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

The analysis of (social) networks

The story of social network analysis started in the 19th century with the seminal works of sociologists:

- the first researchers were Durkheim & Tönnies who studied the link between the individual action with society. (religion, suicide, ...)
- in 1930, Moreno was the first to advocate for the massive use of data in sociology. He in particular studied with network data how small societies (schools, companies, ...) behave.

The analysis of (social) networks

in parallel, the graph theory is something extensively studied in Mathematics for centuries:

- Euler in the 18th century formalized the basis of graph theory.
- it is now a well recognized subfield of Mathematics
- with applications in many scientific fields: Biology, Chemistry, ...

⇒ but networks are not just graphs!

A few examples...

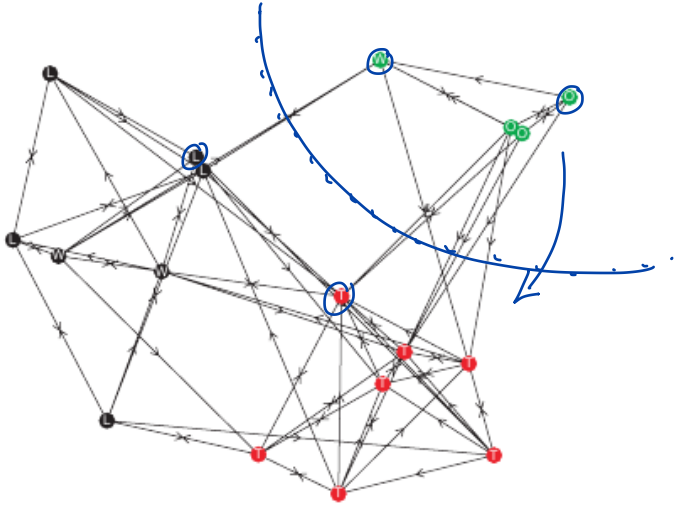


Figure: The Sampson Monks (1969)

A few examples...

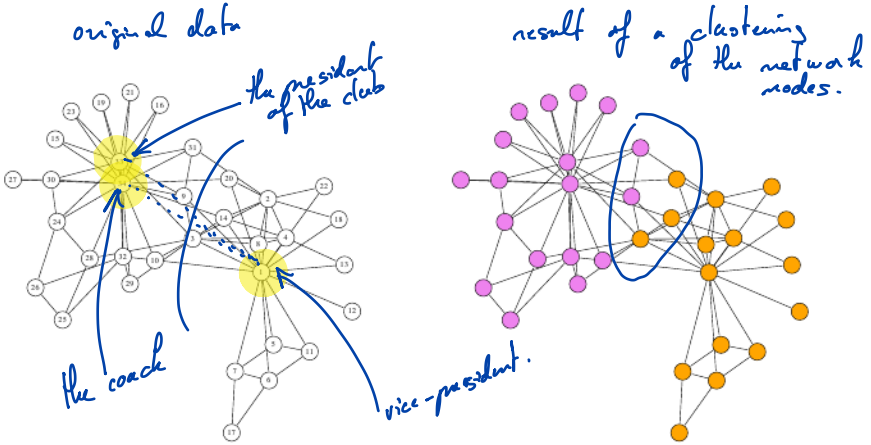


Figure: The Zachary *et al.* karate club (1977)

A few examples...

This work illustrates that most networks are in fact described in several sources / documents.

In this case, there is an important work in modeling / encoding the relationship between the individuals!

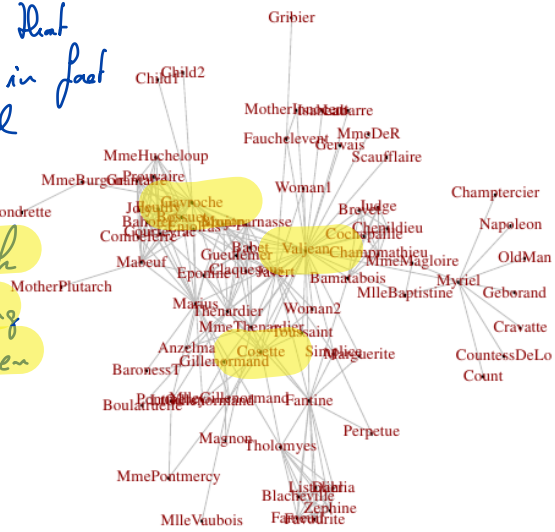


Figure: The network of Les Misérables (Knuth *et al.*, 1993)

A few examples...

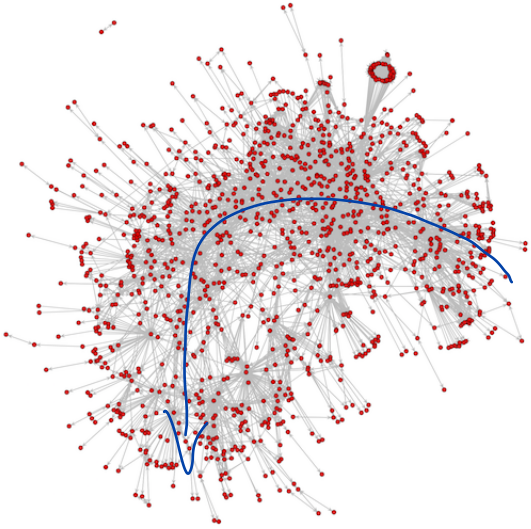


Figure: The Bishop Network (Bouveyron *et al.*, 2015)

A few examples...

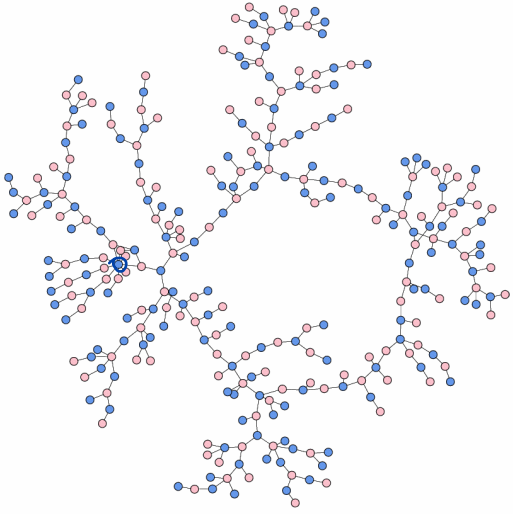


Figure: The dating network (Bearman *et al.*, 2004)

A few examples...

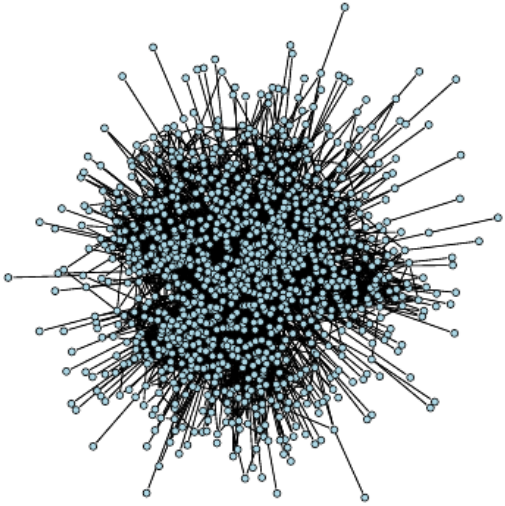


Figure: The Rovira University Email network (Guimera *et al.*, 2003)

From those examples, we can highlight that:

- networks can be directly observed or reconstructed from sources.
- the structure of networks may be extremely different, in particular in term of density.
- network analysis has very different application fields, ranging from sociology, economics to history or medicine.

For which applications?

- medicine : public health, epidemiology, ...
- Biology : modeling of drugs, ...
- social sciences : model and understand some phenomenon.
- Marketing : identification of groups of clients, of influences, ...
- Fraud detection : finance, bank, insurance
- Security : counter-terrorism.

Where to find networks?

Networks can be found under different forms:

- graph (simplest)
- adjacency matrix / socio-matrix (simple as well)
- transactional data (less simple, most situations)
- different sources of different types (most costly)
  - ↳ 1 or several documents
  - ↳ texts, tweets, text messages, images, phone calls.

Some examples:

- social network like twitter → graph
- emails of a campaign → transactional data
- Bishop networks → different sources.

Outline

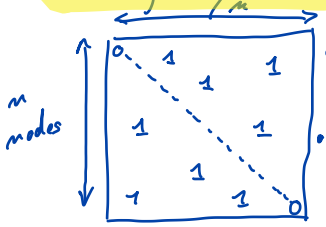
Characterizing networks

- a graph: a text file listing the interactions between the nodes

```
1;2
2;1
1;3
1;4
```

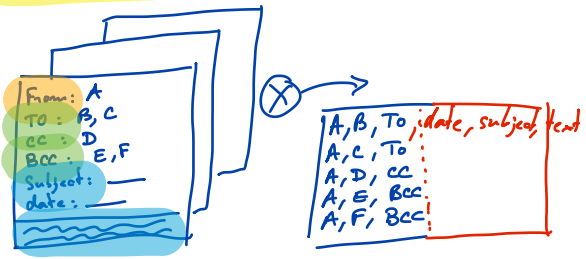
we list here all directed edges.

- an adjacency matrix:



- $A_{ij} = 1$ if i is linked to j;
- $A_{ij} \neq A_{ji}$ if the network is directed, $\forall i,j$

- transactional data:

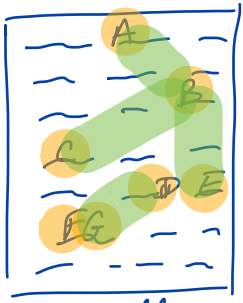


A collection of structured data from which it is clear how to extract the relationships.

⇒ this task needs to write a simple script to transform the transactional data in a graph.

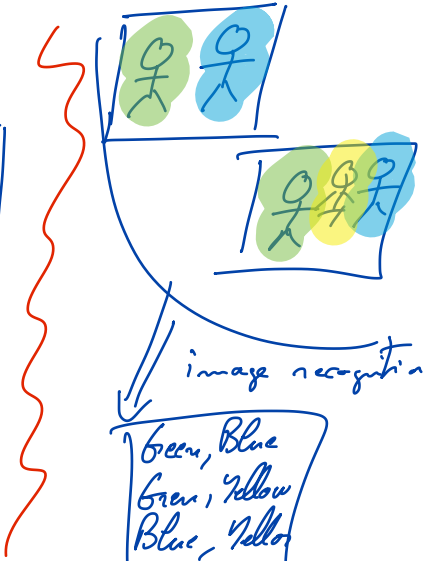
Characterizing networks

- from different sources:



NLP
+ other rules

A, B
B, C
D, F
D, G
B, E



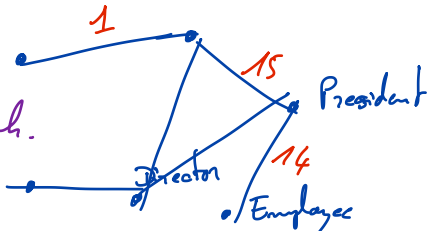
Characterizing networks

A network is composed of:

- nodes (individuals)
- edges (relationships)

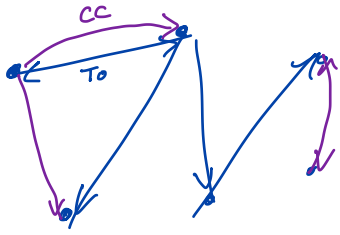
} a graph.

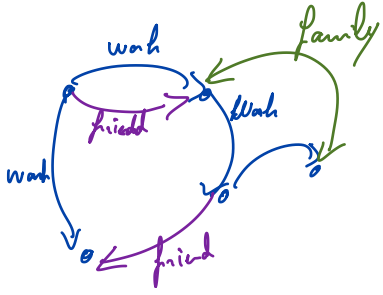
- extra information on nodes or/and edges (covariates)



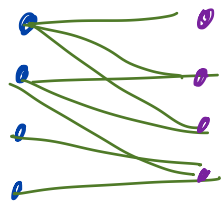
The different types of networks:

- directed and undirected networks.
- dynamic and static networks.
- multiple networks (different types of connections between a set of nodes)
- bipartite networks.





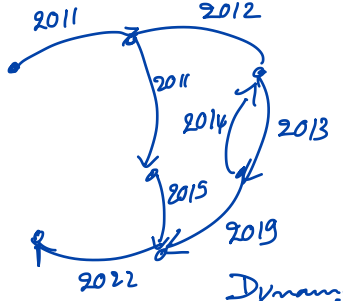
Multiple networks.



patients

Doctors

Bipartite network.



Dynamic network.

Characterizing networks

A first way to characterize a network is to compute general statistics for it:

- degree of a node d_i : it measures the "importance" the centrality of the node in the network

  * $d_i = \sum_{j \neq i} \mathbb{1}\{x_i \sim x_j\} = \sum_{j \neq i} A_{ij} \quad \in \{0, n-1\}$ if undirected

    $= \sum_{j \neq i} A_{ij} + \sum_{j \neq i} A_{ji} \quad \in \{0, 2(n-1)\}$ if directed

$\Rightarrow$ in most of "natural" networks, the distribution of the degrees follows a power law.

Characterizing networks

The notion of density of the network is another way to describe it:

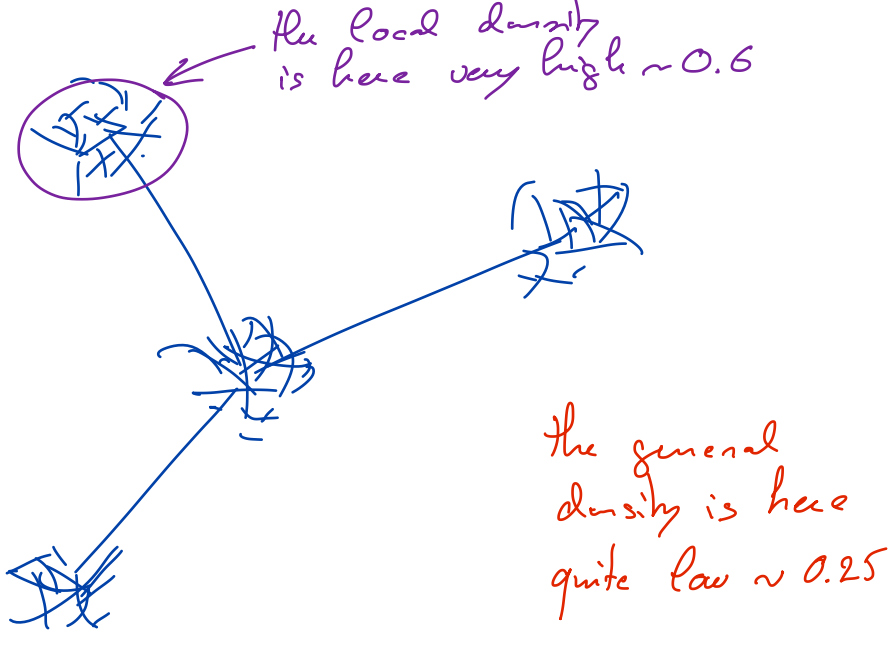
$$d_G = \frac{\sum_{i=1}^{n} \sum_{j=1, j \neq i}^{n} A_{ij}}{n(n-1)}$$

$\sum_{i=1}^{n} \sum_{j=1, j \neq i}^{n} A_{ij}$ ← the total edges in the network

$n(n-1)$ ← maximum nb of connections in a directed network

$$d_G \in [0, 1]$$

Rmk: The density could be also computed for some parts of a network, and the local densities may be very different ⇒ the small world effect


the local density
is here very high ~ 0.6
the general
density is here
quite low ~ 0.25

How to manipulate networks?

To manipulate networks with R :

- igraph
- network
- sna