A probabilistic model for generalized multipartite networks. Application in ecology

Sophie Donnet MIA Paris, INRA, France,
joint work with A. Bar-Hen (CNAM) and Pierre Barbillon (Agroparistech, France)

November 2020
Séminaire Palaisien

Introduction

Modeling a collection of matrix

Estimation and model selection

Application in ecology

Context

Framework

- Networks : fundamental tools in various fields, such as ecological theory to study interactions between agents.
  - Entities / agents = vertices (species for instance)
  - Interactions = edges (pollination for instance)

Context

Framework

- Networks : fundamental tools in various fields, such as ecological theory to study interactions between agents.
  - Entities / agents = vertices (species for instance)
  - Interactions = edges (pollination for instance)
- Statistical objective :
  - Studying the structure of these networks
  - Finding heterogeneity in the way vertices interact

Context

Framework

- Networks : fundamental tools in various fields, such as ecological theory to study interactions between agents.
  - Entities / agents = vertices (species for instance)
  - Interactions = edges (pollination for instance)
- Statistical objective :
  - Studying the structure of these networks
  - Finding heterogeneity in the way vertices interact
- In the recent years : interest for complex networks such as
  - *multiplex networks* –when several types of relations are simultaneously studied on a common set of entities– [Kéfi et al., 2016, Barbillon et al., 2016]
  - *time evolving networks* [Matias and Miele, 2017].

Context

Framework

- Networks : fundamental tools in various fields, such as ecological theory to study interactions between agents.
  - Entities / agents = vertices (species for instance)
  - Interactions = edges (pollination for instance)
- Statistical objective :
  - Studying the structure of these networks
  - Finding heterogeneity in the way vertices interact
- In the recent years : interest for complex networks such as
  - *multiplex networks* –when several types of relations are simultaneously studied on a common set of entities– [Kéfi et al., 2016, Barbillon et al., 2016]
  - *time evolving networks* [Matias and Miele, 2017].
- **In this work** : modeling and inference of multipartite networks

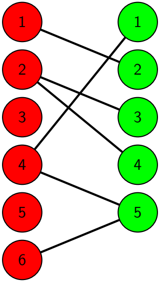
Generalized Multipartite networks

Definition

- Arise when the entities (vertices) at stake can be in advance partitioned into groups defined by their nature.
- Groups will be referred to as *functional groups*.

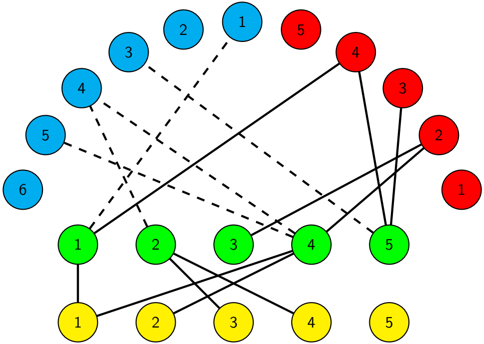
From bipartite networks...

Plants - Pollinators



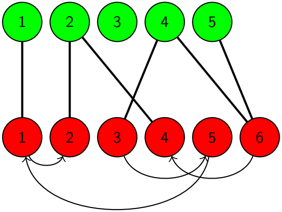
Plants and Pollinators are Functional groups

...to multipartite networks...



For instance : Plants – Pollinators – Seed dispersal birds – Ants

... to Generalized Multipartite networks

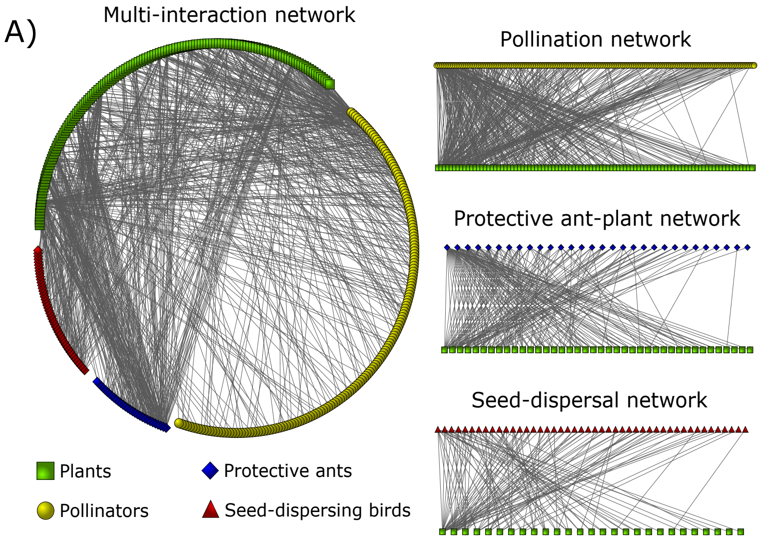


Interactions may also be observed inside Functional Groups.

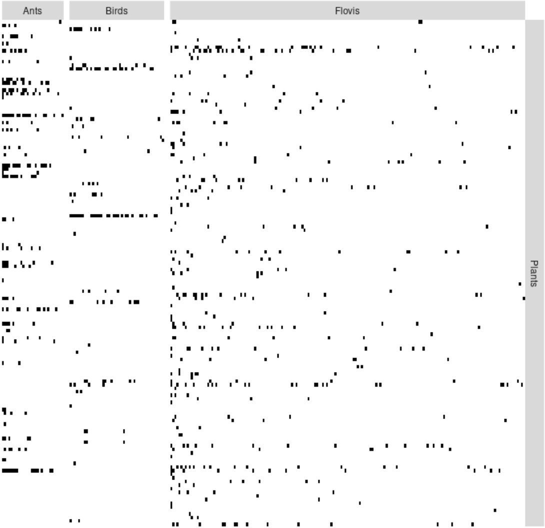
Example of dataset [Dáttilo et al., 2016]

- Observations made along the Mexican Coast by Wesley Dattilo (INECOL, Xalapa, Mexico)
- Entities = living species
- Divided into 4 functional groups : plants, pollinators, ants, seed-dispersing birds
- Edge between plant i and animal j = an individual of animal specie j has been observed at least once in interaction (pollination, protection, eating seeds) with a plant of specie i.

Multipartite network in ecology



Multipartite matrix in ecology



Statistical objectives

Central issue : being able to cluster nodes sharing the same connectivity patterns. Correspond to ecological functions/roles.

Two possible approaches

- Classical metrics detecting pre-specified patterns (e.g. modularity, centrality, nestedness...)
- Probabilistic mixture models : without any prior assumption on the patterns to be found

Statistical objectives

Central issue : being able to cluster nodes sharing the same connectivity patterns. Correspond to ecological functions/roles.

Two possible approaches

- Classical metrics detecting pre-specified patterns (e.g. modularity, centrality, nestedness...)
- Probabilistic mixture models : without any prior assumption on the patterns to be found

Strategy

- New probabilistic mixture model adapted to generalized multipartite networks.
- Multi-clustering : Each functional group is partitioned into clusters gathering nodes sharing the same connection behavior in all the networks they are involved in.

Networks in matrices

- Q functional group : each functional group q of size n_q
- Multipartite network : a collection of networks
- Each network involves one or two functional groups : indexed by pairs (q, q') (q and q' in $[\![1, Q]\!]$).
- $\mathcal{E}$ denotes the list of pairs of observed networks
- Each network encoded in a matrix $X^{qq'}$

$$X^{qq'}_{ii'} = \begin{cases} 1 & \text{if entity } i \text{ of group } q \text{ is in interaction} \\ & \text{with entity } i' \text{ of group } q'. \\ 0 & \text{otherwise} \end{cases}$$

- $\boldsymbol{X} = \left\{ \left(X^{qq'} \right), (q, q') \in \mathcal{E} \right\}.$

Example in ecology

$$X^{1q'}_{ii'} = \begin{cases} 1 & \text{if animal specie } i' \text{ of functional group } q' \text{ has been observed} \\ & \text{in interaction with plant } i \\ 0 & \text{otherwise} \end{cases}$$

$$q' = 2, 3, 4.$$

	Ant 1	…	Ant n_2	Seed dispersing bird 1	…	Seed dispersing bird n_3	Pollinator 1	…	Pollinator n_4
Plant 1			1				1	1	1
Plant 2			1			1			1
⋮		X^{11}_{ij}			X^{12}_{ij}			X^{13}_{ij}	
Plant n_1	1		1			1	1		1

$X^q_{..} \in \{0, 1\}$ to avoid sampling issues

Block model : Mixture model on the $X_{ii'}^{qq'}$

Latent variables

- Each functional group q divided into K_q blocks or clusters
- $$Z_i^q = k$$
  if individual i of functional group q belongs to cluster k.
- Z_i^q are independent random variables :
  $$\mathbb{P}(Z_i^q = k) = \pi_k^q, \tag{1}$$
  with $\sum_{k=1}^{K_q} \pi_k^q = 1$ for any $q = 1, \ldots Q$.
- $\boldsymbol{Z} = (Z_i^q)_{i \in [\![1, n_q]\!], q \in [\![1, Q]\!]}$.

Block model : Mixture model on the $X_{ii'}^{qq'}$

Latent variables

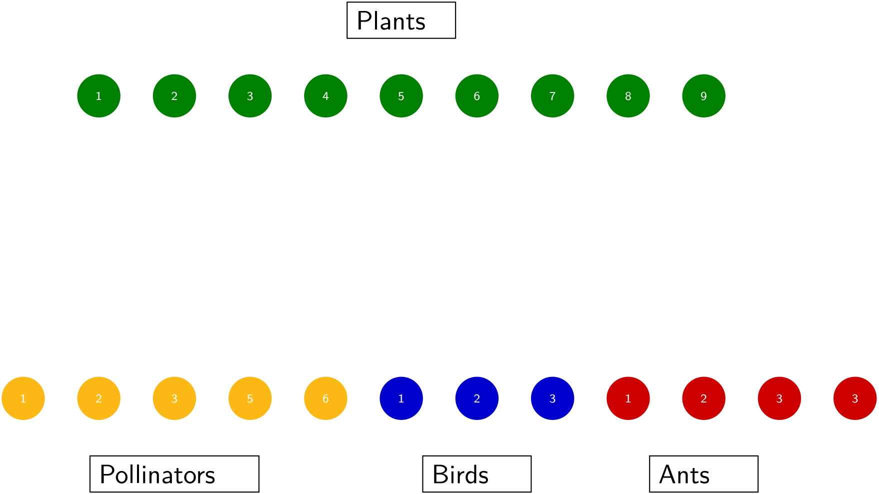
- Each functional group q divided into K_q blocks or clusters
- $$Z_i^q = k$$
  if individual i of functional group q belongs to cluster k.
- Z_i^q are independent random variables :
  $$\mathbb{P}(Z_i^q = k) = \pi_k^q, \tag{1}$$
  with $\sum_{k=1}^{K_q} \pi_k^q = 1$ for any $q = 1, \dots Q$.
- $\boldsymbol{Z} = (Z_i^q)_{i \in [\![1, n_q]\!], q \in [\![1, Q]\!]}$.

Conditionally to the latent variables

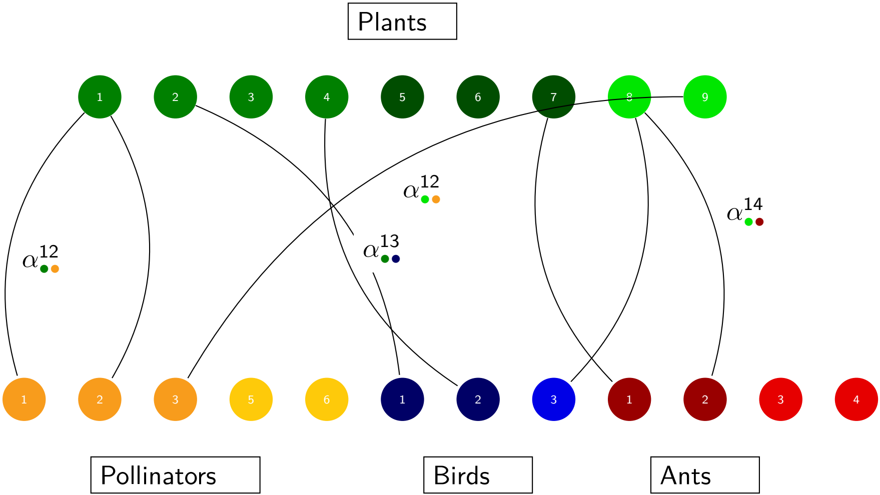
$\forall (i, i', q, q')$, entries of the matrices independant and

$$\mathbb{P}(X_{ii'}^{qq'} = 1 | Z_i^q = k, Z_{i'}^{q'} = k') = \alpha_{kk'}^{qq'} \tag{2}$$

Generative model illustration



Generative model illustration



Remarks

On the model

- Combined extension of SBM and LBM (latent block models)
- Bernoulli may be replaced by any distribution adapted to the data

Remarks

On the model

- Combined extension of SBM and LBM (latent block models)
- Bernoulli may be replaced by any distribution adapted to the data

On the probabilistic dependencies

- Conditionally to $\boldsymbol{Z}$, matrices entries are independent
- But : once $\boldsymbol{Z}$ integrated
  - Dependence between the entries.
  - Different from a standard mixture model.
- Besides, dependence between matrices $X^{qq'}$
- **Consequences** on $\boldsymbol{Z}|\boldsymbol{X}$
  - (Z_i^q) are not independent
  - $\boldsymbol{Z}^q|\boldsymbol{X}$ complicated distribution

Parameters of interest

- Parameters $\boldsymbol{\pi}$, $\boldsymbol{\alpha}$ for given numbers of clusters $K_1, \dots, K_Q$.
- Clustering of the vertices : get the Z_i^q
- Numbers of blocks $K_1, \dots, K_Q$.

Likelihood function

Complete likelihood of $(\boldsymbol{X}, \boldsymbol{Z})$

$$
\begin{aligned}
\ell_c(\boldsymbol{X}, \boldsymbol{Z}; \theta) &= p(\boldsymbol{X}|\boldsymbol{Z}; \boldsymbol{\alpha}) p(\boldsymbol{Z}; \boldsymbol{\pi}) \\
&= \prod_{q,q' \in \mathcal{E}} \prod_{i=1}^{n_q} \prod_{j=1}^{n_{q'}} (\alpha^{qq'}_{Z_i^q, Z_j^{q'}})^{X_{ij}^{qq'}} (1 - \alpha^{q}_{Z_i^q, Z_j^{q'}})^{1 - X_{ij}^{qq'}} && (3) \\
&\quad \times \prod_{q=1}^{Q} \prod_{i=1}^{n_q} \pi^q_{Z_i^q} \, . && (4)
\end{aligned}
$$

Observed likelihood $(\boldsymbol{X})$

$$
\log \ell(\boldsymbol{X}; \theta) = \log \sum_{\boldsymbol{Z} \in \boldsymbol{\mathcal{Z}}} \ell_c(\boldsymbol{X}, \boldsymbol{Z}; \theta) \, . \qquad (5)
$$

Likelihood

$$\log \ell(\boldsymbol{X};\theta) = \log \sum_{\boldsymbol{Z}\in\boldsymbol{\mathcal{Z}}} \ell_c(\boldsymbol{X},\boldsymbol{Z};\theta)\,.$$

Remark

$\boldsymbol{\mathcal{Z}} = \otimes_{q=0\dots Q}\{1,\dots,K_q\}^{n_q} \Rightarrow$ when Q and K_q increase : impossible to calculate

Standard tool to maximise the likelihood : EM algorithm [Dempster et al., 1977]

In this case : variational version of the EM algorithm

Skip VEM

From EM to variational l'EM

EM algorithm

At iteration (t) :

- **Step E** : compute

$$Q(\theta|\theta^{(t-1)}) = \mathbb{E}_{\boldsymbol{Z}|\boldsymbol{X},\theta^{(t-1)}}\left[\log \ell_c(\boldsymbol{X},\boldsymbol{Z};\theta)\right]$$

- **Step M** :

$$\theta^{(t)} = \arg\max_{\theta} Q(\theta|\theta^{(t-1)})$$

Limits

- **Step E** requires calculating $\mathbb{E}_{\boldsymbol{Z}|\boldsymbol{X},\theta^{(t-1)}}\left[\log \ell_c(\boldsymbol{X},\boldsymbol{Z};\theta)\right]$
- Once conditioned by $\boldsymbol{X}$, the $\boldsymbol{Z}$ are not dependent anymore : impossible to compute if $K_1,\dots,K_Q$ et $n_1,\dots,n_Q$ increase.

Variational EM : maximizing a lower bound

Let $\mathcal{R}_{\boldsymbol{X},\boldsymbol{\tau}}$ be any probability distribution on $\boldsymbol{Z}$

Central identity

$$
\begin{aligned}
\mathcal{I}_\theta(\mathcal{R}_{\boldsymbol{X},\boldsymbol{\tau}}) &= \log \ell(\boldsymbol{X};\theta) - \mathbf{KL}[\mathcal{R}_{\boldsymbol{X},\boldsymbol{\tau}}, p(\cdot|\boldsymbol{X};\theta)] \quad \leq \log \ell(\boldsymbol{X};\theta) \\
&= \mathbb{E}_{\mathcal{R}_{\boldsymbol{X},\boldsymbol{\tau}}}[\ell_c(\boldsymbol{X},\boldsymbol{Z};\theta)] - \sum_{\boldsymbol{Z}} \mathcal{R}_{\boldsymbol{X},\boldsymbol{\tau}}(\boldsymbol{Z}) \log \mathcal{R}_{\boldsymbol{X},\boldsymbol{\tau}}(\boldsymbol{Z}) \\
&= \mathbb{E}_{\mathcal{R}_{\boldsymbol{X},\boldsymbol{\tau}}}[\ell_c(\boldsymbol{X},\boldsymbol{Z};\theta)] + \mathcal{H}\left(\mathcal{R}_{\boldsymbol{X},\boldsymbol{\tau}}(\boldsymbol{Z})\right)
\end{aligned}
$$

Variational EM

- Maximization of $\log \ell(\boldsymbol{X};\theta)$ en θ replaced by maximizing the lower bound $\mathcal{I}_\theta(\mathcal{R}_{\boldsymbol{X},\boldsymbol{\tau}})$ in τ and θ.
- **Advantage** : choose $\mathcal{R}_{\boldsymbol{X},\boldsymbol{\tau}}$ such that the we are able to compute the expectation
  - In our case : mean field approximation ; neglect dependencies between the (Z_i^q)
  $$P_{\mathcal{R}_{\boldsymbol{X},\boldsymbol{\tau}}}(Z_i^q = k) = \tau_{ik}^q$$
- [Bickel et al., 2013] : consistency of the variational estimators for SBM. Has been extended to bipartite.

Model selection : penalized likelihood criteria

- Selection of the numbers of blocks $K_1, \dots, K_Q$
- ICL : Integrated Completed Likelihood
- $ICL(\mathcal{M}) = \mathbb{E}_{\boldsymbol{Z}|\boldsymbol{X};\widehat{\theta}\mathcal{M}}\left[\log \ell_c(\boldsymbol{X}, \boldsymbol{Z}; \widehat{\theta}, \mathcal{M})\right] - pen_{\mathcal{M}}$

$$pen_{\mathcal{M}} = \frac{1}{2}\left\{\sum_{q=1}^{Q}(K_q - 1)\log(n_q) + \sum_{(q,q')\in\mathcal{E}} K_{qq'} \log n_{qq'}\right\}$$

[Daudin et al., 2008, Barbillon et al., 2016]

- **In practice** $\tilde{I}CL(\mathcal{M}) = \mathbb{E}_{\mathcal{R}_{\hat{\tau}},z}\left[\log \ell_c(\boldsymbol{X}, \boldsymbol{Z}; \widehat{\theta}, \mathcal{M})\right] + pen_{\mathcal{M}}$
- Stepwize algorithm to select the best model

The dataset

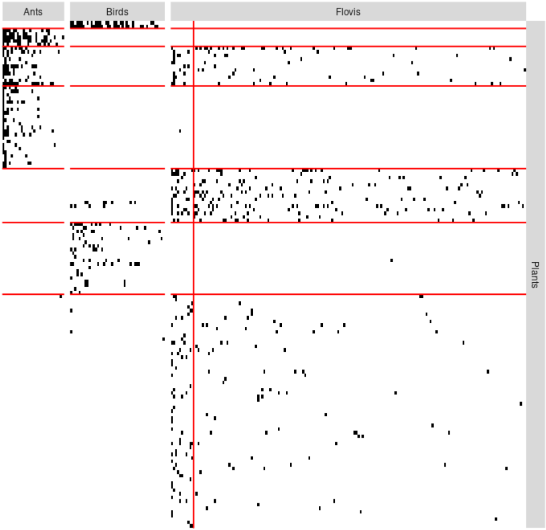
- Weasley Dattilo, Inecol, Jalapa, Mexique [Dáttilo et al., 2016]
-
    - $n_0 = 141$ plants species
    - $n_1 = 30$ ants species
    - $n_2 = 46$ bird species
    - $n_3 = 173$ pollinators species

Results : a mesoscopic view of the network

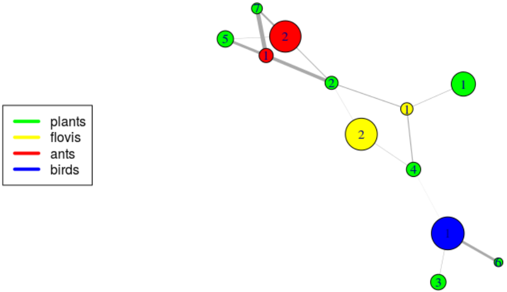
With our model and model selection (a few minutes)

- 7 blocks of plants
- 2 blocks of flower visitors (pollinators)
- 1 block of birds
- 2 blocks of ants

Re-ordered matrix



Mesoscopic view



Now ready for ecological studies such as robustness, comparison of networks, etc...

Conclusion

- Our method : supplies clusterings in multipartite networks without any a priori on the structure
- Accepted for publication in *Statistical Modeling Journal*
- Implemented in two packages
  - original R-package : GREMLINS. On the CRAN
  - Many block models implemented in a new package sbm https://grosssbm.github.io/sbm/ → update including multipartite in a few days
- Handles any structure of generalized multipartite networks, combining binary and weighted interactions.
- Handles missing data.
- OK up to 1000 entities : small networks
- Future : larger networks to apply it to data issued from metabarcoding
- Other structures of multilayer networks

References I

Barbillon, P., Donnet, S., Lazega, E., and Bar-Hen, A. (2016).
Stochastic block models for multiplex networks : An application to a multilevel network of researchers.
Journal of the Royal Statistical Society. Series A : Statistics in Society.

Bickel, P., Choi, D., Chang, X., and Zhang, H. (2013).
Asymptotic normality of maximum likelihood and its variational approximation for stochastic blockmodels.
Ann. Statist., 41(4) :1922–1943.

Dáttilo, W., Lara-Rodríguez, N., Jordano, P., Guimarães, P. R., Thompson, J. N., Marquis, R. J., Medeiros, L. P., Ortiz-Pulido, R., Marcos-García, M. A., and Rico-Gray, V. (2016).
Unravelling darwin's entangled bank : architecture and robustness of mutualistic networks with multiple interaction types.
Proceedings of the Royal Society of London B : Biological Sciences, 283(1843).

References II

Daudin, J. J., Picard, F., and Robin, S. (2008).
A mixture model for random graphs.
Statistics and Computing, 18(2) :173–183.

Dempster, A., Laird, N., and Rubin, D. (1977).
Maximum likelihood from incomplete data via the EM algorithm.
Jr. R. Stat. Soc. B, 39 :1–38.

Kéfi, S., Miele, V., Wieters, E. A., Navarrete, S. A., and Berlow, E. L. (2016).
How structured is the entangled bank ? the surprisingly simple organization of multiplex ecological networks leads to increased persistence and resilience.
PLOS Biology, 14(8) :1–21.

References III

Matias, C. and Miele, V. (2017).
Statistical clustering of temporal networks through a dynamic stochastic block model.
Journal of the Royal Statistical Society : Series B (Statistical Methodology), 79(4) :1119–1141.

Pocock, M. J., Evans, D. M., and Memmott, J. (2012).
The robustness and restoration of a network of ecological networks.
Science, 335(6071) :973–977.