A statistical modeling framework for analyzing tree-indexed data

Application to plant development at microscopic and macroscopic scales

Pierre Fernique[1,2]

*S*upervised by Yann Guédon[2] & Jean-Baptiste Durand[3]

[1]**Université Montpellier 2**, I3M

[2]**Cirad**, UMR AGAP & **Inria**, Virtual Plants

[3]**Université Grenoble Alpes**, LJK & **Inria**, Mistis

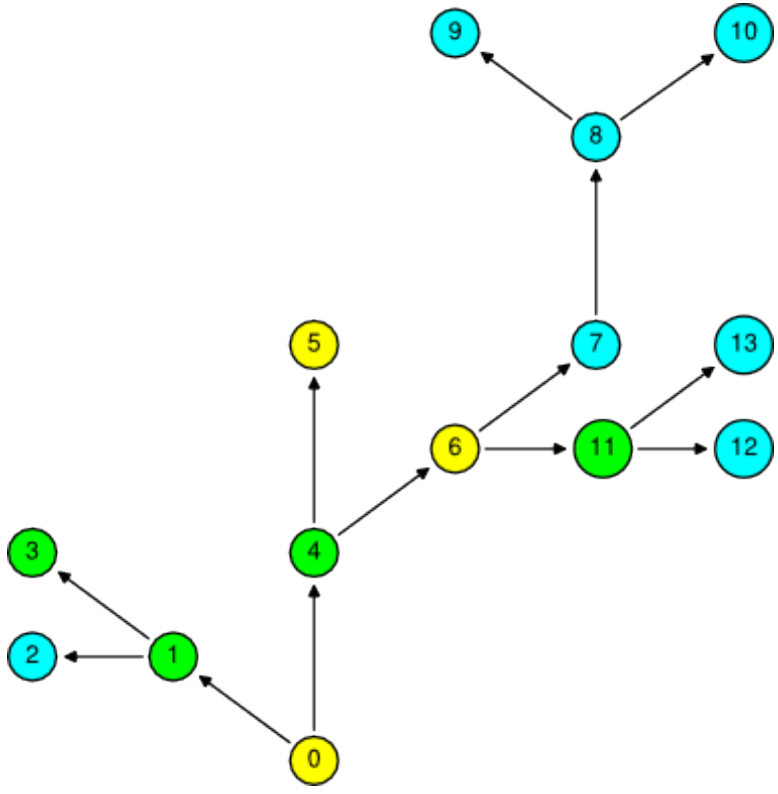
December 10, 2014

Introduction

¬ Tree-indexed data definition

$\mathcal{T} \subset \mathbb{N}$ is the vertex set,
$\mathcal{T} = \{0, \cdots, 13\}$
$\bar{x} = (x_t)_{t \in \mathcal{T}}$ is the data set,
$\bar{x} = (0, 1, 2, 1, 1, 0, \cdots, 2)$
$\bar{n} = (n_t)_{t \in \mathcal{T}}$ is the number of children set,
$\bar{n} = (2, 2, 0, 0, 2, 0, \cdots, 0)$
Problems:

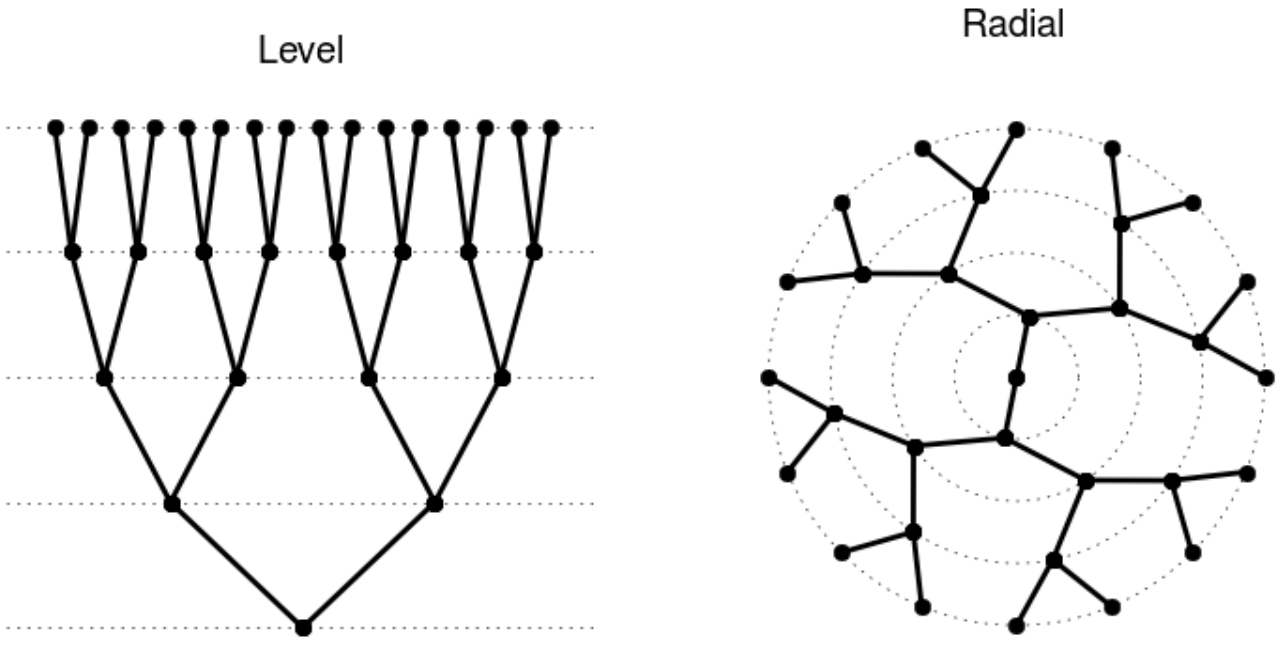
- Motif detection,
- Homogeneous zone detection.



Tree-indexed data representation

Introduction

¬ Tree-indexed data definition



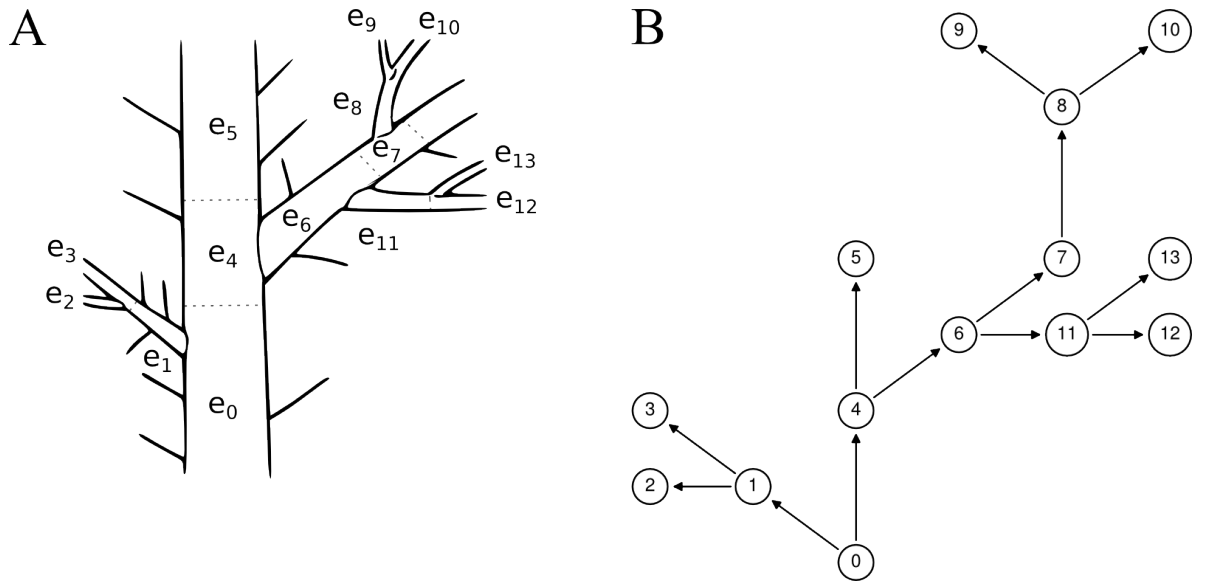
Alternative tree-indexed data representation

Introduction

¬ Tree-indexed data examples

Virtual Plants focuses on plant development and its modulation by environmental and genetic factors:

1. At a macroscopic scale. Each vertex represents a botanical entity and edges encode either the temporal precedence of two botanical entities.



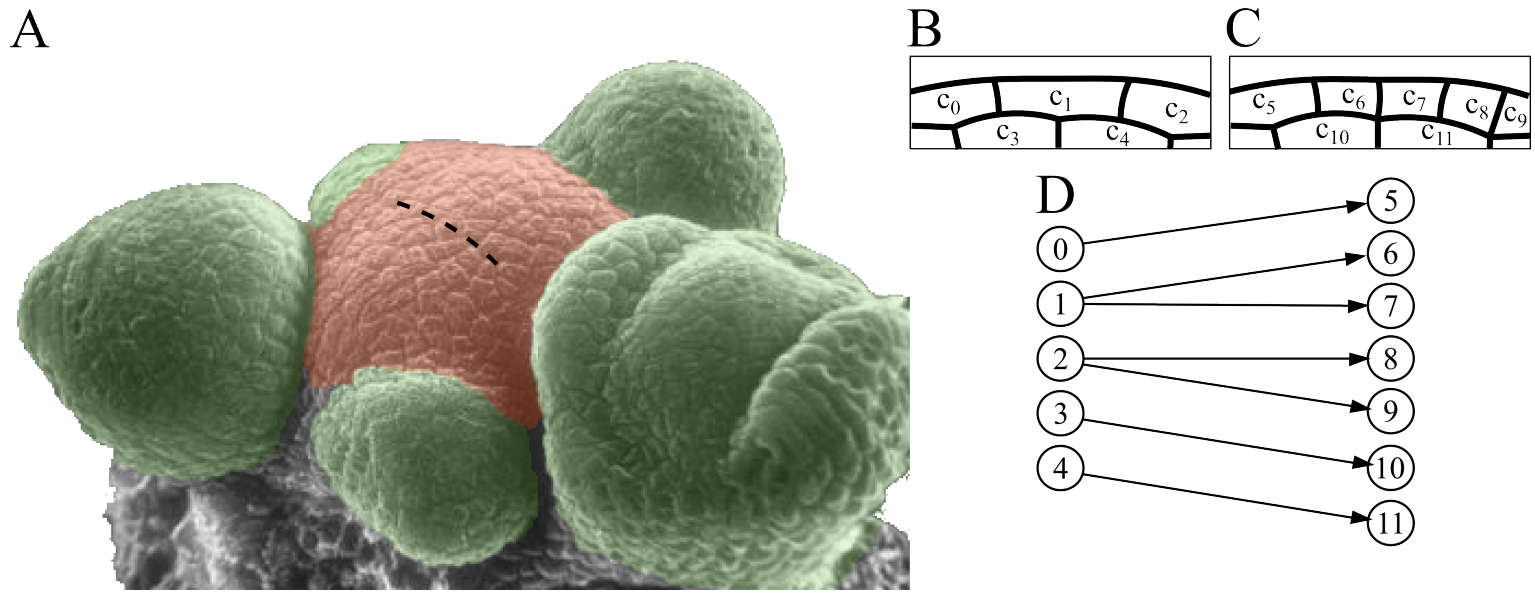
Tree-indexed data extraction from whole plants

Introduction

¬ Tree-indexed data examples

Virtual Plants focuses on plant development and its modulation by environmental and genetic factors:

2. At a microscopic scale. Each vertex represents a cell and edges encode either the tracking of a cell throughout time or the lineage relationships between parent and child cells.



Tree-indexed data extraction from cell lineages

Introduction

¬ Focus on the application at macroscopic scale

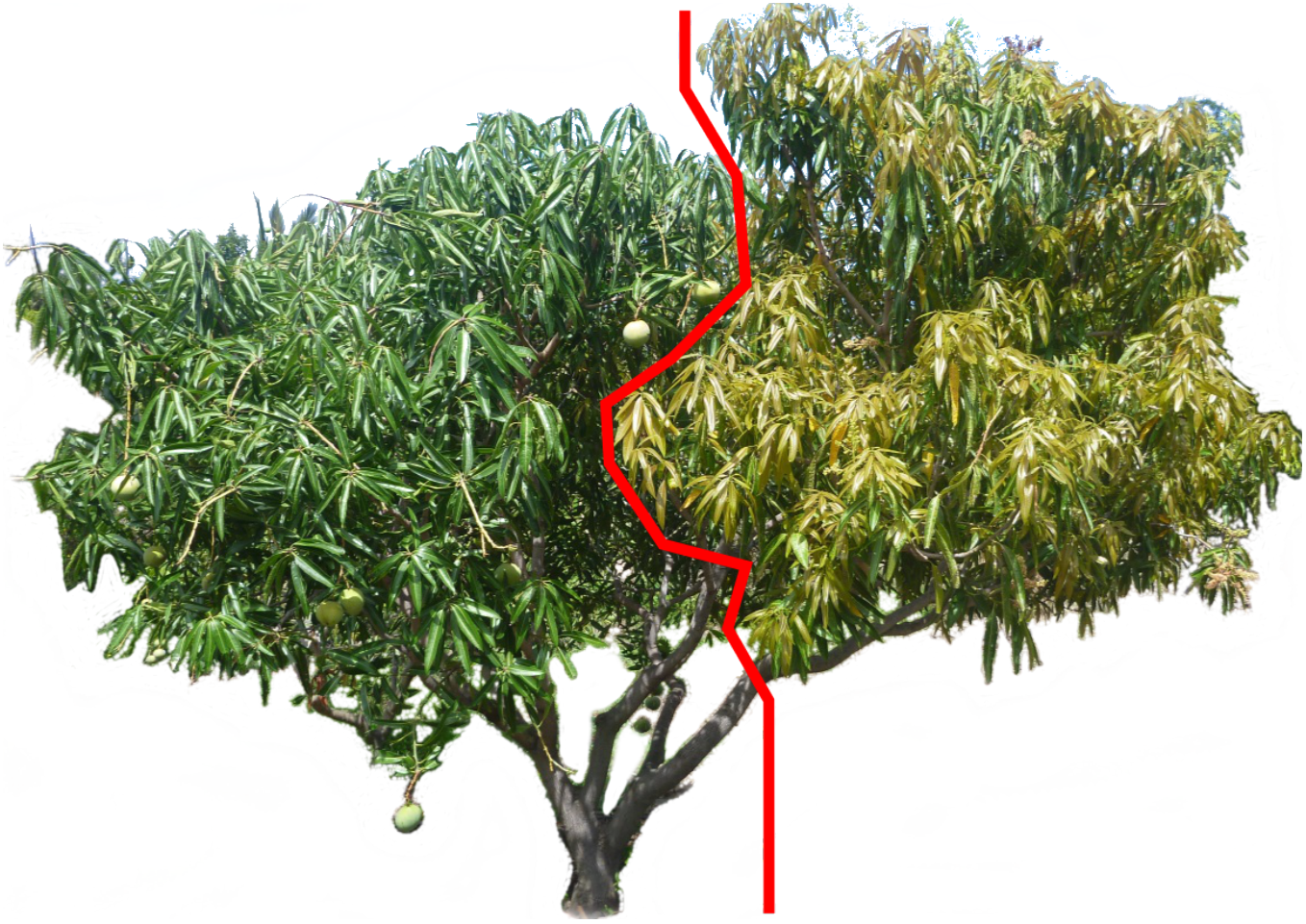
This presentation focuses on mango tree application

A mango tree [Dambreville, 2012]

Introduction

¬ Focus on the application at macroscopic scale

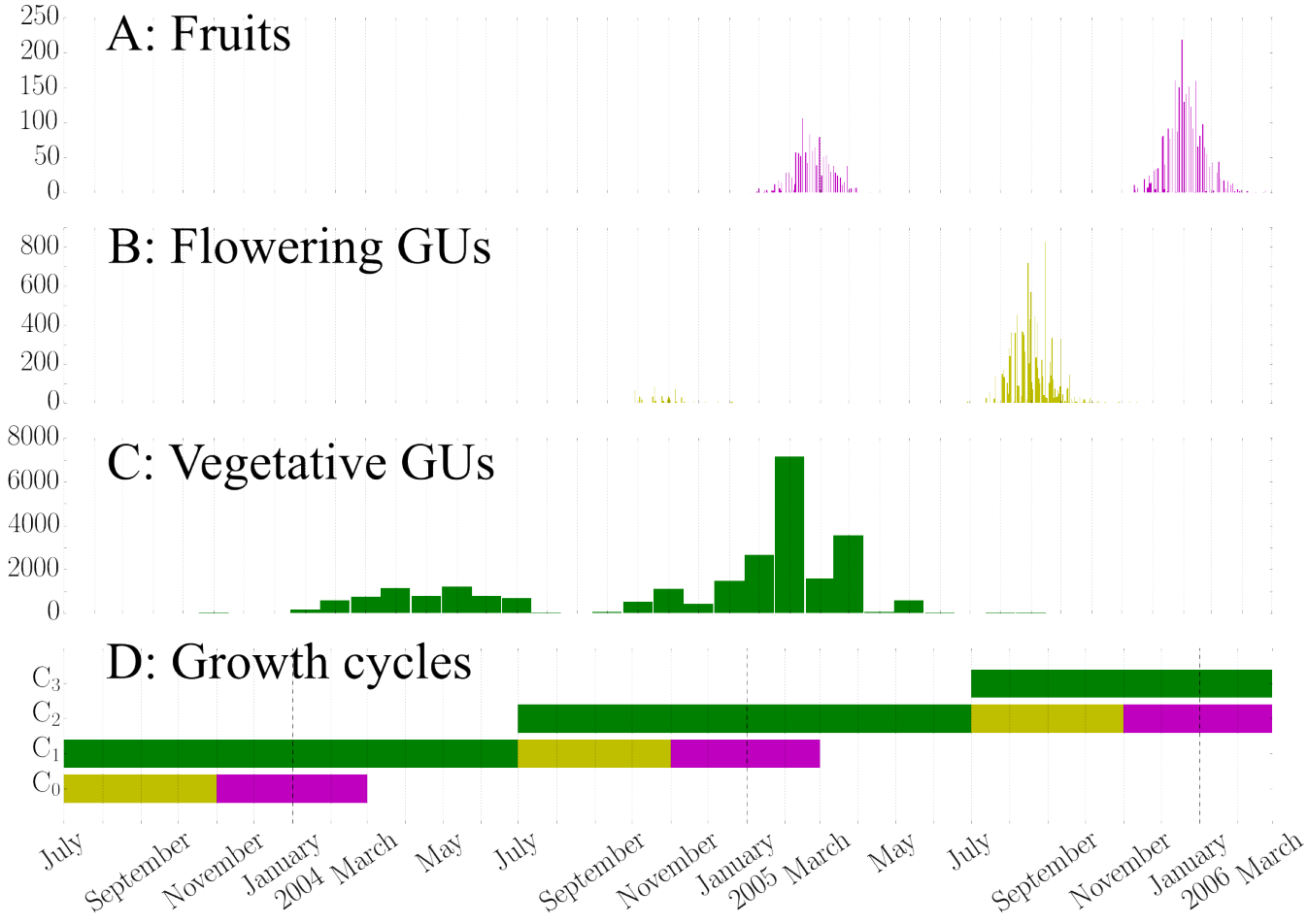
This presentation focuses on mango tree application



A mango tree [Dambreville, 2012]

Introduction

¬ Focus on the application at macroscopic scale

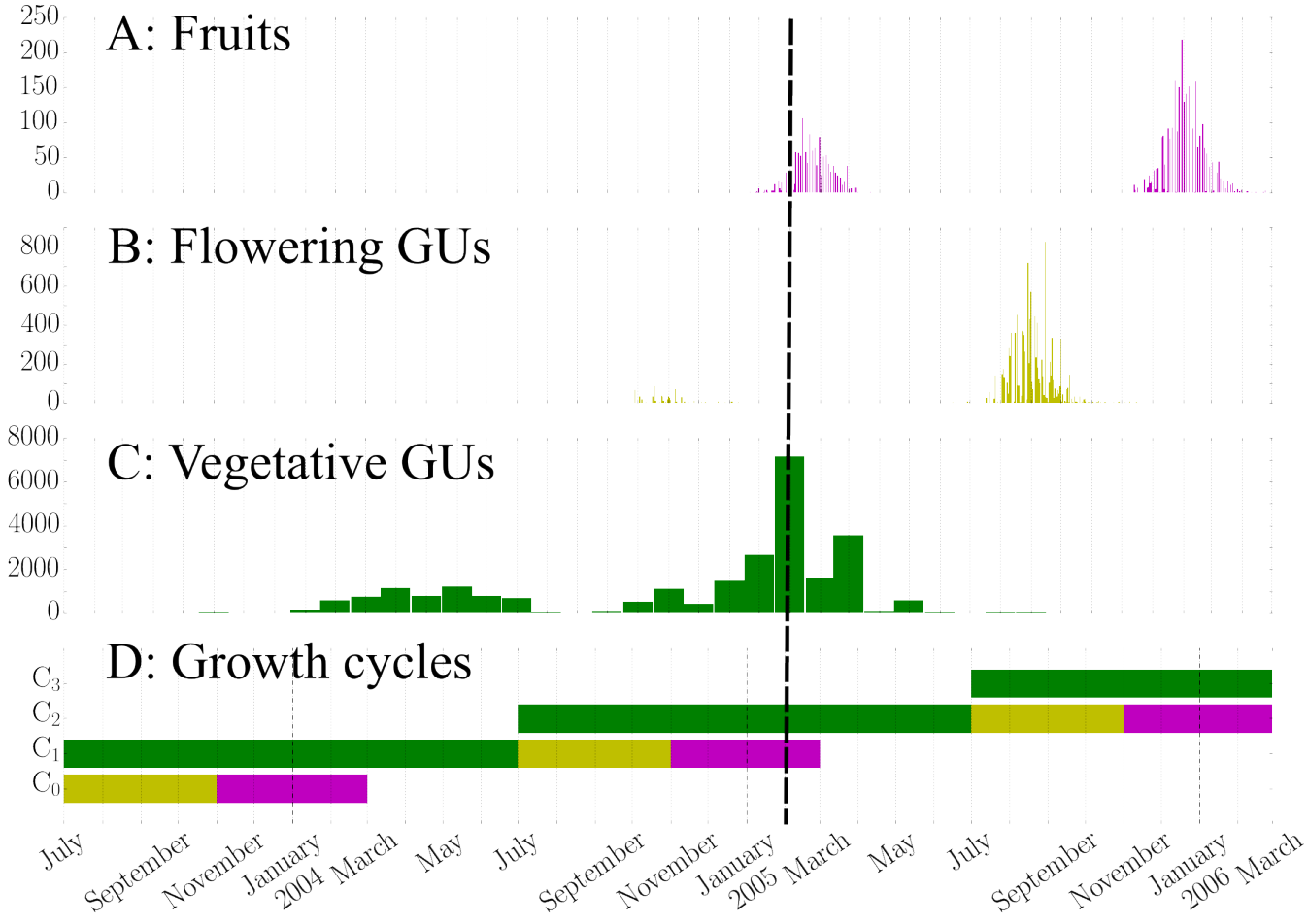


Mango tree growth cycles

GU = Growth Unit

Introduction

¬ Focus on the application at macroscopic scale

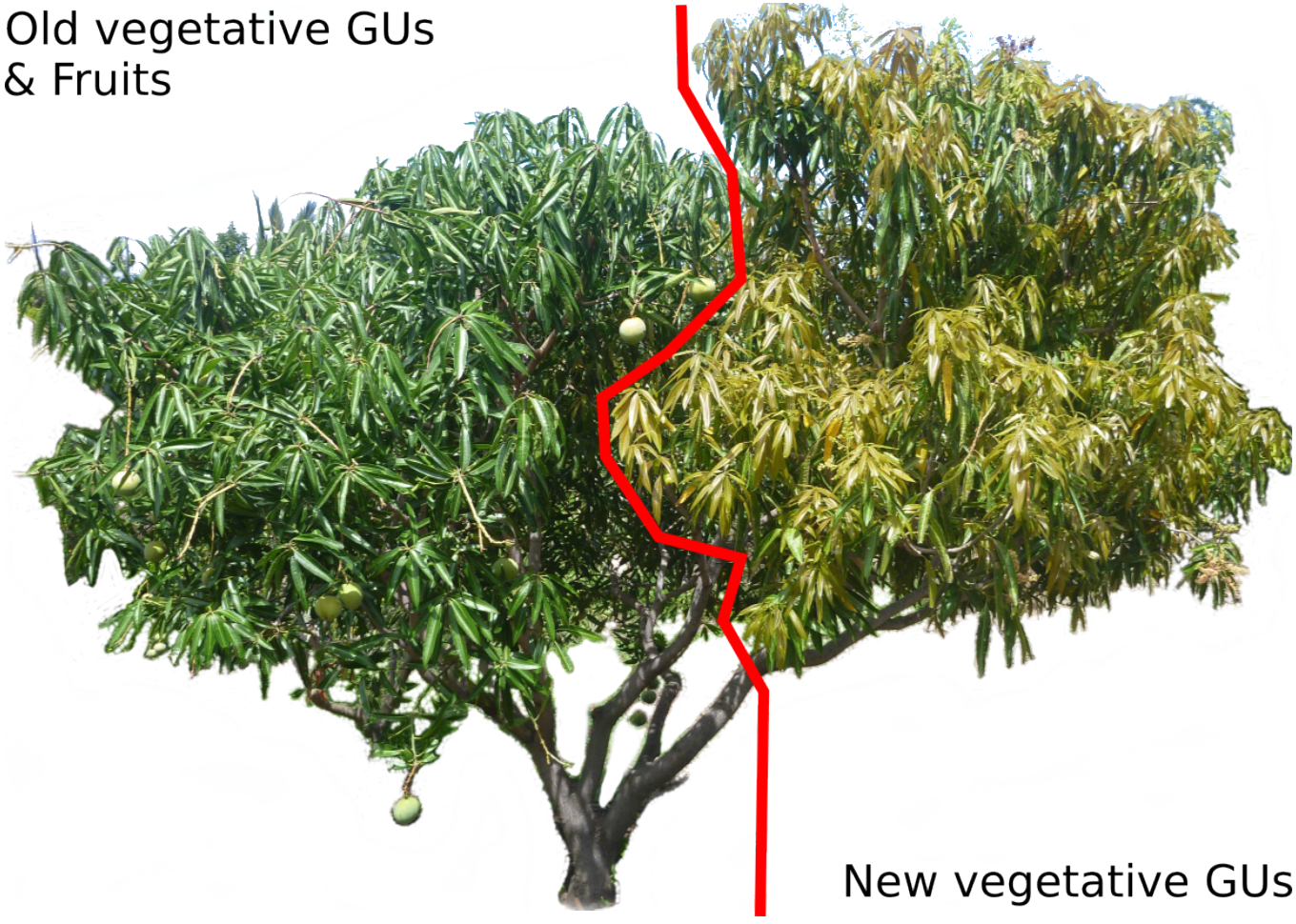


Mango tree growth cycles

GU = Growth Unit

Introduction

¬ Focus on the application at macroscopic scale



Mango tree patchiness illustration [Dambreville, 2012]

Introduction

¬ Focus on the application at macroscopic scale

Patchiness is characterized by clumps of either vegetative or reproductive GUs within the canopy [Chacko, 1986].
Concerns more or less large branching systems and entails various agronomic problems [Ramírez and Davenport, 2010].
Our objective is unfold as follows:

1. Identifying the mechanisms responsible for tree patchiness.
2. Quantifying tree patchiness.

The experimental orchard was located at the Cirad research station in Saint-Pierre, Réunion Island [Dambreville et al., 2013].
7 cultivars, 5 mango trees by cultivar.
Described at the GU scale for 2 complete growth cycles.

Introduction

¬ Focus on the application at macroscopic scale

Patchiness is characterized by clumps of either vegetative or reproductive GUs within the canopy [Chacko, 1986].
Concerns more or less large branching systems and entails various agronomic problems [Ramírez and Davenport, 2010].
Our objective is unfold as follows:

1. Identifying the mechanisms responsible for tree patchiness.
2. Quantifying tree patchiness.

The experimental orchard was located at the Cirad research station in Saint-Pierre, Réunion Island [Dambreville et al., 2013].
7 cultivars, 5 mango trees by cultivar.
Described at the GU scale for 2 complete growth cycles.

Introduction

¬ Focus on the application at macroscopic scale

Patchiness is characterized by clumps of either vegetative or reproductive GUs within the canopy [Chacko, 1986].
Concerns more or less large branching systems and entails various agronomic problems [Ramírez and Davenport, 2010].

Our objective is unfold as follows:

1. Identifying the mechanisms responsible for tree patchiness.
2. Quantifying tree patchiness.

The experimental orchard was located at the Cirad research station in Saint-Pierre, Réunion Island [Dambreville et al., 2013].
7 cultivars, 5 mango trees by cultivar.
Described at the GU scale for 2 complete growth cycles.

Introduction

¬ Overview

Statistical modeling framework:

Markov tree models

- Introduction
- Parametrization of generation distributions
- Inference of generation distributions
- Application

Tree Segmentation/Clustering Models

- Introduction
- Segmentation models
- Clustering models

To deal with 2 different questions:

1. Motif detection,
2. Homogeneous zone detection.

Markov tree models

Statistical modeling framework:

Markov tree models
- Introduction
- Parametrization of generation distributions
- Inference of generation distributions
- Application

Tree Segmentation/Clustering Models
- Introduction
- Segmentation models
- Clustering models

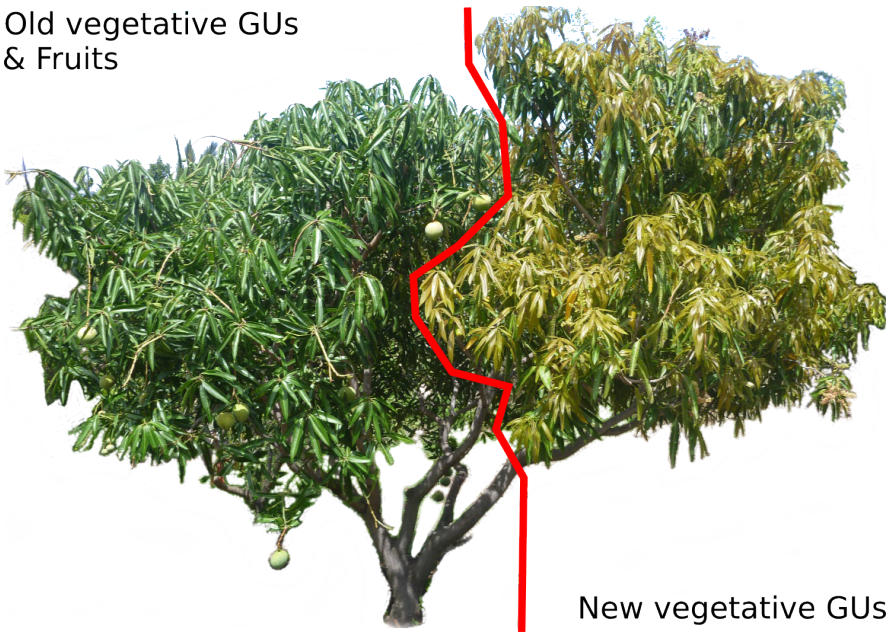
Motif detection in order to:

1. Identify the mechanisms responsible for tree patchiness.

Markov tree models

¬ Introduction – Objectives

1. Identifying the mechanisms responsible for tree patchiness.

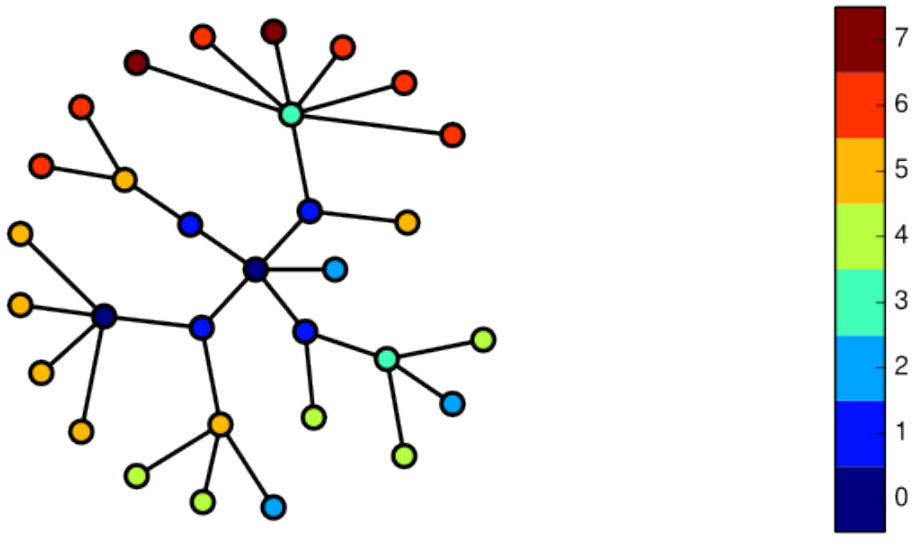


Mango tree growth

Markov tree models

Factorization of

$$P(\bar{X} = \bar{x}, \bar{N} = \bar{n})$$



Tree-indexed data representation

Markov tree models

Factorization of

$$P(\bar{X} = \bar{x}, \bar{N} = \bar{n})$$

Assumptions:

- Markov hypothesis

$$\forall t \in \mathcal{T}, X_t \perp\!\!\!\perp \bar{N}_{\mathrm{nd}(t)\setminus\{\mathrm{pa}(t)\}}, \bar{X}_{\mathrm{nd}(t)\setminus\{\mathrm{pa}(t)\}} \,|\, X_{\mathrm{pa}(t)}$$
$$N_t \perp\!\!\!\perp \bar{N}_{\mathrm{nd}(t)}, \bar{X}_{\mathrm{nd}(t)} \,|\, X_t,$$

- Invariance by permutation,
- Homogeneity.

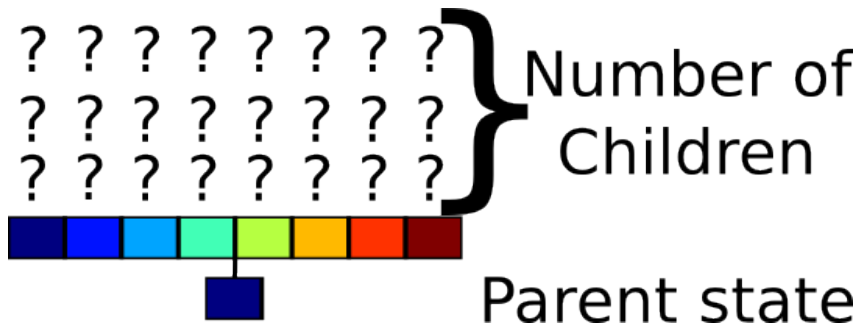
Multi-type branching process [Haccou et al., 2005]:

$$P(\bar{X} = \bar{x}, \bar{N} = n) \propto P(X_0 = x_0) \prod_{t \in \mathcal{T}} P(\boldsymbol{N}_{\mathrm{ch}(t)} = \boldsymbol{n}_{\mathrm{ch}(t)} | X_t = x_t)$$

Markov tree models

Multi-Type Branching Process (MTBP) with K states:

- 1 Initial distribution,
- K generation distributions.



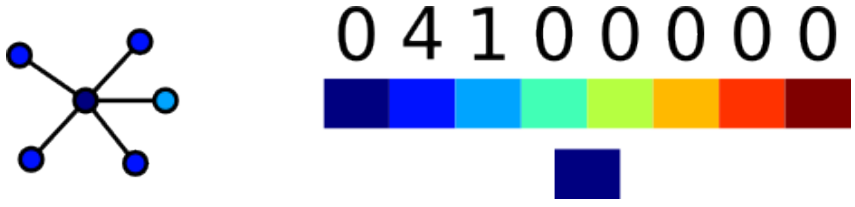
Representation of child state counts using MTBP

Markov tree models

Generating child state counts using MTBP

The initial distribution is not really important but generation distributions are.

Markov tree models

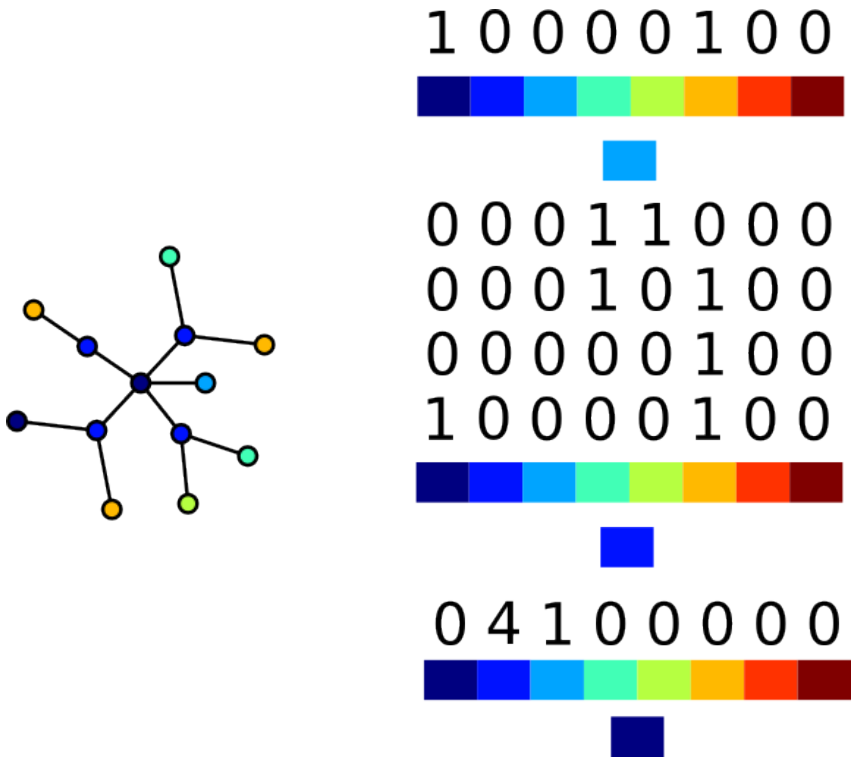


Generating child state counts using MTBP

The outcomes of generation distributions are multivariate counts

Markov tree models

¬ Introduction – Multi-type branching processes



Generating child state counts using MTBP

There are as many generation distribution as states

Markov tree models

¬ Parametrization of generation distributions – Requirements

1. Multivariate parametric distributions have to be used since the combinatorics induced by the variable and high number of children in each state induces a rapid inflation in the number of parameters.

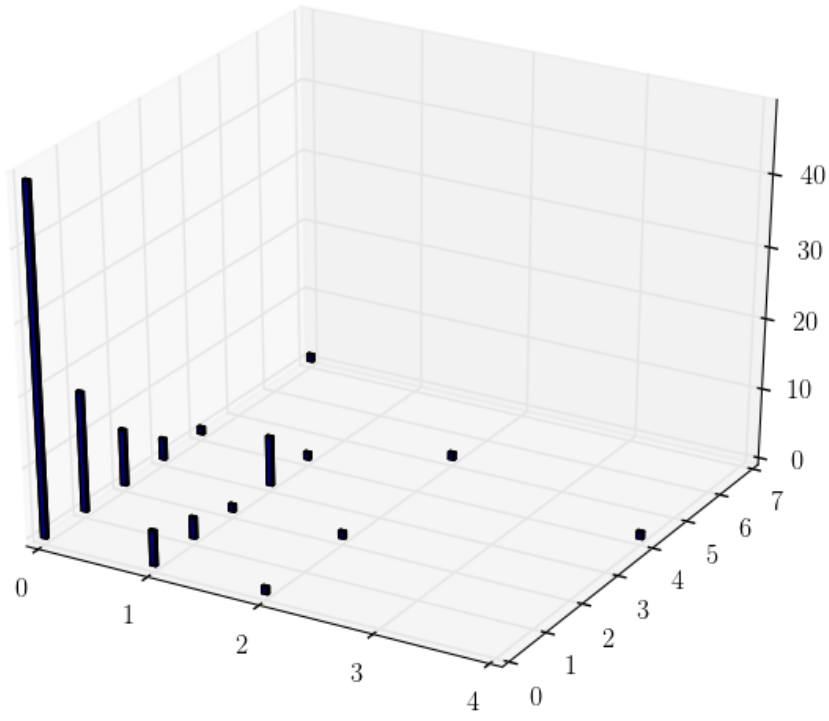
Number	Maximal degree		
of states	2	3	4
2	**11**	**19**	**29**
3	**29**	**59**	**104**
4	**59**	**139**	**279**

Number of parameters of MTBPs as a function of the number of states and the maximal degree.

Markov tree models

¬ Parametrization of generation distributions – Requirements

2. These multivariate parametric distributions can be zero-inflated, right-skewed and have discrete valued marginals.

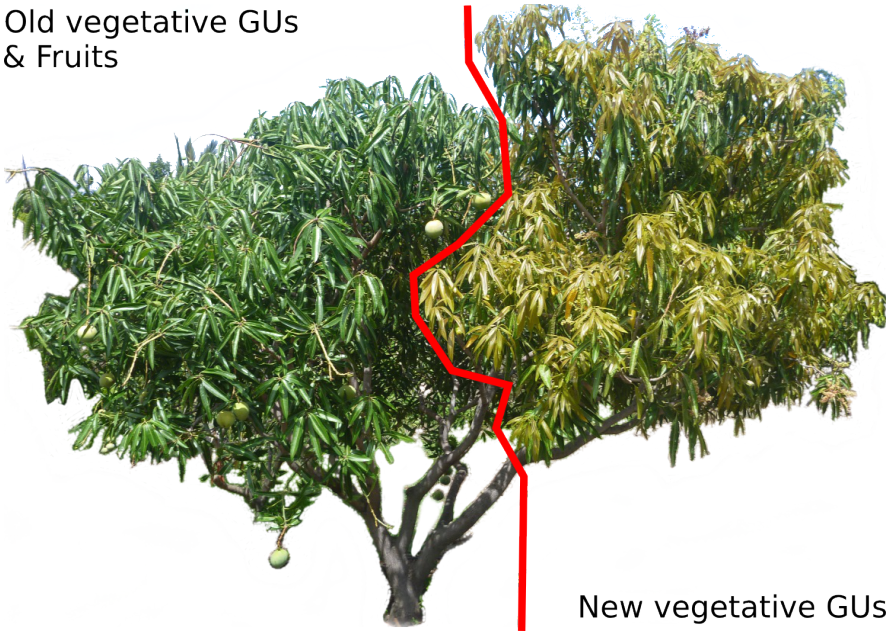


Frequency distribution of the number of children in 2 states given a parent state

Markov tree models

¬ Parametrization of generation distributions – Requirements

3. These multivariate parametric distributions can easily be simulated and probability masses can easily be computed in order to investigate motifs induced by generation distributions and long-range patterns stemming from these generation distributions as trees develop.

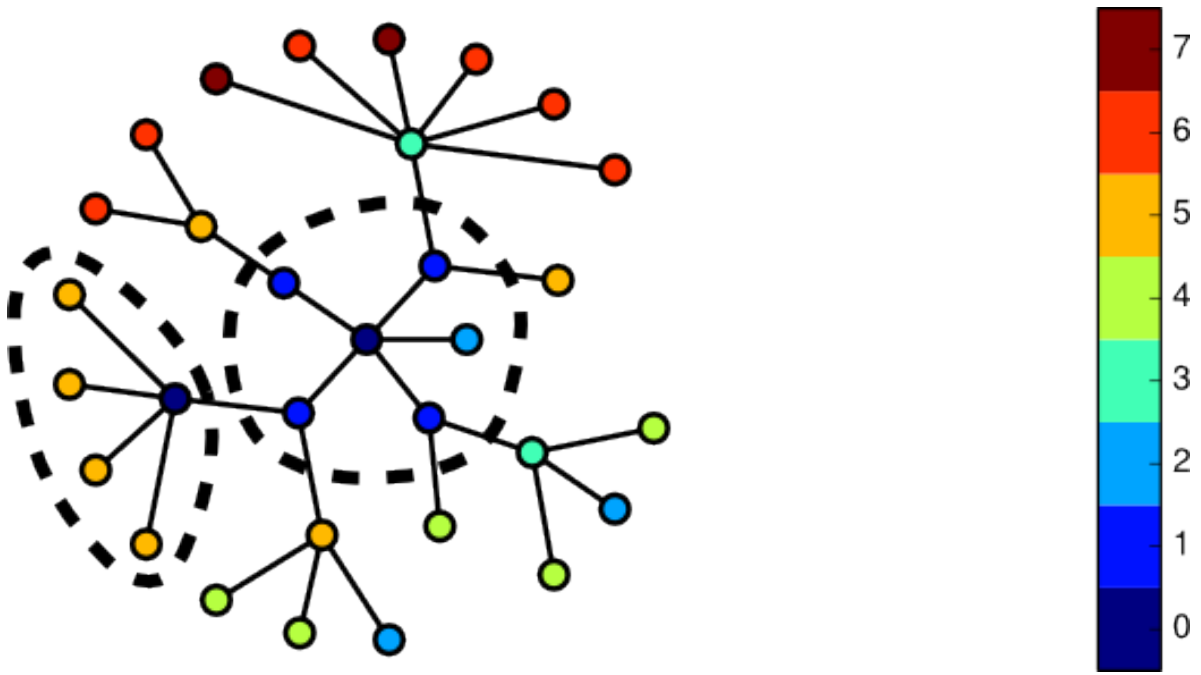


Mango tree growth

Markov tree models

¬ Parametrization of generation distributions – Requirements

4. Since child states tend to appear simultaneously or on the contrary asynchronously, conditional independences in these generation distributions must be inferred.



Associations and competitions in generation distributions

Markov tree models

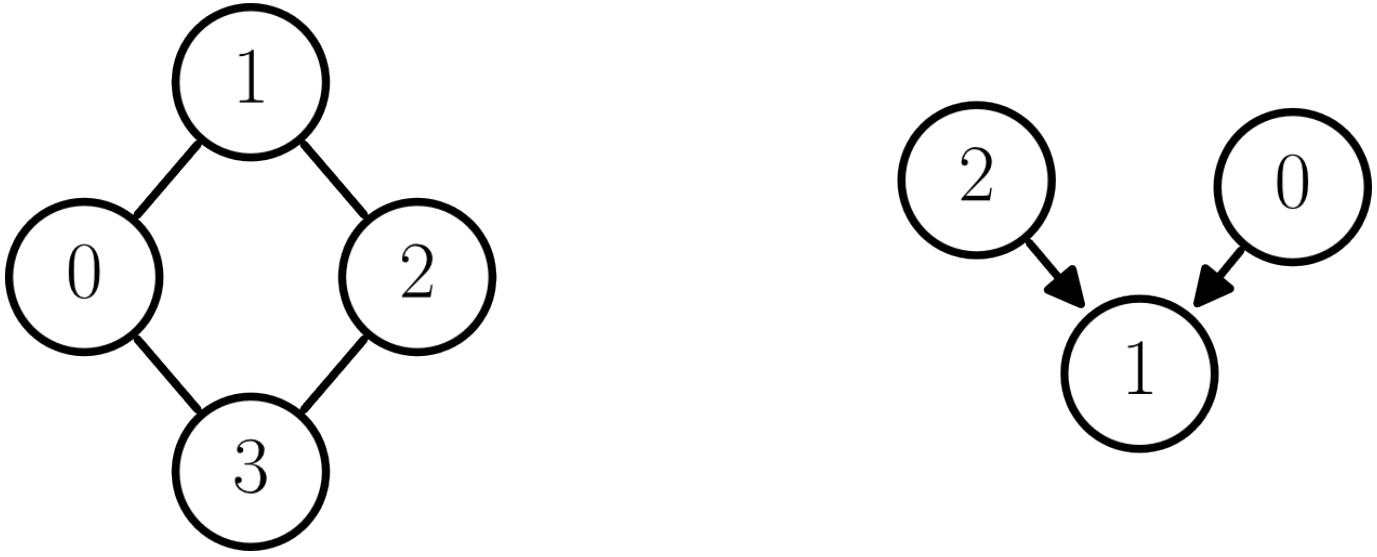
¬ Parametrization of generation distributions – Graphical models

Use of graphical models to represent conditional independence relationships.

Distribution factorizations inducing dependency patterns encoded in graphs [Lauritzen, 1996].

$\mathcal{G} = (\mathcal{V}, \mathcal{E})$ is a graph where

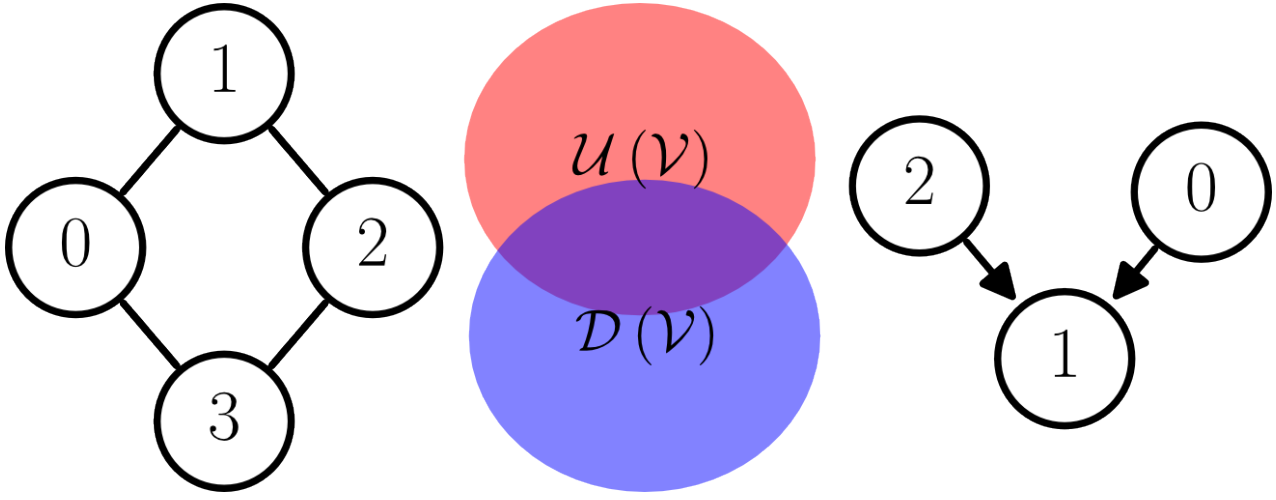
- the vertex set represents variables,
- the edge set represents direct dependencies.



Undirected Graph (UG), $\mathcal{G}_u$, and Directed Acyclic Graph (DAG), $\mathcal{G}_d$

Markov tree models

¬ Parametrization of generation distributions – Graphical models



I-space of UGs, $\mathcal{U}(\mathcal{V})$, and DAGs, $\mathcal{D}(\mathcal{V})$

$\mathcal{I}(\mathcal{G}_u) =$
$3 \perp\!\!\!\perp 1 | 0, 2$
$0 \perp\!\!\!\perp 2 | 3, 1$
diamond shape

$\mathcal{I}(\mathcal{G}_d) =$
$0 \perp\!\!\!\perp 2$
$0 \not\perp\!\!\!\perp 2 | 1$
v-shape

$\mathcal{D}(\mathcal{V}) \cap \mathcal{U}(\mathcal{V})$ contains DGs with no v-shapes and chordal UGs.

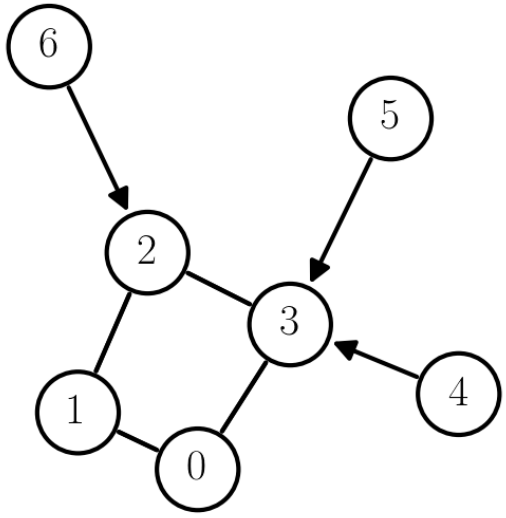
Markov tree models

¬ Parametrization of generation distributions – Graphical models

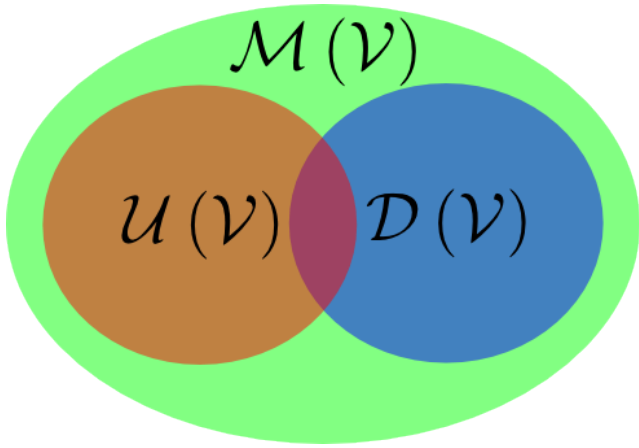
Mixed Acyclic Graphs (MAGs) combining:

- v-shapes (5, 4 and 3),
- diamond shapes (0, 1, 2 and 3).

and introducing u-shapes (6, 5, 2 and 3).



A MAG



I-space of UGs, DAGs and MAGs, $\mathcal{M}(\mathcal{V})$

Markov tree models

¬ Parametrization of generation distributions – Graphical models

Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be a MAG.
Factorization property:

$$P\left[N_0 = n_0, ..., N_{K-1} = n_{K-1}\right] = P[\boldsymbol{N} = \boldsymbol{n}]$$
$$= \prod_{\mathcal{C} \in \mathcal{K}} P[\boldsymbol{N}_{\mathcal{C}} = \boldsymbol{n}_{\mathcal{C}} | \boldsymbol{N}_{pa(\mathcal{C})} = \boldsymbol{N}_{pa(\mathcal{C})}]$$

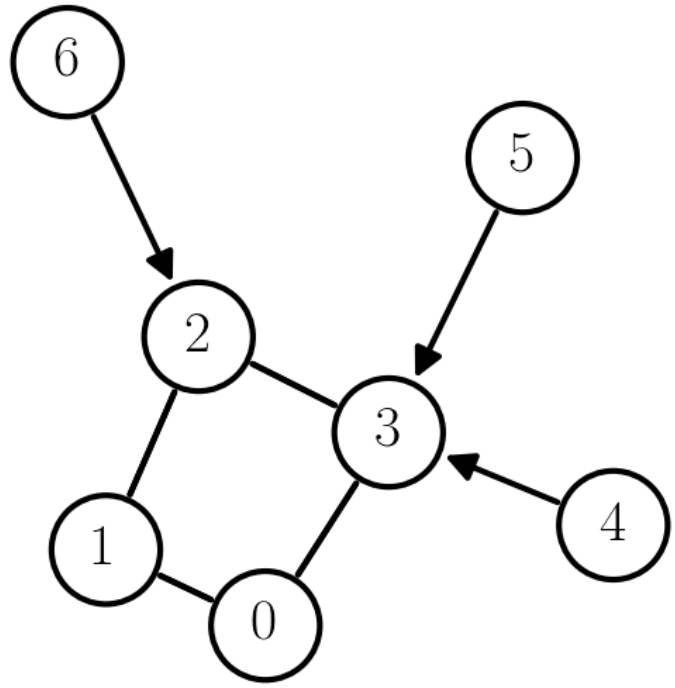
where:

- A chain component $\mathcal{C}$ is a maximal set of vertices connected by undirected edges only,
- $\mathcal{K}$ is the set of chain components,
- $pa(.)$ is the parent set of a chain component.

Acyclicity is the same as in Directed Graph, replacing vertices by chain components.

Markov tree models

¬ Parametrization of generation distributions – MAG models



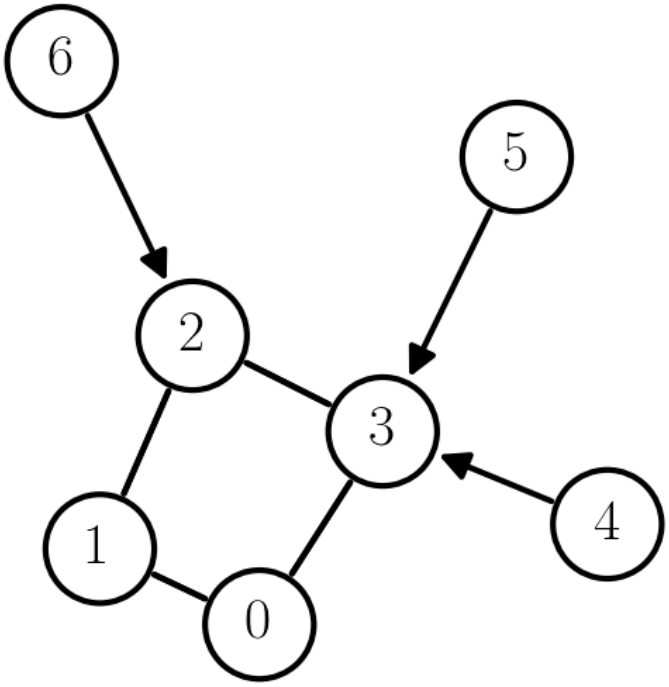
The factorization of MAGs

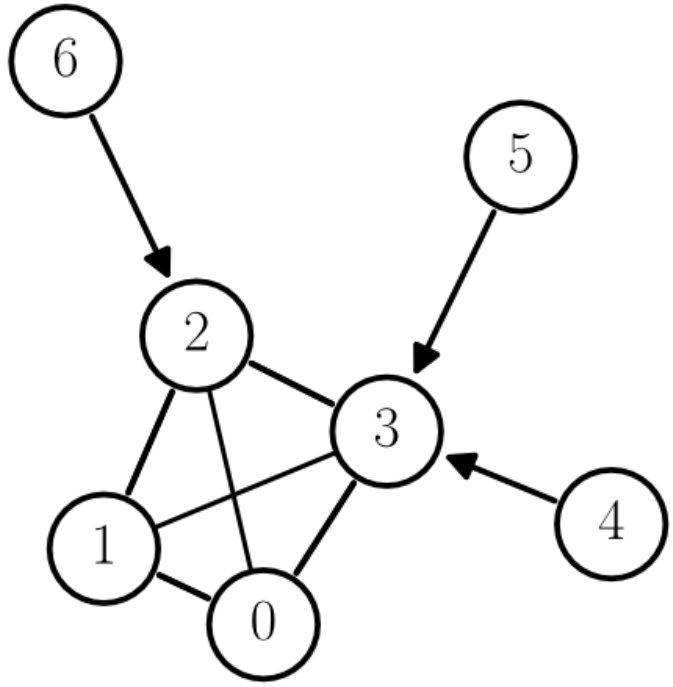
$$\mathcal{K} = \{\{0,1,2,3\},\{4\},\{5\},\{6\}\}$$

$$P\left(\boldsymbol{N}_{\{0,1,2,3\}} = \boldsymbol{n}_{\{0,1,2,3\}} \,\middle|\, \boldsymbol{N}_{\{4,5,6\}} = \boldsymbol{n}_{\{4,5,6\}}\right) \prod_{i \in \{4,5,6\}} P\left(N_i = n_i\right)$$

Markov tree models

¬ Parametrization of generation distributions – MAG models





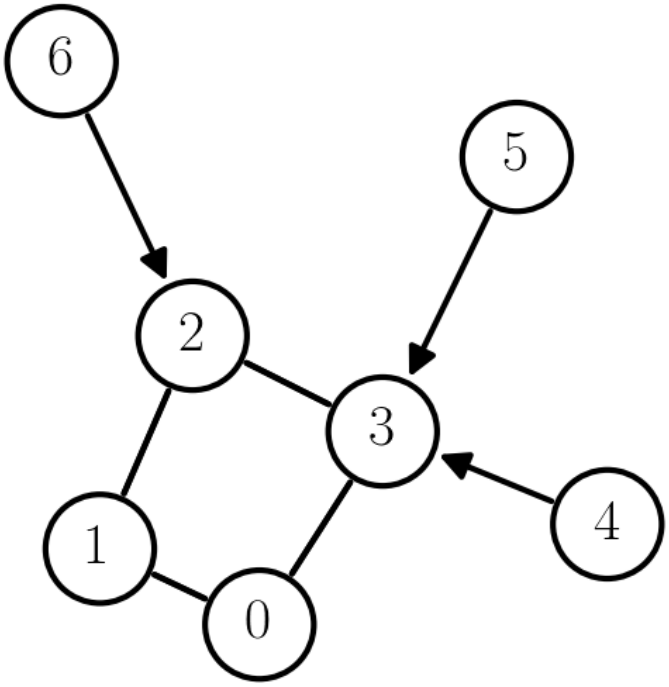
The factorization of MAGs

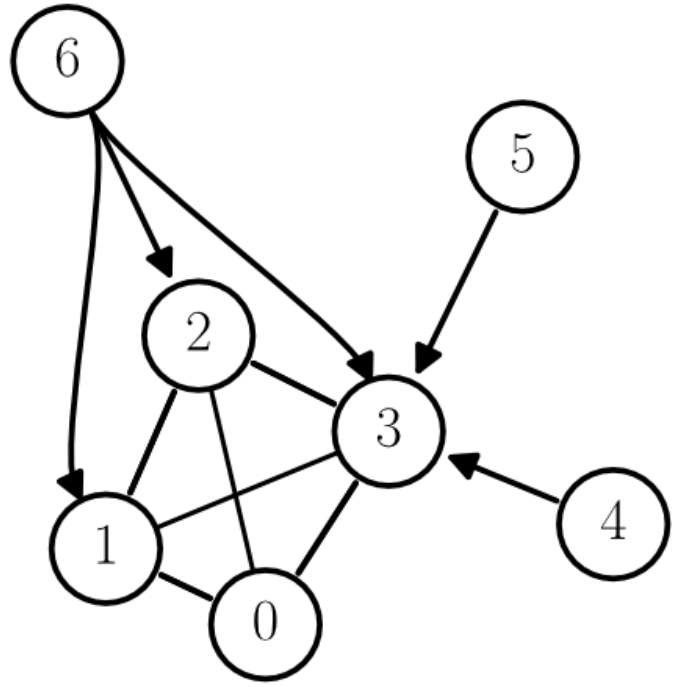
$$\mathcal{K} = \{\{0,1,2,3\},\{4\},\{5\},\{6\}\}$$

$$P\left(\boldsymbol{N}_{\{0,1,2,3\}} = \boldsymbol{n}_{\{0,1,2,3\}} \,\middle|\, \boldsymbol{N}_{\{4,5,6\}} = \boldsymbol{n}_{\{4,5,6\}}\right) \prod_{i\in\{4,5,6\}} P\left(N_i = n_i\right)$$

Markov tree models

¬ Parametrization of generation distributions – MAG models





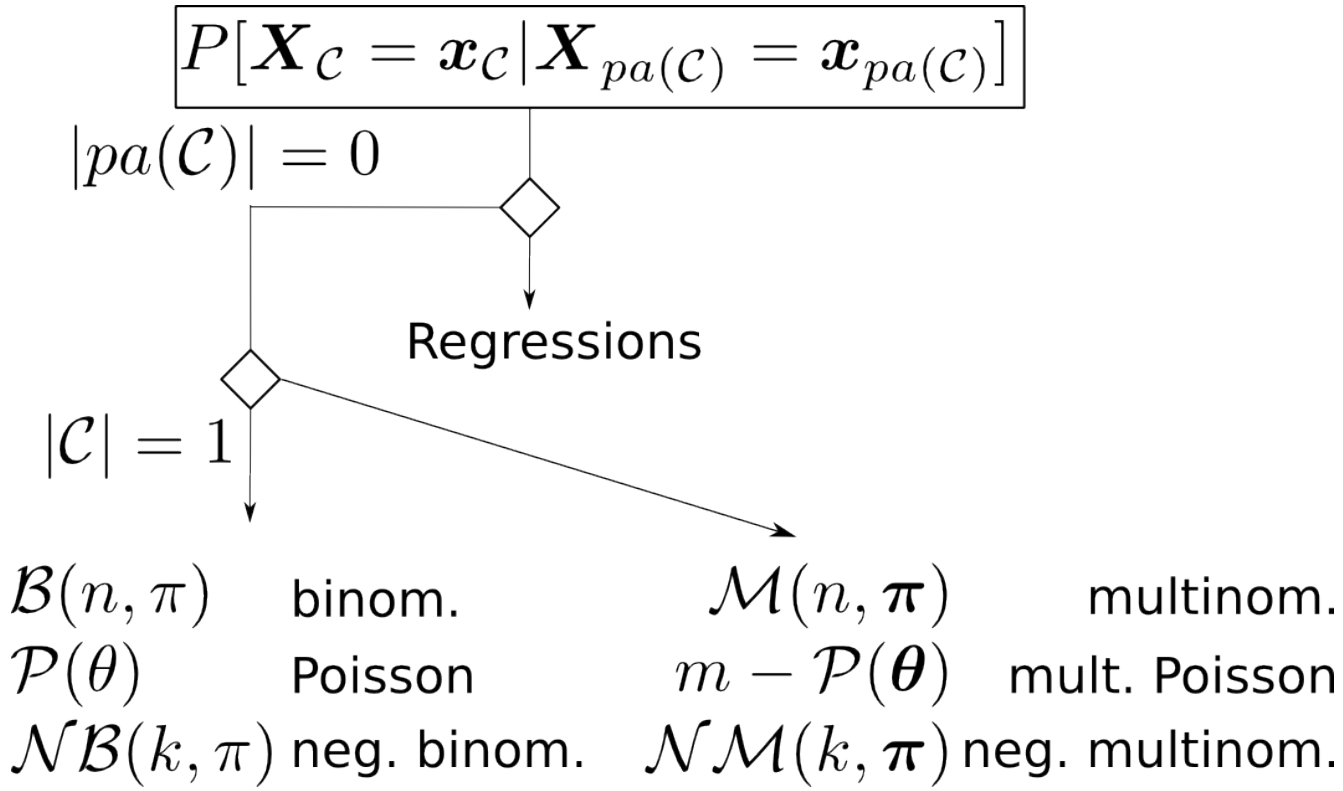
The factorization of MAGs

$$\mathcal{K} = \{\{0,1,2,3\},\{4\},\{5\},\{6\}\}$$

$$P\left(\boldsymbol{N}_{\{0,1,2,3\}} = \boldsymbol{n}_{\{0,1,2,3\}} \,\middle|\, \boldsymbol{N}_{\{4,5,6\}} = \boldsymbol{n}_{\{4,5,6\}}\right) \prod_{i\in\{4,5,6\}} P\left(N_i = n_i\right)$$

Markov tree models

¬ Parametrization of generation distributions – MAG models



Parametric components

Easy [Johnson et al., 1993, Johnson et al., 1997]:

- simulation,
- estimation,

Consequences:

1. Only cliques,
2. In cliques, same parents.

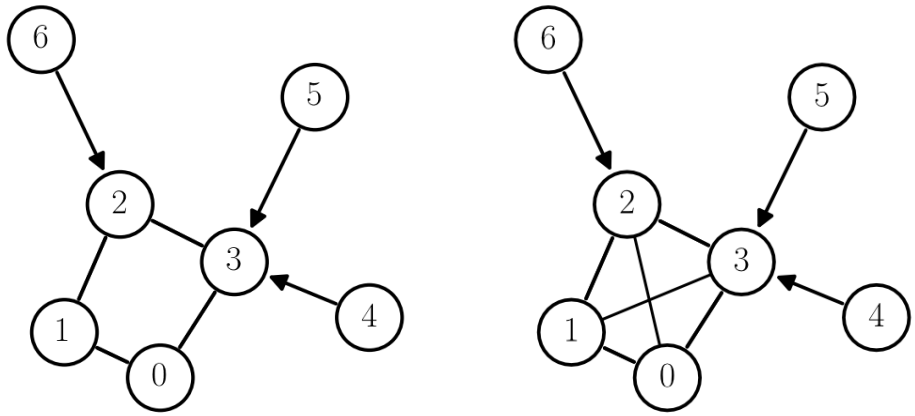
Markov tree models

¬ Inference of generation distributions – Parameter inference

Consequences of the parametrization:

1. Only cliques,
2. In cliques, same parents.

Thus MAGs given are faithful or not to these constraints.

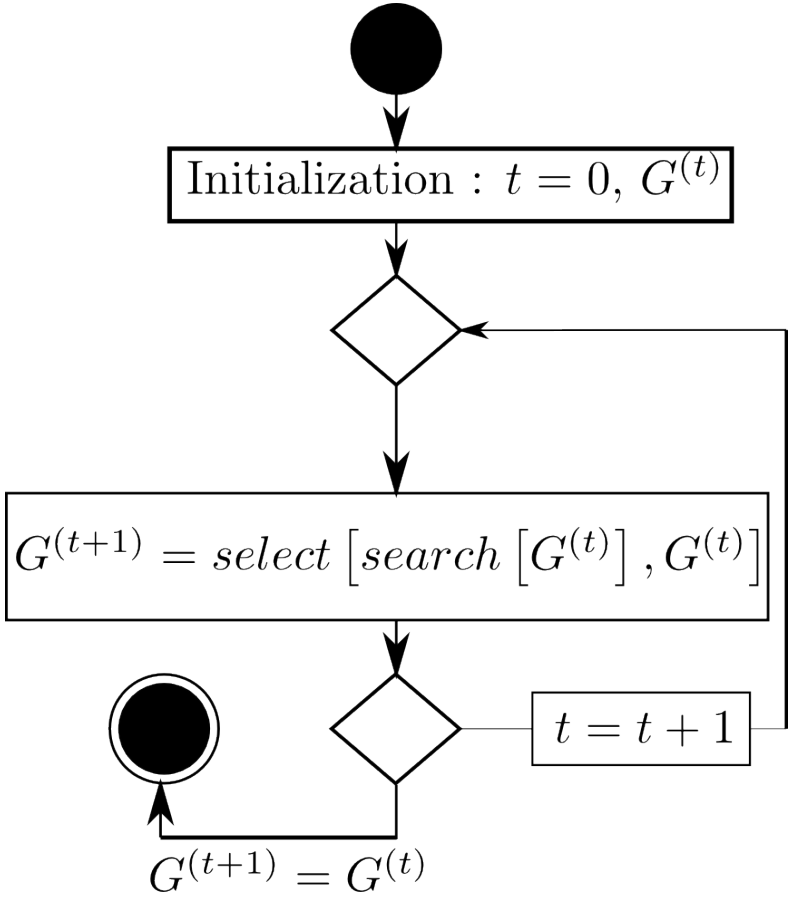


Problematic MAGs

Inference is easy only when graphs are faithful to constraints.

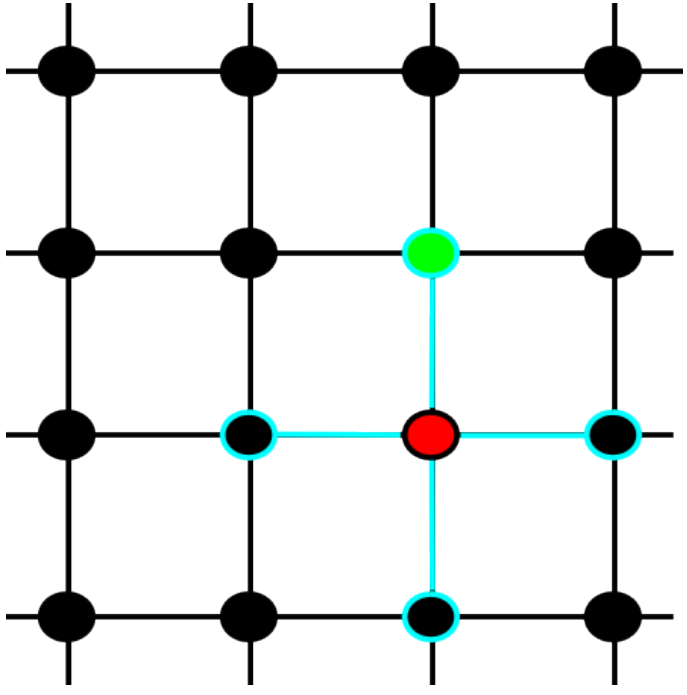
Markov tree models

¬ Inference of generation distributions – Structure inference



Greedy algorithm for DAGs [Koller and Friedman, 2009]:

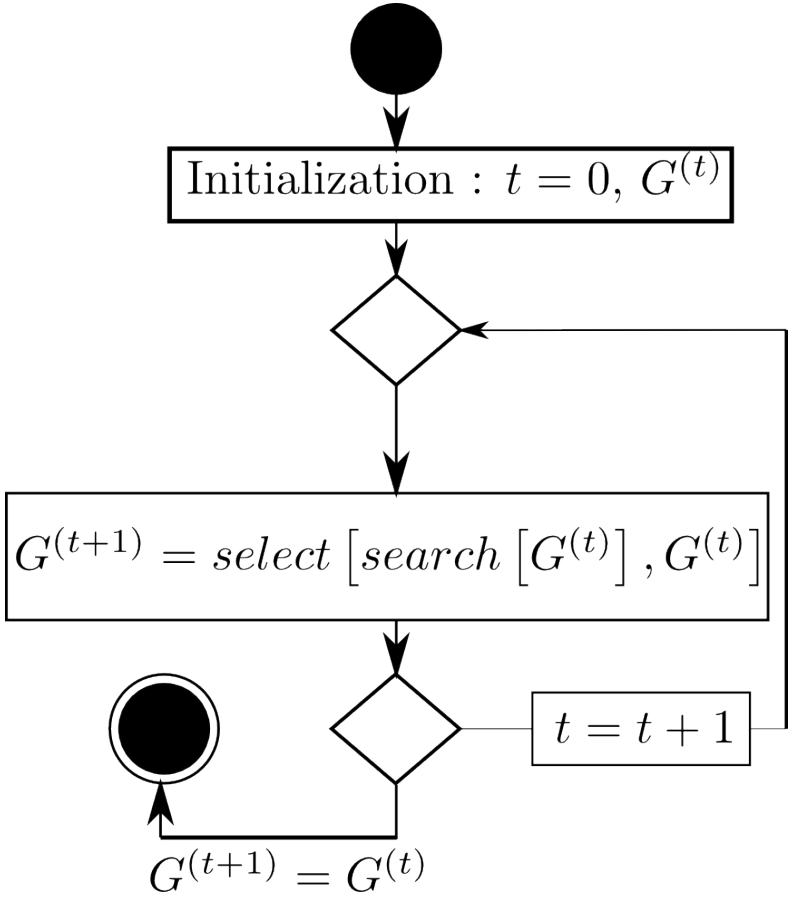
- Starting point $G^{(0)}$
- Search function *search*[.]
- Select function *select*[.]



Search space

Markov tree models

¬ Inference of generation distributions – Structure inference



Greedy algorithm for DAGs [Koller and Friedman, 2009]:

- Starting point $G^{(0)}$
- Search function *search*[.]
- Select function *select*[.]

For search function:

- add an edge,
- remove an edge,
- reverse an edge...

For select function:

- Hill Climbing : arg max using BIC, AIC, Loglikelihood, BDe.
- Simulated Annealing...

Markov tree models

In order to adapt the local search, the search function has to be redefined.
If the MAG search space is easily defined:

- add an edge (directed or not),
- remove an edge,
- reverse an edge,
- orient or disorient an edge.

Since the parametrization induces constraints:

1. Only cliques,
2. In cliques, same parents.

This approach is not relevant !

Markov tree models

In order to adapt the local search, the search function has to be redefined.
If the MAG search space is easily defined:

- add an edge (directed or not),
- remove an edge,
- reverse an edge,
- orient or disorient an edge.

Since the parametrization induces constraints:

1. Only cliques,
2. In cliques, same parents.

This approach is not relevant !

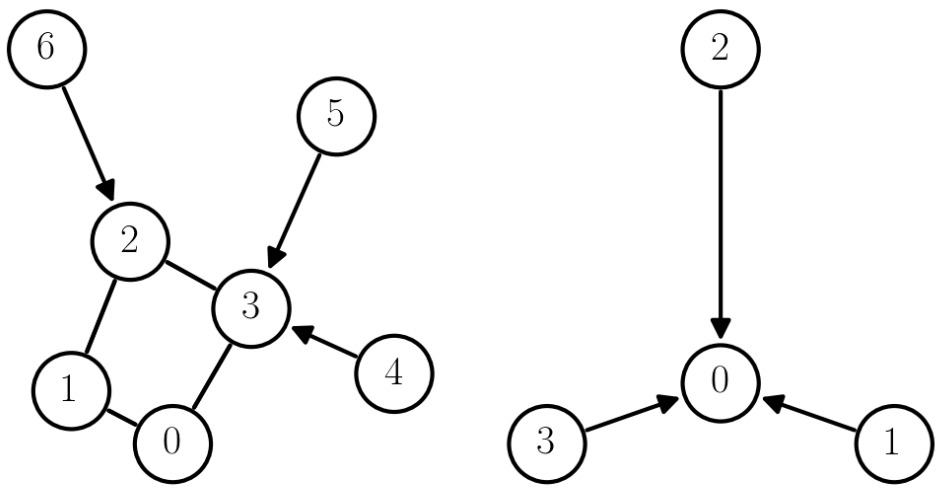
Markov tree models

¬ Inference of generation distributions – Structure inference

Consider a MAG $\mathcal{G} = (\mathcal{V}, \mathcal{E})$.

Let $\mathcal{H} = \left(\Pi, \mathcal{Q}, \tilde{\mathcal{E}}\right)$ be a Quotient Acyclic Graph (QAG) with respect to $\mathcal{G}$ define as follows

- $\Pi = \mathcal{K}$,
- $\mathcal{Q} = \{0, |\mathcal{K}| - 1\}$,
- $\tilde{\mathcal{E}} = \left\{(p, q) \in \mathcal{Q}^2 \,\middle|\, \exists\, (u, v) \in \{\Pi_p \times \Pi_q\} \cap \mathcal{E}\right\}$,



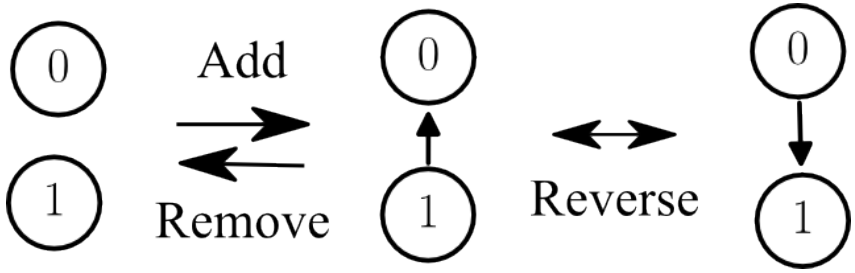
from the MAG $\mathcal{G}$ to its DAG representation $\mathcal{H}$ with chain mapping

$$\Pi = \{\{0, 1, 2, 3\}, \{4\}, \{5\}, \{6\}\}$$

Markov tree models

¬ Inference of generation distributions – Structure inference

Direct use of DAG algorithm on $\mathcal{H}$



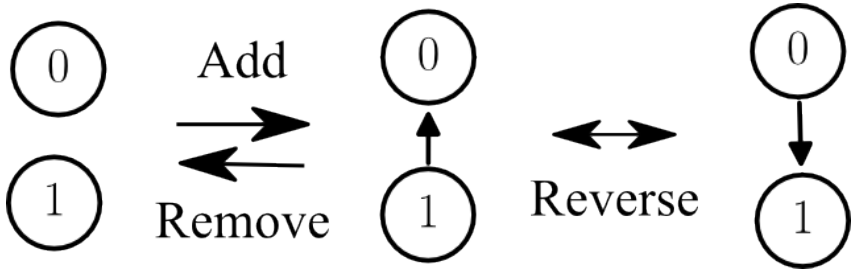
Application of the greedy algorithm on $\mathcal{H}$ with $\Pi = \{\{0, 1, 3\}, \{2\}\}$

And application of resulting modifications on $\mathcal{G}$.

Markov tree models

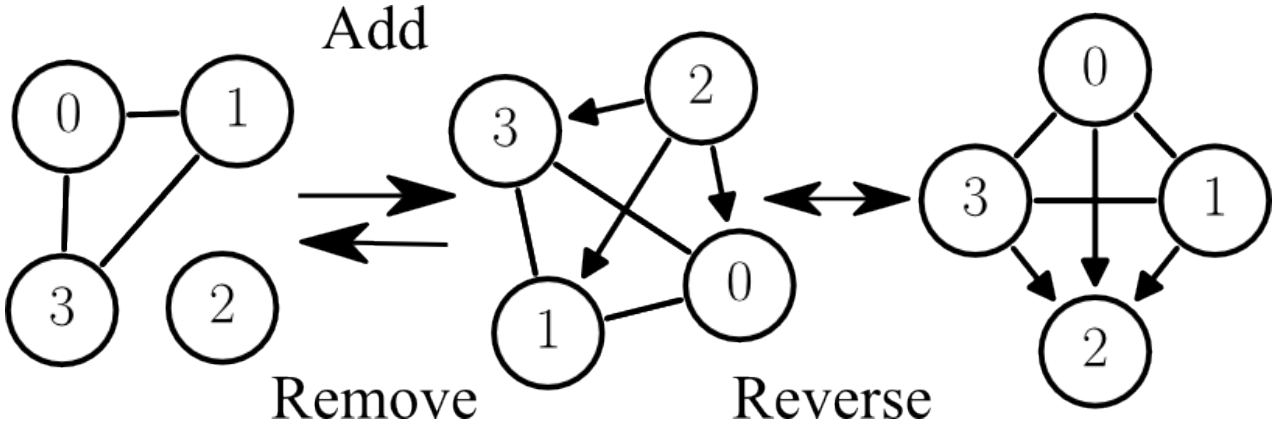
¬ Inference of generation distributions – Structure inference

Direct use of DAG algorithm on $\mathcal{H}$



Application of the greedy algorithm on $\mathcal{H}$ with $\Pi = \{\{0, 1, 3\}, \{2\}\}$

And application of resulting modifications on $\mathcal{G}$.

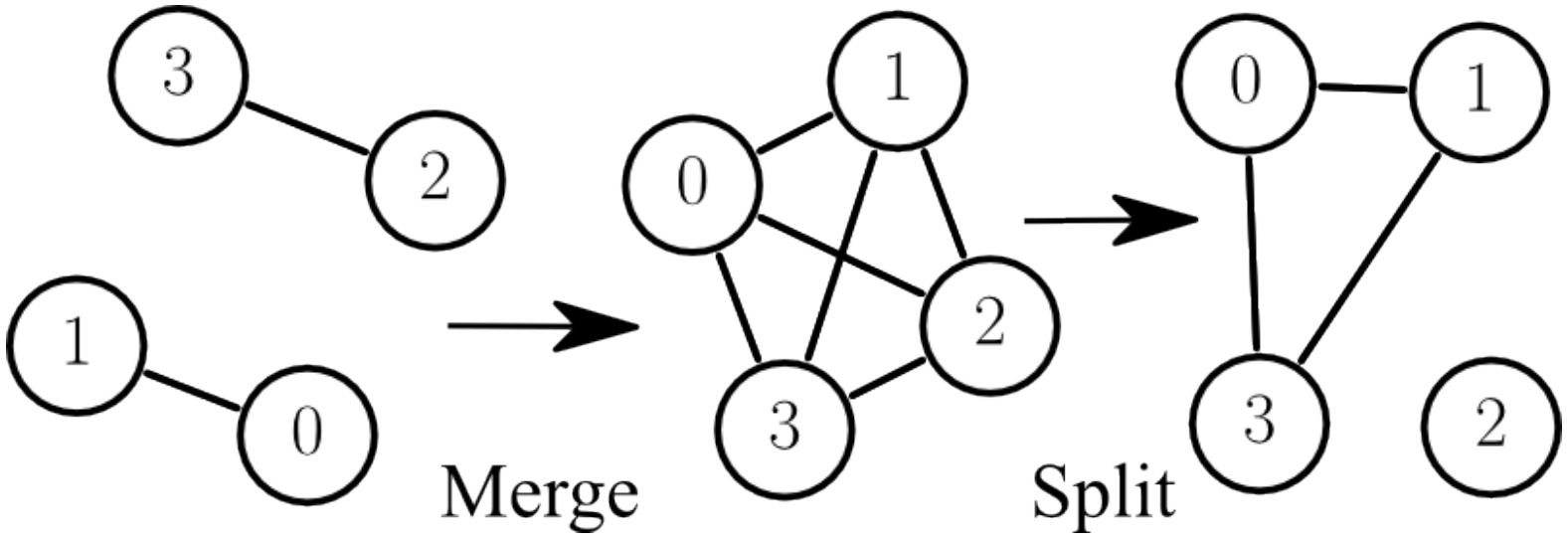


Results of the greedy algorithm on $\mathcal{G}$

Markov tree models

¬ Inference of generation distributions – Structure inference

But the QAG space is not connected using only given edition operators since quotients remain unchanged.



Greedy algorithm

$$\Pi = \{\{0,1\},\{2,3\}\} \underset{merge}{\rightarrow} \{\{0,1,2,3\}\} \underset{split}{\rightarrow} \{\{0,1,3\},\{2\}\}$$

Combining the two set of operators, the QAG space is now connected

Markov tree models

- Modifying parent sets using the DAGs operators.
- Modifying quotients:
  - Increase the number of chain components by removing a vertex from a chain component: $|\Pi^{(t+1)}| = |\Pi^{(t)}| + 1$.
  - Decrease the number of chain components by merging two chain components: $|\Pi^{(t+1)}| = |\Pi^{(t)}| - 1$

Neighborhood search space complexity:

$$O\left(\underbrace{|\mathcal{Q}|^2}_{DAG} + \underbrace{|\mathcal{V}|^2}_{Split} + \underbrace{\binom{|\mathcal{Q}|}{2}}_{merge}\right) \approx O\left(\mathcal{V}^2\right)$$

Same order of magnitude as for DAGs.

Markov tree models

¬ Inference of generation distributions – Structure inference

Time complexity of each step:

$$O\left(|\mathcal{V}|^3\right)$$

Since model estimation for a vertex is considered as constant.

Likelihood (and derived score) is decomposable [Koller and Friedman, 2009]:

$$score[G] = \sum_{\mathcal{C} \in \mathcal{K}} score[\mathcal{C}|pa\left(\mathcal{C}\right)]$$

Therefore, only the scores for changed subgraphs have to be updated:

$$O\left(|\mathcal{V}|^2\right)$$

Moreover only one (or two) clique can change at each step.
Therefore, using score and subgraph caching yields complexity:

$$O\left(|\mathcal{V}|\right)$$

Markov tree models

¬ Inference of generation distributions – Structure inference

Time complexity of each step:

$$O\left(|\mathcal{V}|^3\right)$$

Since model estimation for a vertex is considered as constant.
Likelihood (and derived score) is decomposable
[Koller and Friedman, 2009]:

$$score[G] = \sum_{\mathcal{C} \in \mathcal{K}} score[\mathcal{C}|pa\left(\mathcal{C}\right)]$$

Therefore, only the scores for changed subgraphs have to be updated:

$$O\left(|\mathcal{V}|^2\right)$$

Moreover only one (or two) clique can change at each step.
Therefore, using score and subgraph caching yields complexity:

$$O\left(|\mathcal{V}|\right)$$

Markov tree models

¬ Inference of generation distributions – Structure inference

Time complexity of each step:

$$O\left(|\mathcal{V}|^3\right)$$

Since model estimation for a vertex is considered as constant.
Likelihood (and derived score) is decomposable
[Koller and Friedman, 2009]:

$$score[G] = \sum_{\mathcal{C} \in \mathcal{K}} score[\mathcal{C}|pa\left(\mathcal{C}\right)]$$

Therefore, only the scores for changed subgraphs have to be updated:

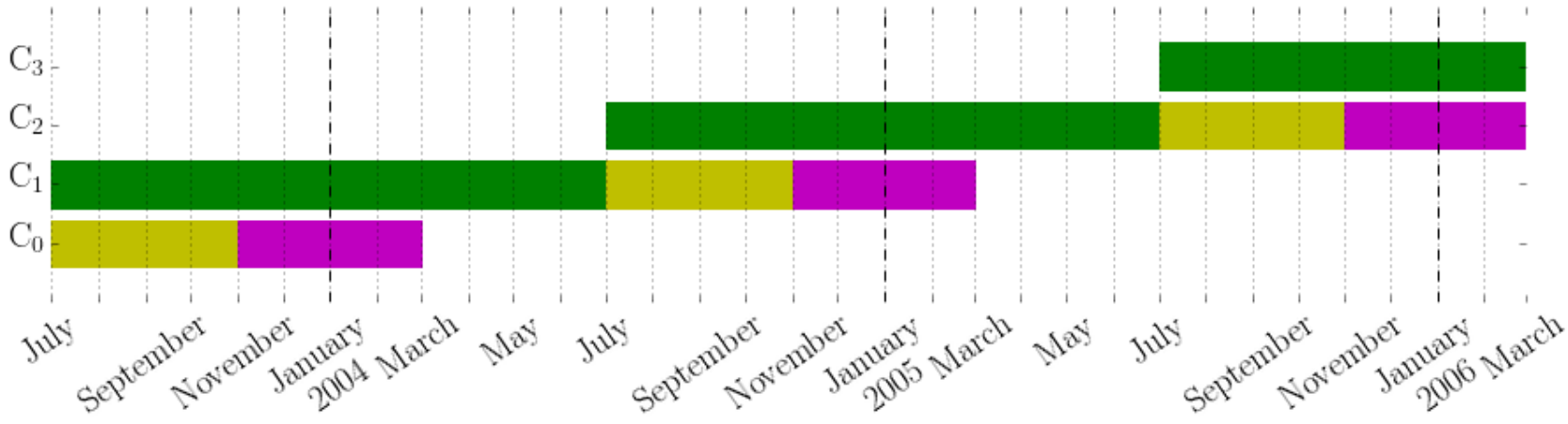
$$O\left(|\mathcal{V}|^2\right)$$

Moreover only one (or two) clique can change at each step.
Therefore, using score and subgraph caching yields complexity:

$$O\left(|\mathcal{V}|\right)$$

Markov tree models

¬ Application – States of mango trees



Mango tree growth cycle

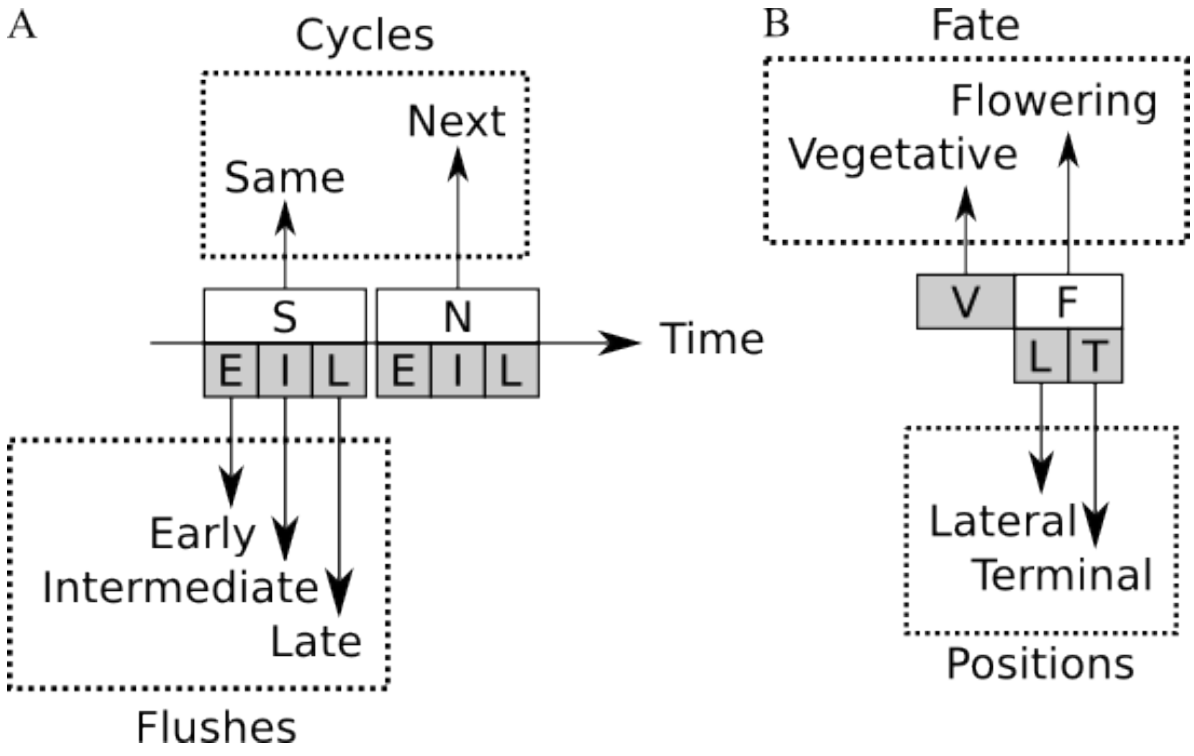
Early flush. Period where the vegetative phase of a growth cycle overlaps the flowering phase of the previous cycle.

Intermediate flush. Period where the vegetative phase of a growth cycle overlaps the fructifying phase of the previous cycle.

Late flush. Period where the vegetative phase of a growth cycle does not overlap the previous or the next cycles.

Markov tree models

¬ Application – States of mango trees



Mango tree states

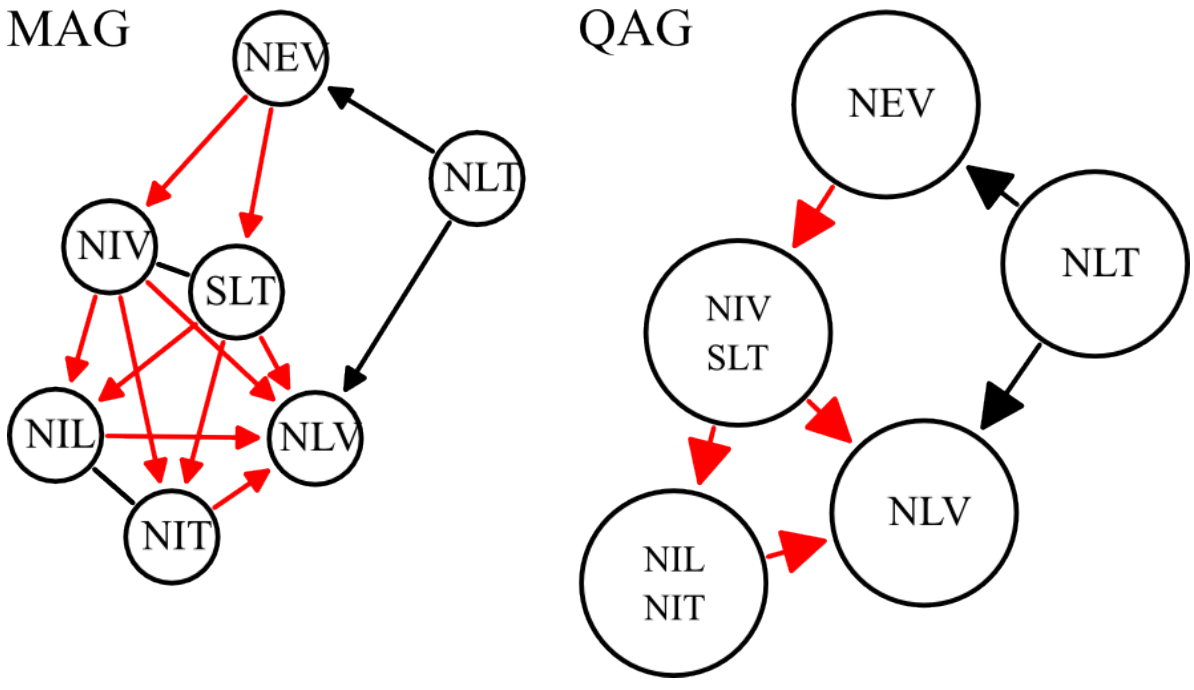
As a consequence, we have the following observation space

$$\mathcal{X} = \{\text{SEV, SLV, NEV, NIV, NLV, SIT, SLT, NIT, NLT, SIL, NIL}\}$$

Markov tree models

¬ Application – Inference of generation distributions

Focus on SIT parent state for Cogshall cultivar: 100 GUs
No children in the same cycle (except few SLT)

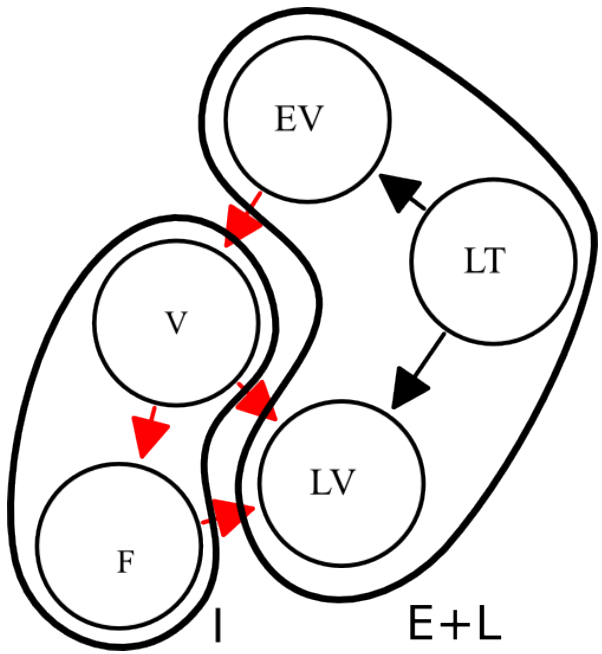


Infered generation distribution for the SIT state

Markov tree models

¬ Application – Inference of generation distributions

Focus on SIT parent state for Cogshall cultivar: 100 GUs
No children in the same cycle (except few SLT)



Infered generation distribution for the SIT state

Markov tree models

¬ Application – Interpretation of generation distributions

Statistical modeling framework:

Markov tree models

- Introduction
- Parametrization of generation distributions
- Inference of generation distributions
- Application

Tree Segmentation/Clustering Models

- Introduction
- Segmentation models
- Clustering models

Motif detection in order to:

1. Identify the mechanisms responsible for tree patchiness.

$\Rightarrow$ Patchiness results from mutual exclusions, at the local scale of sibling GUs, between their burst dates and/or fates.

Tree Segmentation/Clustering Models

Statistical modeling framework:

Markov tree models
- Introduction
- Parametrization of generation distributions
- Inference of generation distributions
- Application

Tree Segmentation/Clustering Models
- Introduction
- Segmentation models
- Clustering models

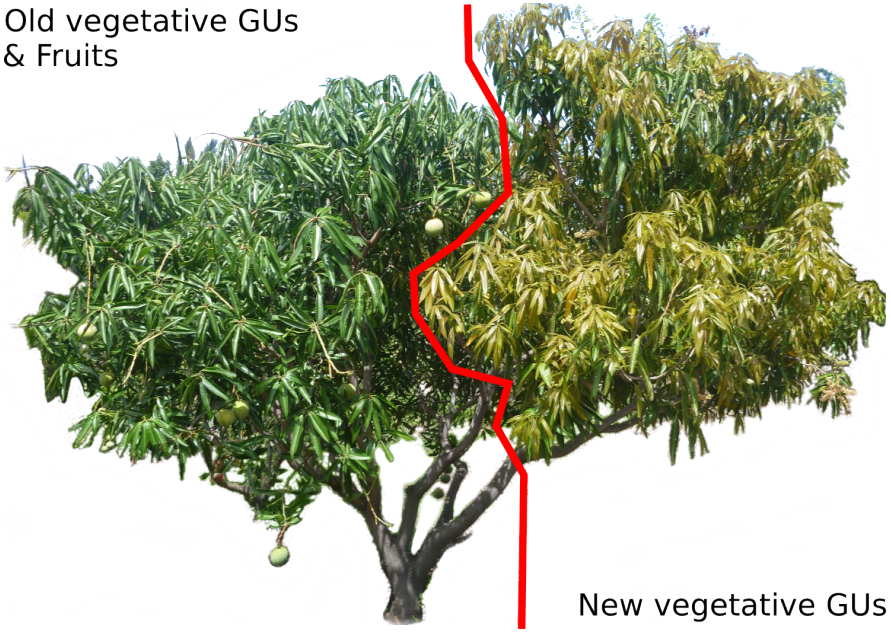
Homogeneous zone detection in order to:

2. Quantifying tree patchiness.

Tree Segmentation/Clustering Models

¬ Introduction – Principle

2. Characterizing tree patchiness.

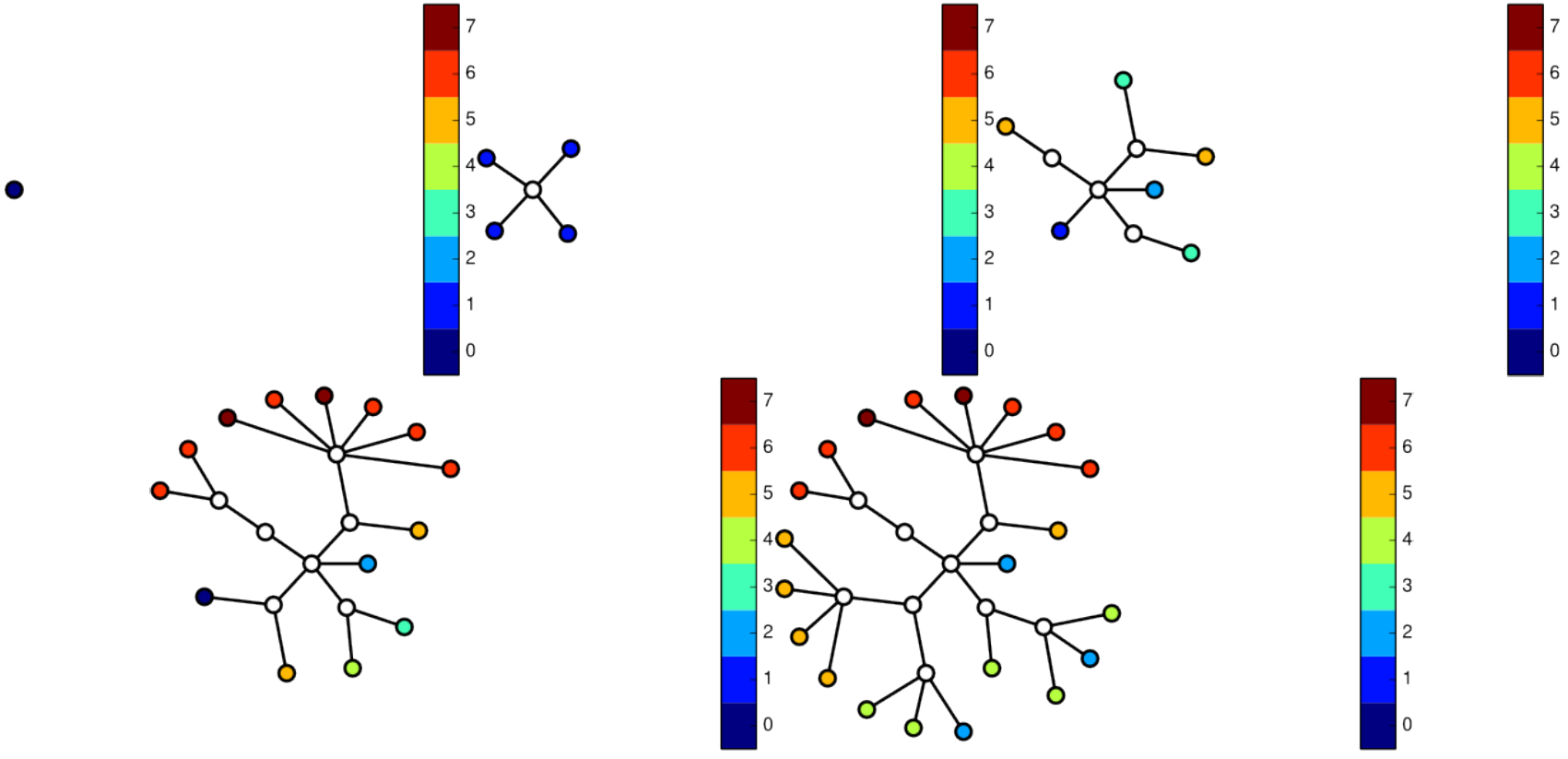


Mango tree growth

Tree Segmentation/Clustering Models

¬ Introduction – Principle

2. Characterizing tree patchiness.

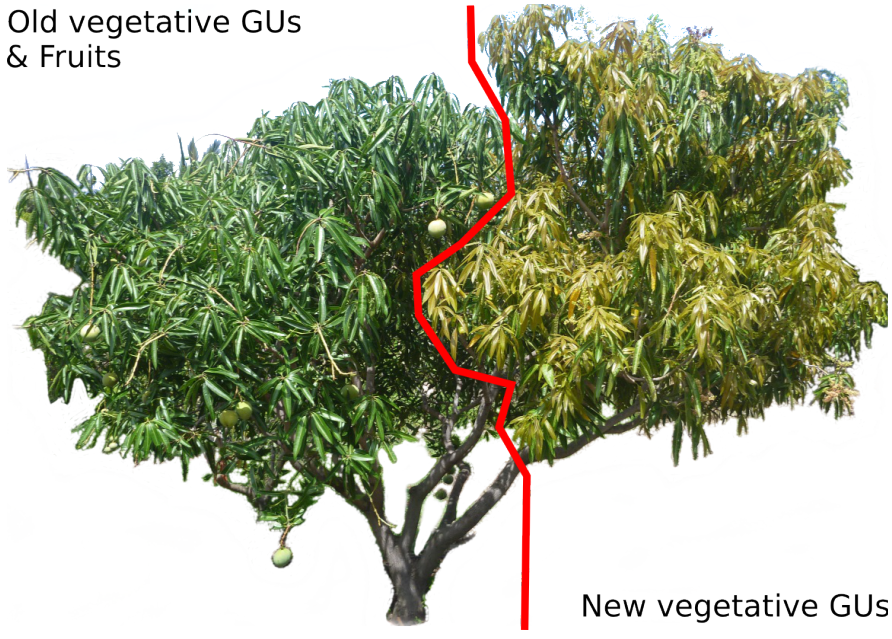


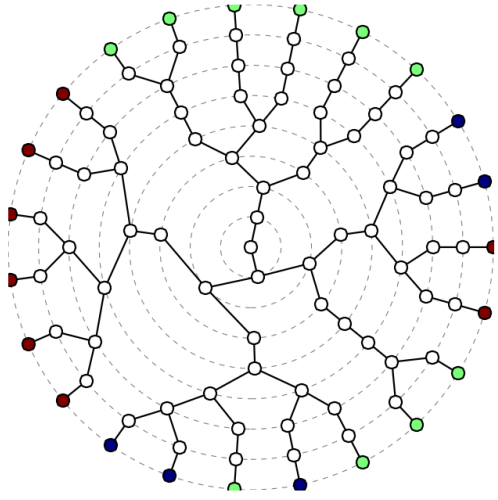
Tree-indexed data extraction from plants

Tree Segmentation/Clustering Models

¬ Introduction – Objective

2. Characterizing tree patchiness.





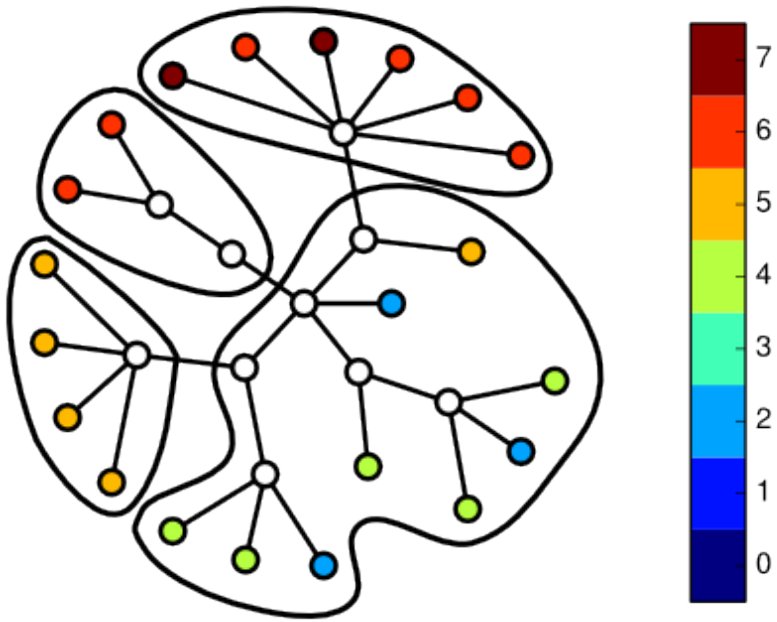
Example of state projection: partitioning tree-indexed data into homogeneous subtrees

Tree Segmentation/Clustering Models

¬ Segmentation models – Definition

A segmentation model is defined by a vertex quotienting Π such that each quotient induces a *tree*.

These quotients can also be identified by the set of their K change points, noted $\mathcal{P}$.



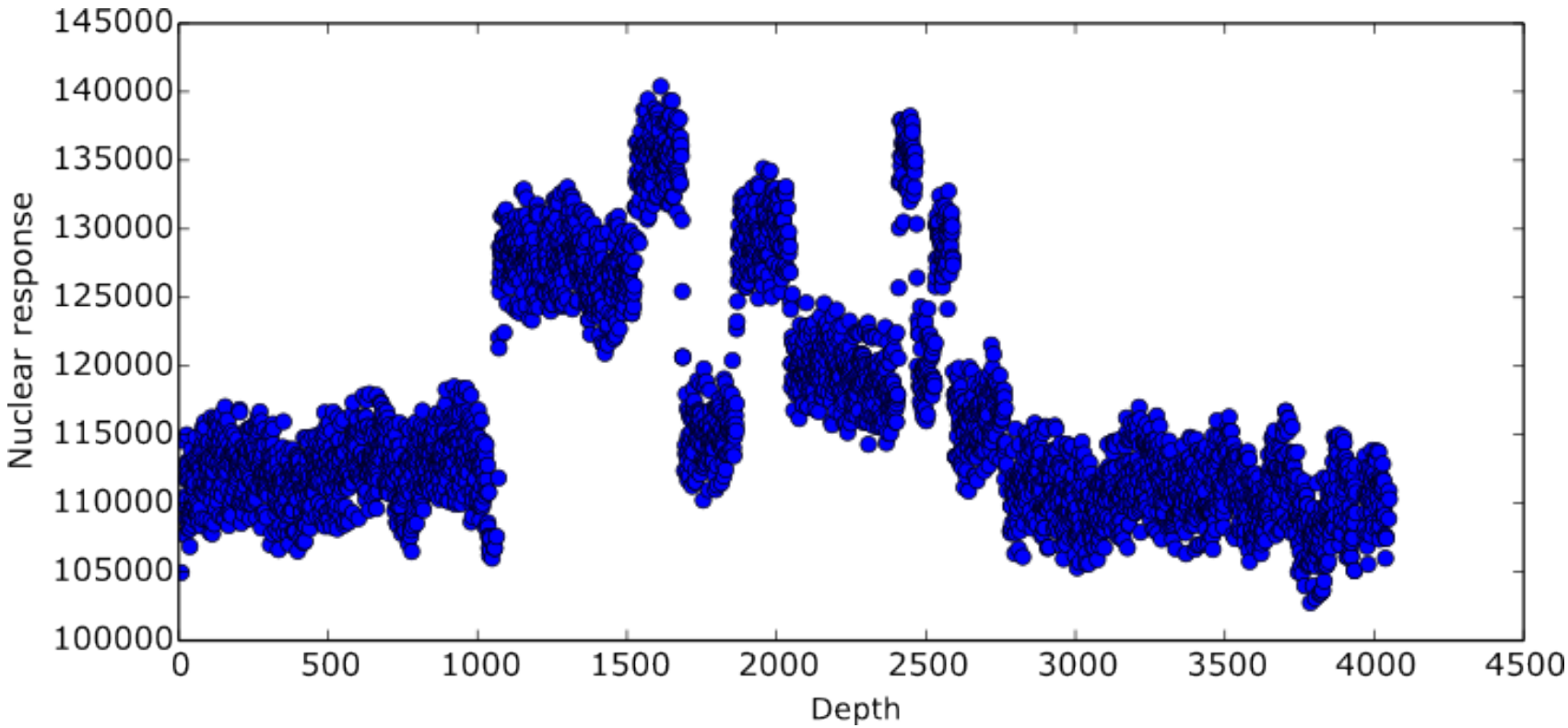
Example of segmentation problem for path-indexed data
[Fearnhead, 2006]

Tree Segmentation/Clustering Models

¬ Segmentation models – Definition

A segmentation model is defined by a vertex quotienting Π such that each quotient induces a *tree*.

These quotients can also be identified by the set of their K change points, noted $\mathcal{P}$.



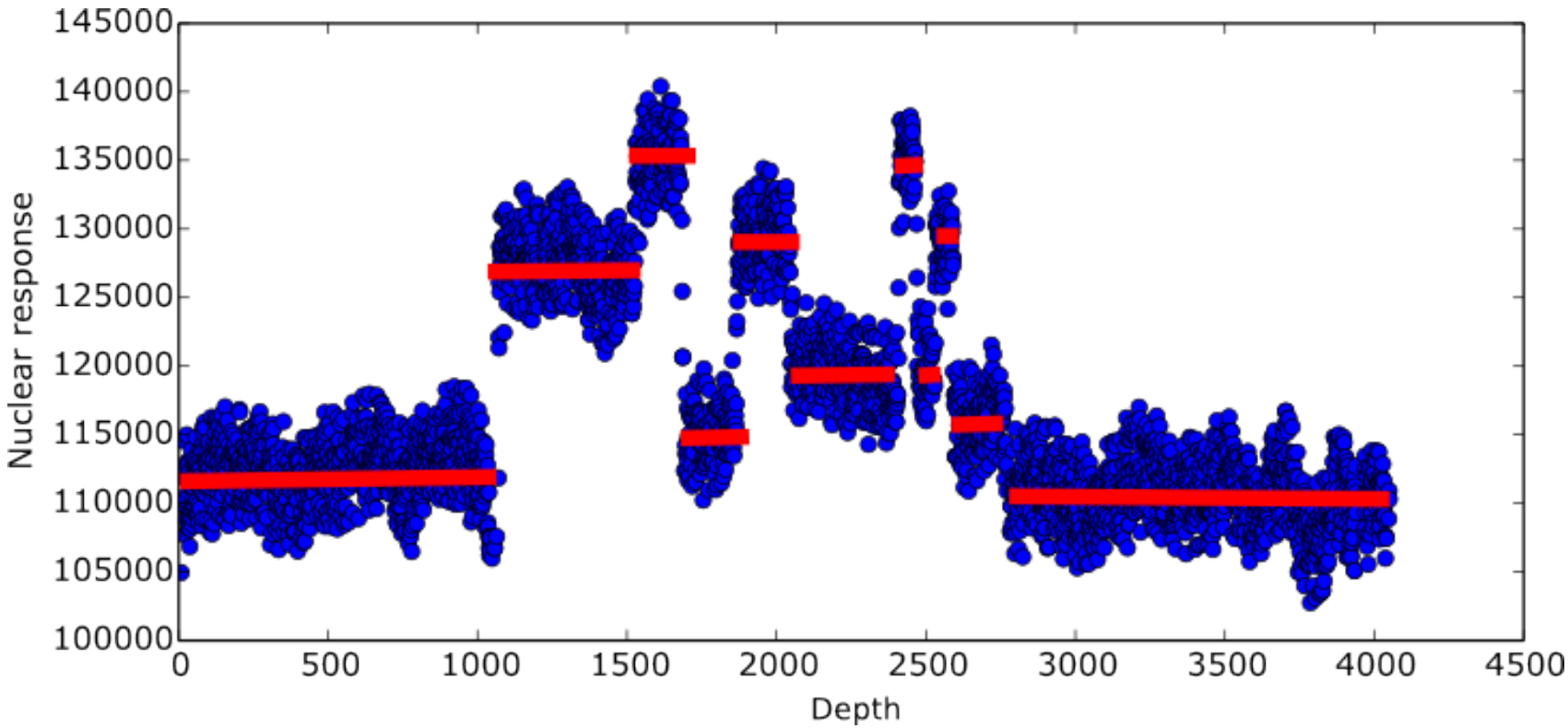
Example of segmentation problem for path-indexed data

[Fearnhead, 2006]

Tree Segmentation/Clustering Models

¬ Segmentation models – Definition

- Given the number of quotients find the best quotienting [Auger and Lawrence, 1989],
- Find the number of quotients [Baudry et al., 2012].

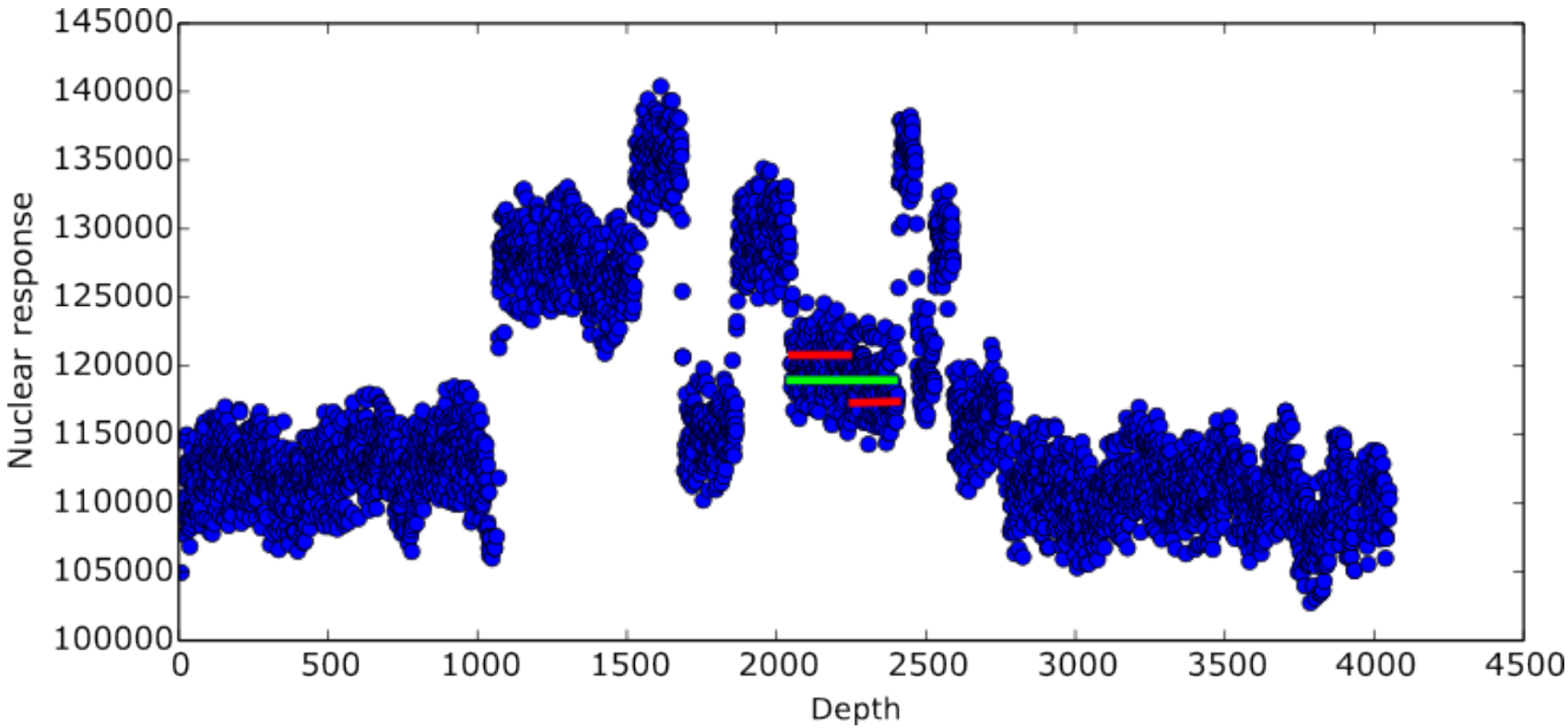


Example of segmentation problem for path-indexed data [Fearnhead, 2006]

Tree Segmentation/Clustering Models

¬ Segmentation models – Definition

- Given the number of quotients find the best quotienting [Auger and Lawrence, 1989],
- Find the number of quotients [Baudry et al., 2012].



Example of segmentation problem for path-indexed data [Fearnhead, 2006]

Tree Segmentation/Clustering Models

¬ Segmentation models – Inference

For tree-indexed data:

- Given Π, inference is a simple Maximum Likelihood inference within each quotient.
- Given K, the best quotienting cannot be found with exact methods [Hawkins, 1976].

By definition:

$$\mathcal{P}^{(0)} = \{r\},$$

and

$$\mathcal{P}^{(1)} = P^{(0)} \cup \left\{ \arg\max_{t \in \mathcal{T}} \left\{ \mathcal{L}\left(\bar{x}; \nu\left(\mathcal{P}^{(0)} \cup \{t\}\right), \theta_{\nu\left(\mathcal{P}^{(0)} \cup \{t\}\right)}\right)\right\}\right\},$$

with

- $\mathcal{L}(\bar{x}; \Pi, \theta_{\Pi})$, the log-likelihood,
- ν, the mapping quotients to change points,

is optimal.

Tree Segmentation/Clustering Models

¬ Segmentation models – Inference

For tree-indexed data:

- Given Π, inference is a simple Maximum Likelihood inference within each quotient.
- Given K, the best quotienting cannot be found with exact methods [Hawkins, 1976].

By definition:

$$\mathcal{P}^{(0)} = \{r\},$$

and

$$\mathcal{P}^{(1)} = P^{(0)} \cup \left\{ \arg\max_{t \in \mathcal{T}} \left\{ \mathcal{L}\left(\bar{x}; \nu\left(\mathcal{P}^{(0)} \cup \{t\}\right), \theta_{\nu\left(\mathcal{P}^{(0)} \cup \{t\}\right)}\right) \right\} \right\},$$

with

- $\mathcal{L}(\bar{x}; \Pi, \theta_\Pi)$, the log-likelihood,
- ν, the mapping quotients to change points,

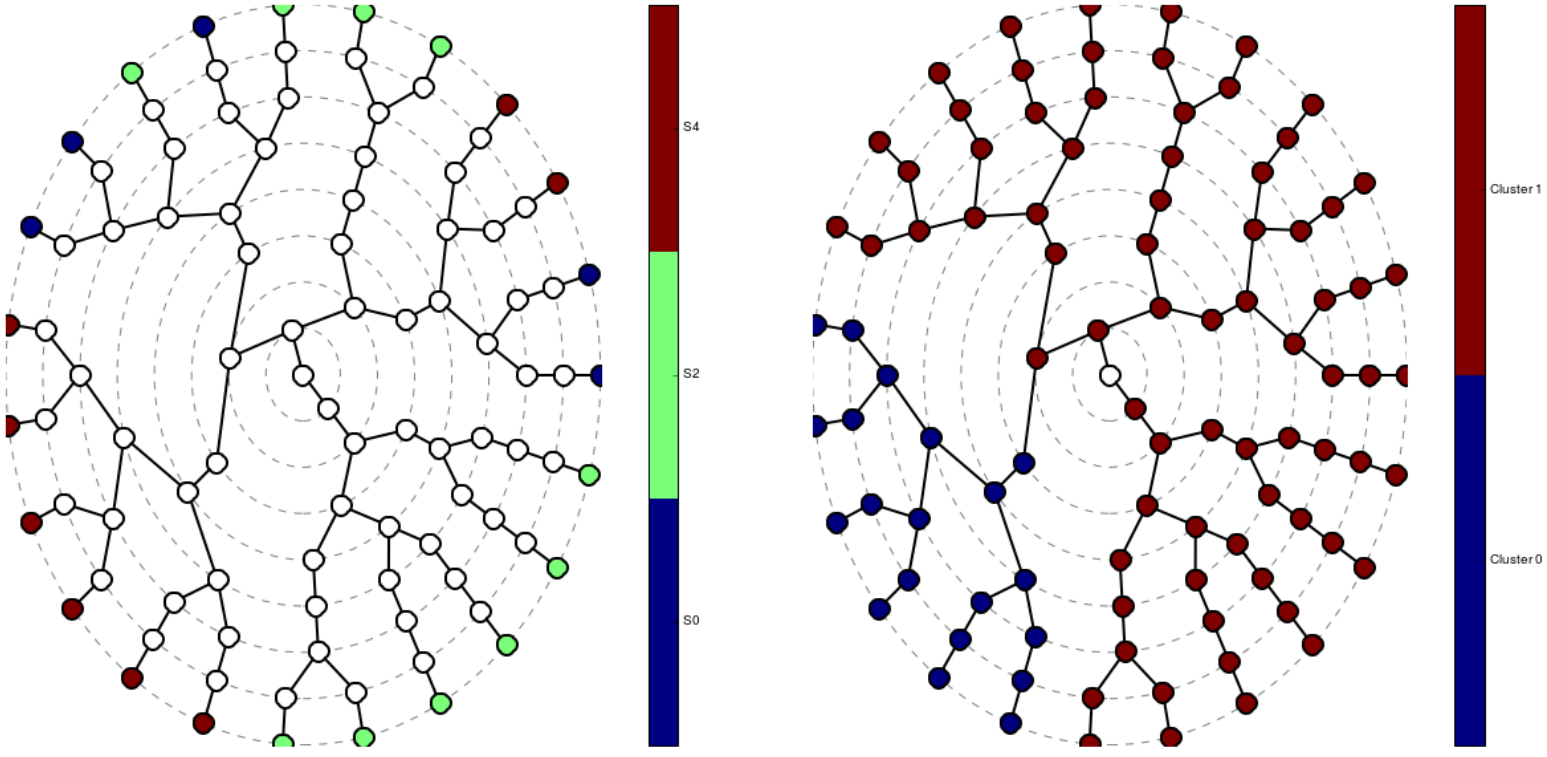
is optimal.

Tree Segmentation/Clustering Models

¬ Segmentation models – Inference

A split approach:

$$\mathcal{P}^{(k)} = P^{(k-1)} \cup \left\{ \arg\max_{t \in \mathcal{T}} \left\{ \mathcal{L}\left(\bar{x}; \nu\left(\mathcal{P}^{(k-1)} \cup \{t\}\right), \theta_{\nu\left(\mathcal{P}^{(k-1)} \cup \{t\}\right)}\right)\right\}\right\},$$



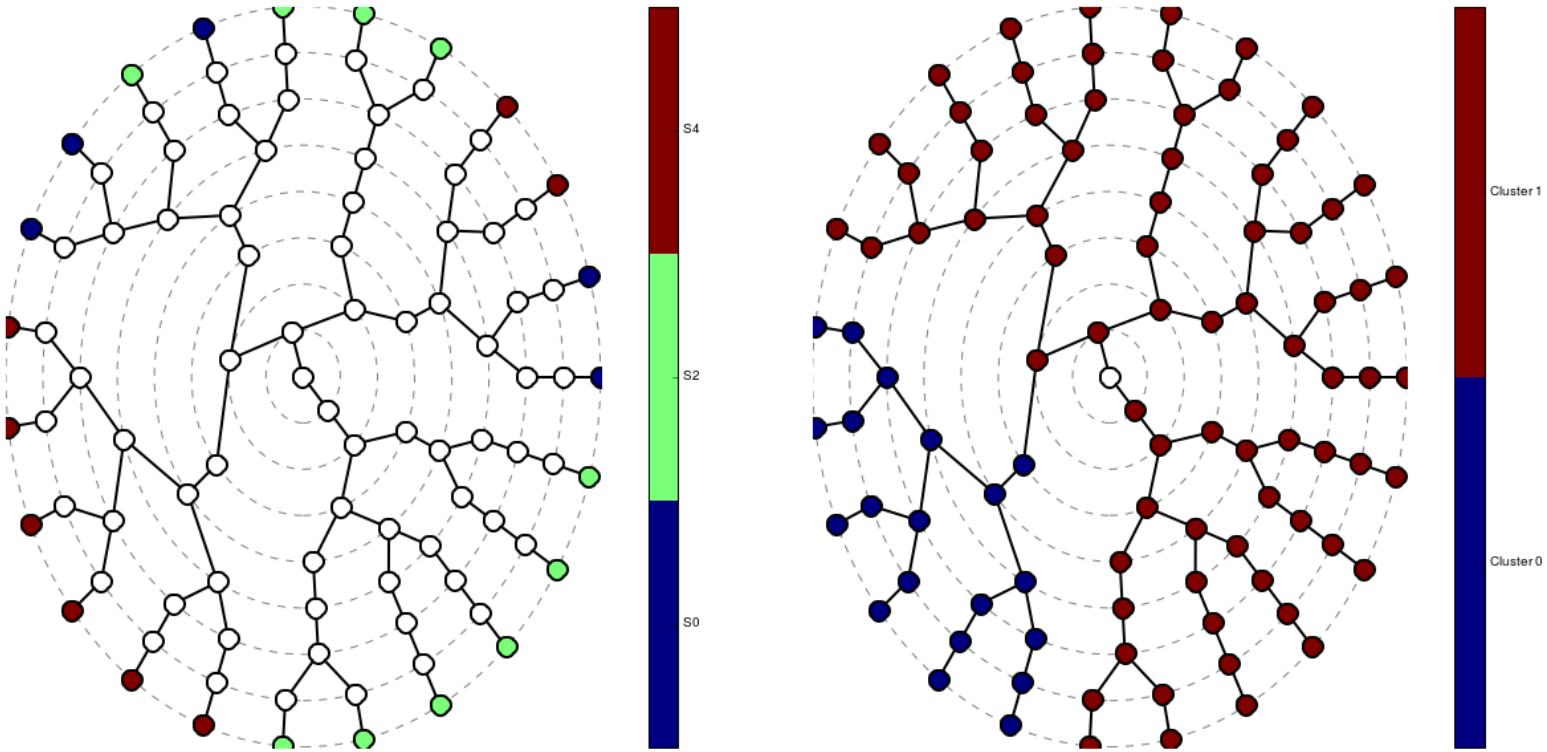
Example of the split approach for segmenting trees

Tree Segmentation/Clustering Models

¬ Segmentation models – Inference

A merge approach:

$$\mathcal{P}^{(k-1)} = P^{(k)} \setminus \left\{ \underset{t \in \mathcal{P}^{(\|)}}{\arg\max} \left\{ \mathcal{L}\left(\bar{x}; \nu\left(\mathcal{P}^{(k)} \setminus \{t\}\right), \theta_{\nu\left(\mathcal{P}^{(k)} \setminus \{t\}\right)}\right)\right\}\right\},$$

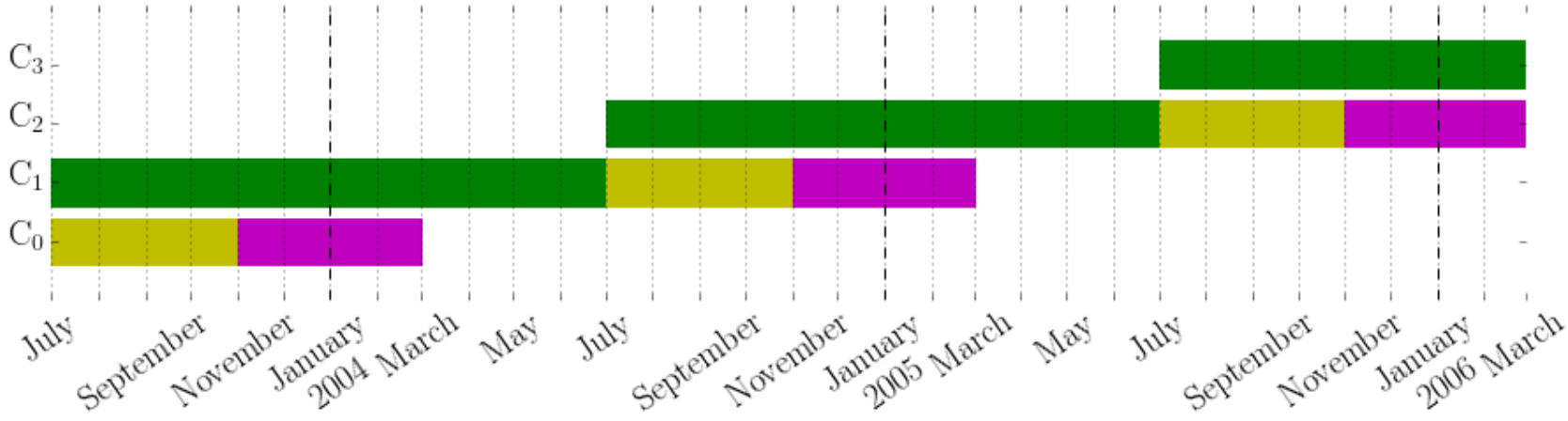


Example of the split-merge approach for segmenting trees

Tree Segmentation/Clustering Models

¬ Segmentation models – Application

2. Characterizing tree patchiness.



Mango tree Growth Cycle (GC)

Early flush. Period where the vegetative phase of a GC overlaps the flowering phase of the previous cycle.

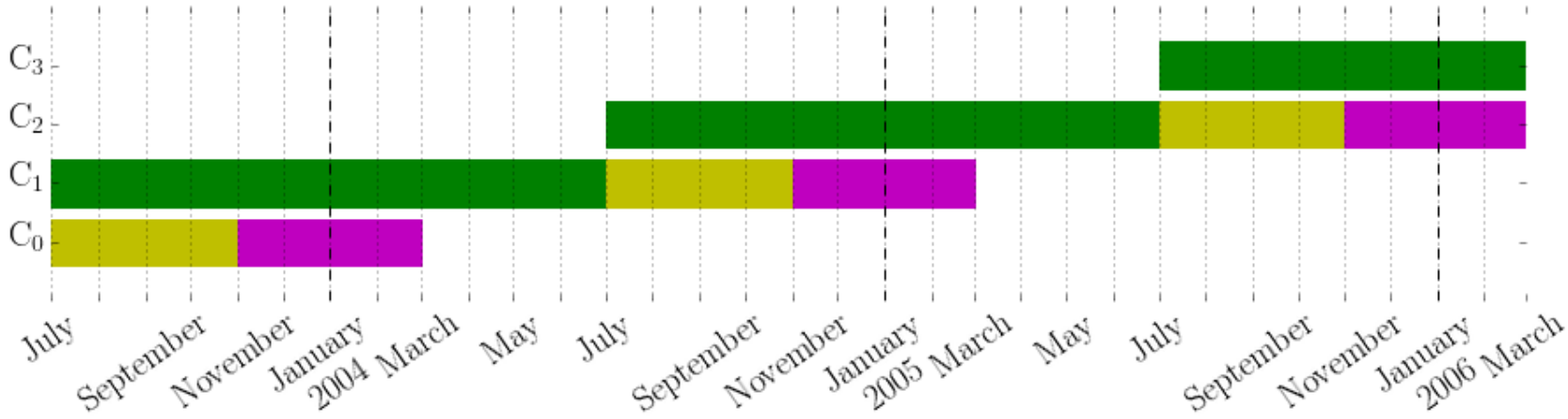
Intermediate flush. Period where the vegetative phase of a GC overlaps the fructifying phase of the previous cycle.

Late flush. Period where the vegetative phase of a GC does not overlap the previous or the next cycles.

Tree Segmentation/Clustering Models

¬ Segmentation models – Application

2. Characterizing tree patchiness.



Mango tree Growth Cycle (GC)

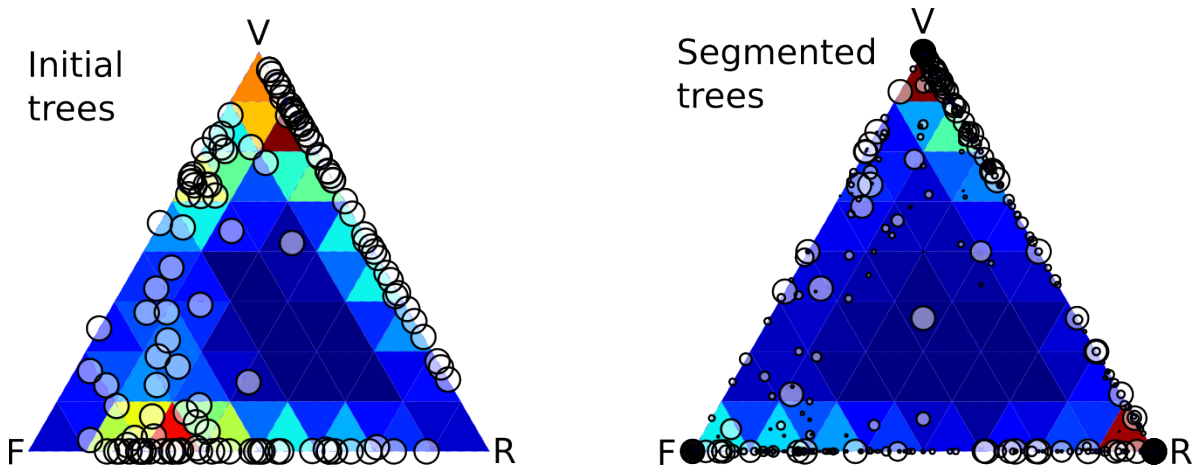
Consider a snapshot of a mango tree for each flush. As a consequence we obtained 181 trees within which mostly leaf vertices were observed with the following observation space

$$\mathcal{X} = \{V, F, R\},$$

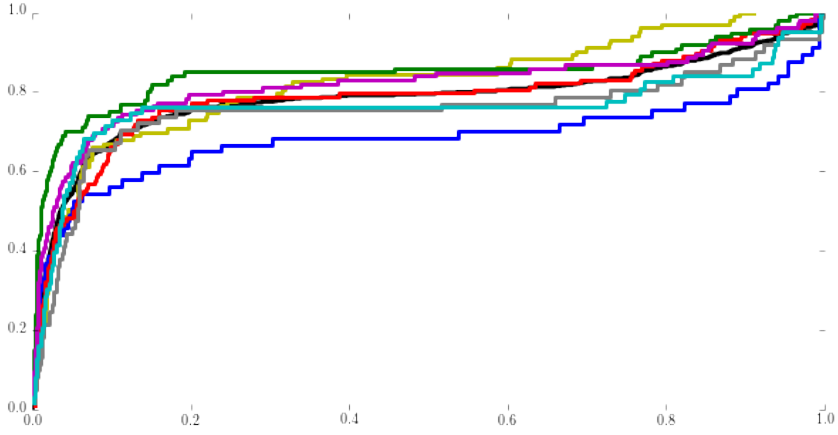
for Flowering, Resting and Vegetative.

Tree Segmentation/Clustering Models

¬ Segmentation models – Application



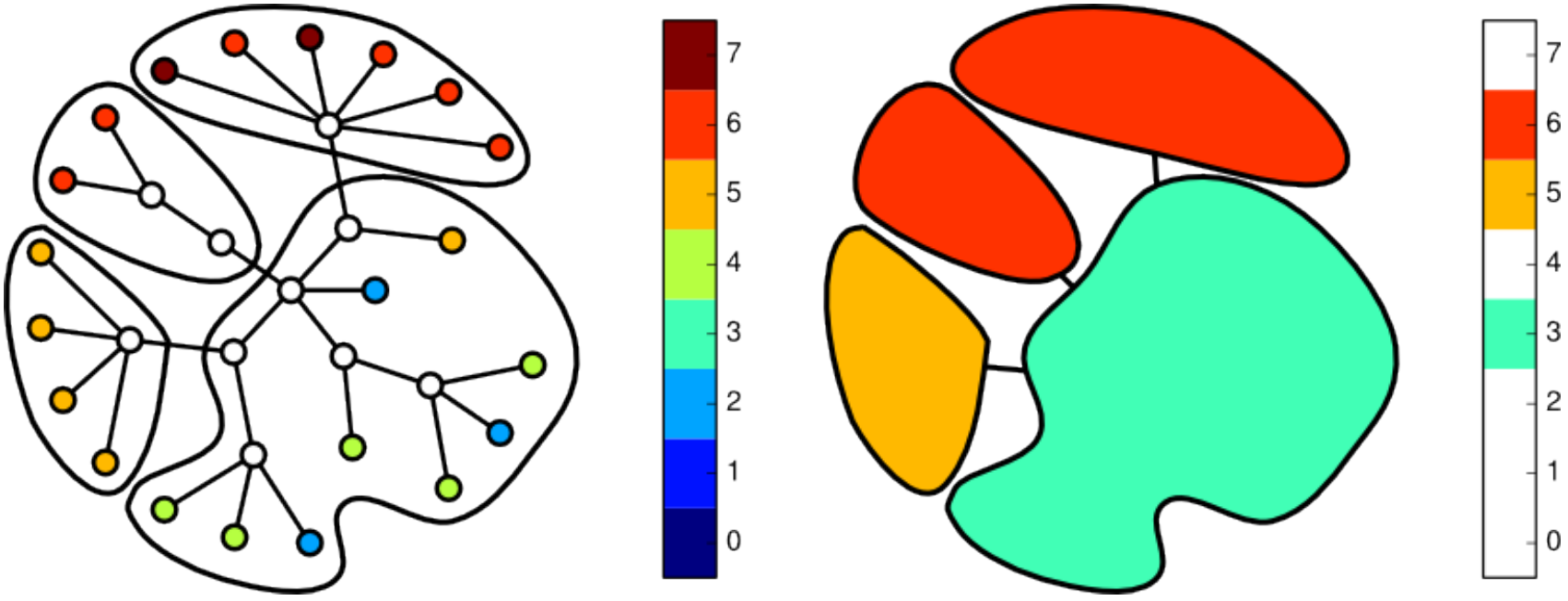
Segmentation of mango trees



Relative size of patches

Tree Segmentation/Clustering Models

¬ Clustering models – Definition

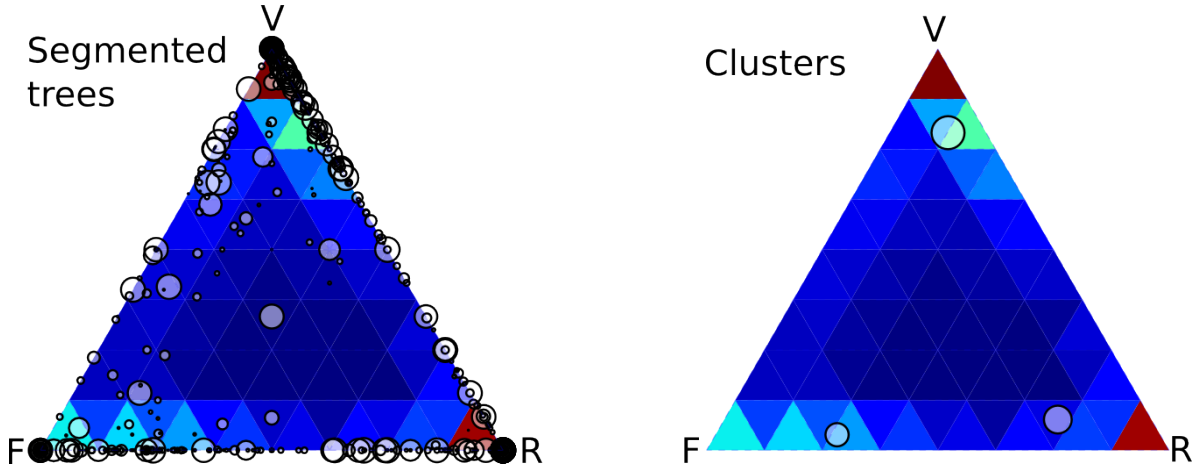


A mixture model for sub-tree clustering

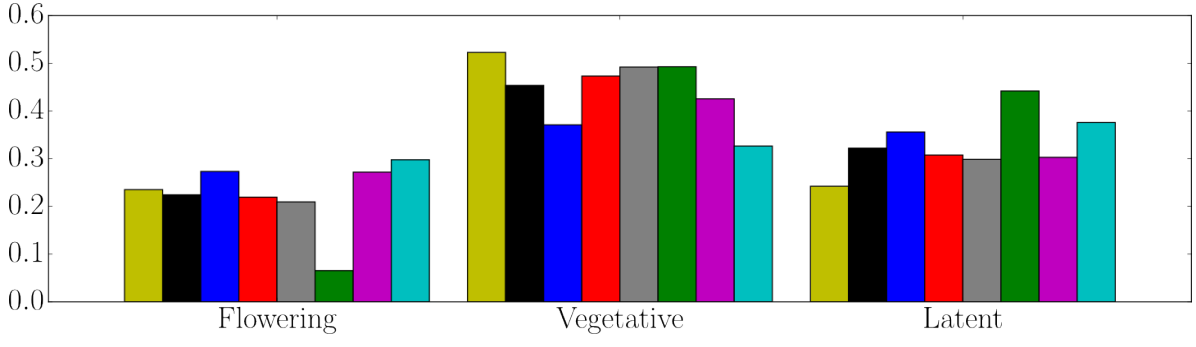
Using EM algorithm and MAP (Maximum A Posteriori) assignment of quotients of standard mixture models [McLachlan and Peel, 2000] such that vertices in same quotient are assigned to the same component [Picard et al., 2005].

Tree Segmentation/Clustering Models

¬ Clustering models – Definition



Clustering of segmented mango trees



MAP assignement

Tree Segmentation/Clustering Models

Statistical modeling framework:

Homogeneous zone detection in order to:

2. Quantifying tree patchiness.

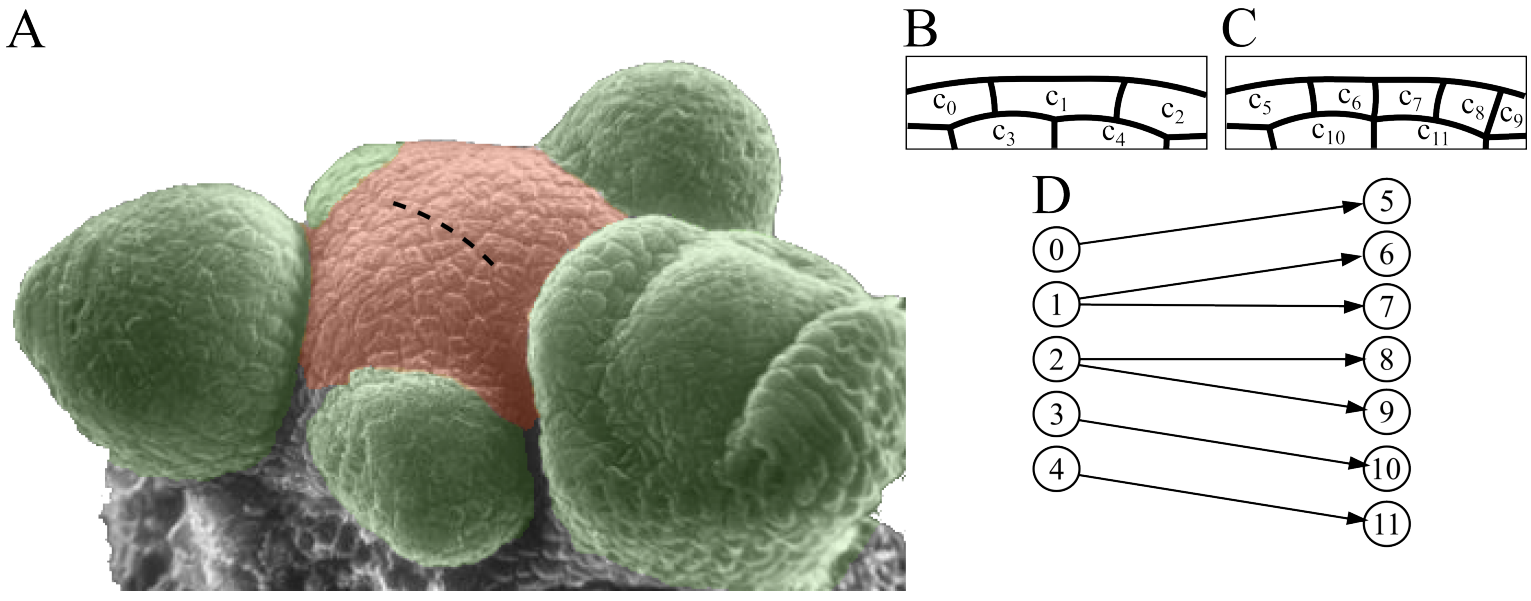
$\Rightarrow$ Identification and characterization but not really quantification but could us tree distances [Ferraro et al., 2003].

And at the microscopic scale...

¬ Particularities

Virtual Plants focus on plant development and its modulation by environmental and genetic factors:

1. At a microscopic scale. Each vertex represents a cell and edges encode either the tracking of a cell throughout time or the lineage relationships among parent and child cells.



Tree-indexed data extraction from cell lineages

And at the microscopic scale...

¬ Particularities

Tree-indexed data
at macroscopic scale:

- General tree

$$\forall t \in \mathcal{T},\ |\mathsf{ch}(t)| \in \mathbb{N}.$$

- Categorical outcomes,
  - Fate.
  - Fate $\times$ Burst.

at macroscopic scale:

- Binary tree

$$\forall t \in \mathcal{T},\ |\mathsf{ch}(t)| \in \{1, 2\}.$$

- Multivariate outcomes,
  - Volume.
  - Surfaces.
  - Curvatures.

And at the microscopic scale...

¬ Particularities

Tree-indexed data
at macroscopic scale:

- General tree

$$\forall t \in \mathcal{T}, |\mathrm{ch}(t)| \in \mathbb{N}.$$

- Categorical outcomes,
  - Fate.
  - Fate $\times$ Burst.

at macroscopic scale:

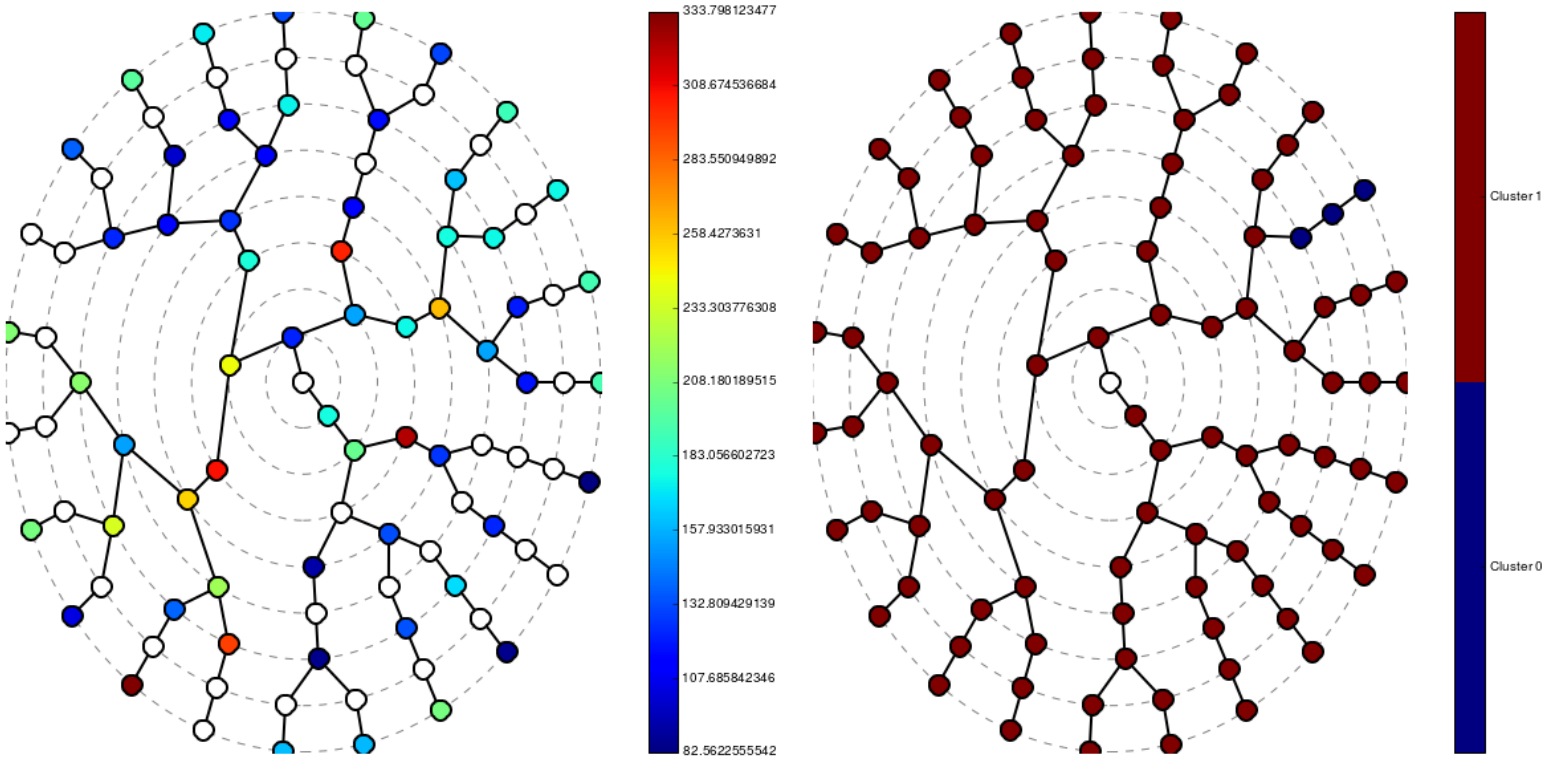
- Binary tree

$$\forall t \in \mathcal{T}, |\mathrm{ch}(t)| \in \{1, 2\}.$$

- Multivariate outcomes,
  - Volume.
  - Surfaces.
  - Curvatures.

And at the microscopic scale...

¬ Tree segmentation/clustering models



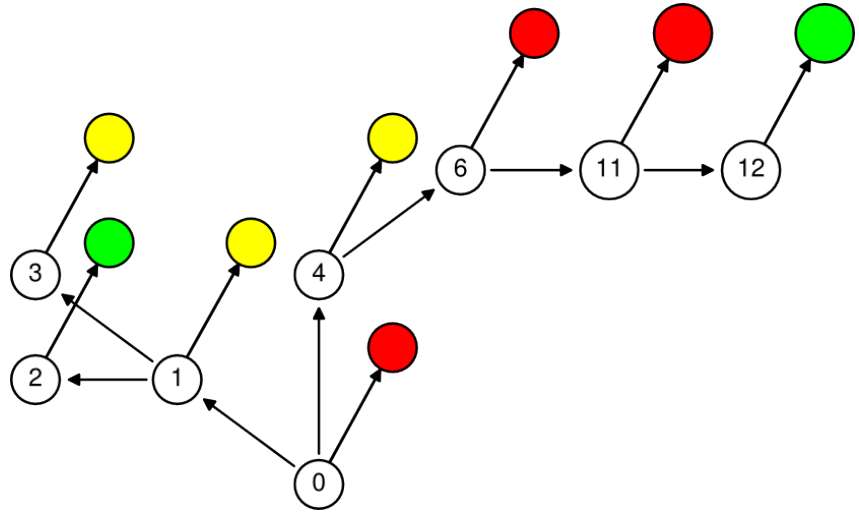
Segmentation of cell lineages considering the volume of cells

Only the ML inference of parameters given the quotienting differs

And at the microscopic scale...

We considered the MTBP [Haccou et al., 2005] since $\mathcal{X} \subset \mathbb{N}$, but here $\mathcal{X} \subset \mathbb{R}^5$.
We introduced the hidden MTBP inspired from [Durand et al., 2004]

- Smoothing algorithm.
- Viterbi algorithm.



Hidden Markov tree model

And at the microscopic scale...

¬ Markov tree models– Parametrization of generation distributions

Number of states	Number of parameters for the non-parametric case	Number of parameters for the Poisson worst case
2	**19**	**11**
3	**59**	**29**
4	**139**	**59**

Number of parameters in non-parametric and worst case Poisson multi-type branching processes as a function of the number of states

Conclusion

- Used MTBPs and defined HMTBPs in order to detect motifs in tree-indexed data with various type of random variables or random vectors.
- Extended parametric DAG models and inference algorithm for MAG models in order to have parsimonious generation distributions in MTBPs and HMTBPs.
- Generalized the multiple change-point model from path-indexed data to tree-indexed data in order to detect homogeneous zones in tree-indexed data.
- Illustrate this framework using two different tree-indexed data types and provided implementation of methods developed in order to make them available to team members and partners.

Conclusion

- Used MTBPs and defined HMTBPs in order to detect motifs in tree-indexed data with various type of random variables or random vectors.
- Extended parametric DAG models and inference algorithm for MAG models in order to have parsimonious generation distributions in MTBPs and HMTBPs.
- Generalized the multiple change-point model from path-indexed data to tree-indexed data in order to detect homogeneous zones in tree-indexed data.
- Illustrate this framework using two different tree-indexed data types and provided implementation of methods developed in order to make them available to team members and partners.

Conclusion

- Used MTBPs and defined HMTBPs in order to detect motifs in tree-indexed data with various type of random variables or random vectors.
- Extended parametric DAG models and inference algorithm for MAG models in order to have parsimonious generation distributions in MTBPs and HMTBPs.
- Generalized the multiple change-point model from path-indexed data to tree-indexed data in order to detect homogeneous zones in tree-indexed data.
- Illustrate this framework using two different tree-indexed data types and provided implementation of methods developed in order to make them available to team members and partners.

Conclusion

- Used MTBPs and defined HMTBPs in order to detect motifs in tree-indexed data with various type of random variables or random vectors.
- Extended parametric DAG models and inference algorithm for MAG models in order to have parsimonious generation distributions in MTBPs and HMTBPs.
- Generalized the multiple change-point model from path-indexed data to tree-indexed data in order to detect homogeneous zones in tree-indexed data.
- Illustrate this framework using two different tree-indexed data types and provided implementation of methods developed in order to make them available to team members and partners.

Perspectives & work in progress

- Generalize the MAG models to continuous multivariate distributions. Under a Gaussian hypothesis, constraints imposed by discrete parametric models could be relaxed: the algorithm can produce results not reachable using only UG or DAG ones combining the Local search and lasso estimators.
- We focused on HMTBPs in this thesis but contrarily to sequences, directed tree are non-symmetrical structures the HMIT models could be studied and results compared.
- Introducing particular mixture of discrete multivariate distributions in order to relax the *soft competition hypothesis*.

Perspectives & work in progress

- Generalize the MAG models to continuous multivariate distributions. Under a Gaussian hypothesis, constraints imposed by discrete parametric models could be relaxed: the algorithm can produce results not reachable using only UG or DAG ones combining the Local search and lasso estimators.
- We focused on HMTBPs in this thesis but contrarily to sequences, directed tree are non-symmetrical structures the HMIT models could be studied and results compared.
- Introducing particular mixture of discrete multivariate distributions in order to relax the *soft competition hypothesis*.

Perspectives & work in progress

- Generalize the MAG models to continuous multivariate distributions. Under a Gaussian hypothesis, constraints imposed by discrete parametric models could be relaxed: the algorithm can produce results not reachable using only UG or DAG ones combining the Local search and lasso estimators.
- We focused on HMTBPs in this thesis but contrarily to sequences, directed tree are non-symmetrical structures the HMIT models could be studied and results compared.
- Introducing particular mixture of discrete multivariate distributions in order to relax the *soft competition hypothesis*.


This is the end...


Auger, I. E. and Lawrence, C. E. (1989).
Algorithms for the optimal identification of segment neighborhoods.
Bulletin of mathematical biology, 51(1):39–54.


Baudry, J.-P., Maugis, C., and Michel, B. (2012).
Slope heuristics: overview and implementation.
Statistics and Computing, 22(2):455–470.


Chacko, E. (1986).
Physiology of vegetative and reproductive growth in mango (*Mangifera indica* L.) trees.
In *Proceedings of the First Australian Mango Research Workshop*, volume 1, pages 54–70. CSIRO Australia, Melbourne.


Dambreville, A. (2012).
Croissance et développement du manguier (Mangifera indica L.) in natura: approche expérimentale et modélisation de l'influence d'un facteur exogène, la température, et de facteurs endogènes architecturaux.
PhD thesis, Université Montpellier II-Sciences et Techniques du Languedoc.


Dambreville, A., Lauri, P.-É., Trottier, C., Guédon, Y., and Normand, F. (2013).
Deciphering structural and temporal interplays during the architectural development of mango trees.
Journal of experimental botany, 64(8):2467–2480.


Durand, J.-B., Goncalvès, P., and Guédon, Y. (2004).
Computational methods for hidden Markov tree models–An application to wavelet trees.
IEEE Transactions on Signal Processing, 52(9):2551–2560.


Fearnhead, P. (2006).
Exact and efficient bayesian inference for multiple changepoint problems.
Statistics and computing, 16(2):203–213.


Ferraro, P., Godin, C., et al. (2003).
An edit distance between quotiented trees.
Algorithmica, 36(1):1–39.


Haccou, P., Jagers, P., and Vatutin, V. A. (2005).
Branching processes: variation, growth, and extinction of populations.
Cambridge University Press.


Hawkins, D. M. (1976).
Point estimation of the parameters of piecewise regression models.
Applied Statistics, 25(1):51–57.


Johnson, N., Kemp, A., and Kotz, S. (1993).
Univariate discrete distributions.
Wiley-Interscience.


Johnson, N., Kotz, S., and Balakrishnan, N. (1997).
Discrete multivariate distributions.
Wiley New York.


Koller, D. and Friedman, N. (2009).
Probabilistic graphical models: principles and techniques.
MIT press.


Lauritzen, S. (1996).
Graphical models, volume 17.
Oxford University Press.


McLachlan, G. and Peel, D. (2000).
Finite mixture models.
Wiley New York.


Picard, F., Robin, S., Lavielle, M., Vaisse, C., and Daudin, J.-J. (2005).
A statistical approach for array CGH data analysis.
BMC bioinformatics, 6(1):27.


Ramírez, F. and Davenport, T. L. (2010).
Mango (*Mangifera indica* L.) flowering physiology.
Scientia Horticulturae, 126(2):65–72.