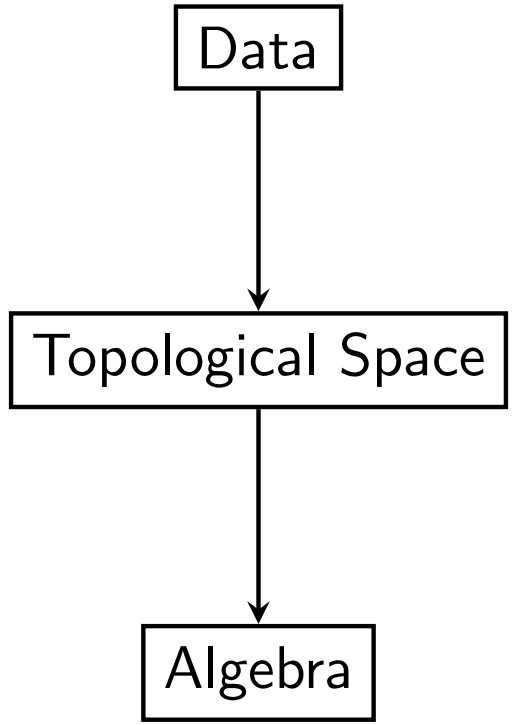
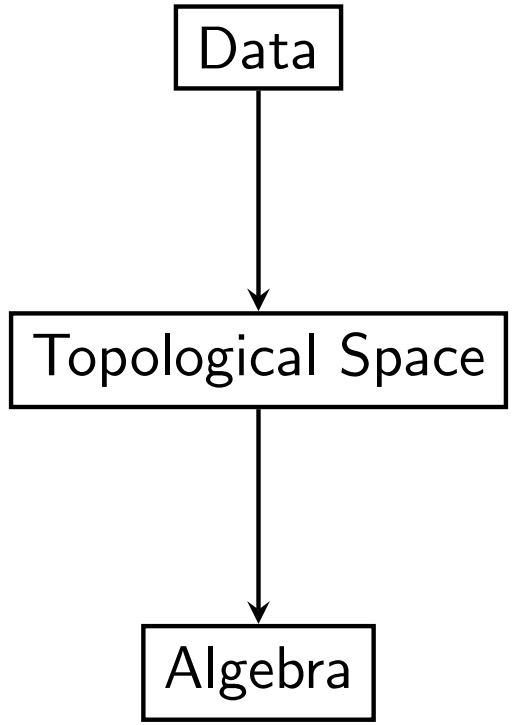
Topological Data Analysis

Graham Ellis
NUI Galway


Data
Topological Space
Algebra



General Aim: Given a finite sample S from an unknown population X we'd like to use algebra to describe/construct a topological space BS that models X.

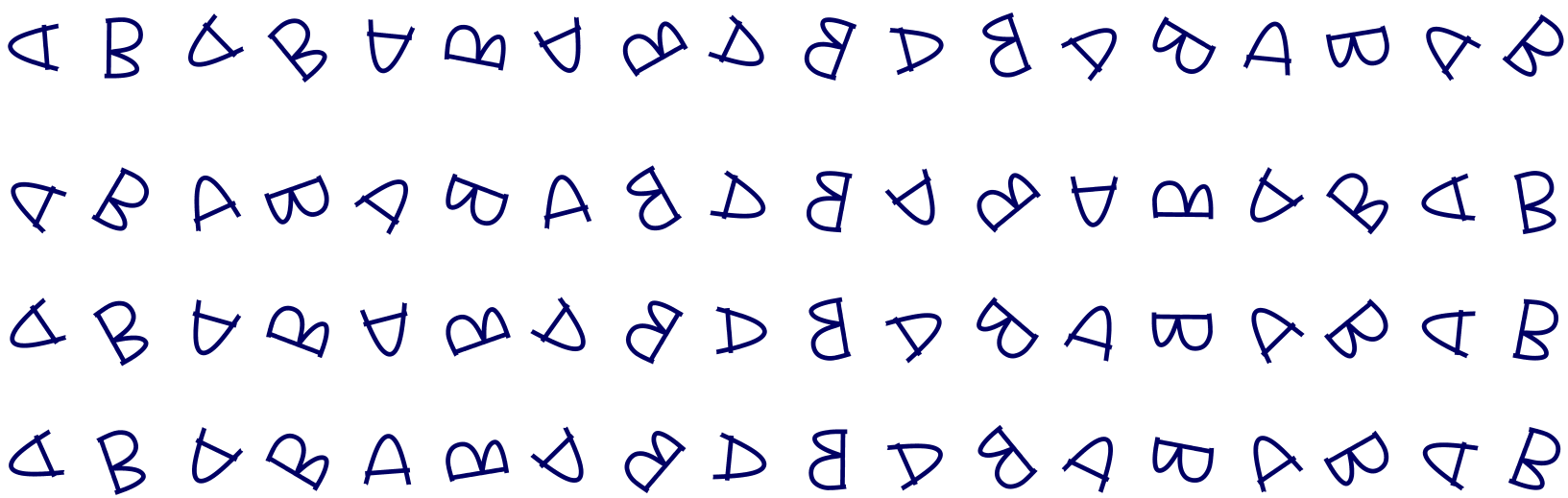
Cluster Analysis

Consider toy data points $S = \{v_1, v_2, \ldots, v_{72}\} \subset \mathbb{R}^{262144}$

Cluster Analysis

Consider toy data points $S = \{v_1, v_2, \ldots, v_{72}\} \subset \mathbb{R}^{262144}$

generated from



Choose real numbers $\epsilon_1 < \epsilon_2 < \cdots < \epsilon_T$ and Euclidean metric.

Choose real numbers $\epsilon_1 < \epsilon_2 < \cdots < \epsilon_T$ and Euclidean metric.

The clique simplicial complex $Y_t = Y(S, \epsilon_t)$ has

- vertex set $S = \{v_1, \ldots, v_{72}\}$.
- n-simplices the subsets $\sigma \subseteq S$ with $n+1$ vertices and $d(v, v') \leq \epsilon_t$ for all $v, v' \in \sigma$.

Choose real numbers $\epsilon_1 < \epsilon_2 < \cdots < \epsilon_T$ and Euclidean metric.

The clique simplicial complex $Y_t = Y(S, \epsilon_t)$ has

- vertex set $S = \{v_1, \ldots, v_{72}\}$.
- n-simplices the subsets $\sigma \subseteq S$ with $n+1$ vertices and $d(v, v') \leq \epsilon_t$ for all $v, v' \in \sigma$.

$\beta_0(Y_t) = \dim(H_0(Y_t, \mathbb{Q})) = \#$ connected components of Y_t.

Choose real numbers $\epsilon_1 < \epsilon_2 < \cdots < \epsilon_T$ and Euclidean metric.

The clique simplicial complex $Y_t = Y(S, \epsilon_t)$ has

- vertex set $S = \{v_1, \ldots, v_{72}\}$.
- n-simplices the subsets $\sigma \subseteq S$ with $n+1$ vertices and $d(v, v') \leq \epsilon_t$ for all $v, v' \in \sigma$.

$\beta_0(Y_t) = \dim(H_0(Y_t, \mathbb{Q})) = \#$ connected components of Y_t.

$\beta_0^{s,t} = rank(H_0(Y_s, \mathbb{Q}) \to H_0(Y_t, \mathbb{Q}))$, $s \leq t$.

Choose real numbers $\epsilon_1 < \epsilon_2 < \cdots < \epsilon_T$ and Euclidean metric.

The clique simplicial complex $Y_t = Y(S, \epsilon_t)$ has

- vertex set $S = \{v_1, \ldots, v_{72}\}$.
- n-simplices the subsets $\sigma \subseteq S$ with $n+1$ vertices and $d(v, v') \leq \epsilon_t$ for all $v, v' \in \sigma$.

$\beta_0(Y_t) = \dim(H_0(Y_t, \mathbb{Q})) = \#$ connected components of Y_t.

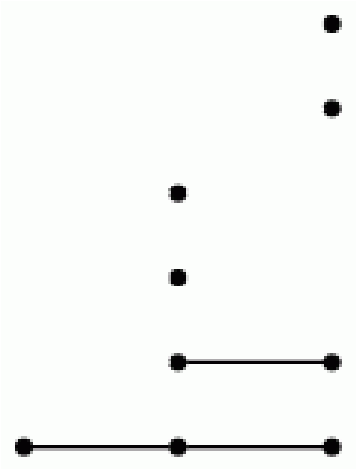
$$\beta_0^{s,t} = rank(H_0(Y_s, \mathbb{Q}) \to H_0(Y_t, \mathbb{Q})), \qquad s \leq t.$$

$$\beta_0^{s,t} = 0, \qquad s > t.$$

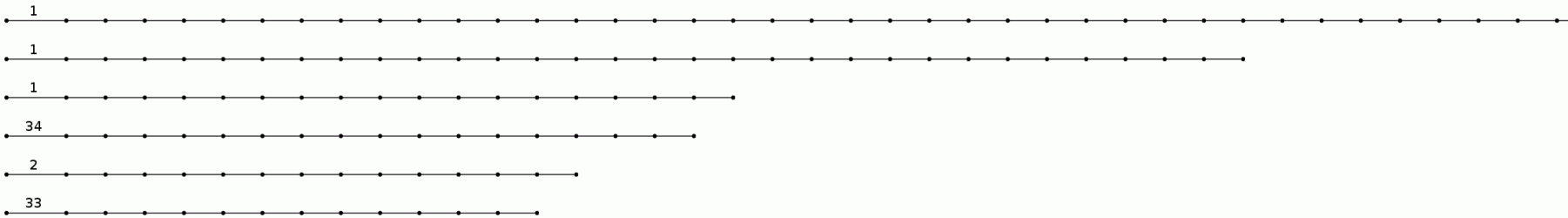
A β_n bar code has

$\beta_n^{s,t}$ *horizontal lines from column s to column t*

$$(\beta_2^{s\,t}) = \begin{pmatrix} 1 & 1 & 1 \\ 0 & 4 & 2 \\ 0 & 0 & 4 \end{pmatrix}$$



β_0 barcode for the toy data S



$\beta_n(Y_t) = \dim(H_n(Y_t, \mathbb{Q}))$ measures n-dimensional 'holes' in Y_t.

$\beta_n(Y_t) = \dim(H_n(Y_t, \mathbb{Q}))$ measures n-dimensional 'holes' in Y_t.

$$\beta_n^{s,t} = rank(H_n(Y_s, \mathbb{Q}) \to H_n(Y_t, \mathbb{Q})), \qquad s \leq t.$$

$$\beta_n^{s,t} = 0, \qquad s > t.$$

$\beta_n(Y_t) = \dim(H_n(Y_t, \mathbb{Q}))$ measures n-dimensional 'holes' in Y_t.

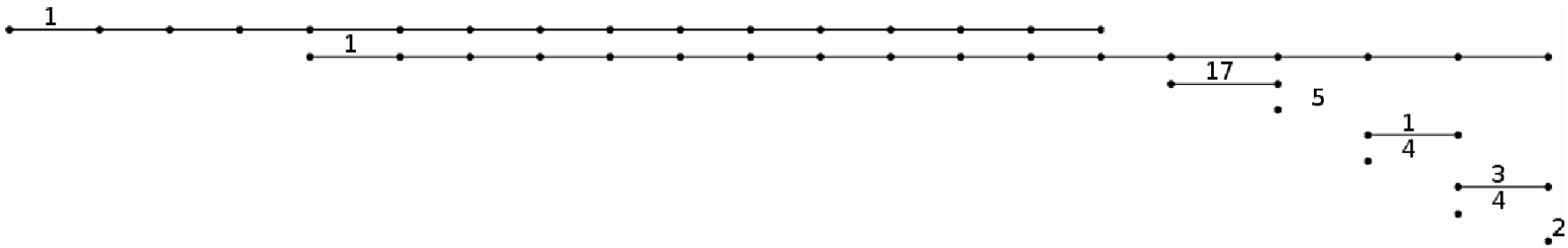
$$\beta_n^{s,t} = rank(H_n(Y_s, \mathbb{Q}) \to H_n(Y_t, \mathbb{Q})), \qquad s \leq t.$$

$$\beta_n^{s,t} = 0, \qquad s > t.$$

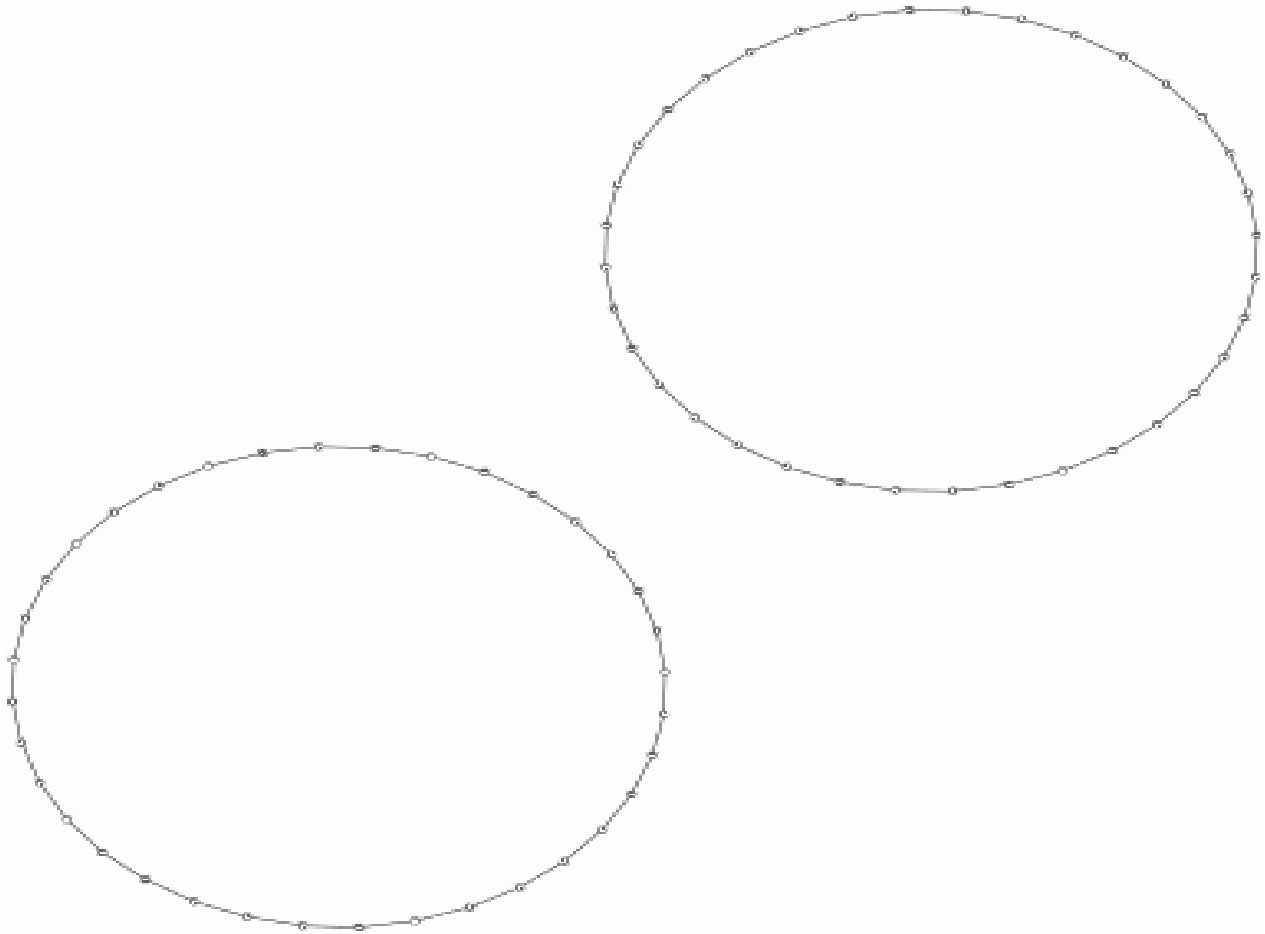
$$H_n(Y, \mathbb{F}) = \ker(\mathbb{F}^{s_n} \xrightarrow{\partial_n} \mathbb{F}^{s_{n-1}}) / \mathrm{im}(\mathbb{F}^{s_{n+1}} \xrightarrow{\partial_{n+1}} \mathbb{F}^{s_n})$$

$s_n =$ number of n-simplices in Y

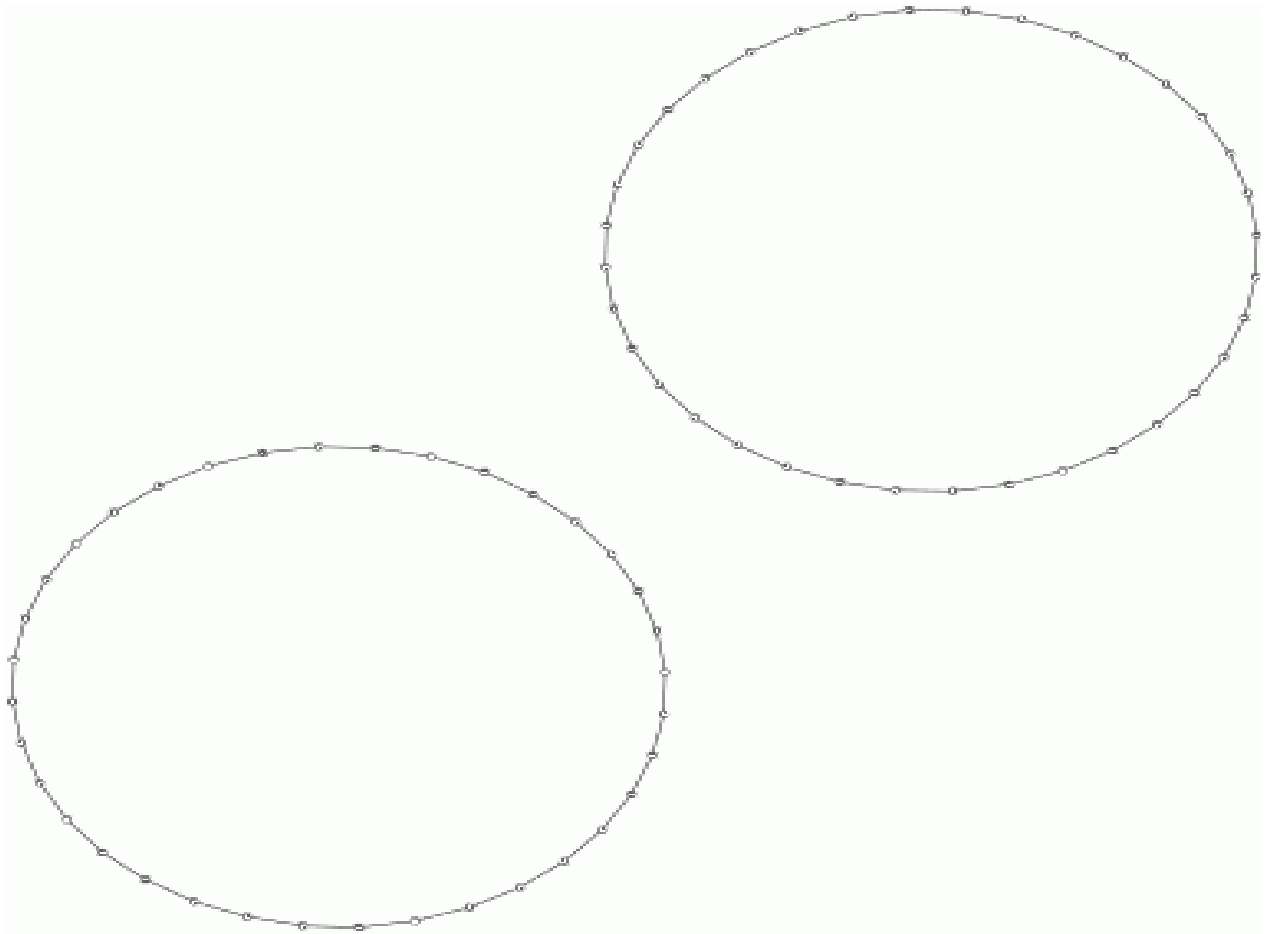
β_1 barcode for the toy data S



Data Model: A homotopy retract $Y \subset Y_{20}$



Data Model: A homotopy retract $Y \subset Y_{20}$



$$Y \simeq S^1 \sqcup S^1$$

Caveat

Kan-Thurston: For any space X there is a map

$$K(G,1) \to X$$

inducing

$$H_*(K(G,1),\mathbb{Z}) \cong H_*(X,\mathbb{Z}).$$

Basic Mapper Cluster Analysis

G. Singh, F. Mémoli & G. Carlsson (2007)

Input:

Distances $d_X(v, v')$ for $v, v' \in S \subset X$, X an unknown metric space.

Basic Mapper Cluster Analysis

G. Singh, F. Mémoli & G. Carlsson (2007)

Input:

Distances $d_X(v, v')$ for $v, v' \in S \subset X$, X an unknown metric space.

Output:

Simplicial complex K which is intended to model X.

Basic Mapper Cluster Analysis

G. Singh, F. Mémoli & G. Carlsson (2007)

Input:

Distances $d_X(v, v')$ for $v, v' \in S \subset X$, X an unknown metric space.

Output:

Simplicial complex K which is intended to model X.

User chooses:

- continuous map $f : X \to Z$.
- open cover $\mathcal{U} = \{U_\alpha\}$ of Z.

Basic Mapper Cluster Analysis

G. Singh, F. Mémoli & G. Carlsson (2007)

Input:

Distances $d_X(v, v')$ for $v, v' \in S \subset X$, X an unknown metric space.

Output:

Simplicial complex K which is intended to model X.

User chooses:

- continuous map $f : X \to Z$.
- open cover $\mathcal{U} = \{U_\alpha\}$ of Z.

Method

$\mathcal{W} = \{W_\alpha = S \cap f^{-1} U_\alpha\}$ is a cover of S.

$\mathcal{V} =$ set of clusters formed by clustering each W_α

Basic Mapper Cluster Analysis

G. Singh, F. Mémoli & G. Carlsson (2007)

Input:

Distances $d_X(v, v')$ for $v, v' \in S \subset X$, X an unknown metric space.

Output:

Simplicial complex K which is intended to model X.

User chooses:

- continuous map $f : X \to Z$.
- open cover $\mathcal{U} = \{U_\alpha\}$ of Z.

Method

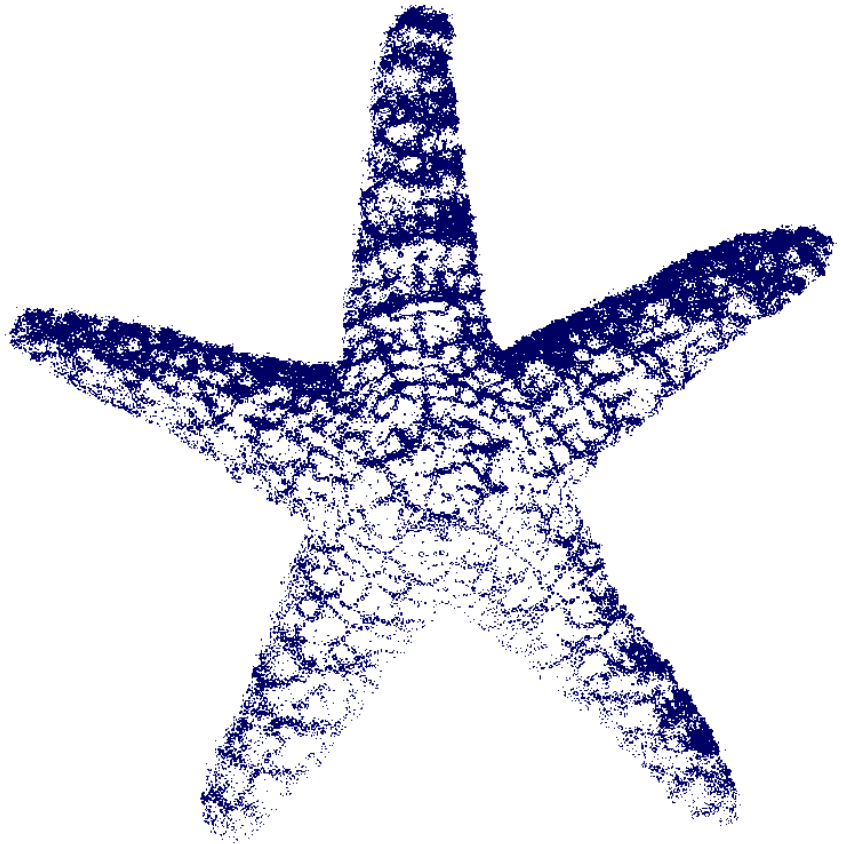
$\mathcal{W} = \{W_\alpha = S \cap f^{-1} U_\alpha\}$ is a cover of S.
$\mathcal{V} =$ set of clusters formed by clustering each W_α

Output

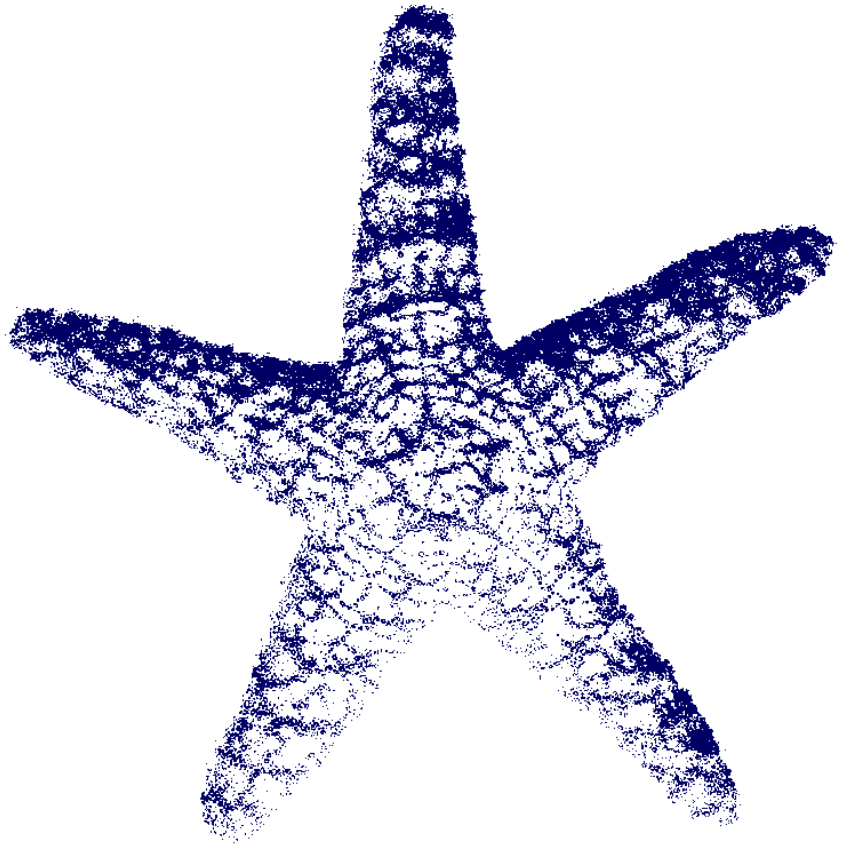
Simplicial nerve of $\mathcal{V}$.

Consider $S = \{v_1, \ldots, v_{200}\} \subset X \subset \mathbb{R}^2$

Consider $S = \{v_1, \ldots, v_{200}\} \subset X \subset \mathbb{R}^2$ where X is



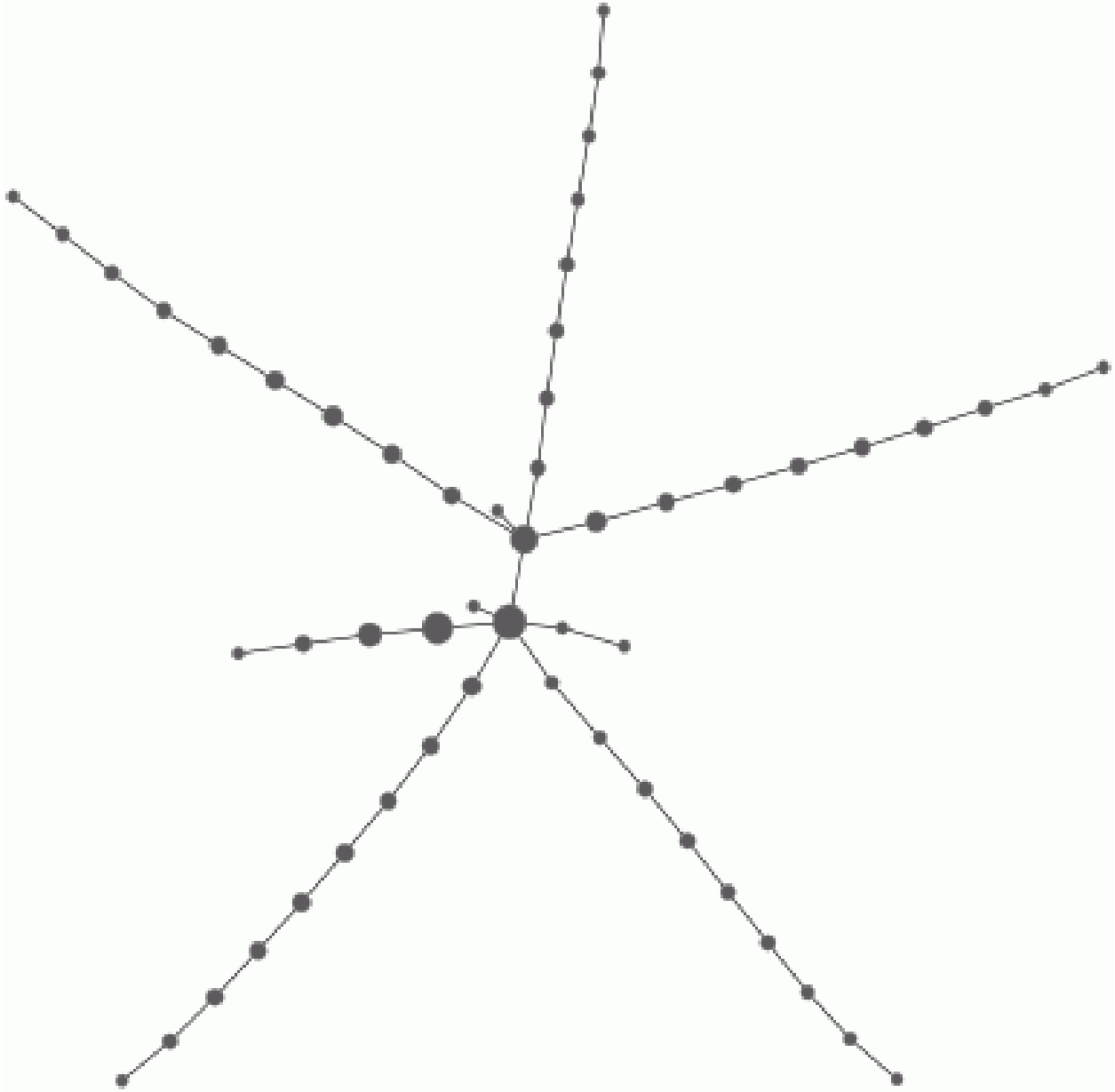
Consider $S = \{v_1, \ldots, v_{200}\} \subset X \subset \mathbb{R}^2$ where X is



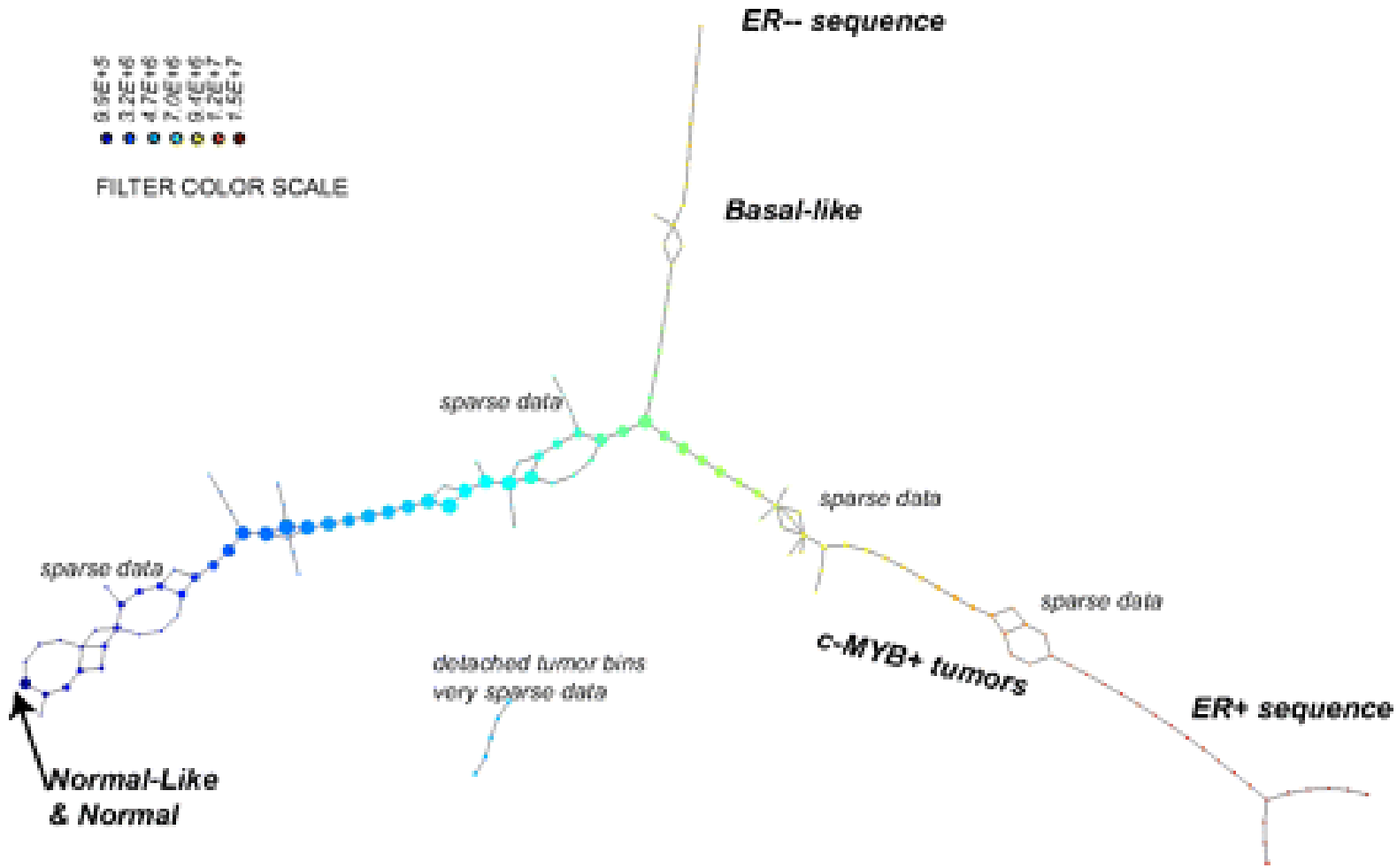
Choose $f : X \to [0, \infty), x \mapsto d(v_1, x)$

and $\mathcal{U}$ an open cover of $Z = [0, \infty)$ with no triple overlaps

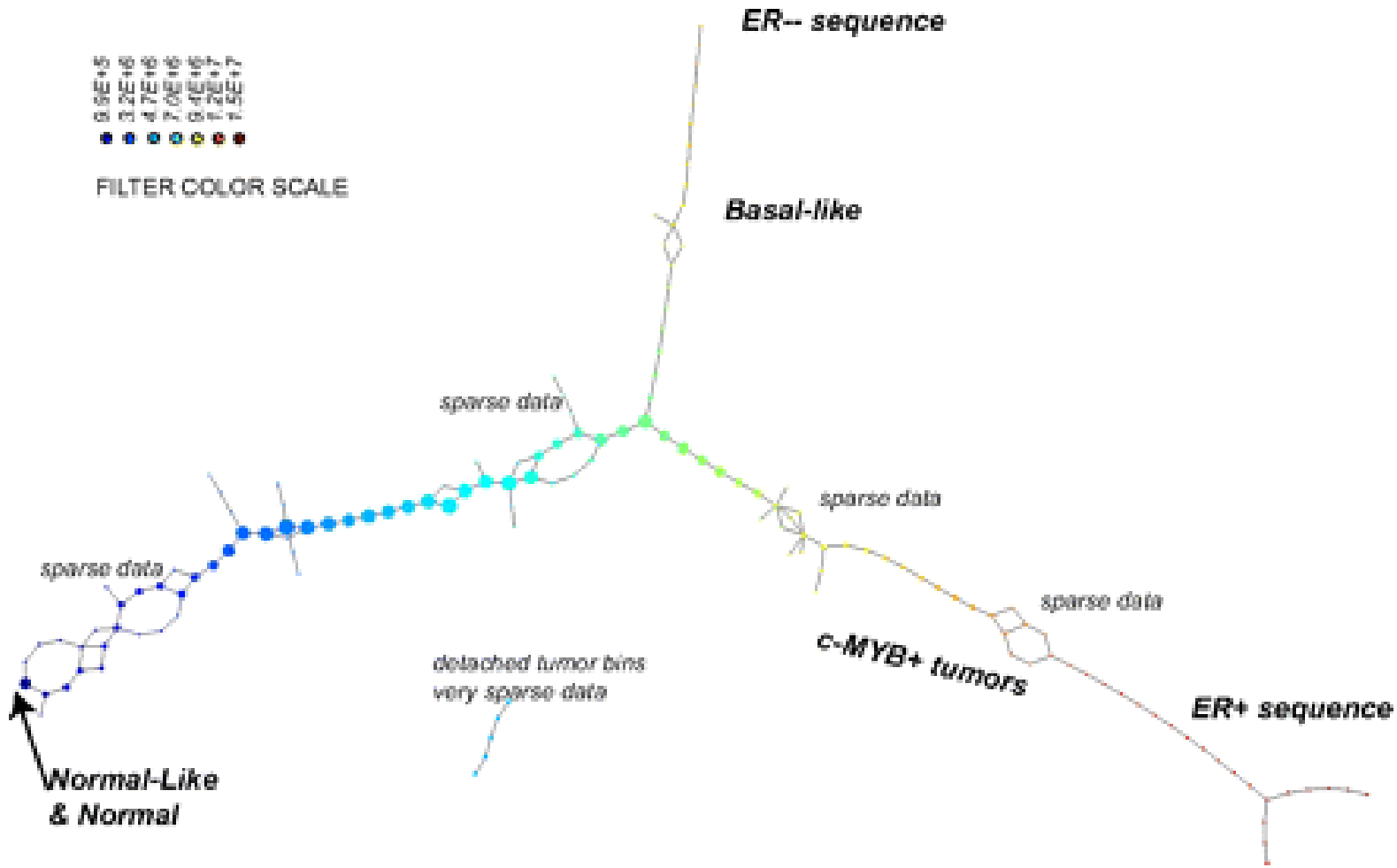
Mapper output for starfish sample



Mapper output for breast cancer microarray gene expression data



Mapper output for breast cancer microarray gene expression data



Nicolau, Levine, Carlsson (PNAS, 2011): identified a subgroup of ER+ breast cancers. These patients exhibit 100% survival.

Enhanced Mapper Cluster Analysis

E + Alokbi

Input:

Distances $d_X(v, v')$ for $v, v' \in S \subset X$, X an unknown metric space.

Enhanced Mapper Cluster Analysis

E + Alokbi

Input:

Distances $d_X(v, v')$ for $v, v' \in S \subset X$, X an unknown metric space.

Output:

Mapper simplicial complex K with

- a simplicial space Y_σ for each $\sigma \in K$,
- an inclusion $Y_\sigma \hookrightarrow Y_\tau$ for each $\tau \subset \sigma$,
- $Y_\emptyset$ a model for X.

Enhanced Mapper Cluster Analysis

E + Alokbi

Input:

Distances $d_X(v, v')$ for $v, v' \in S \subset X$, X an unknown metric space.

Output:

Mapper simplicial complex K with

- a simplicial space Y_σ for each $\sigma \in K$,
- an inclusion $Y_\sigma \hookrightarrow Y_\tau$ for each $\tau \subset \sigma$,
- $Y_\emptyset$ a model for X.

Representation of output: For any functor

$$\Pi\colon \text{topology} \longrightarrow \text{algebra}$$

record $\Pi Y_\sigma \longrightarrow \Pi Y_\tau$.

Consider $S = \{v_1, \ldots, v_{500}\} \subset \mathrm{Im}\phi \subset \mathbb{R}^7$

$$\begin{array}{ccc} \phi\colon \mathbb{R}^3 & \longrightarrow & \mathbb{R}^7 \\ (x, y, z) & \mapsto & (\cos x, 2\sin x, \cos y, \sin y, -2\cos z, \sin z, \cos(x)\sin(y)). \end{array}$$

Consider $S = \{v_1, \ldots, v_{500}\} \subset \mathrm{Im}\phi \subset \mathbb{R}^7$

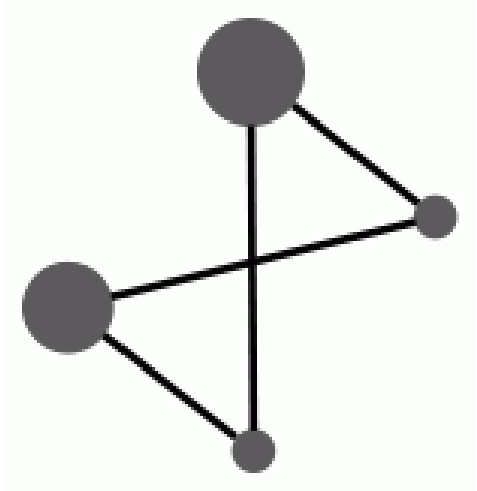
$$\begin{array}{ccc} \phi\colon \mathbb{R}^3 & \longrightarrow & \mathbb{R}^7 \\ (x, y, z) & \mapsto & (\cos x, 2\sin x, \cos y, \sin y, -2\cos z, \sin z, \cos(x)\sin(y)). \end{array}$$

With Euclidean metric on $\mathbb{R}^7$, and filter function

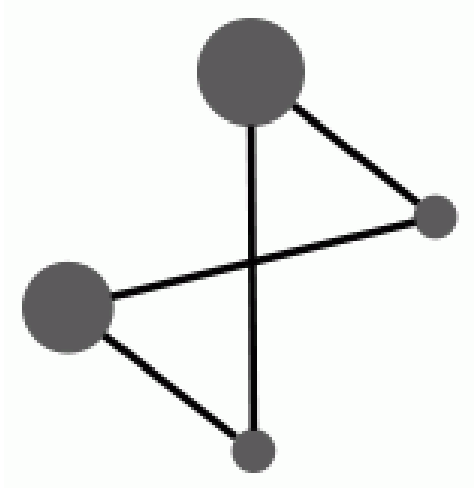
$$\begin{array}{ccc} \mathbb{R}^7 & \xrightarrow{f} & \mathbb{R}^2, \\ (x_1, \ldots, x_7) & \mapsto & (x_1, x_2) \end{array}$$

and open cover $\{U_\alpha\}$ of $Z = \mathbb{R}^2$:

Basic Mapper output K

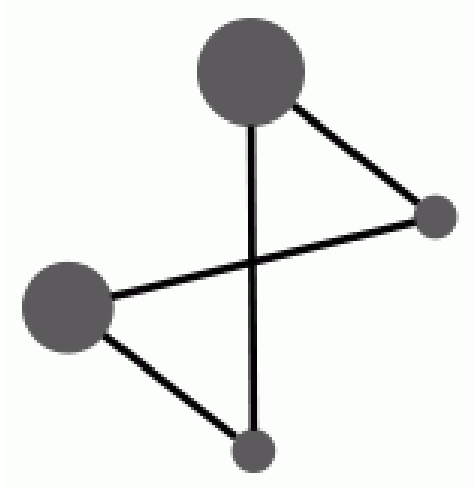


Basic Mapper output K



$\mathcal{W} = \{W_\alpha = S \cap f^{-1} U_\alpha\}$ is a cover of S. Vertices 1,2,3,4 are the clusters formed by clustering each W_α

Basic Mapper output K

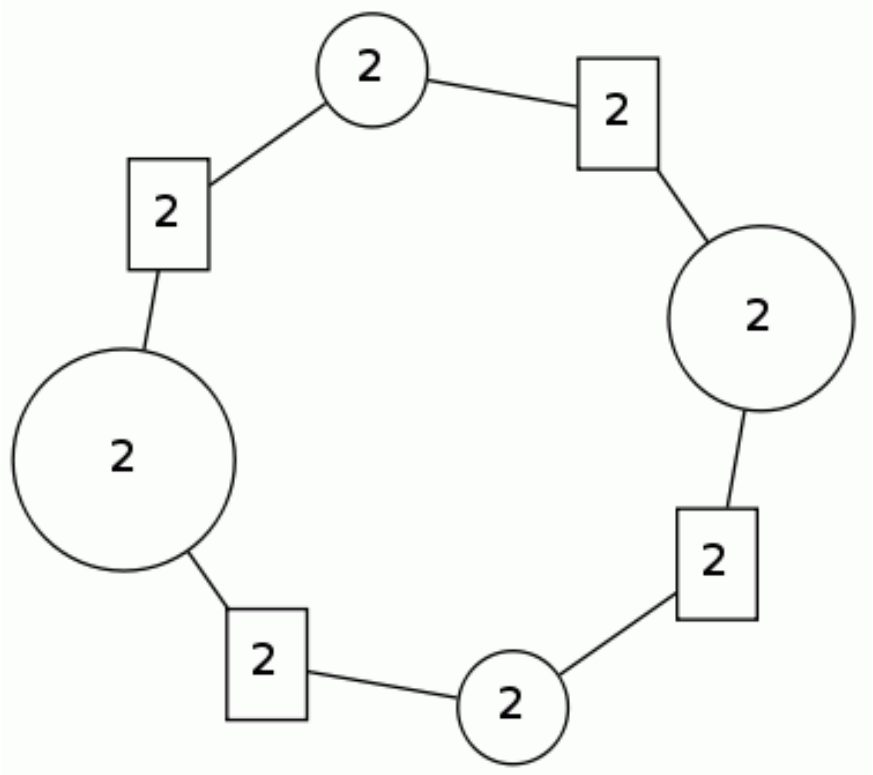


$\mathcal{W} = \{W_\alpha = S \cap f^{-1}U_\alpha\}$ is a cover of S. Vertices 1,2,3,4 are the clusters formed by clustering each W_α

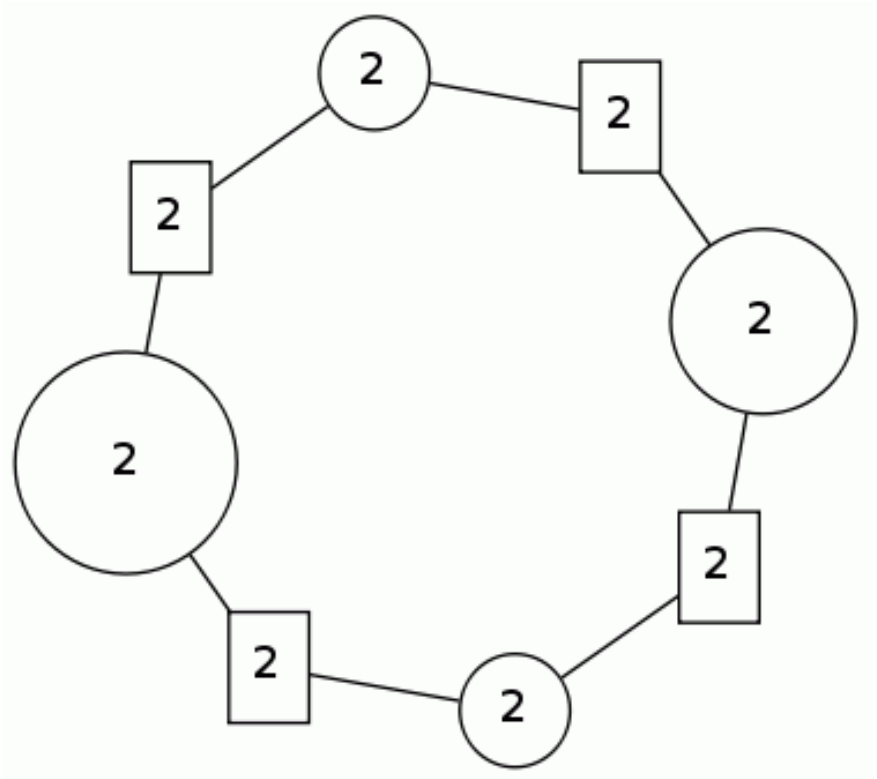
Associate to vertices of K the connected components of clique simplicial complexes $Y(W_\alpha, \epsilon)$, and to simplices $\sigma \in K$

$$Y_\sigma = \bigcap_{v \in \sigma} Y_v, \qquad Y_\emptyset = \bigcup_{v \in \sigma} Y_v.$$

Mapper output enhanced with $\dim H_1(Y_\sigma, \mathbb{Q})$

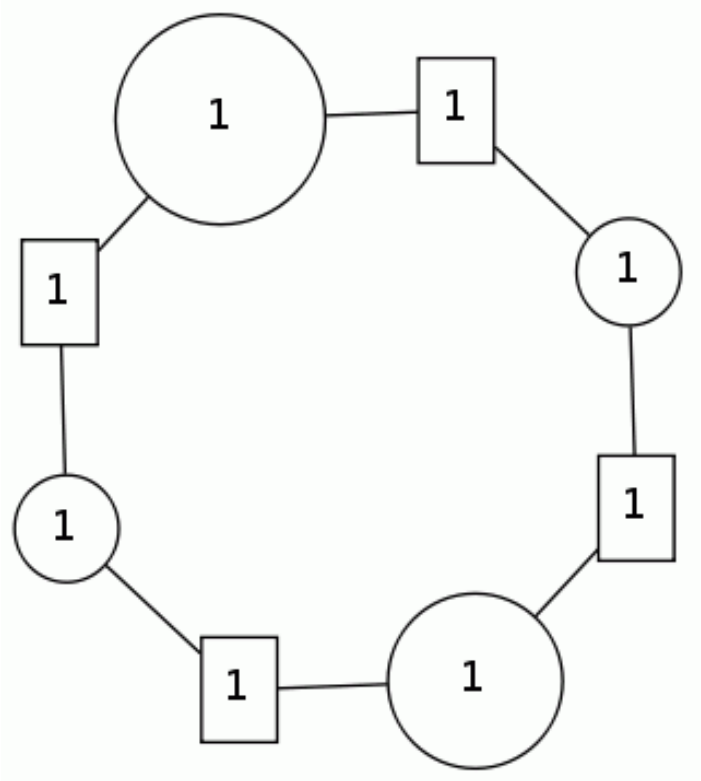


Mapper output enhanced with $\dim H_1(Y_\sigma, \mathbb{Q})$



$$H_1(Y_\emptyset, \mathbb{Q}) = \mathbb{Q} \oplus \mathbb{Q}$$

Mapper output enhanced with $\dim H_2(Y_\sigma, \mathbb{Q})$



$$H_2(Y_\emptyset, \mathbb{Q}) = \mathbb{Q}$$

$Y_\emptyset$ has the same homology as

$$S^1 \times S^1 \times S^1$$

$Y_\emptyset$ has the same homology as

$$S^1 \times S^1 \times S^1 \text{ and } S^1 \vee S^1 \vee S^1 \vee S^2 \vee S^2 \vee S^2 \vee S^3.$$

$Y_\emptyset$ has the same homology as

$$S^1 \times S^1 \times S^1 \text{ and } S^1 \vee S^1 \vee S^1 \vee S^2 \vee S^2 \vee S^2 \vee S^3.$$

Need functors

$$\Pi\colon \text{topology} \longrightarrow \text{algebra}$$

that distinguish between such spaces.

Fundamental group presentations

A finite presentation for

$$\pi_1(Y, y_0) = \{p\colon [0,1] \longrightarrow Y \ : \ p(0) = p(1) = y_0\}/\simeq$$

is implemented in GAP for **small** cellular spaces Y.

Fundamental group presentations

A finite presentation for

$$\pi_1(Y, y_0) = \{p\colon [0,1] \longrightarrow Y \ : \ p(0) = p(1) = y_0\}/\simeq$$

is implemented in GAP for **small** cellular spaces Y. This provides the cup product

$$H^1(Y, \mathbb{Z}) \times H^1(Y, \mathbb{Z}) \longrightarrow H^2(Y, \mathbb{Z}), (\alpha, \beta) \mapsto \alpha \cup \beta$$

induced by

$$Y \longrightarrow Y \times Y, x \mapsto (x, x).$$

Fundamental group presentations

A finite presentation for

$$\pi_1(Y, y_0) = \{p\colon [0,1] \longrightarrow Y \ : \ p(0) = p(1) = y_0\}/\simeq$$

is implemented in GAP for **small** cellular spaces Y. This provides the cup product

$$H^1(Y, \mathbb{Z}) \times H^1(Y, \mathbb{Z}) \longrightarrow H^2(Y, \mathbb{Z}), (\alpha, \beta) \mapsto \alpha \cup \beta$$

induced by

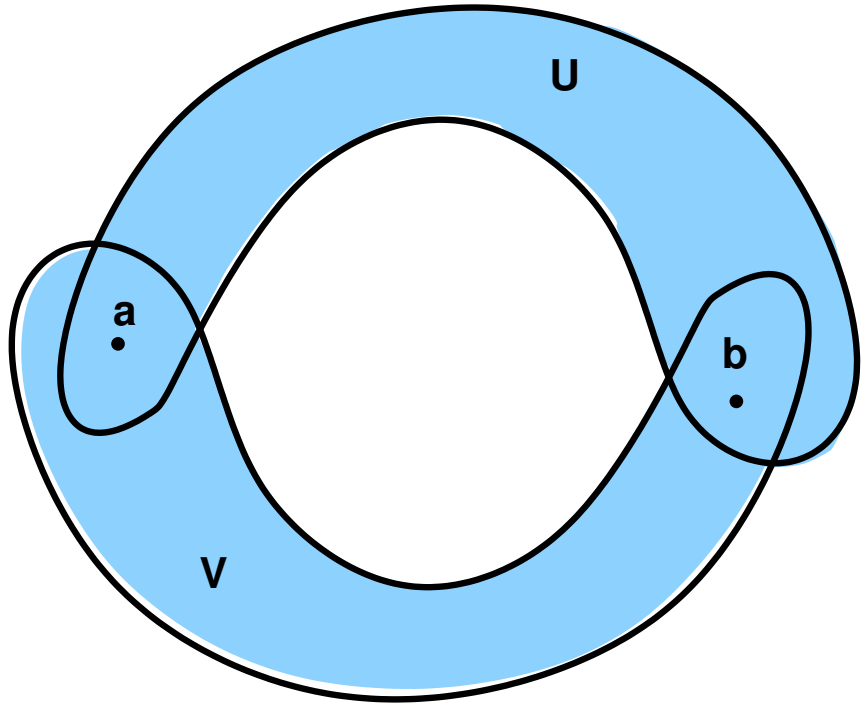
$$Y \longrightarrow Y \times Y, x \mapsto (x, x).$$

$\alpha \cup \beta$ is non-trivial for $Y = S^1 \times S^1 \times S^1$ and trivial for $Y = S^1 \vee S^1 \vee S^1 \vee S^2 \vee S^2 \vee S^2 \vee S^3$.

Parallel computation of $\pi_1 Y$?

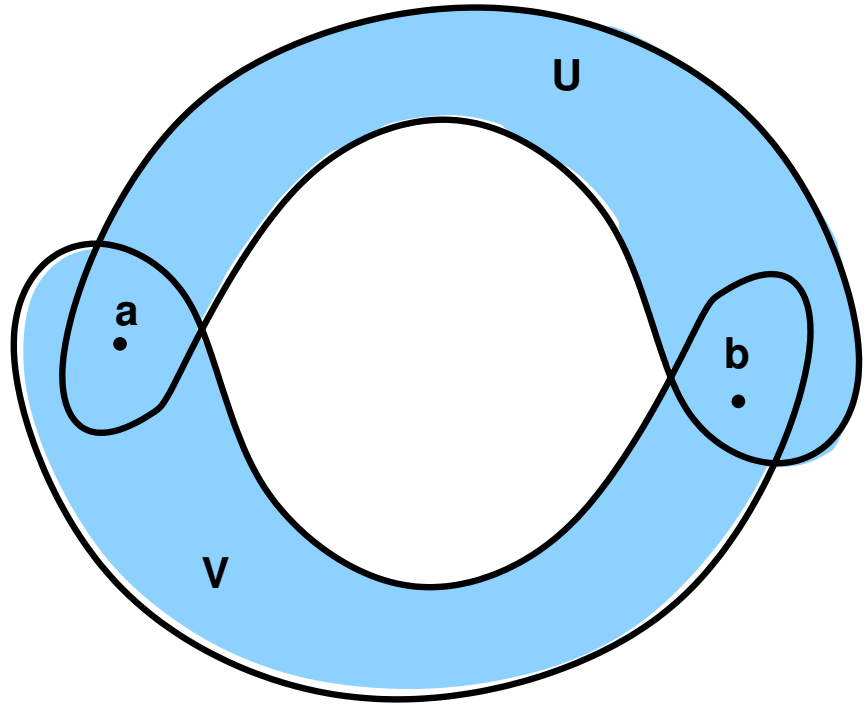
Parallel computation of $\pi_1 Y$?

$Y =$



Parallel computation of $\pi_1 Y$?

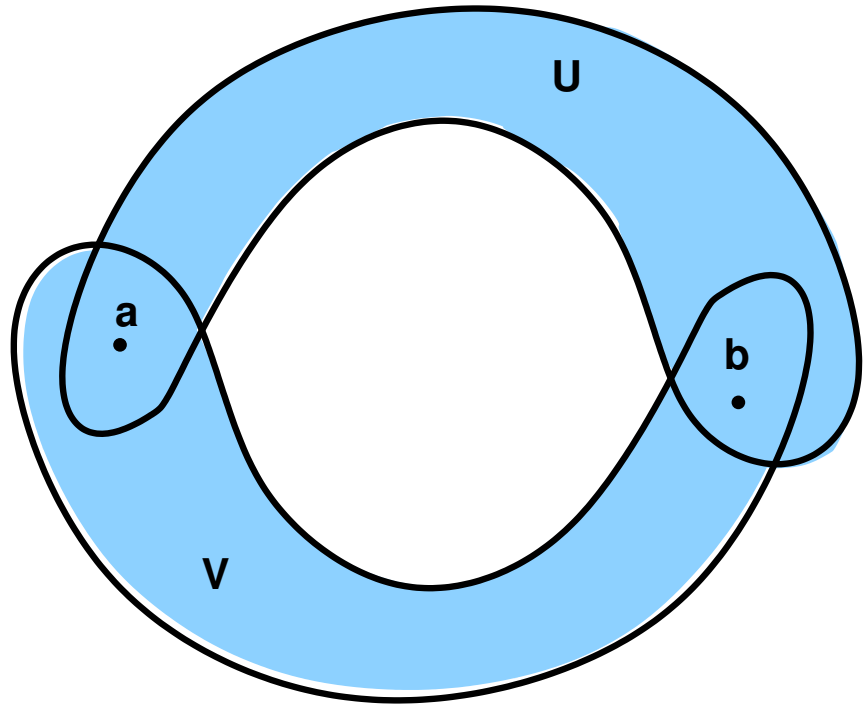
$Y =$



Enter groupoids

$$\pi_1(Y, Y_0) = \{p\colon [0,1] \longrightarrow Y \ : \ p(0), p(1) \in Y_0\}$$

Parallel computation of $\pi_1 Y$?

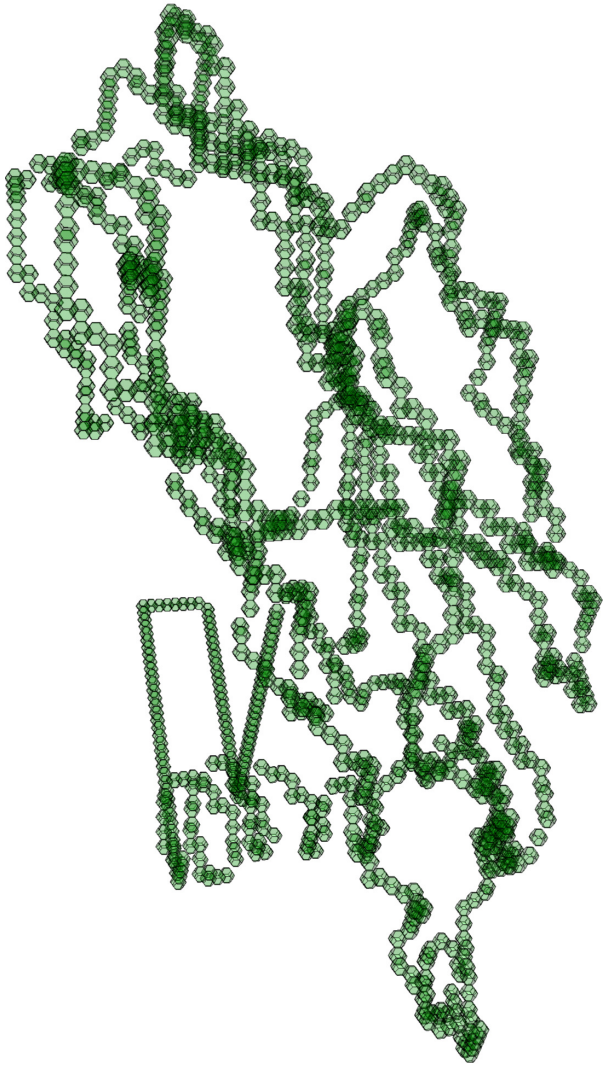
$Y =$ 

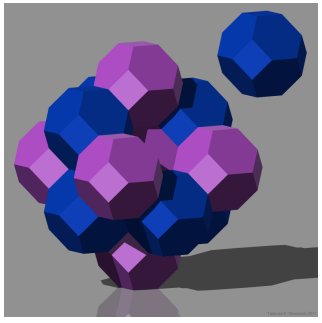
Enter groupoids

$$\pi_1(Y, Y_0) = \{p\colon [0,1] \longrightarrow Y \ : \ p(0), p(1) \in Y_0\}$$

$$\pi_1(U \cup V, \{a,b\}) \cong \pi_1(U, \{a,b\}) *_{\pi_1(U \cap V, \{a,b\})} \pi_1(V, \{a,b\})$$

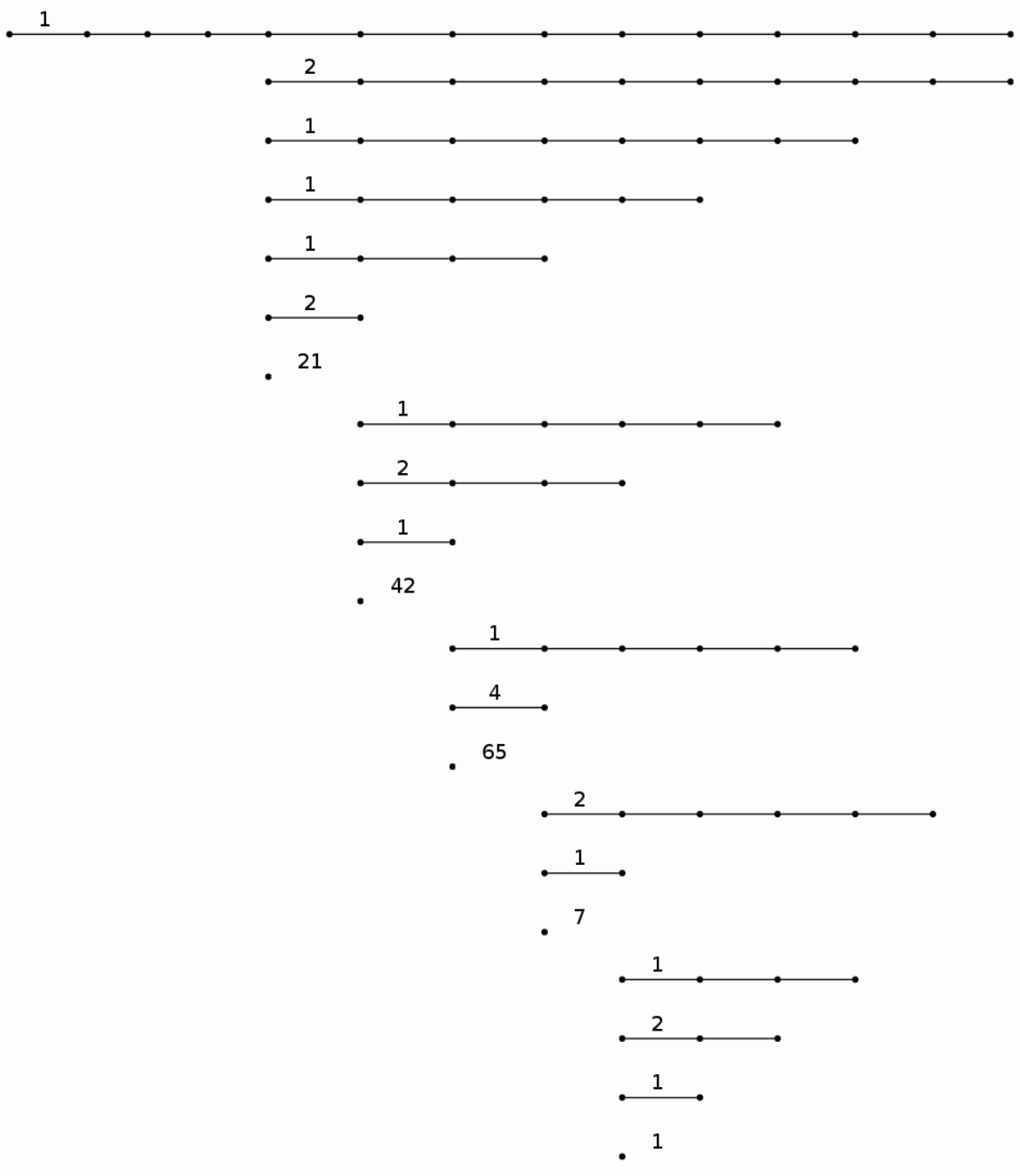
1V2X protein backbone





1V2X protein backbone

Persistent β_1



1V2X protein backbone

Data model: $\mathbb{S}^1 \simeq K \subset \mathbb{R}^3$

1V2X protein backbone

Data model: $\mathbb{S}^1 \simeq K \subset \mathbb{R}^3$

We compute

$$Y = \mathbb{R}^3 \setminus K$$

1V2X protein backbone

Data model: $\mathbb{S}^1 \simeq K \subset \mathbb{R}^3$

We compute

$$Y = \mathbb{R}^3 \setminus K$$

and

$$\pi_1 Y \cong \langle\, x, y \mid yx^{-1}yxy^{-1}x \,\rangle$$

1V2X protein backbone

Data model: $\mathbb{S}^1 \simeq K \subset \mathbb{R}^3$

We compute

$$Y = \mathbb{R}^3 \setminus K$$

and

$$\pi_1 Y \cong \langle\, x, y \mid yx^{-1}yxy^{-1}x \,\rangle$$

But what good is this presentation of the fundamental group?

An isomorphism invariant of finitely presented groups

$$I_n(G) = \{H_{ab} \ : \ H < G \text{ of index } \leq n\}$$

An isomorphism invariant of finitely presented groups

$$I_n(G) = \{H_{ab} \ : \ H < G \text{ of index } \leq n\}$$

$$I_3(\langle x, y | yx^{-1}yxy^{-1}x \rangle) = \{\mathbb{Z},\ \mathbb{Z} \oplus \mathbb{Z}_3,\ \mathbb{Z} \oplus \mathbb{Z},\ \mathbb{Z} \oplus \mathbb{Z}_2 \oplus \mathbb{Z}_2\}$$

An isomorphism invariant of finitely presented groups

$$I_n(G) = \{H_{ab} \ : \ H < G \text{ of index } \leq n\}$$

$$I_3(\langle x, y | yx^{-1}yxy^{-1}x \rangle) = \{\mathbb{Z},\ \mathbb{Z} \oplus \mathbb{Z}_3,\ \mathbb{Z} \oplus \mathbb{Z},\ \mathbb{Z} \oplus \mathbb{Z}_2 \oplus \mathbb{Z}_2\}$$

$I_n(\pi_1(\mathbb{R}^3 \setminus K))$ tested on 1701935 **prime knots** ≤ 14 **crossings**

An isomorphism invariant of finitely presented groups

$$I_n(G) = \{H_{ab} \ : \ H < G \text{ of index } \leq n\}$$

$$I_3(\langle x, y | yx^{-1}yxy^{-1}x \rangle) = \{\mathbb{Z},\ \mathbb{Z} \oplus \mathbb{Z}_3,\ \mathbb{Z} \oplus \mathbb{Z},\ \mathbb{Z} \oplus \mathbb{Z}_2 \oplus \mathbb{Z}_2\}$$

$I_n(\pi_1(\mathbb{R}^3 \setminus K))$ tested on 1701935 **prime knots** ≤ 14 **crossings**

min value of n to distinguish between knots on c crossings

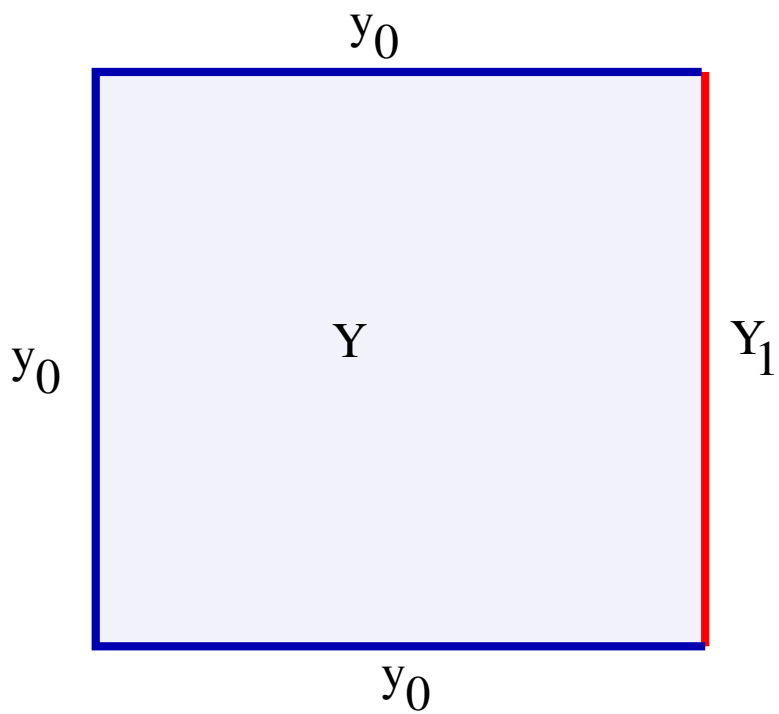
c	3	4	5	6	7	8	9	10	11	12	13	14
n	2	2	3	3	3	3	5	5	6	6	7	7

Brendel, E., Juda, Mrozek

Crossed modules

For $y_0 \in Y_1 \subset Y$ we consider

$$\pi_2(Y, Y_1) = \{p\colon [0,1]^2 \to Y \ : \begin{array}{ll} p(x,y) = y_0 & \text{if } x = 0, \\ p(x,y) = y_0 & \text{if } y = 0 \text{ or } 1 \\ p(x,y) \in Y_1 & \text{if } x = 1\} \end{array} \}/\simeq$$



$G = \pi_1(Y_1)$ acts on $M = \pi_2(Y, Y_1)$ and the group homomorphism

$$\partial \colon M \to G$$

satisfies

- $\partial({}^{g}m) = gmg^{-1}$,
- ${}^{\partial m}m' = mm'm^{-1}$.

$G = \pi_1(Y_1)$ acts on $M = \pi_2(Y, Y_1)$ and the group homomorphism

$$\partial \colon M \to G$$

satisfies

- $\partial({}^g m) = gmg^{-1}$,
- ${}^{\partial m} m' = mm'm^{-1}$.

This algebraic structure $\Pi(Y, Y_1)$ is a **crossed module**.

$G = \pi_1(Y_1)$ acts on $M = \pi_2(Y, Y_1)$ and the group homomorphism

$$\partial \colon M \to G$$

satisfies

- $\partial({}^g m) = gmg^{-1}$,
- ${}^{\partial m} m' = mm'm^{-1}$.

This algebraic structure $\Pi(Y, Y_1)$ is a **crossed module**.

The crossed module $\Pi(Y^2, Y^1)$ is **freely presented** and, given any **finite** crossed module C, the set of homotopy classes of morphisms

$$[\Pi(Y, Y^1), C] = \{\Pi(Y, Y^1) \to C\}/\simeq$$

can be computed and is a homotopy invariant of Y.

The **order** of homotopy 2-type X is the least value of $m = |M||G|$ for a representative crossed module $M \xrightarrow{\partial} G$.

Proposition (E, Le)

The homotopy 2-types of order m are classified up to homotopy for $m \leq 127$, $m \neq 32, 64, 81, 96$ and are distributed with GAP.

$\partial\colon Q \to Aut(Q)$

```
gap> G2:=AutoCrossedModule(DihedralGroup(216));;

gap> Size(G2);
839808

gap> IdQuasiCrossedModule(G2);
[ 72, 68 ]
```

$\partial\colon Q \to Aut(Q)$

```
gap> G2:=AutoCrossedModule(DihedralGroup(216));;

gap> Size(G2);
839808

gap> IdQuasiCrossedModule(G2);
[ 72, 68 ]


gap> G:=SmallQuasiCrossedModule(72,68);
Crossed module
```

$\partial\colon Q \to Aut(Q)$

```
gap> G2:=AutoCrossedModule(DihedralGroup(216));;

gap> Size(G2);
839808

gap> IdQuasiCrossedModule(G2);
[ 72, 68 ]


gap> G:=SmallQuasiCrossedModule(72,68);
Crossed module


gap> Homology(G,5);
[ 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 18 ]
```