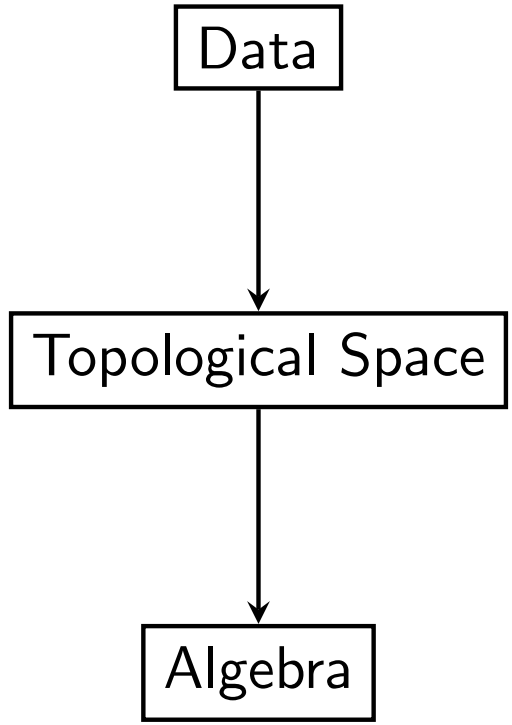
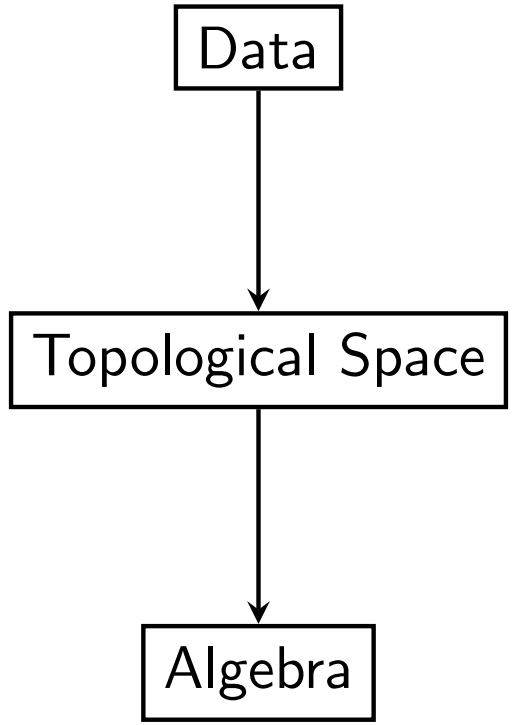
A flavour of topological data analysis

Final MA342 lecture


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.

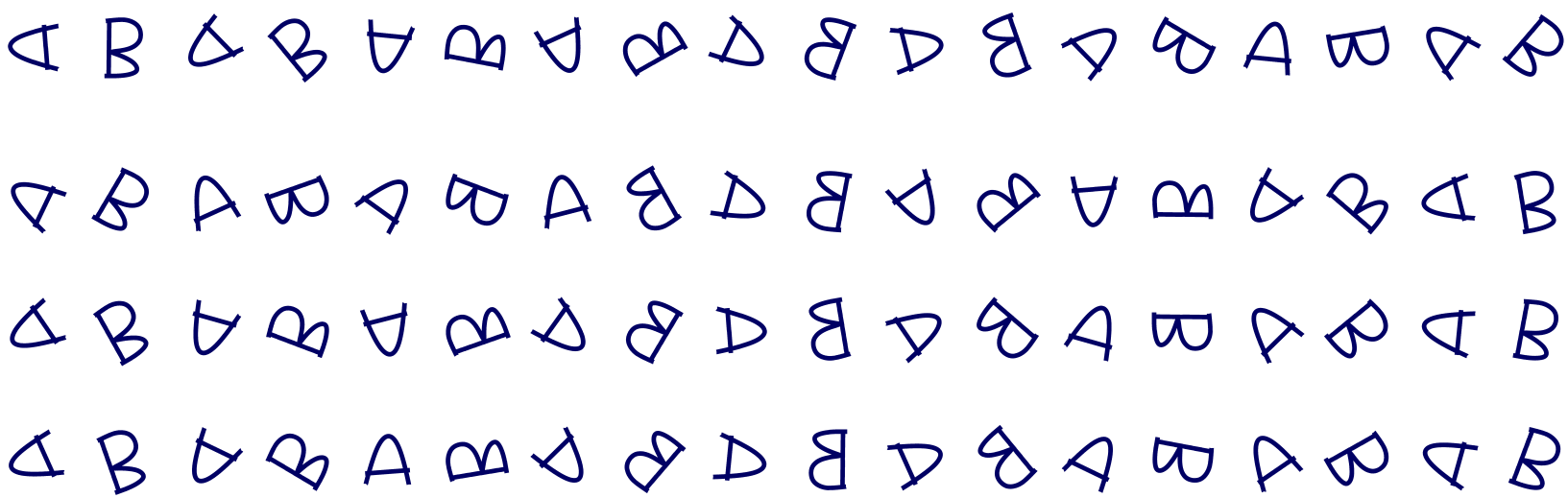
Cluster Analysis

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.

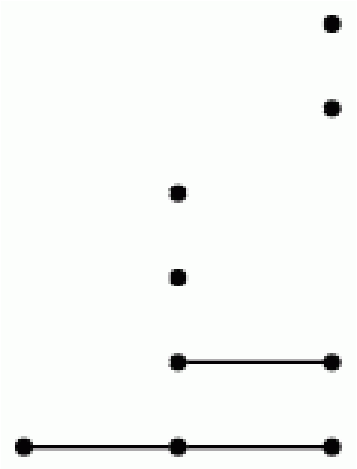
$$\beta_0^{s,t} = rank(H_0(Y_s, \mathbb{Q}) \to H_0(Y_t, \mathbb{Q})), \qquad s \leq t.$$

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

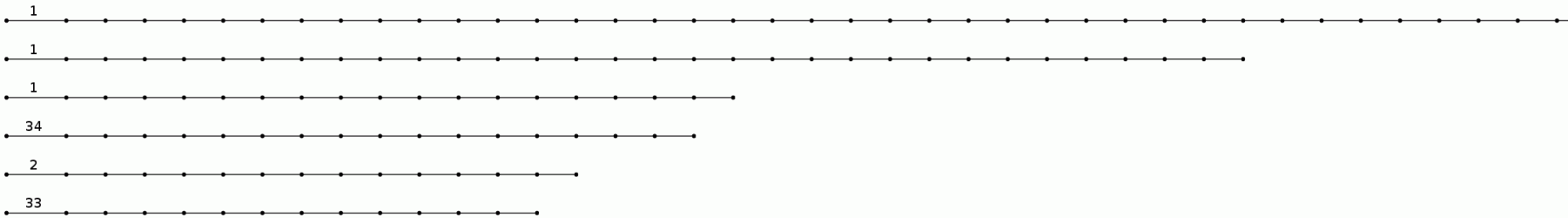
A β_n bar code has

$\beta_n^{s,t}$ *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.

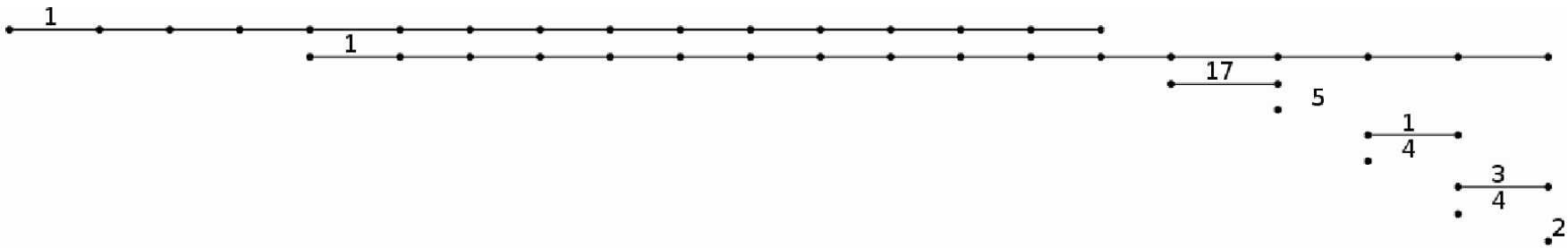
$$\beta_n^{s,t} = rank(H_n(Y_s, \mathbb{Q}) \to H_n(Y_t, \mathbb{Q})), \qquad s \leq t.$$

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

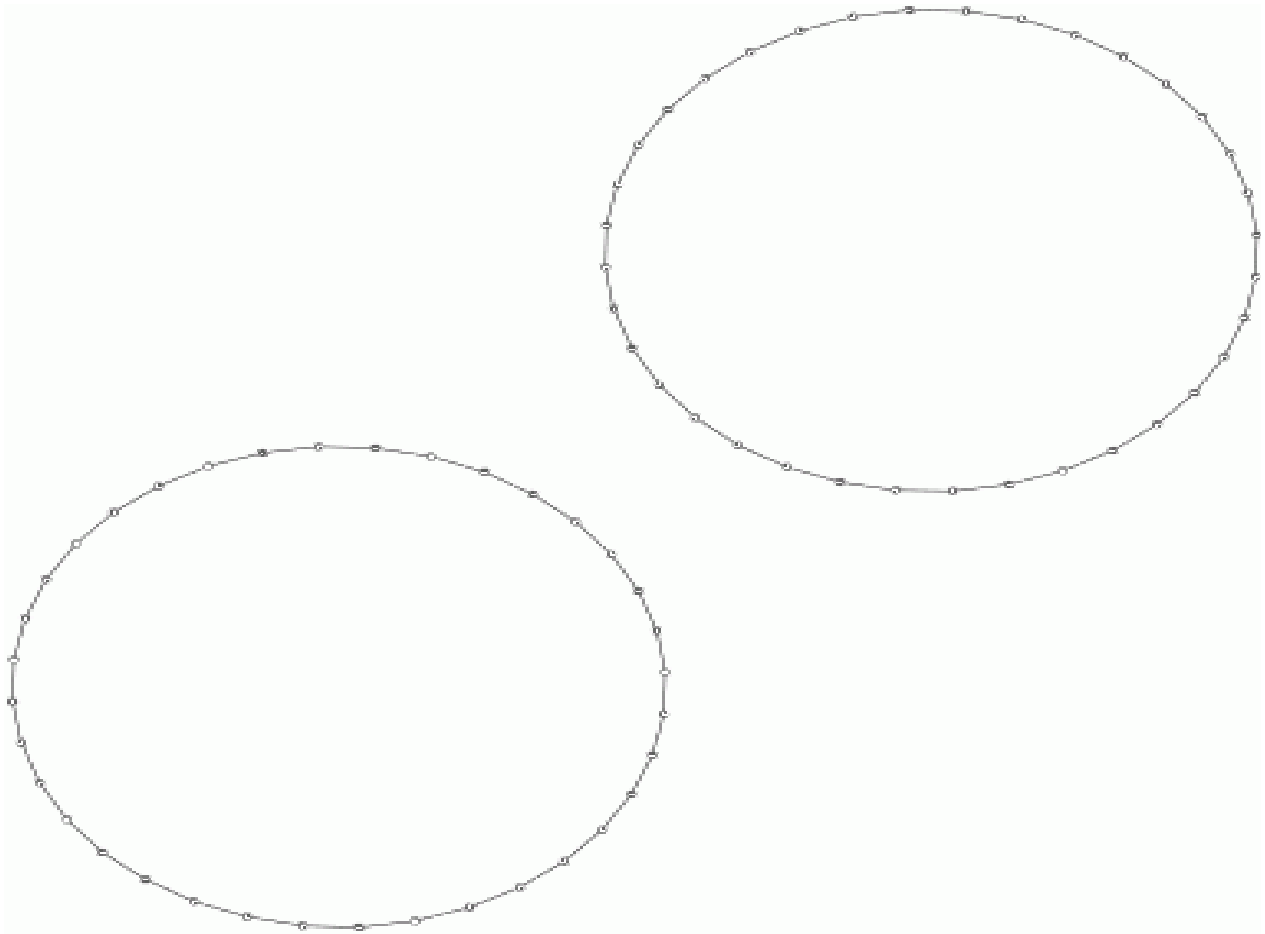
$$H_n(Y, \mathbb{F}) = \ker(\mathbb{F}^{s_n} \xrightarrow{\partial_n} \mathbb{F}^{s_{n-1}}) / \mathrm{im}(\mathbb{F}^{s_{n+1}} \xrightarrow{\partial_{n+1}} \mathbb{F}^{s_n})$$

$s_n =$ number of n-simplices in Y

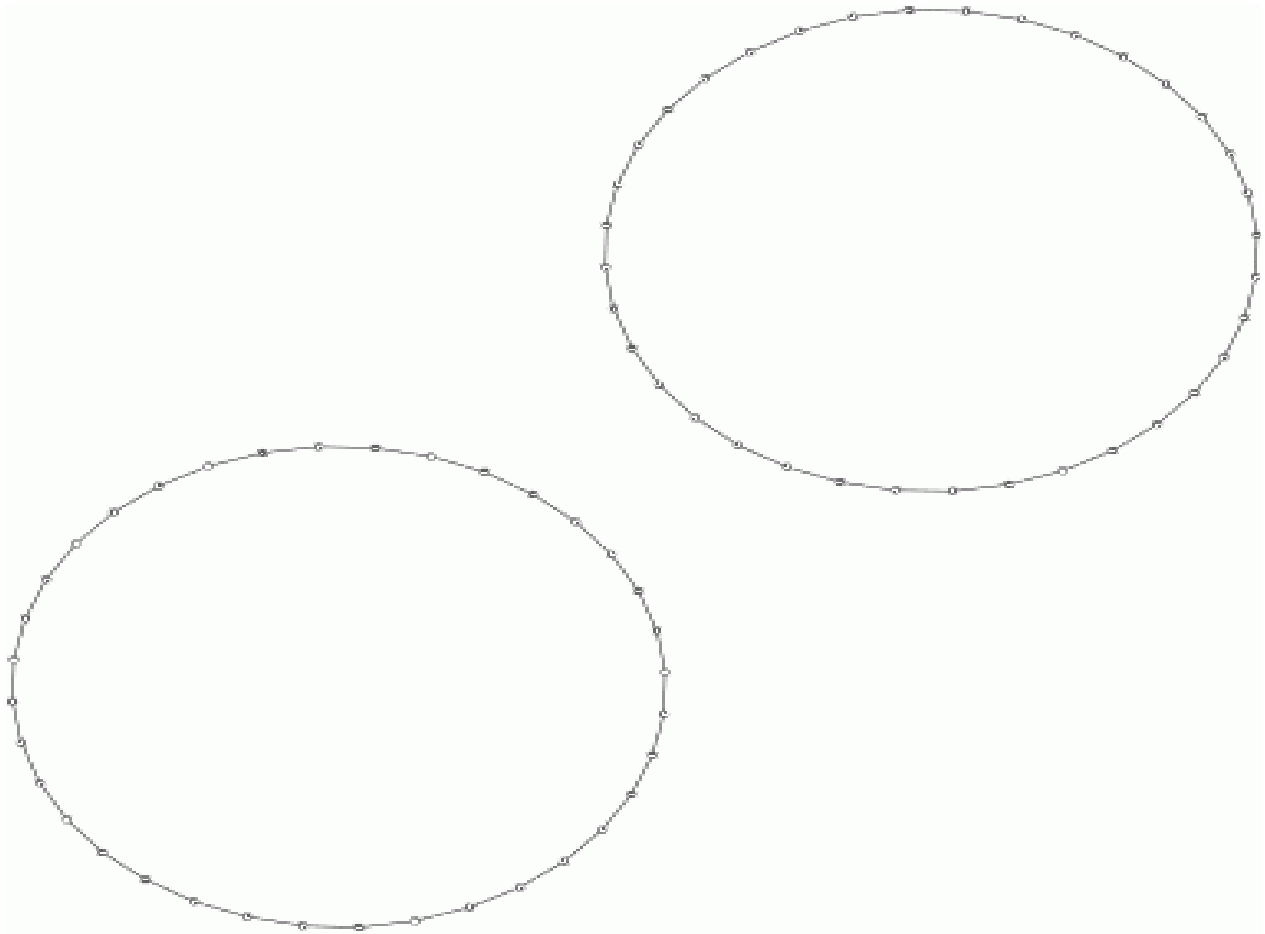
β_1 barcode for the toy data S



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



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



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

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

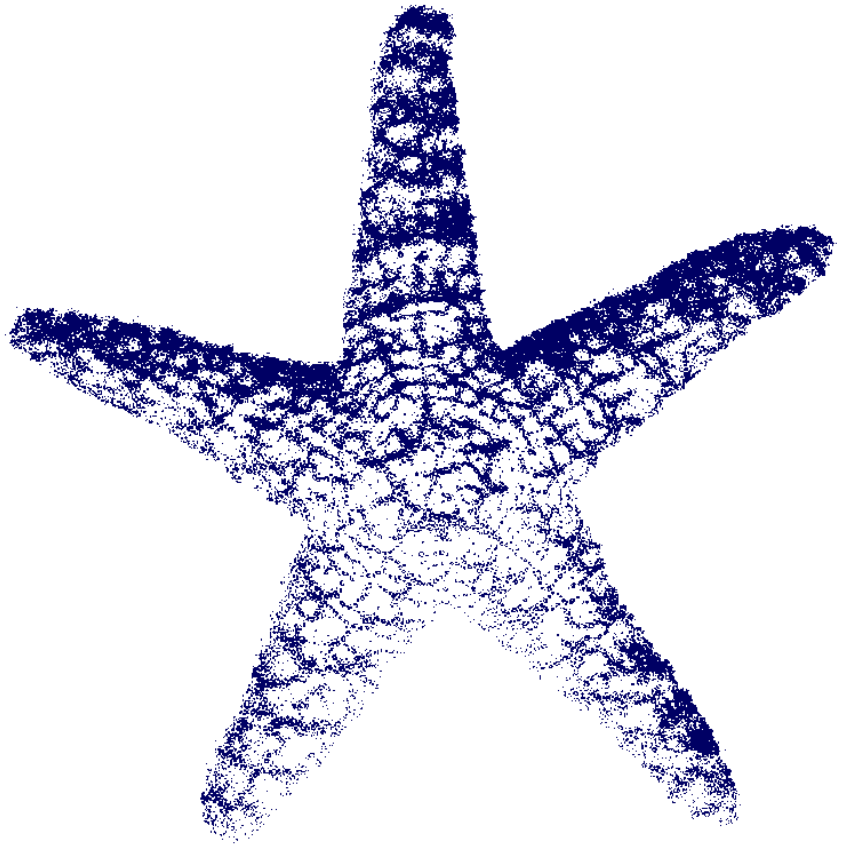
$\mathcal{W} = \{W_\alpha = S \cap f^{-1} U_\alpha\}$ is a cover of S.
$\mathcal{V} =$ set of clusters formed by clustering each W_α

Output

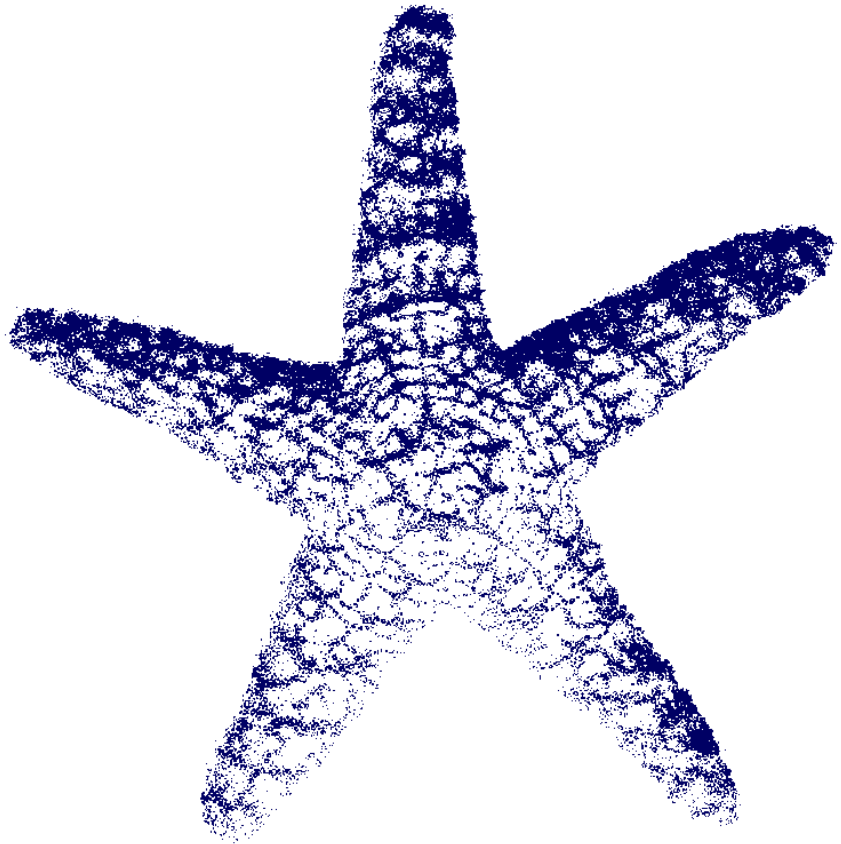
Simplicial nerve of $\mathcal{V}$.

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



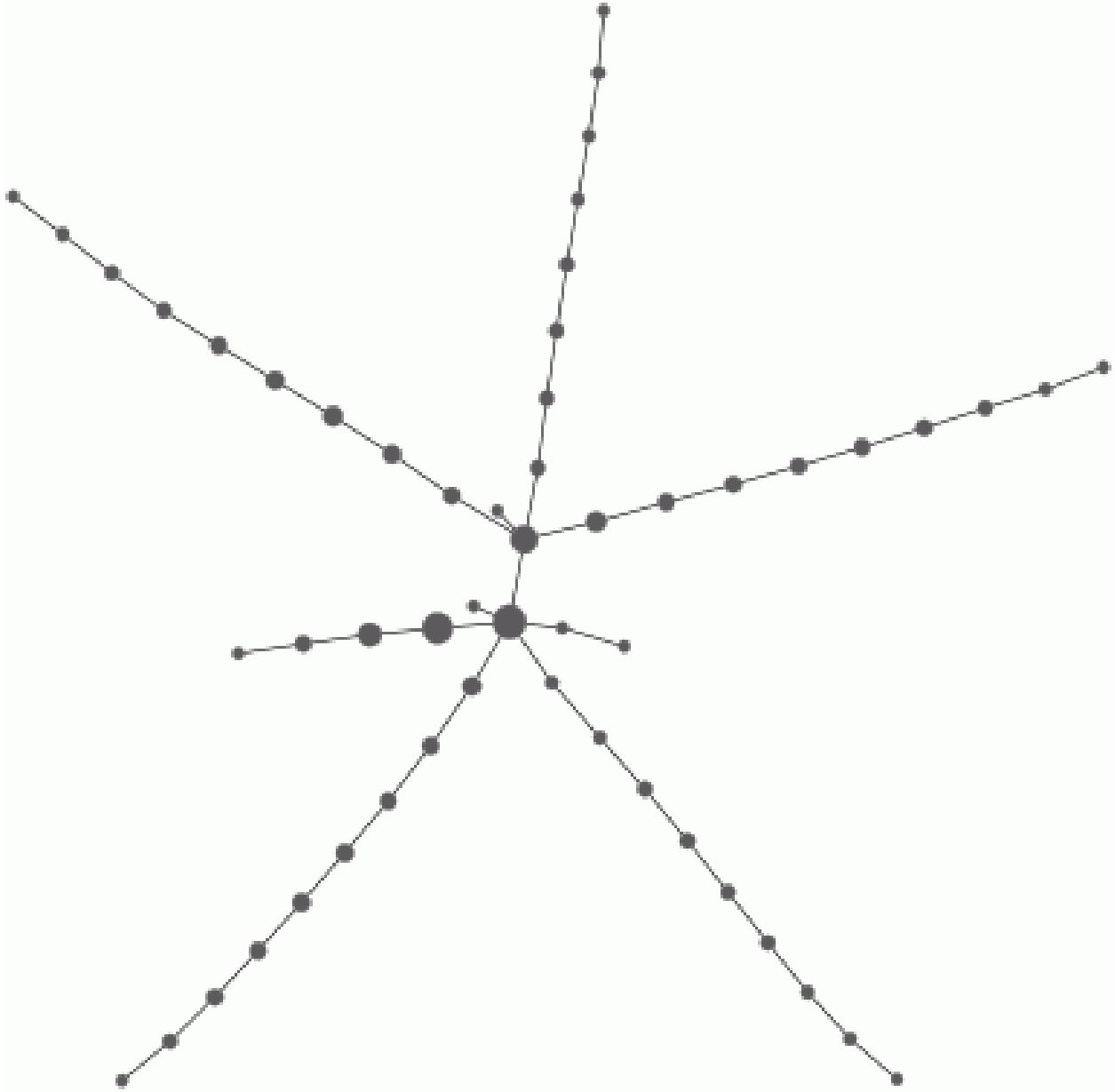
Consider $S = \{v_1, \ldots, v_{200}\} \subset X \subset \mathbb{R}^2$ where X is



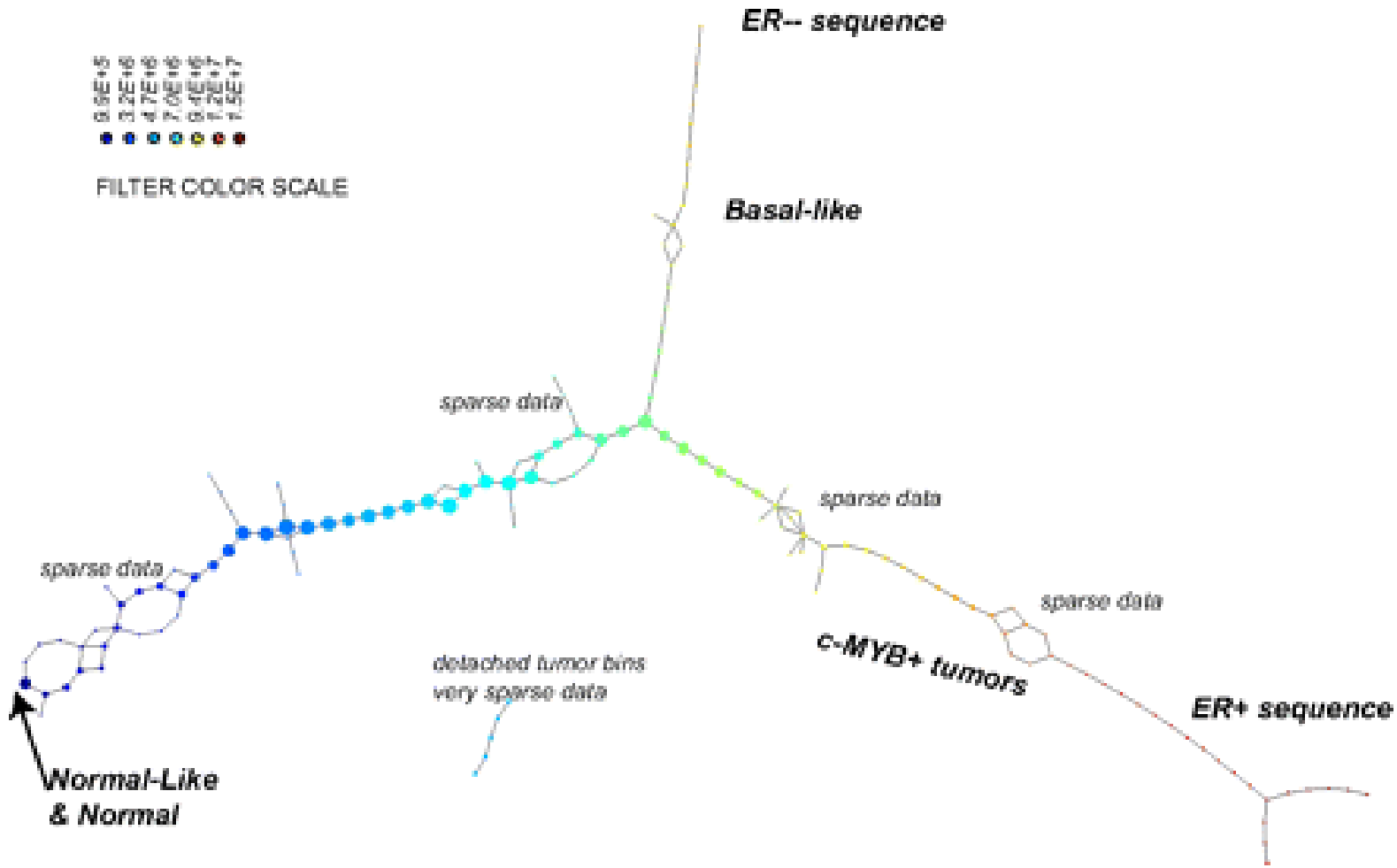
Choose $f : X \to [0, \infty), x \mapsto d(v_1, x)$

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

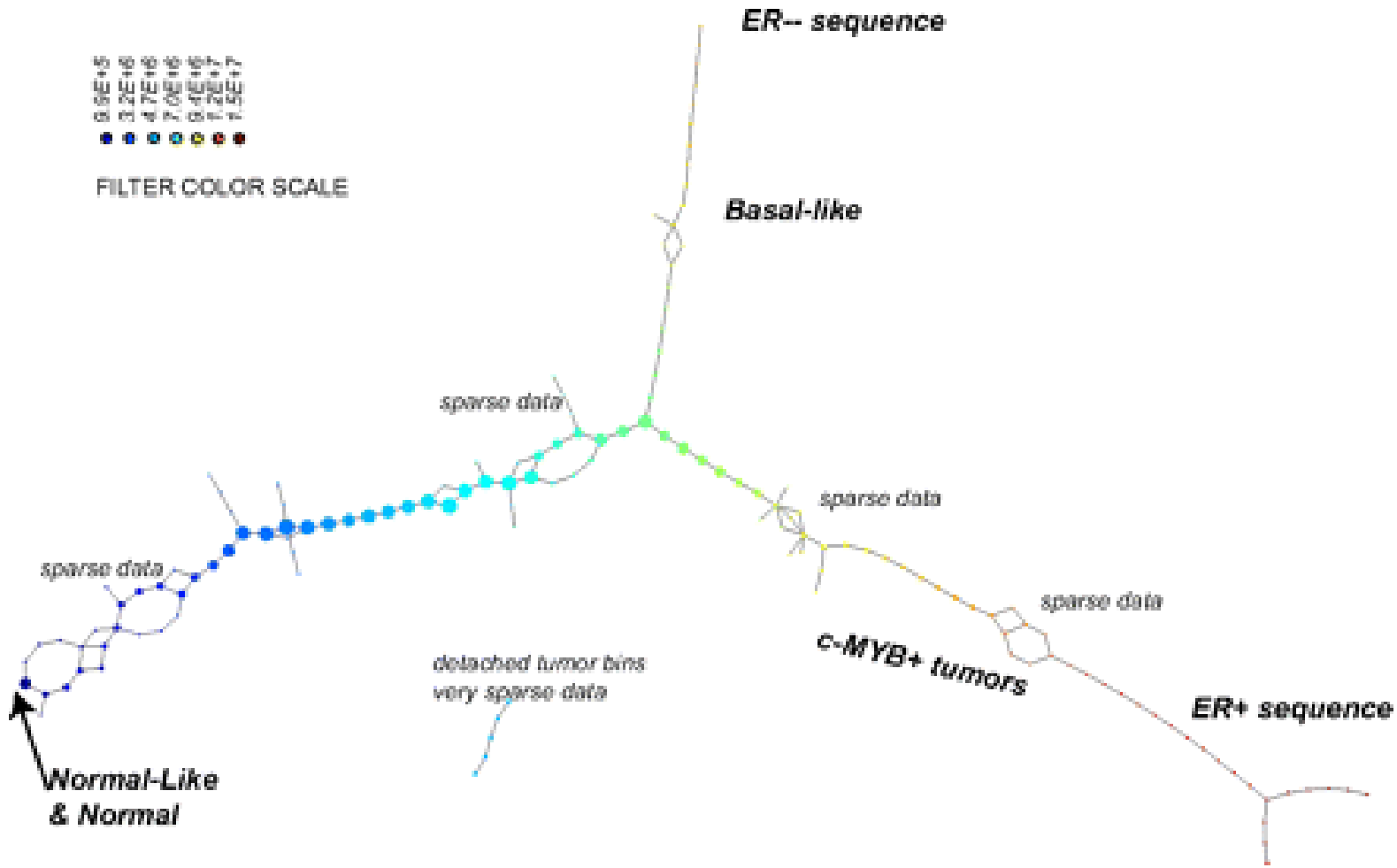
Mapper output for starfish sample



Mapper output for breast cancer microarray gene expression data

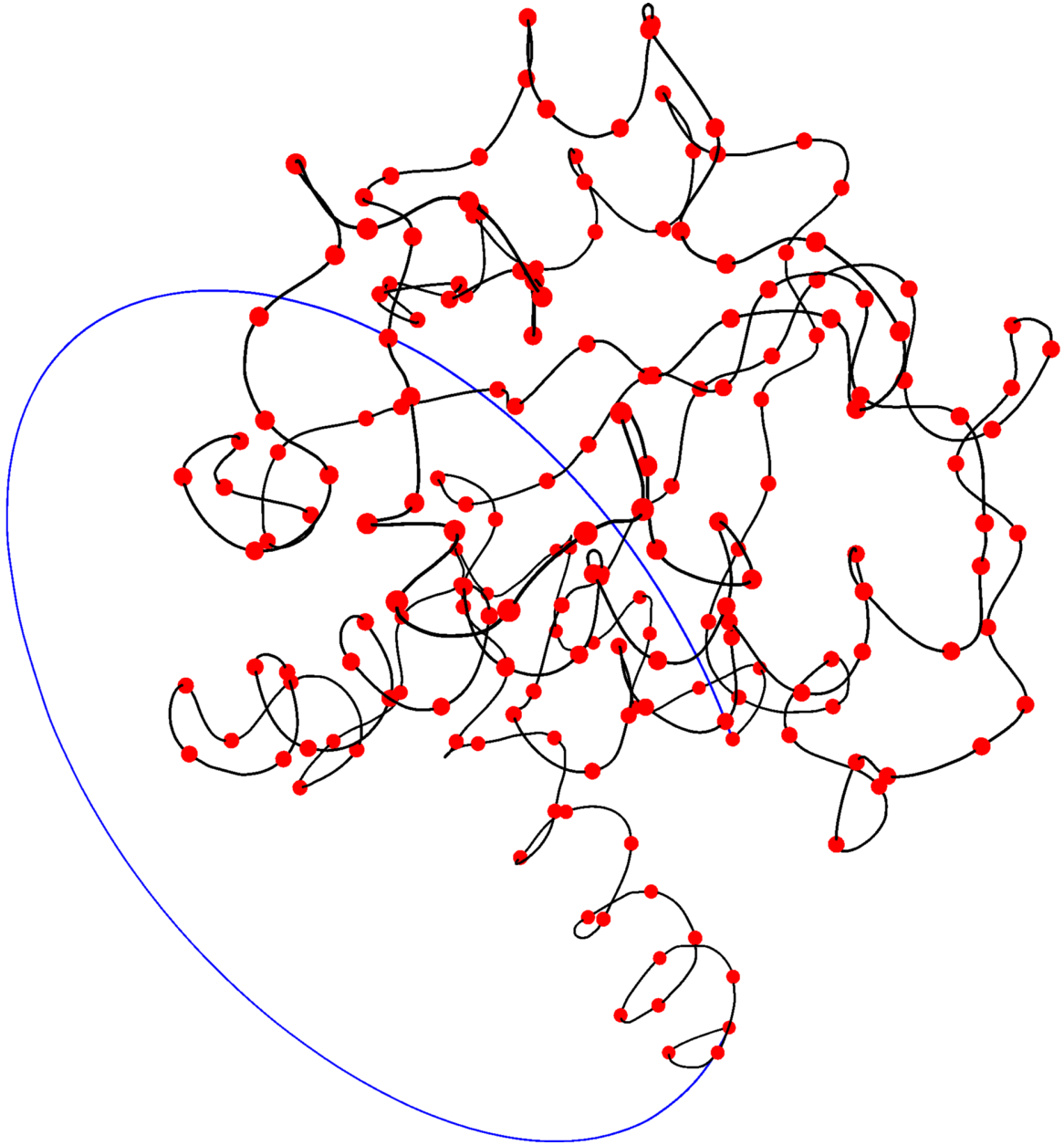


Mapper output for breast cancer microarray gene expression data

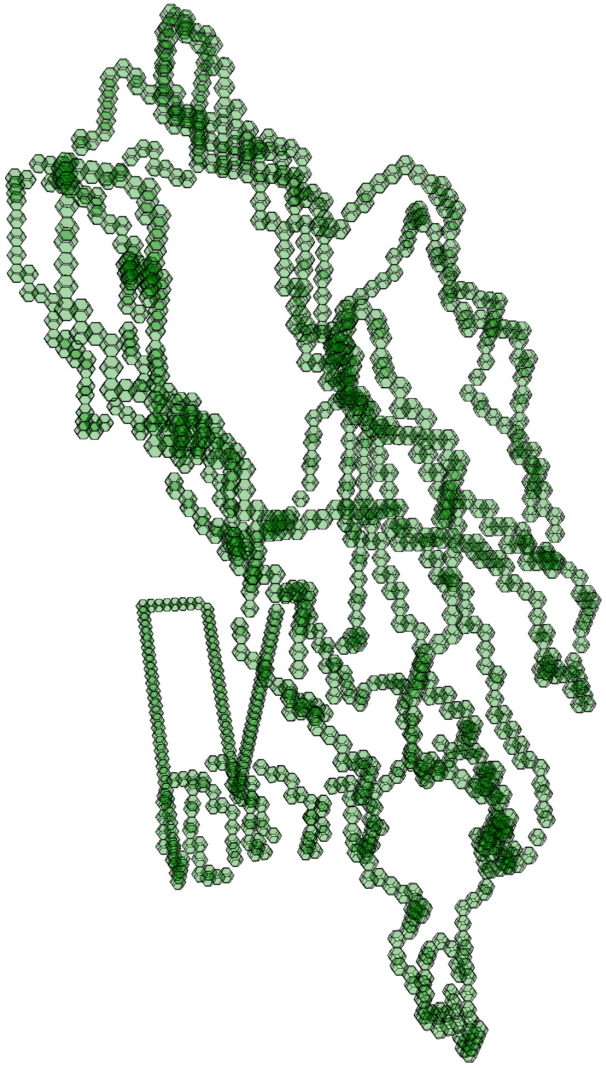


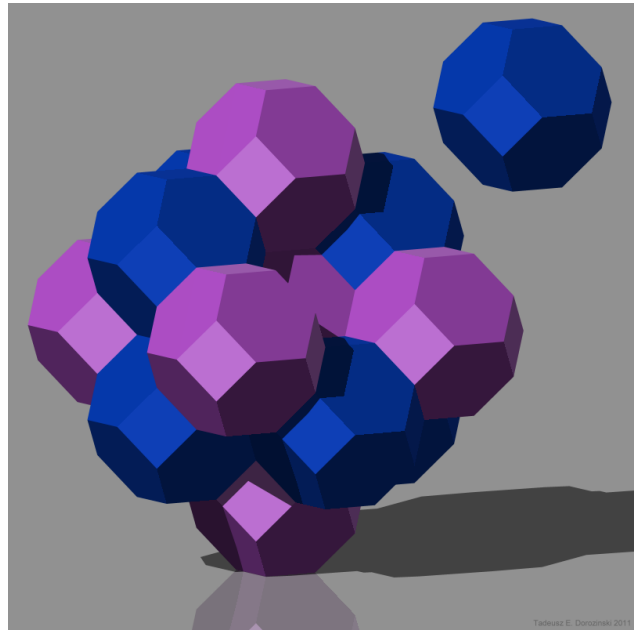
Nicolau, Levine, Carlsson (PNAS, 2011): identified a subgroup of ER+ breast cancers. These patients exhibit 100% survival.

1V2X protein backbone: Is it knotted?



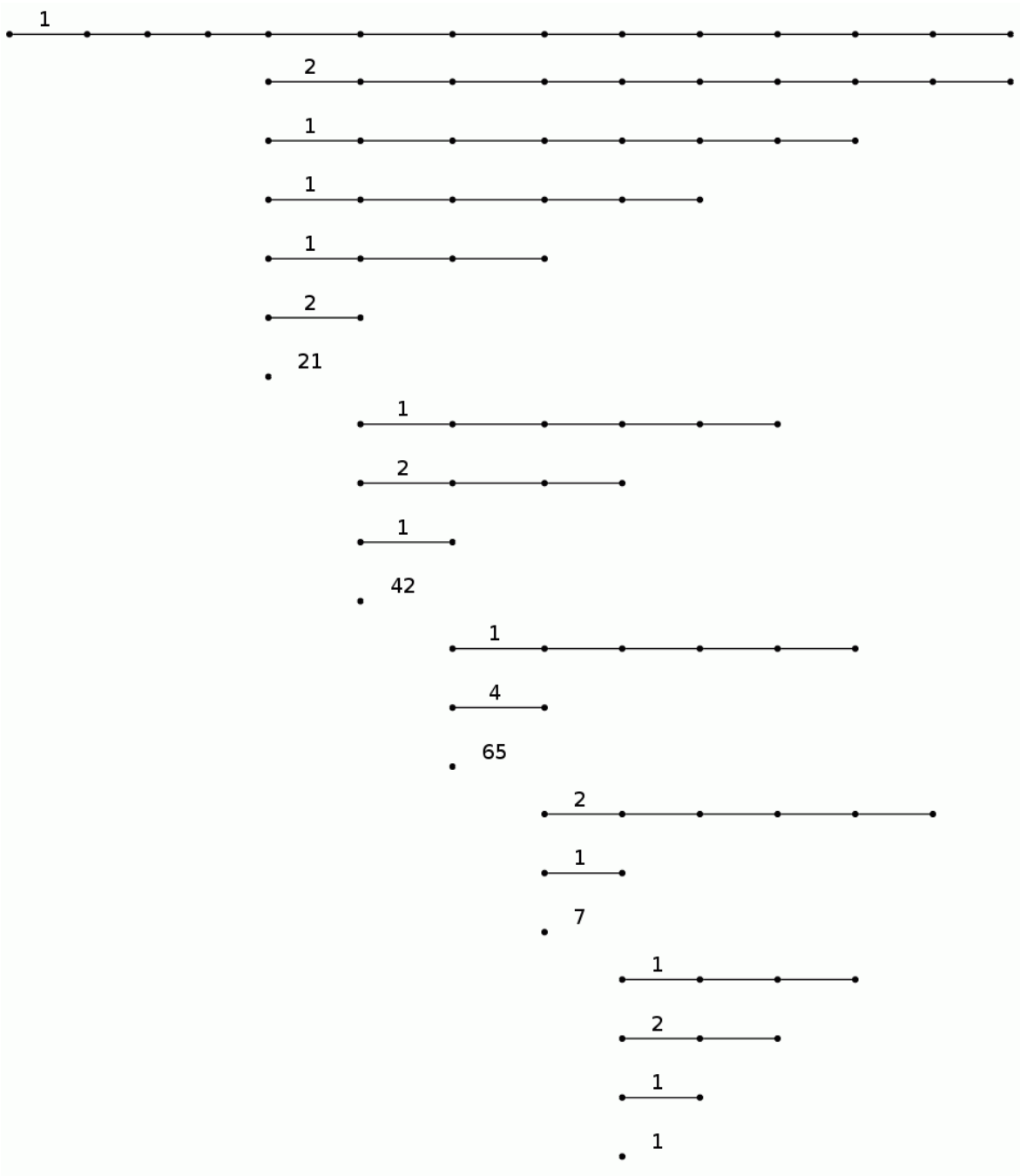
1V2X protein backbone





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

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

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

$$\pi_1(\mathbb{R}^3 \setminus K) \cong \langle x, y | xyx = yxy \rangle$$

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

$$\pi_1(\mathbb{R}^3 \setminus K) \cong \langle x, y | xyx = yxy \rangle$$

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

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

$$\pi_1(\mathbb{R}^3 \setminus K) \cong \langle x, y | xyx = yxy \rangle$$

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

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

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

$$\pi_1(\mathbb{R}^3 \setminus K) \cong \langle x, y | xyx = yxy \rangle$$

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

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

gap> FundamentalGroup(Y);
[ f1^-3*f2*f1^2*f2*f1, f1 ]
```

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