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 late 19th century with the seminal work of sociologists

- the first researchers to consider to consider those kinds of data are Durkheim & Tönnies who studied the link between individual actions in a society (religion, work, suicide, ...)

- in 1930, Moreno was the first sociologist to advocate for the massive use and collection of network data in Sociology. He in particular studied and collected data within small societies (schools, companies, ...)

The analysis of (social) networks

In parallel, in Mathematics, the graph theory is well know and established for centuries:

- Euler in the 18th century formalized the basis of graph theory.
- it is since then a well recognized field of Mathematics.
- with applications in many scientific fields: Biology, chemestry, transportation, ...

⚠ Networks are not just graphs!

A few examples...

- Those data were collected by Sampson during his Ph.D
- within a monastery of 18 monks.
- at the last time step, the 4 green Monks were excluded from the Monastery.

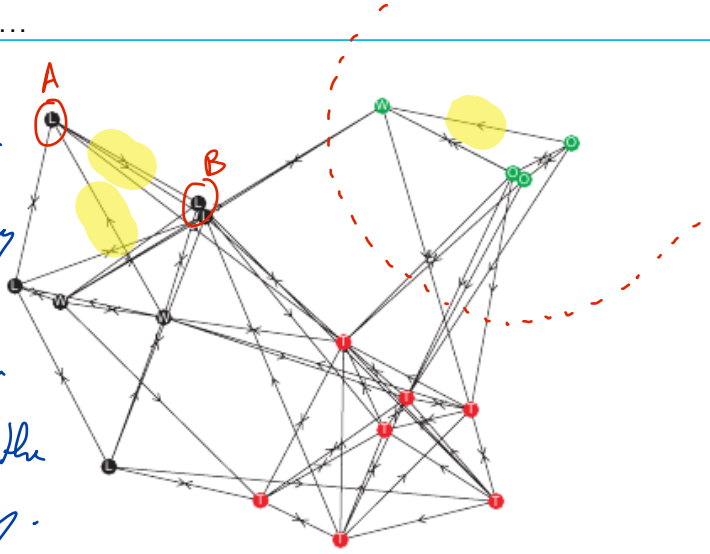


Figure: The Sampson Monks (1969)

A few examples...

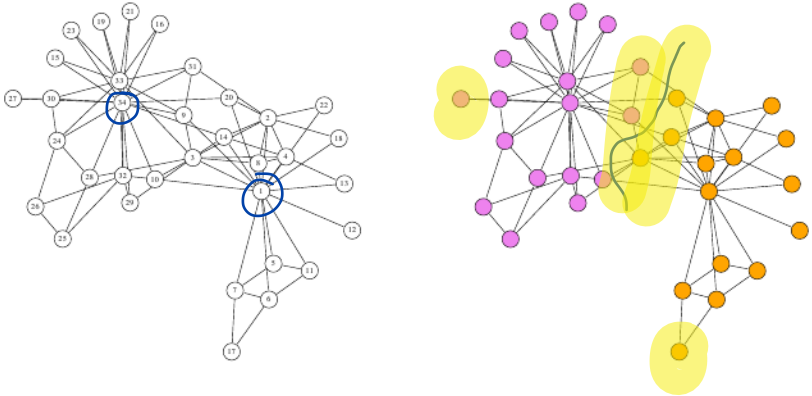


Figure: The Zachary *et al.* karate club (1977)

A few examples...

- This network was reconstructed from the novel.
- There is a link between 2 characters if they appear in the same chapter.

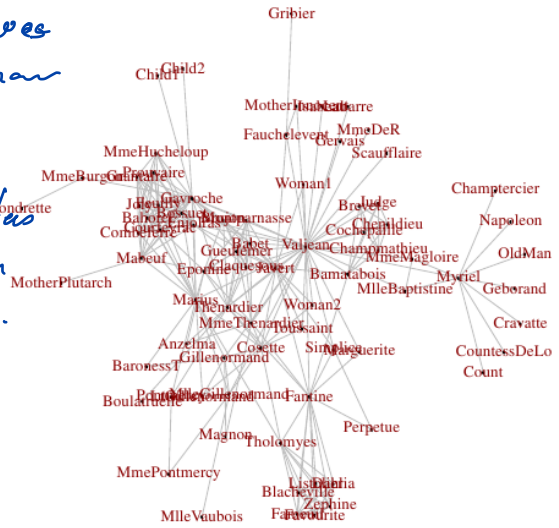


Figure: The network of Les Misérables (Knuth *et al.*, 1993)

A few examples...

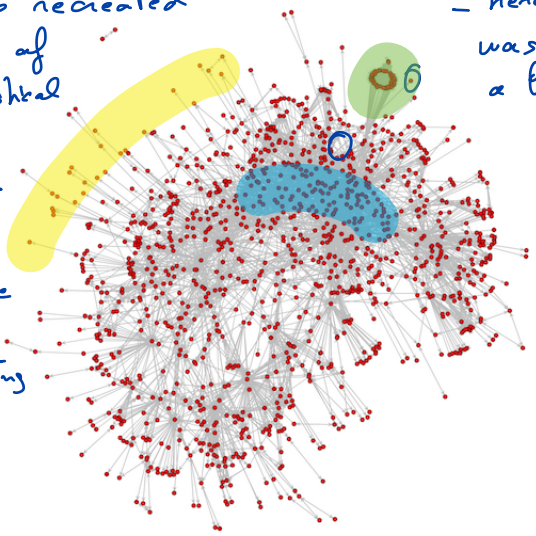


Figure: The Bishop Network (Bouveyron *et al.*, 2015)

A few examples...

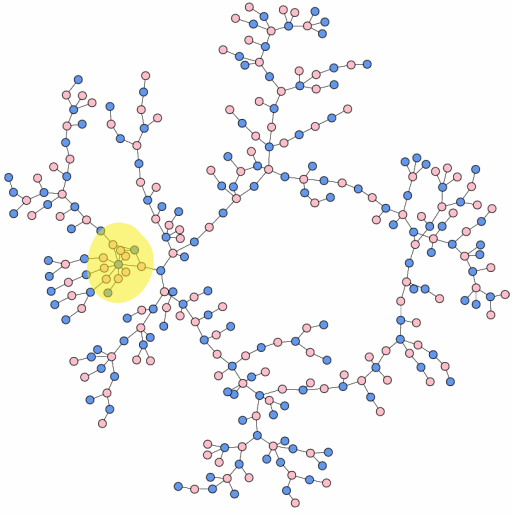


Figure: The dating network (Bearman *et al.*, 2004)

A few examples...

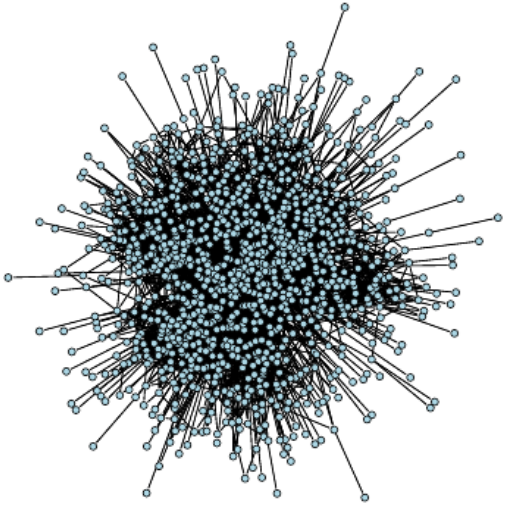


Figure: The Rovira University Email network (Guimera *et al.*, 2003)

For which applications?

Networks can be observed and collected in a variety of contexts to answer different questions:

- Medicine: model some drugs, the spread of a disease, ...
- marketing: identify some groups of clients, of influencers, ...
- Social sciences: model and understand specific phenomenom (religion, suicide, ...)
- Security: contre-terrorism, influence, intelligence, ...
- Fraud detection: finance, insurance, bank,

Where to find networks?

Network data can be found under different forms:

- graph (simplest!)
- adjacency matrix (a socio-matrix, single as well!)
- transactional or relation data (less simple, most of the situations!)
- different unstructured sources of different types (very frequent but naturally very costly!)
  - ↳ 1 or several documents.
  - ↳ texts, tweets, text messages, images, phone calls, ...

Some examples:

- Twitter / Facebook : → graphs
- emails of a University → transactional data
- Bishop networks → different unstructured sources.

Outline

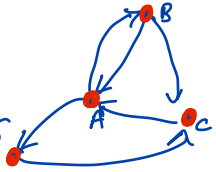
Characterizing networks

- a graph: it is just a texte file listing the interactions between the nodes.

```
A ; B
B ; A
B ; C
C ; A
A ; D
D ; C
```

we list here all directed edges between the nodes

From this graph data, we can get a visualization by using a vis. algorithm.



- an adjacency matrix: it is a square matrix where each row and each column corresponds to an individual, and the entry of the matrix are 0/1 depending on the link between two nodes

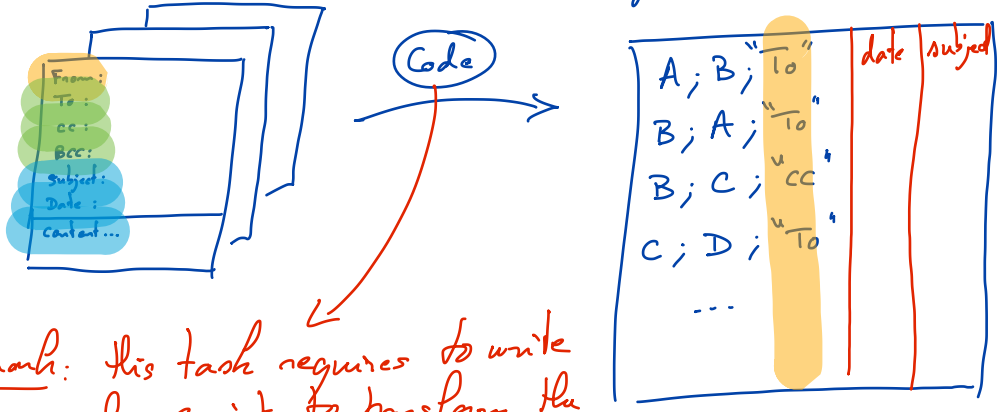
	A	B	C	D	E
A	0	1			
B	1	0	1		
C		0			
D					
E					0

- $A_{ij} = 1$ if $i \to j$
- $A_{ij} = 0$ if $i \not\to j$

⚠ if the network is not directed, then the matrix is symmetric.

Characterizing networks

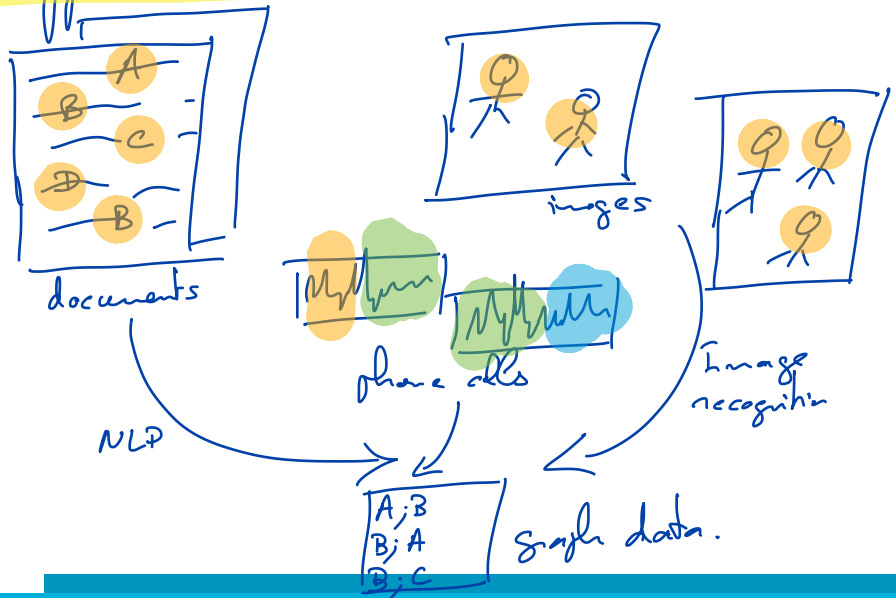
- **transactional data:** a collection of structured data from which we can extract the relationships.



Remark: this task requires to write a simple script to transform the transactional data into a graph data.

Characterizing networks

- from different unstructured sources:



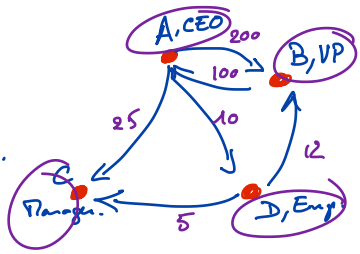
Characterizing networks

A network is composed of:

- nodes (individuals)
- edges (relationships)

} a graph.

+ extra information on both nodes and edges. (covariates).

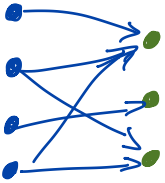


this is a network.

Characterizing networks

We also have different types of networks:

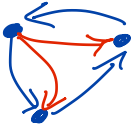
- directed / undirected networks
- normal or bipartite networks
- static or dynamic networks.



clients products

Fig. A bipartite network

- multiple networks



Characterizing networks

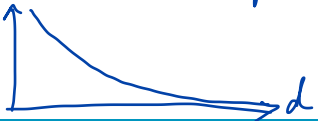
A first way to characterize a network is to compute some general statistics for it:

- degree of a node i: the degree d_i measures the importance of the node in the network.

$$d_i = \sum_{j \neq i} \mathbb{1}\{i \to j\} + \sum_{j \neq i} \mathbb{1}\{i \leftarrow j\} = \sum_{j \neq i} \{A_{ij} + A_{ji}\}$$

$$\in [0, 2(n-1)] \text{ if directed.}$$

Remark: in most "natural" networks, the distribution of the degrees follows a power law



Following this idea, it is possible to derive the notion of density:

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

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

$n(n-1)$ ← the maximum number of connection in a directed network.

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

How to manipulate networks?

To manipulate networks, we can use softwares such as R or Pyhton.

In R, we will make use of:

- igraph
- network
- sna