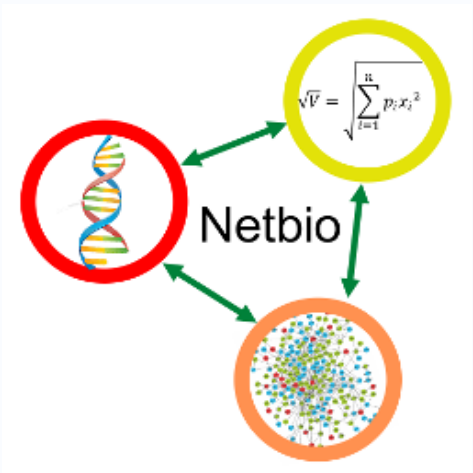
Stochastic Block Model for taxonomic identification in (meta)Barcoding

Mohamed Anwar ABOUABDALLAH[1]
Directed by : Nathalie Peyrard[2] Alain Franc[1] Olivier Coulaud[3]

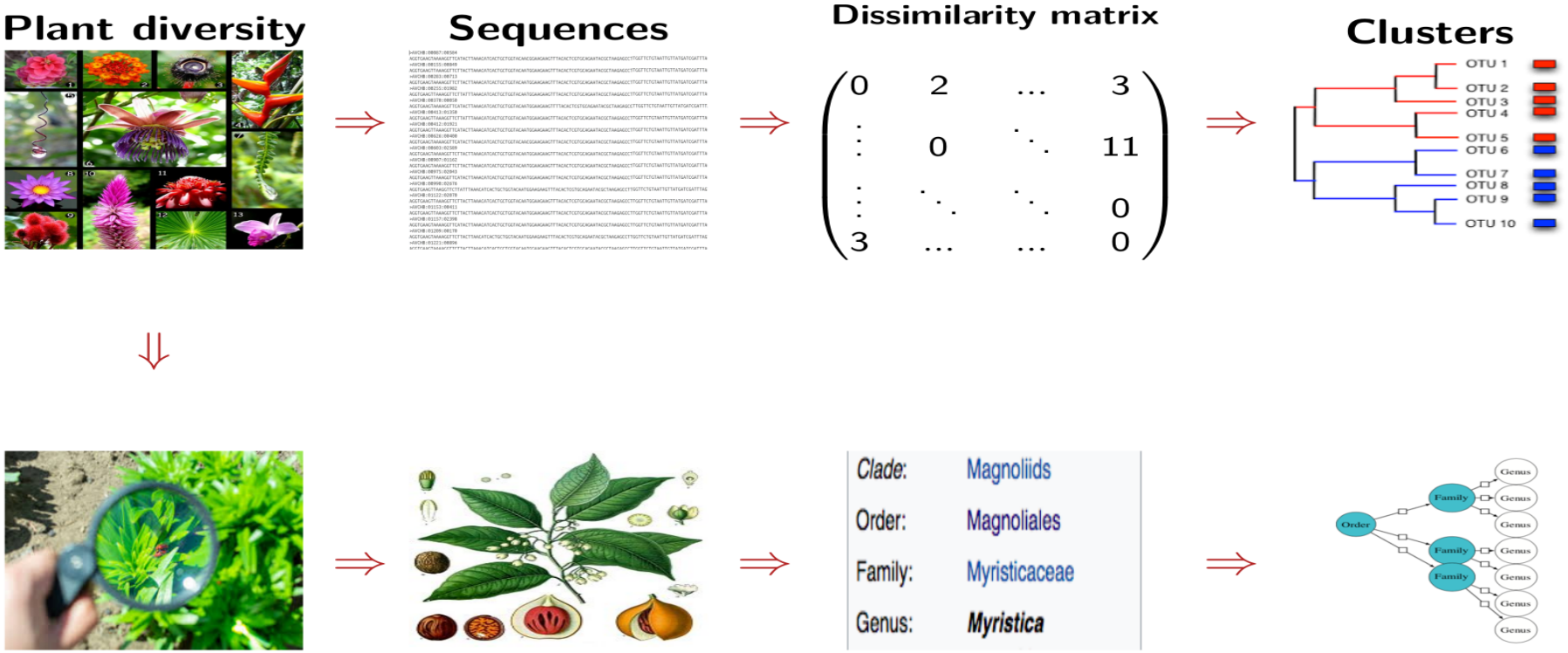
[1] INRAE, UMR BioGeCo, Pierroton & EPC INRAE / Inria Pleiade, Talence, France

[2] INRAE, Unité MIAT, Toulouse, France

[3] Inria, HiePACS, Talence, France



Scientific context



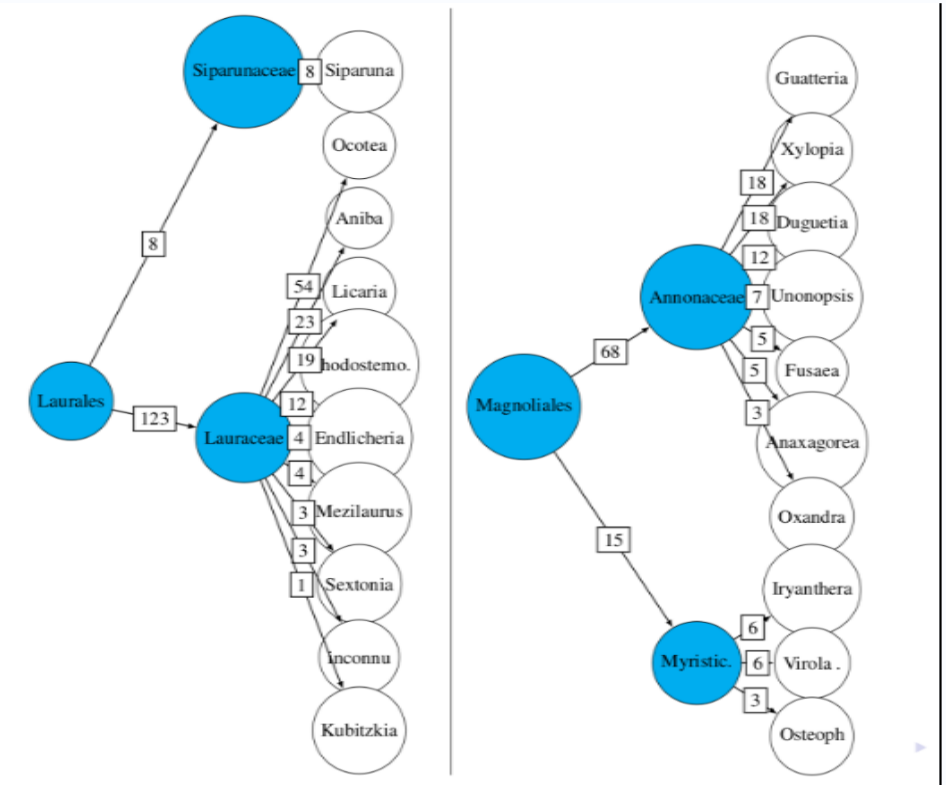
- **Is there an agreement between botanical and molecular classifications on a coarse taxonomic level ?**
- **How to improve SBM model accuracy ?**

Summary

Agreement between botanical and molecular classifications

Data set

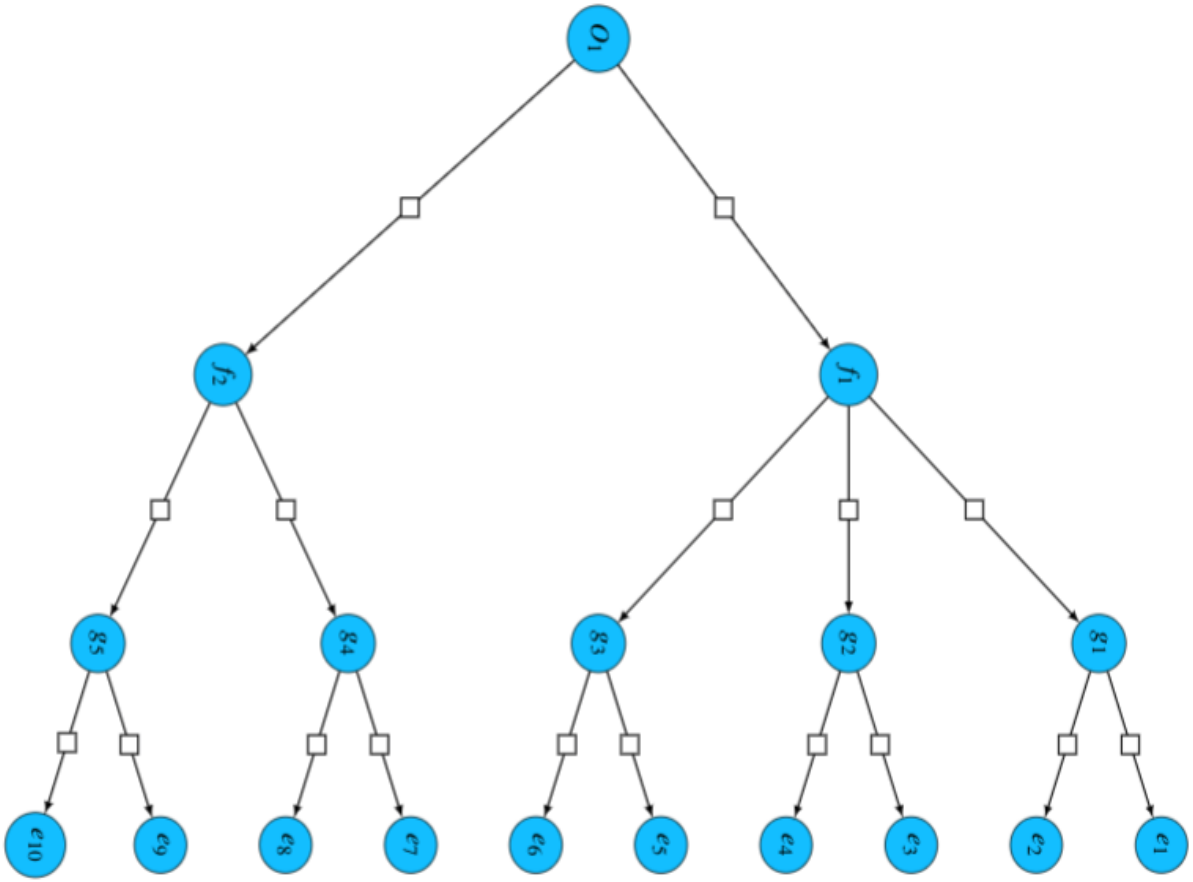
- **1458 trees from an experimental plot in French Guyana.**



Order, family, genus and species of each individual.

```
>AVCHB:00087:00584
AGGTGAAGTAAAAGGTTCATACTTAAACATCACTGCTGGTACAACGGAAGAAGTTTACACTCGTGCAGAATACGCTAAGAGCCTTGGTTCTGTAATTGTTATGATCGATTTA
>AVCHB:00155:00849
AGGTGAAGTTAAAGGTTCTTACTTAAACATCACTGCTGGTACAATGGAAGAAGTTTACACTCGTGCAGAATACGCTAAGAGCCTTGGTTCTGTAATTGTTATGATCGATTTA
>AVCHB:00283:00713
AGGTGAAGTAAAAGGTTCTTACTTAAACATCACTGCTGGTACAATGGAAGAAGTTTACACTCGTGCAGAATACGCTAAGAGCCTTGGTTCTGTAATTGTTATGATCGATTTA
>AVCHB:00255:01982
AGGTGAAGTTAAAGGTTCTTATTTAAACATCACTGCTGGTACAATGGAAGAAGTTTACACTCGTGCAGAATACGCTAAGAGCCTTGGTTCTGTAATTGTTATGATCGATTTA
>AVCHB:00370:00050
AGGTGAAGTAAAAGGTTCATACTTAAACATCACTGCTGGTACAATGGAAGAAGTTTTACACTCGTGCAGAATACGCTAAGAGCCTTGGTTCTGTAATTGTTATGATCGATTT
>AVCHB:00413:01350
AGGTGAAGTTAAAGGTTCTTATTTAAACATCACTGCTGGTACAATGGAAGAAGTTTACACTCGTGCAGAATACGCTAAGAGCCTTGGTTCTGTAATTGTTATGATCGATTTA
>AVCHB:00412:01921
AGGTGAAGTTAAAGGTTCATACTTAAACATCACTGCTGGTACAATGGAAGAAGTTTACACTCGTGCAGAATACGCTAAGAGCCTTGGTTCTGTAATTGTTATGATCGATTTA
>AVCHB:00626:00400
AGGTGAAGTAAAAGGTTCATACTTAAACATCACTGCTGGTACAACGGAAGAAGTTTACACTCGTGCAGAATACGCTAAGAGCCTTGGTTCTGTAATTGTTATGATCGATTTA
>AVCHB:00603:02589
AGGTGAAGTAAAAGGTTCTTACTTAAACATCACTGCTGGTACAATGGAAGAAGTTTACACTCGTGCAGAATACGCTAAGAGCCTTGGTTCTGTAATTGTTATGATCGATTTA
>AVCHB:00907:01162
AGGTGAAGTAAAAGGTTCTTACTTAAACATCACTGCTGGTACAATGGAAGAAGTTTACACTCGTGCAGAATACGCTAAGAGCCTTGGTTCTGTAATTGTTATGATCGATTTA
>AVCHB:00975:02043
AGGTGAAGTTAAAGGTTCATACTTAAACATCACTGCTGGTACAATGGAAGAAGTTTACACTCGTGCAGAATACGCTAAGAGCCTTGGTTCTGTAATTGTTATGATCGATTTA
>AVCHB:00990:02676
AGGTGAAGTTAAGGTTCTTATTTAAACATCACTGCTGGTACAATGGAAGAAGTTTACACTCGTGCAGAATACGCTAAGAGCCTTGGTTCTGTAATTGTTATGATCGATTTAG
>AVCHB:01122:02070
AGGTGAAGTTAAAGGTTCTTACTTAAACATCACTGCTGGTACAATGGAAGAAGTTTACACTCGTGCAGAATACGCTAAGAGCCTTGGTTCTGTAATTGTTATGATCGATTTA
>AVCHB:01153:00411
AGGTGAAGTTAAAGGTTCTTACTTAAACATCACTGCTGGTACAATGGAAGAAGTTTACACTCGTGCAGAATACGCTAAGAGCCTTGGTTCTGTAATTGTTATGATCGATTTA
>AVCHB:01157:02398
AGGTGAAGTAAAAGGTTCATACTTAAACATCACTGCTGGTACAATGGAAGAAGTTTACACTCGTGCAGAATACGCTAAGAGCCTTGGTTCTGTAATTGTTATGATCGATTTA
>AVCHB:01209:00170
AGGTGAAGTAAAAGGTTCTTACTTAACATCACTGCTGGTACAATGGAAGAAGTTTACACTCGTGCAGAATACGCTAAGAGCCTTGGTTCTGTAATTGTTATGATCGATTTAG
>AVCHB:01221:00896
```

DNA sequence of each individual.



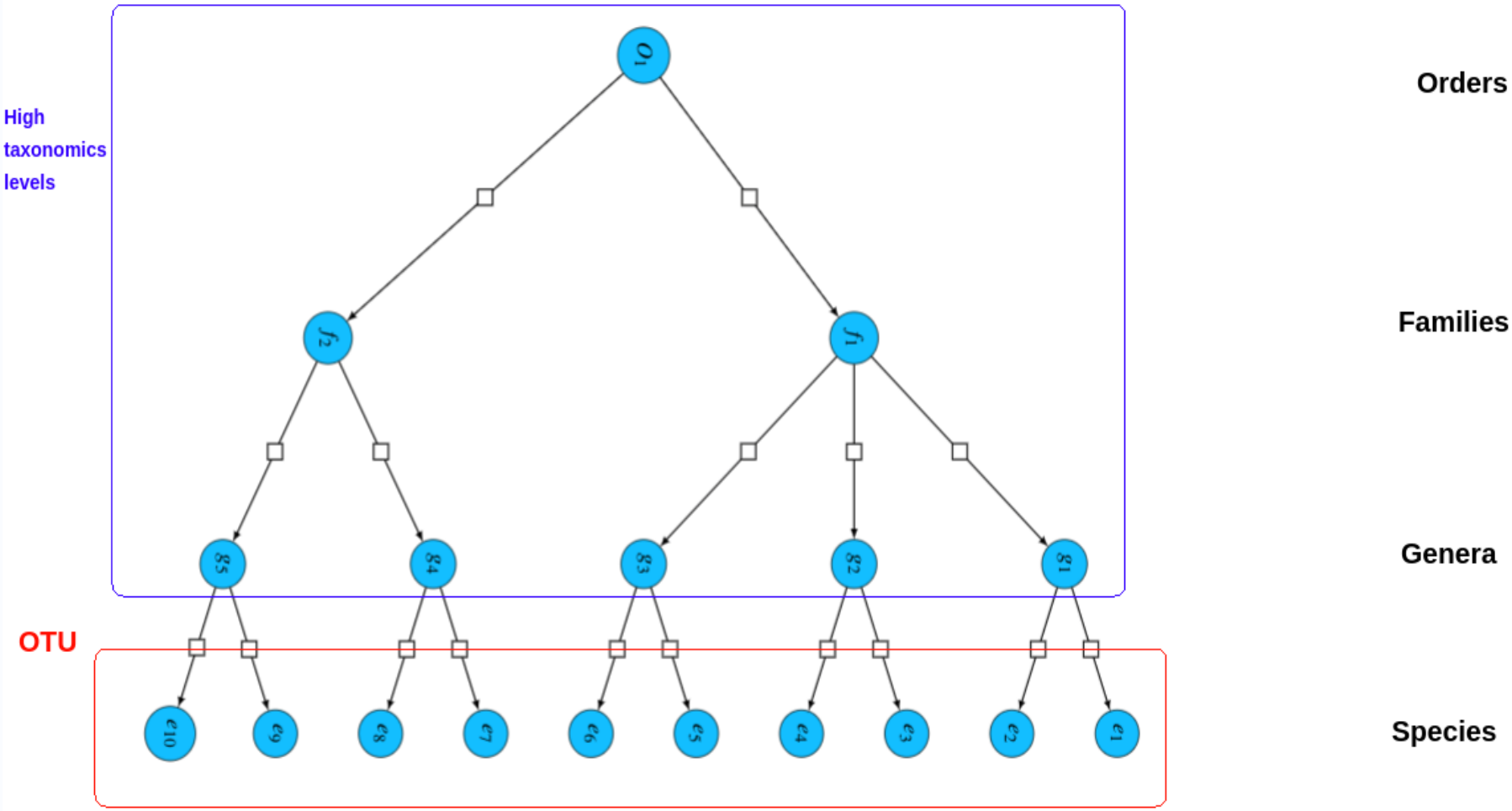
Orders

Families

Genera

OTU

Species


High
taxonomics
levels
OTU
Orders
Families
Genera
Species
o_1
f_2
f_1
g_5
g_4
g_3
g_2
g_1
e_{10}
e_9
e_8
e_7
e_6
e_5
e_4
e_3
e_2
e_1

General approach

The three steps of the approach

- **Step 1 :** Choice of sub-samples to study :
  - Our experiment: Selection

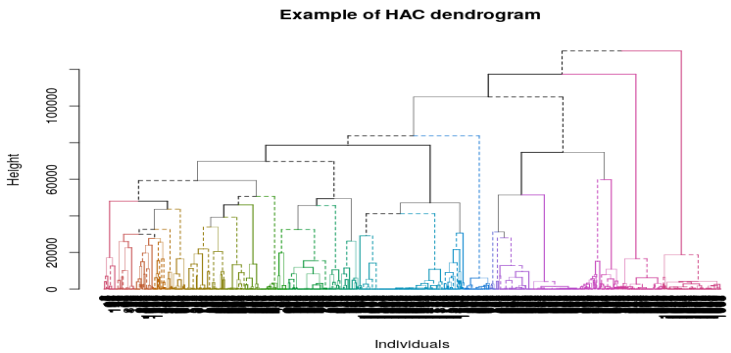
Taxonomic level	Sequences	Number of taxa	Minimal size
Species	313	55	5
Genera	845	36	10
Families	1349	30	10
Orders	1357	11	15

General approach

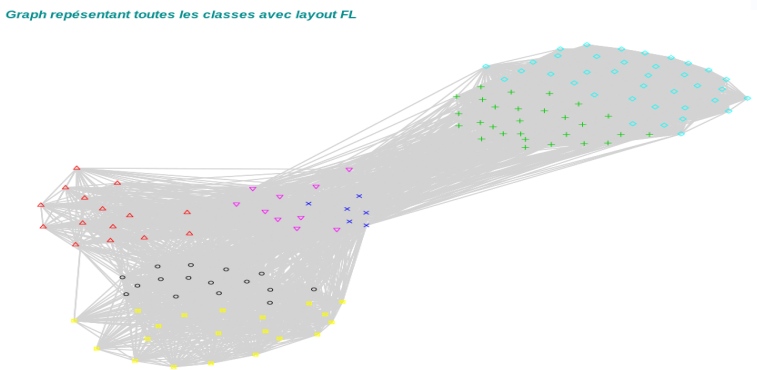
The three steps of the approach

- **Step 2 :** For each sub-sample, building partitions with three methods for each sub-sample and with Smith Waterman and kmer dissimilarities :

  - **M_1 : Agglomerative Hierarchical Clustering (AHC)**



  - **M_2 : Stochastic Block Model (SBM).**



General approach

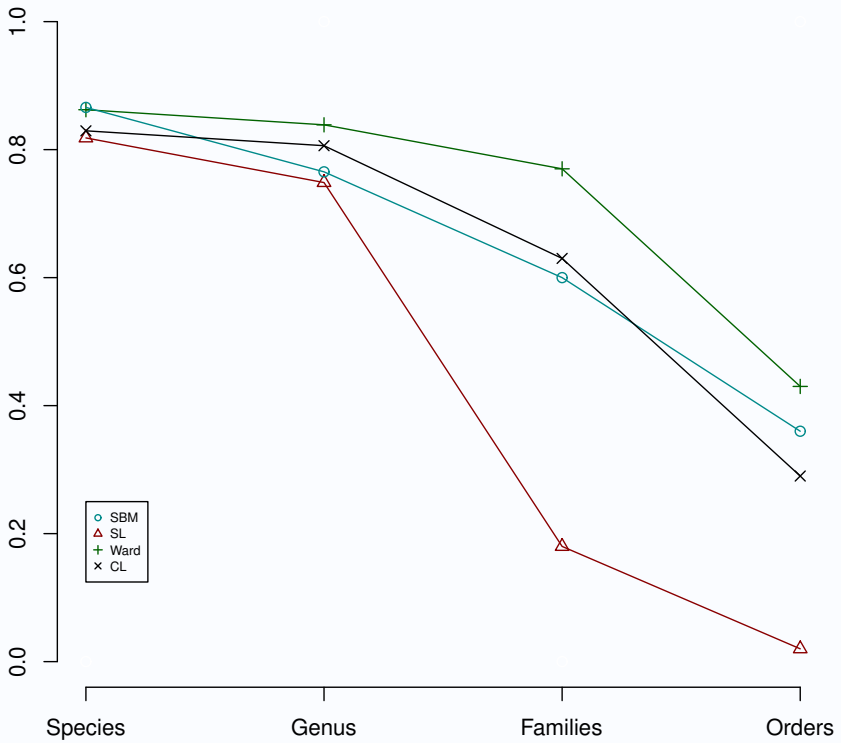
The three steps of the approach

- **Step 3 :** Comparing the classifications two by two
  - Using visual tools
  - Using NMI to characterize the adequacy/independence

Results as a function of taxonomic level

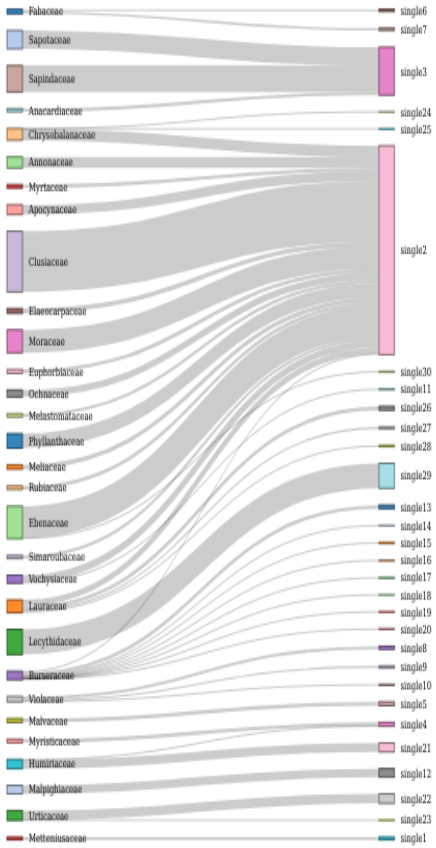
Evolution of NMI index as a function of Taxonomic levels:

kmer based distances

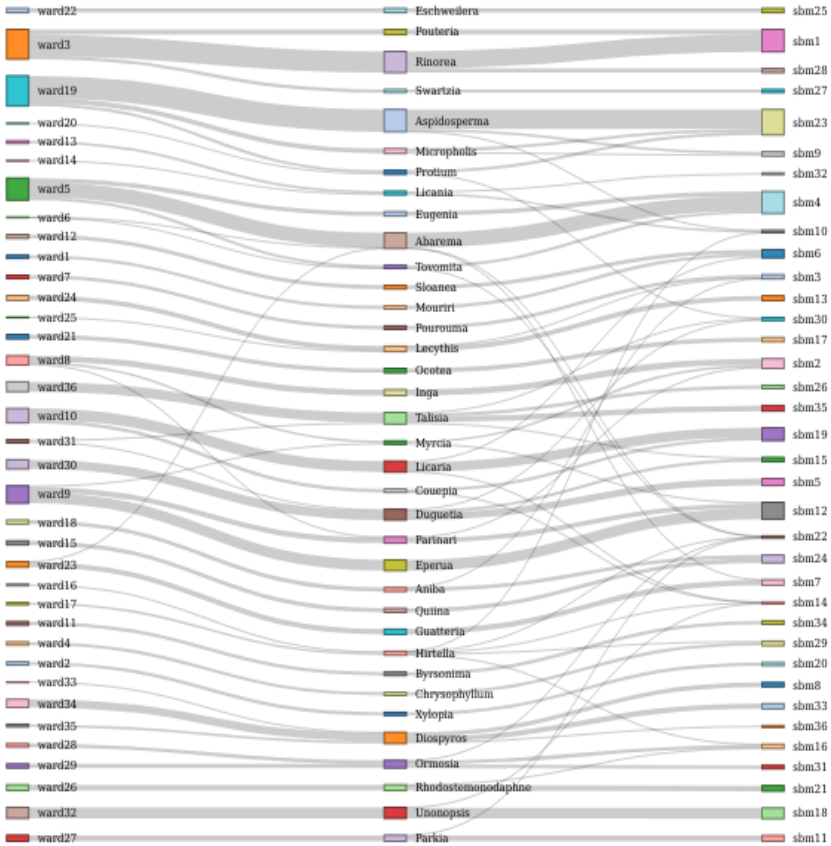


Sankey plots for genera

Botanics WRT SL

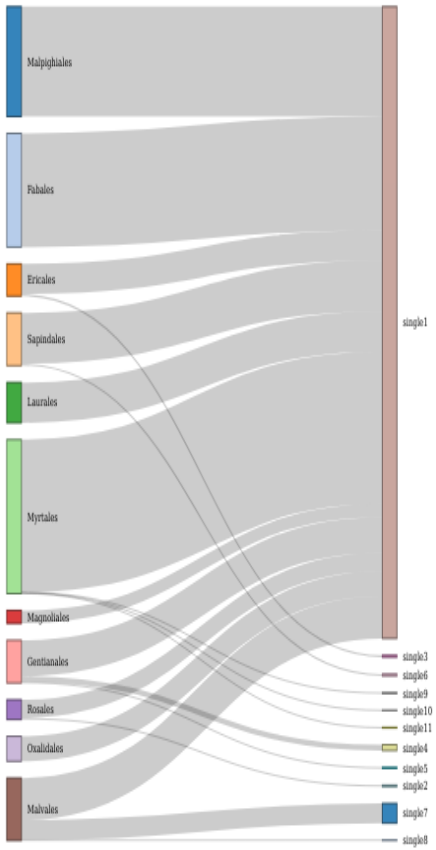


Ward WRT Botanics WRT SBM

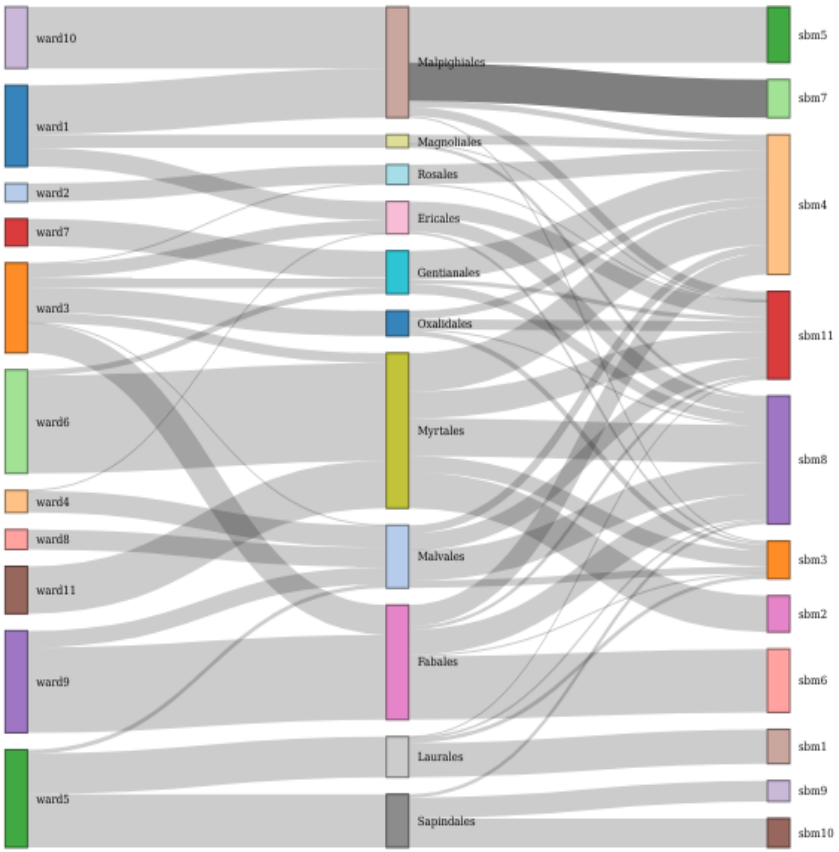


Sankey plots for orders

Botanics WRT SL



Ward WRT Botanics WRT SBM



In conclusion : Interest of SBM models

The main advantage between AHC and SBM :

- AHC produces community w.r.t. SBM produces classes (not necessary communities)
- Outputs of SBM are : Classes and Λ, distance matrix between classes.

Let's talk about Λ :

Case 1 :

$$\Lambda = \begin{pmatrix} 2 & 9 & 11 \\ 6 & 3 & 7 \\ 8 & 5 & 1 \end{pmatrix}$$

There are 3 communities $\Longrightarrow$ SBM $\simeq$ CAH

In conclusion : Interest of SBM models

Case 2 :

$$\Lambda = \begin{pmatrix} 22 & 9 & 11 \\ 6 & 3 & 7 \\ 8 & 5 & 1 \end{pmatrix} \qquad \Lambda = \begin{pmatrix} 2 & 9 & 11 \\ 6 & 23 & 7 \\ 8 & 5 & 1 \end{pmatrix} \qquad \Lambda = \begin{pmatrix} 2 & 9 & 11 \\ 6 & 3 & 7 \\ 8 & 5 & 19 \end{pmatrix}$$

There are 2 communities $\Longrightarrow$ SBM (warning) $\neq$ CAH

Case 3 :

$$\Lambda = \begin{pmatrix} 22 & 9 & 11 \\ 6 & 23 & 7 \\ 8 & 5 & 19 \end{pmatrix}$$

There are no communities $\Longrightarrow$ SBM (warnings) $\neq$ CAH

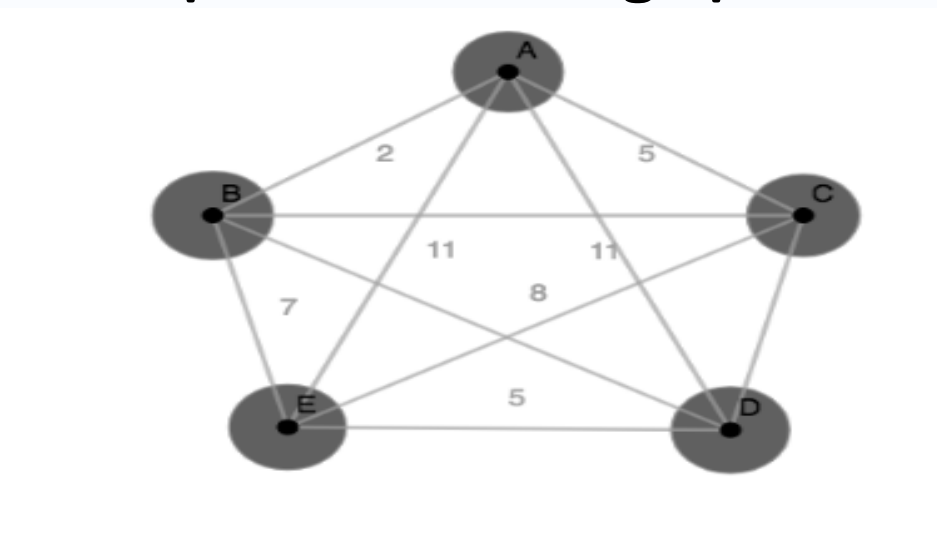
SBM model

Model intuition

Data set

$$\begin{pmatrix} 0 & 2 & 5 & 11 & 11 \\ \vdots & 0 & \ddots & \ddots & 7 \\ \vdots & \vdots & \ddots & \ddots & 8 \\ \vdots & \ddots & \ddots & 0 & 5 \\ \vdots & \dots & \dots & \dots & 0 \end{pmatrix}$$

Representation as graph



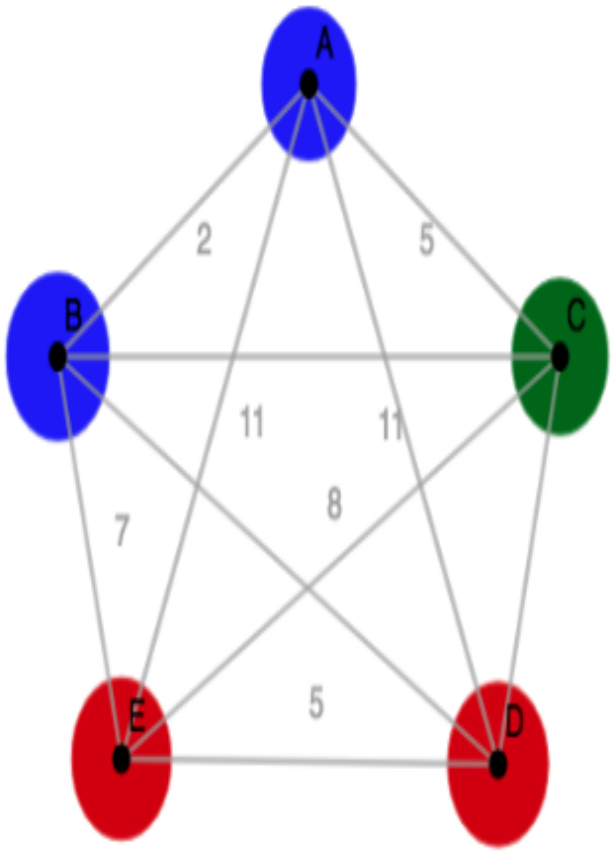
Model parameters

Statistical model : SBM

λ
α

$\Downarrow$

Z



Clusters inference

Let q be the number of classes

$Z = (z_1, \dots z_n)^T \in \mathbb{M}_{n,q}([0,1])$

Model Hypothesis

- H_1 Knowing Z, distances are independent.
- H_2 The latent variables $z_{i,i=1,\dots,n}$ are iid in $\{1, \dots q\}$.

J.-J. Daudin et al, A mixture model for random graphs, Stat Comput (2008)

Parameters presentation

- $D_{i,j}|z_{i,b} = 1, z_{j,b'} = 1 \sim \text{Pois}(\lambda_{b,b'})$
- $\Lambda \in \mathbb{M}_{q,q}, \lambda_{b,b'}$: The parameter of Poisson probability to have a distance d between a vertex of class b and a vertex of class b'.

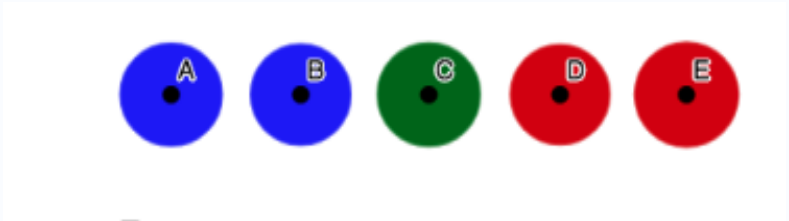
$$\forall b, b' = 1, \ldots, q, \lambda_{b,b'} = z_b^T \Lambda z_{b'}$$

$$\Lambda = \begin{pmatrix} 2 & 9 & 11 \\ 6 & 3 & 7 \\ 8 & 5 & 1 \end{pmatrix}$$

- $\alpha \in [0,1]^q, \alpha_i$: The probability to belong to class i.

EM algorithm

$$Z = (z_1, \dots z_n)^T \in \mathbb{M}_{n,B}([0,1])$$



- estimating Z needs to obtain $\hat{\theta} = (\hat{\alpha}, \hat{\Lambda})$ we proceed by $\hat{\theta}_{mv} = \text{argmax}(P(D|\theta))$
- The most naturel way is the EM algorithm. Each iteration involves two steps :
  - **E-step :** Compute : $Q(\theta, \theta^t) = \mathbb{E}_Z[\log P_\theta(D|Z)|\theta^t, D]$
  - **M-step :** $\theta^{(t+1)} = \text{argmax}_\theta \, Q(\theta, \theta^t)$

⚠ The main difficulty of the EM algorithm is to compute the marginals: It requires q^{n-1} sums.

Dempster et al, "Maximum Likelihood from Incomplete Data via the EM Algorithm". Journal of the Royal Statistical Society (1977)

Estimation approach

There are two main classes of methods :

Monte-Carlo methods : characterize a distribution by randomly sampling values from the distribution.

+ Precision : Accurate
− Computation time : Slow

Estimation approach

There are two main classes of methods :

Monte-Carlo methods : characterize a distribution by randomly sampling valuesfrom of the distribution.

- \+ Precision : Accurate
- − Computation time : Slow

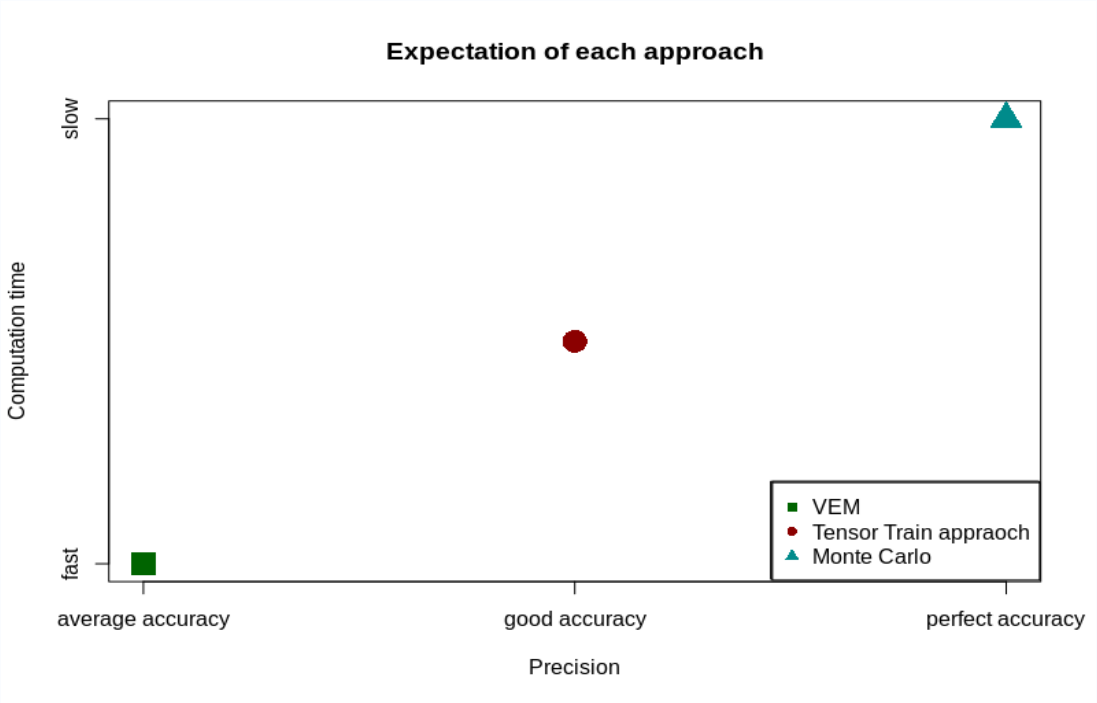
Variational methods: we assume the independence of the nodes knowing the graph to approximate marginal by mean field.

- − Precision : Average accuracy
- \+ Computation time : Fast

Estimation approach

- The ideal approach consists on having good to perfect accuracy and fast computation time

Expectation :



How to get closer of this approach ?

Tensor trains approximation for SBM estimation

The idea

Let's Ψ express a graphical model,

$$\Psi[z_1, \ldots, z_n] = \Pi_{\ell=1,n} \psi_{A_\ell}(Z)$$

Mean field

$$\Psi : [1, q] \longrightarrow \mathbb{R}$$

$$z_1, \ldots, z_n \longrightarrow \Psi[z_1, \ldots, z_n]$$

Tensor Train

$$\Psi \in \left(\mathbb{R}^q\right)^{\otimes n}$$

$$\Psi[z_1, \ldots, z_n] \in \mathbb{R}$$

The idea

- **The mean field** which is approximation by products of scalar (rank $=1$).

Smartness

$$x_{ij} = a_i b_j$$

$$\sum_{i,j} x_{ij} = \sum_{i,j} a_i b_j = \left(\sum_i a_i\right)\left(\sum_j b_j\right)$$

n^2 products kjkjkjkjk 1 product

- **Tensor Train** which is approximation by matrix products (rank >1).

Smartness

$$\sum_{i,j} \mathbf{a}_i \mathbf{b}_j = \left(\sum_i \mathbf{a}_i\right) \cdot \left(\sum_j \mathbf{b}_j\right)$$

$\underline{\mathbf{a}_i}|\mathbf{b}_i$ kjkjkjkjk 1 matrix product

The idea

- **The mean field** which is approximation by products of scalar (rank =1).

Smartness

$$x_{ij} = a_i b_j$$

$$\sum_{i,j} x_{ij} = \sum_{i,j} a_i b_j = \left(\sum_i a_i\right)\left(\sum_j b_j\right)$$

n^2 products kjkjkjkjk 1 product

- **Tensor Train** which is approximation by matrix products (rank >1).

Smartness

$$\sum_{i,j} \mathbf{a}_i \mathbf{b}_j = \left(\sum_i \mathbf{a}_i\right) \cdot \left(\sum_j \mathbf{b}_j\right)$$

$\underline{\mathbf{a}_i}|\mathbf{b}_i$ kjkjkjkjkjk 1 matrix product

The idea

Tensor Train

- The tensor train format of the joint probability distribution :

$$\forall z_1, \ldots, z_n, \Psi[z_1, \ldots, z_n] = A_1[z_1] \cdot A_2[z_2] \ldots A_{n-1}[z_{n-1}] \cdot A_n[z_n]$$

$$\Psi[z_1, \ldots, z_n] = \_\_ \cdot \square \ldots \square .|$$

$A_i[z_i]$: The cores of Ψ (matrix)

- This format allows variable separation, suitable for marginals

How ?

How to compute $A_i[z_i]$ matrix ?

Novikov approach

- compute the TT-approximation of each factor
- then, it uses the Kronecker mixed-product property to compute the TT-approximation of Ψ from the TT approximation of each factor

Novikov et al, Putting MRFs on a Tensor Train (2014)

Improvement of this approach :

- The matrix $A_i[z_i]$ are sizes $O(q^n)$

  A way to deal with this is to use the TT Format that allow us a storage

+ who requires much less memory space
+ who can be used for matrices
+ with efficient operations

Oseledets, Tensor-Train Decomposition SIAM Journal on Scientific Computing (2011)

Computation of marginals

- Partition function :

$$W = \sum_{z_1,\dots,z_n} A_1[z_1] \dots A_n[z_n]$$

$$= \left(\sum_{z_1} A_1[z_1] \right), \dots, \left(\sum_{z_n} A_n[z_n] \right)$$

Let's $B_i = \sum_{z_i} A_i[z_i]$

$$W = B_1 \times \dots \times B_n$$

Computation of marginals

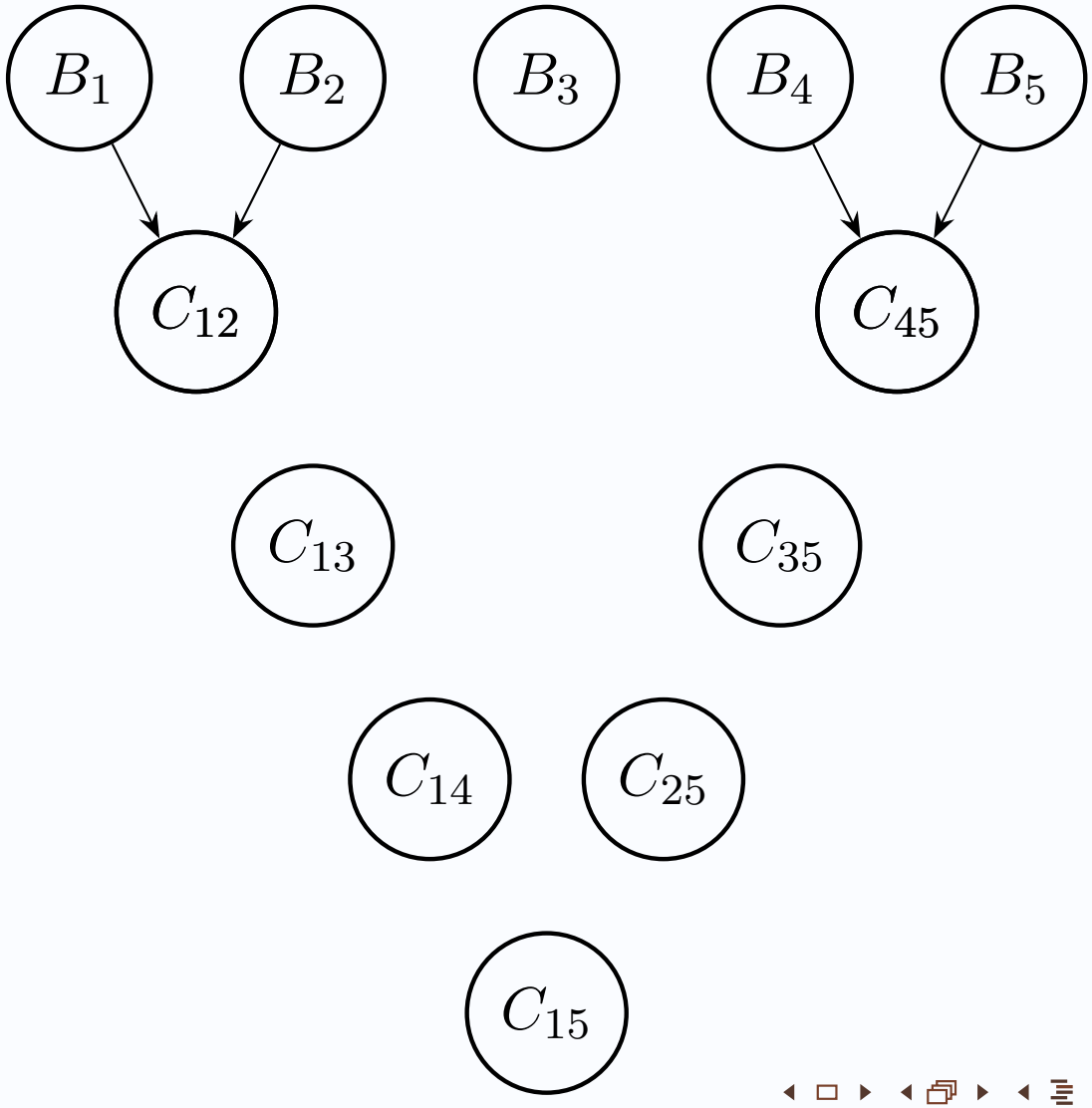
Let's $B_i = \sum_{z_i} A_i[z_i]$

- Binary marginals : $\forall i, j \in [1, n]^2, p_{i,j}(z_i, z_j)$

$$\underbrace{B_1 \times \ldots \times B_{i-1}}_{\text{left}} A_i[z_i] \underbrace{B_{i+1} \times \ldots \times B_{j-1}}_{\text{center}} A_j[z_j] \underbrace{B_{j+1} \times \ldots \times B_n}_{\text{right}}$$

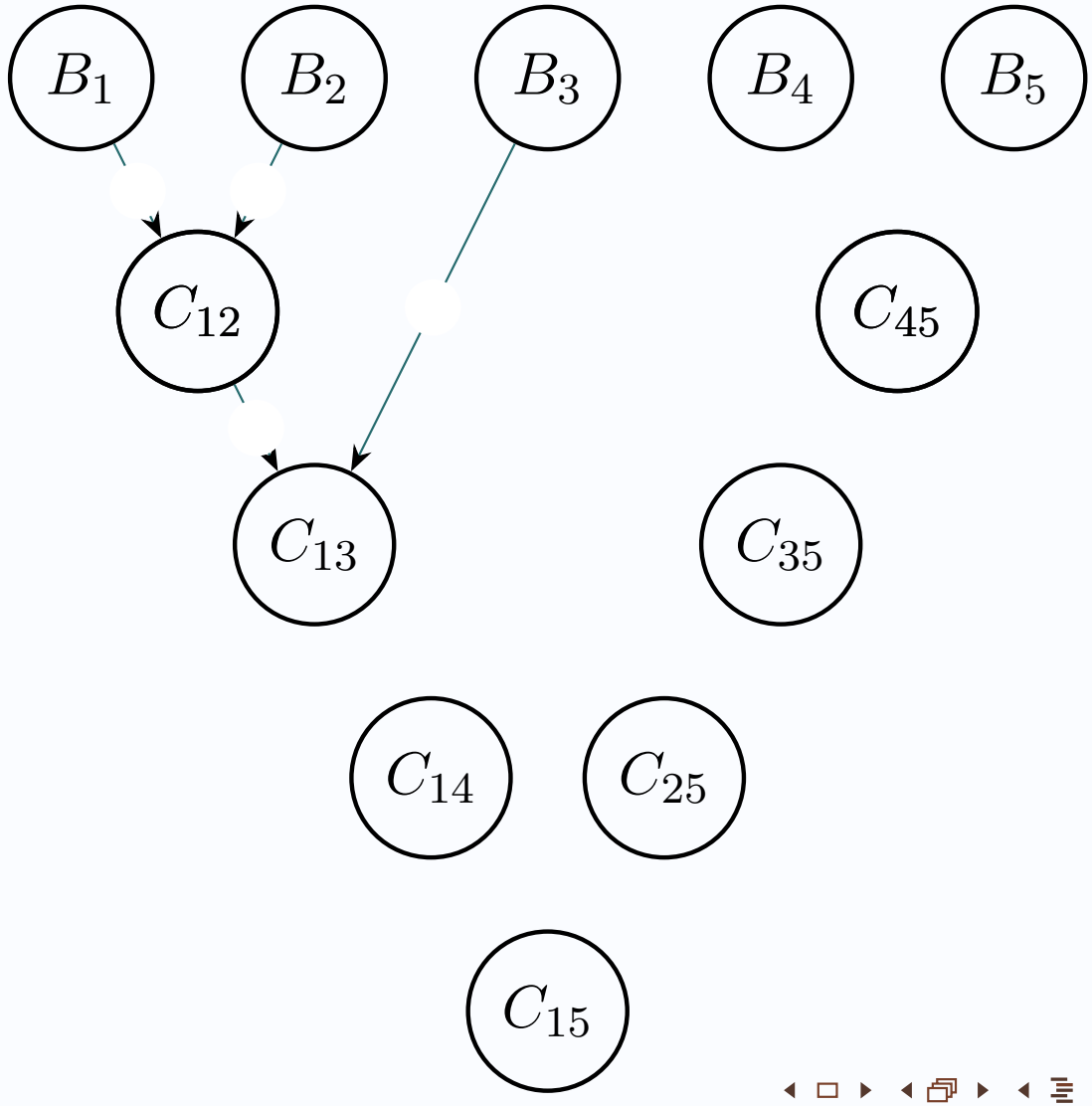
Computation of marginals

- Unary and Binary marginals : First step :



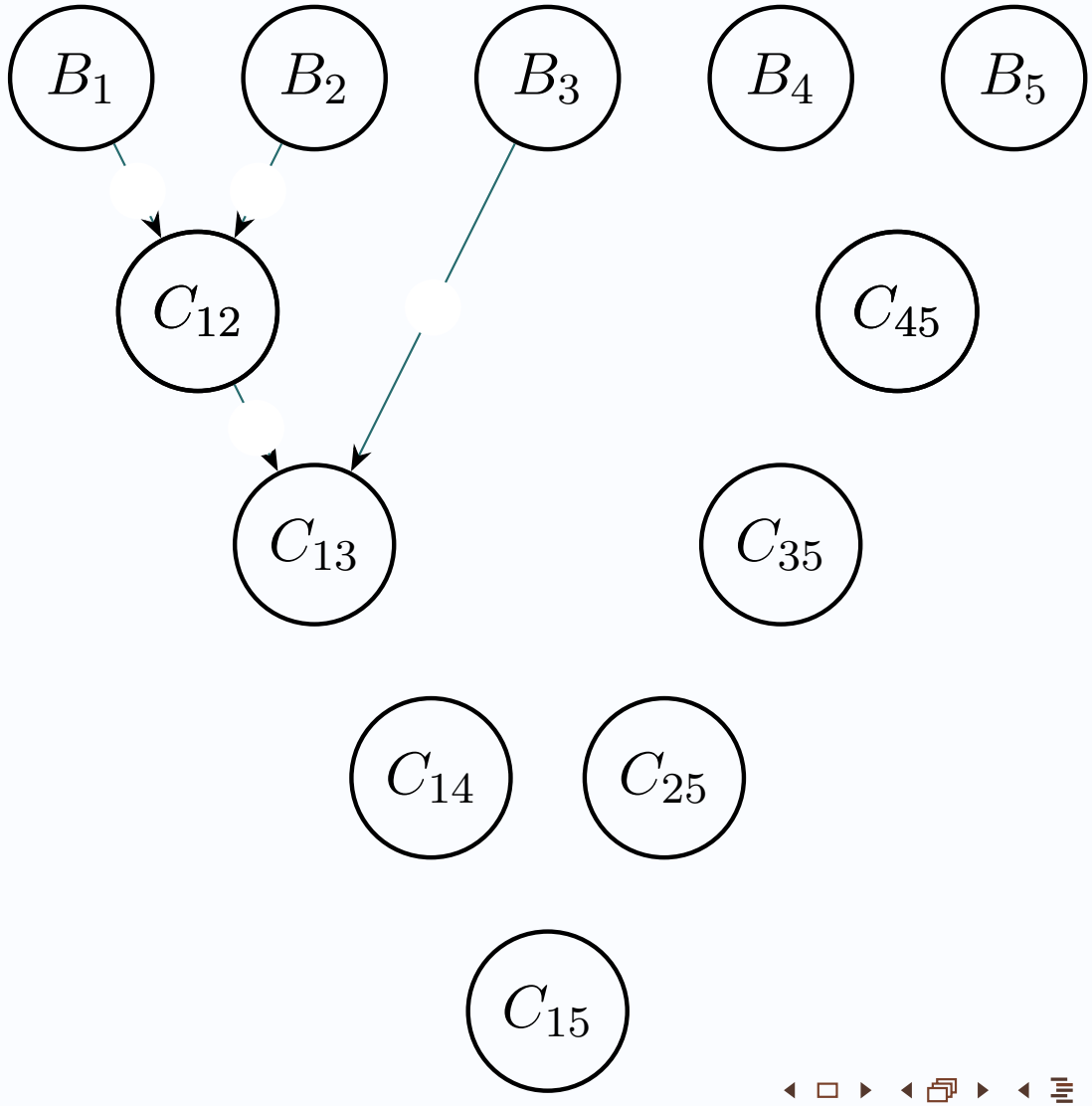
Computation of marginals

- Unary and Binary marginals : Second step : Left side



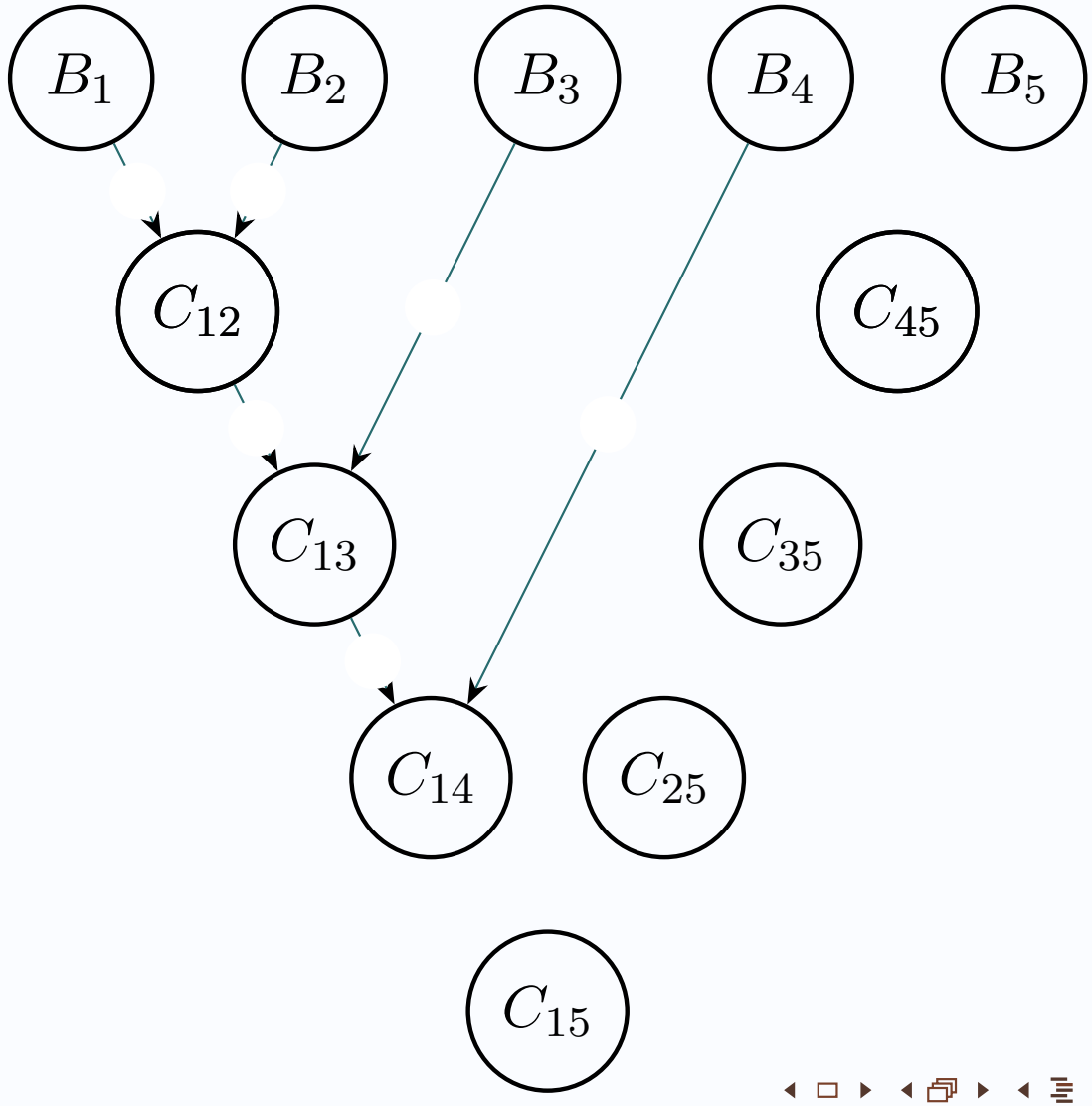
Computation of marginals

- Unary and Binary marginals : Second step : Left side



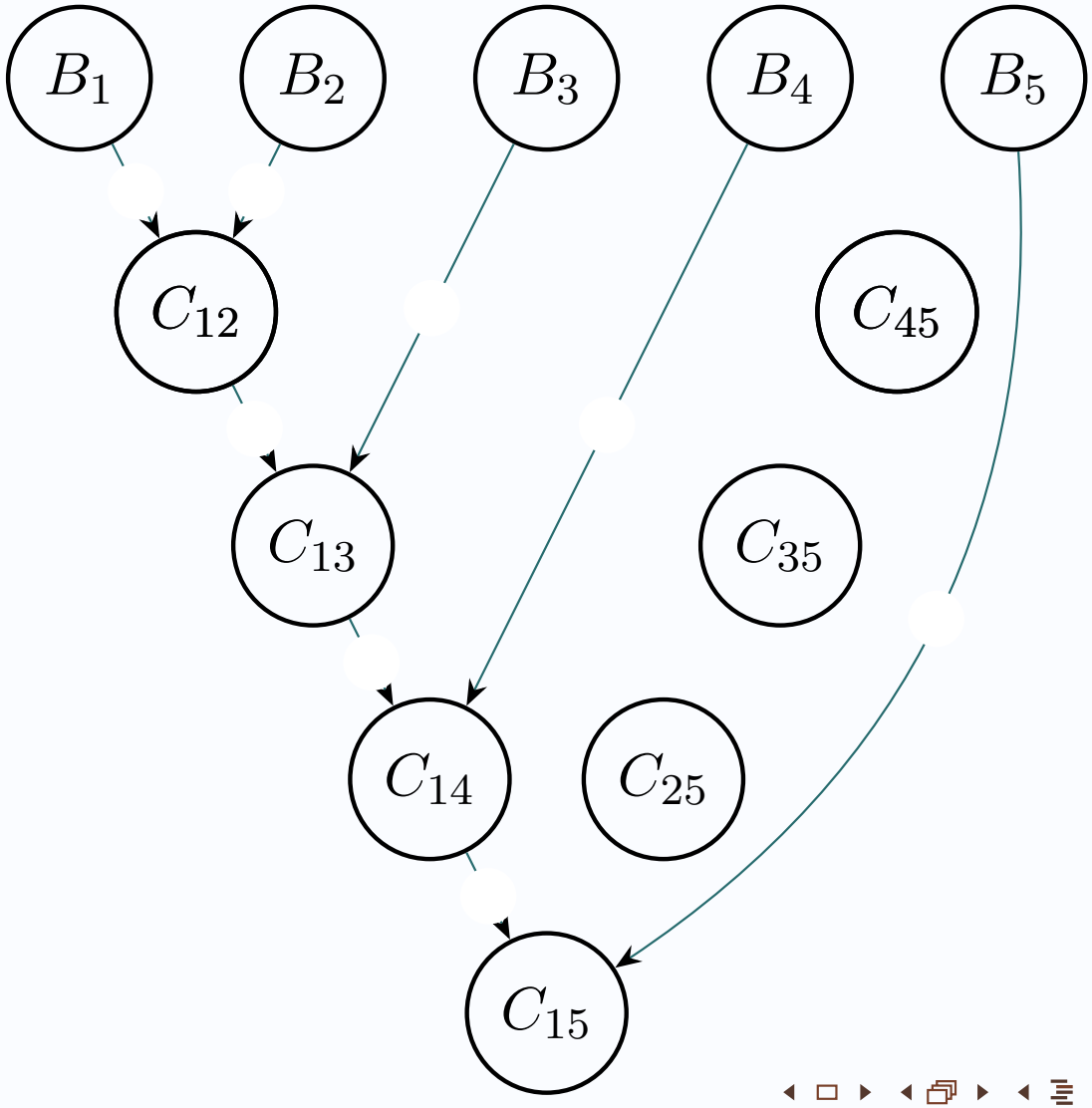
Computation of marginals

- Unary and Binary marginals : Second step : Left side



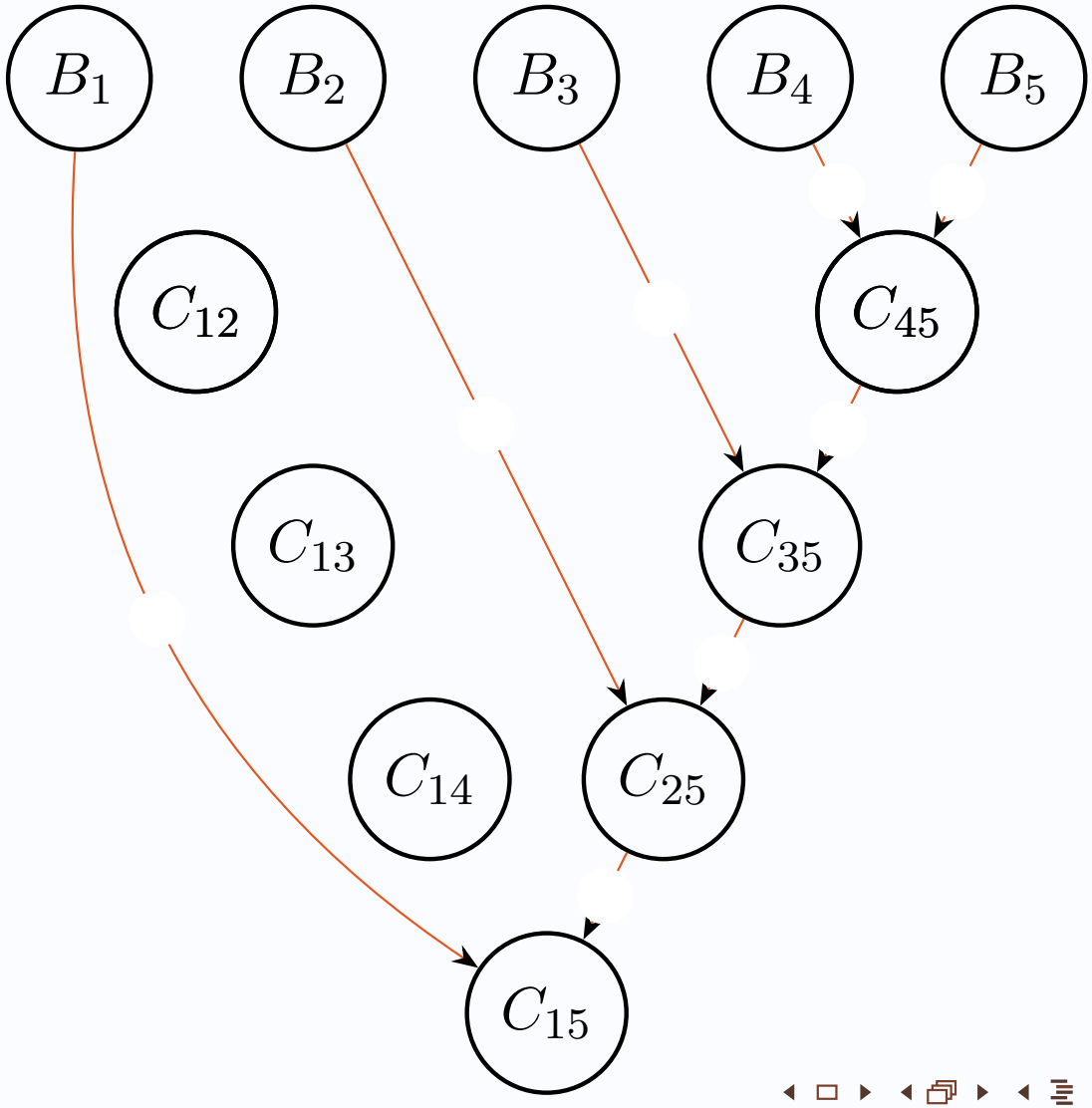
Computation of marginals

- Unary and Binary marginals : Second step : Left side



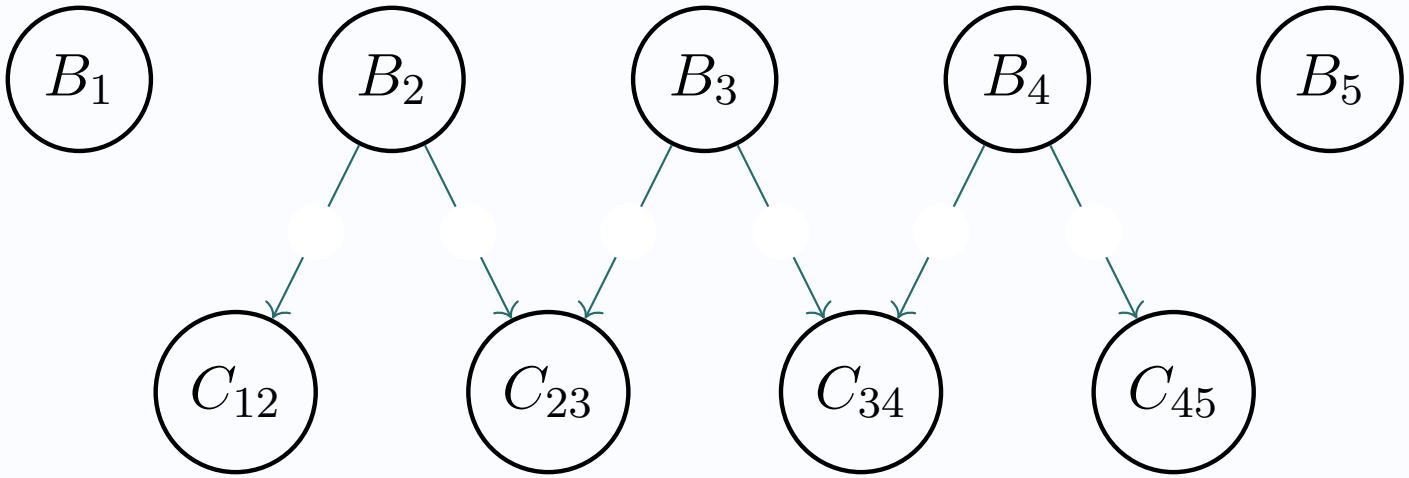
Computation of marginals

- Unary and Binary marginals : Third step : Right side



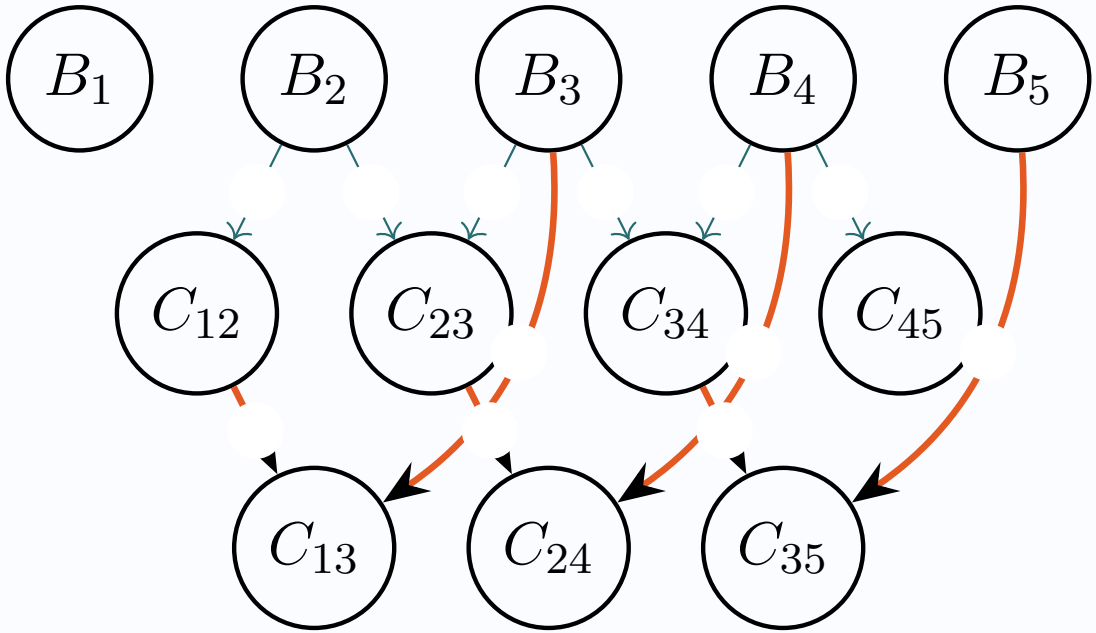
Computation of marginals

- Binary marginals : Fourth step Centers



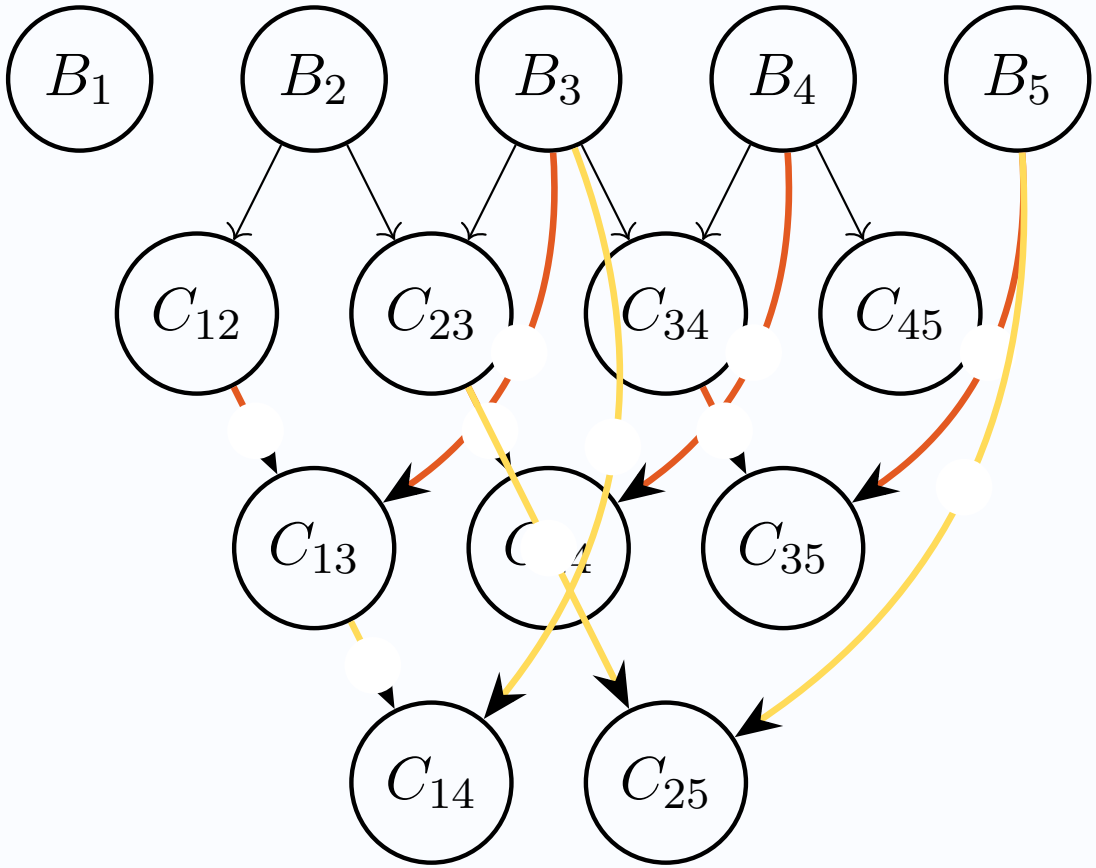
Computation of marginals

- Binary marginals : Fourth step Centers



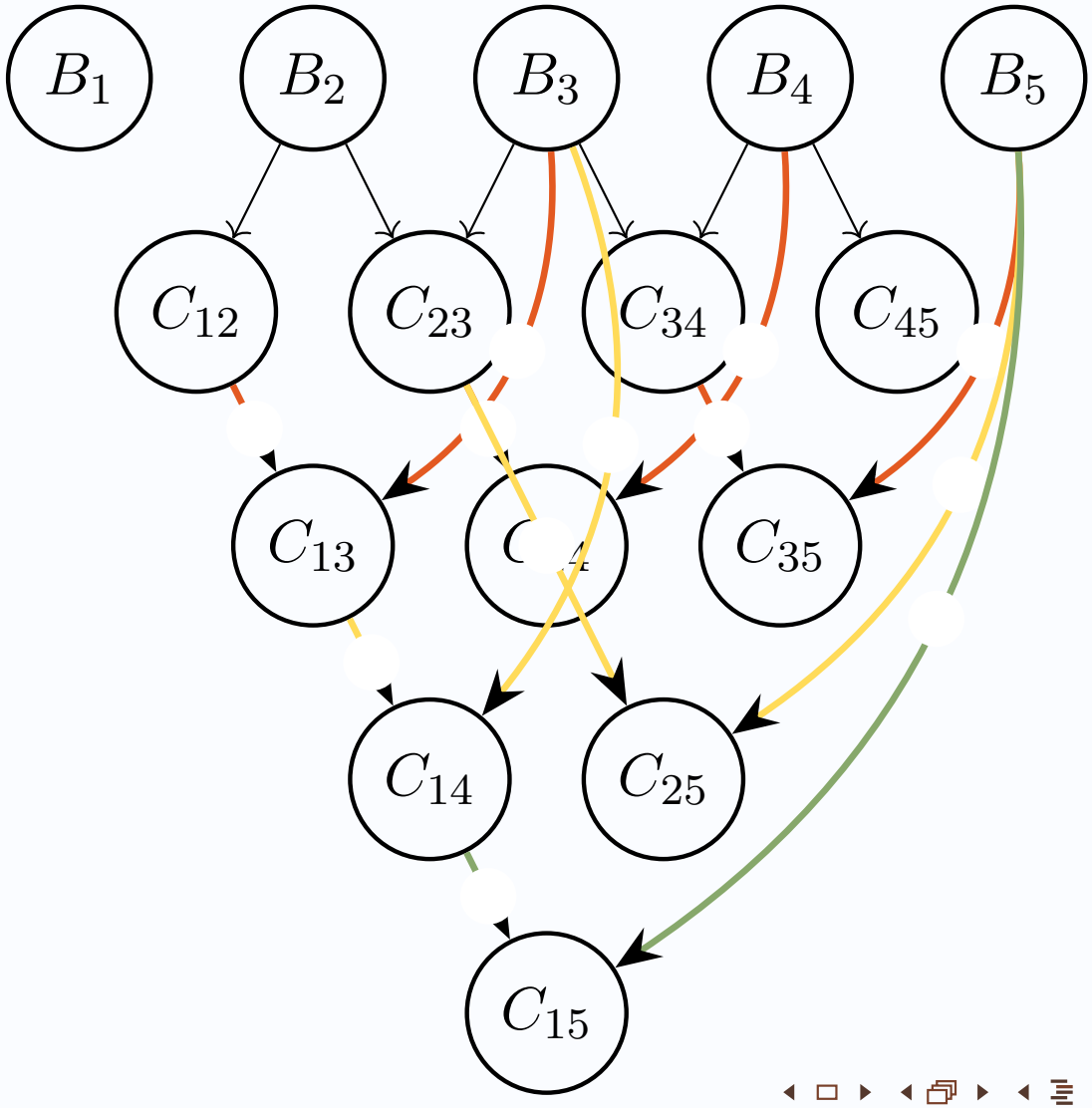
Computation of marginals

- Binary marginals : Fourth step Centers



Computation of marginals

- Binary marginals : Fourth step Centers



Conclusion and prospect

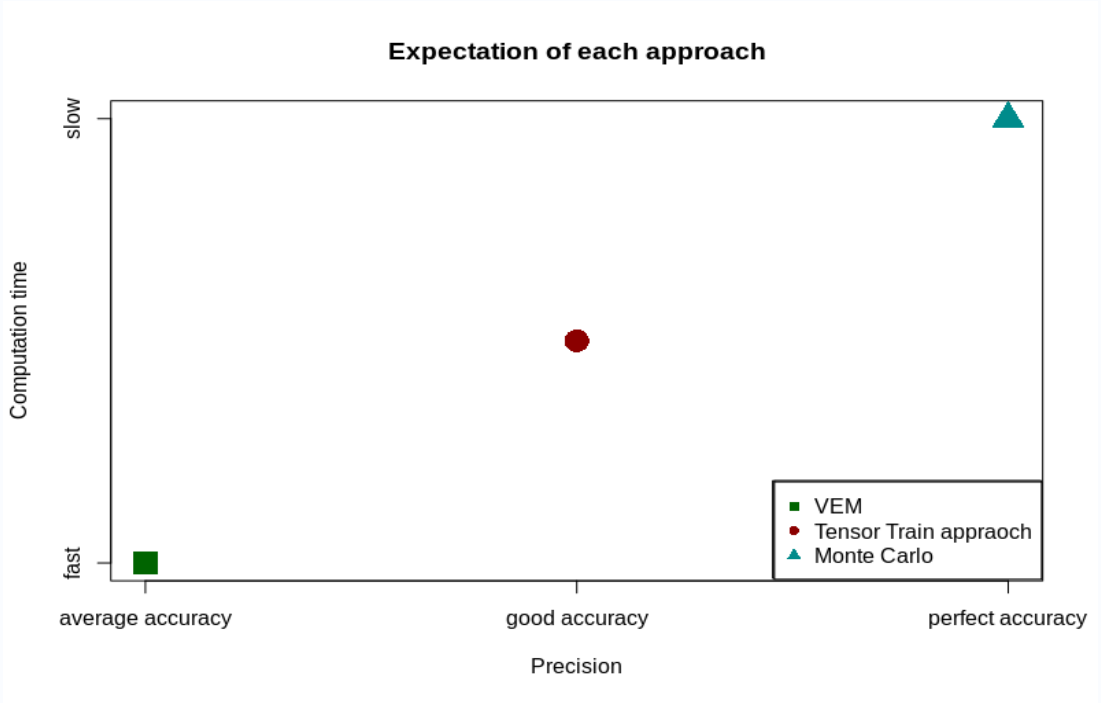
During this presentation, we showed that

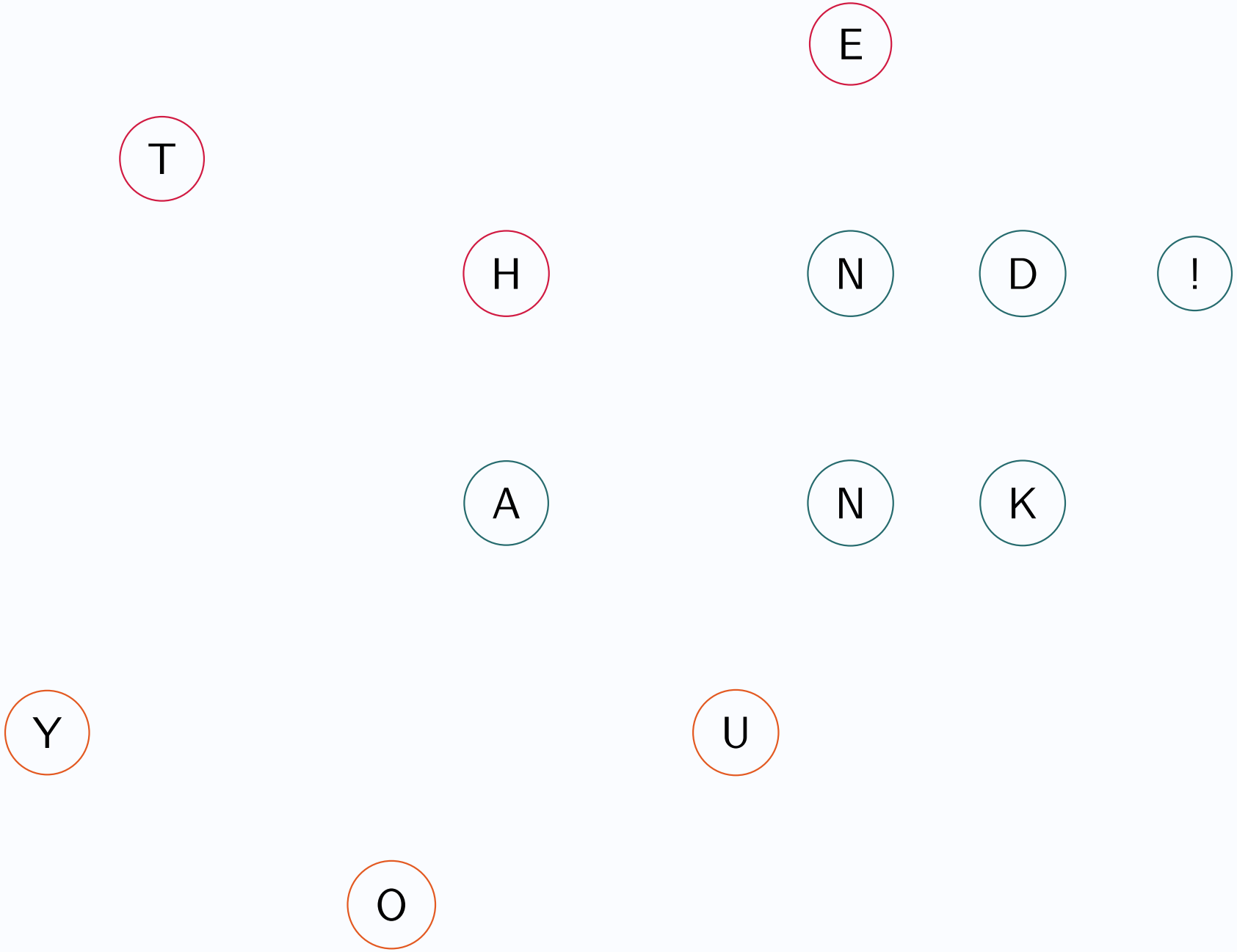
- SBM models can be used for metabarcoding
- that we can potentially improve their accuracy thanks to a tensor based approach

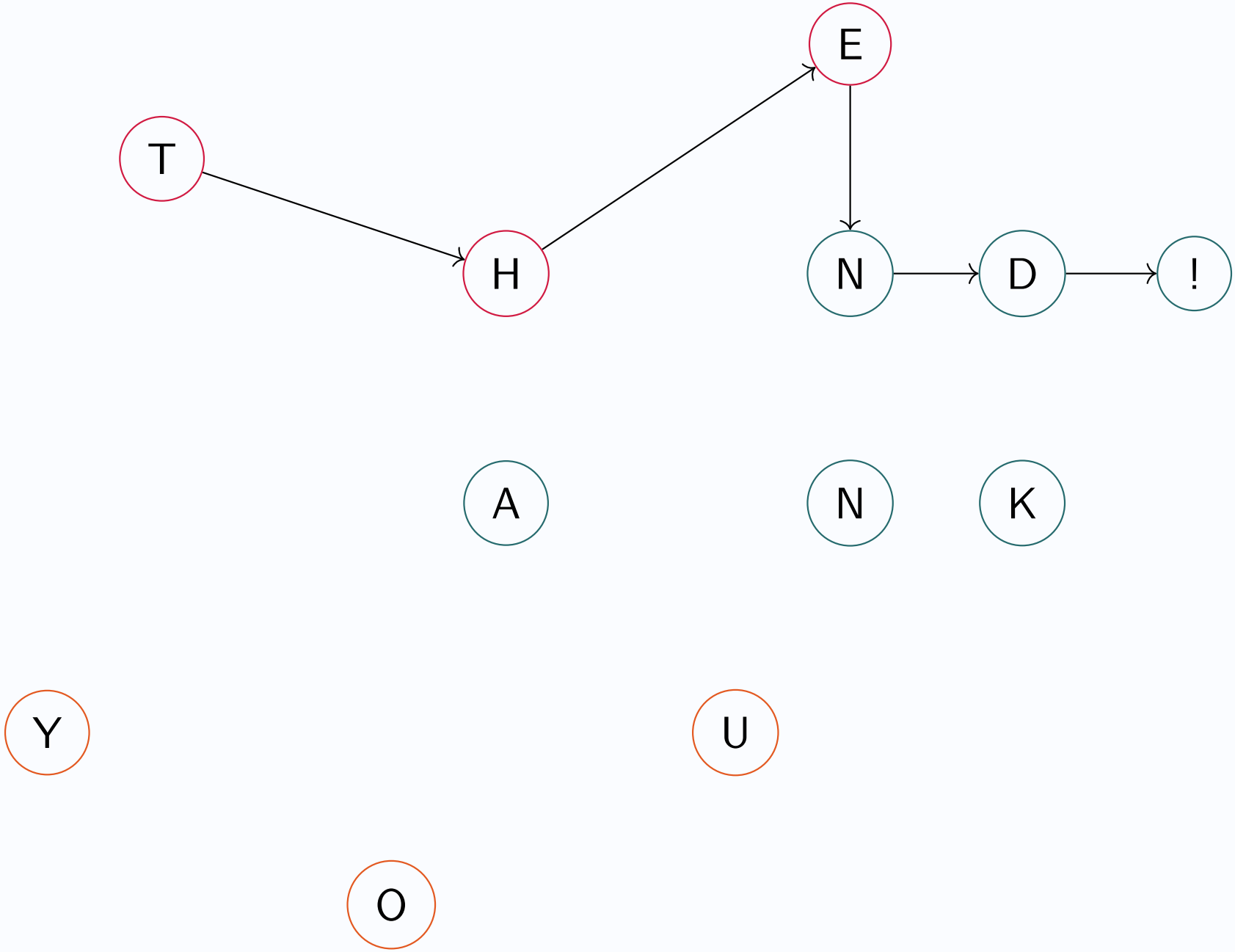
Conclusion and prospect

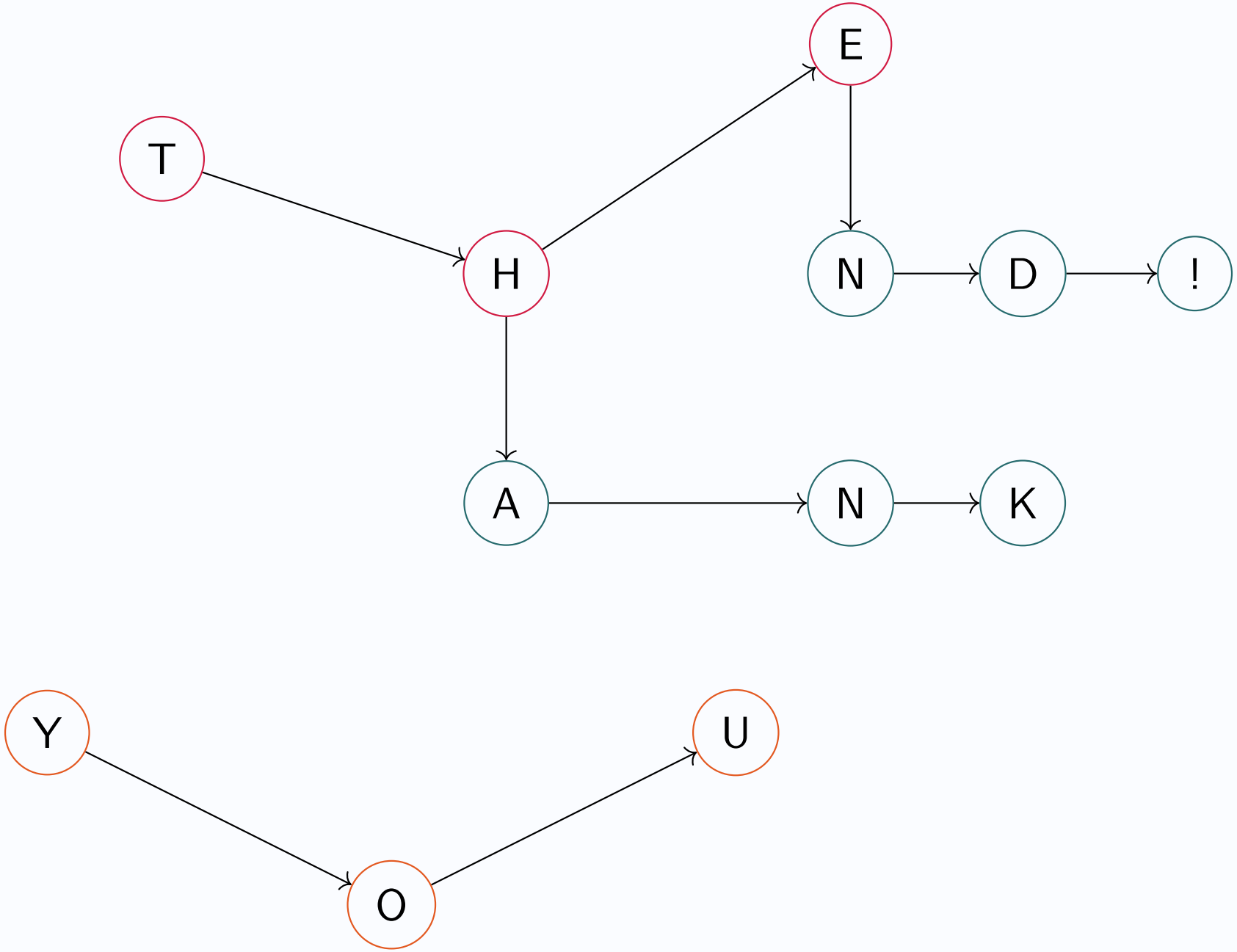
A legitimate question remains unanswered:

How to scale up the SBM model to large datasets ?




E
T
H
N
D
!
A
N
K
Y
U
O


E
T
H
N
D
!
A
N
K
Y
U
O


E
T
H
N
D
!
A
N
K
Y
U
O