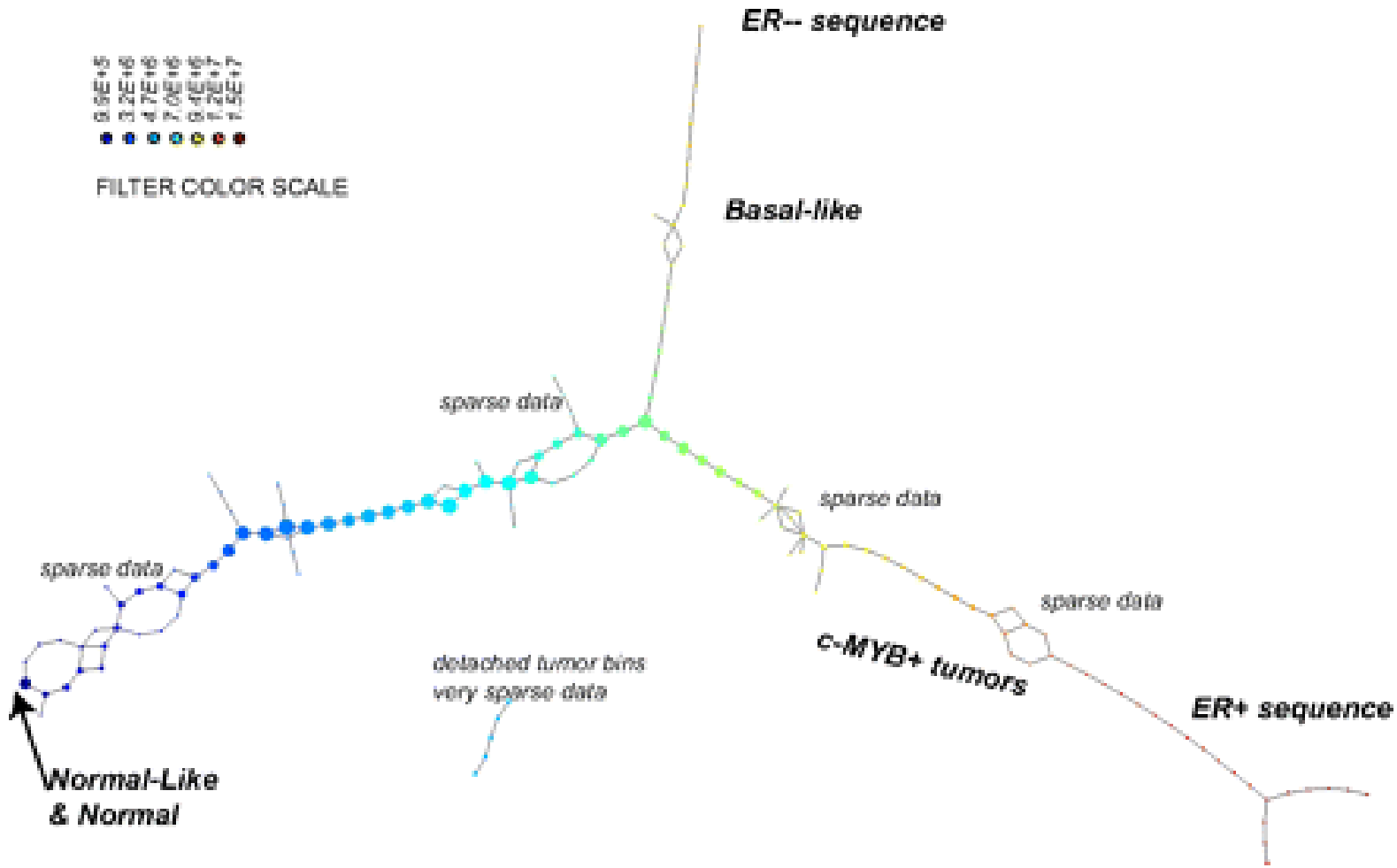
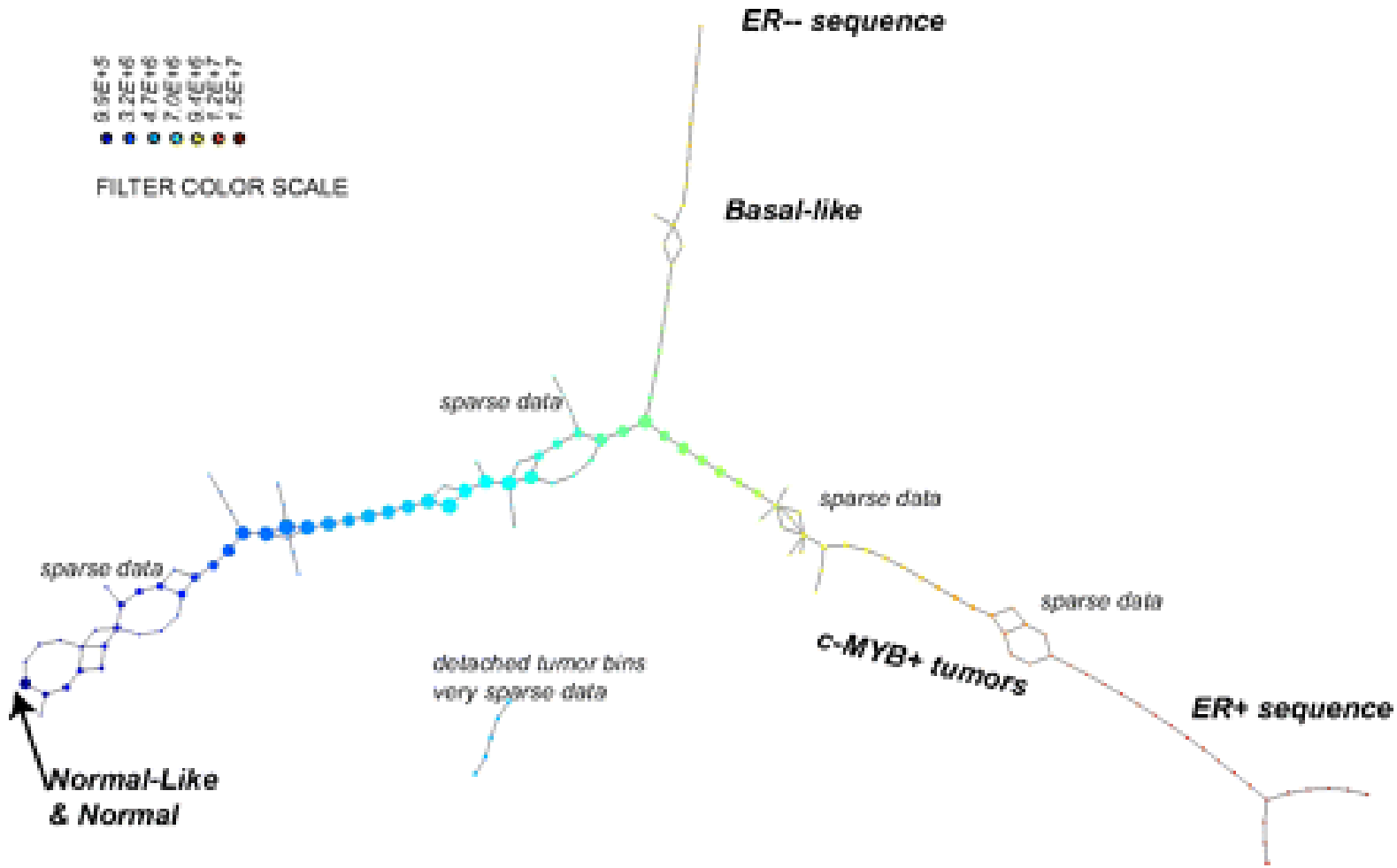
Topological Data Analysis

Graham Ellis
NUI Galway

Breast cancer microarray gene expression data



Breast cancer microarray gene expression data



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

On a population space X choose:

- metric $d_X(x, y)$
- continuous map $f : X \to Z$
- open cover $\{U_\alpha\}_{\alpha \in A}$ of Z
- finite sample $S \subset X$

On a population space X choose:

- metric $d_X(x, y)$
- continuous map $f : X \to Z$
- open cover $\{U_\alpha\}_{\alpha \in A}$ of Z
- finite sample $S \subset X$

Cluster each $f^{-1}(U_\alpha) \cap S \;=\; S_{\alpha,1} \sqcup S_{\alpha,2} \sqcup \cdots \sqcup S_{\alpha,n_\alpha}$

On a population space X choose:

- metric $d_X(x, y)$
- continuous map $f : X \to Z$
- open cover $\{U_\alpha\}_{\alpha \in A}$ of Z
- finite sample $S \subset X$

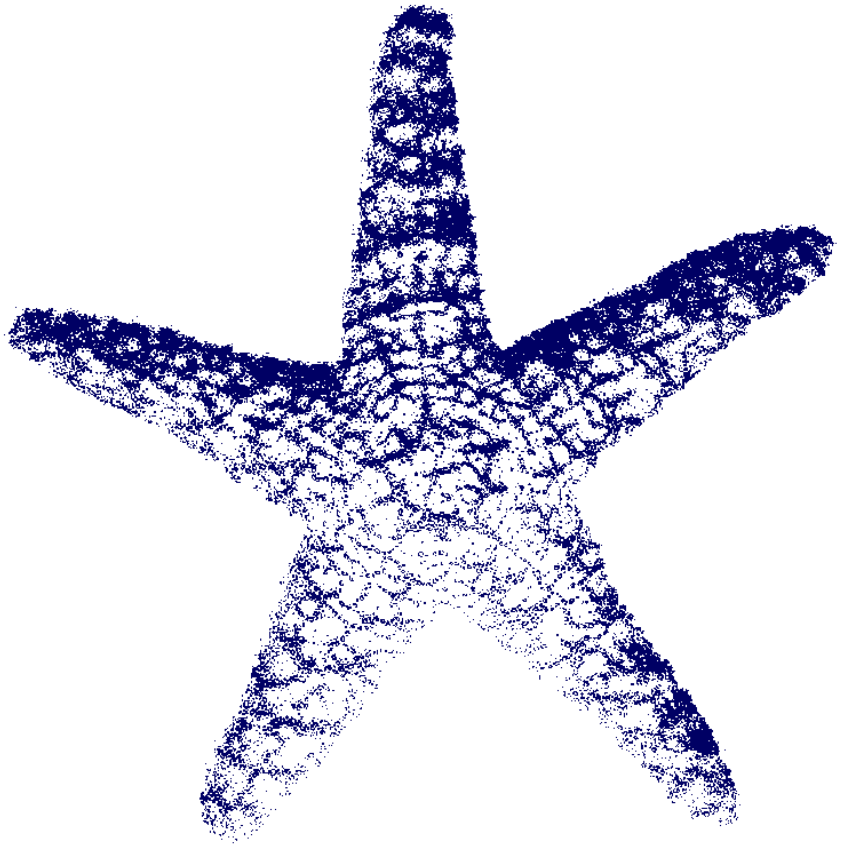
Cluster each $f^{-1}(U_\alpha) \cap S \;=\; S_{\alpha,1} \sqcup S_{\alpha,2} \sqcup \cdots \sqcup S_{\alpha,n_\alpha}$

Output

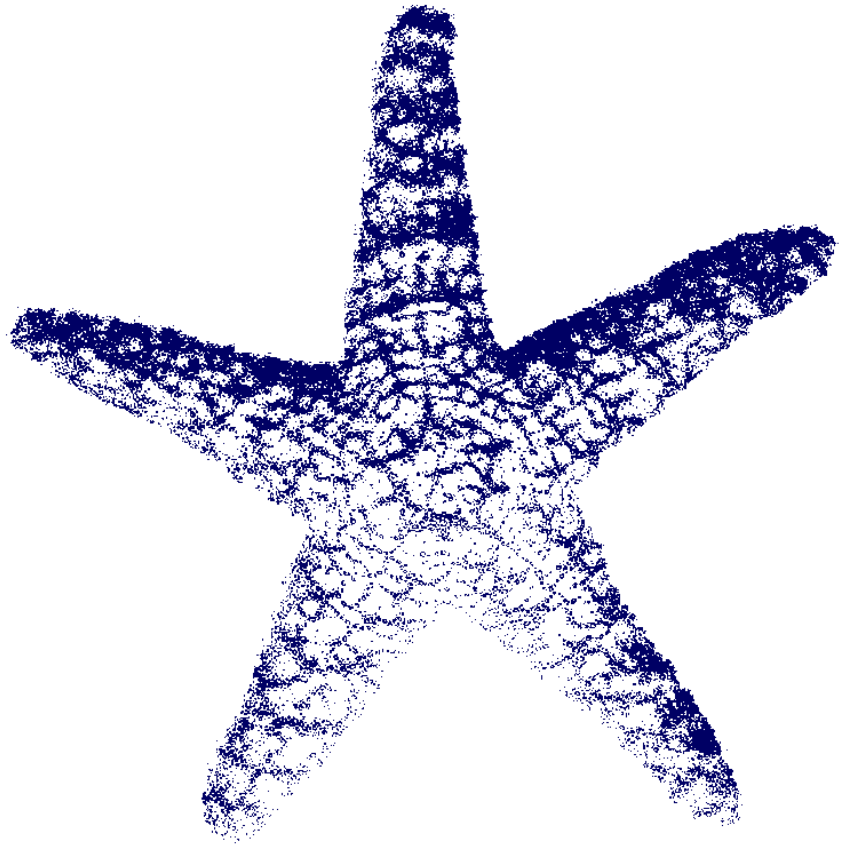
$$K = \mathrm{Nerve}(\ \{S_{\alpha,j}\}_{\alpha \in A, 1 \le j \le n_\alpha}\ )$$

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

$X =$



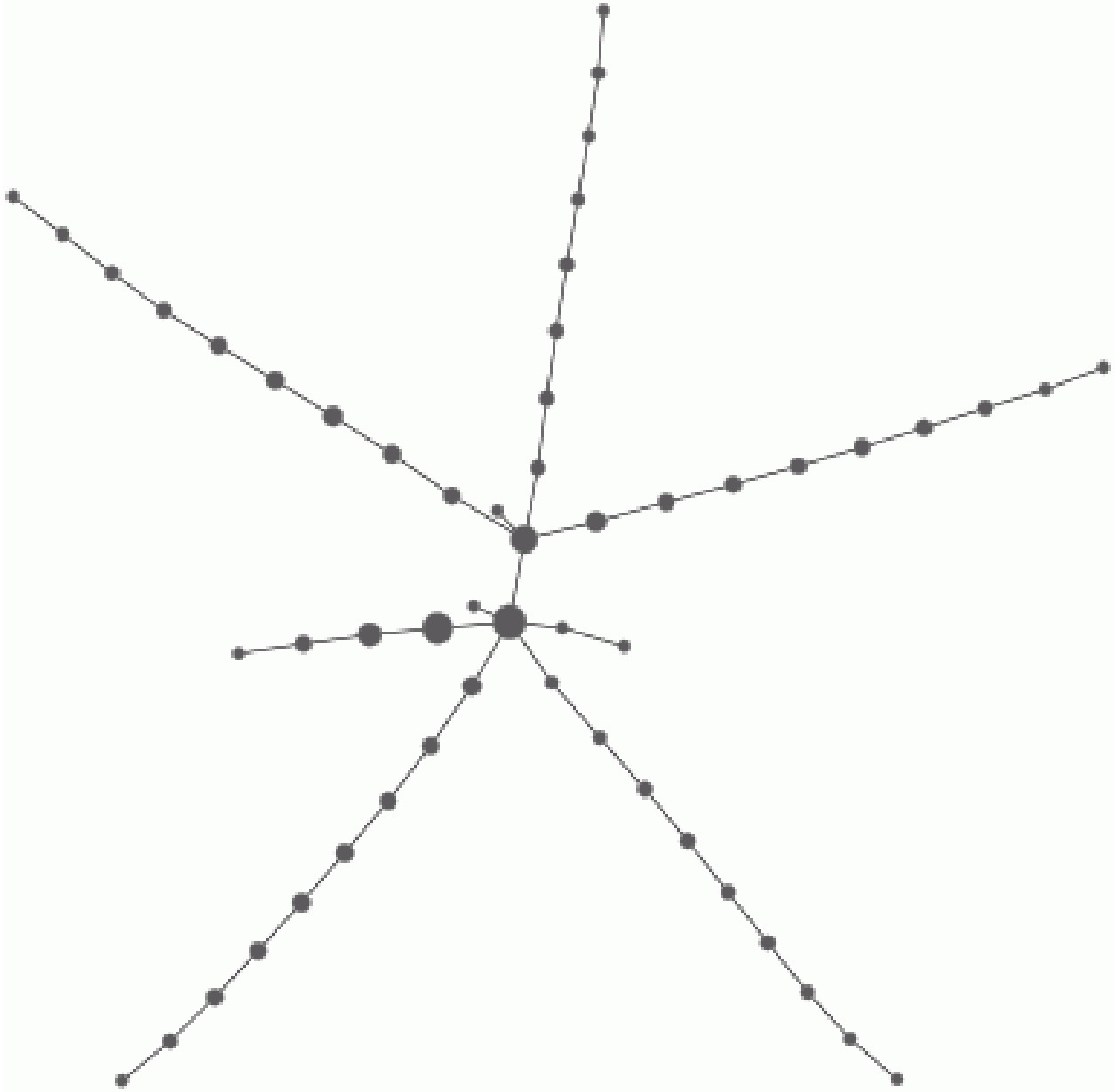
$$\mathbf{S} = \{\mathbf{v_1}, \ldots, \mathbf{v_{200}}\} \subset \mathbf{X} \subset \mathbb{R}^2$$

$X =$ 

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

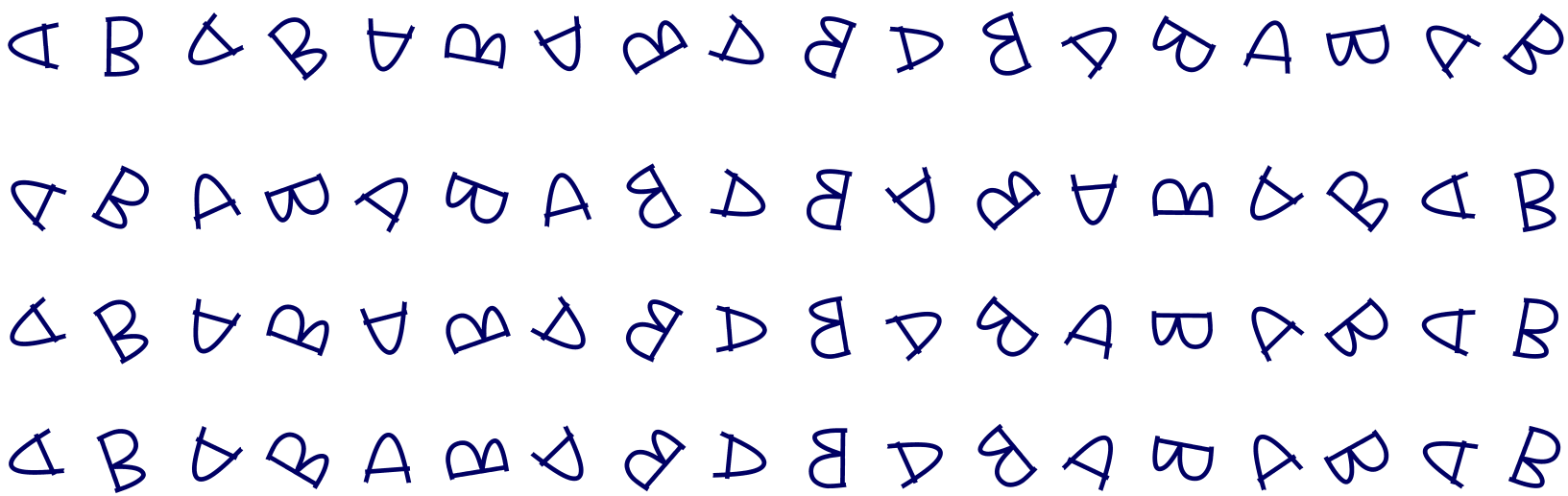
$\{U_\alpha\}_{\alpha \in A}$ an open cover of $Z = [0, \infty)$ with no triple overlaps

Mapper output for starfish sample



$$\mathbf{S} = \{\mathbf{v_1}, \mathbf{v_2}, \ldots, \mathbf{v_{72}}\} \subset \mathbf{X} = \mathbb{R}^{\mathbf{262144}}$$

$$\mathbf{S} = \{\mathbf{v_1}, \mathbf{v_2}, \ldots, \mathbf{v_{72}}\} \subset \mathbf{X} = \mathbb{R}^{\mathbf{262144}}$$



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

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

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 $d_X(v, v') \leq \epsilon_t$ for all $v, v' \in \sigma$.

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

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 $d_X(v, v') \leq \epsilon_t$ 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 d_X.

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 $d_X(v, v') \leq \epsilon_t$ for all $v, v' \in \sigma$.

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

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

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

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 $d_X(v, v') \leq \epsilon_t$ for all $v, v' \in \sigma$.

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

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

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

A β_0 bar code has

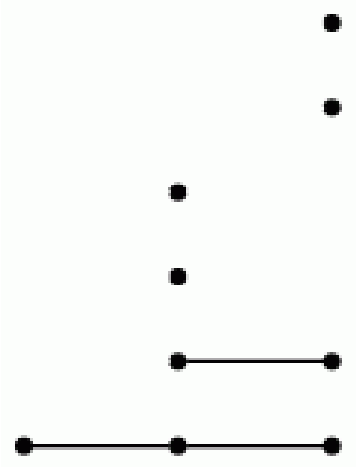
$\beta_0^{s,t}$ *horizontal lines from column s to column t*

$$(\beta_0^{s\,t}) = \begin{pmatrix} 1 & 1 & 1 \\ 0 & 4 & 2 \\ 0 & 0 & 4 \end{pmatrix}$$

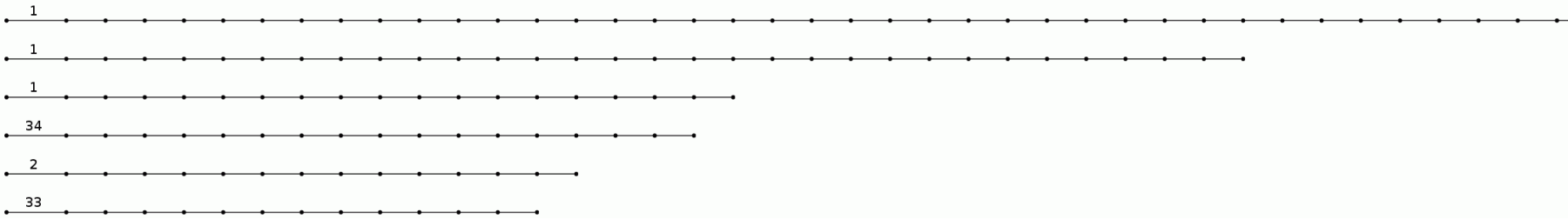
A β_0 bar code has

$\beta_0^{s,t}$ *horizontal lines from column s to column t*

$$(\beta_0^{s\,t}) = \begin{pmatrix} 1 & 1 & 1 \\ 0 & 4 & 2 \\ 0 & 0 & 4 \end{pmatrix}$$



β_0 barcode for 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.

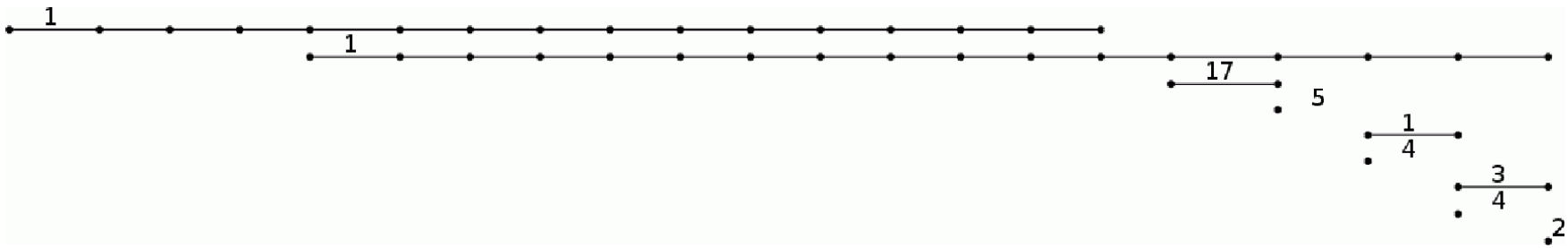
$$\beta_n^{s,t} = rank(\ H_n(Y_s, \mathbb{Q}) \to H_n(Y_t, \mathbb{Q})\ ), \qquad s \leq t.$$

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

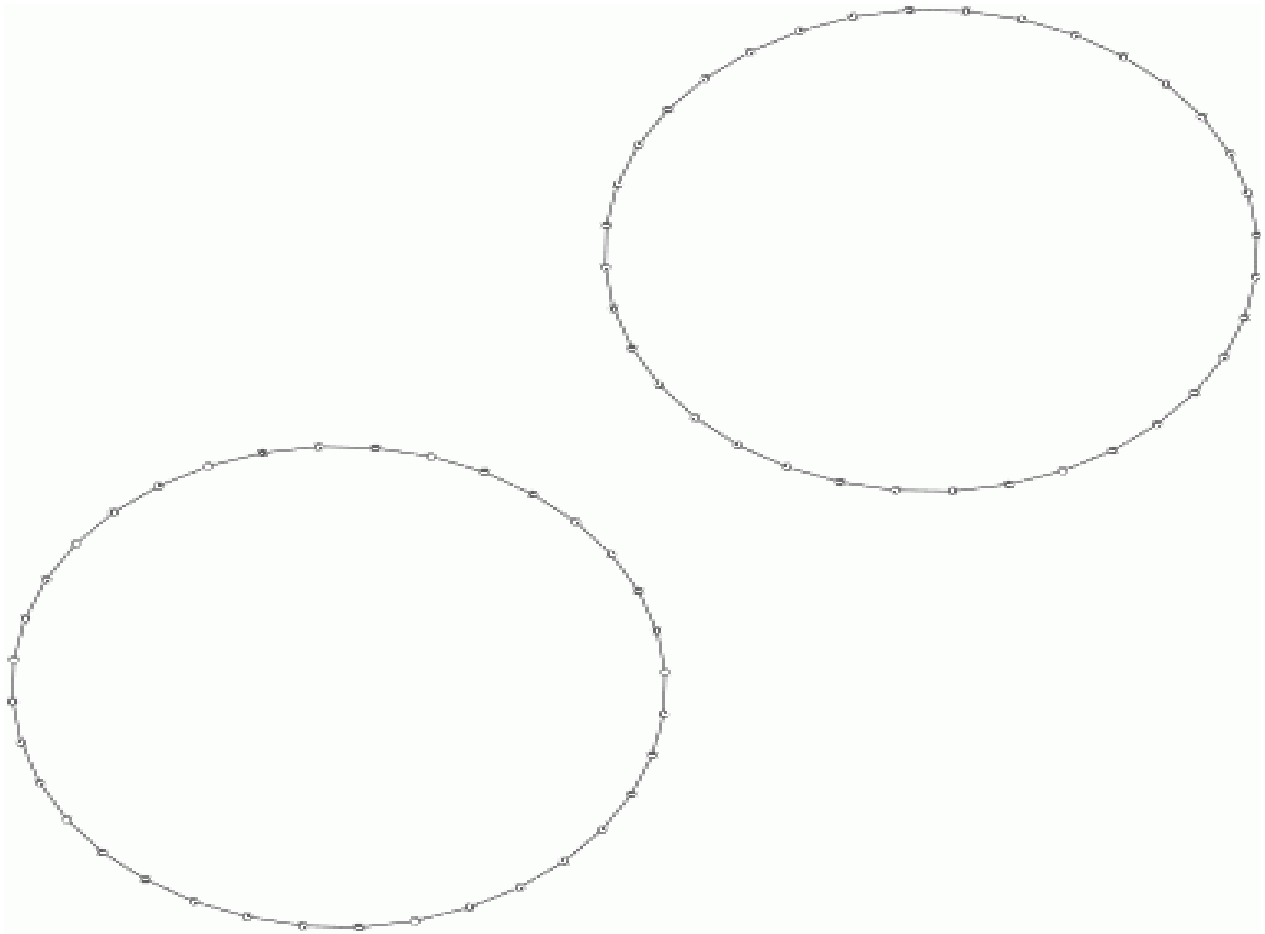
$$H_n(Y, \mathbb{F}) = \ker(\mathbb{F}^{s_n} \xrightarrow{\partial_n} \mathbb{F}^{s_{n-1}}) / \mathrm{im}(\mathbb{F}^{s_{n+1}} \xrightarrow{\partial_{n+1}} \mathbb{F}^{s_n})$$

$s_n =$ number of n-simplices in Y

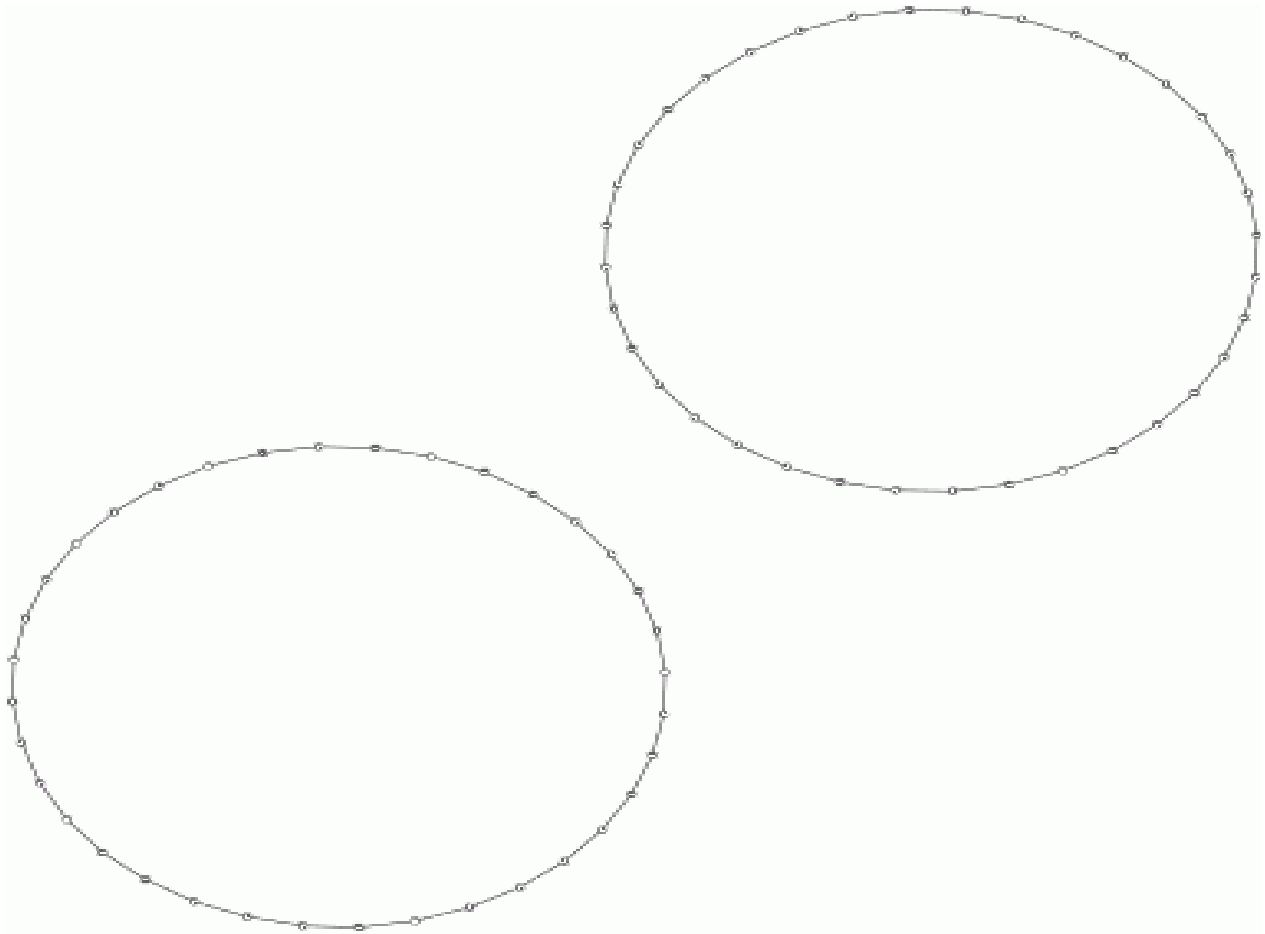
β_1 barcode for the toy data S



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



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



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

Theorem. [Edelsbrunner, Chazal, ...] For two finite metric spaces S, S' and $n \geq 0$ we have

$$d_{BottleNeck}(\beta_n^{**}(S), \beta_n^{**}(S')) \leq d_{GromovHaussdorf}(S, S') \ .$$

Caveat

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

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

inducing

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

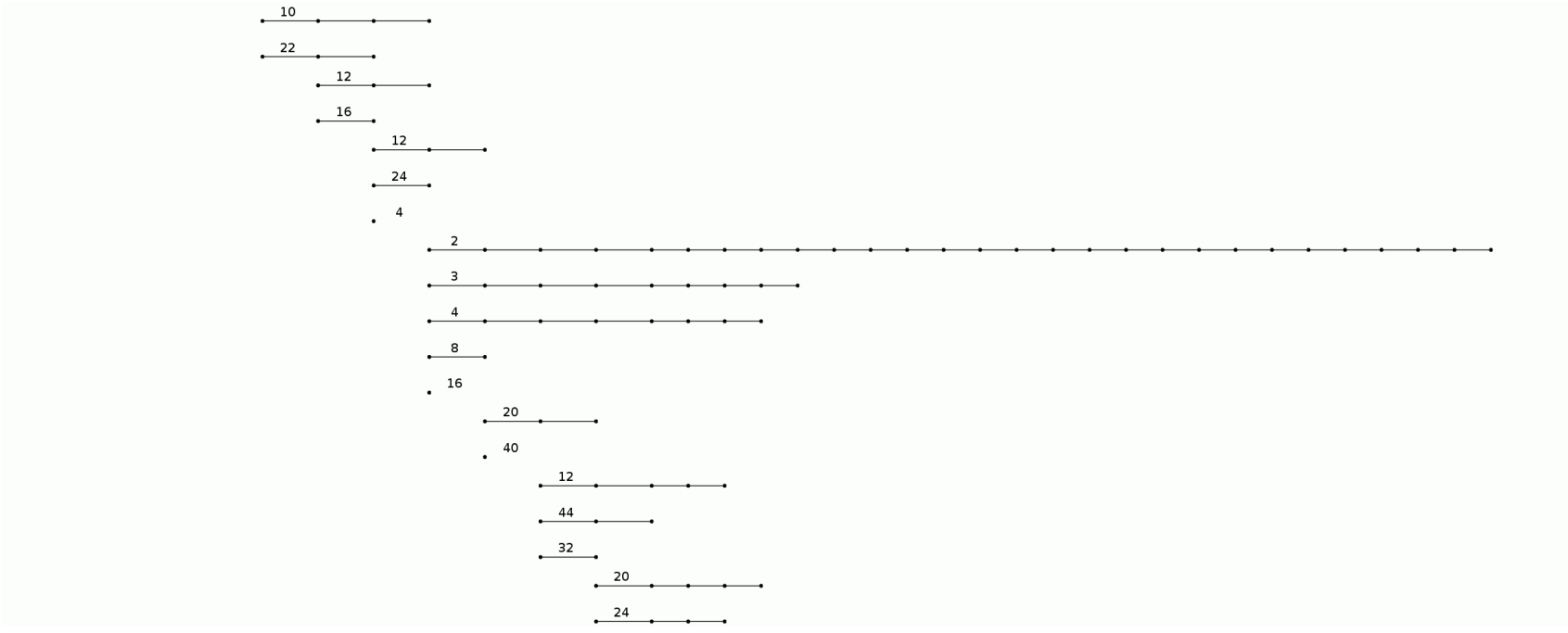
$$\mathbf{S} = \{\mathbf{v_1}, \mathbf{v_2}, \ldots, \mathbf{v_{225}}\} \subset \mathbf{X} = \mathbb{R}^{\mathbf{6}}$$

$$\mathbf{S} = \{\mathbf{v_1}, \mathbf{v_2}, \ldots, \mathbf{v_{225}}\} \subset \mathbf{X} = \mathbb{R}^{\mathbf{6}}$$

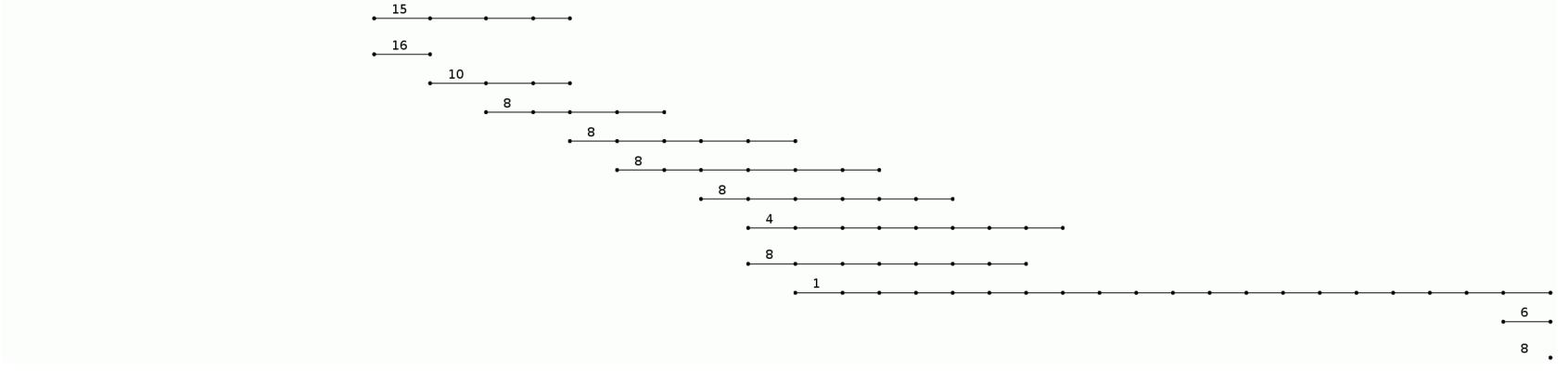
$$S \subset \text{image}(\phi)$$

$$\begin{array}{lcl} \phi\colon \mathbb{R}^2 & \longrightarrow & \mathbb{R}^6 \\ (x, y, z) & \mapsto & (\cos x, \sin x, \cos y, \sin y, \cos(x+y), \sin(x-y)). \end{array}$$

β_1



β_2



$$H_1(\mathbb{S}^1 \times \mathbb{S}^1, \mathbb{Q}) = \mathbb{Q} \oplus \mathbb{Q}$$

$$H_2(\mathbb{S}^1 \times \mathbb{S}^1, \mathbb{Q}) = \mathbb{Q}$$

$$H_1(\mathbb{S}^1 \times \mathbb{S}^1, \mathbb{Q}) = \mathbb{Q} \oplus \mathbb{Q}$$

$$H_2(\mathbb{S}^1 \times \mathbb{S}^1, \mathbb{Q}) = \mathbb{Q}$$

$$H_1(\mathbb{S}^2 \vee \mathbb{S}^1 \vee \mathbb{S}^1, \mathbb{Q}) = \mathbb{Q} \oplus \mathbb{Q}$$

$$H_2(\mathbb{S}^1 \vee \mathbb{S}^1 \vee \mathbb{S}^1, \mathbb{Q}) = \mathbb{Q}$$

$$H_1(\mathbb{S}^1 \times \mathbb{S}^1, \mathbb{Q}) = \mathbb{Q} \oplus \mathbb{Q}$$

$$H_2(\mathbb{S}^1 \times \mathbb{S}^1, \mathbb{Q}) = \mathbb{Q}$$

$$H_1(\mathbb{S}^2 \vee \mathbb{S}^1 \vee \mathbb{S}^1, \mathbb{Q}) = \mathbb{Q} \oplus \mathbb{Q}$$

$$H_2(\mathbb{S}^1 \vee \mathbb{S}^1 \vee \mathbb{S}^1, \mathbb{Q}) = \mathbb{Q}$$

$$\cup\colon H^1(X, \mathbb{Q}) \times H^1(X, \mathbb{Q}) \to H^2(X, \mathbb{Q})$$

distinguishes between the cases.

Finite presentation for

$$\pi_1(X, x_0) = \{p \colon [0,1] \longrightarrow X \ : \ p(0) = p(1) = x_0\}/\simeq$$

implemented in GAP for finite cellular spaces X

Finite presentation for

$$\pi_1(X, x_0) = \{p \colon [0,1] \longrightarrow X \ : \ p(0) = p(1) = x_0\}/\simeq$$

implemented in GAP for finite cellular spaces X

and yields the cup product

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

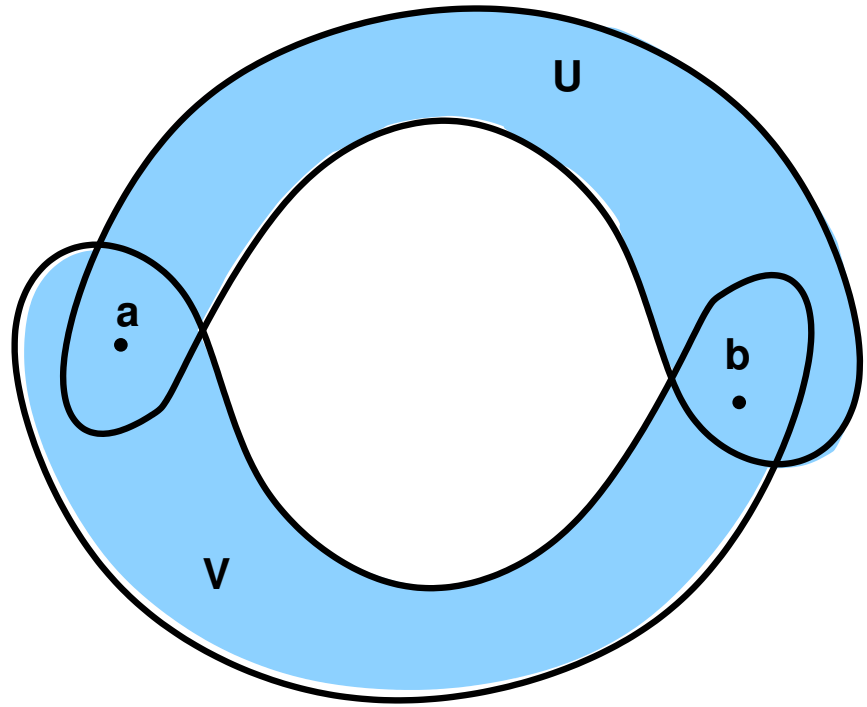
induced by

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

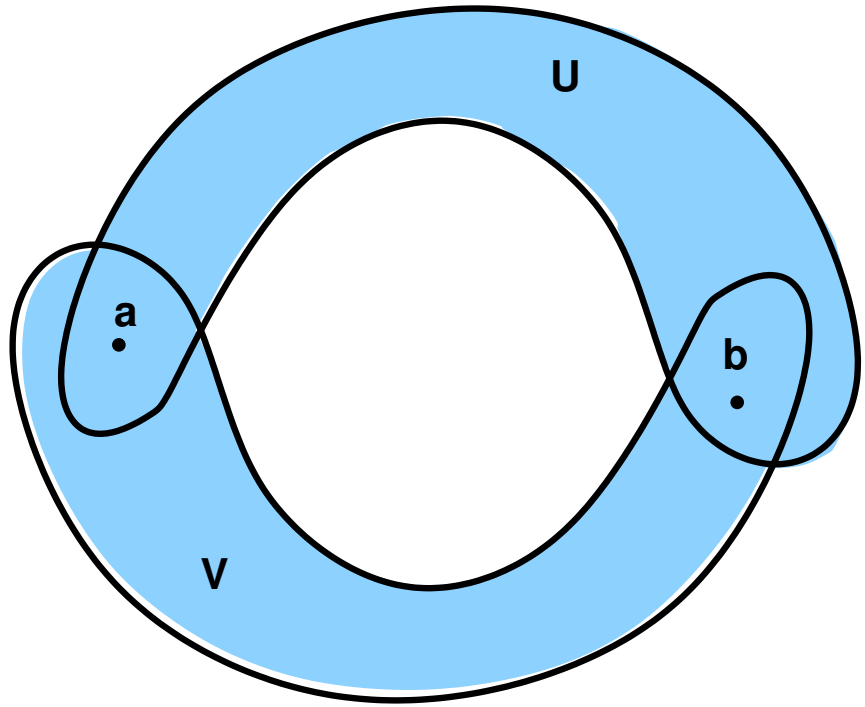
Parallel computation of $\pi_1 X$?

Parallel computation of $\pi_1 X$?

$X =$



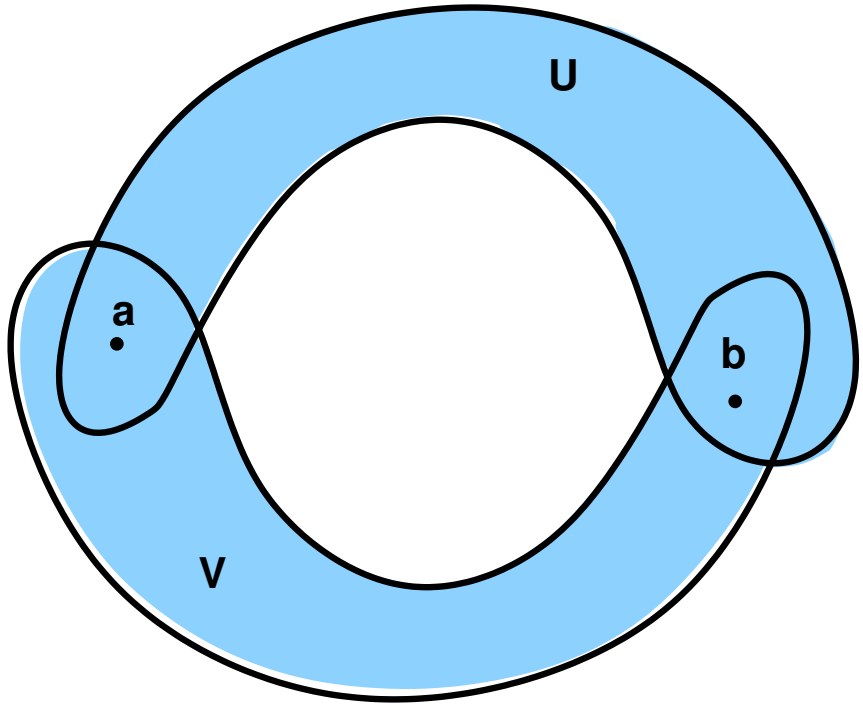
Parallel computation of $\pi_1 X$?

$X =$ 

Enter groupoids

$$\pi_1(X, X_0) = \{p \colon [0,1] \longrightarrow X \ : \ p(0), p(1) \in X_0\}$$

Parallel computation of $\pi_1 X$?

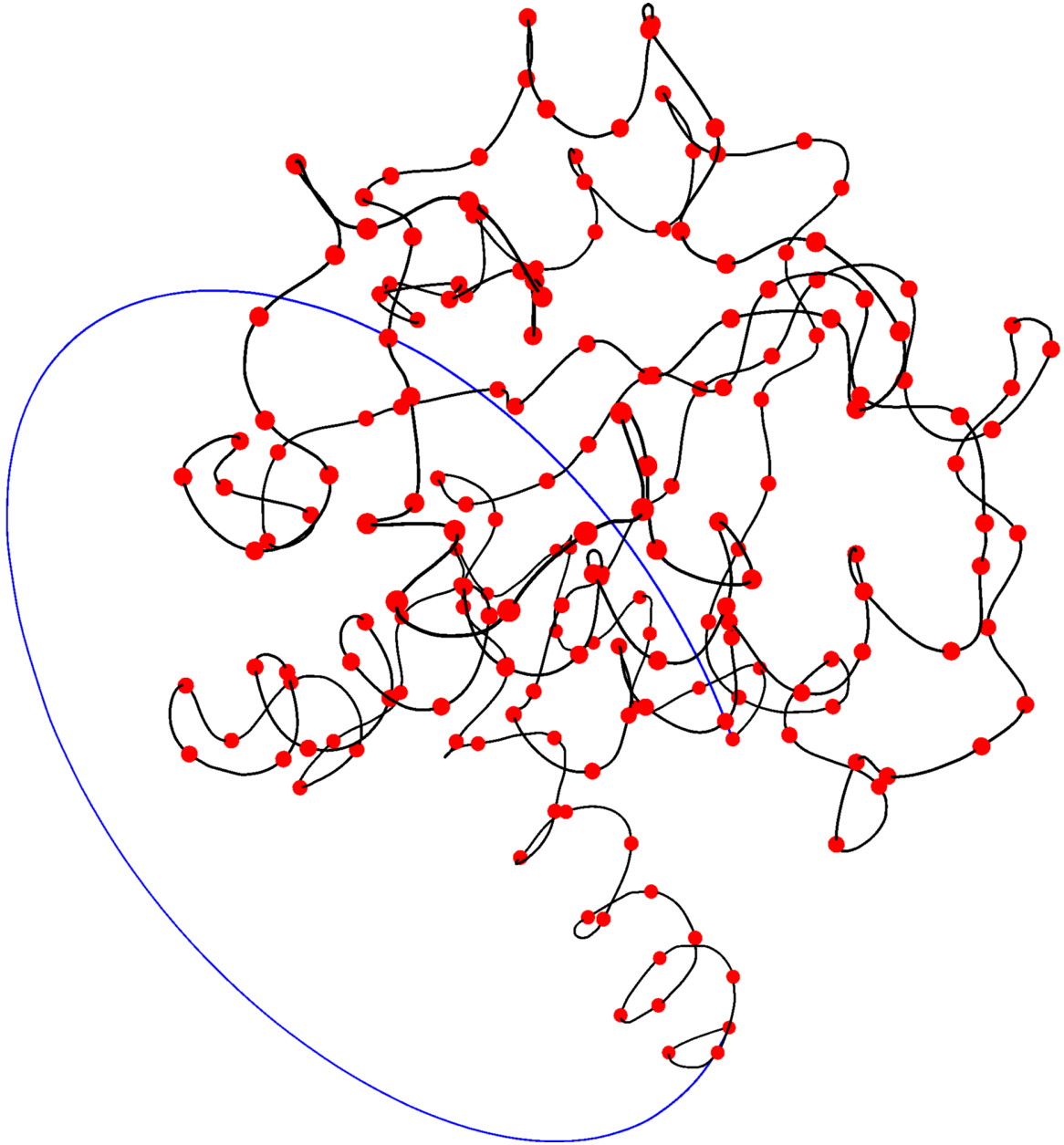
$X =$ 

Enter groupoids

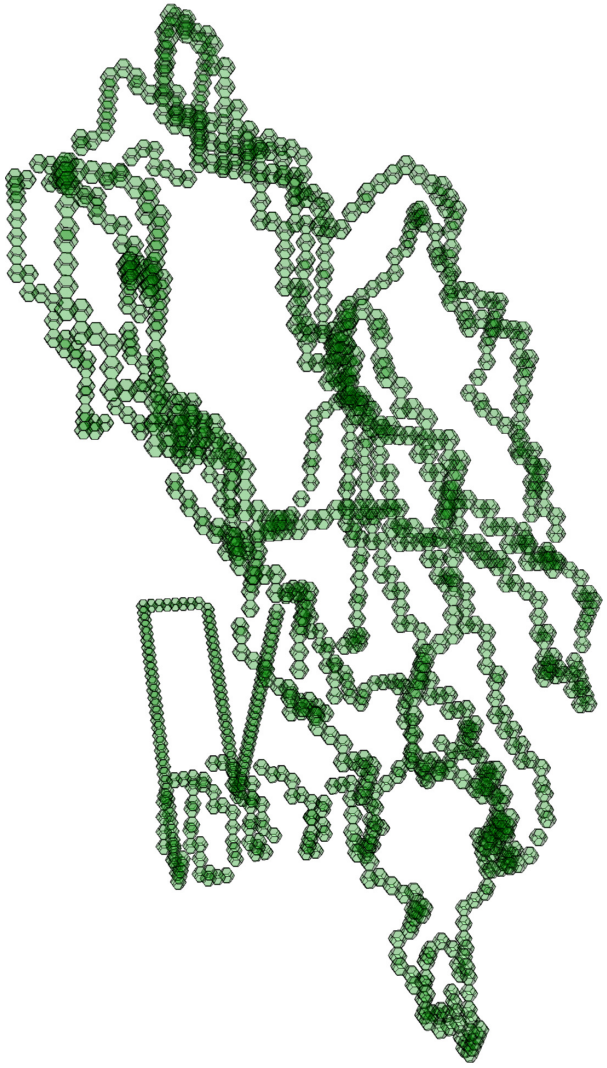
$$\pi_1(X, X_0) = \{p\colon [0,1] \longrightarrow X \ : \ p(0), p(1) \in X_0\}$$

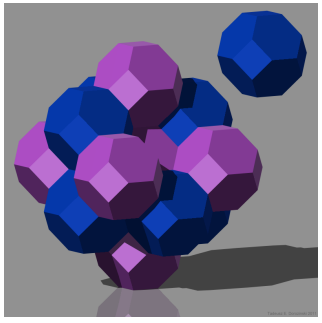
$$\pi_1(U \cup V, \{a,b\}) \cong \pi_1(U, \{a,b\}) *_{\pi_1(U \cap V, \{a,b\})} \pi_1(V, \{a,b\})$$

1V2X protein backbone



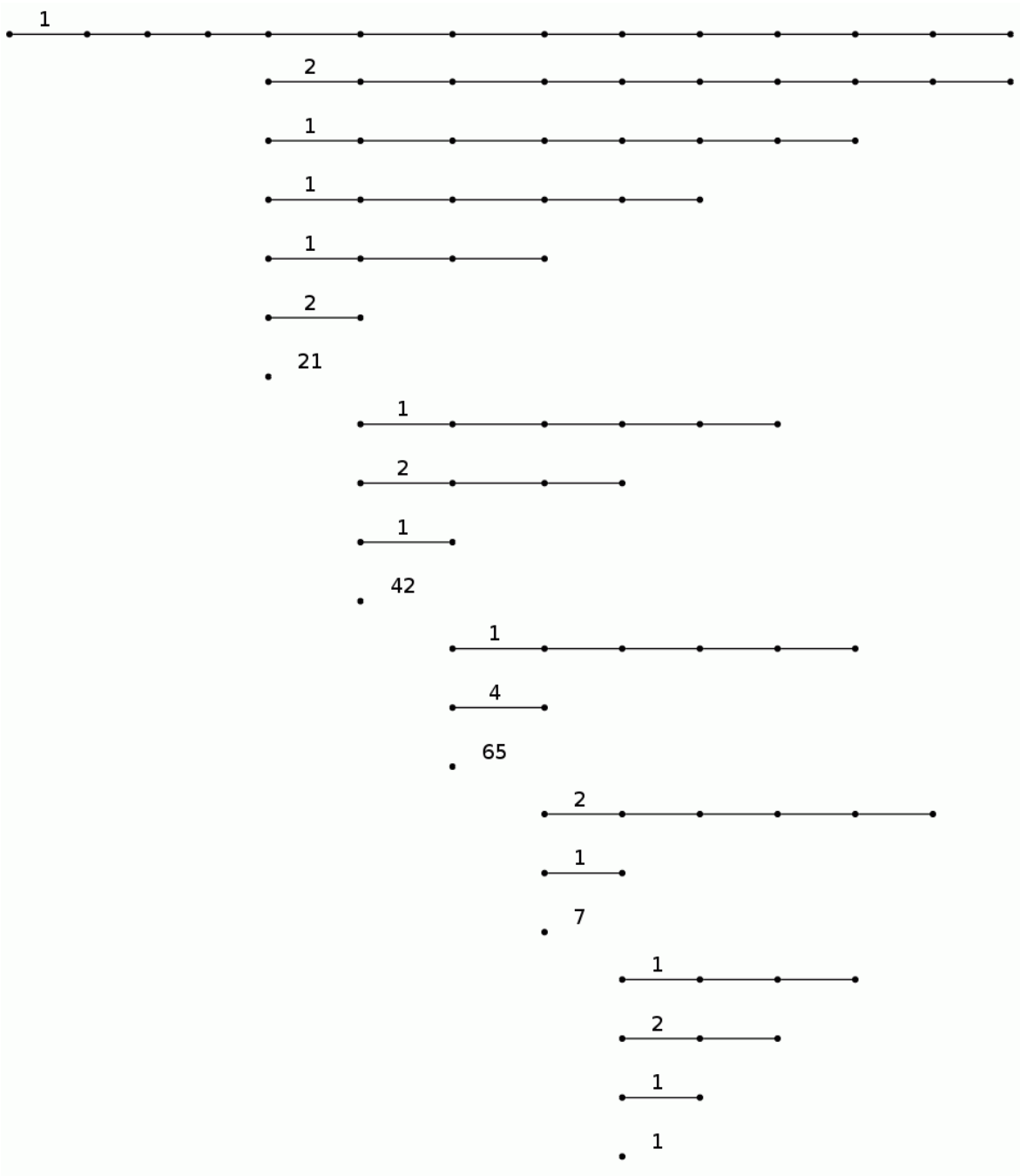
1V2X protein backbone





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

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

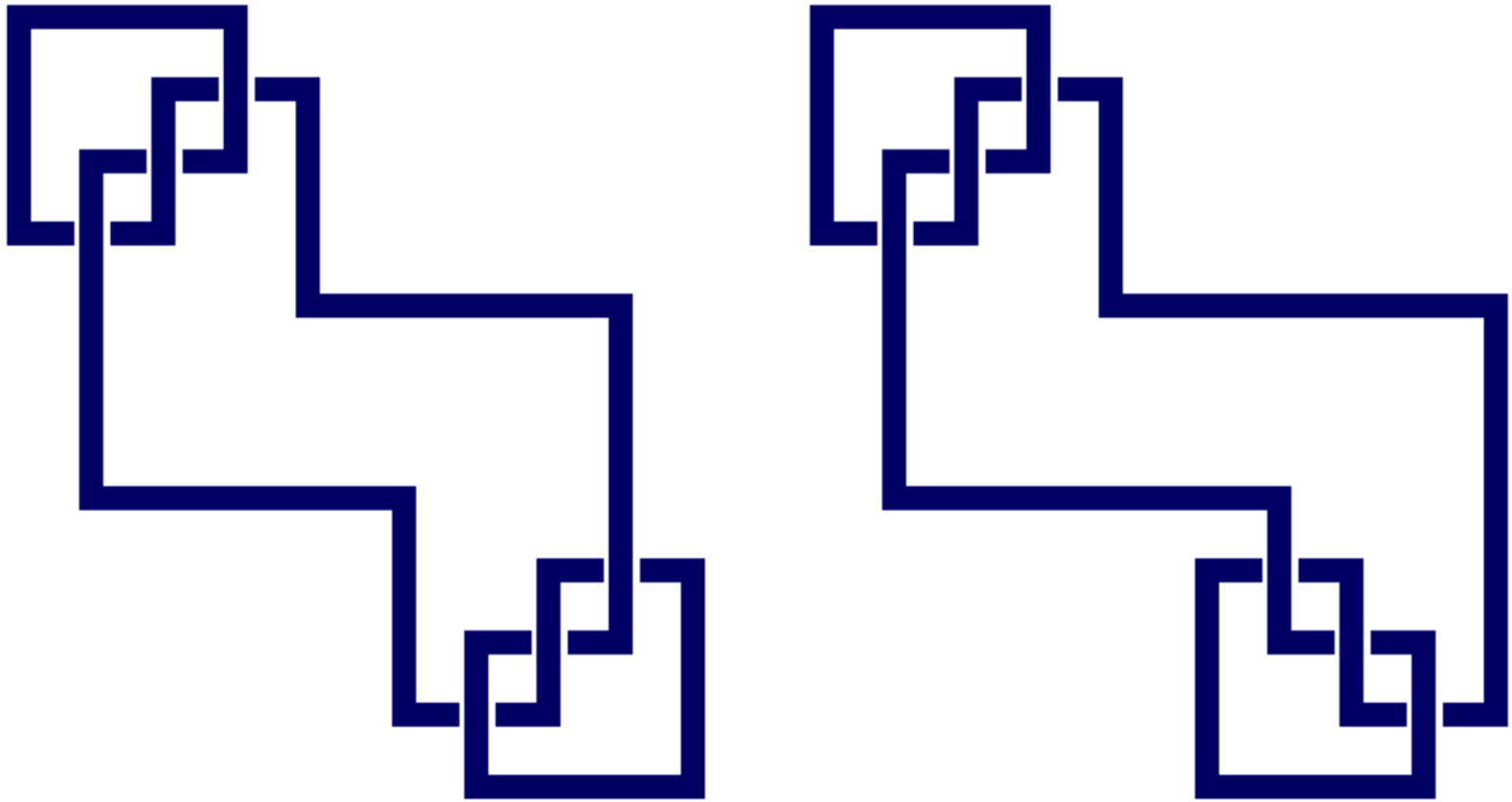
$$I_3(\langle x, y | yx^{-1}yxy^{-1}x \rangle) = \{\mathbb{Z},\ \mathbb{Z} \oplus \mathbb{Z}_3,\ \mathbb{Z} \oplus \mathbb{Z},\ \mathbb{Z} \oplus \mathbb{Z}_2 \oplus \mathbb{Z}_2\}$$

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

min value of n to distinguish between knots on c crossings

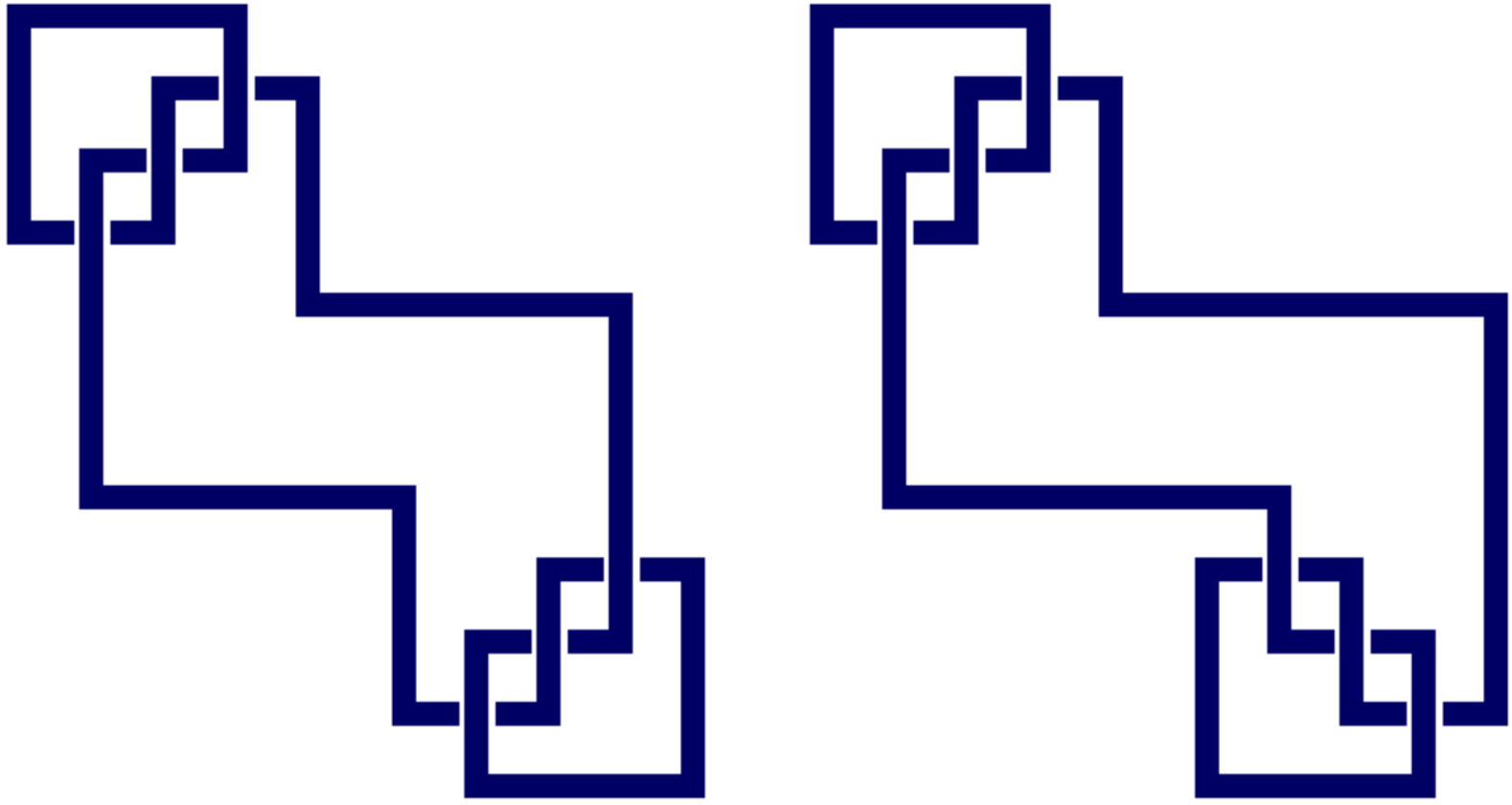
c	3	4	5	6	7	8	9	10	11	12	13	14
n	2	2	3	3	3	3	5	5	6	6	7	7

Brendel, E., Juda, Mrozek



$$\pi_1(\mathbb{R}^3 \setminus (K + K)) \cong \pi_1(\mathbb{R}^3 \setminus (K + L))$$

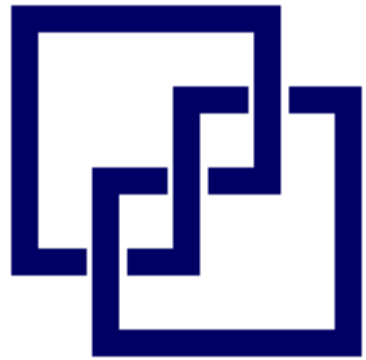
$$K + K \neq K + L$$



$$\pi_1(\mathbb{R}^3 \setminus (K + K)) \cong \pi_1(\mathbb{R}^3 \setminus (K + L))$$

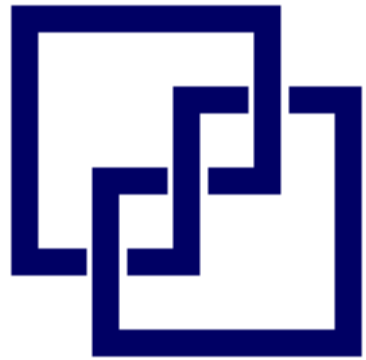
$$K + K \neq K + L$$

$$K + K + K + L \neq K + K + L + L?$$



Proposition: *The alpha carbon atoms of the Thermus Thermophilus protein determine a knot K with peripheral system*

$$\begin{array}{rcl} \pi_1(\partial K) \cong \langle m, \ell | m\ell m^{-1}\ell^{-1} \rangle & \to & \pi_1(\mathbb{R}^3 \setminus K) \cong \langle x, y | xyx = yxy \rangle \\ m & \mapsto & x^{-2}yx^2y \\ \ell & \mapsto & x \end{array} .$$



Proposition: *The alpha carbon atoms of the Thermus Thermophilus protein determine a knot K with peripheral system*

$$\begin{array}{rcl} \pi_1(\partial K) \cong \langle m, \ell | m\ell m^{-1}\ell^{-1}\rangle & \to & \pi_1(\mathbb{R}^3 \setminus K) \cong \langle x, y | xyx = yxy\rangle \\ m & \mapsto & x^{-2}yx^2y \\ \ell & \mapsto & x \end{array} .$$

For $G = \pi_1(\mathbb{R}^3 \setminus K), H = \pi_1(\partial K) < G$ set

$$Q(K) = G/H = \{Hg \ : \ g \in G\}$$

$$(Hg) * (Hf) = H\ell^{-1}gf^{-1}\ell f$$

A **quandle** is a set Q with binary operation $*$ such that

1) $x * x = x$

2) right multiplication $R_x \colon Q \to Q, y \mapsto y * x$ is an automorphism

for all $x \in Q$.

A **quandle** is a set Q with binary operation $*$ such that

1) $x * x = x$

2) right multiplication $R_x \colon Q \to Q, y \mapsto y * x$ is an automorphism

for all $x \in Q$.

For any finite quandle Q we have a knot invariant:

$$|Hom(Q(K), Q)|$$

A **quandle** is a set Q with binary operation $*$ such that

1) $x * x = x$

2) right multiplication $R_x \colon Q \to Q, y \mapsto y * x$ is an automorphism

for all $x \in Q$.

For any finite quandle Q we have a knot invariant:

$$|Hom(Q(K), Q)|$$

Q is **connected** if the group $Inn(Q) = \langle R_x \ : \ x \in Q \rangle$ acts transitively on Q.

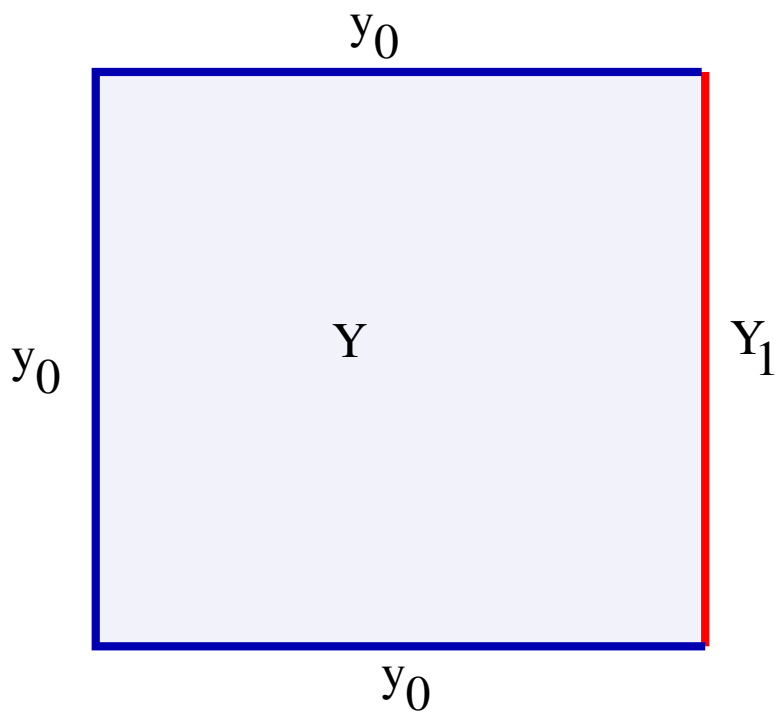
```
gap> K:=ReadPDBfile("1V2X.pdb");
Pure permutahedral complex of dimension 3

gap> Y:=RegularCWComplex(PureComplexComplement(K));;
Regular CW-complex of dimension 3

gap> i:=Boundary(Y);
Map of regular CW-complexes

gap> phi:=FundamentalGroup(i,22495);
[ f1, f2 ] -> [ f1^-3*f2*f1^2*f2*f1, f1 ]
```

```
gap> Q:=ConnectedQuandle(24,17,"import");;
gap> K:=PureCubicalKnot(3,1);;
gap> L:=ReflectedCubicalKnot(K);;
gap> KKKL:=KnotSum(KnotSum(KnotSum(K,K),K),L);;
gap> KKLL:=KnotSum(KnotSum(KnotSum(K,K),L),L);;
gap> A:=PresentationKnotQuandle(KKKL);;
gap> B:=PresentationKnotQuandle(KKLL);;
gap> NumberOfHomomorphisms(A,Q);
1176
gap> NumberOfHomomorphisms(B,Q);
7704
```

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 ]
```