Statistical clustering of temporal networks through a dynamic stochastic block model

Catherine Matias

CNRS - Sorbonne Université - Université de Paris,
Paris, FRANCE
catherine.matias@math.cnrs.fr
`http://cmatias.perso.math.cnrs.fr/`

Joint work with Vincent Miele (LBBE, CNRS, Université Lyon 1, France)

Outline

Introduction: contact networks

Clustering dynamic networks

Simulations

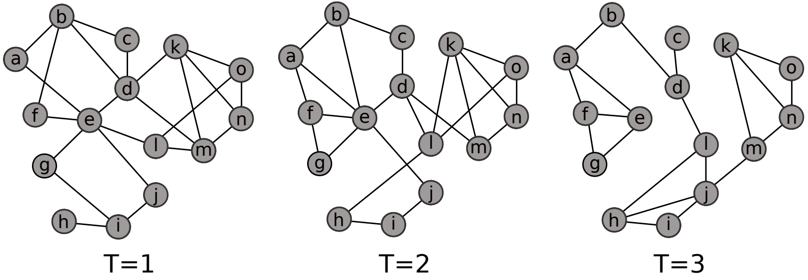
Real data set

Static contact networks

Contact network analysis

- Different ’individuals’ may be at stake : humans, animals, species, ...
- Contact networks may be built from:
  - sensors-based measurements (humans, animals),
  - declarations in surveys (e.g: friendship relations between humans),
  - field observations of associations between animals (e.g: physical proximity) or species (e.g: trophic relationships),
  - trapping data (animals), ...
- With different aims: studying sociability (humans, animals) or Ecology (animals, species)
- Formal definition
  - N individuals $\equiv$ N nodes (vertices)
  - presence/absence of contact or frequency of contact $\equiv$ edges (links)

Dynamic contact networks



Made of

1. snapshots of a contact networks at different time steps (hour, day, week, season...)
2. individuals may be present/absent at each time step

Formal definition:

- T time steps ; N_t individuals $\equiv$ N_t nodes at time step t
- N individuals in total ($N << N_1 + \ldots + N_t$: many individuals stay present across time)
- presence/absence or frequency of contacts $\equiv$ edges at each time step

Studying dynamic contact networks I

- Is there a social structure?
  Understanding if there is a peculiar non-random mixing of individuals that would be a sign for a social organisation.
- What is its dynamics?
- How does it vary with other factors?
  e.g: seasonal changes, breeding season, response to stress, arrival/departure of a peculiar individual,...
- How can we predict how infectious diseases can spread?

Studying dynamic contact networks II

Here we focus on

- Is there a social structure?
- What is its dynamics?
- (How does it vary with other factors?)
- (How can we predict how infectious diseases can spread?)

Our answer:

- moving beyond descriptive statistics and proposing a statistical model for the organisation in these networks.
- We rely on a clustering of the nodes to capture a social structure

Outline

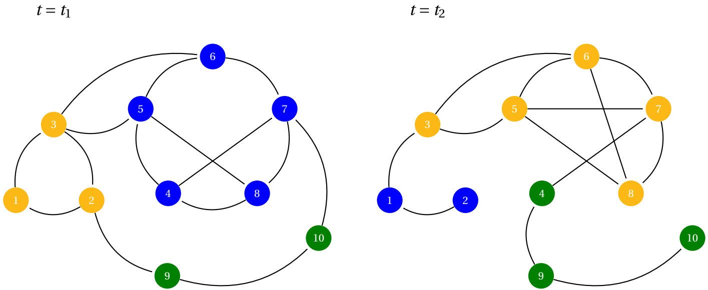
Introduction: contact networks

Clustering dynamic networks

Simulations

Real data set

Clustering dynamic networks I



Issues

- Deal with the label switching across time.
- See the evolution of individual nodes: who is changing group between 2 time points?

Our goal: smooth recovery of the clusters across time.

Our contributions

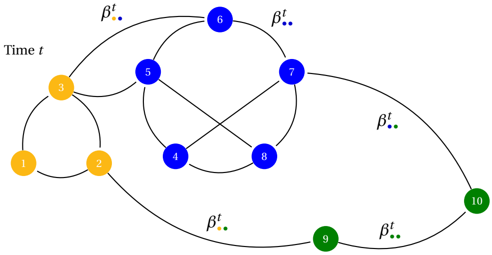
Model: dynamic SBM

- We propose a dynamic version of the Stochastic Block Model;
- The graphs may be directed or undirected, binary or weighted;
- Groups and model parameters may change through time;
- Careful discussion on identifiability conditions on the model.

Inference

- We propose a variational expectation maximisation (`VEM`) algorithm to infer the nodes groups across time and the model parameters;
- We have a model selection criterion (ICL type) to select for the number of groups.

Static part modeling: SBM - binary case



$n = 10, Q = 3,$

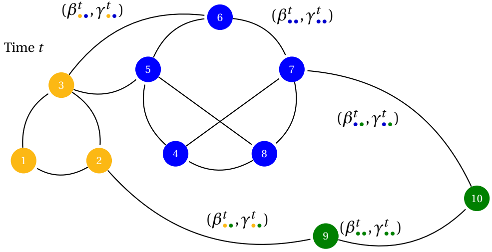
$Z_5^t = \bullet,$

$Y_{12}^t = 1, Y_{15}^t = 0$

Binary case; parameter $\boldsymbol{\beta}^t = (\beta^t_{ql})_{1\le q\le l\le Q}$

- Q groups (=colors $\bullet\bullet\bullet$).
- $\{Z_i^t\}_{1\le i\le n}$ i.i.d. in $\{1,\ldots,Q\}$ not observed.
- Observations: presence/absence of an edge at time t, given through adjacency matrix $\{Y_{ij}^t\}_{1\le i<j\le n}$,
- Conditional on $\{Z_i^t\}$'s, the r.v. Y_{ij}^t are independent $\mathscr{B}(\beta^t_{Z_i^t Z_j^t})$.

Static part modeling: SBM - weighted case



$n = 10, Q = 3,$

$Z_5^t = \bullet,$

$Y_{12}^t \in \mathbb{R}^s, Y_{15}^t = 0$

Weighted case; parameter $(\boldsymbol{\beta}^t, \boldsymbol{\gamma}^t) = (\beta_{ql}^t, \gamma_{ql}^t)_{1 \le q \le l \le Q}$

- Latent variables: *idem*
- Observations: weights Y_{ij}^t, where $Y_{ij}^t = 0$ or $Y_{ij}^t \in \mathbb{R}^s \setminus \{0\}$,
- Conditional on the $\{Z_i^t\}$'s, the random variables Y_{ij}^t are independent with density

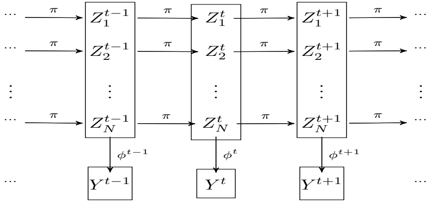
$$\phi(\cdot; \beta_{Z_i^t Z_j^t}^t, \gamma_{Z_i^t Z_j^t}^t) := (1 - \beta_{Z_i^t Z_j^t}^t)\delta_0(\cdot) + \beta_{Z_i^t Z_j^t}^t f(\cdot, \gamma_{Z_i^t Z_j^t}^t),$$

(Assumption: f has continuous cdf at zero).

Dynamics: Markov chain on latent groups

Latent Markov chain

- Across individuals: $(Z_i)_{1 \leq i \leq N}$ iid,
- Across time: Each $Z_i = (Z_i^t)_{1 \leq t \leq T}$ is a stationary Markov chain on $\{1, \ldots, Q\}$ with transition $\boldsymbol{\pi} = (\pi_{qq'})_{1 \leq q, q' \leq Q}$ and initial stationary distribution $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_Q)$.



Goal

Infer the parameter $\theta = (\boldsymbol{\pi}, \boldsymbol{\beta}, \boldsymbol{\gamma})$, recover the clusters $\{Z_i^t\}_{i,t}$ and follow their evolution through time.

Outline

Clustering performances I

Indexes

- Global ARI: Adjusted Rand Index on the whole classification $\{Z_i^t\}_{1 \leq i \leq N, 1 \leq t \leq T}$,
- Averaged ARI: mean value of ARI_t, computed for each t on the classification $\{Z_i^t\}_{1 \leq i \leq N}$. Easier ! Label switching between time steps !

Clustering performances II

Simulations setup

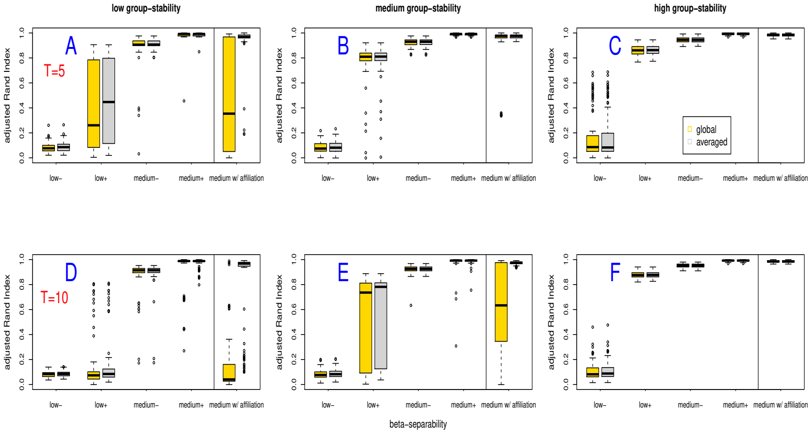
- Binary graphs, $N = 100$ nodes and $T \in \{5; 10\}$, 100 datasets,
- $Q = 2$ latent groups and $\boldsymbol{\pi} \in \{\boldsymbol{\pi}_{low}, \boldsymbol{\pi}_{med}, \boldsymbol{\pi}_{high}\}$

$$\boldsymbol{\pi}_{low} = \begin{pmatrix} 0.6 & 0.4 \\ 0.4 & 0.6 \end{pmatrix}; \boldsymbol{\pi}_{med} = \begin{pmatrix} 0.75 & 0.25 \\ 0.25 & 0.75 \end{pmatrix}; \boldsymbol{\pi}_{high} = \begin{pmatrix} 0.9 & 0.1 \\ 0.1 & 0.9 \end{pmatrix}.$$

- Connectivity parameter $\boldsymbol{\beta}$

Easiness	β_{11}	β_{12}	β_{22}
low-	0.2	0.1	0.15
low+	0.25	0.1	0.2
medium-	0.3	0.1	0.2
medium+	0.4	0.1	0.2
med w/ affiliation	0.3	0.1	0.3

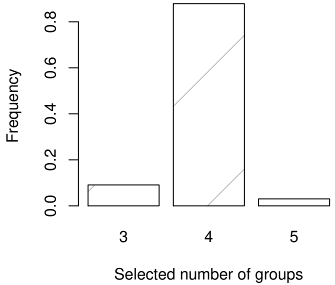
Clustering performances III

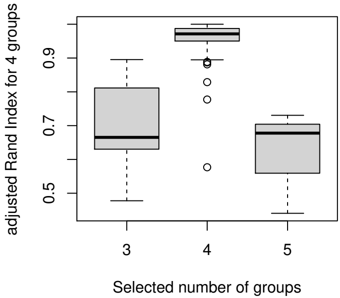


Model selection

Simulation setup

- Binary model, $Q = 4$ groups, $\pi_{qq} = 0.91$ and $\pi_{ql} = 0.03$ for $q \neq l$, 100 datasets
- We draw i.i.d. random variables $\{\epsilon_{ql}\}_{1 \leq q \leq l \leq 4} \in [-1, 1]$ and then choose $\beta_{qq} = 0.4 + \epsilon_{qq} 0.1$ and $\beta_{ql} = 0.1 + \epsilon_{ql} 0.1$ for $q \neq l$.





Outline

Introduction: contact networks

Clustering dynamic networks

Simulations

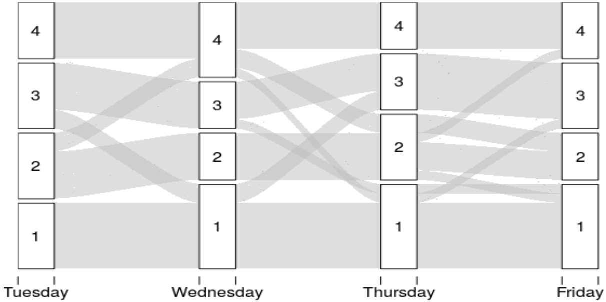
Real data set

Encounters between high school students I

Fournet and Barrat, 2014, `http://www.sociopatterns.org/`

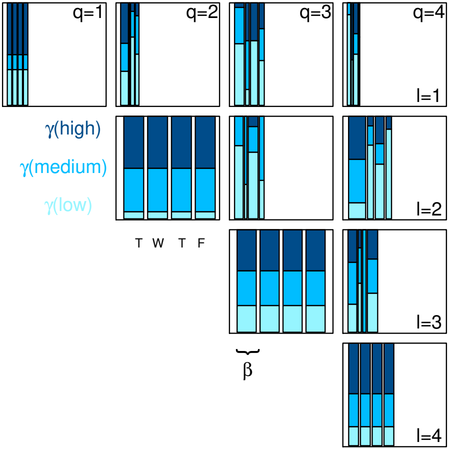
- Face-to-face encounters of high school students (wearable sensors), $T = 4$ days, $N = 27$ students,
- Discrete weight with 3 bins. Selection of $Q = 4$ groups.

Reconstructed dynamics



Encounters between high school students II

Estimated connectivity parameters



Encounters between high school students III

Conclusions on this dataset

- groups 2 and 3 are communities, with resp. 3 and 4 individuals who permanently stay in the groups (social attractors),
- group 4 is also a community, with much less interaction,
- group 2 and 4 exchange students,
- Group 1 stable, low rate of interaction,

A posteriori crossing info with gender

- group 3 has male students only,
- group 1 has a backbone of female students that remain in the group
- females are less likely to change groups than males, a majority of females belongs low interaction groups 1 and 4 and they do not switch between these groups.

Conclusions

DynamicSBM

- Reconstruction of group's evolution through time
- Control of the label switching issue between different time steps
- Models binary or weighted datasets
- Model selection performed through ICL.

R package `dynsbm` available on the CRAN.
Paper published in JRSSB, 2017.
Companion paper (for Ecologists) in Royal Society Open Science, 2017.

Thanks for your attention !