Graph Based Analysis for Gene Segment Interactions In a Scrambled Genome

Masahico Saito
University of South Florida

BIRS, May 7 – 12, 2017

Coauthors: Mustafa Hajij, Nataša Jonoska, Denys Kukushkin
Collaborator: Laura Landweber, Columbia U.

(Supported in part by NIH R01GM109459-01)

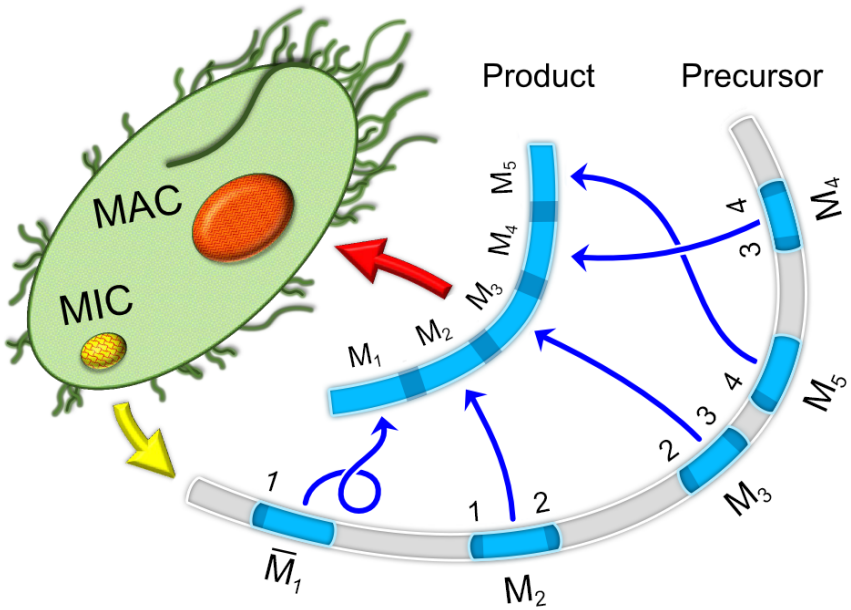
Plan:

1. Representing interactions of gene segments by graphs
2. Examples of graphs thus obtained
3. Applying TDA to a data set consisting of graphs through graph properties
4. Outputs

Outline

Cilliates

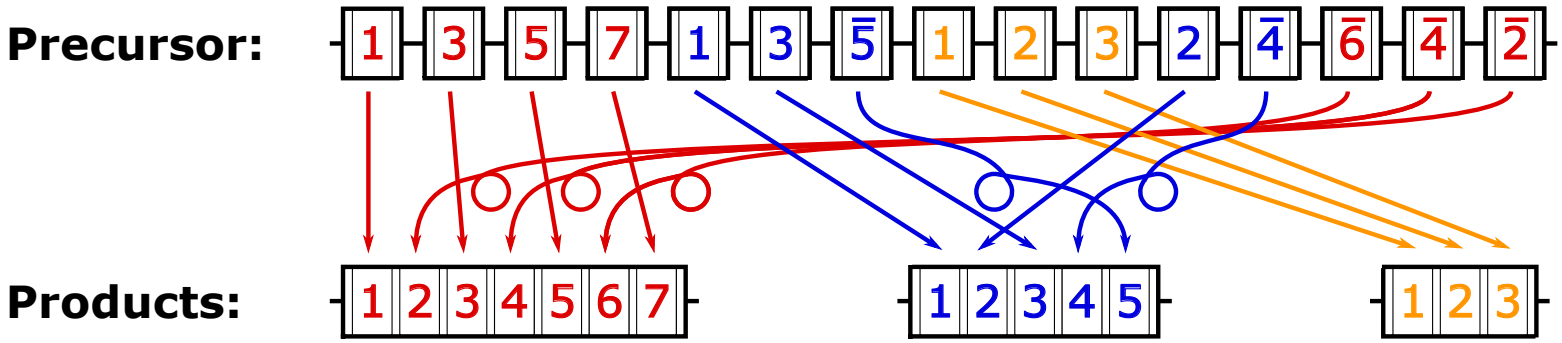
Oxytricha trifallax, a species of ciliate, undergoes massive genome rearrangements.
Landweber Lab sequenced the whole genome of *Oxytricha trifallax*.



Micronucleus (MIC) : used for mating/conjugation
Macronucleus (MAC) : functional
Gene assembly : from MIC to MAC
MDS (MAC Destined Sequence) : segments headed to MAC
IES (Internally Eliminated Sequence) : junk DNA, eliminated

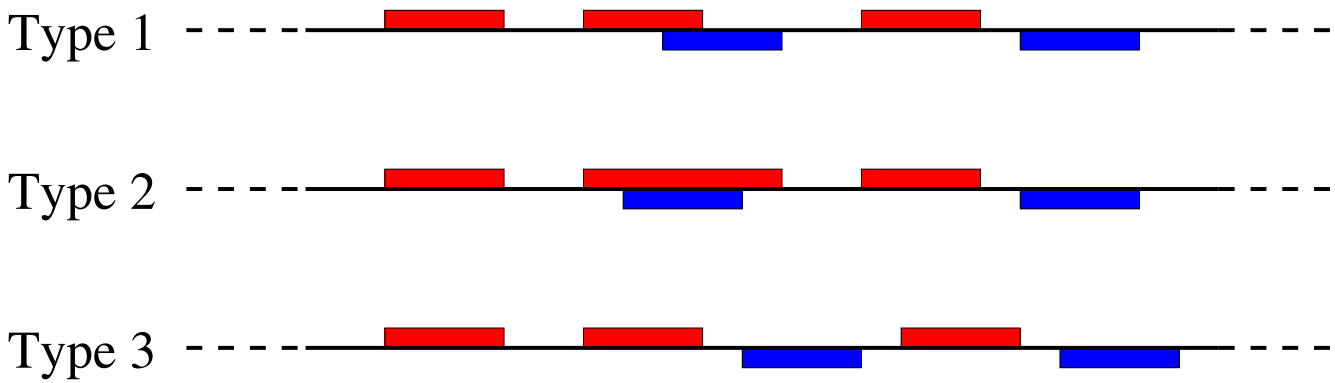
Rearrangement of gene segments in *Oxytricha trifallax*

MIC to MAC:



MDSs from different MAC genes interleave in the MIC contig.

Interaction types of gene segments



Three types of interactions of gene segments:
Type 1: Overlapping
Type 2 : Containment
Type 3 : Interleaving

Seven possible combination types of interaction: Overlapping and interleaving, etc.

Representing gene interactions by graphs

H : a MIC "contig" $\Longleftrightarrow$ $G = G(H) = (V, E)$ a graph

g : a MAC gene $\Longleftrightarrow$ vertex $v(g)$

A (colored and directed) edge e from $v(g_1)$ to $v(g_2)$

$\Longleftrightarrow$ g_1 interacts with g_2

Eg. g_1 is contained in g_2 $\Longleftrightarrow$ an edge from $v(g_1)$ to $v(g_2)$

Colors of edges represent a combination type of interactions
(7 colors)

(About 70 MIC chromosomes, about 283 isomorphism classes of graphs, about 16,000 MAC chromosomes.)

Outline

Examples

Sample graphs in the largest cluster

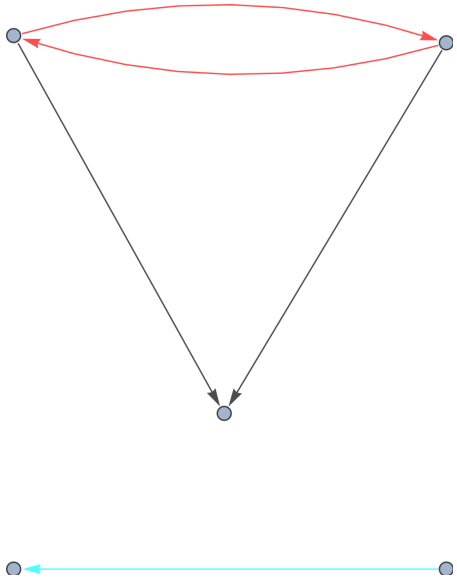


Figure : ctg7180000087289

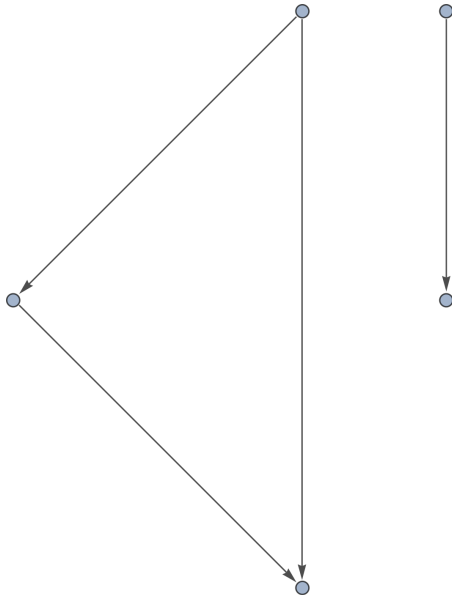


Figure : ctg7180000069936

Examples

Sample graphs in the largest cluster (cont.)

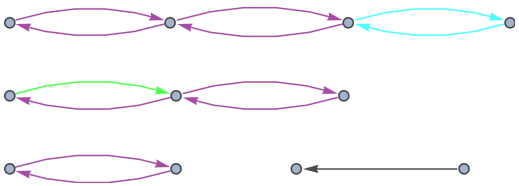


Figure :
ctg7180000087650

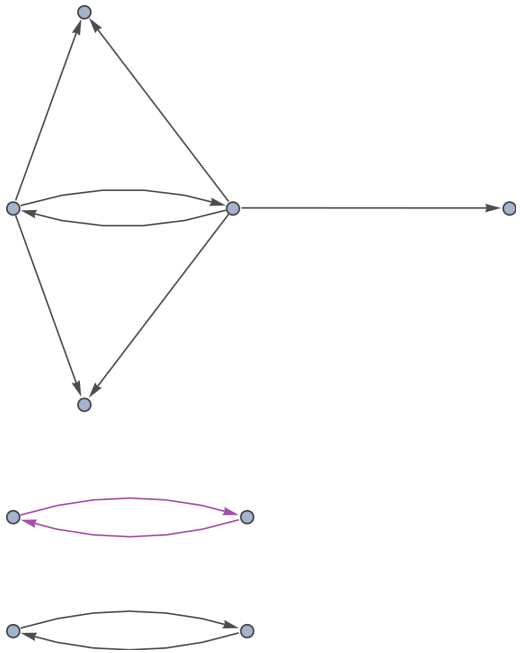


Figure :
ctg7180000069209

Examples

Two in a small component

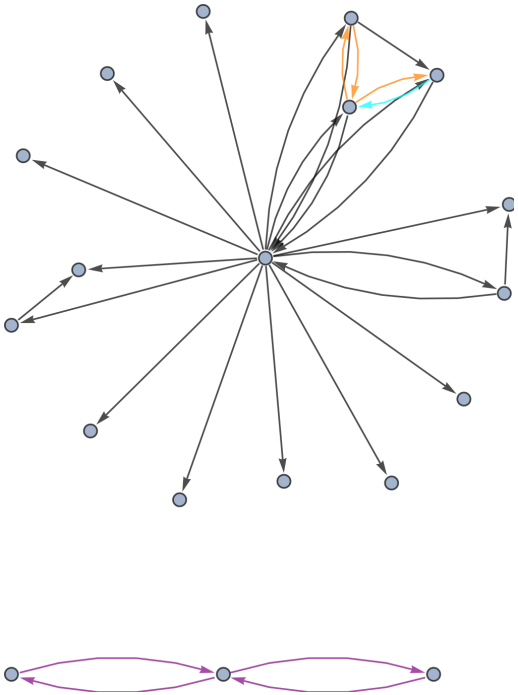


Figure : ctg7180000088096

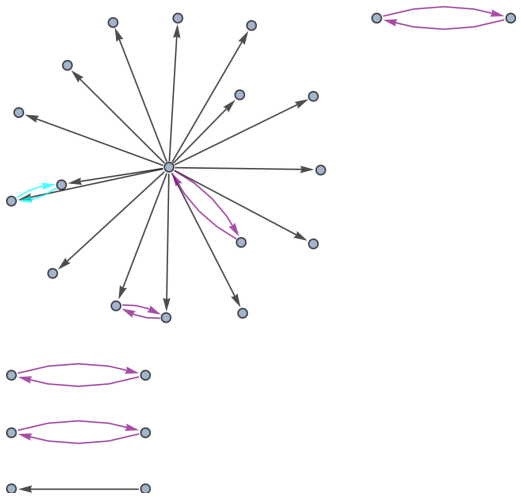


Figure : ctg7180000067742

Examples

Singleton components

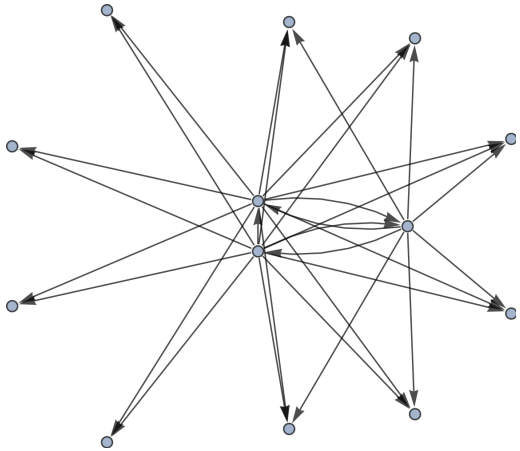


Figure : ctg7180000067761 joins the large component at $d = 11$

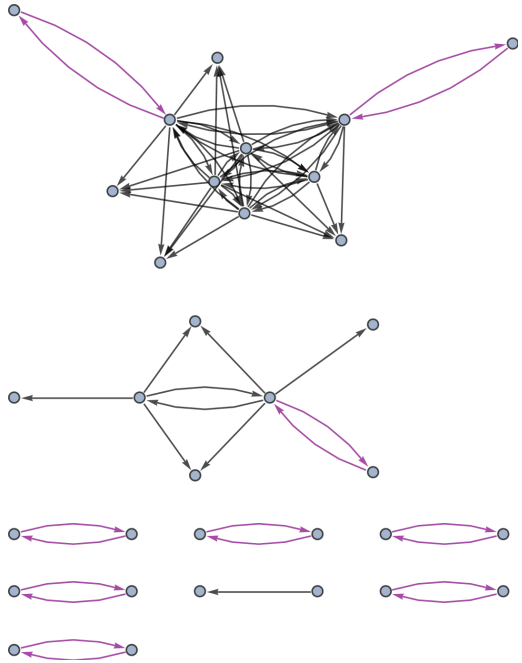
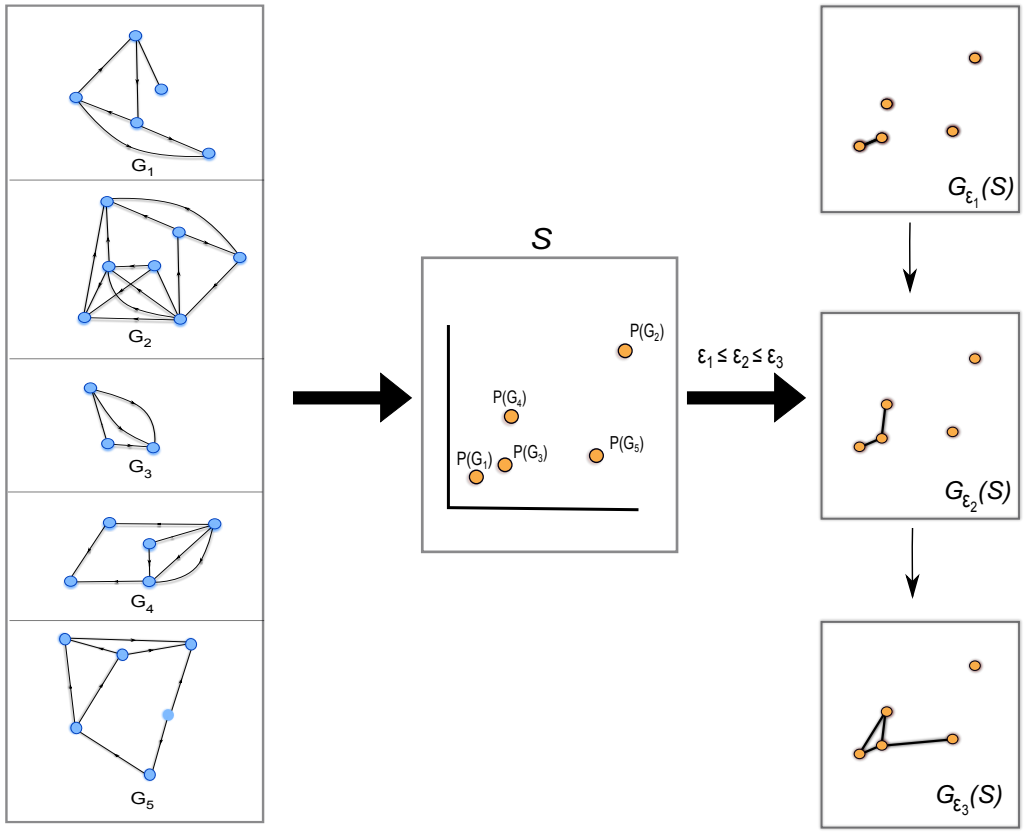


Figure : ctg7180000067223 joins the large comonent last at $d = 24$

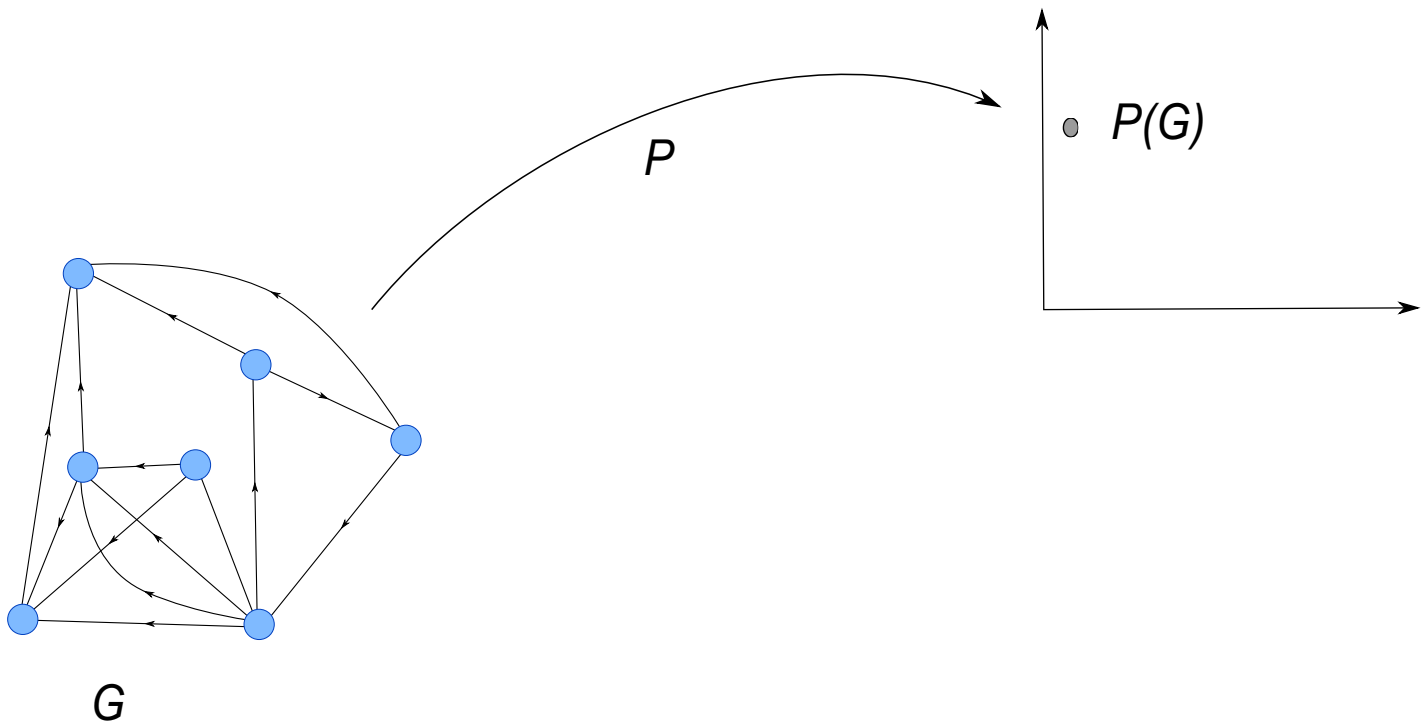
Outline

From graphs to point-cloud to filtration



Left : the set of graphs that represents the contigs.
Middle : Represent graphs as points in a c Euclidean space.
Right : Construct a filtration on the point-cloud.

From graphs to point-cloud



Associate to every graph G a feature vector $P(G)$, a point in a Euclidean space, that represents the graph G.

From graphs to point-cloud

The vector $P(G)$ is defined as follows.

Global Features Vector: $P_g(G) =< V(G), E(G), CN(G) >$
$V(G)$: # of vertices, $E(G)$ # of edges $P_g(G)$
$CN(G)$: the size of the largest clique in G.

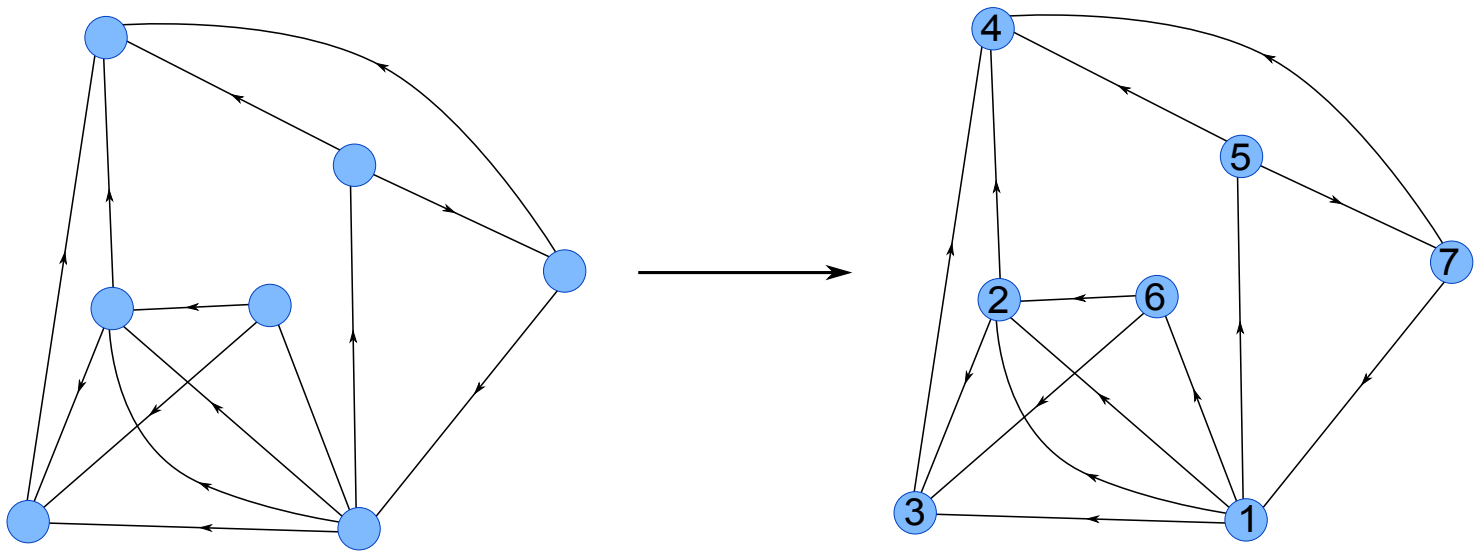
Valence Features Vector: $P_v(G)$: the valency of the vertices ordered decreasingly.

The Clique Vector: $P_c(G)$: # of cliques containing the vertex, in the same order of vertices of $P_v(G)$.

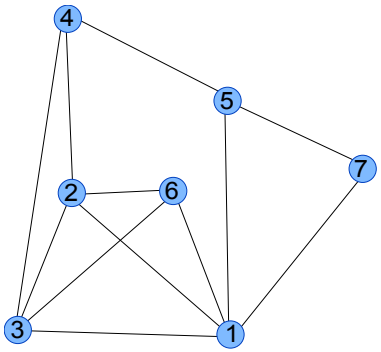
$d = \max(\text{valency})$

Concatenate 0s if $|P_v(G)| < d$

$P(G) \in \mathbb{R}^{2d+3}$: concatenation of $P_g(G), P_v(G), P_c(G)$.

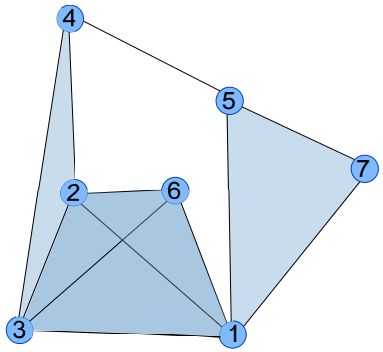


$$P_v(G) =< 6, 5, 4, 4, 3, 3, 3 >$$

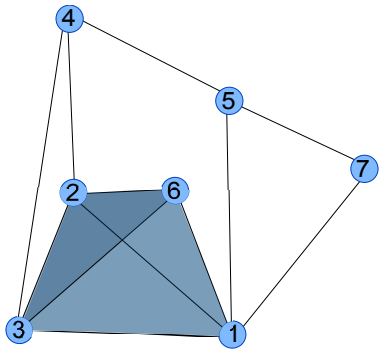


(1,1,1,1,1,1,1)

(5,4,4,3,3,3,2)



(4,3,3,1,1,3,1)

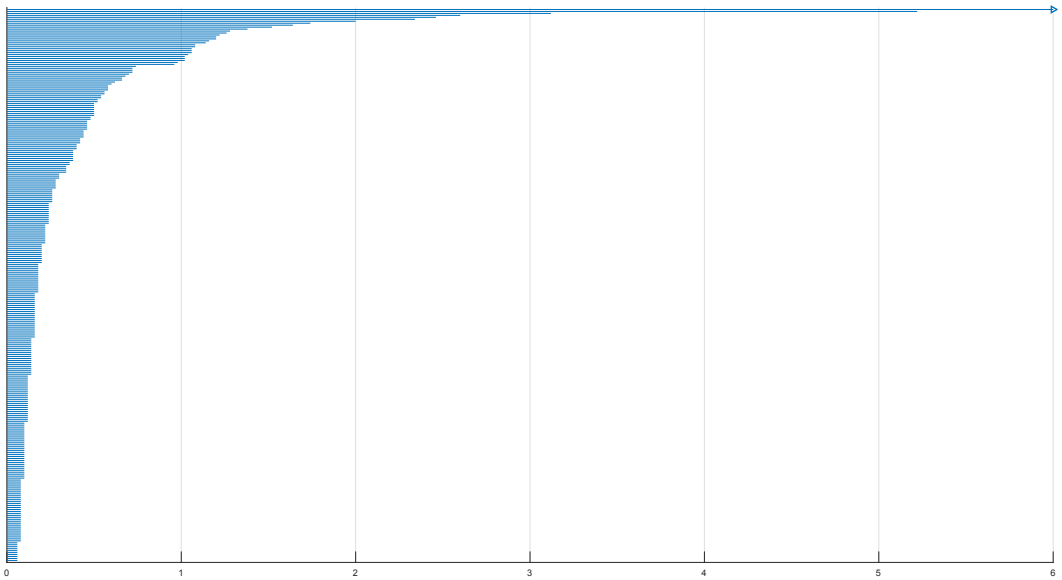


(1,1,1,0,0,1,0)

$$P_c(G) =< 11, 9, 9, 5, 5, 8, 4 >$$

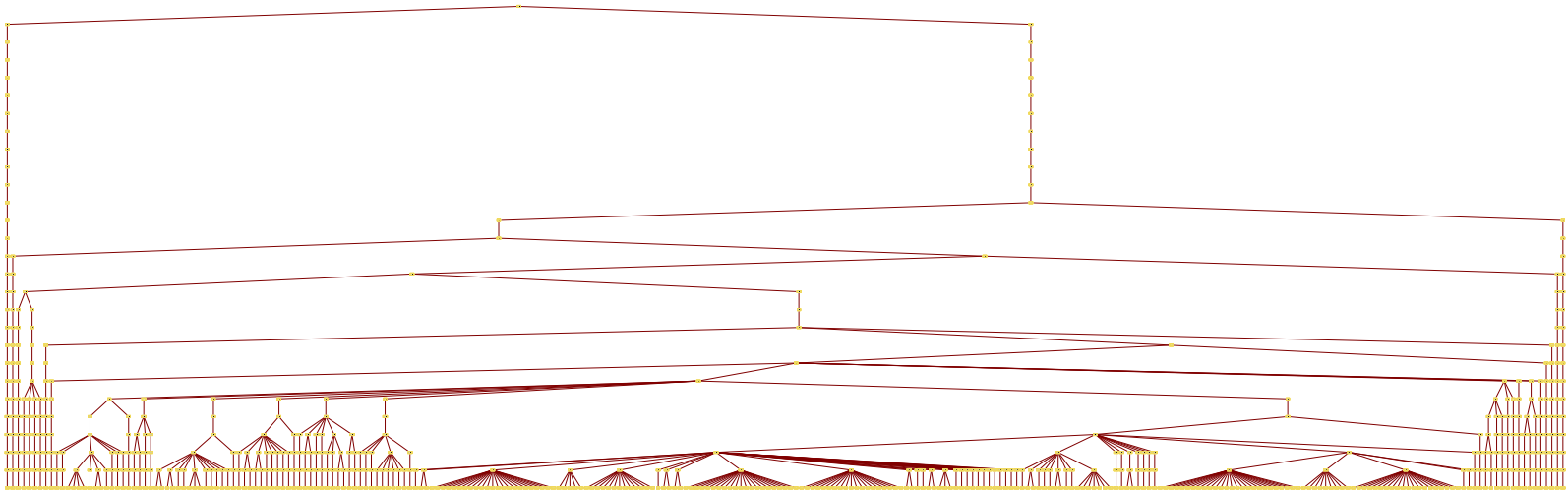
Outline

Barcodes of connected components



The barcode diagram describing the birth and death of the connected components.

Tree diagram of merging components



Tree diagram representing merging components

Thank you !