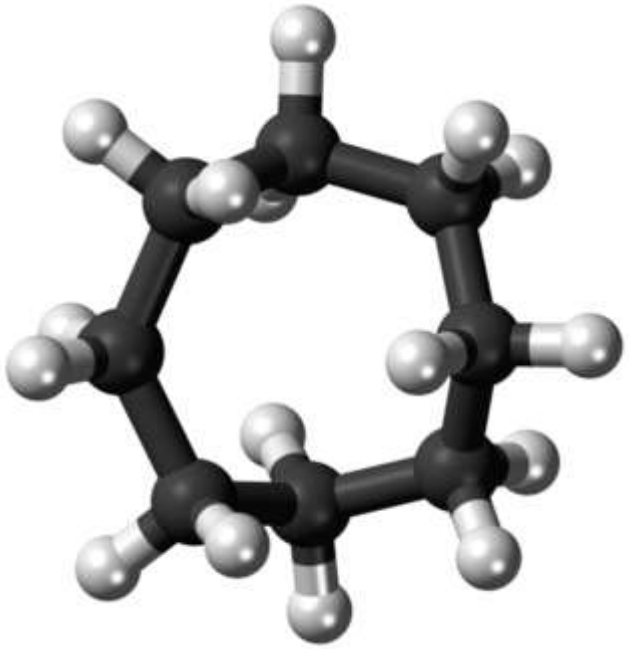
Three geometries …
and Three models of the Universe

Elliptic	Euclidean (flat)	Hyperbolic
K > 0	**K = 0**	**K < 0**
$\alpha + \beta + \gamma > 180$	$\alpha + \beta + \gamma = 180$	$\alpha + \beta + \gamma < 180$

Topology Example -- Cyclooctane

Cyclooctane is molecule with formula C_8H_{16}
To understand molecular motion we need characterize the molecule‘s possible shapes.

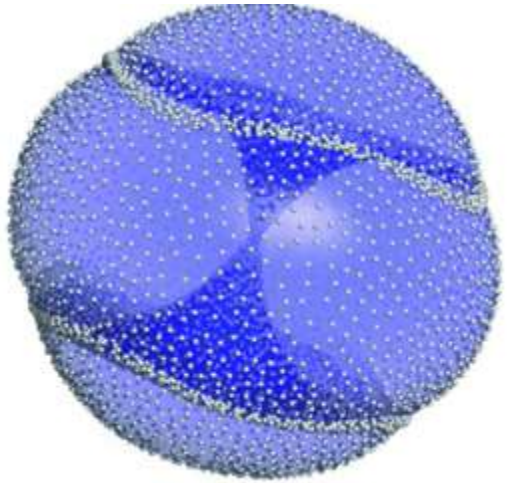
Cyclooctane has 24 atoms and it can be viewd as point in 72 dimensional spaces.



A. Zomorodian. Advanded in Applied and Computational Topology, Proceedings of Symposia in Applied Mathematics, vol 70, AMS, 2012

Topology Example -- Cyclooctane‘s space

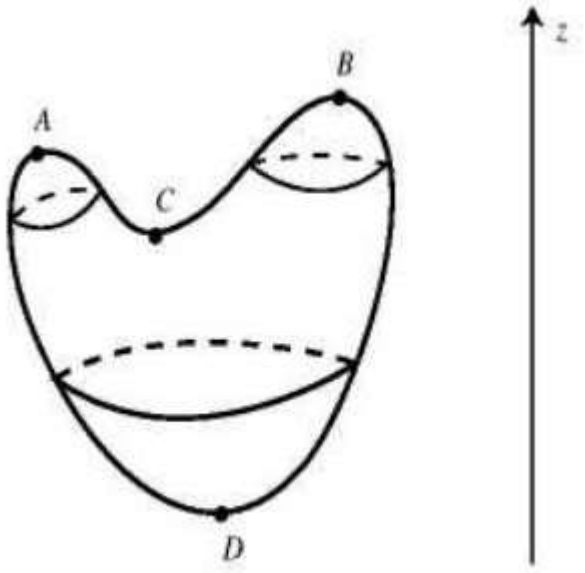
- The conformation space of cyclooctane is a two-dimensional surface with self intersection.



W. M. Brown, S. Martin, S. N. Pollock, E. A. Coutsias, and J.-P. Watson. Algorithmic dimensionality reduction for molecular structure analysis. Journal of Chemical Physics, 129(6):064118, 2008.

Information geometry

- Information geometry is a branch of mathematics that applies the techniques of differential geometry to the field of probability theory. This is done by taking probability distributions for a statistical model as the points of a Riemannian manifold, forming a statistical manifold.



Concept drift as Morse function on a statistical manifold

Shun'ichi Amari, Hiroshi Nagaoka - *Methods of information geometry*, Translations of mathematical monographs; v. 191, American Mathematical Society, 2000

Topology

- **Qualitative information is needed**: One important goal of data analysis is to allow the user to obtain knowledge about the data, i.e. to understand how it is organized on a large scale.
- **Metrics are not theoretically justified**: In physics, the phenomena studied often support clean explanatory theories which tell one exactly hat metric to use. In biological problems, on the other hand, this is much less clear. In the biological context, notions of distance are constructed using some intuitively attractive measures of similarity
- **Coordinates are not natural**: Although we often receive data in the form of vectors of real numbers, it is frequently the case that the coordinates, like the metrics mentioned above.
- **Summaries are more valuable than individual parameter choices:** One method of clustering a point cloud is the so-called *single linkage clustering*, in which a graph is constructed whose vertex set is the set of points in the cloud, and where two such points are connected by an edge if their distance is $\leq \epsilon$, where ϵ is a parameter. Some work in clustering theory has been done in trying to determine the optimal choice of $\leq \epsilon$, but it is now well understood that it is much more informative to maintain the entire *dendogram* of the set, which provides a summary of the behavior of clustering under all possible values of the parameter at once. It is therefore productive to develop other mechanisms in which the behavior of invariants or construction under a change of parameters can be effectively summarized.

Topology

- Topology is exactly that branch of mathematics which deals with qualitative geometric information. This includes the study of what the connected components of a space are, but more generally it is the study of connectivity information, which includes the classification of loops and higher dimensional surfaces within the space. This suggests that extensions of topological methodologies, such as homology, to point clouds should be helpful in studying them qualitatively.
- Topology studies geometric properties in a way which is much less sensitive to the actual choice of metrics than straightforward geometric methods, which involve sensitive geometric properties such as curvature.

Topology

- Topology studies only properties of geometric objects which do not depend on the chosen coordinates, but rather on intrinsic geometric properties of the objects. As such, it is coordinate-free.
- The idea of constructing summaries over whole domains of parameter values involves understanding the relationship between geometric objects constructed from data using various parameter values. The relationships which are useful involve continuous maps between the different geometric objects, and therefore become a manifestation of the notion of *functoriality*, i.e, the notion that invariants should be related not just to objects being studied, but also to the maps between these objects.
  - Functoriality is central in algebraic topology in that the functoriality of homological invariants is what permits one to compute them from local information, and that functoriality is at the heart of most of the interesting applications within mathematics. Moreover, it is understood that most of the information about topological spaces can be obtained through diagrams of discrete sets, via a process of simplicial approximation.

What topology can do?

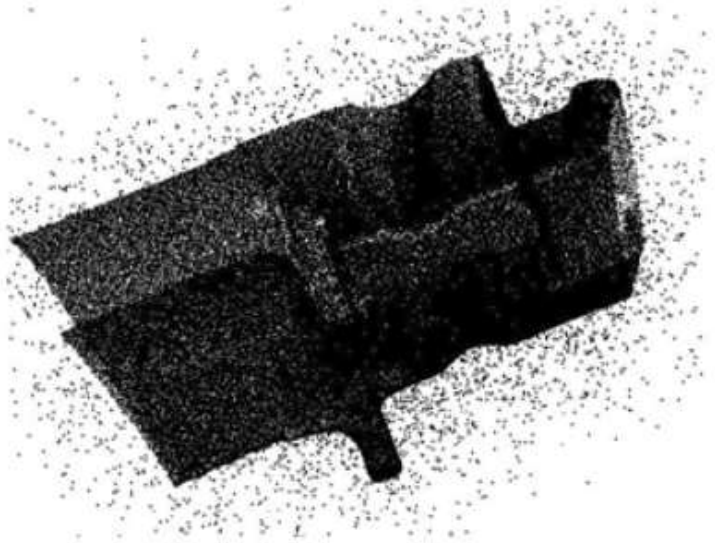
- Characterization: Topological properties encapsulate qualitative signatures e.g. the genus of surface, number of connected components, give global characteristics important to classification.
- Continuation: Topological features are robust. The number of components or holes is not something that changes with a small error of measurement. This is vital to application in scientific disciplines, where data is very noisy.

What topology can do?

- Integration: Topology is the premiere tool for converting local data into global properties. Algebraic topology tools (Homology) integrate local properties to global.
- Obstruction: Topology often provides tools for answering feasibility od certain problems, even the answer to the problems themselves are hard to compute. These characteristics, classes, degrees, indices, or obstruction take the form of algebraic-topological entities.

Topology an Example

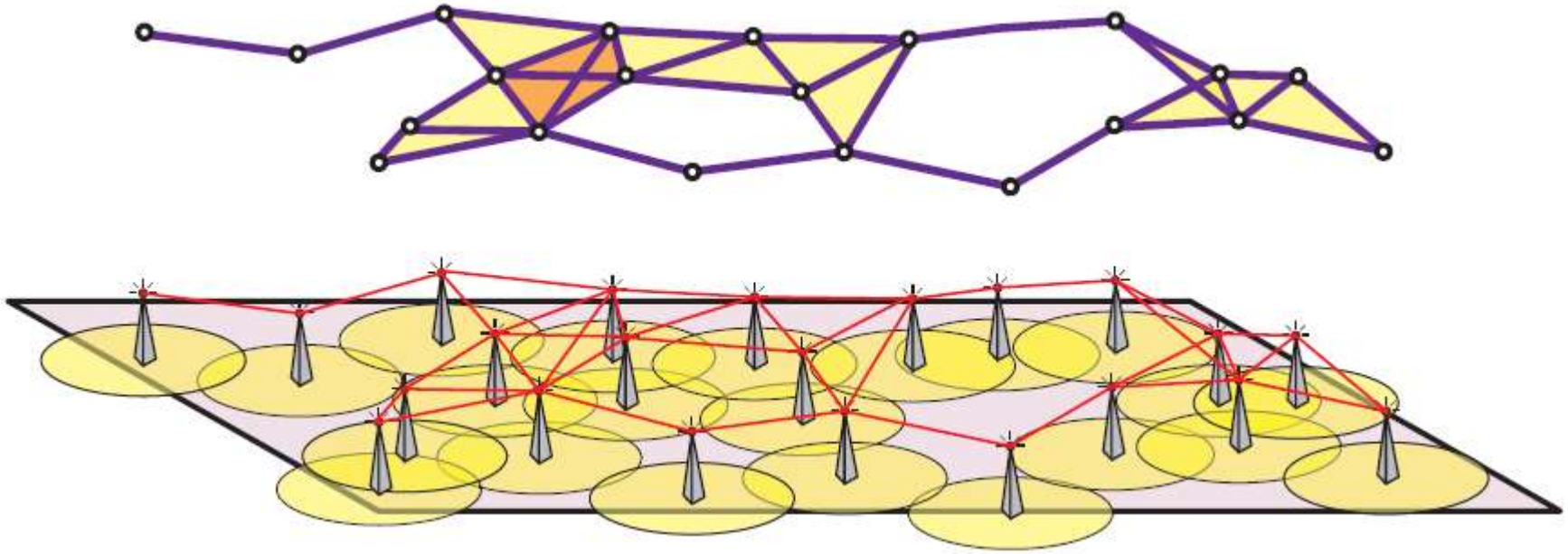
- Input:
  - A set of points P sampled from a probabilistic measure μ on R^d potentially concentrated on a hidden compact (e.g, manifold) X.
- Goal:
  - Approximate topological features of X



When to use Topological Data Analysis (TDA)?

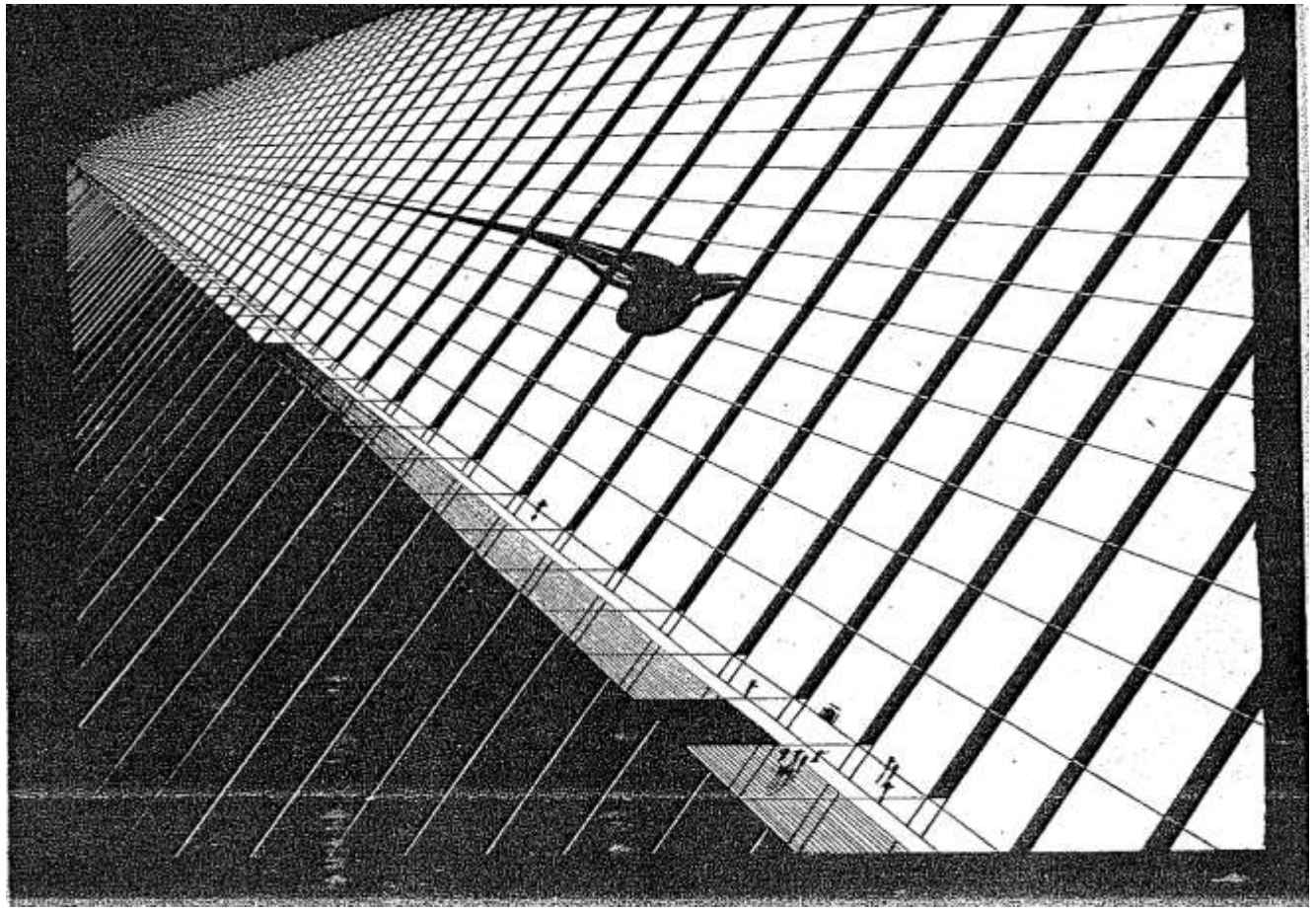
- To study complex high-dimensional data: feature selections are not required in TDA.
- Extracting shapes (patterns) of data.
- Insights qualitative information is needed.
- Summaries are more valuable than individual parameter choices.

Homological Sensor Networks



A network of small, local sensors samples an environment at a set of nodes.
How can one answer global questions from this network of local data?

High dimensional space



Dimensionality of Big data

- Many researchers regard the curse of dimensionality as one aspect of Big Data problems. Indeed, Big Data should not be constricted in data volume, but all take the high-dimension characteristic of data into consideration.
- In fact, processing high-dimensional data is already a tough task in current scientific research.
- The state-of-the-art techniques for handling high-dimensional data intuitively fall into dimension reduction. Namely, we try to map the high-dimensional data space into lower dimensional space with less loss of information as possible.

Dimensionality of Big data

- There are a large number of methods to reduce dimension. Linear mapping methods, such as principal component analysis (PCA) and factor analysis, are popular linear dimension reduction techniques. Non-linear techniques include kernel PCA, manifold learning techniques such as Isomap, locally linear embedding (LLE), Hessian LLE, Laplacian eigenmaps.
- Recently, a generative deep networks, called auto encoder, perform very well as non-linear dimensionality reduction.
- Random projection in dimensionality reduction also have been well-developed.

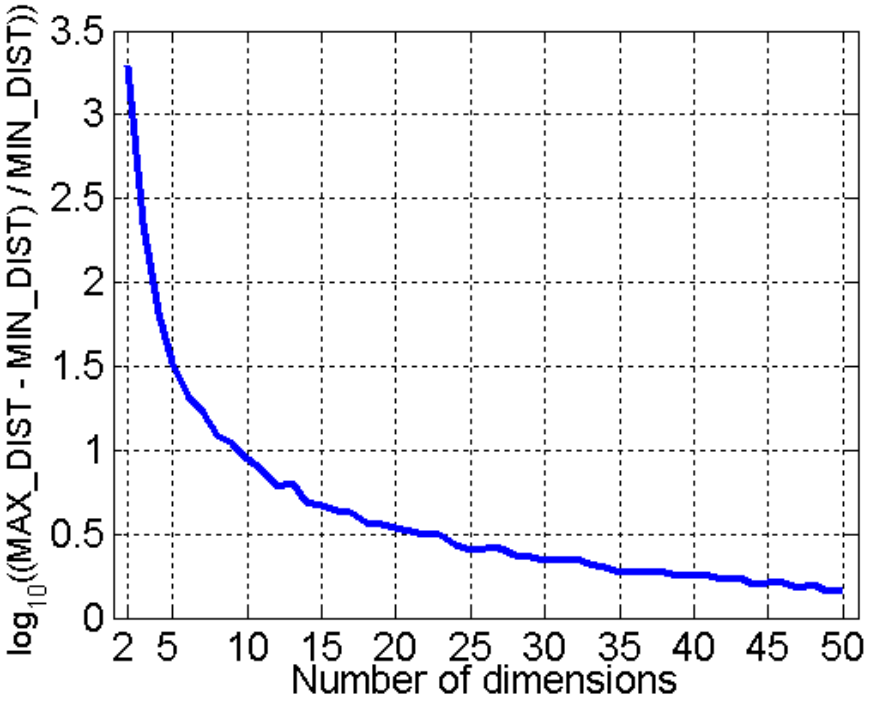
Curse of Dimensionality

- The curse of dimensionality is a term coined by Richard Bellman to describe the problem caused by the exponential increase in volume associated with adding extra dimensions to a space.

- Bellman, R.E. 1957. Dynamic Programming. Princeton University Press, Princeton, NJ.

Curse of Dimensionality

- When dimensionality increases, data becomes increasingly sparse in the space that it occupies
- Definitions of density and distance between points, which is critical for data mining, become less meaningful



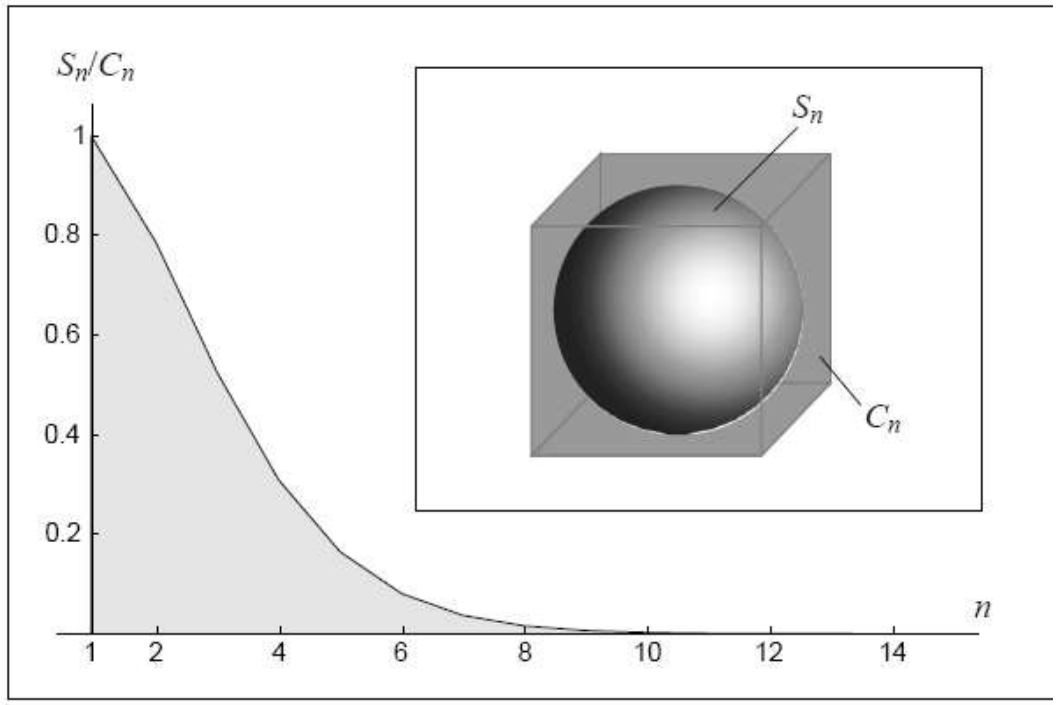
Randomly generate 500 points

Compute difference between max and min distance between any pair of points any pair of points

Curse of Dimensionality

The volume of an n-dimensional sphere with radius r is

$$V_n(r) = \frac{\pi^{\frac{n}{2}} r^n}{\Gamma\left(\frac{n}{2}+1\right)}$$



dimension

Ratio of the volumes of unit sphere and embedding hypercube of side length 2 up to the dimension 14.

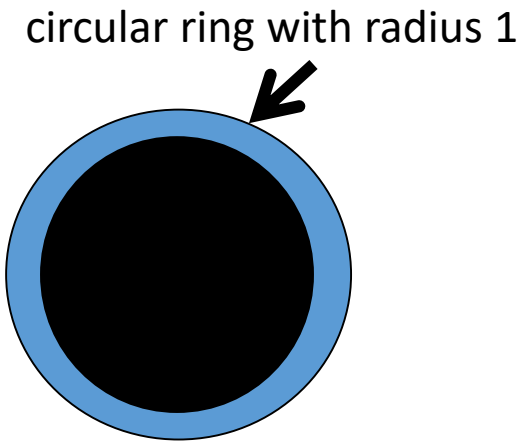
Curse of Dimensionality

The volume of an n-dimensional sphere with radius r is

$$V_n(r) = \frac{\pi^{\frac{n}{2}} r^n}{\Gamma\left(\frac{n}{2} + 1\right)}$$

Ratio of volume of n-dimensional sphere with radius 20 volume of circular ring with radius 1 is

$$R_n(r) = \frac{V_n(r) - V_n(r-1)}{V_n(r)}$$

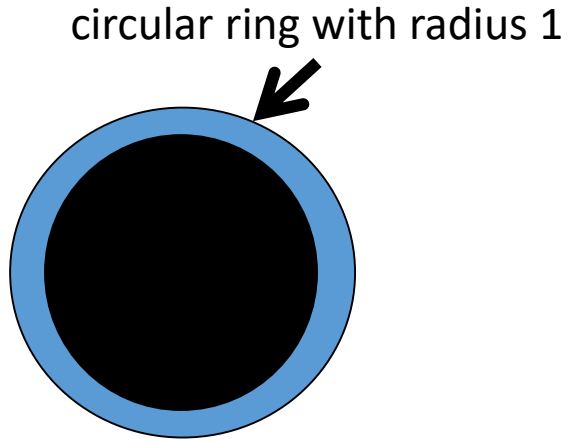


Curse of Dimensionality

2-dimension case

$$R_2(20) = \frac{V_2(20) - V_2(19)}{V_2(20)} = \frac{20^2 - 19^2}{20^2} =$$

$$= \frac{20^2 - (20-1)^2}{20^2} = \frac{1}{10}$$



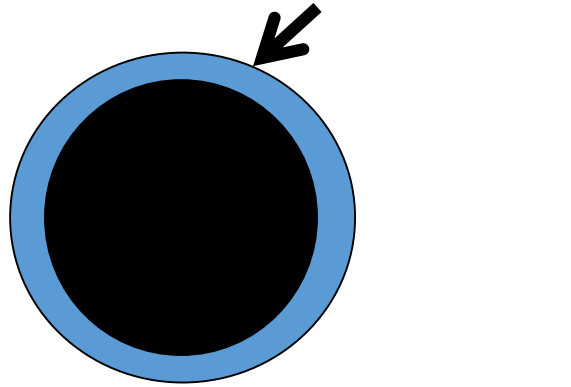
Curse of Dimensionality

20-dimension case

$$R_{20}(20) = \frac{V_{20}(20) - V_{20}(19)}{V_{20}(20)} = \frac{20^{20} - 19^{20}}{20^{20}} =$$

$$= \frac{20^{20} - (20-1)^{20}}{20^{20}} = 1 - \left(1 - \frac{1}{20}\right)^{20}$$

$$\left(1 - \frac{1}{20}\right)^{20} \cong \frac{1}{e} \cong \frac{1}{3} \Rightarrow R_{20}(r) = \frac{2}{3}$$



N-dimensional cube

Problem. What is the maximum or minimum area of an i-dimensional cross section of I^n?

i	1	2	3	4	5	6	7	8	9	10	11	12
$\alpha(3,i)$	$\sqrt{3}$	$\sqrt{2}$	1									
$\alpha(4,i)$	2	2	$\sqrt{2}$	1								
$\alpha(5,i)$	$\sqrt{5}$	??	2	$\sqrt{2}$	1							
$\alpha(6,i)$	$\sqrt{6}$	3	$\sqrt{8}$	2	$\sqrt{2}$	1						
$\alpha(7,i)$	$\sqrt{7}$	??	??	$\sqrt{8}$	2	$\sqrt{2}$	1					
$\alpha(8,i)$	$\sqrt{8}$	4	??	4	$\sqrt{8}$	2	$\sqrt{2}$	1				
$\alpha(9,i)$	3	??	$\sqrt{27}$	??	4	$\sqrt{8}$	2	$\sqrt{2}$	1			
$\alpha(10,i)$	$\sqrt{10}$	5	??	??	$\sqrt{32}$	4	$\sqrt{8}$	2	$\sqrt{2}$	1		
$\alpha(11,i)$	$\sqrt{11}$	??	??	??	??	$\sqrt{32}$	4	$\sqrt{8}$	2	$\sqrt{2}$	1	
$\alpha(12,i)$	$\sqrt{12}$	6	8	9	??	8	$\sqrt{32}$	4	$\sqrt{8}$	2	$\sqrt{2}$	1

α(*n, i*) denote the maximum area

Chuanming Zong, What Is Known About Unit Cubes, Bulletin of The American Mathematical Society, Volume 42, Number 2, Pages 181–211, 2005
Chuanming Zong, The Cube: A Window to Convex and Discrete Geometry, Cambridge University Press 2006

Curse of Dimensionality

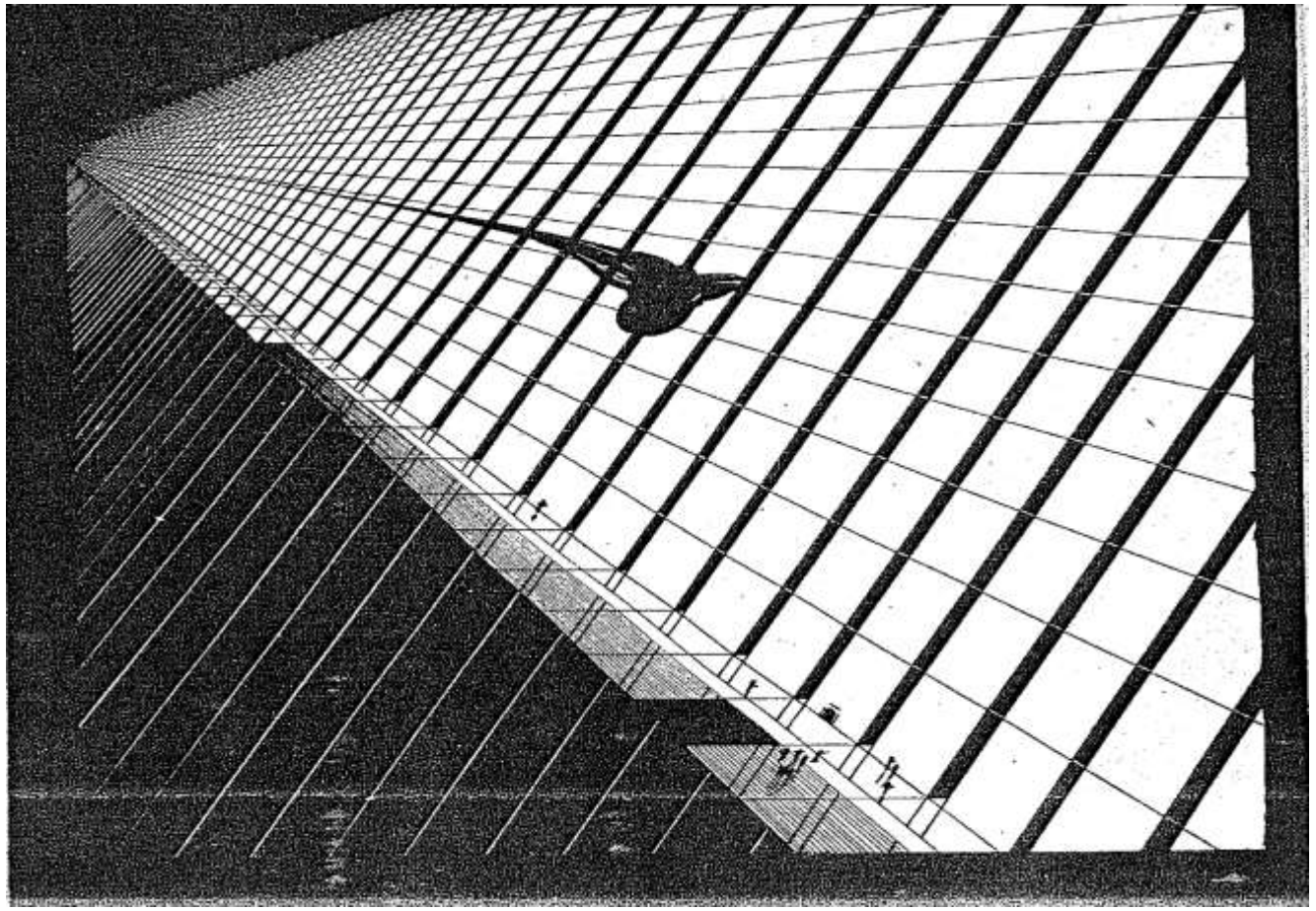
- **The model space is EMPTY!**

  (in huge dimension all volume is in surface)

- **Distribution of data is uniform!**

  (in huge dimension all distance is being uniform)

Ultra metrics



The Ordinary Absolute Value

The ordinary absolute value on $\mathbb{Q}$ is defined as follows:

$$|.| : \mathbb{Q} \to \mathbb{R}_+$$

$$|x| = \begin{cases} x : x \geq 0 \\ -x : x < 0 \end{cases}$$

This satisfied the required conditions.

The Rationals as a Metric Space

$\mathbb{Q}$ forms a metric space with the ordinary absolute value as our distance function.

We write this metric space as $(\mathbb{Q}, |.|)$

If X is a set, then a metric on X is a function d

The metric, d, is defined in the obvious way:

$$d: \mathbb{Q} \times \mathbb{Q} \to \mathbb{R}_+$$
$$d(x, y) = |x - y|$$

Cauchy Sequences

A Cauchy sequence in a metric space is a sequence whose elements become „close“ to each other.

A sequence

$$x_1, x_2, x_3, x_4 \cdots$$

is called Cauchy if for every positive (real) number ε, there is a positive integer N such that for all natural numbers $n, m > N$,

$$d(x_m, x_n) = |x_m, x_n| < \varepsilon$$

Complete Metric Space

We call a metric space (X, d) complete if every Cauchy sequence in (X, d) converges in (X, d)

Concrete example: the rational numbers with the ordinary distance function, $(\mathbb{Q}, |.|)$ is not complete.

Example: $(\sqrt{2})$

$$1, 1.4, 1.41, 1.414, \ldots$$

Completing $\mathbb{Q}$ to get $\mathbb{R}$

If a metric space is not complete, we can complete it by adding in all the „missing“ points.

For $(\mathbb{Q}, |.|)$, we add all the possible limits of all the possible Cauchy sequences.

We obtain $\mathbb{R}$.

It can be proven that the completion of field gives a field.

Since $\mathbb{Q}$ is a field, $\mathbb{R}$ is field.

The p-adic Absolute Value

For each prime p, there is associated p-adic absolute value $|.|_p$ on $\mathbb{Q}$.

Definition. Let p be any prime number. For any nonzero integer a, let $ord_p a$ be the highest power of p which divides a , i.e., the greatest m such that $a \equiv 0 \ (mod\ p^m)$.

$ord_p ab = ord_p a + ord_p b, \qquad ord_p\ a/b = ord_p a - ord_p\ b,$

Examples:

$ord_5 35 = 1, ord_5 77 = 0, ord_2 32 = 5$

The p-adic Absolute Value

Further define absolute value $|.|_p$ on $\mathbb{Q}$ as follows: ($a \in \mathbb{Q}$)

$$|a|_p = \begin{cases} p^{-ord_p a}, & a \neq 0 \\ 0, & a = 0 \end{cases}$$

Proposition. $|.|_p$ is a norm on $\mathbb{Q}$.

Example: $|\frac{968}{9}|_{11} = |11^2.\ \frac{8}{9}|_{11} = 11^{-2}$

Completing $\mathbb{Q}$ a different way

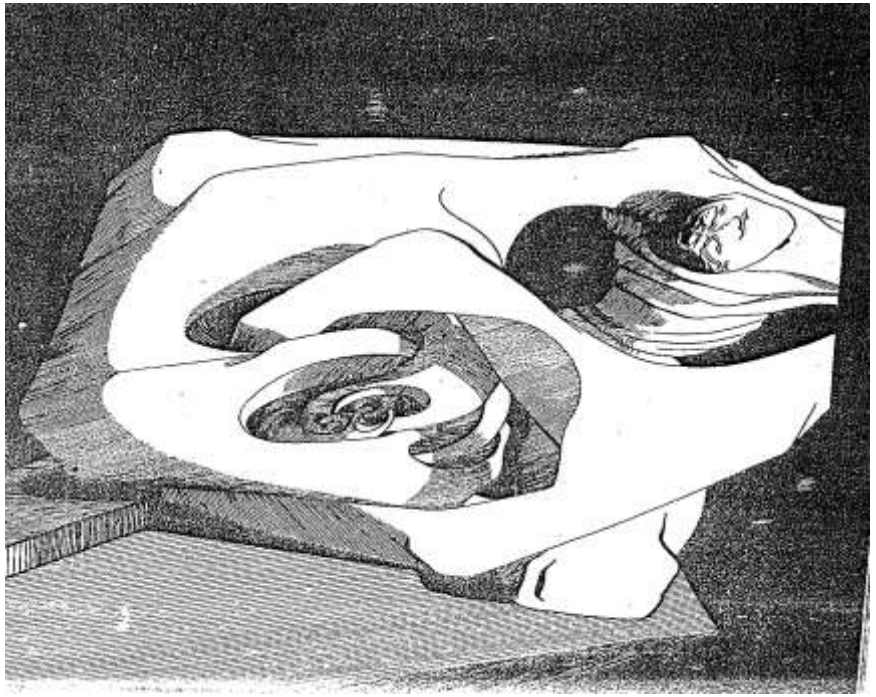
The p-adic absolute value give us a metric on $\mathbb{Q}$ defined by

$$d: \mathbb{Q} \times \mathbb{Q} \to \mathbb{R}_+$$
$$d(x, y) = |x - y|_p$$

When $p = 7$ we have that 7891 and 2 are closer together than 3 and 2

$$|7891 - 2|_7 = |7889|_7 = |7^3 \times 23|_7 = 7^{-3} = 1/343$$
$$|3 - 2|_7 = |1|_7 = |7^0|_7 = 7^0 = 1 > 1/343$$

Completing $\mathbb{Q}$ a different way



The p-adic absolute value give us a metric on $\mathbb{Q}$ defined by

$$d: \mathbb{Q} \times \mathbb{Q} \to \mathbb{R}_+$$
$$d(x, y) = |x - y|_p$$

When $p = 7$ we have that 7891 and 2 are closer together than 3 and 2

$$|7891 - 2|_7 = |7889|_7 = |7^3 \times 23|_7 = 7^{-3} = 1/343$$
$$|3 - 2|_7 = |1|_7 = |7^0|_7 = 7^0 = 1 > 1/343$$

Completing $\mathbb{Q}$ a different way

$\mathbb{Q}$ is not complete with respect to p-adic metric $d(x, y) = |x - y|_p$.

Example:

Let $p = 7$. The infinite sum

$$1 + 7 + 7^2 + 7^3 + 7^4 + 7^5 + \cdots$$

is certainly not element of $\mathbb{Q}$ but sequence

$$1, 1 + 7, 1 + 7 + 7^2, 1 + 7 + 7^2 + 7^3, \ldots$$

is a Cauchy sequence with respect to the 7-adic metric.

Completion of $\mathbb{Q}$ by $|x - y|_p$ gives field $\mathbb{Q}_p$: field of p-adic number.

The p-adic Absolute Value

Definition. A norm is called non-Archimedean if

$$|x + y| \leq \max(|x|, |y|)$$

always holds. A metric is called non-Archimedean if

$$d(x, z) \leq \max(d(x, y), d(y, z))$$

in particular, a metric is non-Archimedean if it is induced by a non-Archimedean norm.

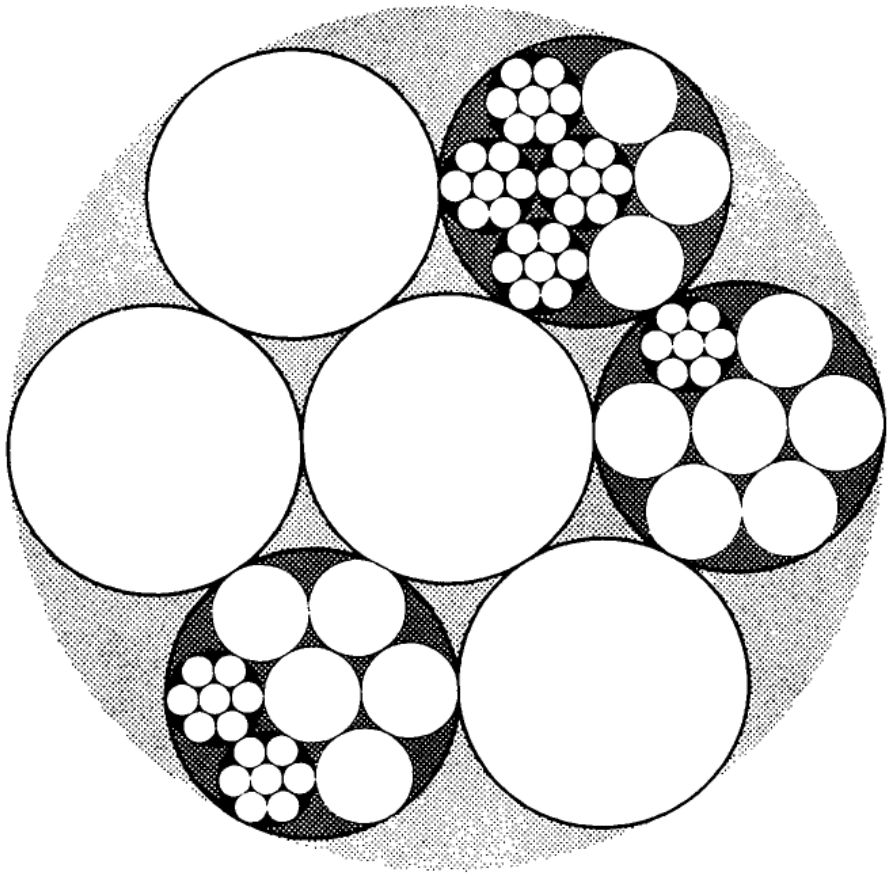
Thus, $|.|_p$ is a non-Archimedean norm on $\mathbb{Q}$.

Theorem (Ostrowski). Every nontrivial norm $|.|$ on $\mathbb{Q}$ is equivalent to $|.|_p$ for some prime p or the ordinary absolute value on $\mathbb{Q}$.

Basic property of a non-Archimedean field

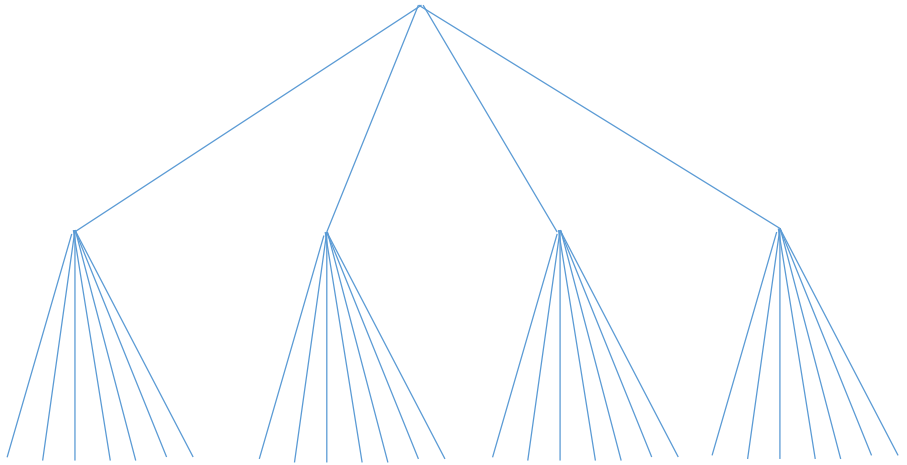
- Every point in a ball is a center!
- Set of possible distances are „small“
  $\{p^n; n \in \mathbb{Z}\}$
- every triangle is isosceles

Balls in $\mathbb{Q}_7$

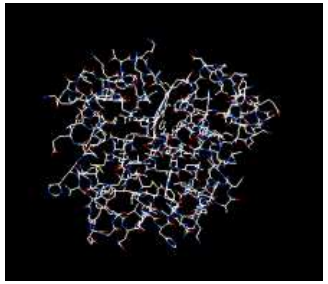


Definition. A metric space (X, d) is an ultrametric space if the metric d satisfies the strong triangle inequality $d(x, z) \leq \max(d(x, y), d(y, z))$.

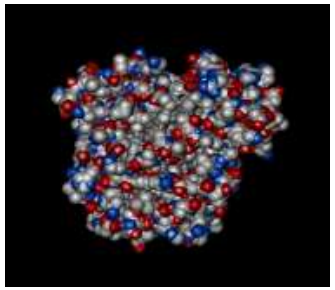
Vizialization of ultrametrics



How to define protein dynamics



Protein is a macromolecule



protein states

Protein states are defined by means of conformations of a protein macromolecule.

A conformation is understood as the spatial arrangement of all "elementary parts" of a macromolecule.

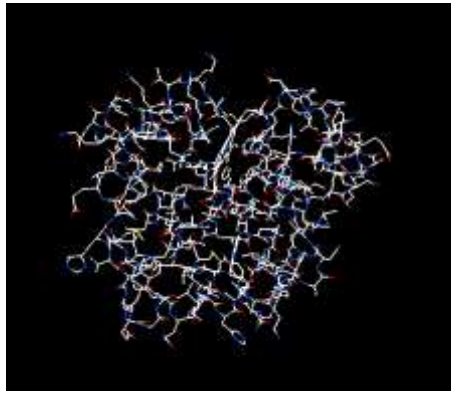
Atoms, units of a polymer chain, or even larger molecular fragments of a chain can be considered as its "elementary parts". Particular representation depends on the question under the study.

protein dynamics

Protein dynamics is defined by means of conformational rearrangements of a protein macromolecule.

Conformational rearrangements involve fluctuation induced movements of atoms, atomic groups, and even large macromolecular fragments.

Protein dynamics



To study protein motions on the subtle scales, say, from ~10^{-9} sec, it is necessary to use the atomic representation of a protein molecule.

Protein molecule consists of ~10^{3} atoms.

Protein conformational states:

number of degrees of freedom : ~ 10^3

dimensionality of (Euclidian) space of states : ~ 10^3

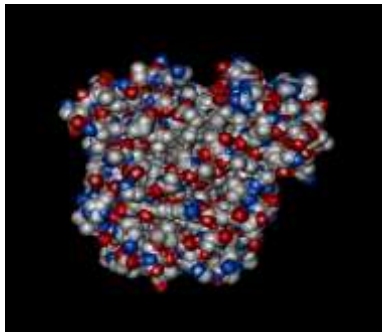
In fine-scale presentation, dimensionality of a space of protein states is very high.

Protein dynamics

Protein dynamics over high dimensional conformational space is governed by complex energy landscape.

protein energy landscape

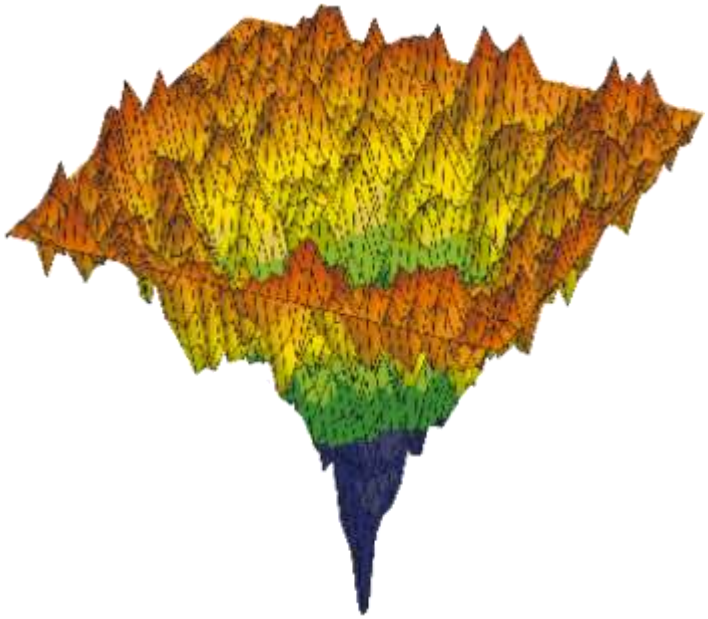
Given the interatomic interactions, one can specify the potential energy of each protein conformation, and thereby define an energy surface over the space of protein conformational states. Such a surface is called the protein energy landscape.



As far as the protein polymeric chain is folded into a condensed globular state, high dimensionality and ruggedness are assumed to be characteristic to the protein energy landscapes

Protein energy landscape: dimensionality: ~ 10^3;
number of local minima ~10^{100}

Protein dynamics



While modeling the protein motions on many time scales (from ~10^{-9} sec up to ~10^{0} sec), we need the simplified description of protein energy landscape that keeps its multi-scale complexity.

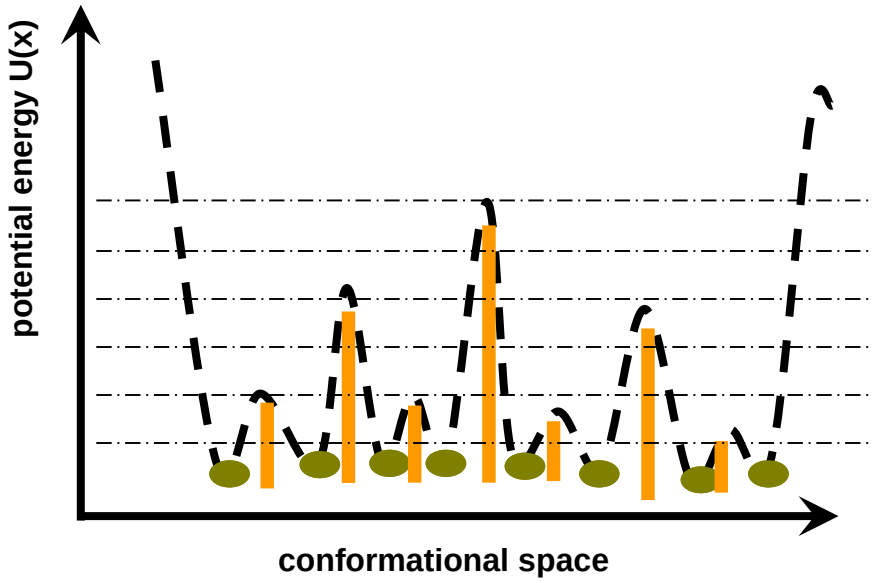
How such model can be constructed?

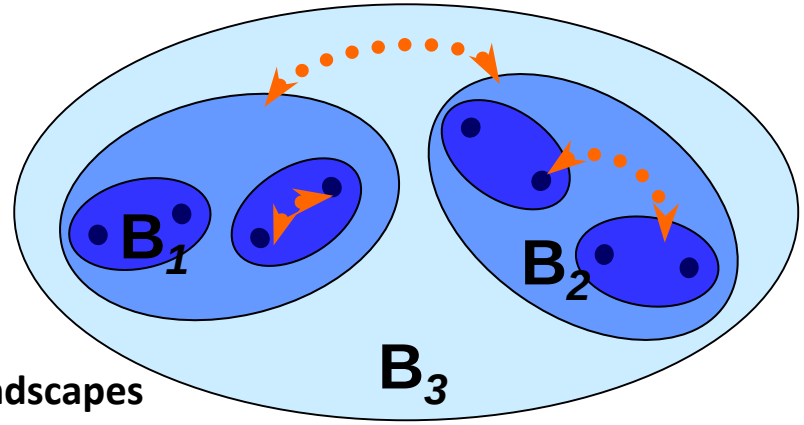
Computer reconstructions of energy landscapes of complex molecular structures suggest some ideas.

Protein dynamics

Method

1. **Computation of local energy minima and saddle points on the energy landscape using molecular dynamic simulation;**
2. **Specification a topography of the landscape by the energy sections;**
3. **Clustering the local minima into hierarchically nested basins of minima.**
4. **Specification of activation barriers between the basins.**



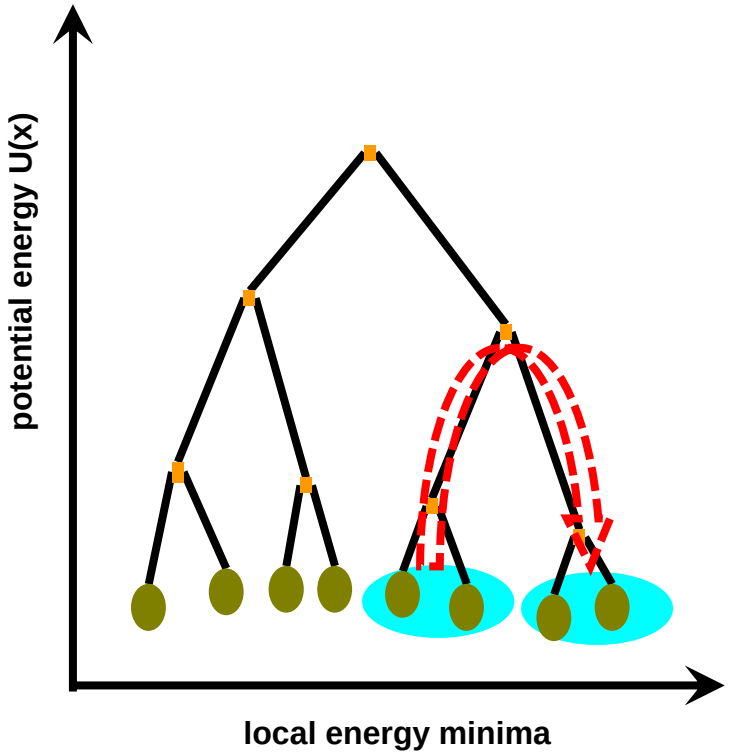


O.M.Becker, M.Karplus, Computer reconstruction of complex energy landscapes *J.Chem.Phys*. 106, 1495 (1997)

Protein dynamics

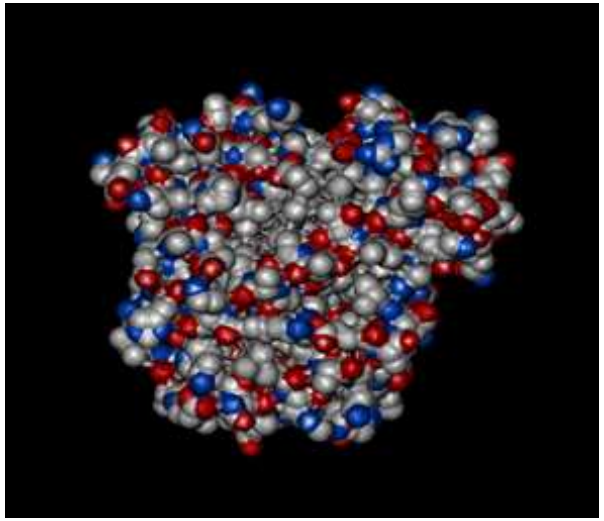
The relations between the basins embedded one into another are presented by a tree-like graph.

Such a tee is interpreted as a "skeleton" of complex energy landscape. The nodes on the border of the tree (the "leaves") are associated with local energy minima (quasi-steady conformational states). The branching vertexes are associated with the energy barriers between the basins of local minima.



O.M.Becker, M.Karplus, Presentation of energy landscapes by tree-like graphs
***J.Chem.Phys*. 106, 1495 (1997)**

Complex energy landscapes : a protein

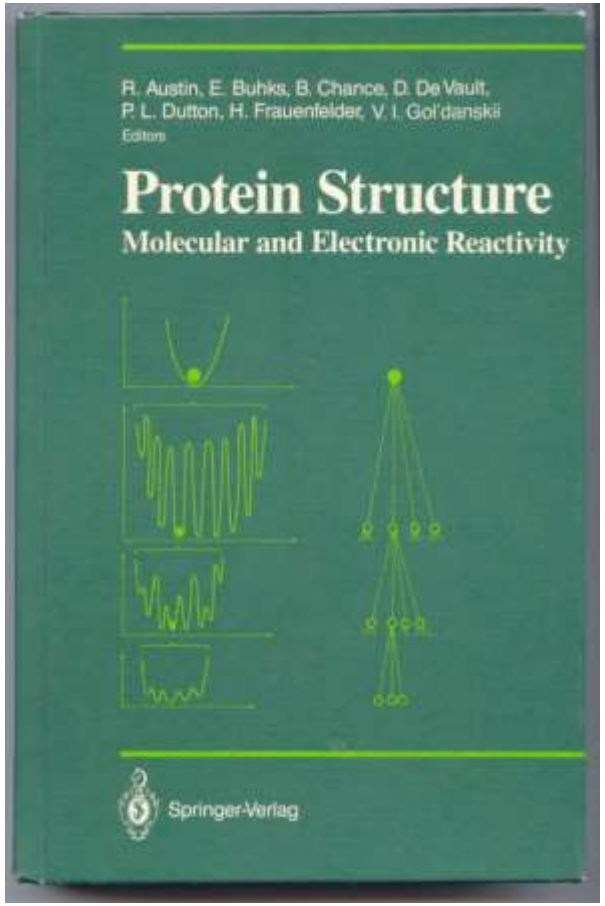


The total number of minima on the protein energy landscape is expected to be of the order of ~10^{100}.

This value exceeds any real scale in the Universe. Complete reconstruction of protein energy landscape is impossible for any computational resources.

Protein Structure

25 years ago, Hans Frauenfelder suggested a tree-like structure of the energy landscape of myoglobin



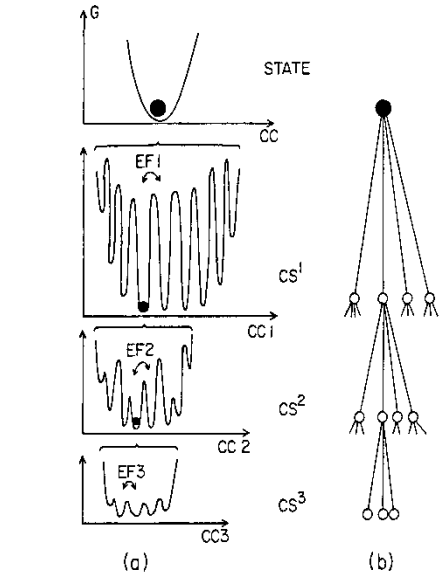


Figure 5. Hierachical arrangement of the conformational substates in myoglobin. (a) Schematized energy surfaces. (b) Tree diagram. G is the Gibbs energy of the protein, CC(1-4) are conformational coordinates. After [31].

Hans Frauenfelder, in *Protein Structure* (N-Y.:Springer Verlag, 1987) p.258.

10 years later, Martin Karplus suggested the same idea

> "In <…> proteins, for example, where individual states are usually clustered in "basins", the interesting kinetics involves basin-to-basin transitions. The internal distribution within a basin is expected to approach equilibrium on a relatively short time scale, while the slower basin-to-basin kinetics, which involves the crossing of higher barriers, governs the intermediate and long time behavior of the system."
>
> Becker O. M., Karplus M. *J. Chem. Phys.*, 1997, 106, 1495

This is exactly the physical meaning of protein ultrameticity !

Persistent homology

Persistent homology is an algebraic method for discerning topological features of data.

More persistent features are detected over a wide range of spatial scales and are considered more likely to represent true features of the underlying space rather than artifacts of sampling, noise, or particular choice of parameters.

To compute the persistent homology of a space, the space must first be represented as a simplicial complex. A distance function on the underlying space corresponds to a filtration of the simplicial complex, that is a nested sequence of increasing subsets.

Computing Persistent Homology

We start with a filtered simplicial complex:

$$\emptyset = K_0 \subset K_1 \subset \cdots \subset K_m = K$$

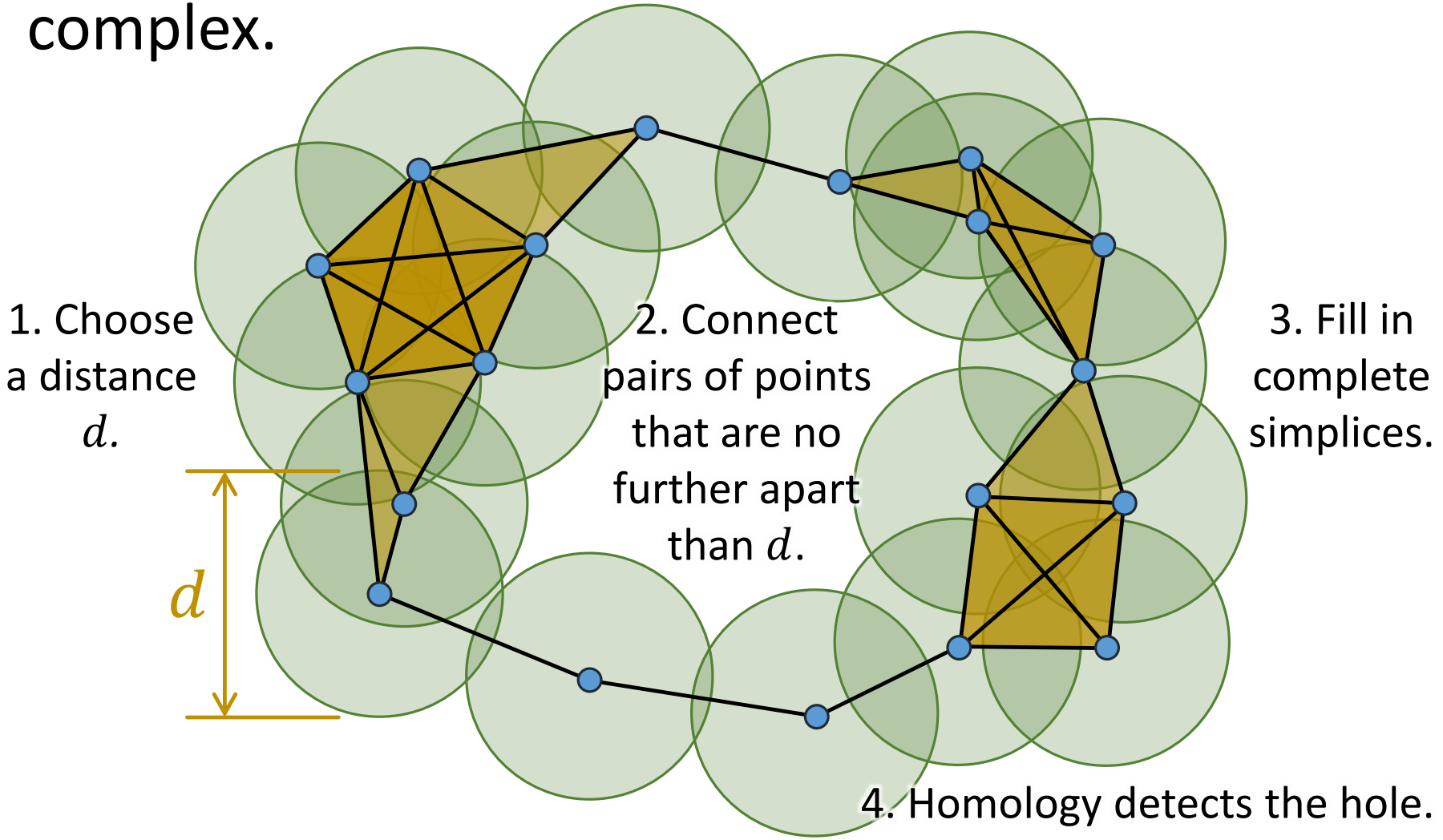
Step 1: Sort the simplices to get a total ordering compatible with the filtration.

Step 2: Obtain a boundary matrix D with respect to the total order on simplices.

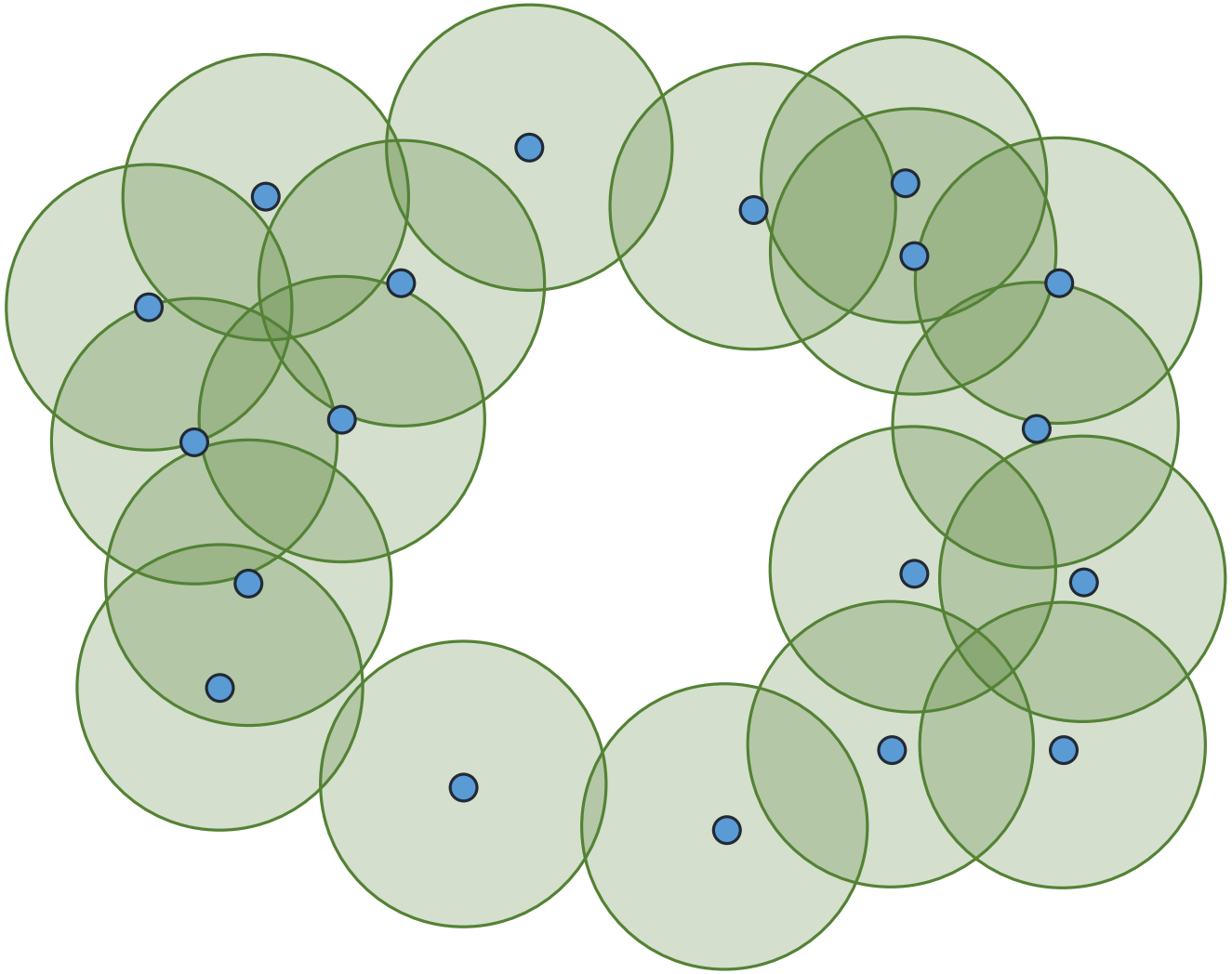
Step 3: Reduce the matrix using column additions, always respecting the total order on simplices.

Step 4: Read the persistence pairs to get the barcode.

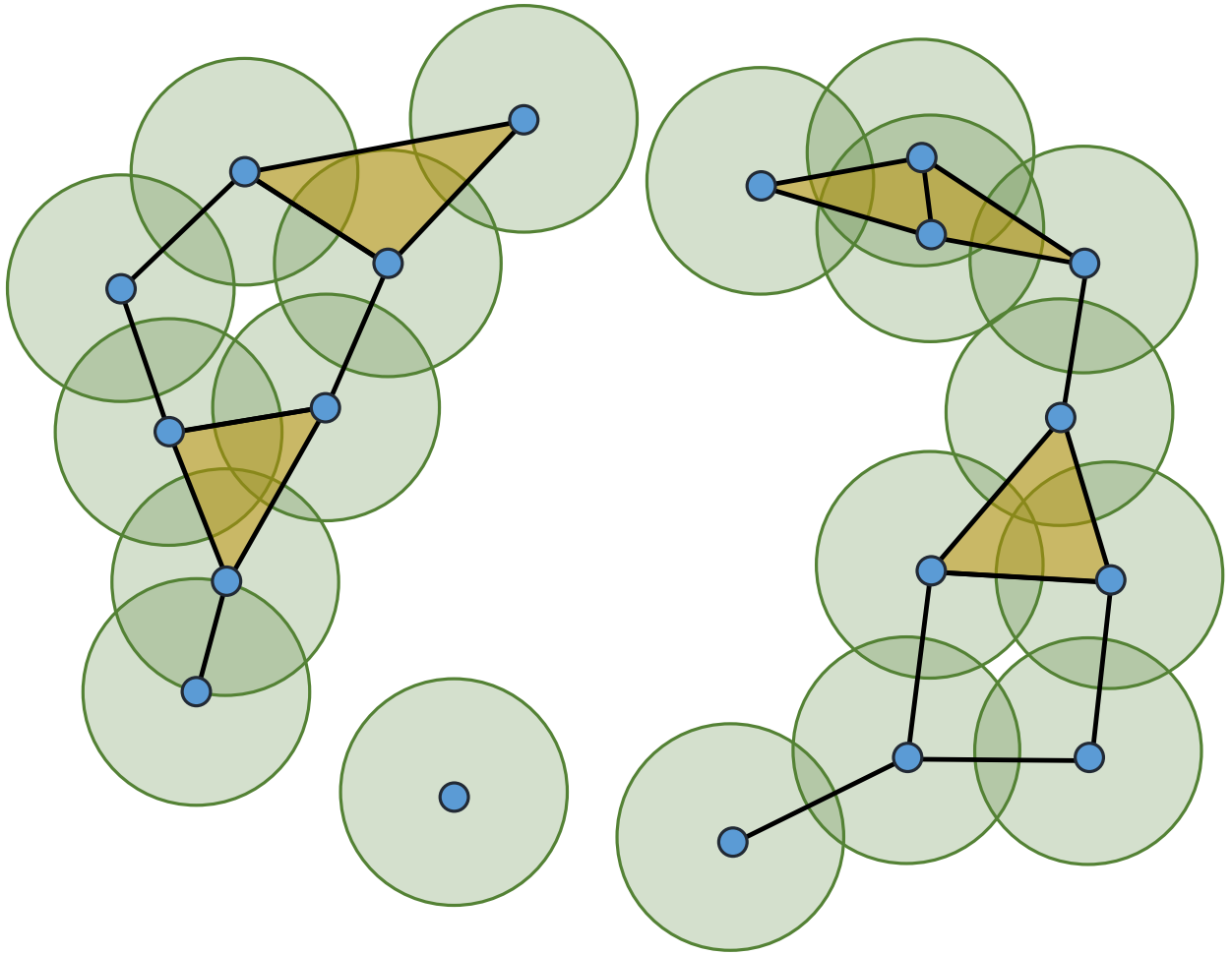
Idea: Connect nearby points, build a simplicial complex.



Problem: How do we choose distance d?

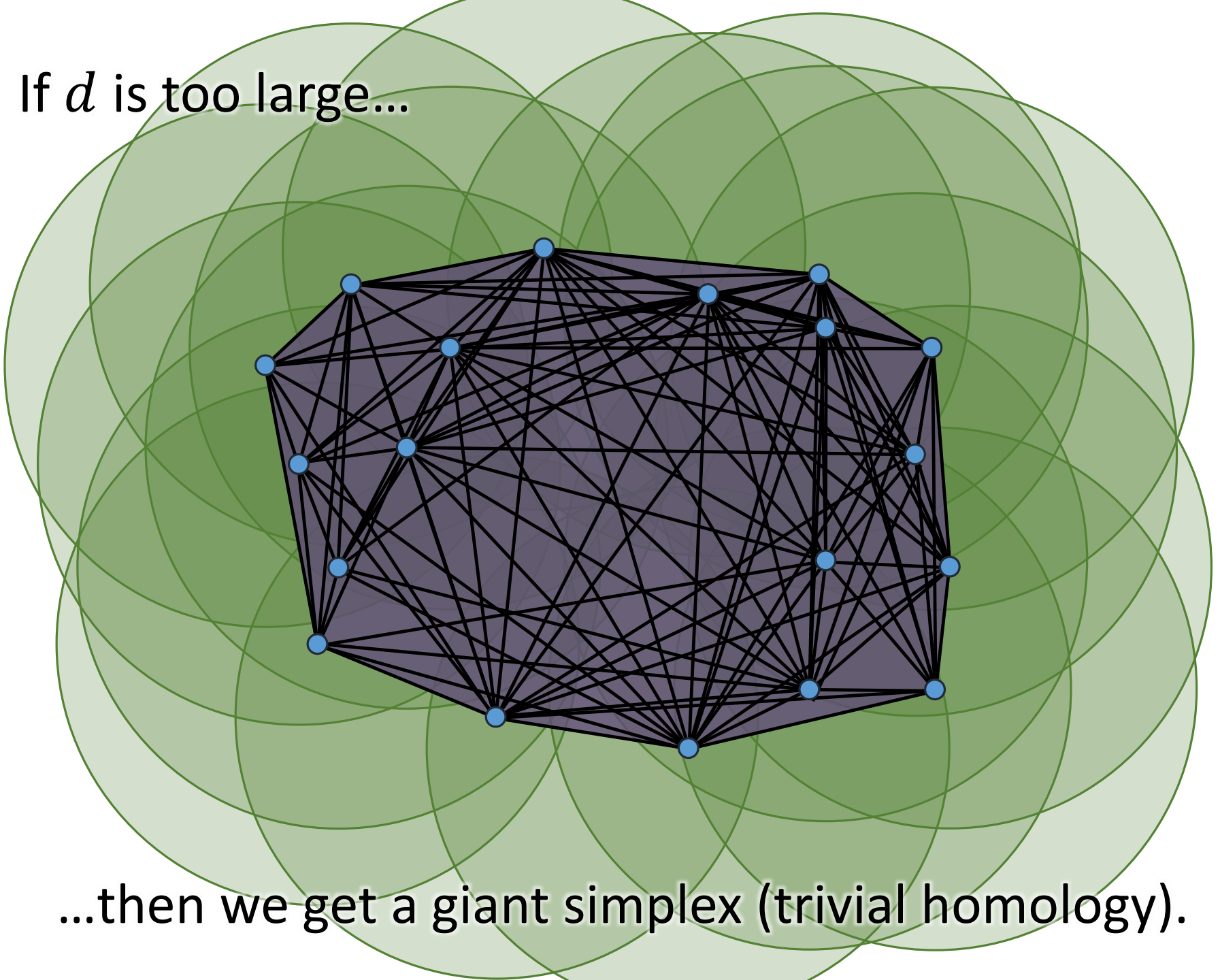


If d is too small…



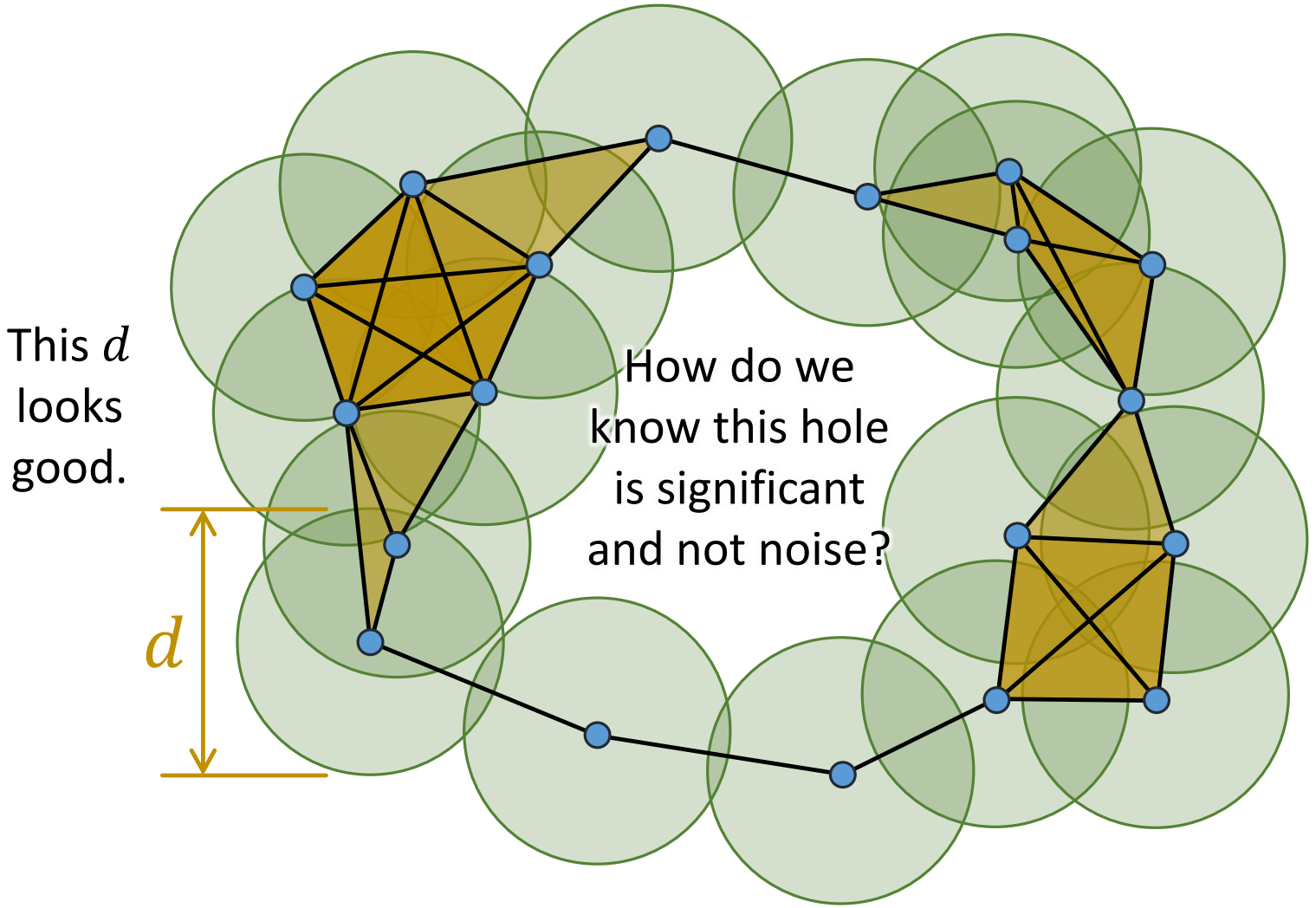
…then we detect noise.

If d is too large…



…then we get a giant simplex (trivial homology).

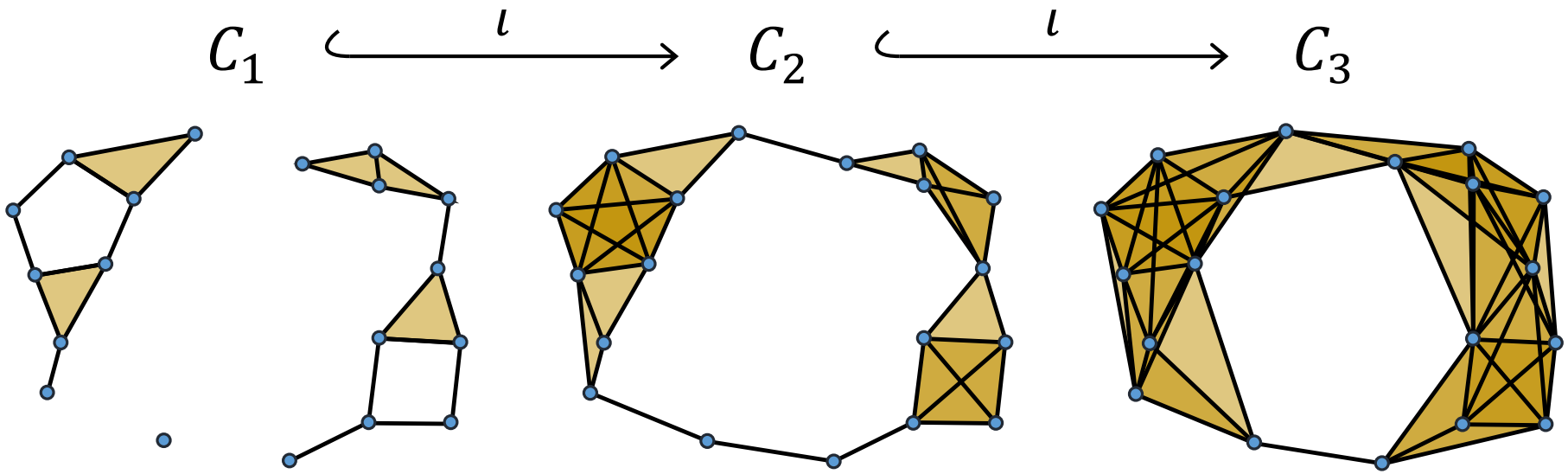
Problem: How do we choose distance d?



Idea: Consider *all* distances d.

A barcode is a visualization of an algebraic structure

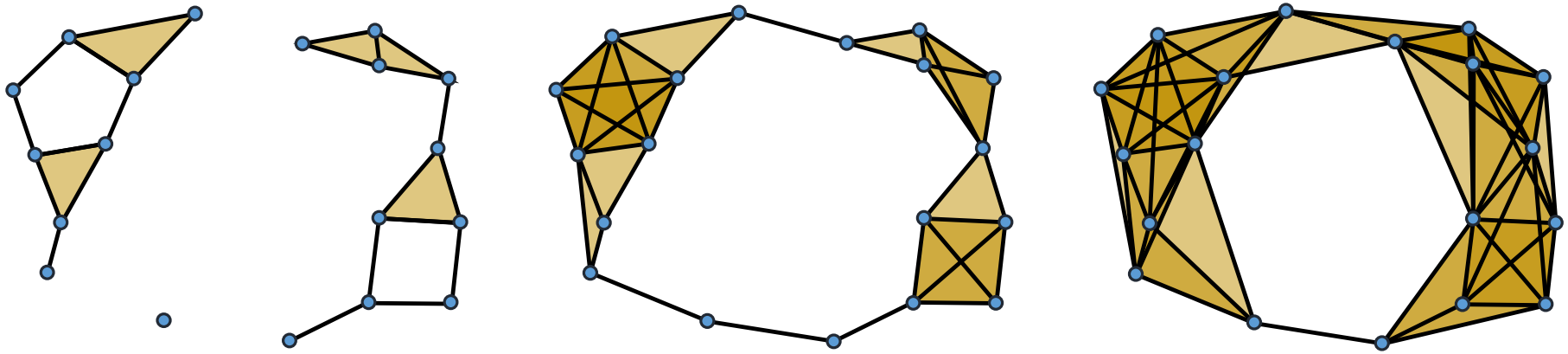
Consider the sequence (C_i) of complexes associated to a point cloud for an sequence of distance values:



A barcode is a visualization of an algebraic structure

Consider the sequence (C_i) of complexes associated to a point cloud for an sequence of distance values:

$$\cdots \hookrightarrow C_1 \hookrightarrow C_2 \hookrightarrow C_3 \hookrightarrow C_4 \hookrightarrow C_5 \hookrightarrow C_6 \hookrightarrow C_7 \hookrightarrow \cdots$$



This sequence of complexes, with maps, is a **filtration**.

A barcode is a visualization of an algebraic structure

Filtration: $C_1 \hookrightarrow C_2 \hookrightarrow \cdots \hookrightarrow C_m$

Homology with coefficients from a field F:

$$H_*(C_1) \to H_*(C_2) \to \cdots \to H_*(C_m)$$

Let $M = H_*(C_1) \oplus H_*(C_2) \oplus \cdots \oplus H_*(C_m)$.

For $i \leq j$, the map $f_i^j : H_*(C_i) \to H_*(C_j)$ is induced by the inclusion $C_i \hookrightarrow C_j$.

Let $F[x]$ act on M by $x^k \alpha = f_i^{i+k}(\alpha)$ for any $\alpha \in H_*(C_i)$.

i.e. x acts as a shift map $x : H_*(C_i) \to H_*(C_{i+1})$

Then M is a graded $F[x]$-module, called a **persistence module**.

Closed Trail Distance in a Biconnected Graph

More interconnected parts of graphs play an essential role in the social and natural sciences.
The formalization of the term "more connected part" can be defined in many ways.
Biconnected components of the graph do not allow good scalability, and their definition is complicated for weighted graphs.
Generalization biconnected components of a graph is based on the limited length cycle.

Vaclav Snasel, Pavla Drazdilova, Jan Platos, Closed trail distance in a biconnected graph, Plos One, 2018. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0202181

Closed trail distance in a undirected graph

Definition 1. *A graph is a k-closed trail connected graph (k-CT) if every two vertices lie on the closed trail (circuit) with a length $\leq k$. A k-CT component of the graph is a maximal k-CT subgraph.*

Definition 2. *A ∞-CT graph is a graph where every two vertices lie on a closed trail of any length.*

Definition 3. *Let $G = (V, E)$ be a graph. Let $d_{ct} : V \times V \to R_0^+$ be defined by the equation*

$$d_{ct}(u, v) = min_{CT(u,v) \subseteq G} |CT(u, v)|,$$

where $CT(u, v)$ is a closed trail that contains the vertices u, v. Then the function d_{ct} is called the closed trail distance (CT-distance).

Theorem 1. *The CT-distance is a metric on the set V.*

Closed trail distance example

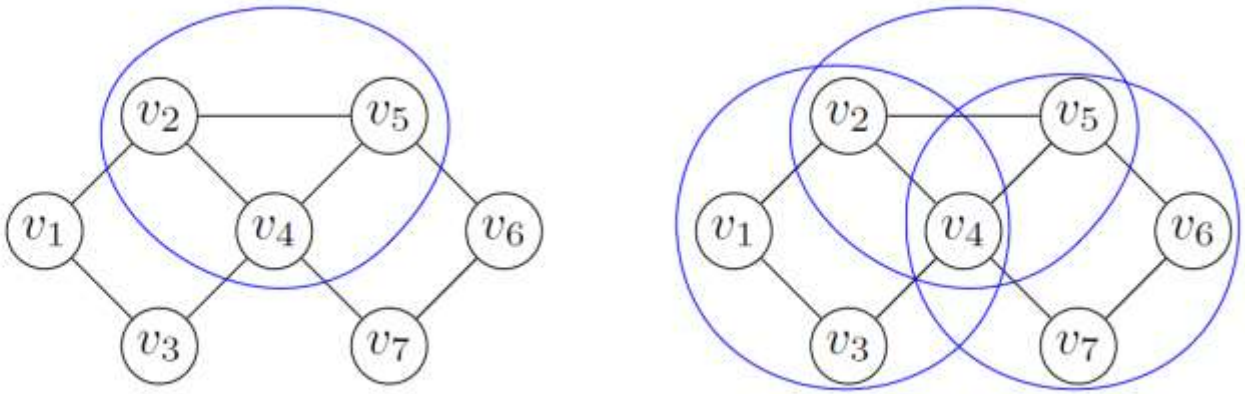


Fig 3. 3-CT and 4-CT components of the graph

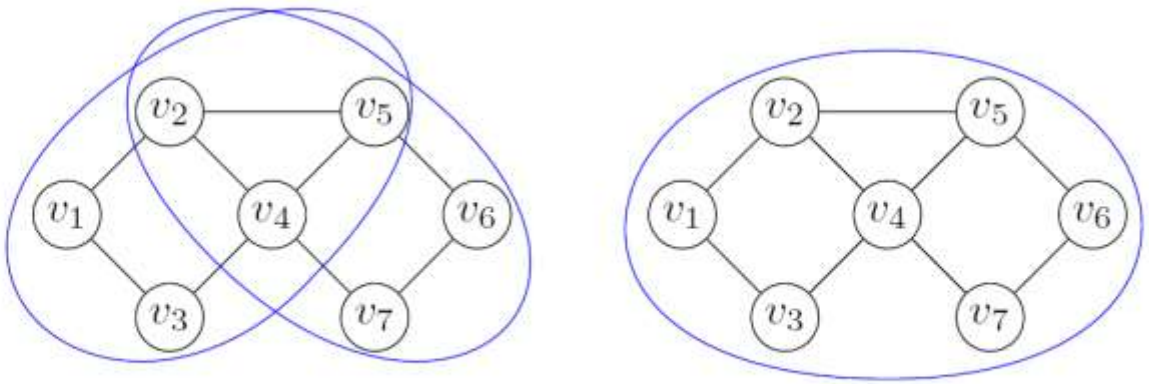


Fig 4. 5-CT and 7-CT components of the graph

Closed trail distance example

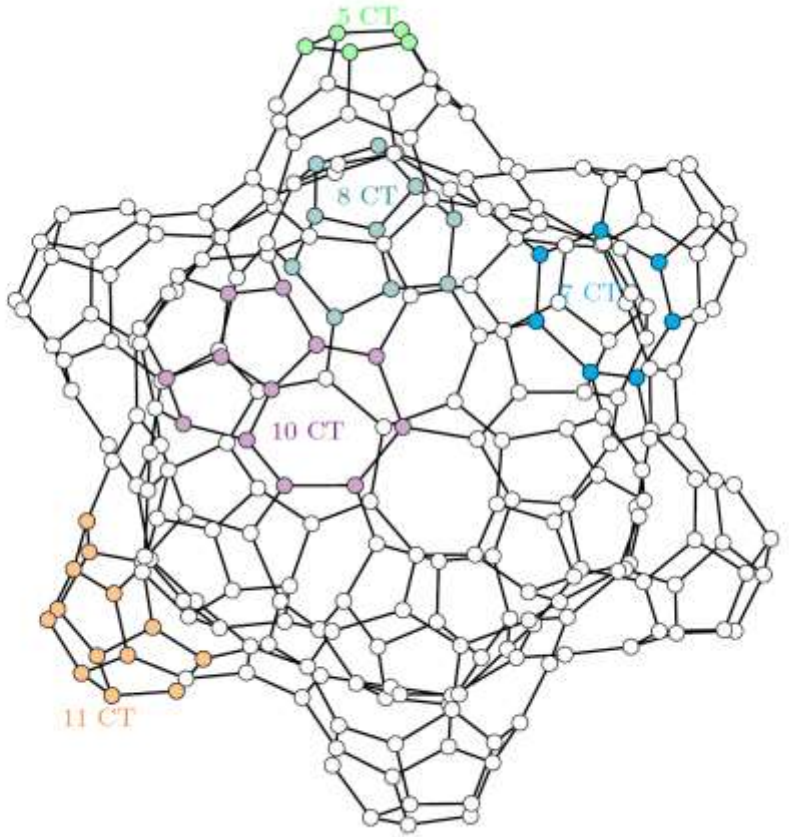


Fig 10. Graph of a fulleroid with a highlighted sample of the k-CT components.

Conclusion