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

With the development of internet, communication networks, social networks, ..., the analysis of network data has become an important field.

The kind of applications which are targeted:

- marketing on social network.
- fight against cyberbullism
- counter-terrorism
- ...

The history of (social) network analysis started with sociologists

The analysis of (social) networks

Among the seminal works, we can cite:

- Durkheim & Tönnies tried to link the individual action with the society life (religion, suicide, ...) ... in theory.
- in 1930, Moreno was the first to advocate for the massive use of network data in Sociology. He in particular studied (and recorded data) in small societies (schools, companies, ...)

In parallel of these similar works, the theory of graphs in Mathematics was present for centuries:

- Euler in the 18th century.
- graph theory is a clean sub-field of Maths.
- applications of graph theory are large: Biology, chemestry.

Networks are not just graphs!

A few examples...

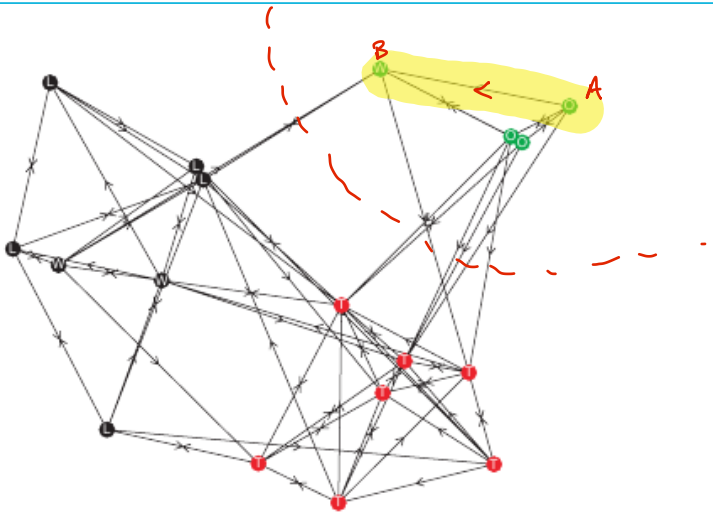


Figure: The Sampson Monks (1969)

A few examples...

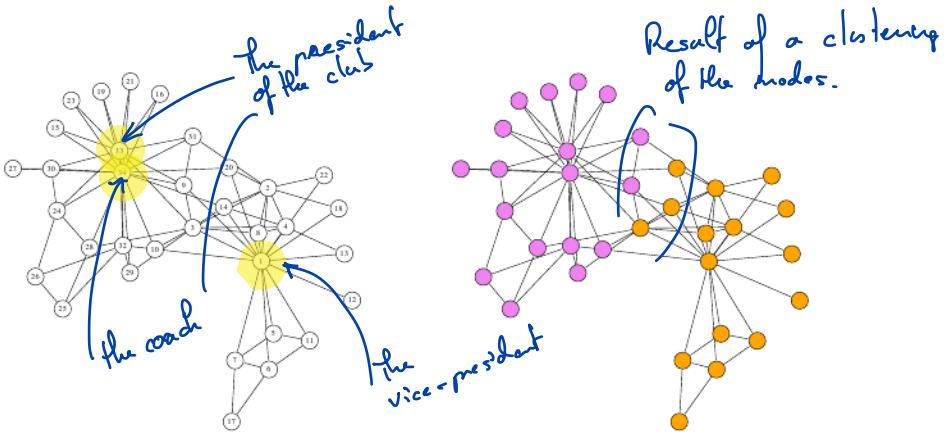


Figure: The Zachary *et al.* karate club (1977)

A few examples...

Rmk: this work illustrates the fact that networks can be reconstructed from various sources / documents.
⇒ Panama papers.

Rmk: in such a case the decision about the definition of the relationships is critical and is the decision of the analysists.

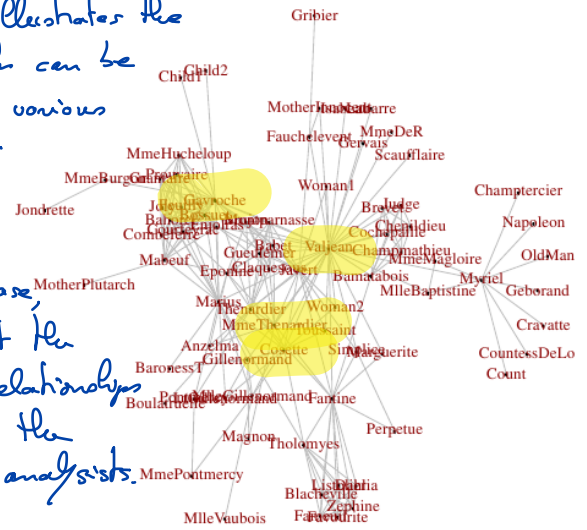


Figure: The network of Les Misérables (Knuth *et al.*, 1993)

A few examples...

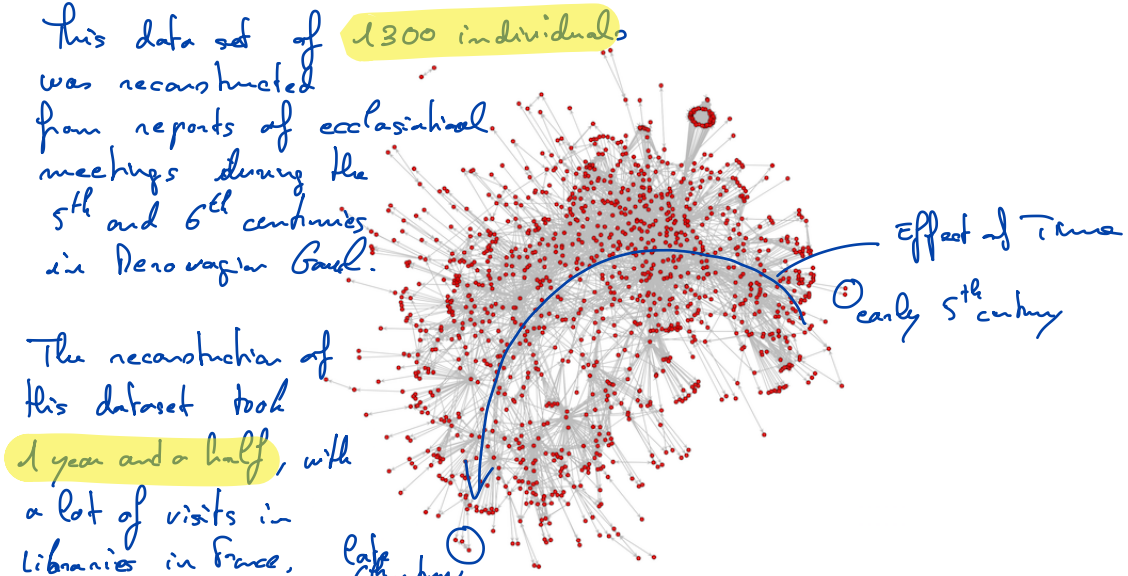


Figure: The Bishop Network (Bouveyron *et al.*, 2015)

A few examples...

From the national US census in high-schools.

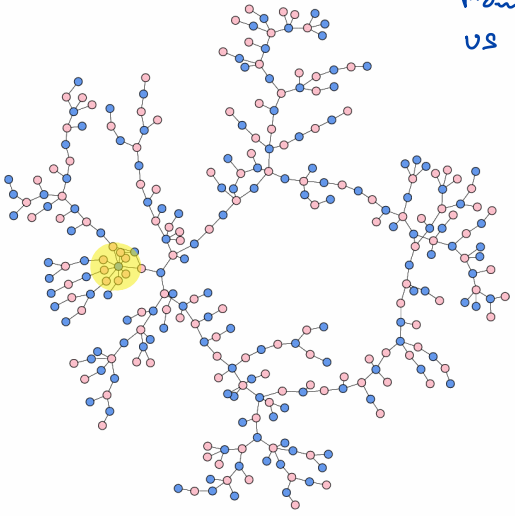


Figure: The dating network (Bearman *et al.*, 2004)

A few examples...

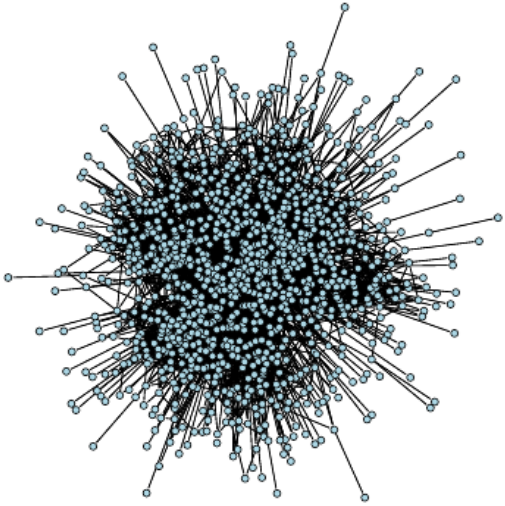


Figure: The Rovira University Email network (Guimera *et al.*, 2003)

For which applications?

Network analysis can be used for instance for:

- Medicine: public health, epidemiology, ...
- Biology: modeling of gene interactions, drugs, ...
- Social Sciences: study of individual and population actions
- Marketing: targeted ads on social networks.
- Fraud detection: Finance, Bank, insurance.
- Security & Defense: cyberbullism, counter-terrorism, ...

Where to find networks?

Networks can be found under different forms:

- a graph (simplest)
- an adjacency matrix (a socio-matrix) (simple as well)
- transactional data (less simple, but oh, most situations!)
- different sources of different types (high cost!)
  - ↳ 1 or several documents
  - ↳ of different types: text, text messages, tweets, phone call records, ...

Some examples:

- a social network like Twitter → graphs.
- emails of a company → transactional data
- Bishop network → different sources in different locations.
- Panama papers

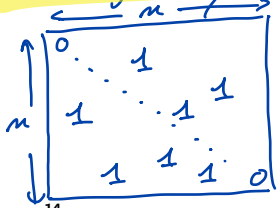
Outline

Characterizing networks

- **a graph**: a text file listing the interactions between the nodes:
  we list here all directed edges between the individuals.

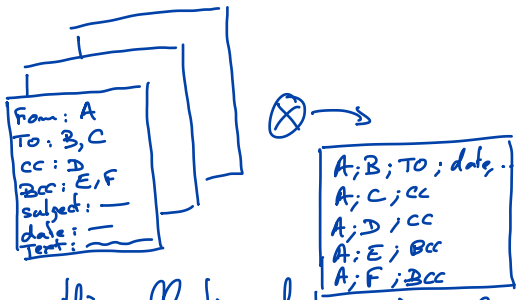
```
A;B
B;A
B;C
C;D
D;A
A;D
```

- **an adjacency matrix**:



- $A_{ij} = 1$ if i is connected with j (and 0 otherwise)
- $A_{ij} \neq A_{ji}$ if the network is directed.

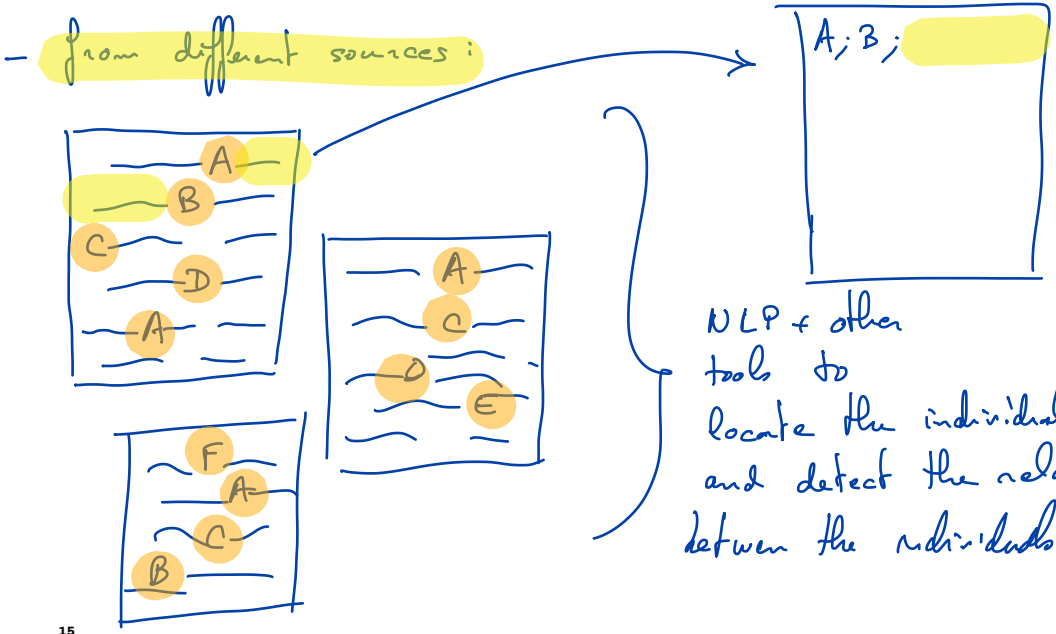
- **the transactional data**:



From this collection of transactional data, it is possible to extract a very rich network.

⇒ Reconstructing a network from this kind of data requires:
1) to define the type of relationship
2) write a code for reconstructing the graph file.

Characterizing networks



Characterizing networks

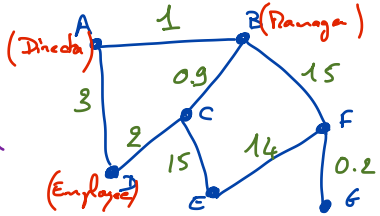
A network is composed of:

- nodes (individuals)
- edges (relationships)

} a graph

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

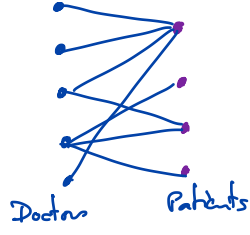
a network.



The different types of networks:

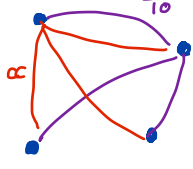
- directed and undirected networks

- bipartite networks



- dynamic or static networks.

- multiple networks (different type of edges)



Characterizing networks

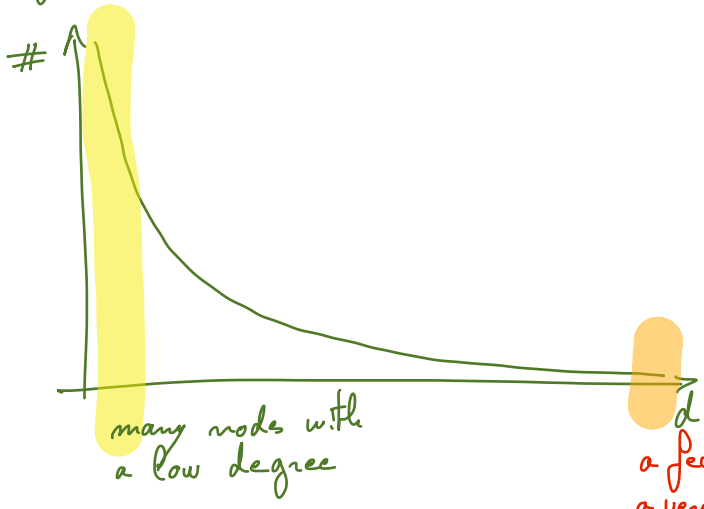
A first way to characterize a network is to compute general statistics for it:

- degree of a node i: it measures its importance, its centrality in the network:

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

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

Rmk: in most "real-world" 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:

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

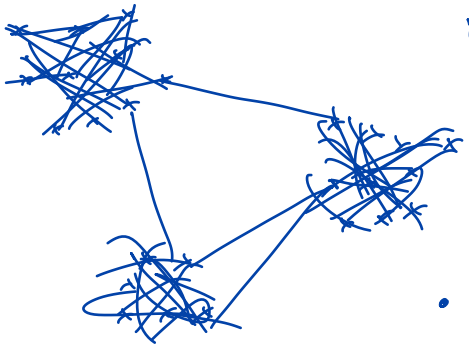
$\sum_{i=1}^{n} \sum_{j=1, j \neq i}^{n} A_{ij}$ ← the total edges in the network

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

$D(G) \in [0, 1]$

Rmk: it could be also interesting to compute the density for subparts of the network. (cf. the small world effect)

Characterizing networks



- The global density is quite low. (≈ 0.2)
- the local density in the communities is quite high (≈ 0.8)

How to manipulate networks?

To manipulate networks with R, we will use several libraries:

- igraph
- network
- sna.

Rmk: at the moment, Python is very lacking of libraries to efficiently manipulate networks

Statistical Learning with Complex Data

Pr. Charles BOUVEYRON

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

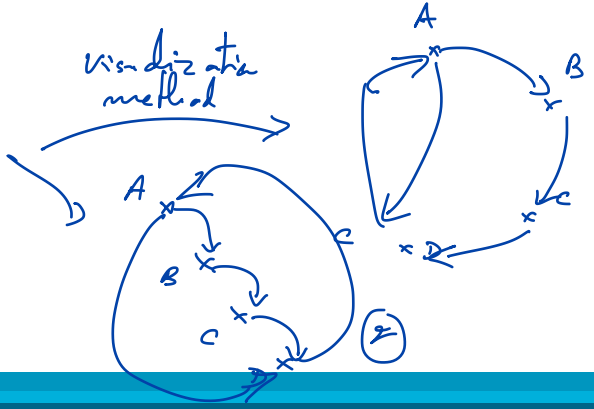
charles.bouveyron@univ-cotedazur.fr
@cbouveyron

Outline

The visualization of networks

First of all, it is important to understand that the visualization of a given network is not a trivial task. It is even a very difficult task if the network is dense.

A → B
B → C
C → D
D → A
A → D



From the previous example, it is obvious that positioning the nodes in a proper way (having a clear visualization of the relationships) is a difficult task.

For visualisation, we have different extensions of existing dimension reduction methods a dedicated statistical methods.

- MDS
- LSM

MDS for visualizing networks

Multidimensional Scaling (MDS) is a method used for the visualization of any kind of data (networks, quantitative, texts, images, ...) for which you are able to define a distance between the observations.

⇒ MDS has been quite popular for the dimension reduction of high-dimensional data.

MDS for visualizing networks

The goal of MDS is to find a low-dimensional representation of the data which is preserving the topology of the original data.

In practice, MDS looks for a positioning of the data points such that the distance between the points in the low-dimensional space are as close as possible than the distances in the original space.

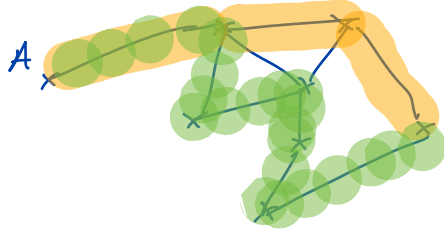
The translation of this problem in equations:

$$\min_{Z} \sum_{i=1}^{n} \sum_{\substack{j \neq i \\ j=1}}^{n} \left\| d(x_i, x_j) - \delta(z_i, z_j) \right\|^2$$

where d and δ are respectively the distances between the observations in the original and representation spaces, and $z_1, \dots, z_n$ are the latent representations of the data points $x_1, \dots, x_n$.

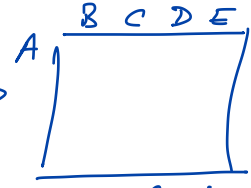
Applying this to a network requires to define:

- d: it could be the shortest-path distance on the graph.

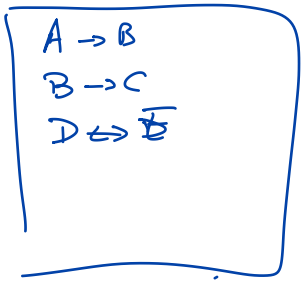


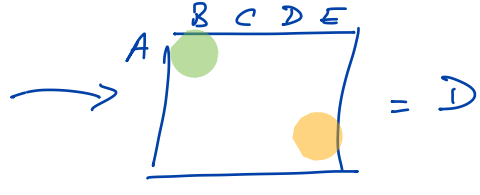
$d(A,B) = \min(3,6,5,4)$

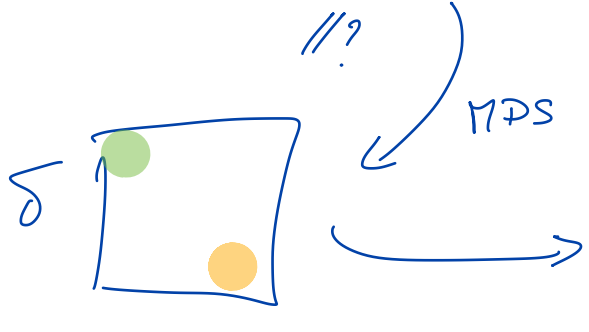
$= 3$



- δ: it could be simply the Euclidean distance in $\mathbb{R}^p$ ($p=2$?)


A → B
B → C
D ↔ E


B C D E
A
= D


MDS
A
B
C
D
E

Summary on MDS:

⊕ it is a generic method, working well for network but also other kinds of data.

⊕ the knowledge of the quality of the representation of the pairs is interesting in practice.

⊖ MDS could be limited for representing very complex networks.

⊖ MDS is not able to model a possible uncertainty on the observed edges.

Exercise: Use TDS to position the following nodes

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

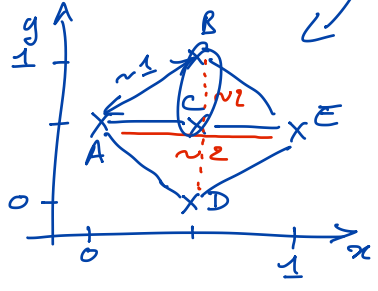
→

	A	B	C	D	E
A	0	1	1	1	2
B		0	2	2	1
C			0	2	1
D				0	1
E					0

= D

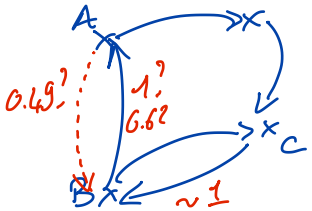
$\delta =$

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



The latent space model (LSM) (Hoff, Handcock and Raftery, 2001)

LSM is the first statistical method ever proposed to visualize and model a network. This method in particular takes into account the possible uncertainty on the observed edges.



The latent space model (LSM)

The goal of the latent space model is two-fold: we would like to find a latent representation of the data points such that:

i) points that are close together should have a high probability to connect

ii) points that a far away should have a low probability to connect.

The latent space model (LSM)

Once again, translating this in equations leads to:

Let suppose that X_{ij} is a random variable such that:

$$\begin{cases} X_{ij} = 1 \text{ if } i \sim j \\ X_{ij} = 0 \text{ if } i \not\sim j \end{cases}$$

In this case, the LSM model assumes that:

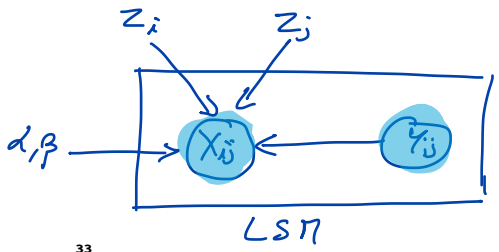
$$\text{logit}\left(P(X_{ij}=1|\theta)\right) = \log\left(\frac{P(X_{ij}=1|\theta)}{P(X_{ij}=0|\theta)}\right) = \alpha + \beta Y_{ij} - \|z_i - z_j\|^2$$

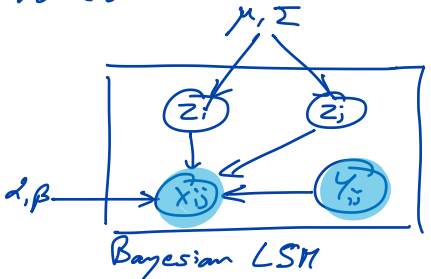
where α is a prior probability to connect, z_i is the latent position of the node i, Y_{ij} is a covariate information on the pair.

The latent space model (LSM)

In this model, the data are the pairs X_{ij}, $i = 1 \dots n$ and $j = 1 \dots n$, (the adjacency matrix) and the parameters of the model are $\alpha, \beta, z_1, \dots, z_n$.

Rmq: even though we have $n+2$ parameters to estimate we can use n^2 edge data to estimate them.





Inferring the LSM or the Bayesian LSM models requires to use either Maximum Likelihood or MCMC methods.

For the LSM model, the Log-Likelihood will have the following form:

$$\log\left(\mathcal{L}(X;\theta)\right) = \sum_{\substack{i,j \\ i \neq j}}^{n} \left[x_{ij}\left(\alpha + \beta y_{ij} - d_{ij}^2\right) - \log\left(1 + \exp\left(\alpha + \beta y_{ij} - d_{ij}^2\right)\right)\right]$$

where $d_{ij}^2 = \|z_i - z_j\|^2$

Unfortunately, as for logistic regression, there is no closed-form solution and we have to rely on a optimization algorithm to maximize this function.

To summarize:

⊕ LSM both model the uncertainty of the edges while providing a visualization of the network.

⊕ LSM offered a first basic statistical model as a basis for a lot of extensions.

⊖ LSM is just able to model communities and not stars.

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

The latent space model (LSM)

Adding covariates:

$$\text{logit}\left(P(X_{ij}|\Theta)\right) = \alpha + \beta Y_{ij} - d(z_i, z_j)$$

The covariate Y_{ij} can be used to provide extra information to the model on the pairs of nodes. For instance:

- Y_{ij} is the nb of years in common in a club/society between i and j.
- a type of relationship (categorial var) $\rightsquigarrow$ $Y_{ij} \in \{1, \dots, K\} \Rightarrow Y_{ij} = (0, 0, 1, 0, 0) \Rightarrow \beta$ is a vector ($\uparrow Y_{ij} = 3$)

Choice of the distance:

Another way to extend this model is to play with the definition of the distance within the latent space

- $d(z_i, z_j) = \|z_i - z_j\|_2$ or $\|z_i - z_j\|_2^2$
- $d(z_i, z_j) = \|z_i - z_j\|_1$ (Manhattan distance)

Modifying the model: A specific and interesting case is the situation of directed networks, in which there are the roles of sender and receiver. It is naturally interesting to model this. A way to do that:

$$\text{logit}\left(P(X_{ij}=1|\theta)\right) = \alpha + \beta Y_{ij} - d(z_i, z_j) + \underbrace{\delta_i + \gamma_j}_{\text{a sender-receiver effect}}$$

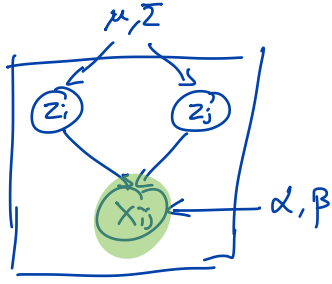
where $\begin{cases} \delta_i \sim \mathcal{N}(0, \sigma_s^2) \\ \gamma_j \sim \mathcal{N}(0, \sigma_r^2) \end{cases}$

← the prior for the propensity to send messages

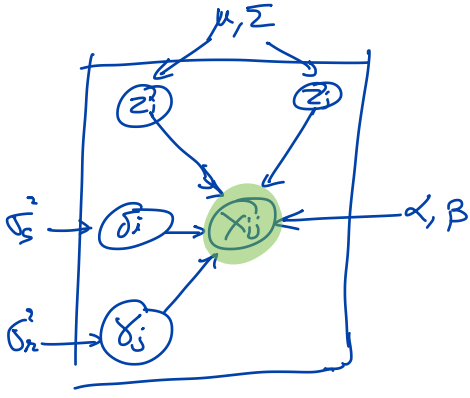
← the prior for receiving messages.

Rmk: this model is highly parametrized: it has $(3n+2)$ parameters to estimate.

Exercise: draw the graphical model for this LSM variant.



Bayesian LSM.



Outline

The clustering of networks

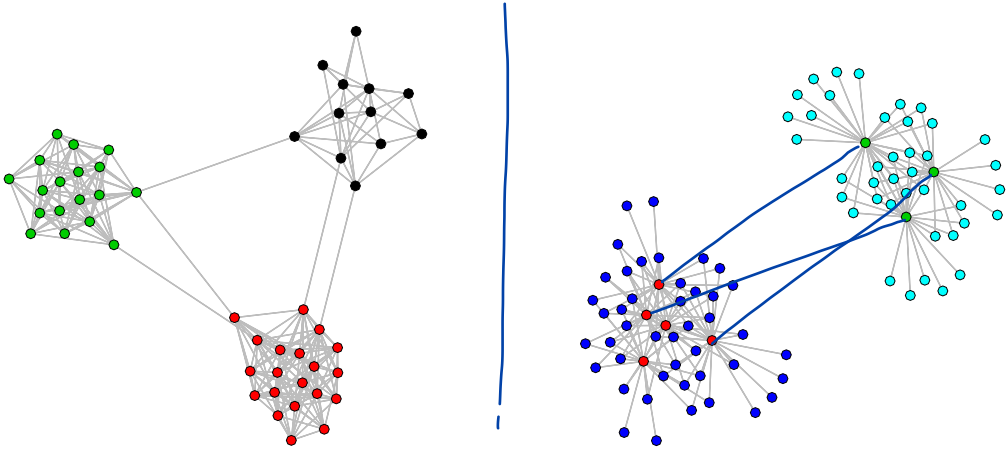
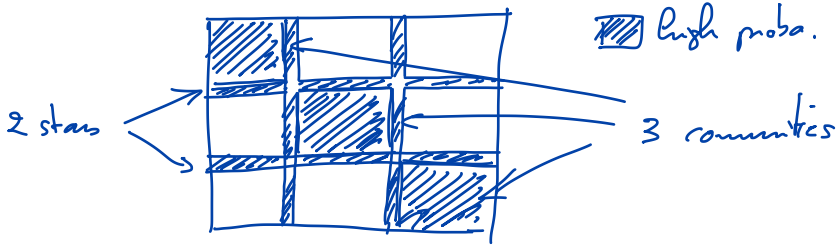


Figure: Clustering of communities vs. stars.

The clustering of networks

Difference between communities and stars:

- in communities, people have a higher probability connection within the community than with other communities
- stars are people that connect less within the group than outside the group.



The latent position cluster model (LPCM)

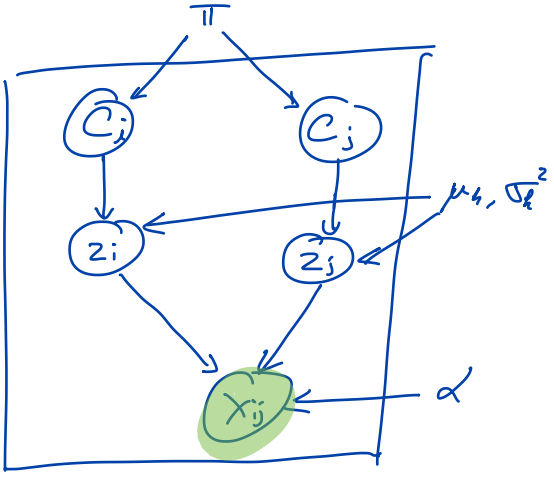
The LPCM extends LSM by adding a clustering structure:

- $\text{logit}\left(P(X_{ij}=1 \mid \theta)\right) = \alpha - d(z_i, z_j)$
- $C_i \sim \mathcal{M}(1; \pi)$ where π_k is the prior probability for cluster k, $k \in \{1, \dots, K\}$
- $Z_i \mid C_{ik} = 1 \sim \mathcal{N}(\mu_k, \sigma_k^2 I)$

$$\Rightarrow Z_i \sim \sum_{k=1}^{K} \pi_k \mathcal{N}(\mu_k, \sigma_k^2 I).$$

The latent position cluster model (LPCM)

The model:



clustering

visualization...

⇒ the inference of this Bayesian model has to be done using MCMC or advanced inference strategies (VBEM)

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

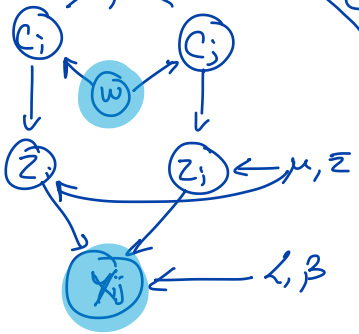
The latent position cluster model (LPCM)

Extension #2: mixture of experts LPCM

This model assumes that some covariates w may have an effect on the clustering

$$p(z_i) = \sum_{k=1}^{K} \pi_k(w_i)\, \mathcal{N}(z_i; \mu_k, \Sigma_k)$$

↑ the prior probability of the groups depends on the individual covariate.



Rmk: This model of course comes with some complications regarding inference

The latent position cluster model (LPCM)

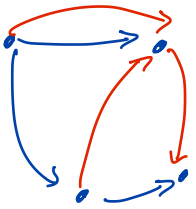
Extension #3: taking into account a dynamic

In order to model real-world networks where interactions may evolve along the time, it is interesting to model this. A way to do that is to assume that the cluster proportions π evolve:

$$\text{logit}(\pi_k(t)) = \alpha_k(t) \sim \mathcal{N}(\alpha_k(t-1), \sigma)$$

this modeling is known as the State Space Model (SSM).

Extension #4: dealing with multi-networks



LSM / LPCM to this situation

$$z_i \sim \sum \pi_k^c \mathcal{N}(\dots)$$

$$\text{logit}(P(X_{ij}^c | \theta)) = \alpha_c - \beta_c |z_i - z_j|$$

The Stochastic Block Model (SBM):

SBM is, at the moment, the most popular and efficient clustering model for networks. SBM has two main interests:

1) it is able to recover both communities and stars at the same time
2) the output of the model can be seen as a network summary (meta-network).

The stochastic block model (SBM)

The SBM model assumes:

- $C_i \sim \mathcal{M}(1; \rho)$
- $X_{ij} \mid C_{ik}=1, C_{jl}=1 \sim B(\pi_{kl})$

where $C_i = (c_{i1}, \dots, c_{iQ})$ with $c_{ik} \in \{0,1\}$ and $\rho = (\rho_1, \dots, \rho_Q)$ are the prior probabilities of the groups.

$C_{jl}=1$ means that j belongs to cluster l.

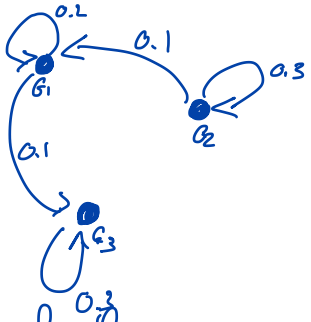
$C_i = (0,0,1,0)$ $\Rightarrow$ i belongs to cluster 3.

π_{kl} is the probability that people from cluster k connect with people from cluster l.

Two remarks:

(i) The matrix Π can be seen as a network between the groups (it adjacency matrix, weighted)

$$\Pi = \begin{pmatrix} 0.2 & 0 & 0.1 \\ 0.1 & 0.3 & 0 \\ 0 & 0 & 0.3 \end{pmatrix}$$

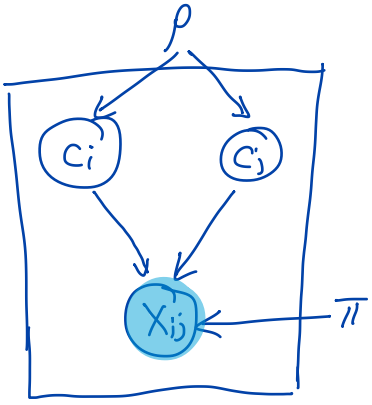


(ii) Π can also indicate if the groups are communities or stars

$$\Pi = \begin{pmatrix} 0.01 & 0. & 0 \\ 0.2 & 0.4 & 0 \\ 0.2 & 0 & 0.5 \end{pmatrix}$$

The stochastic block model (SBM)

The graphical model:



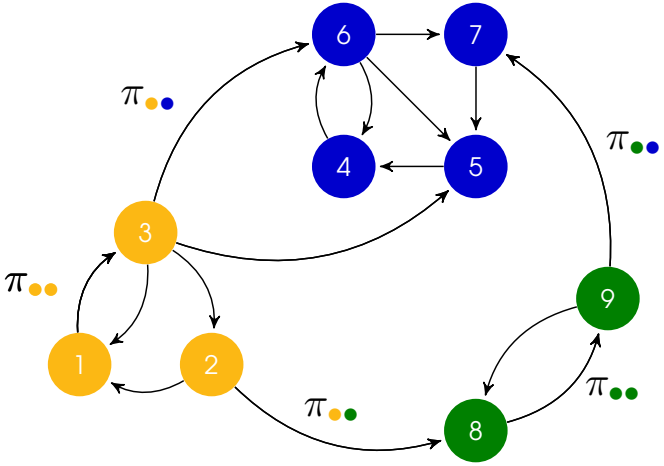
The inference of this model will have to estimate the model parameters, ρ and π, and the latent variables C.

→ Variational EM algorithm.

→ MCMC with a Bayesian version of the model.

The stochastic block model (SBM)

A simple example:



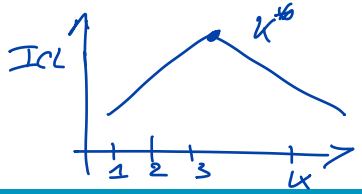
The stochastic block model (SBM)

Choosing the number of clusters:

As for other statistical models, we can rely here on model selection tools:

$$BIC(\mathcal{M}) = \log\left(\mathcal{L}(\hat{\theta})\right) - \frac{\nu(\mathcal{M})}{2}\log(n)$$

$$ICL(\mathcal{M}) = BIC - \sum_i \sum_k c_{ik} \log(c_{ik}).$$



The mixed membership SBM (MMSBM)

The MMSBM extends the SBM as follows; in order to allow people to have different clusters depending on their roles in the network.

- $C_{i\to j} \sim \mathcal{M}(1; \rho_i)$ and $C_{i\leftarrow j} \sim \mathcal{M}(1; \rho_j)$ and $\rho_i \sim \mathrm{Dir}(\alpha)$
- $X_{ij} \mid C_{i\to jk} C_{i\leftarrow jl} \sim \mathcal{B}(\pi_{kl})$