Statistical network clustering: some recent advances and applications to digital humanities

Charles Bouveyron

Laboratoire MAP5, UMR CNRS 8145
Université Paris Descartes

charles.bouveyron@parisdescartes.fr – @cbouveyron

Disclaimer

"Essentially, all models are wrong but some are useful"

George E.P. Box

Outline

Introduction

The analysis of networks:

- is a recent but increasingly important field in statistical learning,
- with applications in domains ranging from biology to history:
  - biology: analysis of gene regulation processes,
  - social sciences: analysis of political blogs,
  - history: visualization of medieval social networks.

Two main problems are currently well addressed:

- visualization of the networks,
- clustering of the network nodes.

Introduction

The analysis of networks:

- is a recent but increasingly important field in statistical learning,
- with applications in domains ranging from biology to history:
  - biology: analysis of gene regulation processes,
  - social sciences: analysis of political blogs,
  - history: visualization of medieval social networks.

Two main problems are currently well addressed:

- visualization of the networks,
- clustering of the network nodes.

Network comparison:

- is a still emerging problem is statistical learning,
- which is mainly addressed using graph structure comparison,
- but limited to binary networks.

Introduction

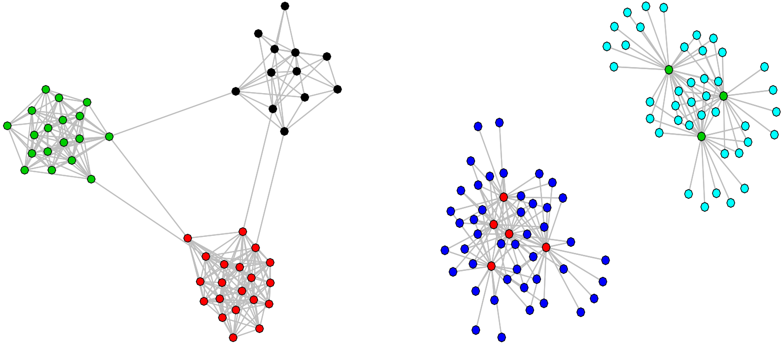


Figure: Clustering of network nodes: communities (left) *vs.* structures with hubs (right).

Introduction

Key works in probabilistic models:

- stochastic block model (SBM) by Nowicki and Snijders (2001),
- latent space model by Hoff, Handcock and Raftery (2002),
- latent cluster model by Handcock, Raftery and Tantrum (2007),
- mixed membership SBM (MMSBM) by Airoldi et al. (2008),
- mixture of experts for LCM by Gormley and Murphy (2010),
- MMSBM for dynamic networks by Xing et al. (2010),
- overlapping SBM (OSBM) by Latouche et al. (2011).

A good overview is given in:

- M. Salter-Townshend, A. White, I. Gollini and T. B. Murphy, "Review of Statistical Network Analysis: Models, Algorithms, and Software", Statistical Analysis and Data Mining, Vol. 5(4), pp. 243–264, 2012.

Introduction: a historical problem

Our colleagues from the LAMOP team were interested in answering the following question:

Was the Church organized in the same way within the different kingdoms in Merovingian Gaul?

Introduction: a historical problem

Our colleagues from the LAMOP team were interested in answering the following question:

Was the Church organized in the same way
within the different kingdoms in Merovingian Gaul?

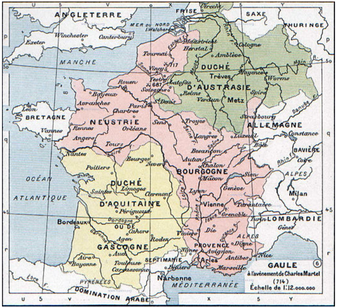
To this end, they have build a relational database:

- from written acts of ecclesiastical councils that took place in Gaul during the 6th century (480-614),
- those acts report who attended (bishops, kings, dukes, priests, monks, ...) and what questions (regarding Church, faith, ...) were discussed,
- they also allowed to characterize the type of relationship between the individuals,
- it took 18 months to build the database.

Introduction: a historical problem

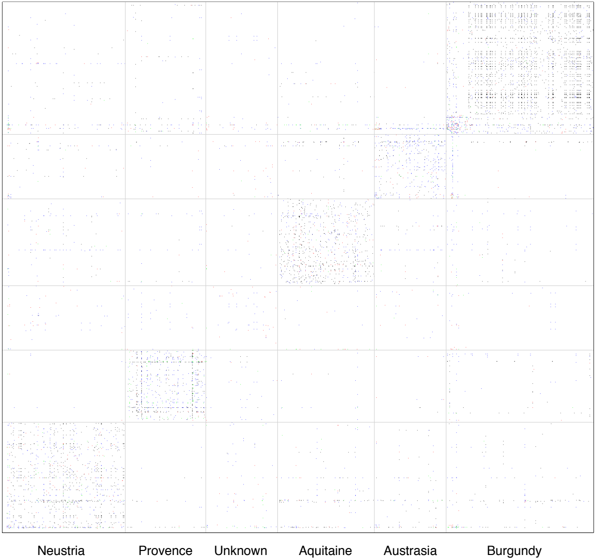
The database contains:

- 1331 individuals (mostly clergymen) who participated to ecclesiastical councils in Gaul between 480 and 614,
- 4 types of relationships between individuals have been identified (positive, negative, variable or neutral),
- each individual belongs to one of the 5 regions of Gaul:
  - 3 kingdoms: Austrasia, Burgundy and Neustria,
  - 2 provinces: Aquitaine and Provence.



- additional information is also available: *social positions*, family relationships, birth and death dates, hold offices, councils dates, ...

Introduction: a historical problem



[9] Figure: Adjacency matrix of the ecclesiastical network (sorted by regions).

Outline

The stochastic block model (SBM)

The SBM (Nowicki and Snijders, 2001) model assumes that the network (represented by its adjacency matrix X) is generated as follows:

- each node i is associated with an (unobserved) group among K according to:

  $$Z_i \sim \mathcal{M}(\alpha),$$

  where $\alpha \in [0,1]^K$ and $\sum_{k=1}^{K} \alpha_k = 1$,

The stochastic block model (SBM)

The SBM (Nowicki and Snijders, 2001) model assumes that the network (represented by its adjacency matrix X) is generated as follows:

- each node i is associated with an (unobserved) group among K according to:

  $$Z_i \sim \mathcal{M}(\alpha),$$

  where $\alpha \in [0,1]^K$ and $\sum_{k=1}^{K} \alpha_k = 1$,

- then, each edge X_{ij} is drawn according to:

  $$X_{ij} | Z_{ik} Z_{jl} = 1 \sim \mathcal{B}(\pi_{kl}),$$

  where $\pi_{kl} \in [0,1]$.

The stochastic block model (SBM)

The SBM (Nowicki and Snijders, 2001) model assumes that the network (represented by its adjacency matrix X) is generated as follows:

- each node i is associated with an (unobserved) group among K according to:

$$Z_i \sim \mathcal{M}(\alpha),$$

  where $\alpha \in [0,1]^K$ and $\sum_{k=1}^{K} \alpha_k = 1$,

- then, each edge X_{ij} is drawn according to:

$$X_{ij} | Z_{ik} Z_{jl} = 1 \sim \mathcal{B}(\pi_{kl}),$$

  where $\pi_{kl} \in [0,1]$.

- this model is therefore a mixture model:

$$X_{ij} \sim \sum_{k=1}^{K} \sum_{\ell=1}^{K} \alpha_k \alpha_\ell \mathcal{B}(\pi_{kl}).$$

The stochastic block model (SBM)

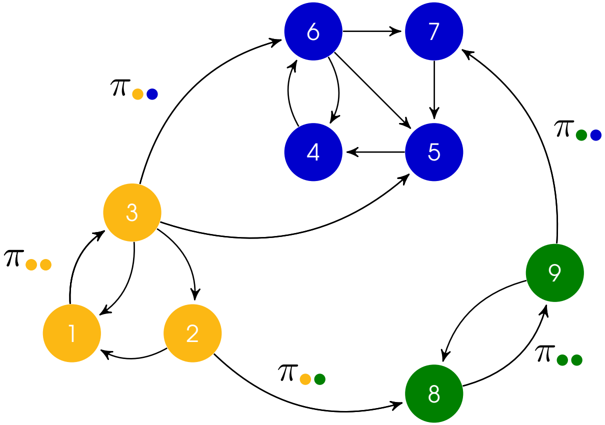


Table: A SBM network.

The stochastic block model (SBM)

Inference of the SBM model (maximum likelihood):

- log-likelihood:

$$\log p(X|\alpha,\Pi) = \log \left\{ \sum_{Z} p(X, Z|\alpha,\Pi) \right\},$$

$\hookrightarrow$ K^N terms!

The stochastic block model (SBM)

Inference of the SBM model (maximum likelihood):

- log-likelihood:

$$\log p(X|\alpha, \Pi) = \log \left\{ \sum_Z p(X, Z|\alpha, \Pi) \right\},$$

$\hookrightarrow$ K^N terms!

- Expectation Maximization (EM) algorithm requires the knowledge of $p(Z|X, \alpha, \Pi)$,
- Problem: $p(Z|X, \alpha, \Pi)$ is not tractable (no conditional independence)!

The stochastic block model (SBM)

Inference of the SBM model (maximum likelihood):

- log-likelihood:

$$\log p(X|\alpha,\Pi) = \log\left\{\sum_Z p(X,Z|\alpha,\Pi)\right\},$$

$\hookrightarrow$ K^N terms!

- Expectation Maximization (EM) algorithm requires the knowledge of $p(Z|X,\alpha,\Pi)$,
- Problem: $p(Z|X,\alpha,\Pi)$ is not tractable (no conditional independence)!

Solutions:

- Variational EM (Daudin et al., 2008) + ICL (Biernacki et al., 2003),
- Variational Bayes EM + $ILvb$ criterion (Latouche et al., 2012).

Outline

The random subgraph model (RSM)

Before the maths, an example of an RSM network:

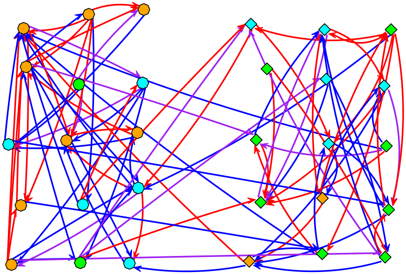


Figure: Example of an RSM network.

We observe:

- the partition of the network into $S = 2$ subgraphs (node form),
- the presence A_{ij} of directed edges between the N nodes,
- the type $X_{ij} \in \{1, ..., C\}$ of the edges ($C = 3$, edge color).

The random subgraph model (RSM)

Before the maths, an example of an RSM network:

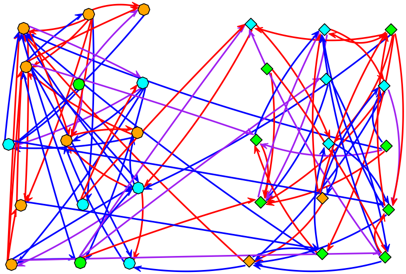


Figure: Example of an RSM network.

We observe:

- the partition of the network into $S = 2$ subgraphs (node form),
- the presence A_{ij} of directed edges between the N nodes,
- the type $X_{ij} \in \{1, ..., C\}$ of the edges ($C = 3$, edge color).

We search:

- a partition of the node into $K = 3$ groups (node color),
- which overlap with the partition into subgraphs.

The random subgraph model (RSM)

The network (represented by its adjacency matrix X) is assumed to be generated as follows:

- the presence of an edge between nodes i and j is such that:

$$A_{ij} \sim \mathcal{B}(\gamma_{s_i s_j})$$

  where $s_i \in \{1, ..., S\}$ indicates the (observed) subgraph of node i,

The random subgraph model (RSM)

The network (represented by its adjacency matrix X) is assumed to be generated as follows:

- the presence of an edge between nodes i and j is such that:

$$A_{ij} \sim \mathcal{B}(\gamma_{s_i s_j})$$

  where $s_i \in \{1, ..., S\}$ indicates the (observed) subgraph of node i,

- each node i is as well associated with an (unobserved) group among K according to:

$$Z_i \sim \mathcal{M}(\alpha_{s_i})$$

  where $\alpha_s \in [0, 1]^K$ and $\sum_{k=1}^{K} \alpha_{sk} = 1$,

The random subgraph model (RSM)

The network (represented by its adjacency matrix X) is assumed to be generated as follows:

- the presence of an edge between nodes i and j is such that:

$$A_{ij} \sim \mathcal{B}(\gamma_{s_i s_j})$$

  where $s_i \in \{1, ..., S\}$ indicates the (observed) subgraph of node i,

- each node i is as well associated with an (unobserved) group among K according to:

$$Z_i \sim \mathcal{M}(\alpha_{s_i})$$

  where $\alpha_s \in [0,1]^K$ and $\sum_{k=1}^K \alpha_{sk} = 1$,

- each edge X_{ij} can be finally of C different (observed) types and such that:

$$X_{ij} | A_{ij} Z_{ik} Z_{jl} = 1 \sim \mathcal{M}(\Pi_{kl})$$

  where $\Pi_{kl} \in [0,1]^C$ and $\sum_{c=1}^C \Pi_{klc} = 1$.

The random subgraph model (RSM)

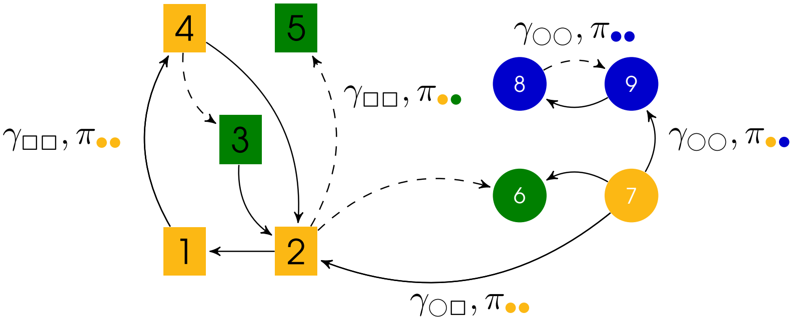


Table: A RSM network.

The random subgraph model (RSM)

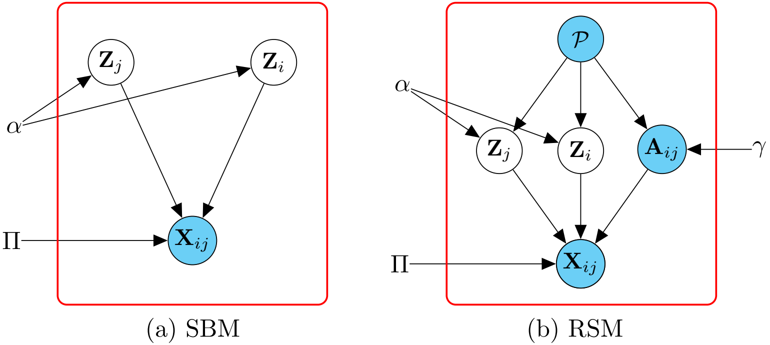


Figure: SBM model *vs.* RSM model.

The random subgraph model (RSM)

Remark 1:

- the RSM model separates the roles of the known partition and the latent clusters,
- this was motivated by historical assumptions on the creation of relationships during the 6th century,
- indeed, the possibilities of connection were preponderant over the type of connection and mainly dependent on the geography.

The random subgraph model (RSM)

Remark 1:

- the RSM model separates the roles of the known partition and the latent clusters,
- this was motivated by historical assumptions on the creation of relationships during the 6th century,
- indeed, the possibilities of connection were preponderant over the type of connection and mainly dependent on the geography.

Remark 2:

- an alternative approach would consist in allowing X_{ij} to directly depend on both the latent clusters and the partition,
- however, this would dramatically increase the number of model parameters ($K^2S^2(C+1)+SK$ instead of S^2+K^2C+SK),
- if $S=6$, $K=6$ and $C=4$, then the alternative approach has 6 516 parameters while RSM has only 216.

The random subgraph model (RSM)

We consider a Bayesian framework:

- the previous model is fully defined by its joint distribution:

$$p(X, A, Z|\alpha, \gamma, \Pi) = p(X|A, Z, \Pi)p(A|\gamma)p(Z|\alpha),$$

- which we complete with conjuguate prior distributions for model parameters:
  - the prior distribution for α is:

$$p(\gamma_{rs}) = Beta(a_{rs}, b_{rs}),$$

  - the prior distribution for γ is:

$$p(\alpha_s) = Dir(\chi_s),$$

  - the prior distribution for Π is:

$$p(\Pi_{kl}) = Dir(\Xi_{kl}).$$

The random subgraph model (RSM)

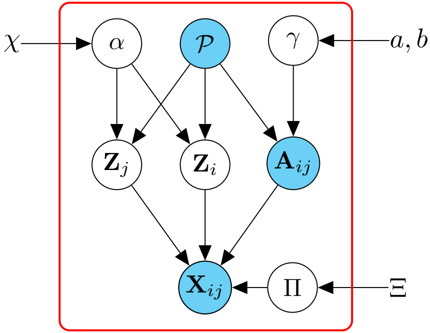


Figure: A graphical representation of the RSM model.

Model inference through a VBEM algorithm

Due to the Bayesian framework introduces above:

- we aim at estimating the posterior distribution $p(Z, \alpha, \gamma, \Pi | X, A)$, which in turn will allow us to compute MAP estimates of Z and (α, γ, Π),
- as expected, this distribution is not tractable and approximate inference procedures are required,
- the use of MCMC methods is obviously an option but MCMC methods have a poor scaling with sample sizes.

Model inference through a VBEM algorithm

Due to the Bayesian framework introduces above:

- we aim at estimating the posterior distribution $p(Z, \alpha, \gamma, \Pi | X, A)$, which in turn will allow us to compute MAP estimates of Z and (α, γ, Π),
- as expected, this distribution is not tractable and approximate inference procedures are required,
- the use of MCMC methods is obviously an option but MCMC methods have a poor scaling with sample sizes.

We chose to use variational approaches:

- because they allow to deal with large networks ($N > 1000$),
- recent theoretical results (Celisse et al., 2012; Mariadassou and Matias, 2013) gave new insights about convergence properties of variational approaches in this context.

The VBEM algorithm

We aim at estimating the posterior distribution $p(Z,\theta|X)$:

- we use the decomposition of the marginal log-likelihood:

$$\log(p(X)) = \mathcal{L}(q(Z,\theta)) + KL(q(Z,\theta)||p(Z,\theta|X)),$$

  where:
  - $\mathcal{L}(q(Z,\theta)) = \sum_Z \int_\theta q(Z,\theta)\log(p(X,Z,\theta)/q(Z,\theta))d\theta$ is a lower bound of the log-likelihood,
  - $KL(q(Z,\theta)||p(Z,\theta|X)) = -\sum_Z \int_\theta q(Z,\theta)\log(p(Z,\theta|X)/q(Z,\theta))d\theta$ is the KL divergence between $q(Z,\theta)$ and $p(Z,\theta|X)$.

- we also assume that q factorizes over Z and θ:

$$q(Z,\theta) = \prod_i q_i(Z_i)q_\theta(\theta).$$

The VBEM algorithm

We aim at estimating the posterior distribution $p(Z, \theta | X)$:

- we use the decomposition of the marginal log-likelihood:

$$\log(p(X)) = \mathcal{L}(q(Z, \theta)) + KL(q(Z, \theta) || p(Z, \theta | X)),$$

  where:
  - $\mathcal{L}(q(Z, \theta)) = \sum_Z \int_\theta q(Z, \theta) \log(p(X, Z, \theta)/q(Z, \theta)) d\theta$ is a lower bound of the log-likelihood,
  - $KL(q(Z, \theta) || p(Z, \theta | X)) = -\sum_Z \int_\theta q(Z, \theta) \log(p(Z, \theta | X)/q(Z, \theta)) d\theta$ is the KL divergence between $q(Z, \theta)$ and $p(Z, \theta | X)$.

- we also assume that q factorizes over Z and θ:

$$q(Z, \theta) = \prod_i q_i(Z_i) q_\theta(\theta).$$

The VBEM algorithm:

- VB-E step: $q_\theta(\theta)$ is fixed and $\mathcal{L}$ is maximized over the q_i
  $\Rightarrow \log q_j^*(Z_j) = E_{i \neq j, \theta}[\log p(X, Z, \theta)] + c$
- VB-M step: all $q_i(Z_i)$ are now fixed and $\mathcal{L}$ is maximized over q_θ
  $\Rightarrow \log q_\theta^*(\theta) = E_Z[\log p(X, Z, \theta)] + c$

Initialization and choice of K

Initialization of the VBEM algorithm:

- the VBEM is known to be sensitive to its initialization,
- we propose a strategy based on several k-means algorithms with a specific distance:

$$d(i,j) = \sum_{h=1}^{N} \delta(X_{ih} \neq X_{jh}) A_{ih} A_{jh} + \sum_{h=1}^{N} \delta(X_{hi} \neq X_{hj}) A_{hi} A_{hj}.$$

Initialization and choice of K

Initialization of the VBEM algorithm:

- the VBEM is known to be sensitive to its initialization,
- we propose a strategy based on several k-means algorithms with a specific distance:

$$d(i,j) = \sum_{h=1}^{N} \delta(X_{ih} \neq X_{jh}) A_{ih} A_{jh} + \sum_{h=1}^{N} \delta(X_{hi} \neq X_{hj}) A_{hi} A_{hj}.$$

Choice of the number K of groups:

- once the VBEM algorithm has converged, the lower bound $\mathcal{L}(q)$ is a good approximation of the integrated log-likelihood $\log p(X, A)$,
- we thus can use $\mathcal{L}(q)$ as a model selection criterion for choosing K,
- if computed right after the M step,

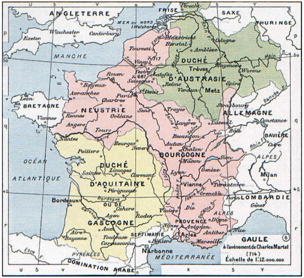
$$\mathcal{L}(q) = \sum_{r,s}^{S} \log\left(\frac{B(a_{rs}, b_{rs})}{B(a_{rs}^0, b_{rs}^0)}\right) + \sum_{s=1}^{S} \log\left(\frac{C(\boldsymbol{\chi}_s)}{C(\boldsymbol{\chi}_s^0)}\right) + \sum_{k,l}^{K} \log\left(\frac{C(\boldsymbol{\Xi}_{kl})}{C(\boldsymbol{\Xi}_{kl}^0)}\right) - \sum_{i=1}^{N} \sum_{k=1}^{K} \tau_{ik} \log(\tau_{ik}).$$

Outline

The ecclesiastical network

The data:

- 1331 individuals (mostly clergymen) who participated to ecclesiastical councils in Gaul between 480 and 614,
- 4 types of relationships between individuals have been identified (positive, negative, variable or neutral),
- each individual belongs to one of the 5 regions (3 kingdoms et 2 provinces).



Our modeling allows a multi-level analysis:

- Z allows to characterize the found clusters through social positions of the individuals,
- parameter Π describes the relations between the found clusters,
- parameter γ describes the connections between the subgraphs,
- parameter α describes the cluster repartition in the subgraphs.

RSM results: the latent clusters

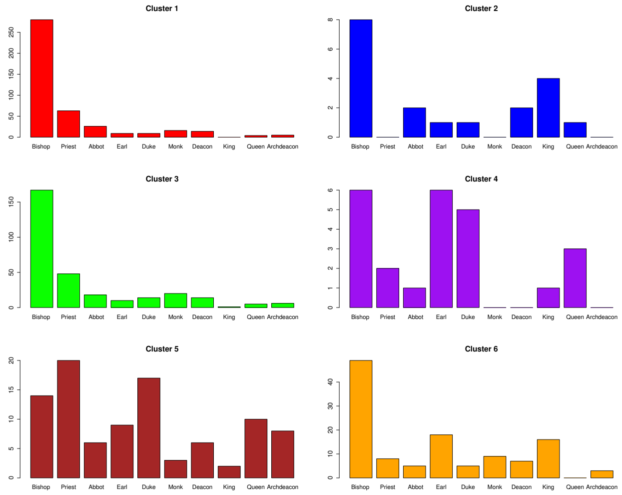


Figure: Characterization of the $K = 6$ clusters found by RSM.

RSM results: the latent clusters

The latent clusters from the historical point of view:

- clusters 1 and 3 correspond to local, provincial of diocesan councils, mostly interested in local issues (ex: council of Arles, 554),
- clusters 2 and 6 correspond to councils dedicated to political questions, usually convened by a king (ex: Orleans, 511),
- clusters 4 and 5 correspond to aristocratic assemblies, where queens and duke and earls are present (ex: Orleans, 529).

RSM results: the relationships between clusters

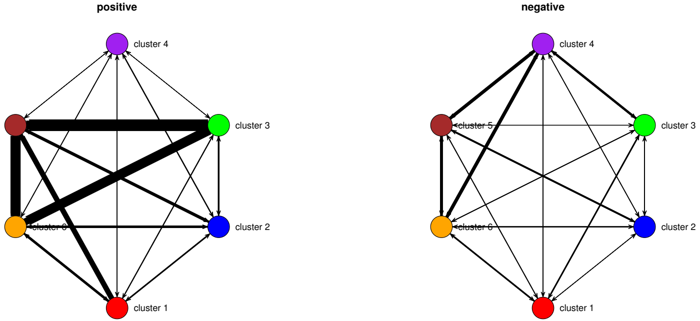


Figure: Characterization of the relationships between clusters (parameter Π).

RSM results: the relationships between clusters

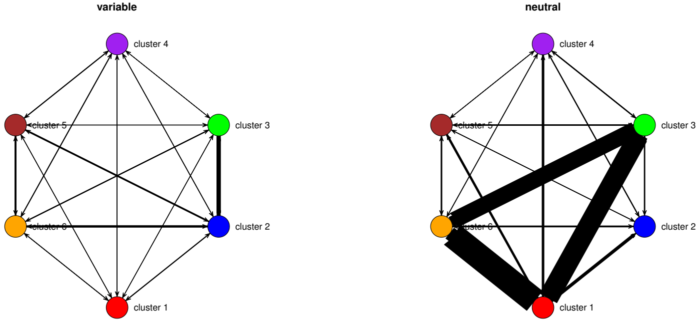


Figure: Characterization of the relationships between clusters (parameter Π).

RSM results: the relationships between clusters

The clusters relationships from the historical point of view:

- positive relations between clusters 3, 5 and 6 mainly corresponds to personal friendships between bishops (source effect),
- negative and variable relations betweens clusters 4, 5 and 6 report the conflicts in the hierarchy of the power,
- neutral relations between clusters 1, 3 and 6 were expected because they deal with different issues (local / political).

RSM results: the relationships between regions

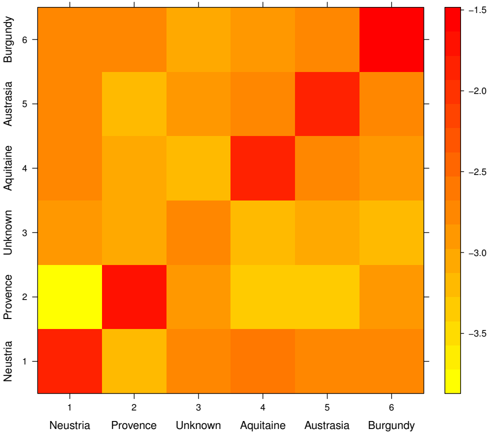


Figure: Characterization of the relationships between the regions (parameter γ in log scale).

RSM results: comparison of the regions

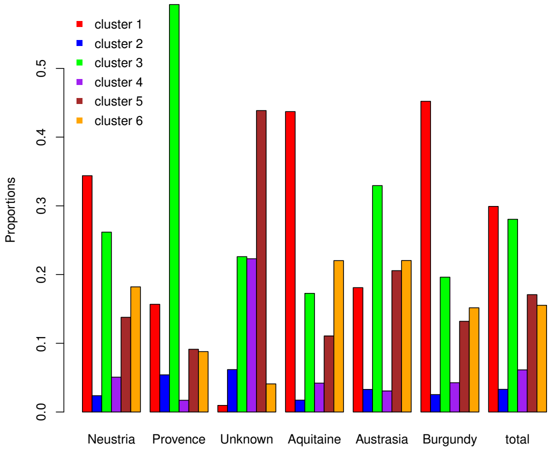


Figure: Characterization of regions through cluster repartition (parameter α).

RSM results: comparison of the regions

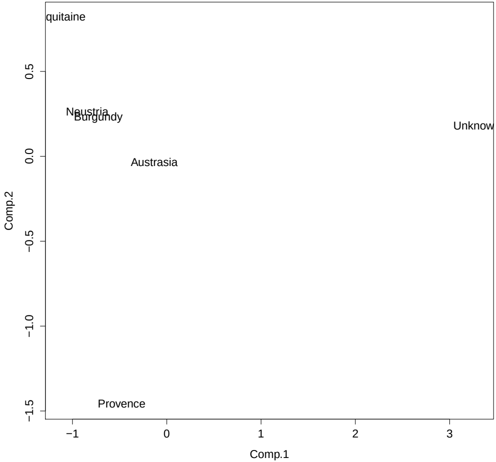


Figure: PCA for compositional data on the parameter α.

Outline

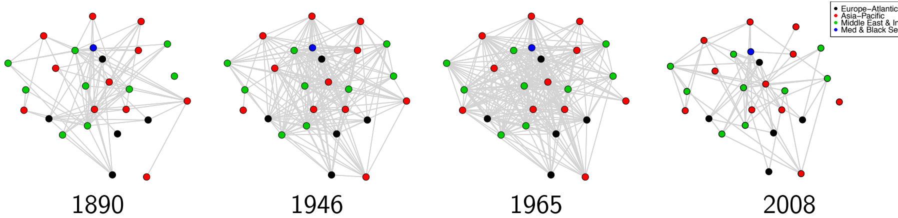
Dynamic networks: a problem in geography

Clustering of dynamic networks is an increasing problem, since most of the observed networks are in fact not static.

Dynamic networks: a problem in geography

Clustering of dynamic networks is an increasing problem, since most of the observed networks are in fact not static.

As an example, we will analyze a maritime flow network from 1870 to 2008:



Dynamic networks: a problem in geography

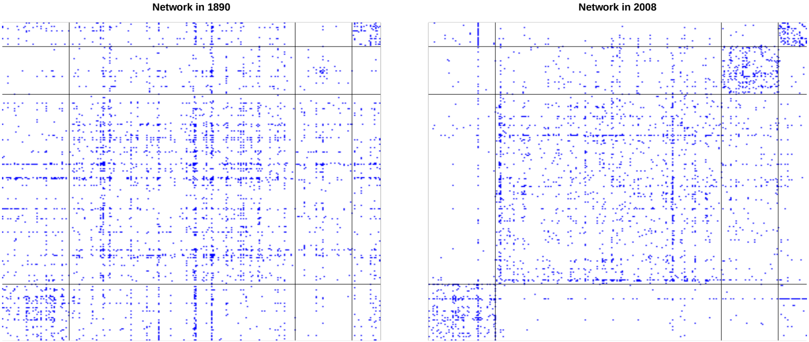


Figure: Adjacency matrix of the maritime flow network organized by subgraph in 1890 and 2008.

Only a few works in the literature

To date, only a few models have been proposed to deal with this kind of networks:

- dynamic MMSBM by Xing *et al.*,
- dynamic SBM by Yang *et al.*,
- another dynamic SBM by Xu *et al.*,
- dynamic LPCM by Sarkar *et al.*,
- and a few others...

Only a few works in the literature

To date, only a few models have been proposed to deal with this kind of networks:

- dynamic MMSBM by Xing *et al.*,
- dynamic SBM by Yang *et al.*,
- another dynamic SBM by Xu *et al.*,
- dynamic LPCM by Sarkar *et al.*,
- and a few others...

Here, we extend the RSM model (Jernite *et al.*, 2012) to be able to deal with dynamic networks with categorical edges and a known partition into subgraphs.

The dRSM model: the model at time t

At time t, the network (represented by its adjacency matrix $X^{(t)}$) is assumed to be generated as follows:

- each node i is associated with an (unobserved) group among K according to:

$$Z_i^{(t)} \sim \mathcal{M}(\alpha_{s_i}^{(t)})$$

  where $\alpha_s^{(t)} \in [0,1]^K$ and $\sum_{k=1}^K \alpha_{sk}^{(t)} = 1$,

- each edge $X_{ij}^{(t)}$ can have $C+1$ different (observed) types (0 denotes the absence of an edge) and such that:

$$X_{ij}^{(t)} | Z_{ik}^{(t)} Z_{jl}^{(t)} = 1 \sim \mathcal{M}(\Pi_{kl})$$

  where $\Pi_{kl} \in [0,1]^{C+1}$ and $\sum_{c=0}^C \Pi_{klc} = 1$.

The dRSM model: modeling the evolution

We rely on a state space model to take into account the dynamic of the network:

- we introduce the latent variable $\gamma_s^{(t)}$ to link the group proportions over the time:

$$\alpha_{sk}^{(t)} = \frac{\exp(\gamma_{sk}^{(t)})}{C(\gamma_s^{(t)})},$$

  where $\gamma_{sK}^{(t)} = 0$ and $C(\gamma_s^{(t)}) = \sum_{\ell=1}^{K} \exp(\gamma_{s\ell}^{(t)})$,
- $\gamma_{s\backslash K}^{(t)}$ is further assumed to be distributed according to a normal distribution with mean $B\nu^{(t)}$ and covariance matrix Σ,

$$\gamma_s^{(t)} \sim \mathcal{N}(B\nu^{(t)}, \Sigma).$$

The dRSM model: modeling the evolution

The reminder of the modeling involves a classical state space model:

- ν^t depends on ν^{t-1} such that:

$$\nu^{(t)} = A\nu^{(t-1)} + \omega^{(t)},$$

  where:

  - $\omega^{(t)} \sim \mathcal{N}(0, \Phi)$,
  - $\nu^1 = \mu_0 + u$,
  - $u \sim \mathcal{N}(0, v^0)$.

The dRSM model: modeling the evolution

The reminder of the modeling involves a classical state space model:

- ν^t depends on ν^{t-1} such that:

$$\nu^{(t)} = A\nu^{(t-1)} + \omega^{(t)},$$

  where:

  - $\omega^{(t)} \sim \mathcal{N}(0, \Phi)$,
  - $\nu^1 = \mu_0 + u$,
  - $u \sim \mathcal{N}(0, v^0)$.

To avoid model identifiability issues, we fixe A, B and v^0 to be equal to the identity matrix I_{K-1} and all components of μ_0 to zero in the numerical experiments.

The dRSM model: modeling the evolution

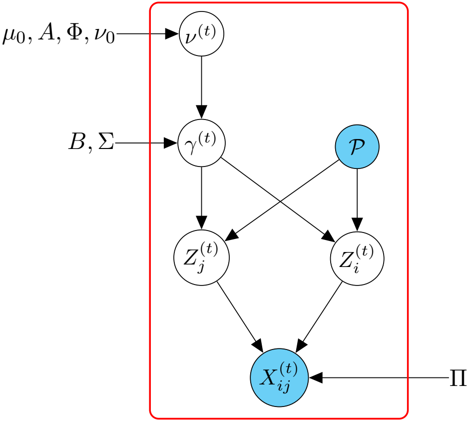


Fig. Graphical representation of the dRSM model.

Analysis of the maritime flow network: The data

We considered the data from Ducruet (2013):

- data from Lloyd's List (Voyage Record) covering the period 1890-2008 at 17 time points,
- huge work to extract from paper versions and complement the lacks (capacity, ...),
- the data contains 176 095 vessels between 4472 ports but we had to reduce to the 286 ports always existing,
- 4 types of relations between ports are considered: liquid bulk, passengers, containers and solid bulk.

Analysis of the maritime flow network: The data

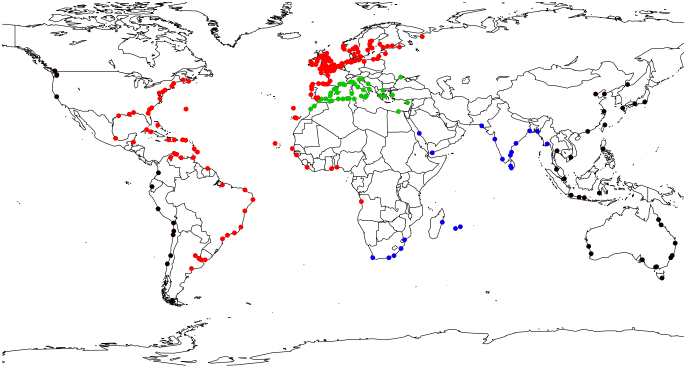


Figure: Map of the ports and their maritime basin.

Analysis of the maritime flow network: The results

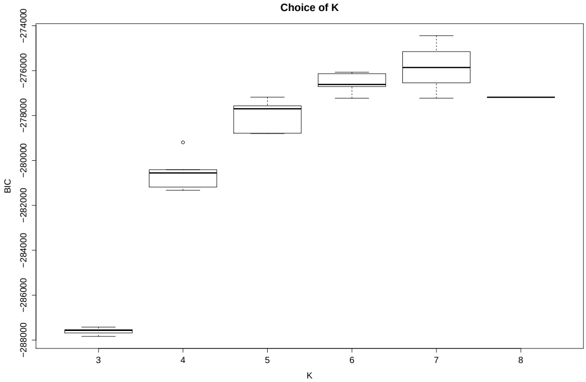


Figure: Choice of the number of groups according to BIC.

Analysis of the maritime flow network: The results

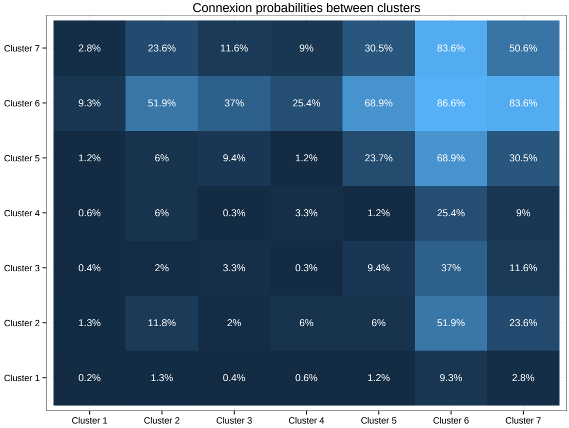
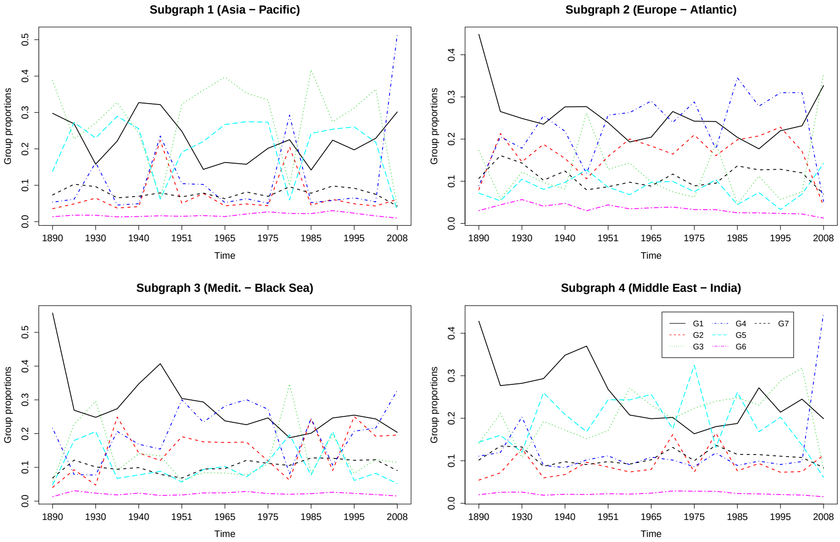


Figure: Estimated values for Π^1_{kl}.

Analysis of the maritime flow network: The results



Outline

Conclusion

Our contributions:

- the RSM model takes into account an existing partition into subgraphs,
- this modeling allows afterward a comparison of the subgraphs,
- the dRSM model allows to deal with evolving networks.

Conclusion

Our contributions:

- the RSM model takes into account an existing partition into subgraphs,
- this modeling allows afterward a comparison of the subgraphs,
- the dRSM model allows to deal with evolving networks.

Software:

package Rambo for the R software is available on the CRAN

Publication:

C. Bouveyron, L. Jegou, Y. Jernite, S. Lamassé, P. Latouche & P. Rivera, *The random subgraph model for the analysis of an ecclesiastical network in merovingian Gaul*, The Annals of Applied Statistics, 8(1), 377-405, 2014.

C. Bouveyron, P. Latouche and R. Zreik, *The Dynamic Random Subgraph Model for the Clustering of Evolving Networks*, Preprint HAL n°01122393, Laboratoire MAP5, Université Paris Descartes, 2015.

C. Bouveyron, C. Ducruet, P. Latouche and R. Zreik, *Cluster Identification in Maritime Flows with Stochastic Methods*, in Maritime Networks: Spatial Structures and Time Dynamics, Routledge, 2015.