The stochastic topic block model

Pierre Latouche (C. Bouveyron)

Journée AppliBUGS, AgroParisTech

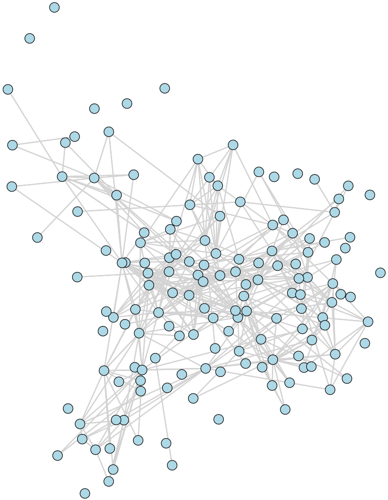
Introduction

- High-dimensional data
  - Sparse regression
  - Sparse probabilistic PCA
  - Biological applications
- Heterogenous data
  - Mix image and text data
  - Curie project. Breast cancer
  - Mix network and text data
- Networks
  - Graphon
  - Applications in social sciences

Plan de l'exposé

Outline

the Enron Email dataset (2001)



Nodes + edges

Introduction

Types of networks: ($\rightarrow$ development of statistical approaches)

- Binary + static edges
- Discrete / continuous / categorical / ...
- Covariates on vertices / edges
- Dynamic edges:
  - Continous time $\rightarrow$ point processes
  - Discrete time $\rightarrow$ Markov,...

Types of clusters: ($\rightarrow$ development of statistical approaches)

- Communities (transitivity)
- Heterogeneous clusters
- Partitions, overlapping clusters, hierarchy

Introduction

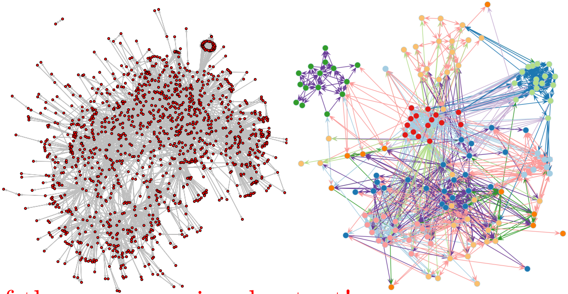
Essentially, two starting points:

- The latent position model [HRH02]
- The stochastic block model [WW87, NS01]

Introduction

Networks can be observed directly or indirectly from a variety of sources:

- social websites (Facebook, Twitter, ...),
- personal emails (from your Gmail, Clinton's mails, ...),
- emails of a company (Enron Email data),
- digital/numeric documents (Panama papers, co-authorships, ...),
- and even archived documents in libraries (digital humanities).



⇒ most of these sources involve text!

Introduction

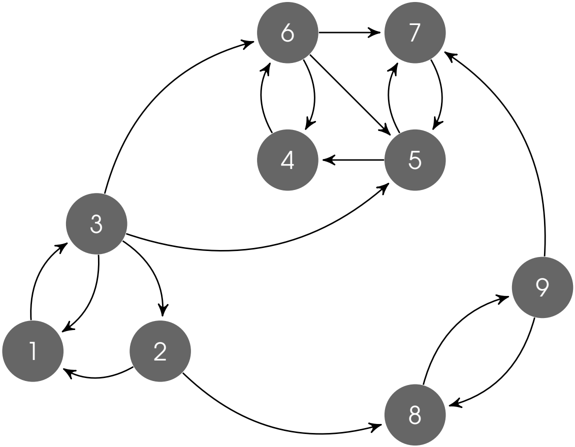


Figure: An (hypothetic) email network between a few individuals.

Introduction

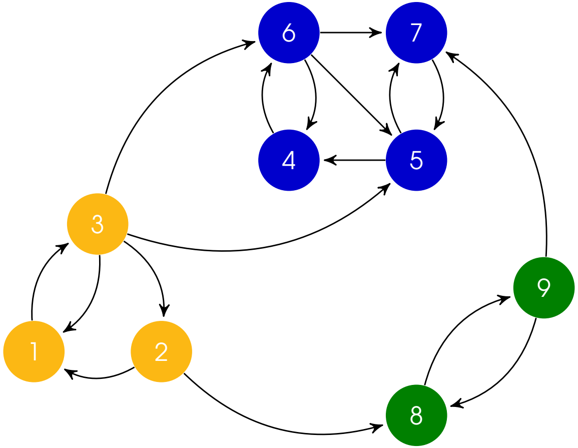


Figure: A typical clustering result for the (directed) binary network.

Introduction

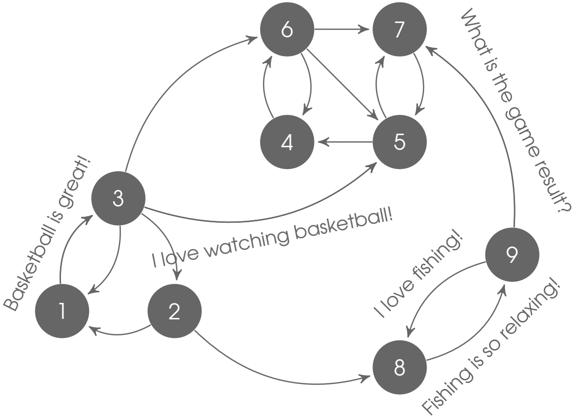


Figure: The (directed) network with textual edges.

Introduction

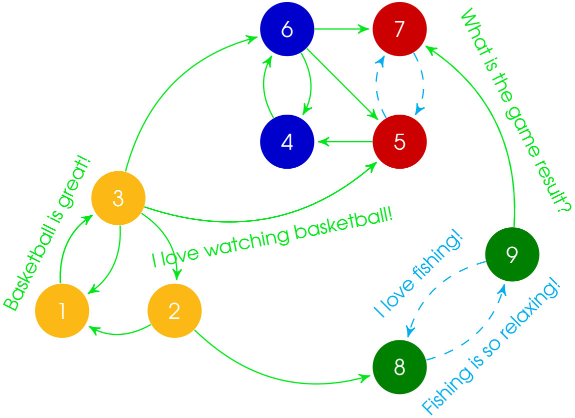


Figure: Expected clustering result for the (directed) network with textual edges.

The stochastic topic block model

the stochastic topic block model (STBM) [BLZ16]:

- generalizes both SBM and LDA models
- allows to analyze (directed and undirected) networks with textual edges.

Outline

Context and notations

We are interesting in clustering the nodes of a (directed) network of M vertices into Q groups:

Context and notations

We are interesting in clustering the nodes of a (directed) network of M vertices into Q groups:

- the network is represented by its $M \times M$ adjacency matrix A:

$$A_{ij} = \begin{cases} 1 & \text{if there is an edge between i and j} \\ 0 & \text{otherwise} \end{cases}$$

Context and notations

We are interesting in clustering the nodes of a (directed) network of M vertices into Q groups:

- the network is represented by its $M \times M$ adjacency matrix A:

$$A_{ij} = \begin{cases} 1 & \text{if there is an edge between i and j} \\ 0 & \text{otherwise} \end{cases}$$

- if $A_{ij} = 1$, the textual edge is characterized by a set of D_{ij} documents:

$$W_{ij} = (W_{ij}^1, ..., W_{ij}^d, ..., W_{ij}^{D_{ij}})$$

Context and notations

We are interesting in clustering the nodes of a (directed) network of M vertices into Q groups:

- the network is represented by its $M \times M$ adjacency matrix A:

$$A_{ij} = \begin{cases} 1 & \text{if there is an edge between i and j} \\ 0 & \text{otherwise} \end{cases}$$

- if $A_{ij} = 1$, the textual edge is characterized by a set of D_{ij} documents:

$$W_{ij} = (W_{ij}^1, ..., W_{ij}^d, ..., W_{ij}^{D_{ij}})$$

- each document W_{ij}^d is made of N_{ij}^d words:

$$W_{ij}^d = (W_{ij}^{d1}, ..., W_{ij}^{dn}, ..., W_{ij}^{dN_{ij}^d}).$$

Modeling of the edges

Let us assume that edges are generated according to a SBM model:

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

$$Y_i \sim \mathcal{M}(1, \rho),$$

where $\rho \in [0,1]^Q$ is the vector of group proportions,

Modeling of the edges

Let us assume that edges are generated according to a SBM model:

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

$$Y_i \sim \mathcal{M}(1, \rho),$$

  where $\rho \in [0,1]^Q$ is the vector of group proportions,

- the presence of an edge A_{ij} between i and j is drawn according to:

$$A_{ij} | Y_{iq} Y_{jr} = 1 \sim \mathcal{B}(\pi_{qr}),$$

  where $\pi_{qr} \in [0,1]$ is the connection probability between clusters q and r.

Modeling of the documents

The generative model for the documents is as follows:

- each pair of clusters (q, r) is first associated to a vector of topic proportions $\theta_{qr} = (\theta_{qrk})_k$ sampled from a Dirichlet distribution:

$$\theta_{qr} \sim \text{Dir}\left(\alpha\right),$$

  such that $\sum_{k=1}^{K} \theta_{qrk} = 1, \forall (q, r)$.

Modeling of the documents

The generative model for the documents is as follows:

- each pair of clusters (q, r) is first associated to a vector of topic proportions $\theta_{qr} = (\theta_{qrk})_k$ sampled from a Dirichlet distribution:
$$\theta_{qr} \sim \mathrm{Dir}\left(\alpha\right),$$
such that $\sum_{k=1}^{K} \theta_{qrk} = 1, \forall (q, r)$.
- the nth word W_{ij}^{dn} of documents d in W_{ij} is then associated to a latent topic vector Z_{ij}^{dn} according to:
$$Z_{ij}^{dn} | \left\{ A_{ij} Y_{iq} Y_{jr} = 1, \theta \right\} \sim \mathcal{M}\left(1, \theta_{qr}\right).$$

Modeling of the documents

The generative model for the documents is as follows:

- each pair of clusters (q, r) is first associated to a vector of topic proportions $\theta_{qr} = (\theta_{qrk})_k$ sampled from a Dirichlet distribution:

$$\theta_{qr} \sim \mathrm{Dir}\left(\alpha\right),$$

  such that $\sum_{k=1}^{K} \theta_{qrk} = 1, \forall (q, r)$.

- the nth word W_{ij}^{dn} of documents d in W_{ij} is then associated to a latent topic vector Z_{ij}^{dn} according to:

$$Z_{ij}^{dn} | \left\{ A_{ij} Y_{iq} Y_{jr} = 1, \theta \right\} \sim \mathcal{M}\left(1, \theta_{qr}\right).$$

- then, given Z_{ij}^{dn}, the word W_{ij}^{dn} is assumed to be drawn from a multinomial distribution:

$$W_{ij}^{dn} | Z_{ij}^{dnk} = 1 \sim \mathcal{M}\left(1, \beta_k = (\beta_{k1}, \dots, \beta_{kV})\right),$$

  where V is the vocabulary size.

Modeling of the documents

- notice that the two previous equations lead to the following mixture model for words over topics:

$$W_{ij}^{dn} | \{Y_{iq}Y_{jr}A_{ij} = 1, \theta\} \sim \sum_{k=1}^{K} \theta_{qrk}\mathcal{M}(1, \beta_k).$$

STBM at a glance...

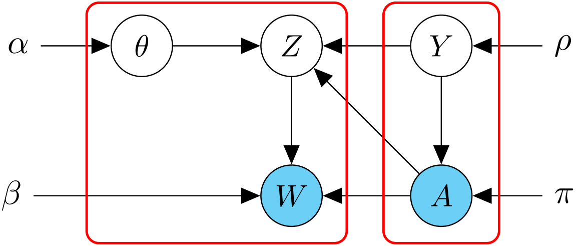


Figure: The stochastic topic block model.

STBM at a glance...

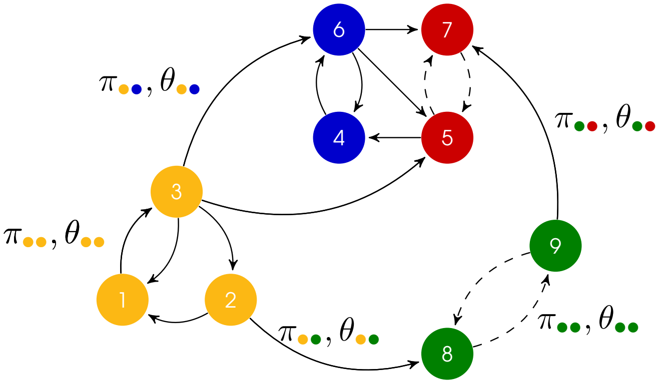


Figure: The stochastic topic block model.

Inference

The full joint distribution of the STBM model is given by:

$$p(A, W, Y, Z, \theta | \rho, \pi, \beta) = p(W, Z, \theta | A, Y, \beta) p(A, Y | \rho, \pi).$$

Inference

The full joint distribution of the STBM model is given by:

$$p(A, W, Y, Z, \theta | \rho, \pi, \beta) = p(W, Z, \theta | A, Y, \beta) p(A, Y | \rho, \pi).$$

A key property of the STMB model:

- let us assume that Y is observed (groups are known),

Inference

The full joint distribution of the STBM model is given by:

$$p(A, W, Y, Z, \theta|\rho, \pi, \beta) = p(W, Z, \theta|A, Y, \beta)p(A, Y|\rho, \pi).$$

A key property of the STMB model:

- let us assume that Y is observed (groups are known),
- it is then possible to reorganize the documents $D = \sum_{i,j} D_{ij}$ documents W such that:

$$W = (\tilde{W}_{qr})_{qr} \text{ where } \tilde{W}_{qr} = \left\{ W_{ij}^{d}, \forall (d, i, j), Y_{iq}Y_{jr}A_{ij} = 1 \right\},$$

Inference

The full joint distribution of the STBM model is given by:

$$p(A, W, Y, Z, \theta | \rho, \pi, \beta) = p(W, Z, \theta | A, Y, \beta) p(A, Y | \rho, \pi).$$

A key property of the STMB model:

- let us assume that Y is observed (groups are known),
- it is then possible to reorganize the documents $D = \sum_{i,j} D_{ij}$ documents W such that:

$$W = (\tilde{W}_{qr})_{qr} \text{ where } \tilde{W}_{qr} = \left\{ W_{ij}^{d}, \forall (d, i, j), Y_{iq} Y_{jr} A_{ij} = 1 \right\},$$

- since all words in $\tilde{W}_{qr}$ are associated with the same pair (q, r) of clusters, they share the same mixture distribution,
- and, simply seeing $\tilde{W}_{qr}$ as a document d, the sampling scheme then corresponds to the one of a LDA model with $D = Q^2$ documents.

Inference

Given the above property of the model, we propose for inference to maximize the complete data log-likelihood:

$$\log p(A, W, Y|\rho, \pi, \beta) = \log \sum_{Z} \int_{\theta} p(A, W, Y, Z, \theta|\rho, \pi, \beta) d\theta,$$

with respect to (ρ, π, β) and $Y = (Y_1, \ldots, Y_M)$.

Inference: the C-VEM algorithm

The C(-V)EM algorithm makes use of a variational decomposition:

$$\log p(A, W, Y|\rho, \pi, \beta) = \mathcal{L}\left(R; Y, \rho, \pi, \beta\right) + \mathrm{KL}\left(R \parallel p(\cdot|A, W, Y, \rho, \pi, \beta)\right),$$

where

$$\mathcal{L}\left(R(\cdot); Y, \rho, \pi, \beta\right) = \sum_{Z} \int_{\theta} R(Z, \theta) \log \frac{p(A, W, Y, Z, \theta|\rho, \pi, \beta)}{R(Z, \theta)} d\theta,$$

and $R(\cdot)$ is assumed to factorize as follows:

$$R(Z, \theta) = R(Z)R(\theta) = R(\theta) \prod_{i \neq j, A_{ij}=1}^{M} \prod_{d=1}^{D_{ij}} \prod_{n=1}^{N_{ij}^{d}} R(Z_{ij}^{dn}).$$

Inference: the C-VEM algorithm

The lower bound is given by:

$$\mathcal{L}\left(R(\cdot); Y, \rho, \pi, \beta\right) = \tilde{\mathcal{L}}\left(R(\cdot); Y, \beta\right) + \log p(A, Y|\rho, \pi),$$

where

$$\tilde{\mathcal{L}}\left(R(\cdot); Y, \beta\right) = \sum_{Z} \int_{\theta} R(Z, \theta) \log \frac{p(W, Z, \theta|A, Y, \beta)}{R(Z, \theta)} d\theta,$$

and $\log p(A, Y|\rho, \pi)$ is the complete data log-likelihood of the SBM model.

Inference: the C-VEM algorithm

The lower bound is given by:

$$\mathcal{L}\left(R(\cdot); Y, \rho, \pi, \beta\right) = \tilde{\mathcal{L}}\left(R(\cdot); Y, \beta\right) + \log p(A, Y|\rho, \pi),$$

where

$$\tilde{\mathcal{L}}\left(R(\cdot); Y, \beta\right) = \sum_{Z} \int_{\theta} R(Z, \theta) \log \frac{p(W, Z, \theta|A, Y, \beta)}{R(Z, \theta)} d\theta,$$

and $\log p(A, Y|\rho, \pi)$ is the complete data log-likelihood of the SBM model.
Algorithm: maximize the lower bound with respect to $R(\cdot), Y, \rho, \pi, \beta$, in turn

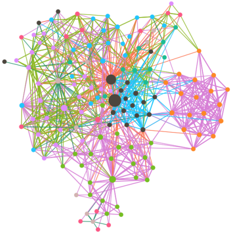
Outline

Conclusion

- STBM : allows to model networks with textual edges
- C-VEM algorithm for inference
- Model selection criterion
- Find clusters of nodes and topics of discussions

Linkage.fr

Linkage   New Job   Jobs   pierre



Innovative and efficient cluster analysis of networks with textual edges

Linkage allows you to cluster the nodes of networks with textual edges while identifying topics which are used in communications. You can analyze with Linkage networks such as email networks or co-authorship networks. Linkage allows you to upload your own network data or to make requests on scientific databases (Arxiv, Pubmed, HAL).

Try Linkage

Linkage.fr

Biblio I

- Charles Bouveyron, Pierre Latouche, and Rawya Zreik, *The stochastic topic block model for the clustering of vertices in networks with textual edges*, Statistics and Computing (2016), 1–21.
- Peter D Hoff, Adrian E Raftery, and Mark S Handcock, *Latent space approaches to social network analysis*, Journal of the american Statistical association **97** (2002), no. 460, 1090–1098.
- K. Nowicki and T.A.B. Snijders, *Estimation and prediction for stochastic blockstructures*, Journal of the American Statistical Association **96** (2001), 1077–1087.
- Y.J. Wang and G.Y. Wong, *Stochastic blockmodels for directed graphs*, Journal of the American Statistical Association **82** (1987), 8–19.