Analysis of complex tables

Sophie Donnet, MIA Paris, INRAE

Update: 16 octobre 2020

Introduction

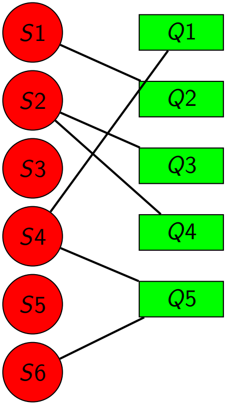
Probabilistic models for bipartite networks

More complex tables

Data of interest

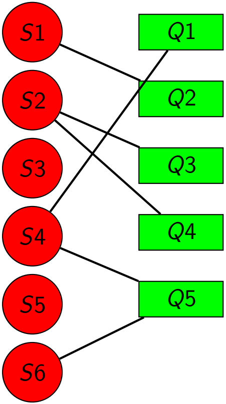
From a table... to a bipartite network

		Questions				
		Q1	Q2	Q3	Q4	Q5
Students	S1	0	1	0	0	0
	S2	0	0	1	1	0
	S3	0	0	0	0	0
	S4	1	0	0	0	1
	S5	0	0	0	0	0
	S6	0	0	0	0	1



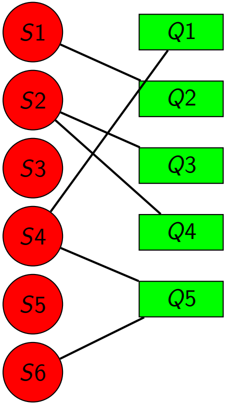
About bipartite networks

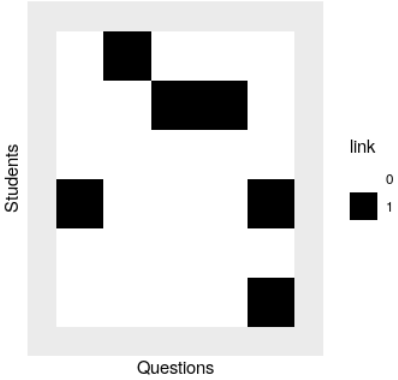
- A bipartite network represents interactions between two sets of nodes



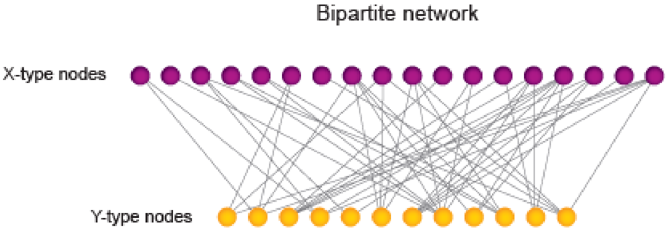
- **Examples**
  - Plant - Pollinators :
  - Farmers - Species : edge is farmer grows plant specie
  - Students - Questions : edge if the student answered to the question

Representation



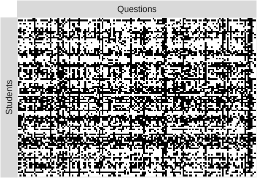


Goals

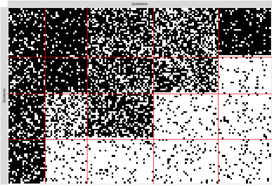


- Unraveling / describing / modeling the network topology.
- Discovering particular structure of interaction between some subsets of nodes.
- Understanding network heterogeneity.

Goals from the tabular point of view



Reordering of rows and cols



How to encode the "tidyness" of a tabular ? ? ? $\Rightarrow$ probabilistic model

Introduction

Probabilistic models for bipartite networks
Latent block model
Statistical inference

More complex tables

Probabilistic approach

- Propose a probabilistic process adapted to our data
- This model will depend on unknown parameters
- Find these parameters adapted to our data

Choosing a model

- Probabilistic point of view : our incidence matrix Y is the realization of a stochastic process.
- Aim : Propose a stochastic process which is able to mimic heterogeneity in the connections.

A first naive model

Erdős–Rényi model

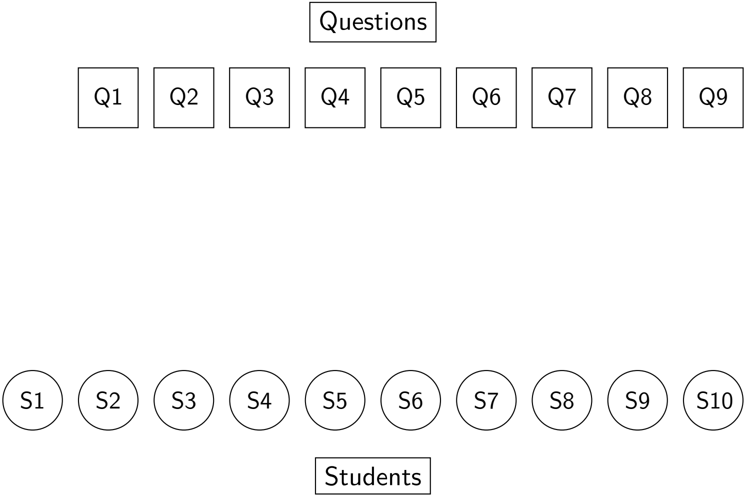
$$\forall (i,j) \in \mathcal{S} \times \mathcal{Q}, \qquad Y_{ij} \sim \mathcal{B}\text{ern}(p)$$

- Homogeneity of the connections : any student i has the same probability to answer well to question j
- No hubs, no community, no nestedness
- All the students have in mean the same number of good answers.

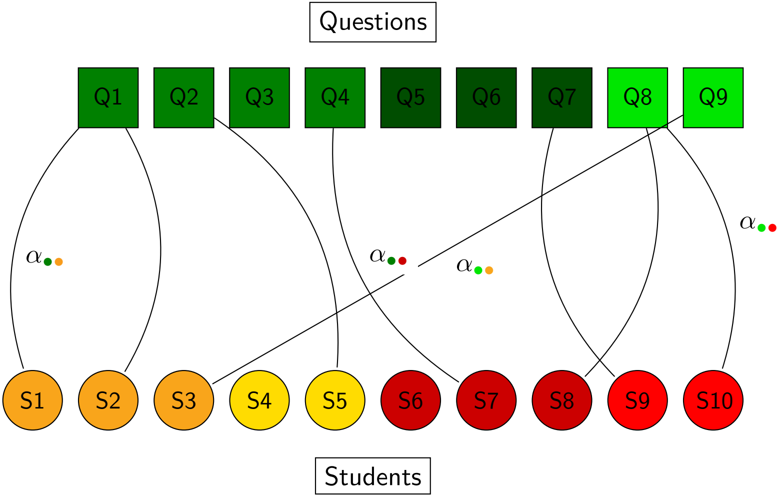
Latent Block Model

- Objective : introduce heterogeneity in the connections
- Tool : introduce blocks of nodes gathering entities that interact roughly similarly in the network

Latent Block Model : a generative model



Latent Block Model : a generative model



Latent Block Model with equations : latent variables I

- Each group of nodes ($\mathcal{S}$ and $\mathcal{Q}$) is divided into blocks / clusters
- $K_{\mathcal{S}}$ number of blocks in $\mathcal{S}$ and $K_{\mathcal{Q}}$ number of blocks in $\mathcal{Q}$
- For any $i \in \{1, \dots, n_{\mathcal{S}}\}$, let $Z_i^{\mathcal{S}}$ be such that

$$Z_i^{\mathcal{S}} = k \quad \text{if entity } i \text{ of group } \mathcal{S} \text{ belongs to cluster } k$$

- For any $j \in \{1, \dots, n_{\mathcal{Q}}\}$, let $Z_j^{\mathcal{Q}}$ be such that

$$Z_j^{\mathcal{Q}} = \ell \quad \text{if entity } j \text{ of group } \mathcal{Q} \text{ belongs to cluster } \ell$$

Latent Block Model with equations : latent variables II

Random latent variables

$(Z_i^{\mathcal{S}})_{i=1\dots n_{\mathcal{S}}}$ and $(Z^{\mathcal{Q}})_{j=1\dots n_{\mathcal{Q}}}$ independent random variables, such that,

$$\begin{aligned}\mathbb{P}(Z_i^{\mathcal{S}}=k) &= \pi_k^{\mathcal{S}},\\ \mathbb{P}(Z_j^{\mathcal{Q}}=\ell) &= \pi_\ell^{\mathcal{Q}}\end{aligned}$$

with $\sum_{k=1}^{K_{\mathcal{S}}}\pi_k^{\mathcal{S}}=1$ and $\sum_{\ell=1}^{K_{\mathcal{Q}}}\pi_\ell^{\mathcal{Q}}=1$

Latent Block Model with equations : connection probability

Conditionally to the latent variables...

$\boldsymbol{Z} = \{Z_i^{\mathcal{S}}, i = 1 \dots n_{\mathcal{S}}, Z_j^{\mathcal{Q}}, j = 1 \dots n_{\mathcal{Q}}\}$:

$$\mathbb{P}(Y_{ij} = 1 | Z_i^{\mathcal{S}} = k, Z_j^{\mathcal{Q}} = \ell) = \alpha_{k\ell} \,.$$

Other emission distributions

- Previous model adapted to 0-1 tables : succeeded or not
- If Y_{ij} is a count : adapted to a score by question (number of points)

$$Y_{ij} | Z_i^{\mathcal{S}} = k, Z_j^{\mathcal{Q}} = \ell \sim \mathcal{P}(\alpha_{k\ell})$$

- If $Y_{ij} \in \mathbb{R}$

$$Y_{ij} | Z_i^{\mathcal{S}} = k, Z_j^{\mathcal{Q}} = \ell \sim \mathcal{N}(\alpha_{k\ell}, \sigma_{k\ell})$$

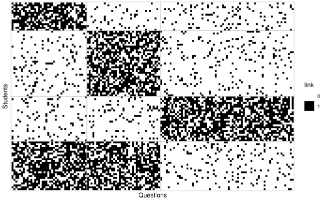
[Govaert and Nadif, 2008]

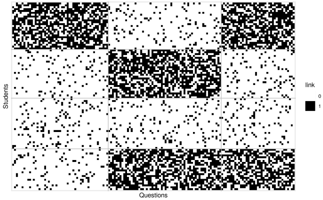
A very flexible model

Why it is reasonable to assume that OUR table is the realization of such a probabilistic model :

- Because it is very flexible
- Depending on the parameters $\pi^{\mathcal{S}}$, $\pi^{\mathcal{Q}}$ and α we can generate very different structures

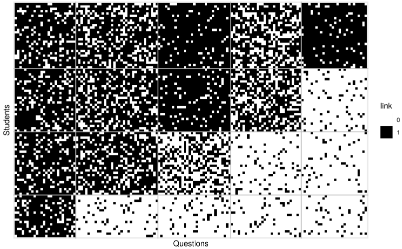
Communities (modules)...





$$\alpha = \begin{pmatrix} 0.60 & 0.09 & 0.09 \\ 0.09 & 0.60 & 0.09 \\ 0.09 & 0.09 & 0.60 \\ 0.60 & 0.60 & 0.09 \end{pmatrix} \quad K_Q\alpha = \begin{pmatrix} 0.60 & 0.09 & 0.60 \\ 0.09 & 0.60 & 0.09 \\ 0.09 & 0.09 & 0.09 \\ 0.09 & 0.60 & 0.60 \end{pmatrix}$$

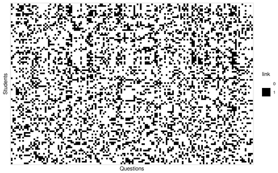
Nested networks



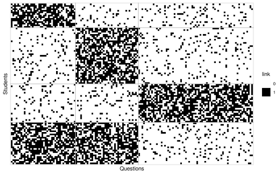
$$\alpha = \begin{pmatrix} 0.80 & 0.70 & 0.90 & 0.60 & 0.90 \\ 0.80 & 0.70 & 0.90 & 0.60 & 0.09 \\ 0.80 & 0.70 & 0.40 & 0.09 & 0.09 \\ 0.80 & 0.09 & 0.09 & 0.09 & 0.09 \end{pmatrix}$$

Inference for LBM

Aim : From an incidence matrix, discovering the clusters



$\Longrightarrow$ Reordering of rows and cols



Remarks

- Looking for the blocks such that, under the assumption that my data come from the LBM model, the observed data Y is most probable (= most likely to occur)
- No specific prior structure
- Entities (students / questions) are gathered because they have similar behavior in the network

Estimation : a difficult task

The "better" clusterings ($Z^{\mathcal{S}}$ and $Z^{\mathcal{Q}}$) have to be found among all the possible clusterings : $K_{\mathcal{Q}}^{n_{\mathcal{Q}}}$, $K_{\mathcal{S}}^{n_{\mathcal{S}}}$

- Complete task from numerical point of view
- Requires well designed algorithm

Model selection

- Selection of the number of blocks $(K_{\mathcal{Q}}, K_{\mathcal{S}})$: in how many rectangles can I organize my tabular ?
- If we only take into account the fit of the model to the data, we will chose as many blocks as students and questions
- But such a model would have many many parameters and this would imply a lot of incertitude on each parameters
- Need for a balance between the fit and the complexity of the model

BIC for observed $\boldsymbol{Z}$

$$ICL(\mathcal{M}) = \log \ell_c(\boldsymbol{X}, \widehat{\boldsymbol{Z}}; \widehat{\theta}, \mathcal{M}) + pen_{\mathcal{M}}$$

where

$$pen_{\mathcal{M}} = -\frac{1}{2}\left\{(K_{\mathcal{S}}-1)\log n_{\mathcal{S}} + (K_{\mathcal{Q}}-1)\log n_{\mathcal{Q}} + K_{\mathcal{S}}K_{\mathcal{Q}}\log(n_{\mathcal{S}}n_{\mathcal{Q}})\right\}$$

Package sbm

Code R

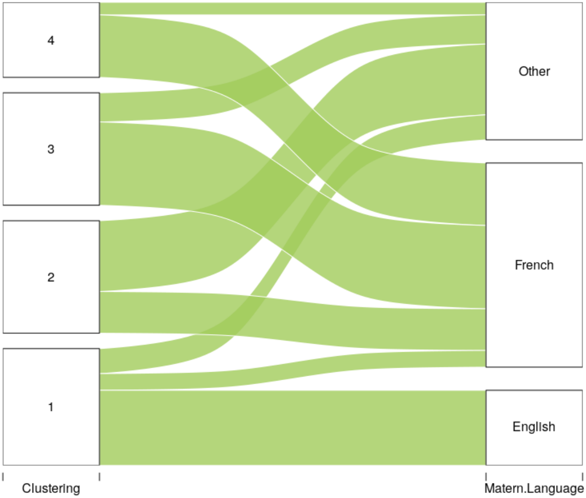
```r
reslbm <- estimateBipartiteSBM(Y,model="bernoulli",dimLabels =
list(row='Students',col='Questions')
```

And after...

What can I do once I have my groups ?

- Have a look at my new tabular globally
- Have a look at the composition of each group
- Sometimes the individuals (students) are described by other items
  - Maternal language
  - Type of cursus
  - ...

Among other plots : the alluvial plots



About covariates

- LBM is modeling connections Y_{ij}.
- If I had covariates on couples (i, j) I could use these elements to explain why student i succeeds in question j.
- However, in that case, my blocks would represent the variability which is not explained by the covariates
- In your case :
  - Covariates on students but not on couples student/question.
  - You want do to groups of similar students and not study why the connections are heterogeneous.
- Integrating covariates would require more thinking.

Introduction

Probabilistic models for bipartite networks

More complex tables
    Modeling a collection of matrices
    Inference

Not one but several tables

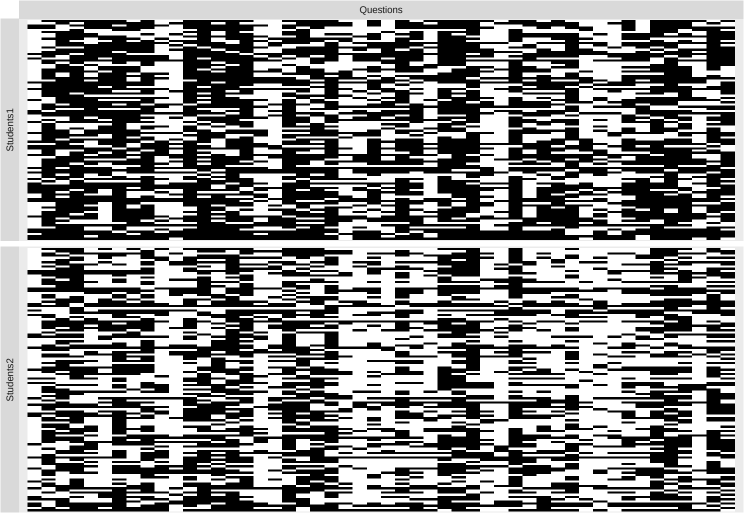
- Sometimes we do not have one table but several tables that are linked together
- I give hereafter a few examples

Several groups of students...

Assume that we have several groups of students (IUT, L1 Staps, L1 Sciences...) who answered to the same questions.

- Students1 : first group of students of size n_1
- Students2 : second group of students of size n_2
- Questions : n_Q. Same questions for all the students

Several groups of students...



We would like to reorganize the two tables at the same time .

Competences on questions...

Assume that each question can be related to a collection of competences

- Students : Students of size n_1
- Questions : n_Q. Same questions for all the students
- Competences : n_C. Each question is related to several competences.

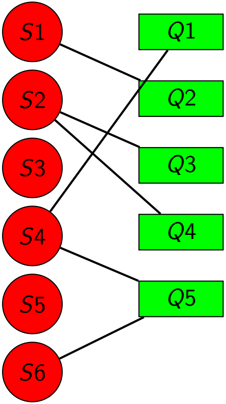
Competences on questions...

We would like to reorganize the two tables at the same time .

Multipartite networks

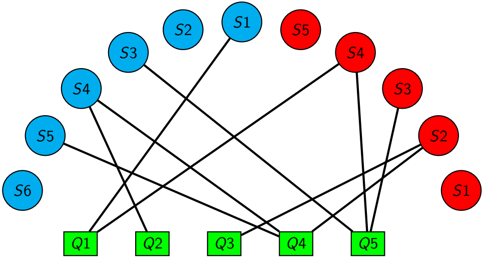
Definition

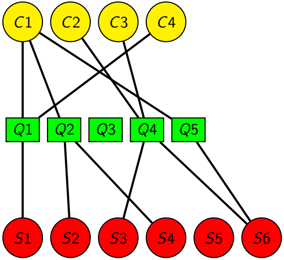
We talk about multipartite network if the vertices are divided into **several** subsets in advance.



From bipartite ...

... to multipartite





Objectives

Aim

Identify subgroups of each functional group sharing the same interaction characteristics and simultaneously taking into account all the matrices.

Existing solutions

- Calculate modularity
  - Detecting communities : making subgroups of individuals who connect more within the subgroup than outside it.
  - In general, people do it separately on each type of interaction and then compare the results between them.

Proposal

Use extensions of the Latent Block Models (LBM) and Stochastic Block Models (SBM) to propose a classification of individuals/agents based on the set of observations.

Data formatting

- P functional groups
- Each functional group p is of size n_p.
- **Data** : a collection of matrices (of adjacency or incidence) representing the relationships within and/or between functional groups :
  - $\mathcal{E}$ = list of pairs (p, p') for which a matrix of interaction between functional groups p and q' is observed.
  - $\boldsymbol{Y} = \{Y^{pp'}, (p, p') \in \mathcal{E}\}$ where $X^{pp'}$ is a matrix of size $n_p \times n_{p'}$.
  - If $p = p'$ matrix of adjacency, symmetrical or not
  - If $p \neq p'$, incidence matrix, bipartite graph

Examples

1. 1 = Students1, 2 = Students 2, 3 = Questions
2. 1 = Students, 2 = Questions, 3 = Comptences

Latent variable probabilistic model

- In the spirit of LBM / SBM : mixing model to model edges
- Each functional group of nodes (or vertices) p is divided into K_p blocks.
- $\forall p = 1 \dots P$, $Z_i^p = k$ if the entity i of the functional group p belongs to the block k.

Latent variables

$(Z_i^p)_{i=1\dots n_p}$ latent, independent random variables : $\forall k = 1 \dots K_p$, $\forall i = 1 \dots n_p$, $\forall p = 1, \dots P$,

$$\mathbb{P}(Z_i^p = k) = \pi_k^p, \tag{1}$$

with $\sum_{k=1}^{K_p} \pi_k^p = 1$ for all $p = 1, \dots P$.

Latent variable probabilistic model

Conditionally...

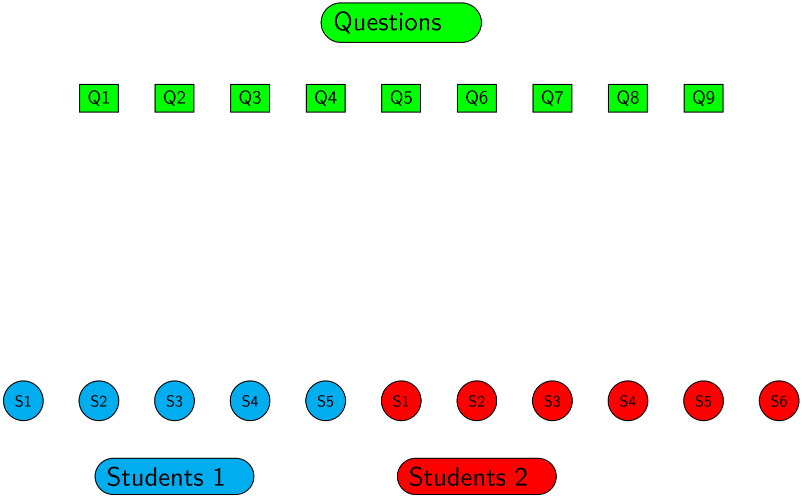
... to latent variables $\boldsymbol{Z} = \{Z_i^p, i = 1 \dots n_p, p = 1 \dots P\}$:

$$P(X_{ij}^{pp'} = 1 | Z_i^p, Z_j^{p'}) = \alpha^{pp'}_{Z_i^p, Z_j^{p'}} . \tag{2}$$

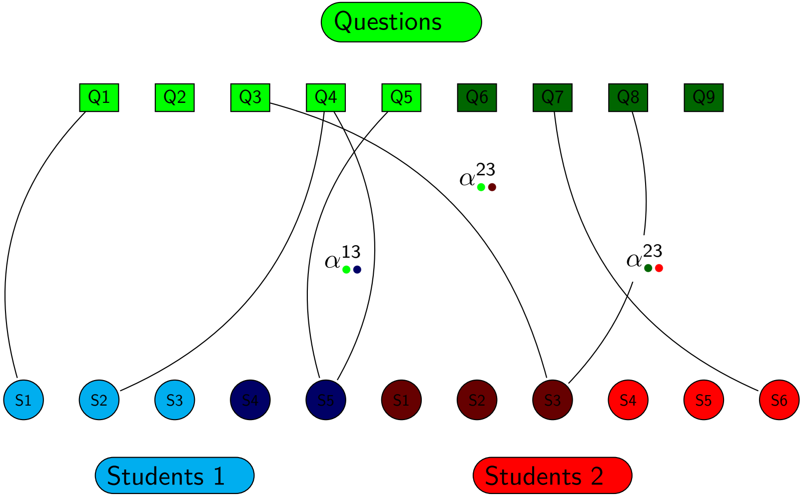
- Law of the interaction phenomenon depends on the i and j membership groups
- In the examples, adapted to binary interactions. We could also consider scores.

[Bar-Hen et al., pear]

Synthetic scheme for Students1 - Students2 - Questions



Synthetic scheme for plants/insects networks



Dependencies between matrices

- If $K_p = 1$ for all p then all the entries of all the matrices are independent random variables : homogeneous connection.
- Otherwise, integration of the random variables $\Rightarrow$ dependence between the elements of the matrices
- Dependence between matrices
- **Consequences** on $\boldsymbol{Z}^p | \boldsymbol{X}$
  - The obtained clustering depends on all interaction matrices.
  - Few simplifications possible

Estimation and model selection

- Likelihood maximized by an adapted version of the VEM algorithm
- Numbers of blocks $(K_1, \dots ; K_P)$ chosen with an adapted ICL criterion (penalized likelihood)
- Method implemented in R package sbm

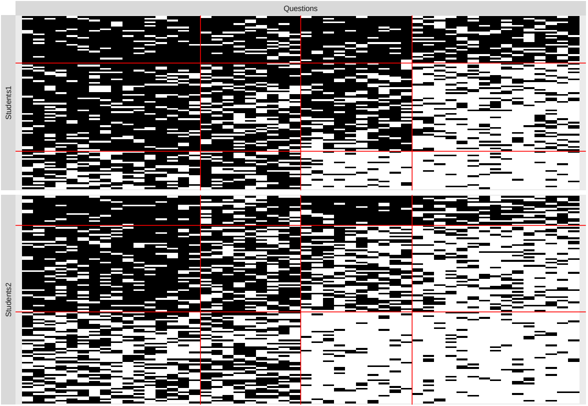
In R, formatting the data. Example 1

```
net1 <- defineSBM(tableStu1Questions, model='bernoulli',
type='bipartite', dimLabels = list(row = 'Students1',col='Questions'))
net2 <- defineSBM(tableStu2Questions, model='bernoulli',
type='bipartite', dimLabels = list(row = 'Students2',col='Questions'))
```

Inference. Example 1

```
resEstimMBM <- estimateMultipartiteSBM(list(net1,net2))
```

Results. Example 1



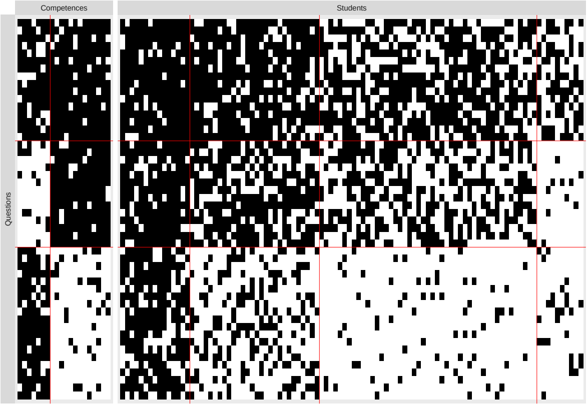
In R, formatting the data. Example 2

```
net1 <- defineSBM(tableStudentsQuestions, model='bernoulli',
type='bipartite', dimLabels = list(row = 'Students',col='Questions'))
net2 <- defineSBM(tableQuestionsCompetences, model='bernoulli',
type='bipartite', dimLabels = list(row =
'Questions',col='Competences'))
```

Inference. Example 2

```
resEstimMBM <- estimateMultipartiteSBM(list(net1,net2))
```

Results. Example 2



Conclusions

- Tool to analyse several tables at the same time
- Interesting if one wants to do groups of entities (students, questions...) coherent accross the tables.
- Adapted to any type of achitecture between the matrices.
- Everything is adapted to scores (in all the matrices or not).

Références I

Bar-Hen, A., Barbillon, P., and Donnet, S. (To appear).
Block models for multipartite networks.applications in ecology and ethnobiology.
Statistical Modelling.

Govaert, G. and Nadif, M. (2008).
Block clustering with bernoulli mixture models : Comparison of different approaches.
Computational. Statistics and Data Analysis, 52(6) :3233–3245.