# Topological Data Analysis on Data With Non-Symmetric Distances

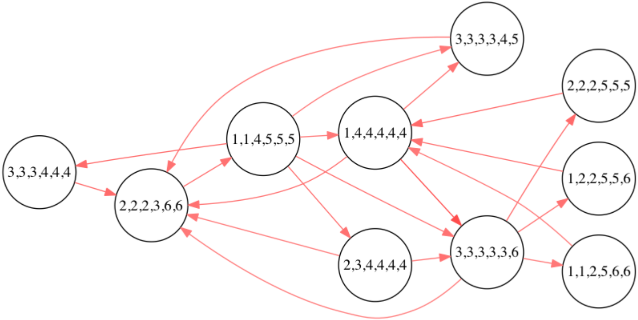
YTM 2015

Scott Balchin

# Motivation

# Motivation

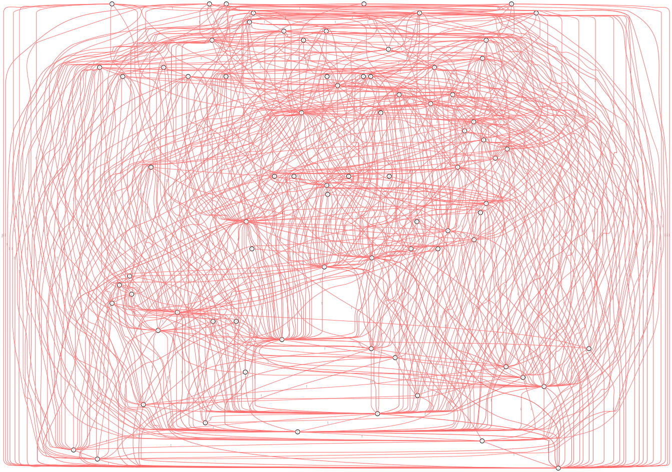
How can we study the structure of a space such as the following:



# Motivation

# Motivation

What about when we move a single dimension up?



# Motivation

Is there a way to study these spaces and the relation between them using topological data analysis?

# General Philosophy of Topological Data Analysis

# General Philosophy of Topological Data Analysis

- Point cloud of data in $\mathbb{R}^n$.
- Convert this point cloud into a family of topological spaces.
- Tools such as persistent homology.

# Vietoris-Rips Complex

# Vietoris-Rips Complex

**Definition (Vietoris-Rips Complex)**

Given a finite collection of points $\{x_\alpha\}$ in $\mathbb{R}^n$ endowed with some metric, the *Vietoris-Rips complex* $\mathcal{R}_\epsilon$ is the abstract simplicial complex whose $k$-simplices correspond to unordered $(k+1)$-tuples of points $\{x_\alpha\}_0^k$ that are pairwise within distance $\epsilon$.

# Vietoris-Rips Complex

## Definition (Vietoris-Rips Complex)

Given a finite collection of points $\{x_\alpha\}$ in $\mathbb{R}^n$ endowed with some metric, the *Vietoris-Rips complex* $\mathcal{R}_\epsilon$ is the abstract simplicial complex whose $k$-simplices correspond to unordered $(k+1)$-tuples of points $\{x_\alpha\}_0^k$ that are pairwise within distance $\epsilon$.

The Vietoris-Rips complex is a *flag complex*, this means that its structure is determined solely by its edge structure.

# Persistent Homology

# Persistent Homology

Note that for the Vietoris-Rips complexes constructed, we have inclusions $\mathcal{R}_\epsilon \subseteq \mathcal{R}_\delta$ for $\epsilon < \delta$.

# Persistent Homology

Note that for the Vietoris-Rips complexes constructed, we have inclusions $\mathcal{R}_\epsilon \subseteq \mathcal{R}_\delta$ for $\epsilon < \delta$.

**Definition (Persistent Homology)**

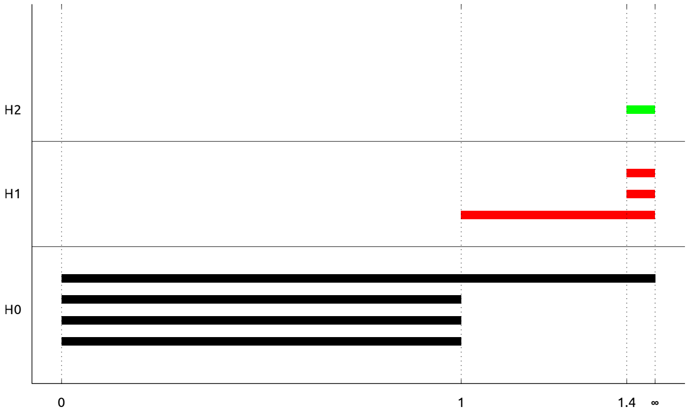
Denote by $\mathcal{R}$ the full collection of Rips complexes constructed for a point cloud of data. For $i < j$, the $(i, j)$-*persistent homology* of $\mathcal{R}$ is the image of the induced homomorphism in homology $H_*(\mathcal{R}^i) \to H_*(\mathcal{R}^j)$.

In short, persistent homology allows us to track when homology generators are born and die as we vary our parameter $\epsilon$.

# Barcodes

# Barcodes

We can display the persistent homology data as a *persistence diagram* or *barcode* to visualise the results.



# Outline of the Problem

# Outline of the Problem

- There are interesting data sets which don't come in the form of point cloud data.

# Outline of the Problem

- There are interesting data sets which don't come in the form of point cloud data.
- Not all data comes equipped with some canonical metric.

# Outline of the Problem

- There are interesting data sets which don't come in the form of point cloud data.
- Not all data comes equipped with some canonical metric.
- Some data cannot be equipped with a metric without throwing away some sort of information.

# Outline of the Problem

- There are interesting data sets which don't come in the form of point cloud data.
- Not all data comes equipped with some canonical metric.
- Some data cannot be equipped with a metric without throwing away some sort of information.
- Main motivation : Directed graphs.

# Outline of the Problem

- There are interesting data sets which don't come in the form of point cloud data.
- Not all data comes equipped with some canonical metric.
- Some data cannot be equipped with a metric without throwing away some sort of information.
- Main motivation : Directed graphs.
- How to construct a simplicial complex from a data set with a non-symmetric distance, which captures this non-symmetric features? (Joint work with Etienne Pillin)

# Non-Symmetric Complex Construction

# Non-Symmetric Complex Construction

From now on we will assume that $\mathcal{X}$ is a data set with some distance $d$ between all points $X$ and $Y$ (possibly $\infty$). Without loss of generality assume

$$d_u(X, Y) = d(X, Y) \geqslant d(Y, X) = d_l(X, Y)$$

and let the *disparity* $\delta_{X,Y} = d_u(X, Y) - d_l(X, Y)$.

# Non-Symmetric Complex Construction

From now on we will assume that $\mathcal{X}$ is a data set with some distance $d$ between all points $X$ and $Y$ (possibly $\infty$). Without loss of generality assume

$$d_u(X, Y) = d(X, Y) \geqslant d(Y, X) = d_l(X, Y)$$

and let the *disparity* $\delta_{X,Y} = d_u(X, Y) - d_l(X, Y)$.

Let $F(a, b) : \mathbb{R}^+ \times \mathbb{R}^+ \to \mathbb{R}^+$ be an increasing positive function in both variables, such that $F(a, 0) = 0$ for all $a$. For our purpose $a$ will be the dimension of a simplex and $b$ will take the values $\delta$.

# Non-Symmetric Complex Construction

# Non-Symmetric Complex Construction

## Definition (Non-Symmetric Complex With Respect To $F$)

The *non-symmetric simplex of $\mathcal{X}$ with respect to $F$* of distance $\epsilon$, $\mathcal{N}_\epsilon^F$, is constructed as follows:

# Non-Symmetric Complex Construction

## Definition (Non-Symmetric Complex With Respect To $F$)

The *non-symmetric simplex of $\mathcal{X}$ with respect to $F$* of distance $\epsilon$, $\mathcal{N}_\epsilon^F$, is constructed as follows:

- We add a 1-simplex between points $X_1$ and $X_2$ if $d_l(X_1, X_2) \leqslant \epsilon$.

# Non-Symmetric Complex Construction

**Definition (Non-Symmetric Complex With Respect To $F$)**

The *non-symmetric simplex of $\mathcal{X}$ with respect to $F$* of distance $\epsilon$, $\mathcal{N}_{\epsilon}^{F}$, is constructed as follows:

- We add a 1-simplex between points $X_1$ and $X_2$ if $d_l(X_1, X_2) \leqslant \epsilon$.
- We add a 2-simplex between points $X_1$, $X_2$ and $X_3$ if all $d_u(X_i, X_j) \leqslant \epsilon$.

# Non-Symmetric Complex Construction

## Definition (Non-Symmetric Complex With Respect To $F$)

The *non-symmetric simplex of $\mathcal{X}$ with respect to $F$* of distance $\epsilon$, $\mathcal{N}_\epsilon^F$, is constructed as follows:

- We add a 1-simplex between points $X_1$ and $X_2$ if $d_l(X_1, X_2) \leqslant \epsilon$.
- We add a 2-simplex between points $X_1$, $X_2$ and $X_3$ if all $d_u(X_i, X_j) \leqslant \epsilon$.
- We add a 3-simplex between points $X_1, \ldots, X_4$ if all $d_u(X_i, X_j) + F(1, \delta_{X_i, X_j}) \leqslant \epsilon$.

$$\vdots$$

# Non-Symmetric Complex Construction

## Definition (Non-Symmetric Complex With Respect To $F$)

The *non-symmetric simplex of $\mathcal{X}$ with respect to $F$* of distance $\epsilon$, $\mathcal{N}_\epsilon^F$, is constructed as follows:

- We add a 1-simplex between points $X_1$ and $X_2$ if $d_l(X_1, X_2) \leqslant \epsilon$.
- We add a 2-simplex between points $X_1$, $X_2$ and $X_3$ if all $d_u(X_i, X_j) \leqslant \epsilon$.
- We add a 3-simplex between points $X_1, \ldots, X_4$ if all $d_u(X_i, X_j) + F(1, \delta_{X_i,X_j}) \leqslant \epsilon$.

$$\vdots$$

- We add an $i$-simplex between points $X_1, \ldots, X_{i+1}$ if all $d_u(X_i, X_j) + F(i-2, \delta_{X_i,X_j}) \leqslant \epsilon$.

# Properties of the Non-Symmetric Complex

# Properties of the Non-Symmetric Complex

If $d$ is symmetric, we will retrieve the classical Rips complex as we will have $\delta_{X,Y} = 0$ for all $X$ and $Y$, and because we asked for $F(a, 0) = 0$.

The complex is constructed so that if there is a large disparity $\delta_{X,Y}$, then the higher dimensional complexes involving the points $X$ and $Y$ will not be filled in until $\epsilon$ large. Can find near-symmetric nodes using this.

In the case that $\delta_{X,Y} = \infty$, then there will only ever be 1-simplicies whenever $X$ and $Y$ are involved.

We still get inclusions $\mathcal{N}_{\epsilon}^{F} \subset \mathcal{N}_{\delta}^{F}$ for $\epsilon < \delta$, which means we can do persistent homology.

# Non-Symmetric Excess

# Non-Symmetric Excess

- In the Rips complex construction we finish when we reach $\epsilon$ being the maximum distance between two points.
- This is not the case in the non-symmetric complex.
- If all distances involved are finite then we will have an $\epsilon_{\mathsf{max}}$ which gives us a fully connected simplicial complex.

# Non-Symmetric Excess

- In the Rips complex construction we finish when we reach $\epsilon$ being the maximum distance between two points.
- This is not the case in the non-symmetric complex.
- If all distances involved are finite then we will have an $\epsilon_{\text{max}}$ which gives us a fully connected simplicial complex.

**Definition (Non-Symmetric Excess)**

Let $\delta_{\text{max}}$ be the maximum finite disparity between data points in $\mathcal{X}$. Then we define the *non-symmetric excess* $\mathcal{E}$ on a non-symmetric complex with respect to $F$ to be

$$\mathcal{E} = F(|\mathcal{X}| - 2, \delta_{\text{max}})$$

Where $|\mathcal{X}|$ is the number of data points.

# Computational Downfall

# Computational Downfall

The complex that we construct is no longer a flag complex

# Computational Downfall

The complex that we construct is no longer a flag complex

One way to overcome this would be truncating the construction, and allowing the higher dimensional simplicies be defined by the structure of the $i$ simplicies. This would be an *i-flag complex*.

# A Specific Case of $F$

# A Specific Case of $F$

We set $F(a, b) = ab$, which is increasing in the domain of where we will be using it, and $F(a, 0) = a \cdot 0 = 0$, so is a valid such function.

# A Specific Case of $F$

We set $F(a, b) = ab$, which is increasing in the domain of where we will be using it, and $F(a, 0) = a \cdot 0 = 0$, so is a valid such function.

This means that the $i$-simplicies are formed between $i + 1$-tuples of points where we have $d_u(X_i, X_j) + (i - 2)\delta_{X_i,X_j} \leqslant \epsilon$ for all pairs.

# A Specific Case of $F$

We set $F(a, b) = ab$, which is increasing in the domain of where we will be using it, and $F(a, 0) = a \cdot 0 = 0$, so is a valid such function.

This means that the $i$-simplicies are formed between $i + 1$-tuples of points where we have $d_u(X_i, X_j) + (i - 2)\delta_{X_i, X_j} \leqslant \epsilon$ for all pairs.

Other possible $F$ include:

- $F(a, b) = a^n b^m$ where $m, n \geqslant 1$
- $F(a, b) = b^a$
- $F(a, b) = a^b - 1$

# Social Network Analysis

# Social Network Analysis

- "Social network analysis (SNA) is a strategy for investigating social structures through the use of network and graph theories."

# Social Network Analysis

- "Social network analysis (SNA) is a strategy for investigating social structures through the use of network and graph theories."
- It is a tool being adapted to sociology, anthropology, psychology, management, health, defence, etc.

# Social Network Analysis

- "Social network analysis (SNA) is a strategy for investigating social structures through the use of network and graph theories."
- It is a tool being adapted to sociology, anthropology, psychology, management, health, defence, etc.
- "It characterizes networked structures in terms of nodes (individual actors, people, or things within the network) and the ties or edges (relationships or interactions) that connect them."

# Directed Graphs via Twitter

# Directed Graphs via Twitter

- Twitter has a naturally non-symmetric relation built in with its follower feature.
- We can represent this as a directed graph.
- We then can consider the shortest directed path between two people and let the distance between them be the length of this path.
- If no such path exists then we say the distance between the two people is $\infty$.

# Twitter's API

# Twitter's API

- Twitter has a very convenient and practical API (Application Program Interface).
- This means we can actually create the graphs that we described with real data.
- We can start from an initial seed person and build their network (with some truncation).

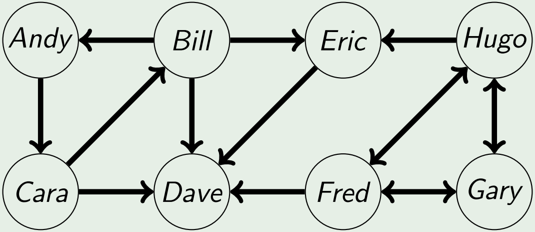
# A Toy Example

# A Toy Example

The following made-up example highlights some of the key features one might wish to identify in a social network.

# A Toy Example

The following made-up example highlights some of the key features one might wish to identify in a social network.



# A Toy Example

# A Toy Example

The distance matrix with respect to the shortest path for this directed graph reads as below, where the entry $(i, j)$ is the distance from $i$ to $j$.

$$
\begin{array}{c} \\ A \\ B \\ C \\ D \\ E \\ F \\ G \\ H \end{array}
\begin{array}{c}
\begin{array}{cccccccc} A & B & C & D & E & F & G & H \end{array} \\
\begin{pmatrix}
0 & 2 & 1 & 2 & 3 & \infty & \infty & \infty \\
1 & 0 & 2 & 1 & 1 & \infty & \infty & \infty \\
2 & 1 & 0 & 1 & 2 & \infty & \infty & \infty \\
\infty & \infty & \infty & 0 & \infty & \infty & \infty & \infty \\
\infty & \infty & \infty & 1 & 0 & \infty & \infty & \infty \\
\infty & \infty & \infty & 1 & 2 & 0 & 1 & 1 \\
\infty & \infty & \infty & 2 & 2 & 1 & 0 & 1 \\
\infty & \infty & \infty & 2 & 1 & 1 & 1 & 0
\end{pmatrix}
\end{array}
$$

# A Toy Example

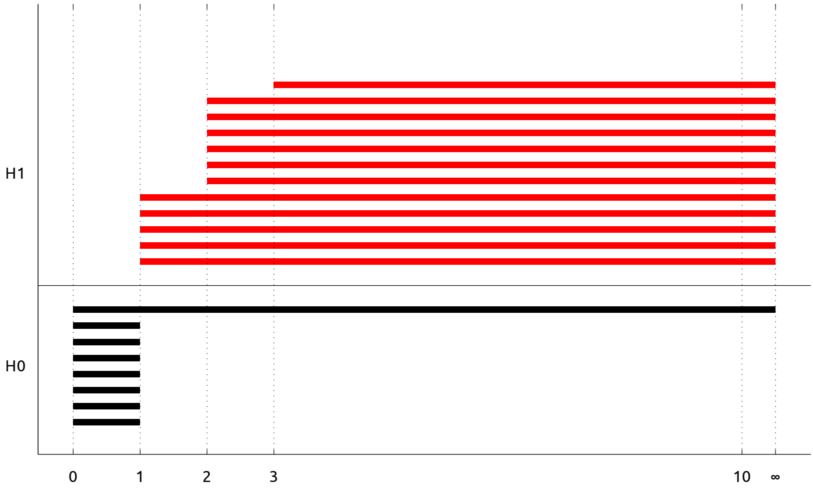
The distance matrix with respect to the shortest path for this directed graph reads as below, where the entry $(i, j)$ is the distance from $i$ to $j$.

$$
\begin{array}{c} \\ A \\ B \\ C \\ D \\ E \\ F \\ G \\ H \end{array}
\begin{array}{c}
\begin{array}{cccccccc} A & B & C & D & E & F & G & H \end{array} \\
\begin{pmatrix}
0 & 2 & 1 & 2 & 3 & \infty & \infty & \infty \\
1 & 0 & 2 & 1 & 1 & \infty & \infty & \infty \\
2 & 1 & 0 & 1 & 2 & \infty & \infty & \infty \\
\infty & \infty & \infty & 0 & \infty & \infty & \infty & \infty \\
\infty & \infty & \infty & 1 & 0 & \infty & \infty & \infty \\
\infty & \infty & \infty & 1 & 2 & 0 & 1 & 1 \\
\infty & \infty & \infty & 2 & 2 & 1 & 0 & 1 \\
\infty & \infty & \infty & 2 & 1 & 1 & 1 & 0
\end{pmatrix}
\end{array}
$$

The only non-zero and non-infinite distance disparities are $\delta(A, B), \delta(B, C)$ and $\delta(A, C)$ which are all equal 1.

# A Toy Example

# A Toy Example



# A Toy Example

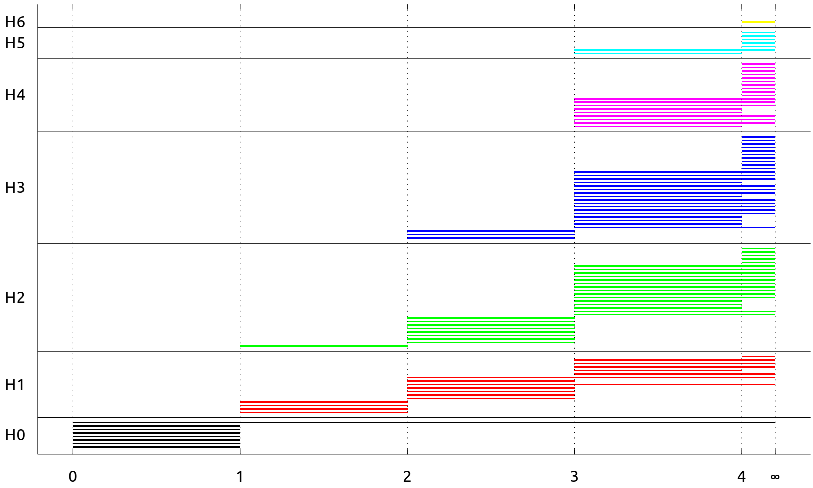
# A Toy Example

Compare this with the matrix we would get by disregarding the directions.

$$\begin{array}{c} \\ A \\ B \\ C \\ D \\ E \\ F \\ G \\ H \end{array} \begin{array}{c} \begin{array}{cccccccc} A & B & C & D & E & F & G & H \end{array} \\ \begin{pmatrix} 0 & 1 & 1 & 1 & 2 & 3 & 4 & 3 \\ 1 & 0 & 1 & 1 & 1 & 2 & 3 & 2 \\ 1 & 1 & 0 & 1 & 2 & 2 & 3 & 3 \\ 1 & 1 & 1 & 0 & 1 & 1 & 2 & 2 \\ 2 & 1 & 2 & 1 & 0 & 2 & 2 & 1 \\ 3 & 2 & 2 & 1 & 2 & 0 & 1 & 1 \\ 4 & 3 & 3 & 2 & 2 & 1 & 0 & 1 \\ 3 & 2 & 3 & 2 & 1 & 1 & 1 & 0 \end{pmatrix} \end{array}$$

# A Toy Example

# A Toy Example



# A Brief Introduction

## Definition (Non-Transitive Dice)

# A Brief Introduction

## Definition (Non-Transitive Dice)

- An $n$-side dice is an $n$-tuple $X = [d_1, \dots, d_n]$, $d_i \in [1, n]$.

# A Brief Introduction

**Definition (Non-Transitive Dice)**

- An $n$-side dice is an $n$-tuple $X = [d_1, \ldots, d_n]$, $d_i \in [1, n]$.
- A dice $X$ *beats* dice $Y$ if $\mathbb{P}(X > Y) > \frac{1}{2}$, we denote this $X \gg Y$.

# A Brief Introduction

**Definition (Non-Transitive Dice)**

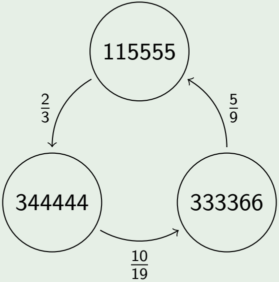
- An $n$-side dice is an $n$-tuple $X = [d_1, \ldots, d_n]$, $d_i \in [1, n]$.
- A dice $X$ *beats* dice $Y$ if $\mathbb{P}(X > Y) > \frac{1}{2}$, we denote this $X \gg Y$.
- A *cycle of length $r$ of non-transitive dice* is an ordered collection of dice $(X_1, \ldots, X_r)$ such that:
  1. $X_i \gg X_{i+i}$ $\forall 1 \leqslant i \leqslant r - 1$
  2. $X_r \gg X_1$

# A Brief Introduction

## Definition (Non-Transitive Dice)

- An $n$-side dice is an $n$-tuple $X = [d_1, \dots, d_n]$, $d_i \in [1, n]$.
- A dice $X$ *beats* dice $Y$ if $\mathbb{P}(X > Y) > \frac{1}{2}$, we denote this $X \ggg Y$.
- A *cycle of length r of non-transitive dice* is an ordered collection of dice $(X_1, \dots, X_r)$ such that:
  1. $X_i \ggg X_{i+i}$ $\forall 1 \leqslant i \leqslant r - 1$
  2. $X_r \ggg X_1$
- A dice $X$ is *triangular* if $d_1 + \cdots + d_n = T(n)$.

# A Brief Introduction

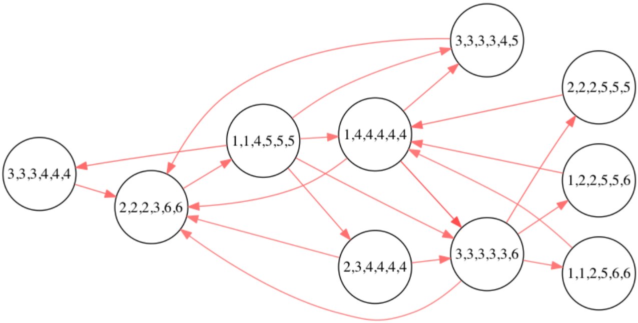


# Directed Graphs from Dice

# Directed Graphs from Dice

- Take some dice set $\mathcal{D}$.
- Consider all dice which appear in a non-transitive cycle.
- Plot with the dice being nodes, and the non-transitive relations being the directed edges.
- These graphs are always strongly connected.

# Triangular 6-Sided Dice



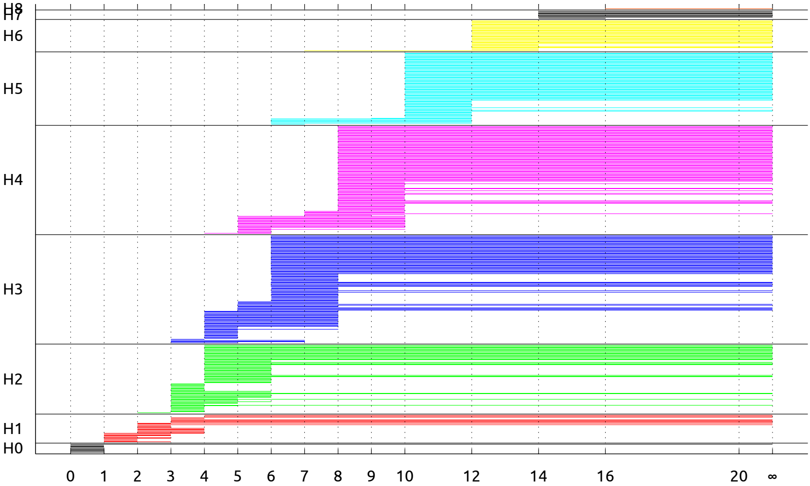
# Without Disregarding Distances

$$\begin{pmatrix} 0 & 3 & 3 & 3 & 4 & 2 & 4 & 3 & 1 & 4 \\ 3 & 0 & 3 & 3 & 4 & 2 & 4 & 3 & 1 & 4 \\ 3 & 3 & 0 & 3 & 1 & 2 & 1 & 2 & 1 & 1 \\ 3 & 3 & 1 & 0 & 2 & 2 & 2 & 3 & 1 & 2 \\ 4 & 2 & 2 & 4 & 0 & 3 & 3 & 1 & 2 & 3 \\ 1 & 1 & 1 & 1 & 2 & 0 & 2 & 1 & 2 & 2 \\ 4 & 2 & 2 & 4 & 3 & 3 & 0 & 1 & 2 & 3 \\ 3 & 1 & 1 & 3 & 2 & 2 & 2 & 0 & 1 & 2 \\ 2 & 2 & 2 & 2 & 3 & 1 & 3 & 2 & 0 & 3 \\ 4 & 2 & 2 & 4 & 3 & 3 & 3 & 1 & 2 & 0 \end{pmatrix}$$

# Without Disregarding Distances

- In this case we have no infinite distances.
- $\delta_{\text{max}} = 2$
- $\mathcal{E} = F(10 - 2, 2) = 8 \times 2 = 16$
- Therefore we should expect $\epsilon_{\text{max}} = 4 + 16 = 20$

# Without Disregarding Distances



# Required Development

# Required Development

- What can we ascertain from the results, what are they actually telling us?

# Required Development

- What can we ascertain from the results, what are they actually telling us?
- Is there a complex construction which approximates the non-symmetric one homotopically, but is easier to compute?

# Required Development

- What can we ascertain from the results, what are they actually telling us?
- Is there a complex construction which approximates the non-symmetric one homotopically, but is easier to compute?
- How does our choice of the function $F$ affect the results?

# Required Development

- What can we ascertain from the results, what are they actually telling us?
- Is there a complex construction which approximates the non-symmetric one homotopically, but is easier to compute?
- How does our choice of the function $F$ affect the results?
- Code improvements - Bug squashing, parallelising.

# Required Development

- What can we ascertain from the results, what are they actually telling us?
- Is there a complex construction which approximates the non-symmetric one homotopically, but is easier to compute?
- How does our choice of the function $F$ affect the results?
- Code improvements - Bug squashing, parallelising.