

Statistical Learning with Complex Data



Pr. Charles BOUVEYRON

Professor of Statistics
Chair of the Institut 3IA Côte d'Azur
Université Côte d'Azur & Inria

✉ charles.bouveyron@univ-cotedazur.fr
🐦 @cbouveyron

Outline

1. Networks

2. Texts

3. Images

The analysis of (social) networks

- The first analyses of (social) networks were done by sociologists → Poreno (the end of 19th c.)
 - ↳ he was working on suicide, religion, ...
- real collected data by Poreno and sucesors in 1930's
 - schools
 - companies
- Radcliffe-Brown asked his colleagues to build data base systematically.

The analysis of (social) networks

Sociology → Economists, ... Humanities

↳ many fields move to networks
and data

↳ epidemiology, physics, biology, ...

Mathematics: graph theory (18th century).

↳ Euler

Applications: Chemistry, Historical Sciences, ...

A few examples...

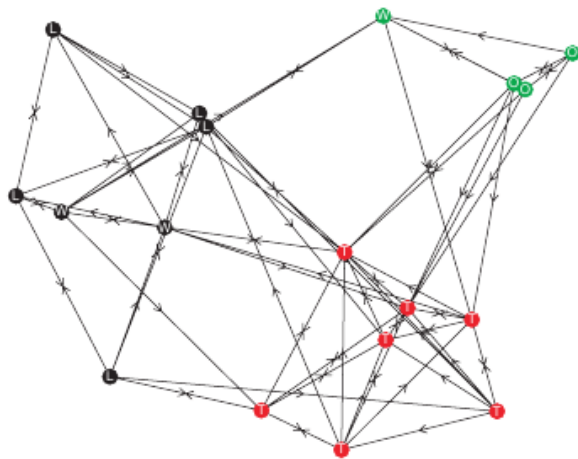


Figure: The Sampson Monks (1969)

A few examples...

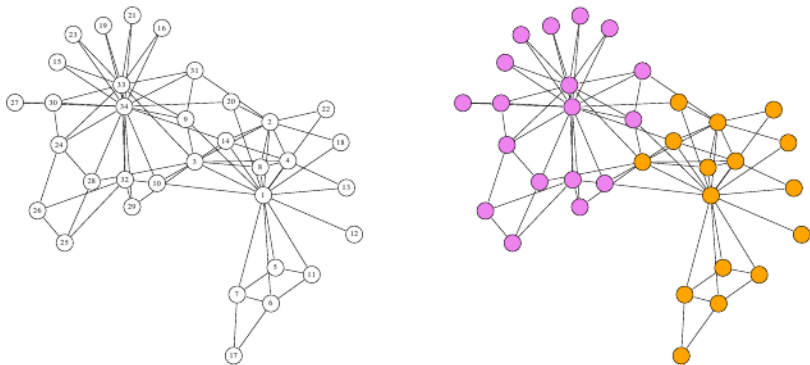


Figure: The Zachary *et al.* karate club (1977)

A few examples...

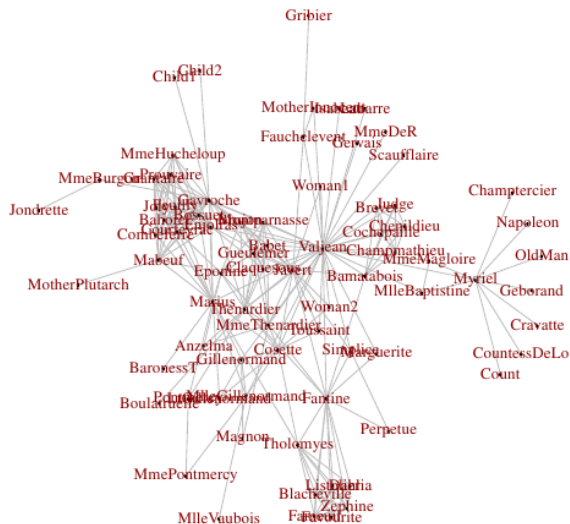


Figure: The network of *Les Misérables* (Knuth *et al.*, 1993)

A few examples...

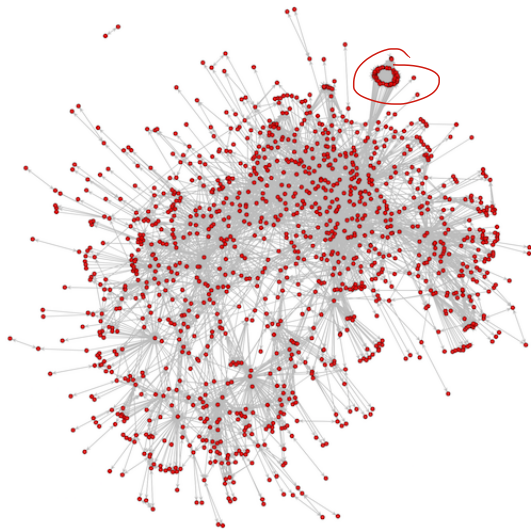


Figure: The Bishop Network (Bouveyron *et al.*, 2015)

A few examples...

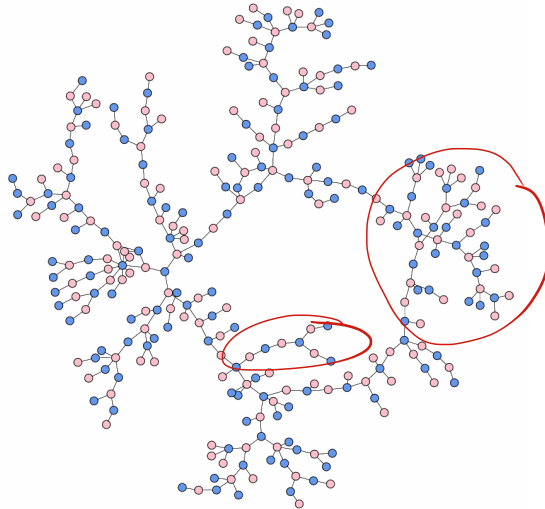


Figure: The dating network (Bearman *et al.*, 2004)

A few examples...

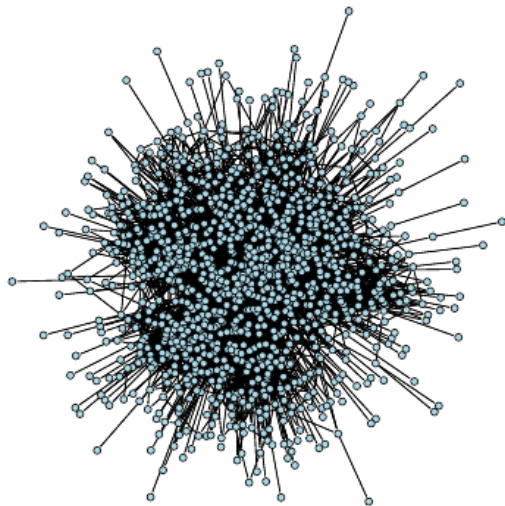


Figure: The Rovira University Email network (Guimera *et al.*, 2003)

For which applications?

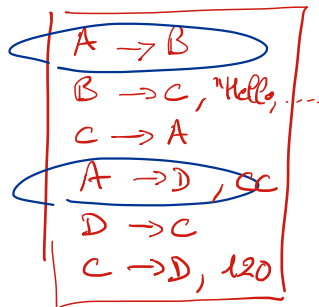
- sociology / economic / health studies
- ads / commercial studies / actions
- defense and security
- biology / epidemiology
- ...

Where to find networks?

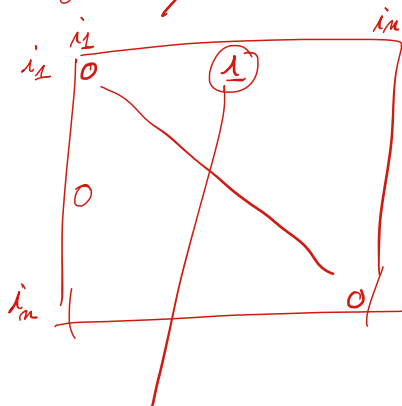
- social network (Twitter, Facebook, ...)
- interaction data (Emails, text messages, phone calls, logs, IOT)
- various documents (Les Misérables, Panama Papers, ...)

Characterizing networks

- a graph:



- adjacency matrix

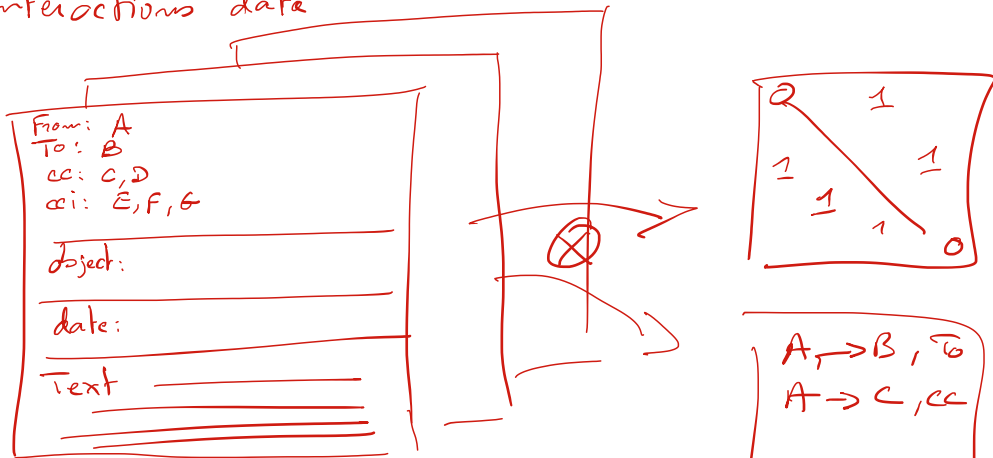


$A_{ij} = 1$ if $i \rightarrow j$

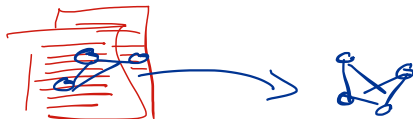
$A_{ij} \neq A_{ji}$

Characterizing networks

— interactions data



— Text documents



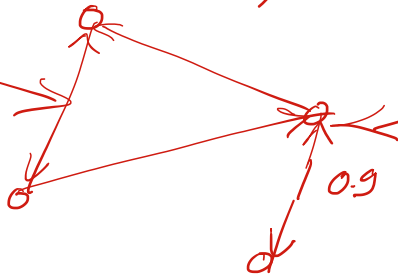
(very expensive!)

Characterizing networks

- nodes
- edges (maybe with extra informations)
(maybe directed)

Types of networks :

- undirected
- directed
- dynamic network
- multi-networks
- bipartite network

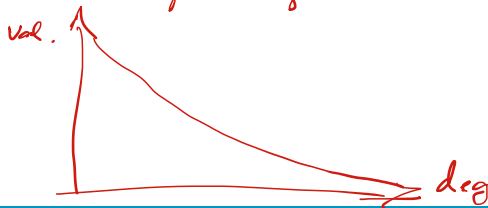
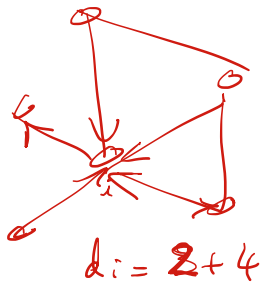


Characterizing networks

- degree distribution:

Def: the degree of a node

$$d_i = \underbrace{\sum_{j \neq i} A_{ij}}_{\text{output degree}} + \underbrace{\sum_{j \neq i} A_{ji}}_{\text{input degree}}$$



Characterizing networks


Density of a network:

$$D_G = \frac{\sum_{i \neq j=1}^n A_{ij}}{n(n-1)} \in [0, 1]$$


\Rightarrow it is interesting to also evaluate the density locally, on small part of the network, to detect communities.

Characterizing networks

- cliques / stars

 2-cliques

 3-cliques

 4-cliques

 $\rightarrow (369, 28, 4, \textcircled{2})$



2-star



3-star



4-star



5-star

How to manipulate networks?

In R, there are several libraries:

- For manipulating:

- igraph
- network
- sna

- For visualisation and clustering

- latentnet

- mixer

- lda

install.packages(—)