Probabilistic models for ecological networks

Sophie Donnet, Pierre Barbillon & Avner Bar-Hen

14/06/17

Outline

Network data

Networks can account for

- Food web,
- Co-existence networks,
- Host-parasite interactions,
- Plant-pollinator interactions,
- ...

Networks may be or not bipartite: Interactions between nodes belonging to the same or to different functional group(s).

Terminology

A network consists in:

- nodes/vertices which reprensent individuals / species which may interact or not,
- links/edges/connections which stand for an interaction between a pair of nodes / dyad.

A network may be

- directed / oriented (e.g. food web...),
- symmetric / undirected (e.g. coexistence network),
- with or without loops.

This distinction only makes sense for simple networks (not bipartite).

Available data and goal

Available data:

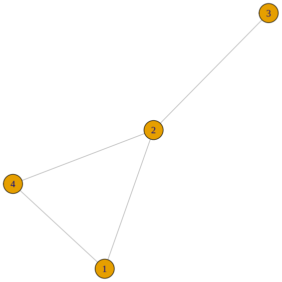
- the network provided as:
  - an adjacency matrix (for simple network) or an incidence matrix (for bipartite network),
  - a list of pair of nodes / dyads which are linked.
- some additional covariates on nodes, dyads which can account for sampling effort.

Goal:

- Unraveling / describing / modeling the network topology.
- Discovering particular structure of interaction between some subsets of nodes.
- Understanding network heterogeneity.
- Not inferring the network !

Network representation and adjacency matrix

$$X = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \end{pmatrix}$$



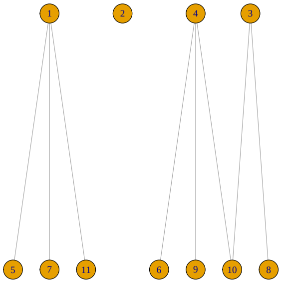
- n rows and n columns,
- symmetric or not.

Bipartite network and incidence matrix

$$X = \begin{pmatrix} 1 & 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 & 1 & 0 \end{pmatrix}$$

- n rows and m columns, rectangular matrix.
- corresponding adjacency matrix $(n+m) \times (n+m)$:

$$\begin{pmatrix} 0 & X \\ X^T & 0 \end{pmatrix}$$



Some common features studied on networks

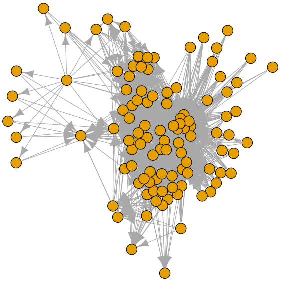
- Degree distribution, can be viewed as a measure of heterogeneity,
- Nestedness: a network is said to be nested when its nodes that have the smallest degree, are connected to nodes with the highest degree, Rodríguez-Gironés & Santamaria (2006)
- Betweenness centrality: for a node, numbers of shortest paths between any pair of nodes passing through this node. Freeman (1979)
- Modularity: is a measure for a given partition of its tendency of favoring intra-connection over inter-connection. $\Rightarrow$ Finding the best partition with respect to modularity criterion. Clauset, Newman & Moore (2004)

All this criterion shall be adapted to:

- directed network,
- bipartite network.

R packages: igraph, sna, vegan.

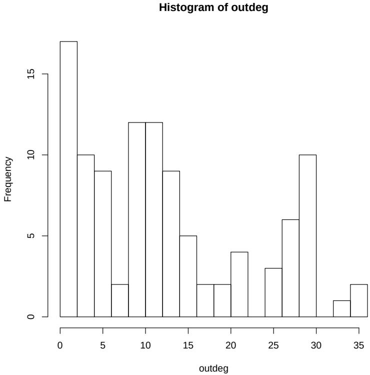
Example Chilean food web

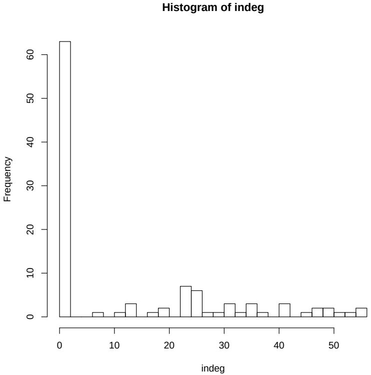


- $n = 106$ species / nodes,
- density of edges: 12.1%.

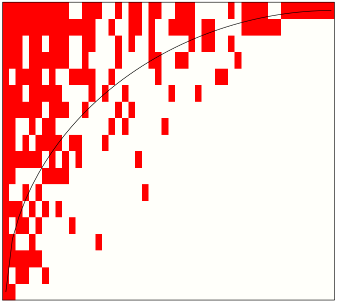
Kéfi, Miele, Wieters, Navarrete & Berlow (2016)

Degree distribution



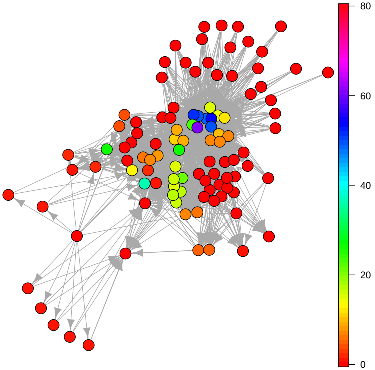


Nestedness



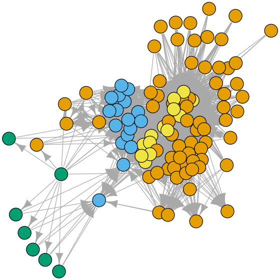
- more generally used on incidence matrices,
- significance of the nestedness index computed by random permutations of the matrix,
- this food web is found to be nested.

Betweenness



Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.000	0.000	0.000	6.604	6.929	59.570

Modularity



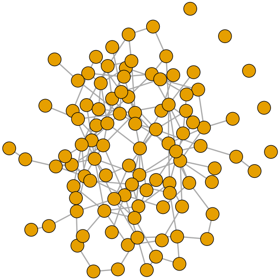
1	2	3	4
69	17	7	13
- very low modularity.

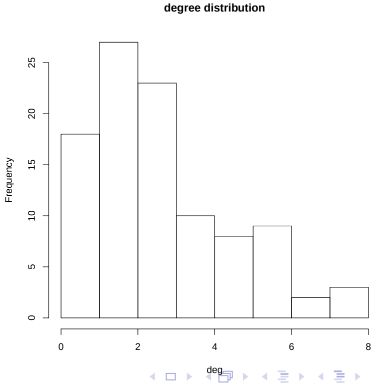
A first random graph model for network: Null model

Erdős-Rényi (1959) Model for n nodes

$$\forall 1 \leq i,j \leq n, \quad X_{ij} \overset{i.i.d.}{\sim} b(p),$$

where b is the Bernoulli distribution and $p \in [0,1]$ a probability for a link to exist.



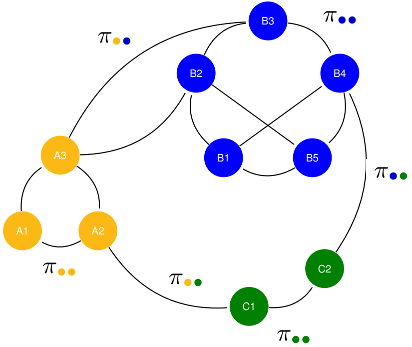


Limitations of an ER graph to describe real networks

- Degree distribution too concentrated, no high degree nodes,
- all nodes are equivalent (no nestedness...),
- no modularity.

Outline

Stochastic Block Model



Stochastic Block Model

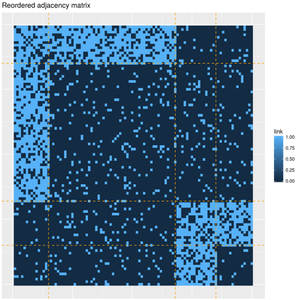
Let n nodes divided into

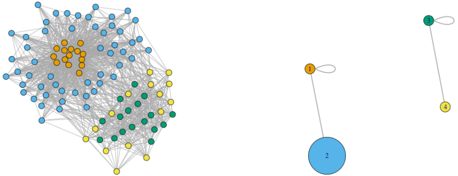
- $\mathcal{Q} = \{{\color{orange}\bullet}, {\color{blue}\bullet}, {\color{green}\bullet}\}$ classes
- $\alpha_{\bullet} = \mathbb{P}(i \in \bullet), \bullet \in \mathcal{Q}, i = 1, \dots, n$
- $\pi_{{\color{orange}\bullet}{\color{blue}\bullet}} = \mathbb{P}(i \leftrightarrow j | i \in {\color{orange}\bullet}, j \in {\color{blue}\bullet})$

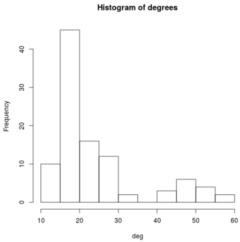
$$Z_i = \mathbf{1}_{\{i \in \bullet\}} \sim^{\text{iid}} \mathcal{M}(1, \alpha), \quad \forall \bullet \in \mathcal{Q},$$

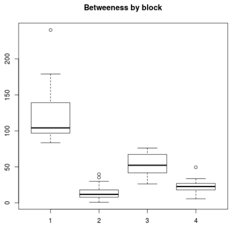
$$X_{ij} \mid \{i \in {\color{orange}\bullet}, j \in {\color{blue}\bullet}\} \sim^{\text{ind}} \mathcal{B}(\pi_{{\color{orange}\bullet}{\color{blue}\bullet}})$$

Some remarkable structure generated with SBM : networks with hubs

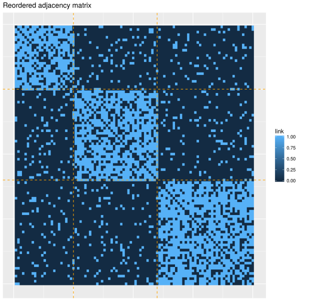


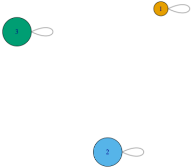


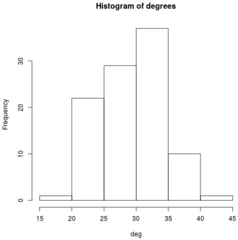




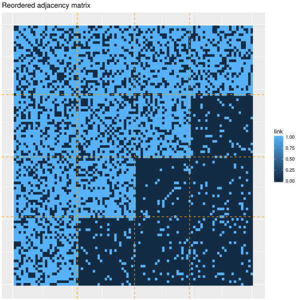
Some remarkable structure generated with SBM : community network

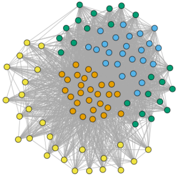


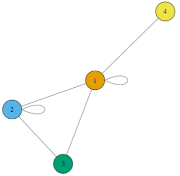


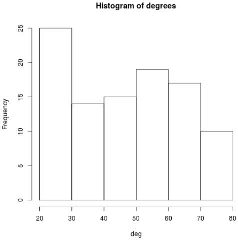


Some remarkable structure generated with SBM : nestedness

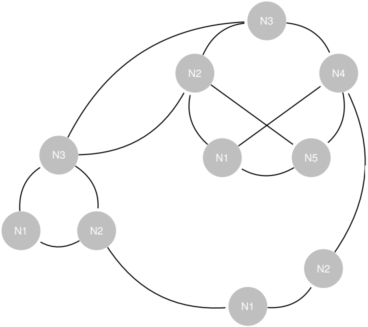








Statistical inference



Stochastic Block Model

Let *n* nodes divided into

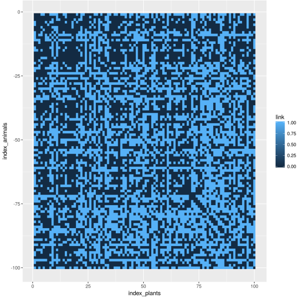
- $\mathcal{Q} = \{\textcolor{orange}{\bullet}, \textcolor{blue}{\bullet}, \textcolor{green}{\bullet}\}$, card($\mathcal{Q}$) known
- $\alpha_{\bullet} = ?$,
- $\pi_{\textcolor{orange}{\bullet}\textcolor{blue}{\bullet}} = ?$

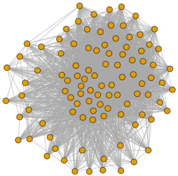
Nowicki, & Snijders (2001), Daudin et al. (2008)

R package: blockmodels.

Statistical inference

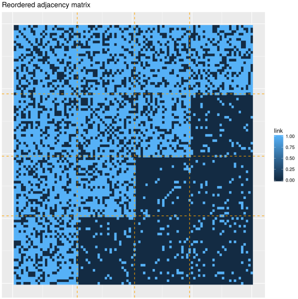
From....

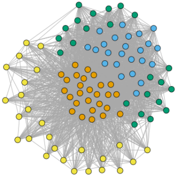




Statistical inference

... to

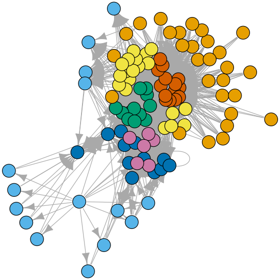




Statistician job

- Find the clusters
- Find the number of clusters
- Practical implementation
- Theoretical results

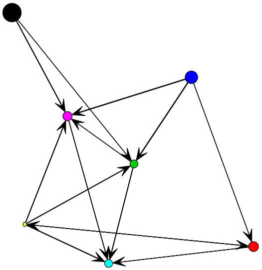
Application to the Chilean food web



- 7 groups/blocks/clusters found,
-

1	2	3	4	5	6	7
28	15	12	19	12	14	6

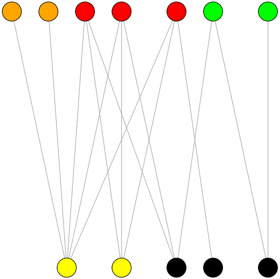
Application to the Chilean food web



Another example for food web in Allesina & Pascual (2009).

Outline

Latent Block Model



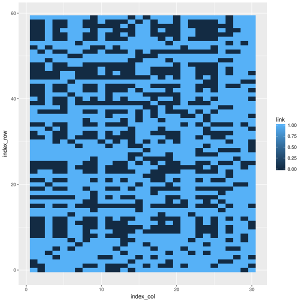
Latent Block Model

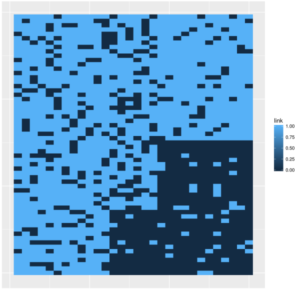
- n row nodes $\mathcal{Q}_1 = \{\color{red}\bullet, \color{orange}\bullet, \color{green}\bullet\}$ classes
- $\alpha_\bullet = \mathbb{P}(i \in \bullet), \bullet \in \mathcal{Q}_1, i = 1, \ldots, n$
- m column nodes $\mathcal{Q}_2 = \{\color{yellow}\bullet, \bullet\}$ classes
- $\beta_\bullet = \mathbb{P}(j \in \bullet), \bullet \in \mathcal{Q}_2, j = 1, \ldots, m$
- $\pi_{\color{red}\bullet\color{yellow}\bullet} = \mathbb{P}(i \leftrightarrow j | i \in \color{red}\bullet, j \in \color{yellow}\bullet)$

$$Z_i = \mathbf{1}_{\{i \in \bullet\}} \sim^{\text{iid}} \mathcal{M}(1, \alpha), \quad \forall \bullet \in \mathcal{Q}_1,$$
$$W_j = \mathbf{1}_{\{j \in \bullet\}} \sim^{\text{iid}} \mathcal{M}(1, \beta), \quad \forall \bullet \in \mathcal{Q}_2,$$
$$X_{ij} \mid \{i \in \color{red}\bullet, j \in \color{yellow}\bullet\} \sim^{\text{ind}} \mathcal{B}(\pi_{\color{red}\bullet\color{yellow}\bullet})$$

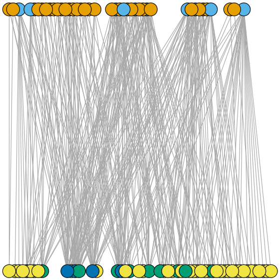
Govaert & Nadif (2008) and R package: blockmodels as well.

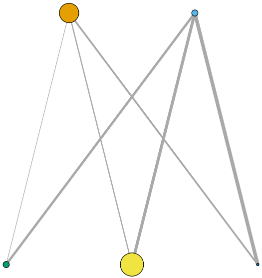
Incidence matrix point of view





LBM for ant-plant data





- 2 blocks found over the 41 ant species,
- 3 blocks found over the 51 plant species.

Blüthgen, Stork & Fiedler (2004)

Outline

Valued-edge networks or multiplex-edge networks

Information on edges can be something different from presence/absence. It can be:

1. a count of the number of observed interactions,
2. a quantity interpreted as the interaction strength,
3. several kind of interactions between nodes (Multiplex networks).

Natural extensions of SBM and LBM for these three cases:

1. Poisson distribution: $X_{ij} \mid \{i \in \bullet, j \in \bullet\} \sim^{\text{ind}} \mathcal{P}(\lambda_{\bullet\bullet})$,
2. Gaussian distribution: $X_{ij} \mid \{i \in \bullet, j \in \bullet\} \sim^{\text{ind}} \mathcal{N}(\mu_{\bullet\bullet}, \sigma^2)$,
3. Bivariate Bernoulli: $(X_{ij}, X'_{ij}) \mid \{i \in \bullet, j \in \bullet\} \sim^{\text{ind}} \mathcal{B}^2(\pi_{\bullet\bullet})$.

Remark: a particular case of multiplex network is dynamic network, Matias & Miele (2015).

Multipartite networks

Incidence matrix

$$X = \left(\begin{array}{c|c|c} X_1 & X_2 & X_3 \end{array} \right),$$

where

- X_1, X_2, X_3 correspond to the bipartite networks with the same functional groups in rows,
- for instance, X_1 is plant-pollinator network, X_2 is plant-ant network and X_3 is plant-seed dispersers network.

- LBM is for bipartite networks,
- When there are more than two functional groups involved in interactions $\Rightarrow$ Multipartite networks.
- Extension of LBM but choice of the number of blocks is more challenging.

Taking into account covariates

Sometimes covariates are available. They may be on:

- nodes,
- edges,
- both.

1. They can be used a posteriori to explain blocks inferred by SBM.
2. Extension of the SBM which takes into account covariates. Blocks are structure of interaction which is not explained by covariates !

If covariates are sampling conditions, case 2 may more interesting.

Probabilistic model for networks in a nutshell

SBM/LBM

- generative models,
- flexible,
- comprehensive models which can be linked to a lot of classical descriptors.

Extension of the binary SBM model are quite natural:

- all the one presented above,
- missing data in the network,
- multi-layers ?