The stochastic blockmodel for clustering nodes in graphs and hypergraphs

Catherine Matias (Sorbonne Université, Université Paris Cité, CNRS)

A popular way of extracting information from heterogeneous data is clustering. In the graphs context, stochastic blockmodels (SBMs) were introduced in the early eighties and have flourished in many directions. These models assume that nodes are clustered into groups and the connection probabilities between nodes are driven by their groups memberships. Variants handling weighted graphs and degree corrected versions have been developed among others. SBMs are widely used for clustering nodes in graphs and may produce more general clusters than community detection methods.

In this talk, I will present this class of models, an inference method based on variational Expectation-Maximization algorithm and recent extensions to handle dynamic data as well as higher-order interactions.

The stochastic blockmodel for clustering interacting entities

Catherine Matias

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

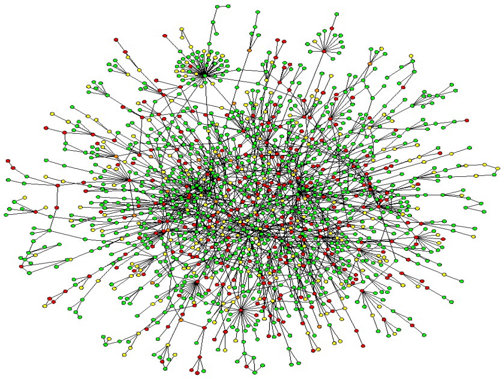
statistical learning on LARge scale GRaphs (LARGR)
March 2023

Outline

- Graphs for modeling networks
- Graphs clustering: different approaches
- The stochastic block model
  - Model definition
  - SBM vs community detection
  - Inference in SBMs
- Dynamic Random Graphs
  - Dynamic networks data
  - Clustering graphs sequences
  - Clustering links streams (with no duration)
- Higher order interactions
  - Hypergraphs: what it is and why you may need it
- Conclusions

Data: networks, their properties and beyond I

Some networks characteristics



- Potentially large number n of interacting entities,
- Potentially sparse networks: number of edges $\ll O(n^2)$,
- Scale-free property : Degree distribution has a power law $\mathbb{P}(D_i = k) = ck^{-\gamma}, (\gamma > 0)$,
- Small world property: shortest path length is small on average (less than 6),
- Transitivity/clustering property: is there a large amount of triangles?
- ...

Data: networks, their properties and beyond II

Some challenges

- Go beyond these (local) descriptors and capture higher-level structures, such as topological patterns, cliques, nodes groups, etc,
- Propose relevant models that will capture those structures without any a priori information on which structures we are looking for,
- From static to dynamic models,
- From pairwise to higher-order interactions,
- ...

Beware: Issues with sampling

The graph at stake is a sample (or to be sampled) from a larger, not observed graph.

- Does the sampled graph have the same characteristics than the larger unobserved one?
- How should we sample from the larger unobserved graph to ensure good properties on the sample?

These are difficult questions on which very few is known.

Outline

- Graphs for modeling networks
- Graphs clustering: different approaches
- The stochastic block model
  - Model definition
  - SBM vs community detection
  - Inference in SBMs
- Dynamic Random Graphs
  - Dynamic networks data
  - Clustering graphs sequences
  - Clustering links streams (with no duration)
- Higher order interactions
  - Hypergraphs: what it is and why you may need it
- Conclusions

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 inter-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 (or modules),
  - Modularity-based approaches;
  - Random walk algorithms;
  - Spectral clustering;
  - 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.

Model-based approaches for clustering

Let $A = (A_{ij})_{1 \leq i,j \leq n}$ denote the adjacency matrix of the graph.

Common principle of latent variable models

- for each node i there exists some latent random variable Z_i that drives the nodes interactions,
- More precisely, given the Z_i's, the random variables A_{ij}'s are independent,
- and the conditional distribution of A_{ij} depends only on Z_i, Z_j.

These approaches include

- Hoff *et al.*'s model
- the Stochastic block model.

Outline

Graphs for modeling networks

Graphs clustering: different approaches

The stochastic block model

- Model definition
- SBM vs community detection
- Inference in SBMs

Dynamic Random Graphs

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

Higher order interactions

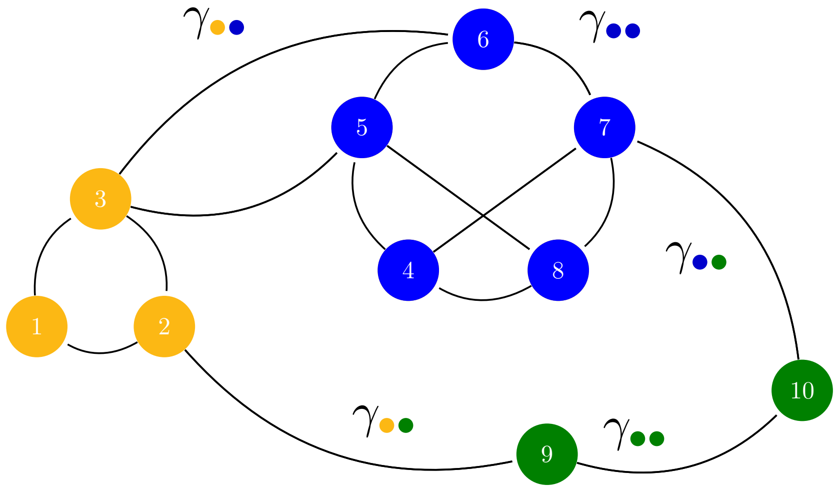
- Hypergraphs: what it is and why you may need it

Conclusions

Outline

- Graphs for modeling networks
- Graphs clustering: different approaches
- The stochastic block model
  - Model definition
  - SBM vs community detection
  - Inference in SBMs
- Dynamic Random Graphs
  - Dynamic networks data
  - Clustering graphs sequences
  - Clustering links streams (with no duration)
- Higher order interactions
  - Hypergraphs: what it is and why you may need it
- Conclusions

Stochastic block model (binary graphs)

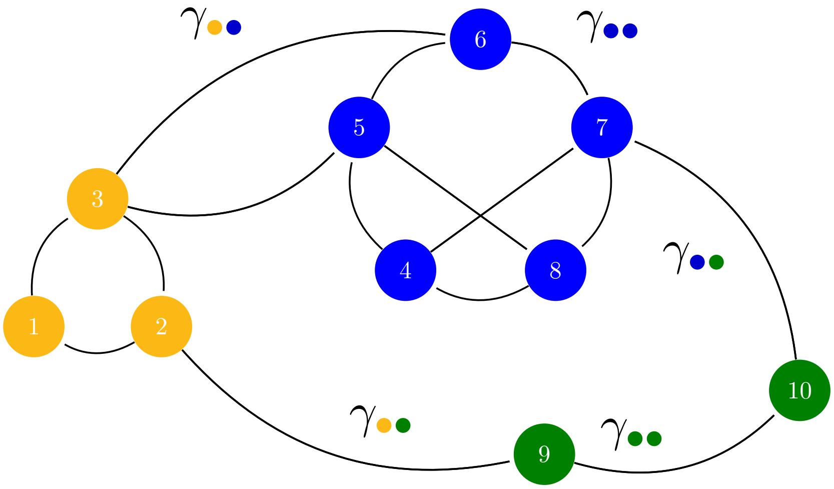


$n = 10, Z_{5●} = 1$
$A_{12} = 1, A_{15} = 0$

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 $\mathcal{B}(\gamma_{Z_iZ_j})$.

Stochastic block model (weighted graphs)


$\gamma_{\bullet\bullet}$
$\gamma_{\bullet\bullet}$
$\gamma_{\bullet\bullet}$
$\gamma_{\bullet\bullet}$
$\gamma_{\bullet\bullet}$
1
2
3
4
5
6
7
8
9
10

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

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

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

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

Outline

Graphs for modeling networks

Graphs clustering: different approaches

The stochastic block model

- Model definition
- SBM vs community detection
- Inference in SBMs

Dynamic Random Graphs

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

Higher order interactions

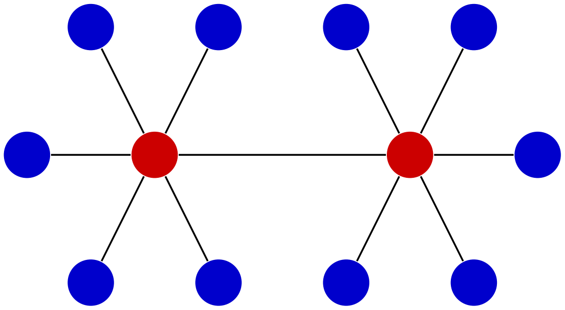
- Hypergraphs: what it is and why you may need it

Conclusions

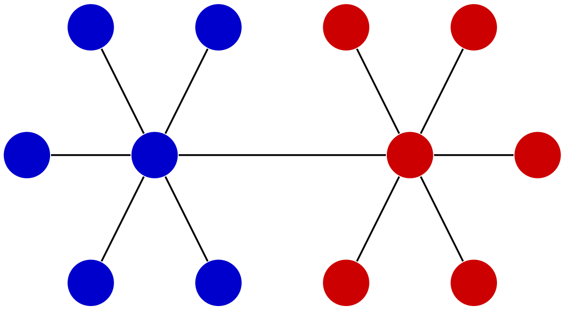
SBM classification vs community detection

Toy example

- Nodes classification induced by the model reflects a common connectivity behaviour;
- Most clustering methods try to group nodes that belong to the same module/community (called community detection)
- Toy example



SBM clusters



Community detection or SBM

SBM : particular cases and generalisations I

Particular case: Affiliation model

$$\boldsymbol{\gamma} = \begin{pmatrix} \alpha & \dots & \beta \\ \vdots & \ddots & \vdots \\ \beta & \dots & \alpha \end{pmatrix}$$

- When $\alpha \gg \beta \implies$ community detection
  - Called planted partition when groups have same size
- When $\alpha \ll \beta \implies$ multi-partite structures (heterophily)

SBM : particular cases and generalisations II

Some generalisations

- Graph setting:
  - Overlapping groups [Latouche et al.(2011), Airoldi et al.(2008)] for binary graphs;
  - SBM with covariates [Zanghi et al.(2010)];
  - Degree-corrected SBM [Karrer and Newman(2011)];
  - Missing edges [Barbillon et al.(2022)] ;
  - Latent block models (LBM), for array data or bipartite graphs [Govaert and Nadif(2003)] and more general multi-partite models [Bar-Hen et al.(2020)]
  - see `GroßBM` `https://github.com/GrossSBM/` that implements inference for many and more variants of SBM;
  - Nonparametric SBM (graphon);
- Dynamic SBMs [M. & Miele(2017), M. et al.(2018)];
- Hypergraph SBM [Brusa and M.(2022)];

Outline

Graphs for modeling networks

Graphs clustering: different approaches

The stochastic block model

- Model definition
- SBM vs community detection
- Inference in SBMs

Dynamic Random Graphs

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

Higher order interactions

- Hypergraphs: what it is and why you may need it

Conclusions

Overview of algorithms

Goal is MLE. Likelihood computation is untractable unless n is 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 [Biernacki et al.(2000), 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)]

Outline

Graphs for modeling networks

Graphs clustering: different approaches

The stochastic block model
- Model definition
- SBM vs community detection
- Inference in SBMs

Dynamic Random Graphs
- Dynamic networks data
- Clustering graphs sequences
- Clustering links streams (with no duration)

Higher order interactions
- Hypergraphs: what it is and why you may need it

Conclusions

Outline

- Graphs for modeling networks
- Graphs clustering: different approaches
- The stochastic block model
  - Model definition
  - SBM vs community detection
  - Inference in SBMs
- Dynamic Random Graphs
  - Dynamic networks data
  - Clustering graphs sequences
  - Clustering links streams (with no duration)
- Higher order interactions
  - Hypergraphs: what it is and why you may need it
- Conclusions

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

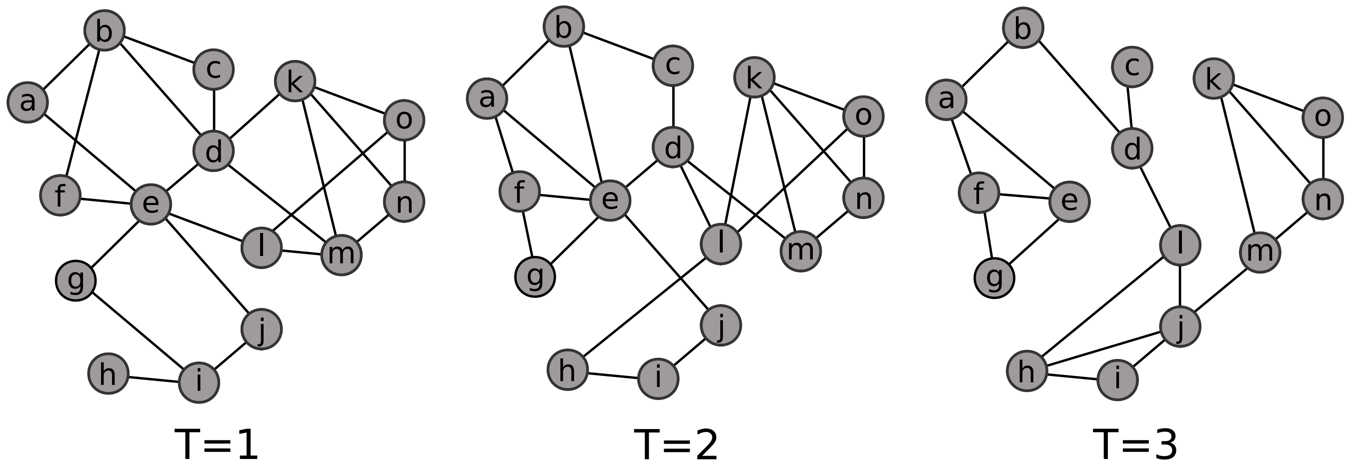


FIGURE 1.3 – Exemple de série de graphes sur trois intervalles de temps.

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

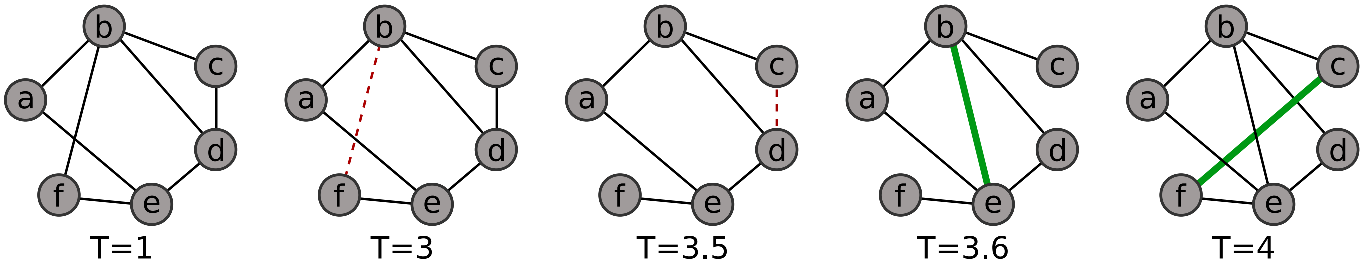
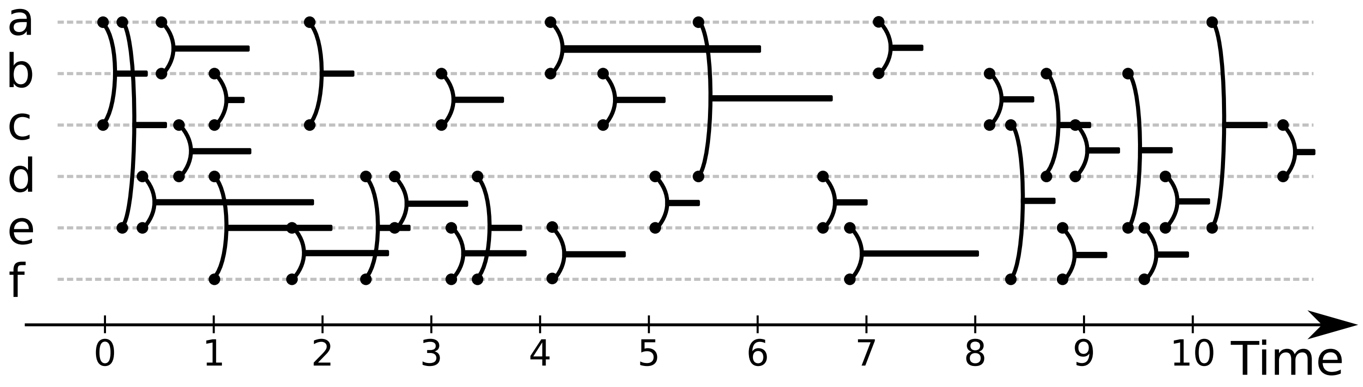


FIGURE 1.5 – Graphe temporel avec des ajouts de lien représentés en traits épais verts et des suppressions de lien représentées par des liens pointillés rouges.

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.(2018)]



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

- Graphs for modeling networks
- Graphs clustering: different approaches
- The stochastic block model
  - Model definition
  - SBM vs community detection
  - Inference in SBMs
- Dynamic Random Graphs
  - Dynamic networks data
  - Clustering graphs sequences
  - Clustering links streams (with no duration)
- Higher order interactions
  - Hypergraphs: what it is and why you may need it
- Conclusions

`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.

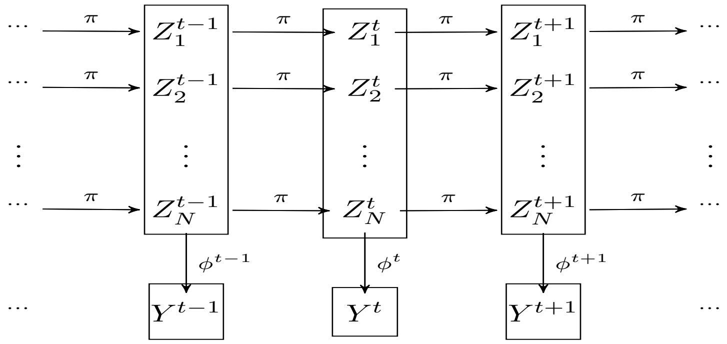
Inference

- `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 $\{1,\dots,Q\}$ with transition $\boldsymbol{\pi} = (\pi_{qq'})_{1\leq q,q'\leq Q}$ and initial stationary distribution $\boldsymbol{\alpha} = (\alpha_1,\dots,\alpha_Q)$.



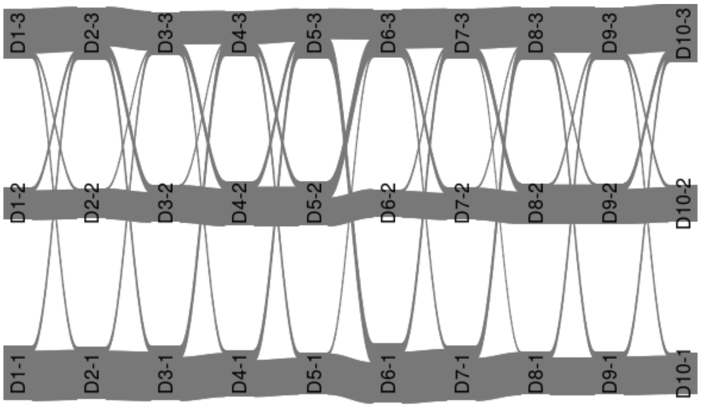
Goal

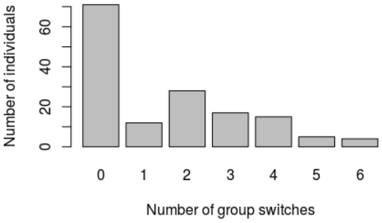
Infer the parameter $\theta = (\boldsymbol{\pi},\boldsymbol{\beta},\boldsymbol{\gamma})$, recover the clusters $\{Z_i^t\}_{i,t}$ and follow their evolution through time.

Ecological networks [Miele & M.(2017)] I

Ants dataset[Mersch et al.(2013)]

T=10, N=152



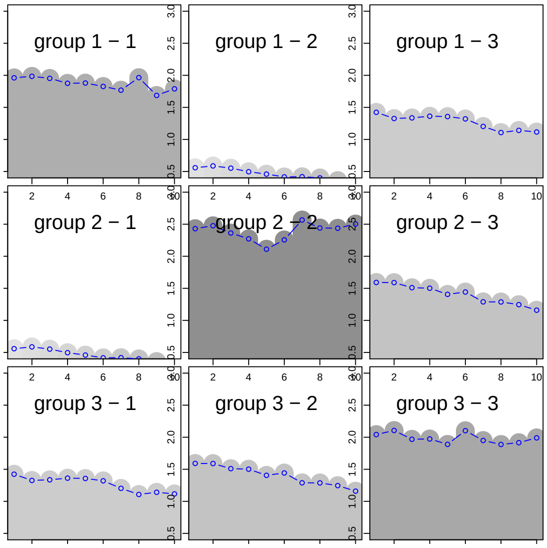


Selection of 3 social groups.

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

No group switches between groups 1 and 2.

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*

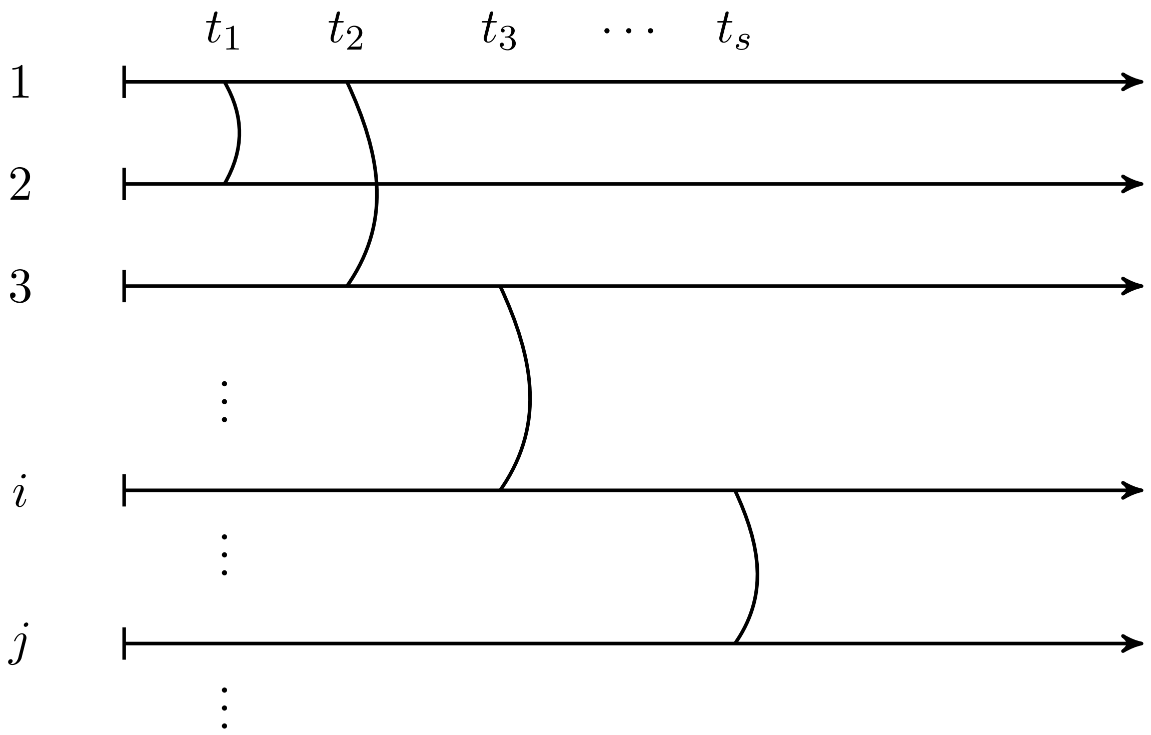
	nurses	foragers	cleaners
1	42	0	0
2	0	29	2
3	4	1	29

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

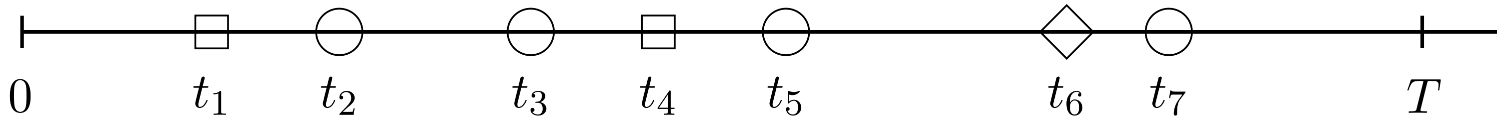
Outline

- Graphs for modeling networks
- Graphs clustering: different approaches
- The stochastic block model
  - Model definition
  - SBM vs community detection
  - Inference in SBMs
- Dynamic Random Graphs
  - Dynamic networks data
  - Clustering graphs sequences
  - Clustering links streams (with no duration)
- Higher order interactions
  - Hypergraphs: what it is and why you may need it
- Conclusions

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. et al.(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 $\mathcal{Z} = (Z_1, \dots, Z_n)$ and estimate the intensities per groups pairs $\{\alpha^{(q,l)}(\cdot)\}_{1 \le q < l \le Q}$ 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

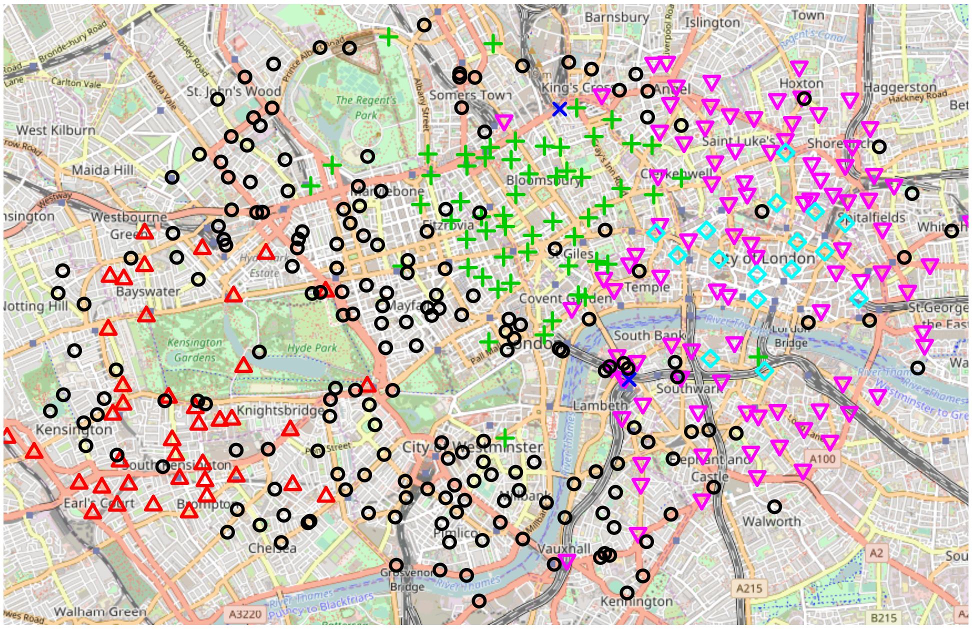
Data

- 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.

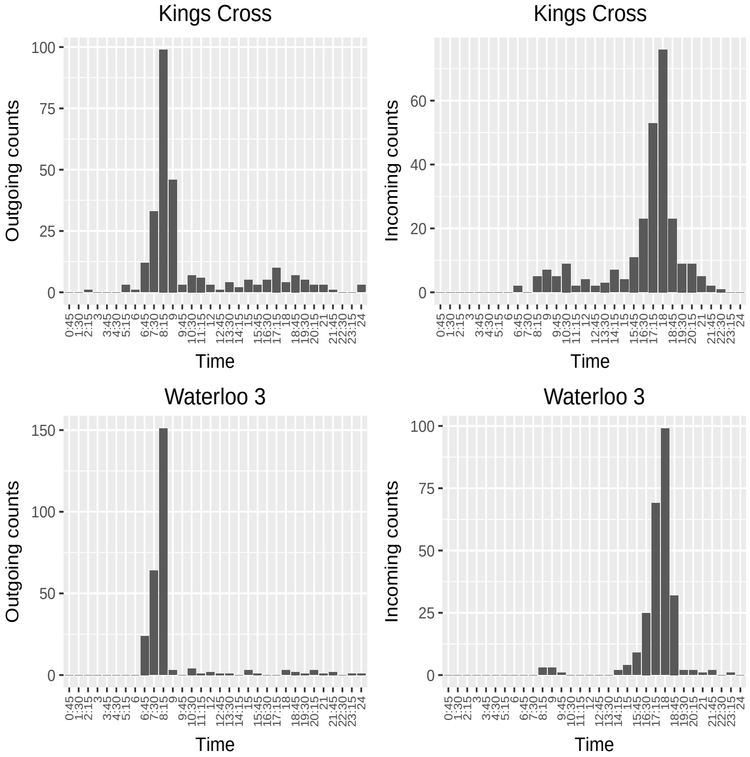
London Santander cycles: geographical projection of the clusters



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 ◇ that corresponds to business city center.

Outline

Graphs for modeling networks

Graphs clustering: different approaches

The stochastic block model

- Model definition
- SBM vs community detection
- Inference in SBMs

Dynamic Random Graphs

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

Higher order interactions

- Hypergraphs: what it is and why you may need it

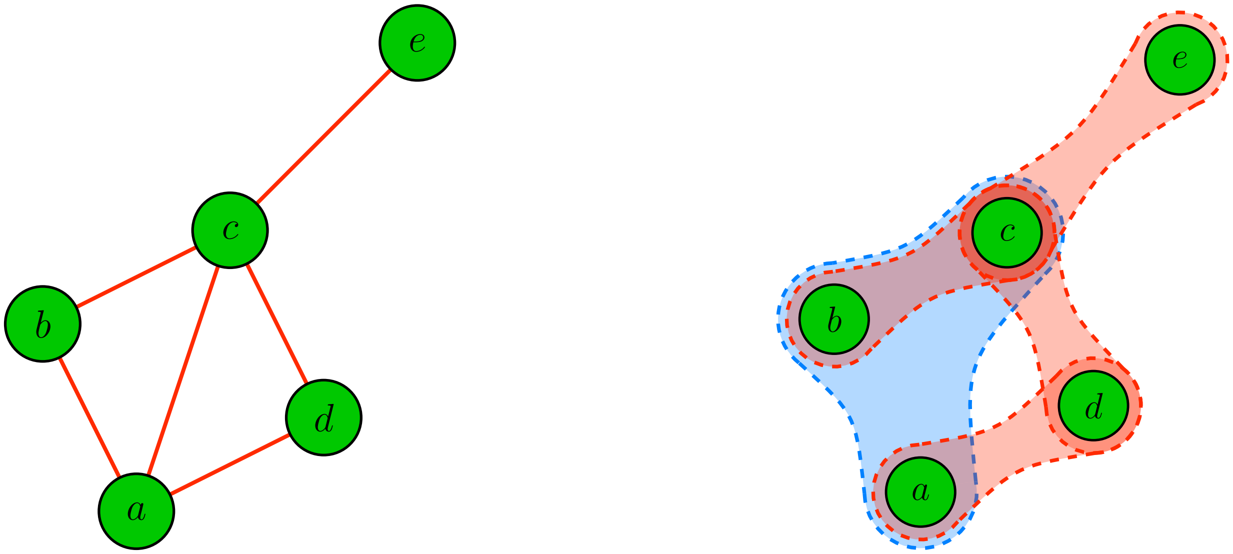
Conclusions

Outline

- Graphs for modeling networks
- Graphs clustering: different approaches
- The stochastic block model
  - Model definition
  - SBM vs community detection
  - Inference in SBMs
- Dynamic Random Graphs
  - Dynamic networks data
  - Clustering graphs sequences
  - Clustering links streams (with no duration)
- Higher order interactions
  - Hypergraphs: what it is and why you may need it
- Conclusions

Graphs vs Hypergraphs

Hypergraphs represent higher-order interactions, e.g. $\{[a, b, c], [a, d], [c, d], [c, e]\}$



(a) Graph representation

(b) Hypergraph representation

The need for modelling higher-order interactions

Why higher-order interactions?

- Social networks: triadic and larger groups (as early as Simmel, 1950)
- Scientific co-authorship,
- Interactions between more than two species in ecological systems,
- Higher-order interactions between neurons in brain networks,
- Metabolites in chemical reactions,
- etc

These interactions **CAN NOT** be represented by a graph.

Simple hypergraphs

Definition

A (simple) hypergraph $\mathcal{H} = (\mathcal{V}, \mathcal{E})$ is defined as a set of nodes $\mathcal{V} \neq \emptyset$ and a set of hyperedges $\mathcal{E}$. Each hyperedge is a non-empty collection of m distinct nodes $(2 \leq m \leq M)$ taking part within an interaction.

- Hypergraphs naturally include the entity of graphs, by simply considering hyperedges of size $m = 2$;
- A hypergraph can contain a size-3 hyperedge $[a, b, c]$ without any requirement on the existence of the size-2 hyperedges $[a, b]$, $[a, c]$, and $[b, c]$.

HyperSBM formulation

- $\mathcal{H} = (\mathcal{V}, \mathcal{E})$, with $\mathcal{V} = \{1, \dots, n\}$ nodes and $\mathcal{E}$ hyperedges;
- For each $2 \leq m \leq M$, let $\mathcal{V}^{(m)} = \big\{\{i_1, \dots, i_m\} : i_1, \dots, i_m \in \mathcal{V} \text{ and } i_1 \neq \dots \neq i_m\big\}$, set of unordered node tuples of size m;
- Observations: At each $\{i_1, \dots, i_m\} \in \mathcal{V}^{(m)}$, we observe indicator variable $Y_{i_1, \dots, i_m} = 1\{\{i_1, \dots, i_m\} \in \mathcal{E}\}$;
- Latent clusters: $Z_1, \dots, Z_n$ iid in $\{1, \dots, Q\}$ with $\pi_q = \mathbb{P}(Z_i = q)$;
- Conditional independence assumption: $\{Y_{i_1, \dots, i_m}\}_{\{i_1, \dots, i_m\} \in \mathcal{V}^{(m)}} | \{Z_1, \dots, Z_n\}$ are independent with $Y_{i_1, \dots, i_m} | \{Z_1 = q_1, \dots, Z_m = q_m\} \sim \text{Bern}(B^{(m)}_{q_{i_1}, \dots, q_{i_m}})$.

Computational complexity - and considerations over the choice of M

- Focusing on *single* hypergraphs has a high price: we need to explore all the $\binom{n}{m}$ tuples of nodes for all $2 \leq m \leq M$;
- Our algorithm has a complexity of $O(n\binom{n}{M}Q^M)$, which is large;
- Current modularity approaches avoid this issue by working with multisets-hypergraphs, because there the summations over multisets of nodes $\sum_{i_1,\ldots,i_m}$ factorize into m independent sums (no constraint that the nodes be different), and this further simplifies the expression of the modularity;
- Again, this is inappropriate on some datasets;
- As a consequence: we recommend to use a reasonable value of M: indeed M is not necessarily the largest observed hyperedge size (e.g. co-authorship dataset);

Co-authorship dataset I

Dataset description

- Available at `http://vlado.fmf.uni-lj.si/pub/networks/data/2mode/Sandi/Sandi.htm`
- Bipartite author/article graph transformed into hypergraph of authors where hyperedges link the authors of a same paper;
- We choose $M = 4$ and consider the induced largest connected component: 79 authors and 76 hyperedges (68.5% of which have size 2, while 29% have size 3 and 2.5% have size 4).

Co-authorship dataset II

Analysis through HyperSBM

- ICL selects $Q = 2$ groups, the first has only 8 authors;
- Our first group is made of authors (among) the most collaborative ones, which are also (among) the most prolific ones.
- None of these groups is a community (the first co-publishes with all, the second has low intra-group connectivity).

Comparison with hypergraph spectral clustering (HSC)

- HSC with $Q = 2$ gives a group of size 24 and one of size 55
- These groups are neither characterized by the number of co-authors nor their degrees in the bipartite graph
- Very different from our results because: spectral clustering tends to: i) extract communities ; ii) favor groups of similar size.

Outline

Graphs for modeling networks

Graphs clustering: different approaches

The stochastic block model

- Model definition
- SBM vs community detection
- Inference in SBMs

Dynamic Random Graphs

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

Higher order interactions

- Hypergraphs: what it is and why you may need it

Conclusions

Conclusions

- Stochastic Blockmodels are powerful tools for clustering entities in interaction
- Parameter estimation and nodes clustering may be performed through VEM algorithm
- ICL criterion is used to select the number of groups
- Try the different softwares !

Any questions ?

References I

- E.M. Airoldi, D.M. Blei, S.E. Fienberg, and E.P. Xing.
  Mixed-membership stochastic blockmodels.
  J Mach Learn Res, 9:1981–2014, 2008.

- C. Ambroise and C. Matias.
  New consistent and asymptotically normal parameter estimates for random graph mixture models.
  J Roy Statist Soc B, 74(1):3–35, 2012.

- P. Barbillon, J. Chiquet, and T. Tabouy
  missSBM: An R Package for Handling Missing Values in the Stochastic Block Model.
  J of Statist Software, 101(1):1–32, 2022.

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

References II

C. Biernacki, G. Celeux and G. Govaert
Assessing a Mixture Model for Clustering with the Integrated Completed Likelihood.
IEEE Trans. Pattern Anal. Machine Intel., 22(7):719–725, 2000.

L. Brusa and C. Matias.
Model-based clustering in simple hypergraphs through a stochastic blockmodel.
hal-03811678, 2022.

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 III

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

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

A. Bar-Hen, P. Barbillon and S. Donnet
Block models for multipartite networks. Applications in ecology and ethnobiology.
Statistical Modelling, 22(4):273–296, 2020.

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

References IV

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

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.

B. Karrer and M. E. J. Newman.
Stochastic blockmodels and community structure in networks.
Phys. Rev. E, 83:016107, 2011.

References V

- Matthieu Latapy, Tiphaine Viard, Clémence Magnien
  Stream Graphs and Link Streams for the Modeling of Interactions over Time
  Soc Net Anal mining, 8: 61, 2018.

- 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.

- C. Matias and V. Miele.
  Statistical clustering of temporal networks through a dynamic stochastic block model.
  J Roy Statist Soc B, 79(4), 1119–1141, 2017

References VI

C. Matias, T. Rebafka, and F. Villers.
A semiparametric extension of the stochastic block model for longitudinal networks.
Biometrika, 105(3): 665-680, 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.

V. Miele and C. Matias.
Revealing the hidden structure of dynamic ecological networks.
Roy Soc Open Sc, 4(6), 170251, 2017

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

References VII

H. Zanghi, S. Volant, and C. Ambroise.
Clustering based on random graph model embedding vertex features.
Pattern Recognition Letters, 31:830–836, 2010.