Statistical clustering of temporal networks through a dynamic stochastic block model

Catherine Matias and Vincent Miele

CNRS - Université Pierre et Marie Curie, Paris
catherine.matias@math.cnrs.fr
http://cmatias.perso.math.cnrs.fr/

Rencontres de Statistique Avignon/Marseille
Juin 2016

Outline

Introduction and model

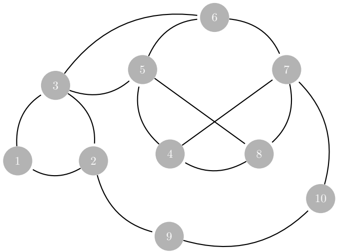
Inference

Simulations

Real data set

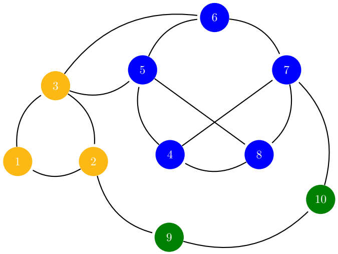
Clustering dynamic networks I

$t = t_1$

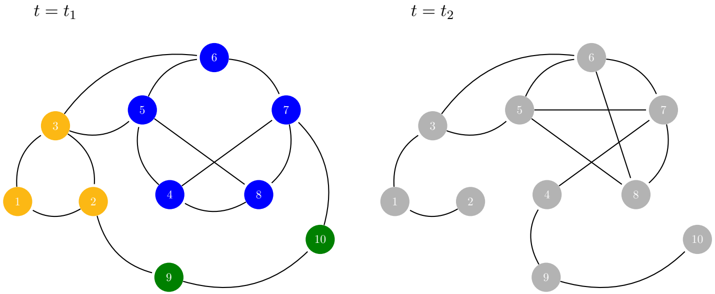


Clustering dynamic networks I

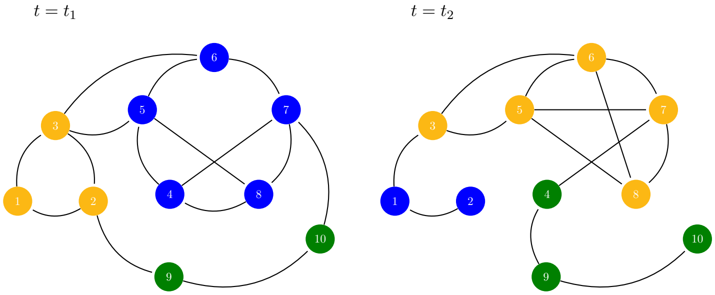
$t = t_1$



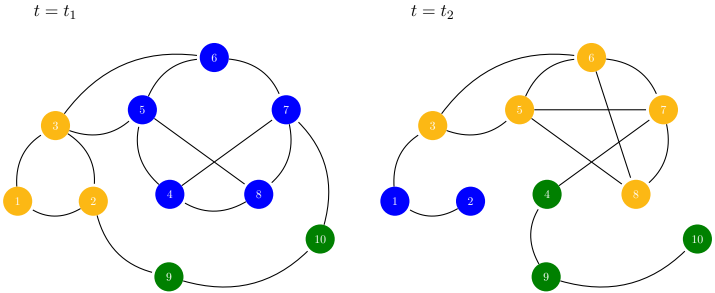
Clustering dynamic networks I



Clustering dynamic networks I



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.

Clustering dynamic networks II

Discrete time networks

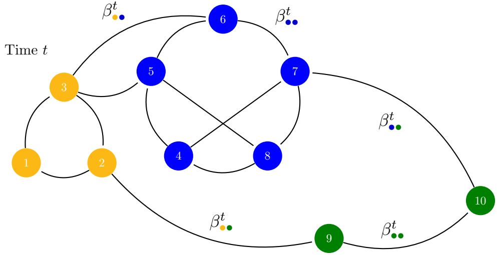
- Observe $Y^1, \dots, Y^T$ adjacency matrices (graphs snapshots),
- $\forall t, Y^t = (Y^t_{ij})_{1 \leq i,j \leq N_t}$ may contain either binary, discrete or continuous values (contact information)
- Individuals may be present/absent at each time step t.

Nodes clustering

- Clusters model heterogeneity in nodes interactions,
- They summarize information through a finite number of behaviors.
- Many different approaches: spectral algorithms, community detection (e.g. based on modularity criterion), model-based clustering (e.g. latent space models, SBM)

Here, we choose to focus on the Stochastic block model (SBM) for undirected graphs, with no self-loops.

Static part modeling: SBM - binary case

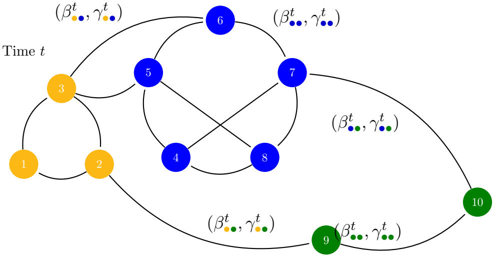


$n = 10, Q = 3,$
$Z_5^t = \bullet,$
$Y_{12}^t = 1, Y_{15}^t = 0$

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

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

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 $(\boldsymbol{\beta}^t, \boldsymbol{\gamma}^t) = (\beta_{ql}^t, \gamma_{ql}^t)_{1 \leq q \leq l \leq Q}$

- 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

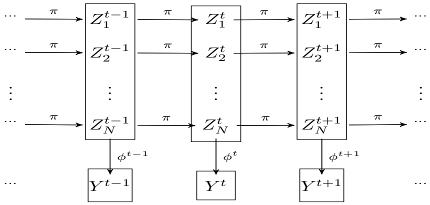
$$\phi(\cdot; \beta_{Z_i^t Z_j^t}^t, \gamma_{Z_i^t Z_j^t}^t) := (1 - \beta_{Z_i^t Z_j^t}^t)\delta_0(\cdot) + \beta_{Z_i^t Z_j^t}^t f(\cdot, \gamma_{Z_i^t Z_j^t}^t),$$

(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.

Other very close works

[Yang *et al., 2011*] and [Xu and Hero, 2014] propose very close models (in the binary setup).
Main differences with our work

- We allow for both groups and parameters to vary with time and discuss valid assumptions for parameters' identifiability;
- We model binary as well as weighted graphs;
- We propose a model selection criterion for the number of clusters;
- We discuss a proper clustering index for measuring the classification performances taking into account label switching across time.

Identifiability: the problem

If both $(\beta^t, \gamma^t)_t$ and $(Z^t)_t$ can change, the parameters are not identifiable.

Toy example with 3 groups {hub, community, periphery}

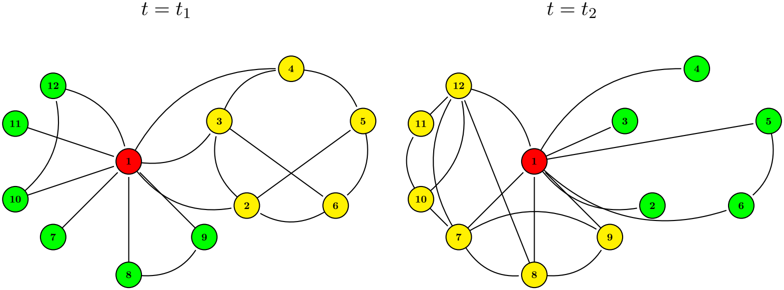
First scenario:

- $\forall 2 \leq i \leq 6, Z_i^1 =$ community$, Z_i^2 =$ periphery
- $\forall 7 \leq i \leq 12, Z_i^1 =$ periphery$, Z_i^2 =$ community
- $(\beta^{t_1}, \gamma^{t_1}) = (\beta^{t_2}, \gamma^{t_2})$

Identifiability: the problem

If both $(\beta^t, \gamma^t)_t$ and $(Z^t)_t$ can change, the parameters are not identifiable.

Toy example with 3 groups {hub, community, periphery}



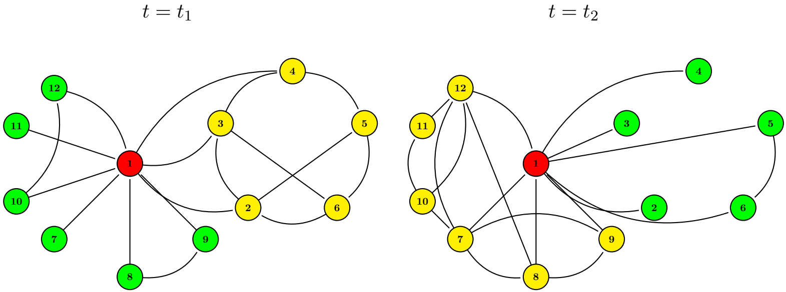
First scenario:

- $\forall 2 \leq i \leq 6, Z_i^1 =$ community, $Z_i^2 =$ periphery
- $\forall 7 \leq i \leq 12, Z_i^1 =$ periphery, $Z_i^2 =$ community
- $(\beta^{t_1}, \gamma^{t_1}) = (\beta^{t_2}, \gamma^{t_2})$

Identifiability: the problem

If both $(\beta^t, \gamma^t)_t$ and $(Z^t)_t$ can change, the parameters are not identifiable.

Toy example with 3 groups {hub, community, periphery}



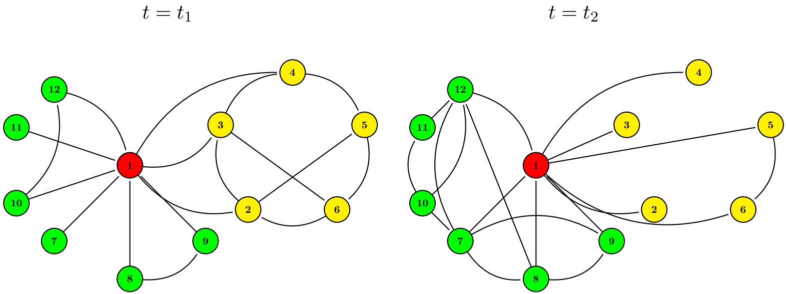
First scenario:

- $\forall 2 \leq i \leq 6, Z_i^1 =$ community, $Z_i^2 =$ periphery
- $\forall 7 \leq i \leq 12, Z_i^1 =$ periphery, $Z_i^2 =$ community
- $(\beta^{t_1}, \gamma^{t_1}) = (\beta^{t_2}, \gamma^{t_2})$

Identifiability: the problem

If both $(\beta^t, \gamma^t)_t$ and $(Z^t)_t$ can change, the parameters are not identifiable.

Toy example with 3 groups {hub, community, periphery}



Second scenario:

- $\forall i, Z_i^1 = Z_i^2$
- For $t = t_1$, $(\beta^{t_1}_{\color{green}\bullet,\bullet}, \gamma^{t_1}_{\color{green}\bullet,\bullet})$= periphery and $(\beta^{t_1}_{\color{olive}\bullet,\bullet}, \gamma^{t_1}_{\color{olive}\bullet,\bullet})$= community ,
- For $t = t_2$, $(\beta^{t_2}_{\color{green}\bullet,\bullet}, \gamma^{t_2}_{\color{green}\bullet,\bullet})$= community and $(\beta^{t_2}_{\color{olive}\bullet,\bullet}, \gamma^{t_2}_{\color{olive}\bullet,\bullet})$= periphery.

Identifiability

If both $(\beta^t, \gamma^t)_t$ and $(Z^t)_t$ can change, the parameters are not identifiable.

Main Assumption: Fixed diagonal connectivity parameters

$\forall q \in \mathcal{Q}, \forall t, t'$, we assume that

$$\begin{cases} \text{Binary case:} & \beta^t_{qq} = \beta^{t'}_{qq}, \\ \text{Weighted case:} & \gamma^t_{qq} = \gamma^{t'}_{qq}. \end{cases}$$

Results

- Under the above assumption (plus other classical assumptions), we prove identifiability (up to *a global* label switching) of the model's parameters.
- We underly that in the affiliation case, no current method can avoid label switching between time steps ! The parameters are not identifiable.

Identifiability

If both $(\beta^t, \gamma^t)_t$ and $(Z^t)_t$ can change, the parameters are not identifiable.

Main Assumption: Fixed diagonal connectivity parameters

$\forall q \in \mathcal{Q}, \forall t, t'$, we assume that

$$\begin{cases} \text{Binary case:} & \beta^t_{qq} = \beta^{t'}_{qq}, \\ \text{Weighted case:} & \gamma^t_{qq} = \gamma^{t'}_{qq}. \end{cases}$$

Results

- Under the above assumption (plus other classical assumptions), we prove identifiability (up to *a global* label switching) of the model's parameters.
- We underly that in the affiliation case, no current method can avoid label switching between time steps ! The parameters are not identifiable.

Outline

Variational Expectation Maximization (VEM) I

Complete data log-likelihood (here $Z_i^t = (Z_{i1}^t, \ldots, Z_{iQ}^t)$).

$$\log \mathbb{P}_\theta(\mathbf{Y}, \mathbf{Z}) = \sum_{i=1}^{N} \sum_{q=1}^{Q} Z_{iq}^1 \log \alpha_q + \sum_{t=2}^{T} \sum_{i=1}^{N} \sum_{1 \leq q, q' \leq Q} Z_{iq}^{t-1} Z_{iq'}^t \log \pi_{qq'}$$
$$+ \sum_{t=1}^{T} \sum_{1 \leq i < j \leq N} \sum_{1 \leq q, l \leq Q} Z_{iq}^t Z_{jl}^t \log \phi(Y_{ij}^t; \beta_{ql}^t, \gamma_{ql}^t).$$

- Conditional expectation of latent $\mathbf{Z}$, given observations $\mathbf{Y}$ may not be exactly computed,
- Use instead a variational approximation

$$\mathbb{Q}_\tau(\mathbf{Z}) = \prod_{i=1}^{N} \mathbb{Q}_\tau(Z_i) = \prod_{i=1}^{N} \mathbb{Q}_\tau(Z_i^1) \prod_{t=2}^{T} \mathbb{Q}_\tau(Z_i^t | Z_i^{t-1}).$$

Variational Expectation Maximization (VEM) II

Let

$$J(\theta, \tau) := \mathbb{E}_{\mathbb{Q}_\tau}(\log \mathbb{P}_\theta(\mathbf{Y}, \mathbf{Z})) + \mathcal{H}(\mathbb{Q}_\tau)$$

and note that

$$\log \mathbb{P}_\theta(\mathbf{Y}) = J(\theta, \tau) + \mathcal{KL}(\mathbb{Q}_\tau \| \mathbb{P}_\theta(\mathbf{Z}|\mathbf{Y})).$$

VEM principle

Iterate the following steps

- `VE`-step: Compute $\tau^{(k+1)} = \text{Argmax}_\tau J(\theta^{(k)}, \tau)$,
- `M`-step: Compute $\theta^{(k+1)} = \text{Argmax}_\theta J(\theta, \tau^{(k+1)})$.

More details can be found in the paper ...

Model selection

ICL criterion

$$ICL(Q) = \log \mathbb{P}_{\hat{\theta}_Q}(\mathbf{Y}, \hat{\mathbf{Z}}) - \frac{1}{2}Q(Q-1)\log(NT) - pen(N, T, \boldsymbol{\beta}, \boldsymbol{\gamma}),$$

- the second penalty $pen(N, T, \boldsymbol{\beta}, \boldsymbol{\gamma})$ depends on the distribution ϕ ; we give expressions for classical cases (Bernoulli, Poisson, Gaussian, ...)
- Groups parameters $\boldsymbol{\pi}$ and connectivity parameters $(\boldsymbol{\beta}, \boldsymbol{\gamma})$ are not penalized in the same way (count the number of observations corresponding to these parameters).

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

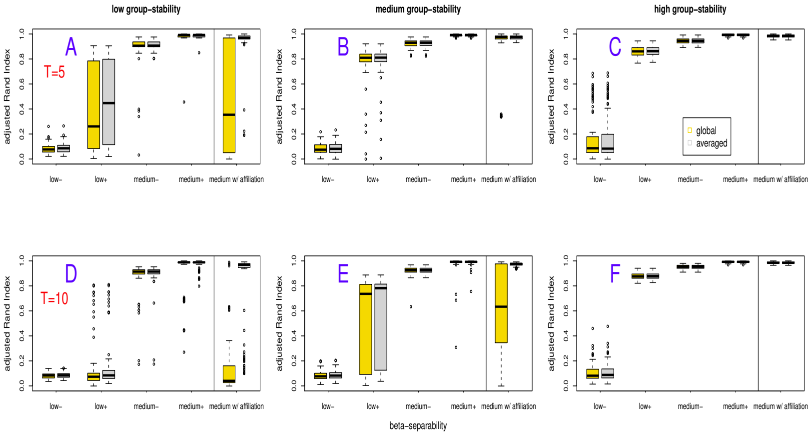
- Binary graphs, $N = 100$ nodes and $T \in \{5; 10\}$, 100 datasets,
- $Q = 2$ latent groups and $\boldsymbol{\pi} \in \{\boldsymbol{\pi}_{low}, \boldsymbol{\pi}_{med}, \boldsymbol{\pi}_{high}\}$

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

- Connectivity parameter $\boldsymbol{\beta}$

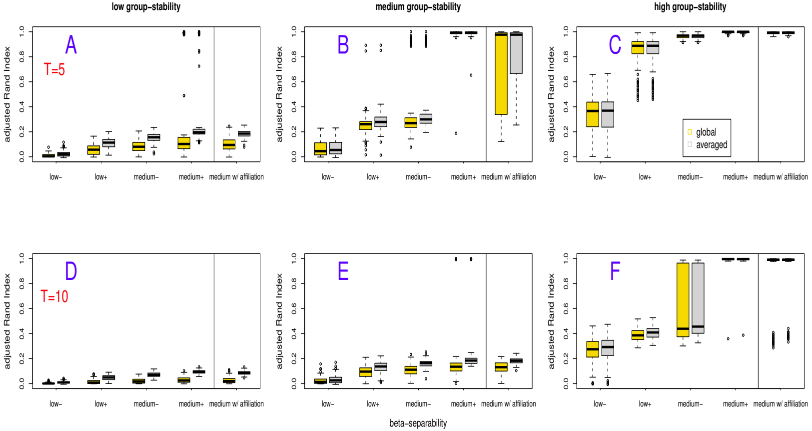
Difficulty	β_{11}	β_{12}	β_{22}
low-	0.2	0.1	0.15
low+	0.25	0.1	0.2
medium-	0.3	0.1	0.2
medium+	0.4	0.1	0.2
med w/ affiliation	0.3	0.1	0.3

Clustering performances III



Clustering performances IV

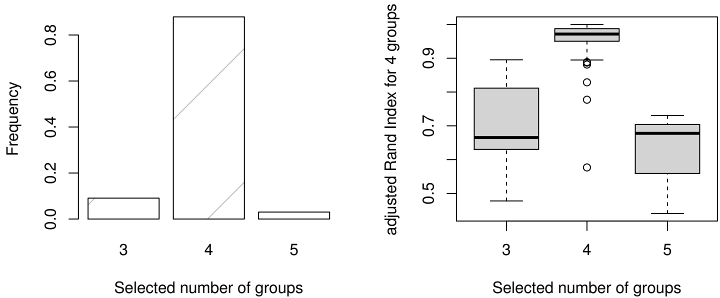
Yang *et al.*'s method with our initialization strategy



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 and model

Inference

Simulations

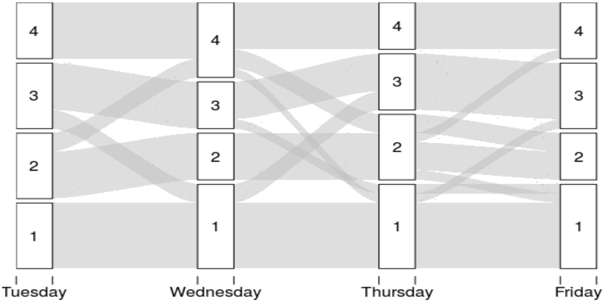
Real data set

Encounters between high school students I

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

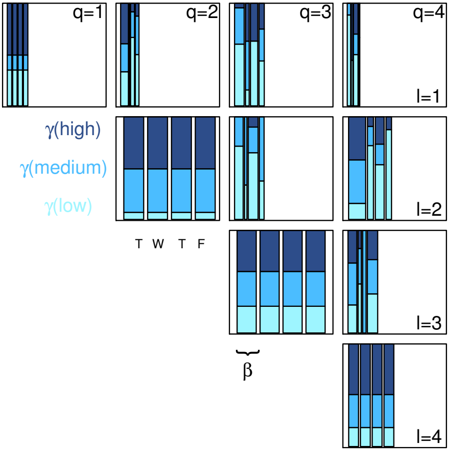
- Face-to-face encounters of high school students (wearable sensors), $T = 4$ days, $N = 27$ students,
- Discrete weight with 3 bins. Selection of $Q = 4$ groups.

Reconstructed dynamics



Encounters between high school students II

Estimated connectivity parameters



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 available at `http://lbbe.univ-lyon1.fr/dynsbm` and soon on the CRAN.
Preprint available at `http://arxiv.org/abs/1506.07464`

Thanks for your attention !

Extra short biblio

Xu, K. and A. Hero.
Dynamic stochastic blockmodels for time-evolving social networks.
Selected Topics in Signal Processing, IEEE Journal of 8(4), 552–562, 2014.

Yang, T., Y. Chi, S. Zhu, Y. Gong, and R. Jin.
Detecting communities and their evolutions in dynamic social networks—a Bayesian approach.
Machine Learning 82(2), 157–189, 2011.