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: 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

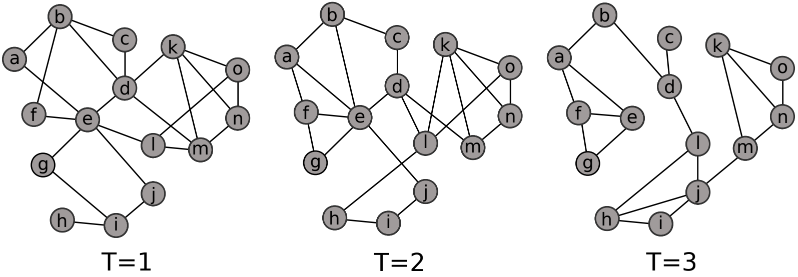


FIGURE 1.3 – Exemple de série de graphes sur trois intervalles de temps.

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

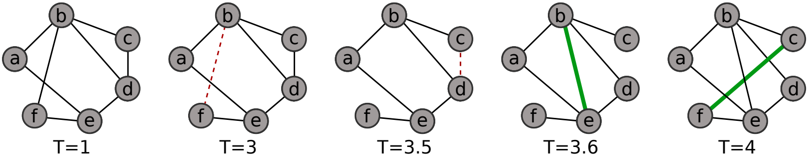
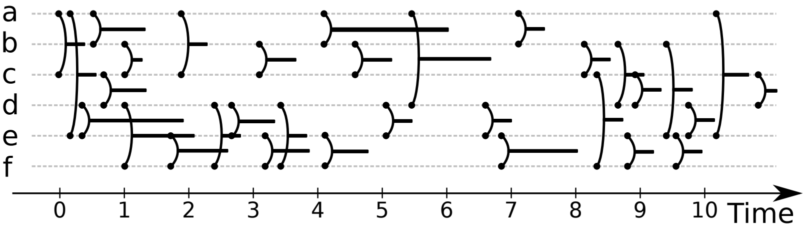


FIGURE 1.5 – Graphe temporel avec des ajouts de lien représentés en traits épais verts et des suppressions de lien représentées par des liens pointillés rouges.

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.

Links streams [Latapy et al.(2017)]



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

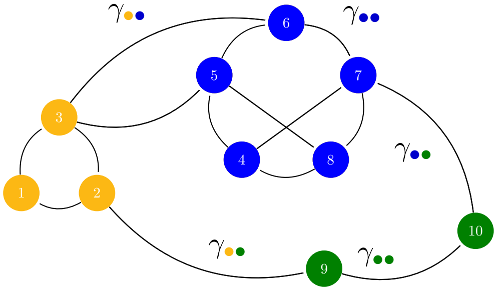
Graphs clustering: different approaches

The stochastic block model

Clustering dynamic networks

- Clustering graphs sequences
- Clustering links streams (with no duration)

Stochastic block model (binary graphs)



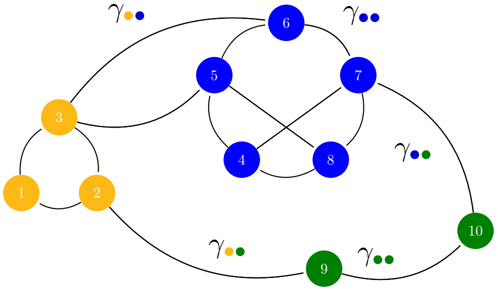
$n = 10, Z_{5\bullet} = 1$

$A_{12} = 1, A_{15} = 0$

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

- K groups (=colors •••).
- $\{Z_i\}_{1 \leq i \leq n}$ i.i.d. vectors $Z_i = (Z_{i1}, \ldots, Z_{iK}) \sim \mathcal{M}(1, \boldsymbol{\pi})$, with $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)$ 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 $\mathcal{B}(\gamma_{Z_i Z_j})$.

Stochastic block model (weighted graphs)



$n = 10, Z_{5\bullet} = 1$

$A_{12} \in \mathbb{R}, A_{15} = 0$

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

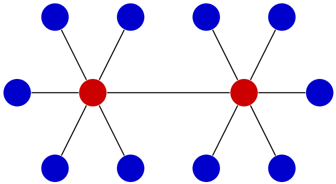
- Latent variables: *idem*
- Observations: 'weights' A_{ij} , where $A_{ij} = 0$ or $A_{ij} \in \mathbb{R}^s \setminus \{0\}$,
- Conditional on the $\{Z_i\}$'s, the random variables A_{ij} are independent with distribution

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

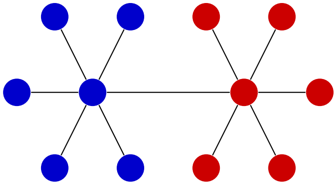
SBM classification vs community detection

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)

$$\boldsymbol{\gamma} = \begin{pmatrix} \alpha & \dots & \beta \\ \vdots & \ddots & \vdots \\ \beta & \dots & \alpha \end{pmatrix} \qquad (\alpha \gg \beta \implies \text{community detection})$$

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: $\mathbb{P}(\{Z_i\}_i|\{A_{ij}\}_{i,j}) \neq \prod_i \mathbb{P}(Z_i|\{A_{ij}\}_{i,j})$
- Alternatives:
  - Gibbs sampling
  - Variational approximation to `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

$\mathcal{L}_{\mathbf{A}}(\boldsymbol{\theta}) := \log \mathbb{P}(\mathbf{A}; \boldsymbol{\theta}) = \log \mathbb{P}(\mathbf{A}, \mathbf{Z}; \boldsymbol{\theta}) - \log \mathbb{P}(\mathbf{Z}|\mathbf{A}; \boldsymbol{\theta})$ and for any distribution $\mathbb{Q}$ on $\mathbf{Z}$,

$$\mathcal{L}_{\mathbf{A}}(\boldsymbol{\theta}) = \mathbb{E}_{\mathbb{Q}}(\log \mathbb{P}(\mathbf{A}, \mathbf{Z}; \boldsymbol{\theta})) + \mathcal{H}(\mathbb{Q}) + \mathcal{KL}(\mathbb{Q} \| \mathbb{P}(\mathbf{Z}|\mathbf{A}; \boldsymbol{\theta}))$$

`em` principle

- `e`-step: maximise the quantity $\mathbb{E}_{\mathbb{Q}}(\log \mathbb{P}(\mathbf{A}, \mathbf{Z}; \boldsymbol{\theta}^{(t)})) + \mathcal{H}(\mathbb{Q})$ with respect to $\mathbb{Q}$. This is equivalent to minimizing $\mathcal{KL}(\mathbb{Q} \| \mathbb{P}(\mathbf{Z}|\mathbf{A}; \boldsymbol{\theta}^{(t)}))$ with respect to $\mathbb{Q}$.
- `m`-step: keeping now $\mathbb{Q}$ fixed, maximize the quantity $\mathbb{E}_{\mathbb{Q}}(\log \mathbb{P}(\mathbf{A}, \mathbf{Z}; \boldsymbol{\theta})) + \mathcal{H}(\mathbb{Q})$ with respect to $\boldsymbol{\theta}$ and update the parameter value $\boldsymbol{\theta}^{(t+1)}$ to this maximiser. This is equivalent to maximizing the conditional expectation $\mathbb{E}_{\mathbb{Q}}(\log \mathbb{P}(\mathbf{A}, \mathbf{Z}; \boldsymbol{\theta}))$ w.r.t. $\boldsymbol{\theta}$.

Variational approximation principle II

Variational `em`

- `e`-step: search for an optimal $\mathbb{Q}$ within a restricted class $\mathcal{Q}$, e.g. class of factorized distr.

$$\mathbb{Q}(\mathbf{Z}) = \prod_{i=1}^{n} \mathbb{Q}(Z_i), \quad \mathbb{Q}^\star = \operatorname*{argmin}_{\mathbb{Q}\in\mathcal{Q}} \mathcal{KL}(\mathbb{Q}\|\mathbb{P}(\mathbf{Z}|\mathbf{A};\boldsymbol{\theta}^{(t)}))$$

- `m`-step: unchanged, *i.e.* $\theta^{(t+1)} = \operatorname{argmax}_\theta \mathbb{E}_{\mathbb{Q}^\star}(\log \mathbb{P}(\mathbf{A},\mathbf{Z};\boldsymbol{\theta}))$
- A consequence of $\mathcal{KL} \geq 0$ is the lower bound

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

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

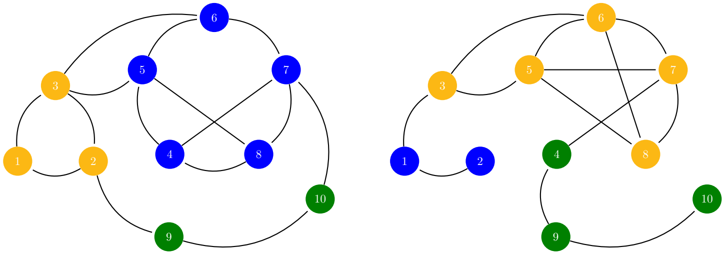
The stochastic block model

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

Follow the groups through time

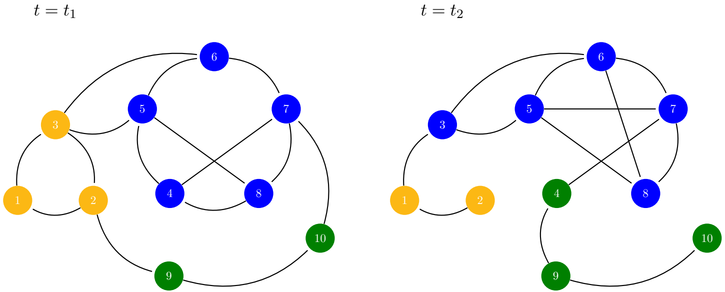
Label switching issue in the dynamic context

$t = t_1$ $t = t_2$



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.

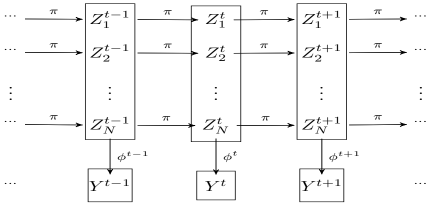
Inference

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



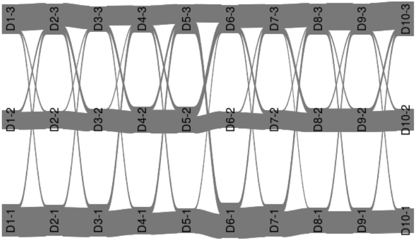
Goal

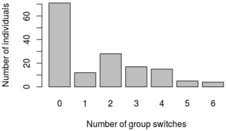
Infer the parameter $\theta = (\boldsymbol{\pi}, \boldsymbol{\beta}, \boldsymbol{\gamma})$, recover the clusters $\{Z_i^t\}_{i,t}$ and follow their evolution through time.

Application on ecological networks [Miele & M.(2017)] I

Ants dataset[Mersch et al.(2013)]

T=10, N=152



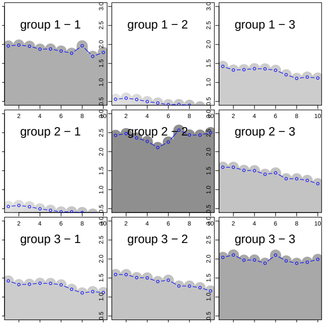


Selection of 3 social groups.

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

No group switches between groups 1 and 2.

Application on ecological networks [Miele & M.(2017)] II



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*

	nurses	foragers	cleaners
1	42	0	0
2	0	29	2
3	4	1	29

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

Outline

Dynamic Random Graphs: the data

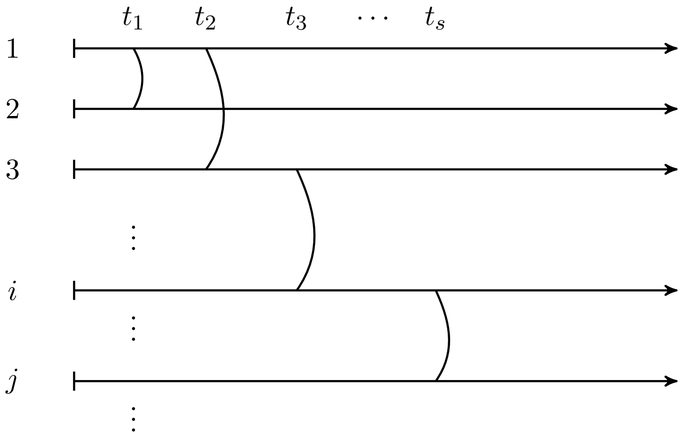
Graphs clustering: different approaches

The stochastic block model

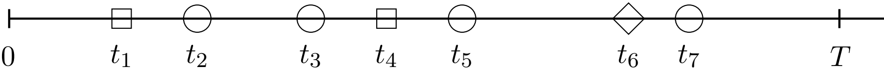
Clustering dynamic networks

- Clustering graphs sequences
- Clustering links streams (with no duration)

Longitudinal interaction networks = Stream links view



Longitudinal interaction networks = point process view



□ interactions between individuals i, j

○ interactions between individuals i, k

◇ interactions between individuals k, l

- 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

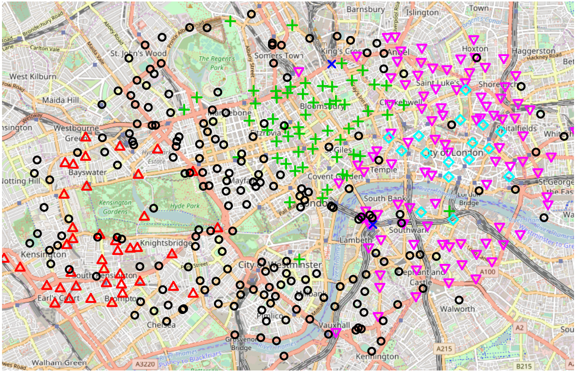
Data

- 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)
- 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.

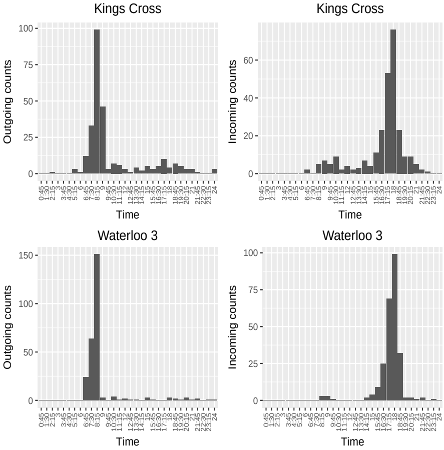
London Santander cycles: geographical projection of the clusters



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 $(N_{i\cdot}(\cdot))$ 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

E.M. Airoldi, D.M. Blei, S.E. Fienberg, and E.P. Xing.
Mixed-membership stochastic blockmodels.
Journal of Machine Learning Research, 9:1981–2014, 2008.

C. Ambroise and C. Matias.
New consistent and asymptotically normal parameter estimates for random graph mixture models.
JRSSB, 74(1):3–35, 2012.

P. Bickel, A. Chen and E. Levina
The method of moments and degree distributions for network models
Ann. Statist., 39(5):2280—2301, 2011.

A. Channarond, J.-J. Daudin, and S. Robin.
Classification and estimation in the Stochastic Blockmodel based on the empirical degrees.
Electron. J. Statist., 6:2574—2601, 2012.

References II

E. Côme and P. Latouche.
Model selection and clustering in stochastic block models based on the exact integrated complete data likelihood.
Statistical Modelling, 2015.

J.-J. Daudin, F. Picard, and S. Robin.
A mixture model for random graphs.
Stat. Comput., 18(2):173–183, 2008.

N. Gaumont.
Groupes et communautés dans les flots de liens : des données aux algorithmes.
PhD thesis, Université Pierre et Marie Curie, 2016.

G. Govaert and M. Nadif.
Clustering with block mixture models.
Pattern Recognition, 36(2):463–473, 2003.

References III

Hoff, P., A. Raftery, and M. Handcock (2002).
Latent space approaches to social network analysis.
J. Amer. Statist. Assoc. 97(460), 1090–98.

P. Holme.
Modern temporal network theory: a colloquium.
Eur. Phys. J. B, 88(9):234, 2015.

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

P. Latouche, E. Birmelé, and C. Ambroise.
Overlapping stochastic block models with application to the French political blogosphere.
Ann. Appl. Stat., 5(1):309–336, 2011.

References IV

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.
To appear in Biometrika, 2018.

D. P. Mersch, A. Crespi, and L. Keller.
Tracking individuals shows spatial fidelity is a key regulator of ant social organization.
Science, 340(6136):1090–1093, 2013.

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
The Annals of Applied Statistics 7(2), 1010–1039, (2013).

J. Wyse and N. Friel.
Block clustering with collapsed latent block models.
Statistics and Computing, 22(2):415–428, 2012.