Clustering dynamic random graphs

Catherine Matias

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

GDR Statistique et Santé
Oct 2019
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: pone call, physical encounter, ...): temporal networks or stream links

For a nice review, see [Holme(2015)].
Pictures that follow are from [Gaumont(2016)].
**Graphs sequences**

![Graphs sequences](image)

**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

![Temporal graphs](image)

**Figure 1.5** – Graphe temporel avec des ajouts de lien représentés en traits épais verts et des suppressions de lien représentées par des liens pointillés rouges.

**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.
Dans un graphe temporel, une structure de graphe existe à chaque instant. Il est donc possible de calculer après chaque modification l'évolution d'une métrique. Par exemple, il est possible de calculer après l'ajout d'un lien le nouveau degré interne des nœuds concernés par ce changement. En fonction de l'évolution de cette métrique, on décide alors d'ajouter ou de retirer un nœud voire de fusionner deux communautés. Latapy et al. [LHB+12] se basent sur le nombre de liens que partage un nœud avec les communautés environnantes. Ainsi, un nœud est toujours dans la communauté avec laquelle il partage le plus de liens. Shang et al. [SLX+14], Cordeiro et al. [CSG16] et Sun et al. [SHZ+14] se basent sur l'évolution de la modularité. Cependant, ces approches ne permettent pas de reproduire l'ensemble des évolutions de communauté possibles, en particulier l'apparition d'une nouvelle communauté. C'est pourquoi l'évolution de la structure courante peut mener à une structure ayant une faible qualité. Une autre approche a été proposée par Cazabet et al. [CAH10] afin d'améliorer l'évolution de la partition. Ils utilisent une métrique locale basée sur le nombre de chemins de longueur 2 existant entre un nœud et une communauté. Après chaque modification, ils considèrent également la possibilité de créer une nouvelle communauté sous la forme d'une petite clique. Ainsi, ils assurent une meilleure qualité de la partition au cours de l'évolution.

**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, …
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)

$n = 10, Z_5 = 1$
$A_{12} = 1, A_{15} = 0$

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

- $K$ groups (=colors ••).
- $\{Z_i\}_{1 \leq i \leq n}$ 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 $\mathcal{B}(\gamma_{Z_iZ_j})$. 

Diagram:
- Nodes: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10
- Edges: (1, 2), (1, 3), (2, 4), (3, 5), (4, 6), (5, 7), (6, 8), (7, 9), (8, 10), (9, 10)
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

$$
\mu_{Z_i Z_j}(\cdot) = \gamma^{(1)}_{Z_i Z_j} f(\cdot, \gamma^{(2)}_{Z_i Z_j}) + (1 - \gamma^{(1)}_{Z_i Z_j}) \delta_0(\cdot)
$$

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

$A_{12} \in \mathbb{R}, A_{15} = 0$
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](image1.png) ![Community detection or SBM](image2.png)
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

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

- Alternatives:
  - Gibbs sampling
  - Variational approximation to \textit{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 \( Z \),

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

**em principle**

- **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 \( Q \).

- **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. \( \theta \).
Variational approximation principle II

Variational em

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

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

- 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)
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 $Z_i = (Z_t^i)_{1 \leq t \leq T}$ 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 $\mathcal{Z} = (Z_1, \ldots, Z_n)$ 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)

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 1

- 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.
- Try our R packages: `dynsbm` for sequences of graphs and `ppsbm` for stream links.

Thank you for your attention!
References I

E.M. Airoldi, D.M. Blei, S.E. Fienberg, and E.P. Xing.
Mixed-membership stochastic blockmodels.

C. Ambroise and C. Matias.
New consistent and asymptotically normal parameter estimates for random graph mixture models.

P. Bickel, A. Chen and E. Levina
The method of moments and degree distributions for network models

Classification and estimation in the Stochastic Blockmodel based on the empirical degrees.
E. Côme and P. Latouche.
Model selection and clustering in stochastic block models based on the exact integrated complete data likelihood. 
*Statist Model*, 2015.

A mixture model for random graphs. 

N. Gaumont.
*Groupes et communautés dans les flots de liens : des données aux algorithmes.*

G. Govaert and M. Nadif.
Clustering with block mixture models. 


C. Matias and V. Miele.  
Statistical clustering of temporal networks through a dynamic stochastic block model.  

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.  
