Détection de structures à l'aide de modèles probabilistes sur les graphes

Introduction

Pierre Barbillon

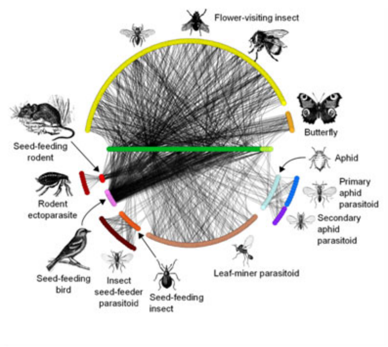
21 juin 2024

Sociology

- **Nodes**: individuals or organizations
- **Edges**: advice, competition, ...
- **Examples of objectives**: characterizing the role of individuals in the network, link their role to covariates

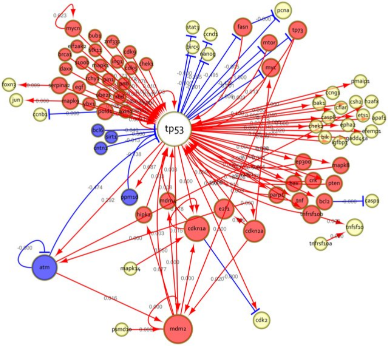
Ecology

- **Nodes:** species (plants or animals)
- **Edges:** predation, pollination, competition...
- **Examples of objectives** : characterizing the structure of the network because it conditions their robustness to the disappearance of species.



Biology

- **Nodes:** genes, metabolites, proteines,
- **Edges:** Regulation, co-expression, reactions,
- **Examples of objectives:** Determine groups of genes co-expressed together under some stresses.



Networks

Graph $G = (V, E, W)$ with

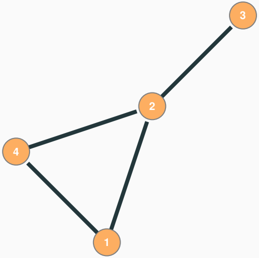
- a set of nodes $V = \{1, \dots, N\}$,
- a set of edges $E \subset V^2$, particular cases: (un)directed, with(out) loop,...
- additional information on edges, $w \in W$ containing weights (number of interactions, positive or negative interaction,...)

Attributes of:

- **nodes**, for any $i \in V$, X_i attributes of a node (taxon, gender, age, social group,...), or information derived from the edges: degree of i,
- **edges**, for any $e = (i, j) \in E$, the edges may have an attribute coming from the two nodes (difference of ages, same gender...,) or particular attribute (date of interaction,...)
- **network**, global attribute derived from the edges mean connectivity, diameter, or an associated variable.

Network encoding/representation

Simple network

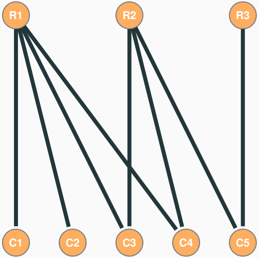


Adjacency matrix:

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

edge list:
E={(1,2),(2,3),(1,4),(2,4)}

Bipartite network



Incidence matrix:

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

edge list:

$$E = \{(R1, C1), (R1, C2), (R1, C3)...\}$$

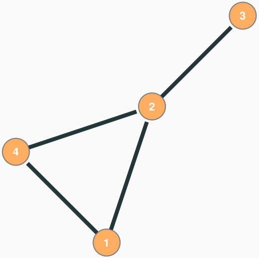
Outline

Objectives

Visualisation and descriptive statistics

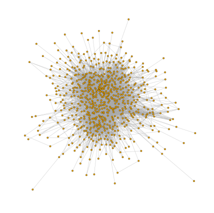
Topics:

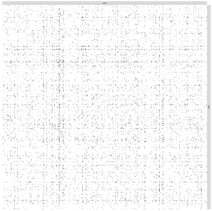
- Network inference, from nodes information determine

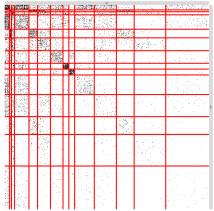


  for $i = 1, \ldots, N$ features $\mathbf{X}_i = (X_{i1}, \ldots, X_{ip}) \rightarrow$

- from the observation of a network determine structure





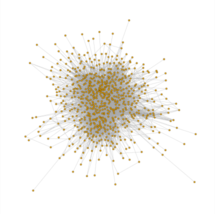


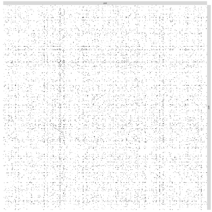
Topics:

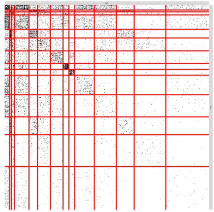
- Network inference, from nodes information determine

for $i = 1, \ldots, N$ features $\mathbf{X}_i = (X_{i1}, \ldots, X_{ip}) \rightarrow$

- from the observation of a network determine structure



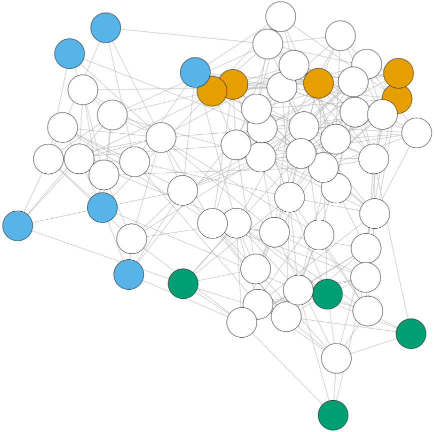




Other Topics: Semi-supervised learning

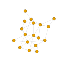
Data: $G = (V, E)$ and labels in $\{1, \ldots, K\}$ for a subset of V,

- learn $f : i \in V \mapsto \{1, \ldots, K\}$,
- leverage the network structure E.



Other Topics: Classification of graphs or regression on graphs

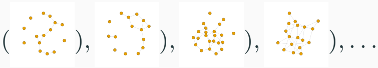
Data:

( $, y_1),$  $, y_2),$  $, y_3),$  $, y_4), \ldots$

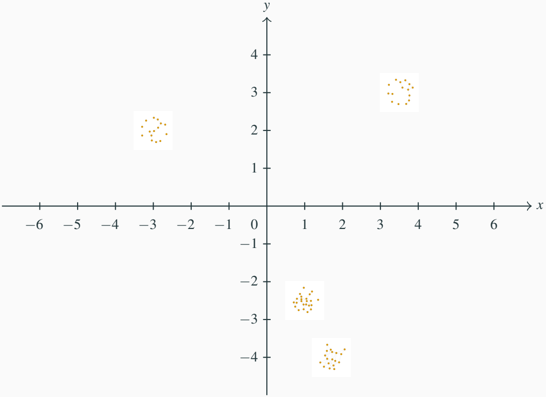
Goal: learn $f : G = (V, E) \mapsto y \in \{1, \ldots, K\}$ or $f : G = (V, E) \mapsto y \in \mathbb{R}$.

Other Topics: Clustering of graphs / Embeddings

Data:



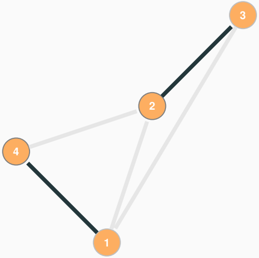
Goal: learn a partition of graphs , learn an embedding:



Other Topics: Predict of dyads, missing links

Data: a graph G with missing or incomplete data.

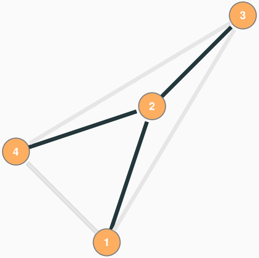
missing data



Adjacency matrix:

$$A = \begin{pmatrix} 0 & \mathtt{NA} & \mathtt{NA} & 1 \\ \mathtt{NA} & 0 & 1 & \mathtt{NA} \\ \mathtt{NA} & 1 & 0 & 0 \\ 1 & \mathtt{NA} & 0 & 0 \end{pmatrix}$$

incomplete data



Adjacency matrix:

$$A = \begin{pmatrix} 0 & 1 & \mathtt{NA} & \mathtt{NA} \\ 1 & 0 & 1 & 1 \\ \mathtt{NA} & 1 & 0 & \mathtt{NA} \\ \mathtt{NA} & 1 & \mathtt{NA} & 0 \end{pmatrix}$$

Goal: Predict NA to $\{0, 1\}$ or predict most likely existing links.

Outline

Objectives

Visualisation and descriptive statistics

Different visualisations of the same graph i

Warning: Visualisation can be misleading!

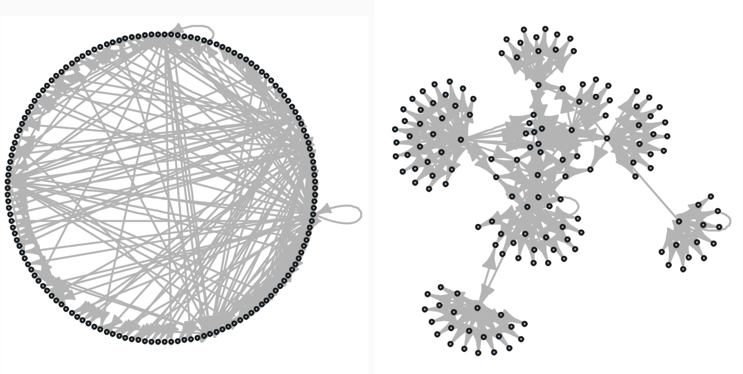


Figure 1: 2 representations of the same blogs network [Kolaczyk and Csárdi, 2014].

Different visualisations of the same graph ii

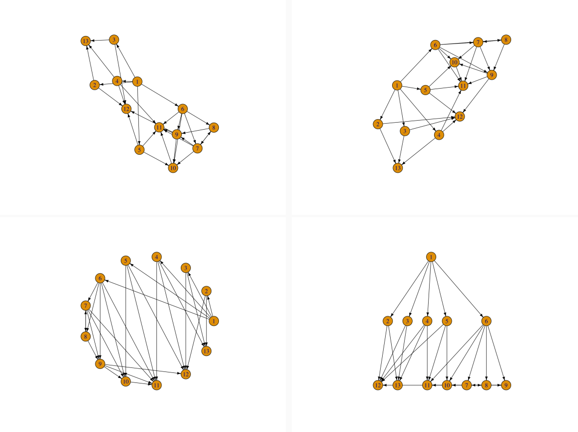
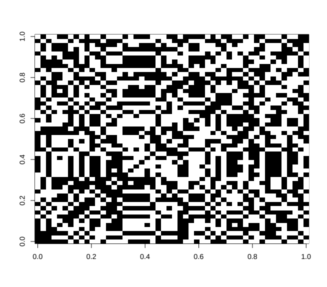


Figure 2: Different visualisations a the food web



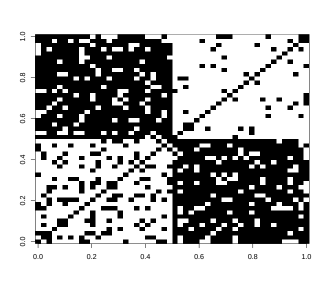
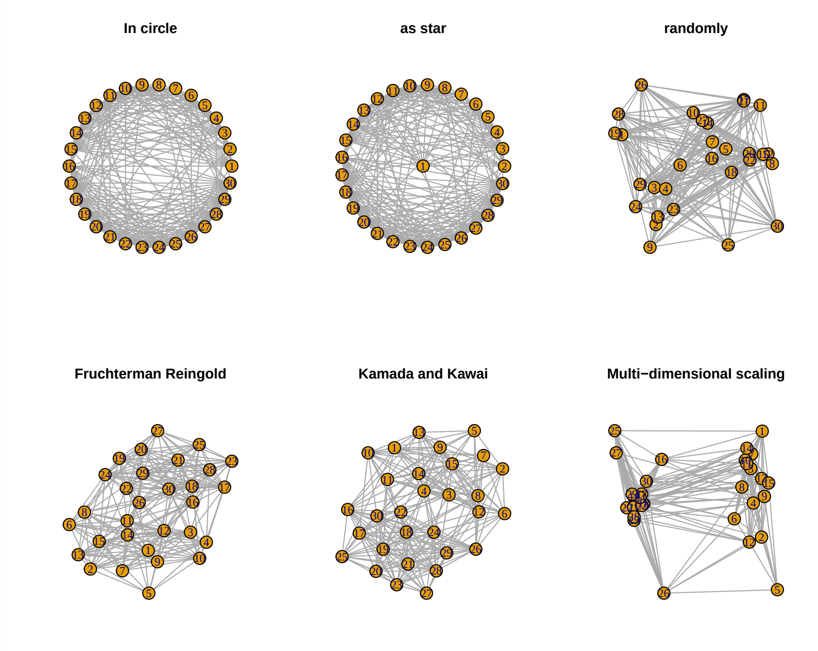


Figure 3: Dotplot representation of a graph: random node numbering (left) and specific permutation of the nodes (right)

Examples of representations



Density / Connectance

A simple binary graph has at most $\binom{n}{2} = n(n-1)/2$ edges.

Its density or connectance is:

$$\text{den}(G) = \frac{|E|}{\binom{n}{2}} = \frac{|E|}{n(n-1)/2}.$$

- the complete graph K_n is the undirected graph with n nodes that contains all possible $\binom{n}{2}$ edges; it has density 1.
- a clique is a complete subgraph in a graph

Neighbors and degrees i

- Neighbors of node $i \in V$ are $\mathcal{N}_i = \{j \in V, j \neq i, \{i,j\} \in E\}$: nodes connected to i in the graph
- Degree of node i is the number of its neighbours $d_i = |\mathcal{N}_i| = \sum_{j \neq i} A_{ij} = \sum_{j \neq i} A_{ji}$
- In directed graphs, one may define indegrees and outdegrees: $d_i^{out} = \sum_{j \neq i} A_{ij}$ and $d_i^{in} = \sum_{j \neq i} A_{ji}$
- Degrees are obtained as rowSums or colSums of adjacency matrix
- We always have $\sum_{i=1}^{n} d_i = 2|E|$
- Average degree $\bar{d} = n^{-1} \sum_{i=1}^{n} d_i$
- a d-regular graph has constant degree d (ex infinite grid)
- Hubs (informal) a hub is a large degree node in a graph

Degree distributions only loosely characterize graphs

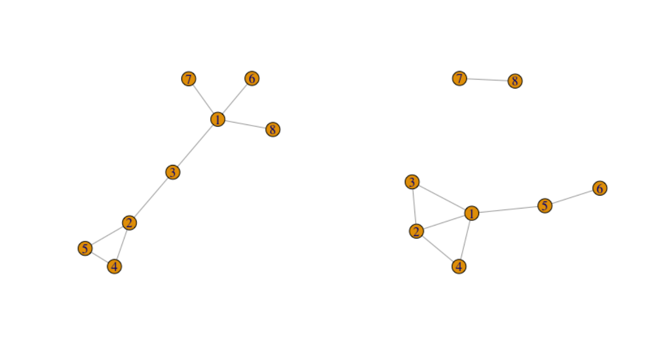
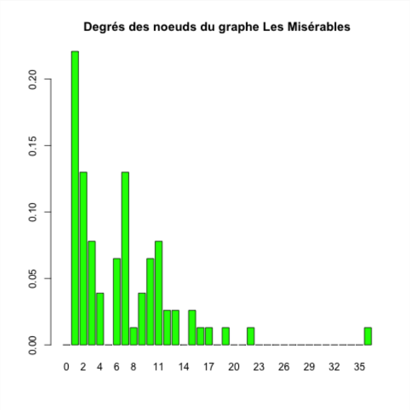


Figure 4: Example of 2 graphs with same degree sequence.

Neighbors and degrees iii

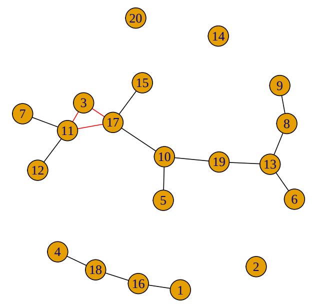
Graphs often show degree distributions with heavy tails, such as scale-free distributions ($\mathbb{P}(d_i = k) = c/k^{\gamma}$, where $\gamma > 0$ is the exponent of the power law)



Paths, connectivity, diameter i

Paths

- A path between nodes $i, j \in V$ is a sequence of edges $e_1, \ldots, e_k \in E$ such that e_t and e_{t+1} share a node, $i \in e_1$ and $j \in e_k$. Its length is k;
- A cycle is a path that connects a node to itself; (ex: a self-loop is a cycle of length 1)



Connectivity

- A set of nodes $C = \{v_1, \ldots, v_k\} \in V$ such that there exists a path between any 2 nodes $v_i, v_j \in C$ is a connected component (cc);
- Any graph may be decomposed into a unique collection of maximal cc;
- An isolated node forms a (maximal) cc;
- There are at most $n - |E|$ such maximal cc;
- When there is a unique cc, the graph is connected;
- Giant component (informal): In a sequence of graphs G_n each with n nodes, let C_n be the largest mcc in G_n. We say that C_n is a giant component if its relative size $|C_n|/n$ does not tend to 0 as n increases;

Diameter

- the distance ℓ_{ij} between 2 nodes $i, j \in V$ is the length of the shortest path between i, j (and $+\infty$ if the nodes are not in the same cc)
- the average distance in the graph is $\bar{\ell} = 1/(n(n-1)) \sum_{i,j} \ell_{ij}$
- diameter diam$(G) = \max\{\ell_{ij}; i, j \in V\}$;
- It's finite only if the graph is connected;
- Small-world property (informal): a graph has the small-world property whenever $\bar{\ell}$ is of the order of $\log(n)$;
- See the small-world experiment by Stanley Milgram; and its modern version: three and a half degrees of separation (see https://research.facebook.com/blog/2016/2/three-and-a-half-degrees-of-separation/).

Clustering coefficients, transitivity, centrality i

- Let H_i be the subgraph induced by the neighbors of node $i \in V$, i.e. $H_i = (\mathcal{N}_i, E_i)$ where $\mathcal{N}_i$ is the set of neighbors and E_i set of edges $\{j, k\} \in E$ st $j, k \in \mathcal{N}_i$.
- Clustering coefficient C_i is the number of edges $|E_i|$ between neighbors of node i divided by the maximum of such number $d_i(d_i - 1)/2$; i.e.

$$C_i = \begin{cases} \frac{2|E_i|}{d_i(d_i-1)} & \text{if } d_i \geq 2, \\ 0 & \text{otherwise} \end{cases}$$

- It is the connectance of the subgraph induced by the neighbors of i; thus $C_i \in [0, 1]$
- the average clustering coefficient is $\bar{C} = \frac{1}{|V|} \sum_{i \in V} C_i$
- Transitivity is

$$T = \frac{\text{Nb of triangles}}{\text{Nb of triplets of connected nodes}}$$

Clustering coefficients, transitivity, centrality ii

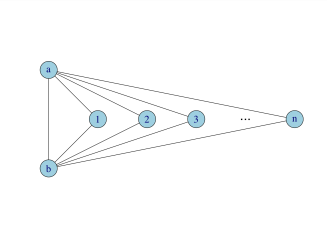


Figure 5: Here $C_i = 1$ for all nodes except a, b and thus $\bar{C}$ tends to 1. However T tends to 0.

Centrality of nodes

- Degree centrality $C_D(i) = d_i$
- Closeness centrality $C_P(i) = \left(\sum_{j \in V} \ell_{ij}\right)^{-1}$, where ℓ_{ij} is the distance between i, j
- (Node) Betweenness centrality $C_B(i) = \sum_{j,k: j \neq k \neq i} \frac{g_{jk}(i)}{g_{jk}}$, where g_{jk} is the number of shortest paths from j to k, and $g_{jk}(i)$ is the number of shortest paths from j to k that go through i;

Beware that those quantities are not normalised and strongly depend on the order of the graph.

Edge betweenness

$C_B(e) = \sum_{j,k:j \neq k \neq i} \frac{g_{jk}(e)}{g_{jk}}$, where g_{jk} is the number of shortest paths from j to k, and $g_{jk}(e)$ is the number of shortest paths from j to k that go through edge e.

This quantity is linked to modularity.

Kolaczyk, E. D. and Csárdi, G. (2014).
Statistical analysis of network data with R, volume 65.
Springer.