Clustering dynamic random graphs

Catherine Matias

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

Journées de Statistique
Saclay, May 2018
Outline

Dynamic Random Graphs: the data

Graphs clustering: different approaches

The stochastic block model

Clustering dynamic networks
  Clustering graphs sequences
  Clustering links streams (with no duration)
Dynamic interactions data

Types of data and their representation

One should distinguish between

- **Long time relations** (eg social relations, physical wiring of routers, ...): graphs sequences
- **Short time interactions** (eg: phone call, physical encounter, ...): temporal networks or stream links

For a nice review, see [Holme(2015)].
Pictures that follow are from [Gaumont(2016)].
Chapitre 1. Etat de l'art en fonction des contacts qui existent entre les personnes. Il est donc tout naturel de manipuler une série de graphes.

Plus formellement, une série de graphe est définie par 
\[ G = \{ G_i \} \text{ i<T o} \text{u T est un entier (voir la figure 1.3 ).} \]

Graphs sequences

Figure 1.3 – Exemple de série de graphes sur trois intervalles de temps.

**Remarks**

- In practice, there could be small variations in the individuals present at each time step,
- These data are sometimes obtained through aggregation
  - possible loss of information
  - problem of choosing the time window for aggregation.
Temporal networks

Remarks

- Again, variations in node presence/absence is possible,
- Here, there is no loss of information.
- Ideal setup in the sense that most of the time, we do not have all this knowledge.
Remarks

- Here, there is no underlying graph!
- One could add in the data (and in its visualisation) the info that one individual is not present during some time periods,
- Again, no loss of information.
Outline

Dynamic Random Graphs: the data

Graphs clustering: different approaches

The stochastic block model

Clustering dynamic networks
  Clustering graphs sequences
  Clustering links streams (with no duration)
Graph clustering: why and how? I

Why?

- Networks are intrinsically heterogeneous: need to account for different nodes behaviours,
- Summarise network information through a higher-level view (zoom-out the network),
- Some networks exhibit modularity: modules or communities are groups of nodes with high number of intra-connections and low number of outer-connections;
- Other structures might be of interest: hierarchical groups, hubs, periphery nodes, homophilic/heterophilic structures, …
Graph clustering: why and how? II

How?
Many methods, with different aims

- Searching for **communities**,
  - Modularity-based approaches;
  - Random walk algorithms;
  - Spectral clustering (NB: absolute spectral clust. also captures heterophilic struct.);
  - Latent space models by [Hoff et al.(2002)].

- Searching for groups, without any a priori on their structure: **Stochastic block models (SBMs)**. SBMs search for groups of nodes with a similar connectivity behaviour towards the other groups.

- Recently, mixtures of ERGMs [Vu et al.(2013)].
Outline

Dynamic Random Graphs: the data

Graphs clustering: different approaches

The stochastic block model

Clustering dynamic networks
  Clustering graphs sequences
  Clustering links streams (with no duration)
Stochastic block model (binary graphs)

Binary case (parametric model with $\theta = (\pi, \gamma)$)

- $K$ groups (=colors $\bullet$).
- $\{Z_i\}_{1 \leq i \leq n}$ i.i.d. vectors $Z_i = (Z_{i1}, \ldots, Z_{iK}) \sim \mathcal{M}(1, \pi)$, with $\pi = (\pi_1, \ldots, \pi_K)$ groups proportions. $Z_i$ not observed (latent).
- Observations: presence/absence of an edge $\{A_{ij}\}_{1 \leq i < j \leq n}$.
- Conditional on $\{Z_i\}$'s, the r.v. $A_{ij}$ are independent $\mathcal{B}(\gamma_{Z_iZ_j})$.

$n = 10, Z_5. = 1$

$A_{12} = 1, A_{15} = 0$
Stochastic block model (weighted graphs)

Weighted case (parametric model with $\theta = (\pi, \gamma^{(1)}, \gamma^{(2)})$)

- Latent variables: *idem*
- Observations: 'weights' $A_{ij}$, where $A_{ij} = 0$ or $A_{ij} \in \mathbb{R}^s \setminus \{0\}$,
- Conditional on the $\{Z_i\}$'s, the random variables $A_{ij}$ are independent with distribution

$$
\mu_{Z_iZ_j}(\cdot) = \gamma_{Z_iZ_j}^{(1)} f(\cdot, \gamma_{Z_iZ_j}^{(2)}) + (1 - \gamma_{Z_iZ_j}^{(1)}) \delta_0(\cdot)
$$

$n = 10, Z_{5.} = 1$

$A_{12} \in \mathbb{R}, A_{15} = 0$
SBM classification vs community detection

SBM classification

- Nodes classification induced by the model reflects a common connectivity behaviour;
- Community detection methods focus on communities;
- Toy example

SBM clusters

Community detection or SBM
Particular cases and generalisations

Particular case: Affiliation model (planted partition)

\[ \gamma = \begin{pmatrix} \alpha & \ldots & \beta \\ \vdots & \ddots & \vdots \\ \beta & \ldots & \alpha \end{pmatrix} \quad (\alpha \gg \beta \implies \text{community detection}) \]

Some generalisations

- Overlapping groups
  [Latouche et al. (2011), Airoldi et al. (2008)] for binary graphs; SBM with covariates; Degree corrected SBM;...

- Latent block models (LBM), for array data or bipartite graphs [Govaert and Nadif (2003)];

- Nonparametric SBM (graphon);

- Dynamic SBM
Overview of algorithms

Goal is MLE. Likelihood computation is untractable for \( n \) not small.

Parameter estimation

- **em algorithm** not feasible because latent variables are not independent conditional on observed ones:
  \[ P(\{Z_i\}_i|\{A_{ij}\}_{i,j}) \neq \prod_i P(Z_i|\{A_{ij}\}_{i,j}) \]

- Alternatives:
  - Gibbs sampling
  - Variational approximation to em.
  - Ad-hoc methods: Composite likelihood or Moment methods
    [Ambroise and M. (2012), Bickel et al. (2011)]; Degrees
    [Channarond et al. (2012)];
Variational approximation principle I

Log-likelihood decomposition

\[ \mathcal{L}_A(\theta) := \log \mathbb{P}(A; \theta) = \log \mathbb{P}(A, Z; \theta) - \log \mathbb{P}(Z|A; \theta) \]
and for any distribution \( Q \) on \( Z \),

\[ \mathcal{L}_A(\theta) = \mathbb{E}_Q(\log \mathbb{P}(A, Z; \theta)) + \mathcal{H}(Q) + \mathcal{KL}(Q\|\mathbb{P}(Z|A; \theta)) \]

\text{em principle}

\begin{itemize}
  \item e-step: maximise the quantity \( \mathbb{E}_Q(\log \mathbb{P}(A, Z; \theta^{(t)})) + \mathcal{H}(Q) \)
  with respect to \( Q \). This is equivalent to minimizing \( \mathcal{KL}(Q\|\mathbb{P}(Z|A; \theta^{(t)})) \)
  with respect to \( Q \).
  \item m-step: keeping now \( Q \) fixed, maximize the quantity \( \mathbb{E}_Q(\log \mathbb{P}(A, Z; \theta)) + \mathcal{H}(Q) \)
  with respect to \( \theta \) and update the parameter value \( \theta^{(t+1)} \) to this maximiser. This is equivalent to maximizing the conditional expectation
  \( \mathbb{E}_Q(\log \mathbb{P}(A, Z; \theta)) \) w.r.t. \( \theta \).
\end{itemize}
Variational approximation principle II

Variational em

- **e-step**: search for an optimal $Q$ within a restricted class $Q$, e.g. class of factorized distr.

$$Q(Z) = \prod_{i=1}^{n} Q(Z_i), \quad Q^* = \arg\min_{Q \in Q} KL(Q \parallel P(Z|A; \theta(t)))$$

- **m-step**: unchanged, i.e.

$$\theta^{(t+1)} = \arg\max_{\theta} \mathbb{E}_{Q^*} (\log P(A, Z; \theta))$$

- A consequence of $KL \geq 0$ is the lower bound

$$\mathcal{L}_A(\theta) \geq \mathbb{E}_{Q}(\log P(A, Z; \theta)) + \mathcal{H}(Q)$$

So that the variational approximation consists in maximizing a lower bound on the log-likelihood. Why does it make sense?
Model selection

How do we choose the number of groups $K$?

Frequentist setting

- Maximal likelihood is not available (thus neither AIC or BIC),
- ICL criterion is used [Daudin et al.(2008)] (no consistency result on that).

Bayesian setting

- MCMC approach to select number of LBM groups [Wyse and Friel(2012)].
- Exact ICL requires greedy search optimization [Côme and Latouche(2015)].
(Some) SBMs packages/codes

VEM implementations

- **MixNet**
  http://www.math-evry.cnrs.fr/logiciels/mixnet is a C/C++ code and MixeR R package on the CRAN: for binary SBM, directed or not;

- **OSBM** R package R for Overlapping SBM,
  http://www.math-evry.cnrs.fr/logiciels/osbm

- **Blockmodels** R package binary/valued SBM, possibly with covariates
Outline

Dynamic Random Graphs: the data

Graphs clustering: different approaches

The stochastic block model

Clustering dynamic networks
  Clustering graphs sequences
  Clustering links streams (with no duration)
Outline

Dynamic Random Graphs: the data

Graphs clustering: different approaches

The stochastic block model

Clustering dynamic networks
   Clustering graphs sequences
   Clustering links streams (with no duration)
Follow the groups through time

Label switching issue in the dynamic context

\[ t = t_1 \]

\[ t = t_2 \]
Follow the groups through time

Label switching issue in the dynamic context

If the 2 classifications are constructed independently, then it’s impossible to follow the groups evolution. It’s thus mandatory to do a joint clustering of the graphs.
Dynsbm: a dynamic stochastic blockmodel

Model [M. & Miele(2017)]

- We simply combine a latent Markov chain with weighted SBMs;
- Our graphs may be directed or undirected, binary or weighted; some individuals can appear or disappear;
- Groups and model parameters may change through time;
- Careful discussion on identifiability conditions on the model.

Inference

- **VEM** algorithm to infer the nodes groups across time and the model parameters;
- Model selection criterion (ICL type) to select for the number of groups.
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 Markov chain on \(\{1, \ldots, Q\}\) with transition \(\pi = (\pi_{qq'})_{1\leq q,q'\leq Q}\) and initial stationary distribution \(\alpha = (\alpha_1, \ldots, \alpha_Q)\).

Goal
Infer the parameter \(\theta = (\pi, \beta, \gamma)\), recover the clusters \(\{Z_i^t\}_{i,t}\) and follow their evolution through time.
Application on ecological networks [Miele & M. (2017)]

Ants dataset [Mersch et al. (2013)]

T = 10, N = 152

Selection of 3 social groups.

Low turnover: 47% of ants do not switch group.

No group switches between groups 1 and 2.
Group 2: a community.

Group 3: contacts with all ants from any groups.

Group 1: avoid contacts with group 2.

Perfect match with the three functional category groups: nurses, foragers and cleaners

<table>
<thead>
<tr>
<th></th>
<th>nurses</th>
<th>foragers</th>
<th>cleaners</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>42</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>29</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>1</td>
<td>29</td>
</tr>
</tbody>
</table>

(75% of ants, staying at least 8/10 steps in same group)
Outline

Dynamic Random Graphs: the data

Graphs clustering: different approaches

The stochastic block model

Clustering dynamic networks
  Clustering graphs sequences
    Clustering links streams (with no duration)
Longitudinal interaction networks = Stream links view
Longitudinal interaction networks = point process view

We observe a marked point process: the mark is a pair of individuals $(i, j)$ that interact at time $t$.

Goal: cluster the individuals $i$ (not the processes $N_{ij}$ !)
ppsbm: a dynamic point process SBM

Model characteristics [M., Rebafka, Villers(2018)]

- Pointwise interactions with no duration only; Individuals are always present;
- Groups are constant through time;
- Conditional on the latent groups $Z_i, Z_j$, the point process $N_{ij}$ is a non-homogeneous point process with (nonparametric) intensity $t \mapsto \alpha_{Z_i, Z_j}(t)$.
- Recover latent groups $Z = (Z_1, \ldots, Z_n)$ and estimate the intensities per groups pairs $\{\alpha^{(q,l)}(\cdot)\}_{1 \leq q < l \leq Q}$ with VEM

Inference characteristics

- Procedure is semi-parametric: intensities may either be estimated through histograms (with adaptive selection of the partition), or kernels.
- ICL to select the number of groups $Q$. 
London Santander cycles

Data

- Cycles journeys from the Santander cycles hiring stations: departure station, arrival station, time of journey start.
- 1st dataset from Wed. February 1st, 2012, with \( n = 415 \) stations (=individuals), and \( M = 17,631 \) journeys (time points)
- 2nd dataset from Thursday February 2nd, 2012: \( n = 417 \) stations, \( M = 16,333 \) journeys.

Model selection of the number of groups \( Q \)

ICL selects 6 groups for both days.
London Santander cycles: geographical projection of the clusters

Clustering for 1st dataset.
The smallest cluster $x I$

- Contains only 2 bike stations, located at Waterloo and King’s Cross
- among the stations with highest activities

Barplots of outgoing $(N_i(\cdot))$ and incoming $(N_{\cdot i}(\cdot))$ processes from the 2 stations $i$ in the smallest cluster: volumes of connections to all other stations during day 1.

The cluster is composed of ’outgoing’ stations in the morning and ’ingoing’ stations in the evening.
The smallest cluster $x$ II

- Stations close to Victoria and Liverpool Street stations also have high activity but not the same temporal profile so they cluster differently,
- This cluster $x$ is due to a specific temporal profile, that would not be captured through a snapshot approach.
- The cluster has strong connections with cluster $\diamond$ that corresponds to business city center.
Conclusions

Dynamic modeling of interactions is still in its early developments, lot of things to improve.
Thank you for your attention!
References I


Matthieu Latapy, Tiphaine Viard, Clémence Magnien Stream Graphs and Link Streams for the Modeling of Interactions over Time *arXiv:1710.04073*

C. Matias and V. Miele. 
Statistical clustering of temporal networks through a dynamic stochastic block model. 
*JRSSB*, 79(4), 1119–1141, 2017

C. Matias, T. Rebafka, and F. Villers. 
A semiparametric extension of the stochastic block model for longitudinal networks. 

D. P. Mersch, A. Crespi, and L. Keller. 
Tracking individuals shows spatial fidelity is a key regulator of ant social organization. 
References V

V. Miele and C. Matias.  
Revealing the hidden structure of dynamic ecological networks.  
*Royal Society Open Science, 4*(6), 170251, 2017

Vu, Hunter & Schweinberger  
Model-based clustering of large networks  

J. Wyse and N. Friel.  
Block clustering with collapsed latent block models.  