Statistical Clustering of Temporal Networks through a Dynamic SBM

Written by Catherine Matias and Vincent Miele

Presenters: Rong, Trang, Simon, Konstantinos

Introduction

- We use a model that combines SBM for its static part with independent Markov chains for the evolution of the nodes groups through time
- Outline:
  - Model description
  - Parameters identifiability
  - Algorithm
  - Synthetic experiments
  - Application to a real dataset

Model description

- N vertices, Q groups, T time steps
- **Y**: adjacency matrix (TxNxN)
- **Z**: latent group membership that may vary through time (TxNxQ)
- For each vertex i, Z_i is an aperiodic stationary Markov chain with transition matrix π (QxQ) and initial stationary distribution α (1xQ)
- For fixed time t, random graph Y^t follows a stochastic block model. In other words, for each time t, conditional on Z^t, random variables Y_{ij}^t are independent and the distribution of each Y_{ij}^t only depends on Z_i^t, Z_j^t

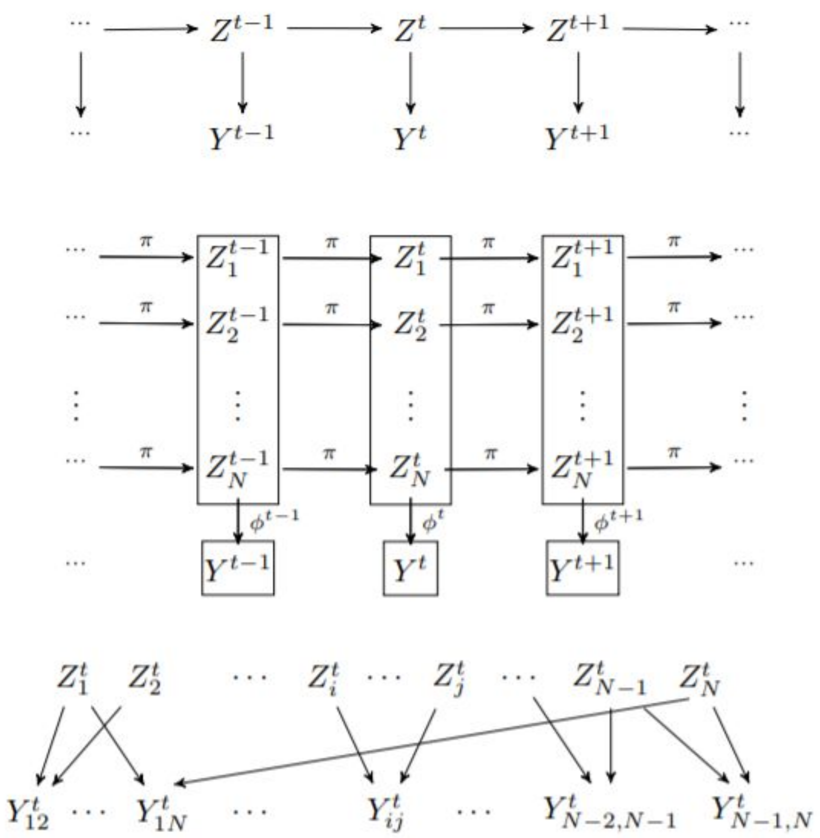


Figure 1. Dependency structures of the model. Top: general view corresponding to hidden Markov model (HMM) structure; Middle: details on latent structure organization corresponding to N different iid Markov chains $Z_i = (Z_i^t)_{1\leq t\leq T}$ across individuals; Bottom: details for fixed time point t corresponding to SBM structure.

Model description

- We assume a very generic form for distribution of Y_{ij}^t:

$$Y_{ij}^t | \{Z_{iq}^t Z_{jl}^t = 1\} \sim (1 - \beta_{ql}^t)\delta_0(\cdot) + \beta_{ql}^t F(\cdot, \gamma_{ql}^t),$$

- δ_0 is a Dirac mass at 0
- F is a parametric family of distributions with no point mass at 0
- γ is the parameter of F, indicating strength of interaction
- β ∈ [0,1] is sparsity parameter, with 1 corresponding to a complete graph
- The model is parameterized by:

$$\theta = (\boldsymbol{\pi}, \boldsymbol{\beta}, \boldsymbol{\gamma}) = (\boldsymbol{\pi}, \{\beta^t, \gamma^t\}_{1 \le t \le T}) = (\{\pi_{qq'}\}_{1 \le q, q' \le Q}, \{\beta_{ql}^t, \gamma_{ql}^t\}_{1 \le t \le T, 1 \le q \le l \le Q}) \in \Theta,$$

Varying connectivity parameters vs group membership

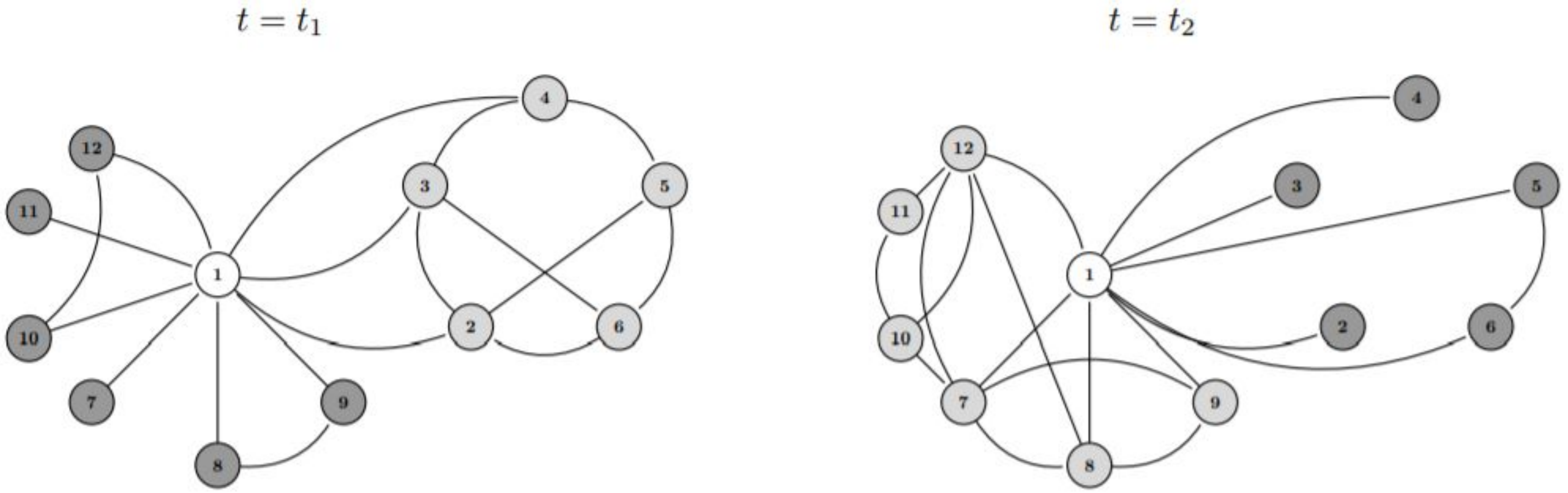


Figure 2. Connectivity parameters or group membership variation: a toy example.

Varying connectivity parameters vs group membership

- Two ways to interpret figure 2
- We choose to characterize groups by their stable within group connectivity parameter
- It is impossible to let both connectivity parameters and group membership vary through time without entering into label switching issues

Parameters identifiability

Definition(Identifiability up to label switching):

$$\forall \theta, \tilde{\theta} \in \Theta, \quad \mathbb{P}_{\theta}^{Y} = \mathbb{P}_{\tilde{\theta}}^{Y} \Rightarrow \exists \sigma \in \mathfrak{S}_{Q}, \theta = \sigma(\tilde{\theta}).$$

Note: Without further assumptions on the parameters **θ = (π,β,γ)** the parameters may not be recovered up to label switching.

Example(Non identifiability)

- $Q=2$(groups),$2T$ time steps,N individuals
- $\theta=(\pi,\beta,\gamma)$ where π is the identity 2x2 matrix and (β,γ) constant on time .
- $\underline{\theta}=(\underline{\pi},\underline{\beta},\underline{\gamma})$ another set of parameters.

Parameter Identifiability(Cont'ed)

Thus,we choose groups that are characterized through their stable within-group connectivity parameter,i.e we impose the following constraints on the parameter θ :

$$\forall q \in \mathcal{Q}, \forall t, t' \in \{1, \ldots, T\}, \quad \begin{cases} \text{binary case, } \beta_{qq}^{t} = \beta_{qq}^{t'} := \beta_{qq}, \\ \text{weighted case, } \gamma_{qq}^{t} = \gamma_{qq}^{t'} := \gamma_{qq}. \end{cases}$$

Cont’ed

Furthermore, we assume(Assumption 1) :

- **For any t , all the entries of γ are <u>distinct</u>.(Guarantees that the groups don’t look alike too much,no assumptions on the sparsity parameter β).**

Proposition 1. Considering the distribution $\mathbb{P}_\theta^Y$ on the set of observations and assuming constraint (2), the parameter $\theta = (\pi, \beta, \gamma)$ satisfies the following conditions.

(a) *Binary case*: θ is generically identified from $\mathbb{P}_\theta^Y$, up to label switching, as soon as N is not too small with respect to Q.

(b) *Weighted case*: under additional assumption 1, the parameter θ is identified from $\mathbb{P}_\theta^Y$, up to label switching, as soon as $N \geqslant 3$.

Algorithm

Variational Expectation-Maximization (VEM):

Complete-data log-likelihood of the model:

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

Dependence Structure of the conditional distribution P(Z|Y):

$$\mathbb{P}_\theta(\mathbf{Z}|\mathbf{Y}) = \mathbb{P}_\theta(Z^1|Y^1)\prod_{t=2}^{T}\mathbb{P}_\theta(Z^t|Z^{t-1},Y^t).$$

Algorithm

However, the distribution $\mathbb{P}_\theta(Z^t|Z^{t-1}, Y^t) = \mathbb{P}_\theta\{(Z_i^t)_{1\leqslant i\leqslant N}|Z^{t-1}, Y^t\}$ cannot be further factored. Indeed, for any $i \neq j$, the variables Z_i^t and Z_j^t are not independent when conditioned on Y^t. Our variational approximation naturally considers the following class of probability distributions $\mathbb{Q} := \mathbb{Q}_\tau$ parameterized by τ:

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

$$= \prod_{i=1}^{N} \left\{ \prod_{q=1}^{Q} \tau(i,q)^{Z_{iq}^1} \right\} \times \prod_{t=2}^{T} \prod_{1\leqslant q,q'\leqslant Q} \tau(t,i,q,q')^{Z_{iq}^{t-1} Z_{iq'}^t},$$

where, for any values (t,i,q,q'), we have $\tau(i,q)$ and $\tau(t,i,q,q')$ both belong to the set $[0,1]$ and are constrained by $\Sigma_q \tau(i,q) = 1$ and $\Sigma_{q'} \tau(t,i,q,q') = 1$. This class of probability distributions $\mathbb{Q}_\tau$ corresponds to considering independent laws through individuals, whereas, for each $i \in \{1,\ldots,N\}$, the distribution of Z_i under $\mathbb{Q}_\tau$ is the distribution of a Markov chain (through time t), with inhomogeneous transition $\tau(t,i,q,q') = \mathbb{Q}_\tau(Z_i^t = q'|Z_i^{t-1} = q)$ and initial distribution $\tau(i,q) = \mathbb{Q}_\tau(Z_i^1 = q)$.

Algorithm

We shall need the marginal components of $\mathbb{Q}_\tau$, namely $\tau_{\text{marg}}(t,i,q) := \mathbb{Q}_\tau(Z_i^t = q)$. These quantities are computed recursively by

$$\tau_{\text{marg}}(1,i,q) = \tau(i,q) \text{ and } \forall t \geqslant 2,$$

$$\tau_{\text{marg}}(t,i,q) = \sum_{q'=1}^{Q} \tau_{\text{marg}}(t-1,i,q')\tau(t,i,q',q).$$

Algorithm

$$
\begin{aligned}
J(\theta,\tau) :=& \mathbb{E}_{\mathbb{Q}_\tau}[\log\{\mathbb{P}_\theta(\mathbf{Y},\mathbf{Z})\}] + \mathcal{H}(\mathbb{Q}_\tau) \\
=& \sum_{i=1}^{N}\sum_{q=1}^{Q} \tau(i,q)[\log(\alpha_q) - \log\{\tau(i,q)\}] \\
&+ \sum_{t=2}^{T}\sum_{i=1}^{N}\sum_{1\leqslant q,q'\leqslant Q} \tau_{\mathrm{marg}}(t-1,i,q)\,\tau(t,i,q,q')[\log(\pi_{qq'}) - \log\{\tau(t,i,q,q')\}] \\
&+ \sum_{t=1}^{T}\sum_{1\leqslant i<j\leqslant N}\sum_{1\leqslant q,l\leqslant Q} \tau_{\mathrm{marg}}(t,i,q)\,\tau_{\mathrm{marg}}(t,j,l)\log\{\phi_{ql}^{t}(Y_{ij}^{t})\}.
\end{aligned}
$$

It consists of iterating the following two steps. At the kth iteration, with current parameter value $(\tau^{(k)},\theta^{(k)})$, we do the following steps.

(a) VE step: compute $\tau^{(k+1)} = \arg\max_\tau J(\theta^{(k)},\tau)$.
(b) M-step: compute $\theta^{(k+1)} = \arg\max_\theta J(\theta,\tau^{(k+1)})$.

Algorithm - Optimization

Proposition 2. The value $\hat{\tau}$ that maximizes in τ the function $J(\theta,\tau)$ satisfies the fixed point equation

$$\forall t \geqslant 2, \forall i \geqslant 1, \forall q, q' \in \mathcal{Q}, \quad \hat{\tau}(t,i,q,q') \propto \pi_{qq'} \prod_{j=1}^{N} \prod_{l'=1}^{Q} \phi_{q'l'}^{t}(Y_{ij}^{t})^{\hat{\tau}_{\mathrm{marg}}(t,j,l')}$$

- Update this value: $\forall i \geqslant 1, \forall q \in \mathcal{Q}, \quad \hat{\tau}(i,q) \propto \alpha_q \prod_{j=1}^{N} \prod_{l=1}^{Q} \phi_{ql}^{1}(Y_{ij}^{1})^{\hat{\tau}(j,l)}.$

- Parameter α is not obtained from maximizing J as it is not a free parameter but rather the stationary distribution associated with transition π. Thus, α is obtained from the empirical mean of the marginal distribution over all data points:

$$\forall q \in \mathcal{Q}, \quad \hat{\alpha}_q = \frac{1}{NT} \sum_{t=1}^{T} \sum_{i=1}^{N} \hat{\tau}_{\mathrm{marg}}(t,i,q).$$

Algorithm - Optimization

- Other parameter values that maximize J:

$$\forall (q,q') \in \mathcal{Q}^2, \quad \hat{\pi}_{qq'} \propto \sum_{t=2}^{T} \sum_{i=1}^{N} \tau_{\text{marg}}(t-1,i,q)\,\tau(t,i,q,q'),$$

$$\forall t, \forall q \neq l \in \mathcal{Q}^2, \quad \hat{\beta}^t_{ql} = \frac{\sum_{i,j} \tau_{\text{marg}}(t,i,q)\,\tau_{\text{marg}}(t,j,l)\,\mathbf{1}_{Y^t_{ij} \neq 0}}{\sum_{i,j} \tau_{\text{marg}}(t,i,q)\,\tau_{\text{marg}}(t,j,l)},$$

$$\forall q \in \mathcal{Q}, \quad \hat{\beta}_{qq} = \frac{\sum_{t} \sum_{i,j} \tau_{\text{marg}}(t,i,q)\,\tau_{\text{marg}}(t,j,q)\,\mathbf{1}_{Y^t_{ij} \neq 0}}{\sum_{t,i,j} \tau_{\text{marg}}(t,i,q)\,\tau_{\text{marg}}(t,j,q)}.$$

- Finally, optimization with respect to γ depends on the choice of the parametric family.

Model Selection

- Number of groups Q is important.
- Maximizing an integrated classification likelihood (ICL).
- For any number of groups Q1, let $\hat{\theta}_Q$ be the estimated parameter value with Q groups and $\hat{\mathbf{Z}}$ the corresponding maximum a posteriori classification $\hat{\theta}_Q$. In our case, the general form of the ICL is given by

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

- The first penalization term accounts for transition matrix π and pen(N, T ,β, γ) is a penalizing term for the connectivity parameters (β, γ).

Algorithm Initialization

- Initialize our VEM procedure by running k-means on the rows of a concatenated data matrix containing all the adjacency time step matrices Yt stacked in consecutive column blocks.
- Initial clustering of the individuals is constant across time.
- A consequence of this choice is that this initialization works well when the group memberships do not vary too much across time.
- Increasing T increases the probability for an individual to change group at some point in time and thus, starting with a constant-in-time clustering of the individuals, it becomes more difficult to infer the group memberships at each time point correctly.

Adjusted Rand Index(ARI)

Definition: Rand Index(RI) is a measure of the similarity between two data clusterings. ARI is the corrected-for-chance version of RI.

Clustering Performance of the Algorithm using ARI:

Evaluation of the estimated and true latent structure using two different criteria:

1. **The averaged value of ARI over all the time steps.**
2. **Global ARI value which compares the clustering of the nodes for all time steps.**

Cont’ed

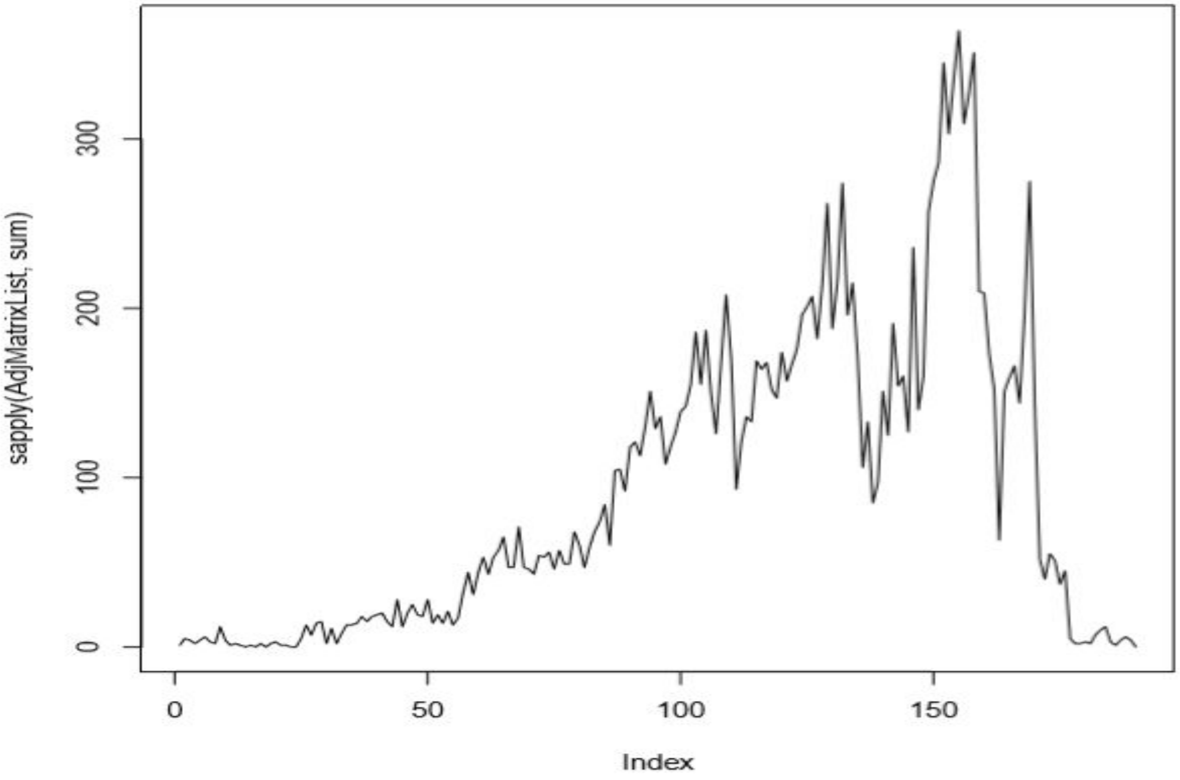
Cons:

- **The averaged ARI doesn’t give any information about the group membership trajectories along time.**
- **Using the global ARI is more difficult to obtain good performance.**
- **Both suffer when either the within-group stability is too low or when the groups look alike (Same β , not sufficiently distinct γ between groups)**

Pros:

- **The averaged ARI outperforms global ARI in the planted partition model.**
- **Both work well in identifiable cases.**

An application on real dataset



An application on real dataset

- 184 nodes, 190 weeks
- Very sparse networks -> truncate to get the middle section: week 81-170
- Divide these 90 weeks into 10 bins, thus each bin is roughly 2 months
- Combine adjacency matrices in each bin
- Symmetrize these 10 adjacency matrices
- Delete absent nodes

Graph Characteristics

1) Unweighted (Binary)

$$Y_{ij}^{t} \mid \{Z_{iq}^{t}\ Z_{jl}^{t} = 1\} \sim (1 - \beta_{ql}^{t})\ \delta_0(\ y\ ) + \beta_{ql}^{t}\ F(\ y\ ,\ \gamma_{ql}^{t}\ )$$

$$\downarrow$$

$$Y_{ij}^{t} \mid \{Z_{iq}^{t}\ Z_{jl}^{t} = 1\} \sim (1 - \beta_{ql}^{t})\ \delta_0(\ y\ ) + \beta_{ql}^{t}\ \delta_1(\ y\ )$$

$$\longrightarrow \quad Y_{ij}^{t} \sim \text{Bernoulli}(\beta_{ql}^{t})$$

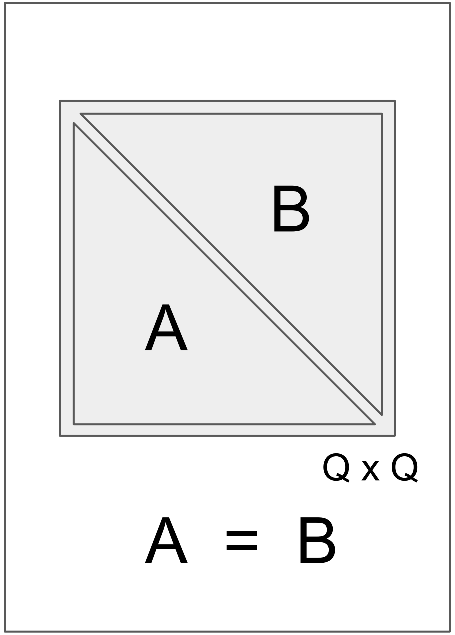
Graph Characteristics

2) Undirected (Symmetric)

Parameters β_{ql}^{t} and γ_{ql}^{t} satisfy

$$\beta_{ql}^{t} = \beta_{lq}^{t} \text{ and } \gamma_{ql}^{t} = \gamma_{ql}^{t}$$

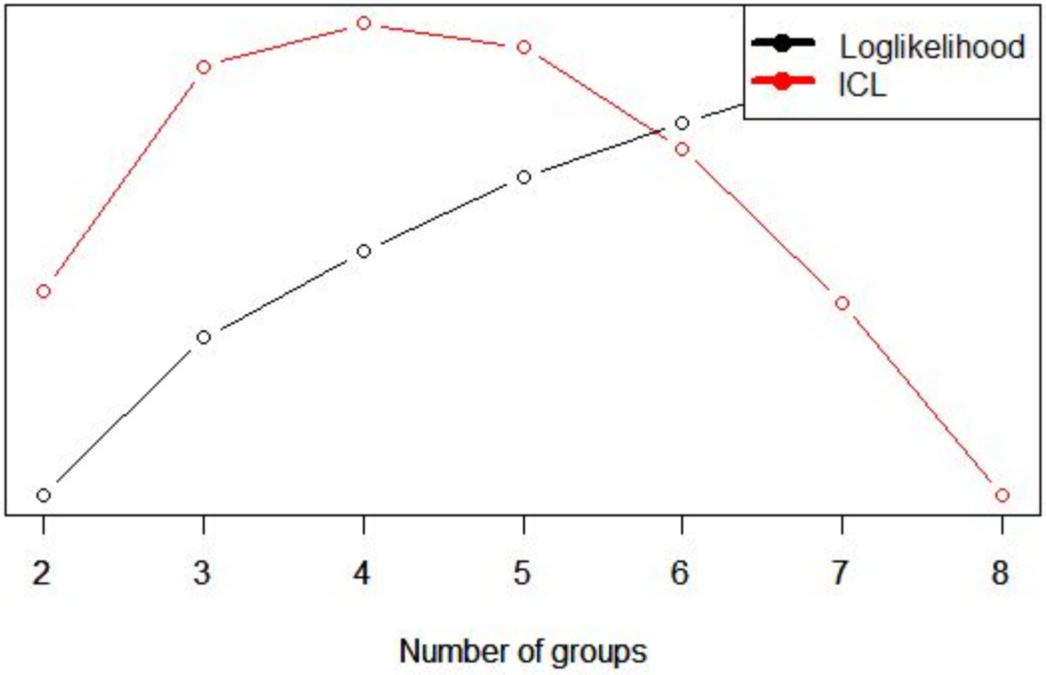
for all $1 \leq q, l \leq Q$.



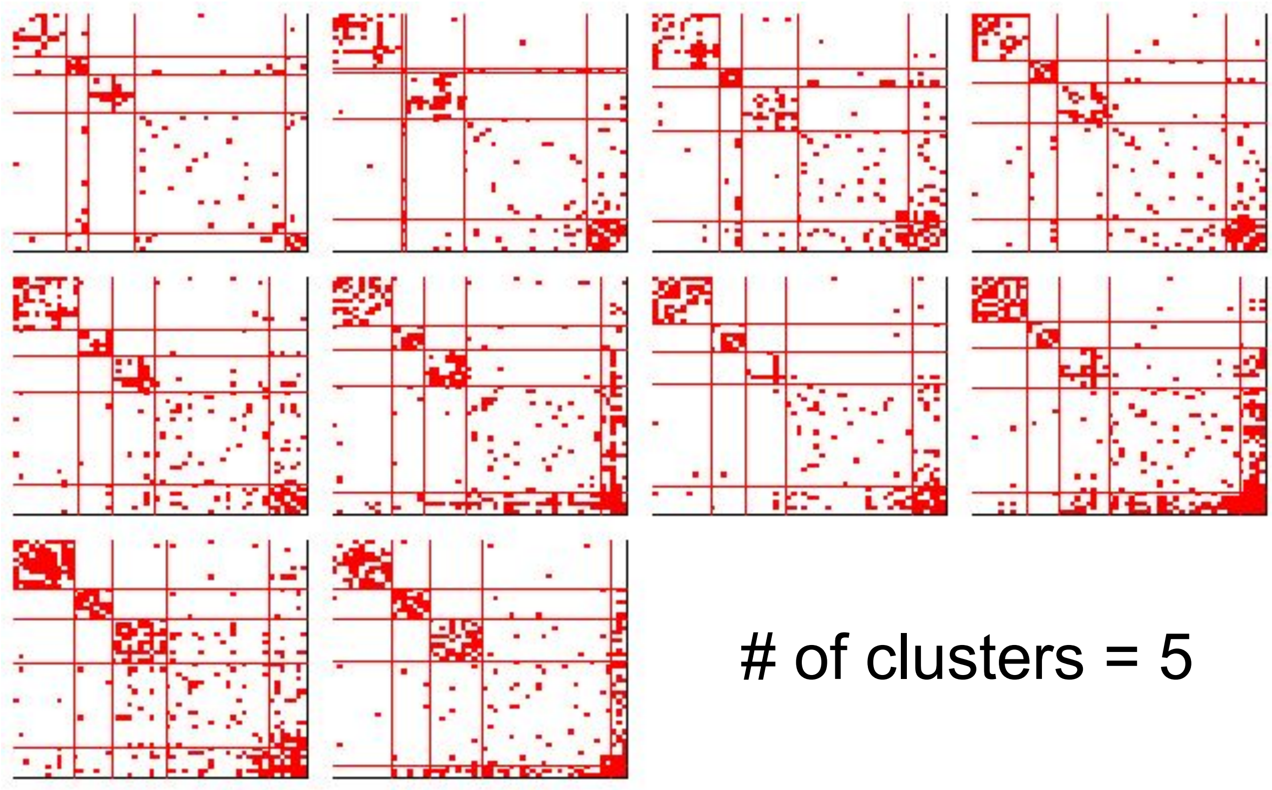
Model Selection

Loglikelihood - Measures how likely the parameters are, given the data

ICL - Penalized form of Loglikelihood

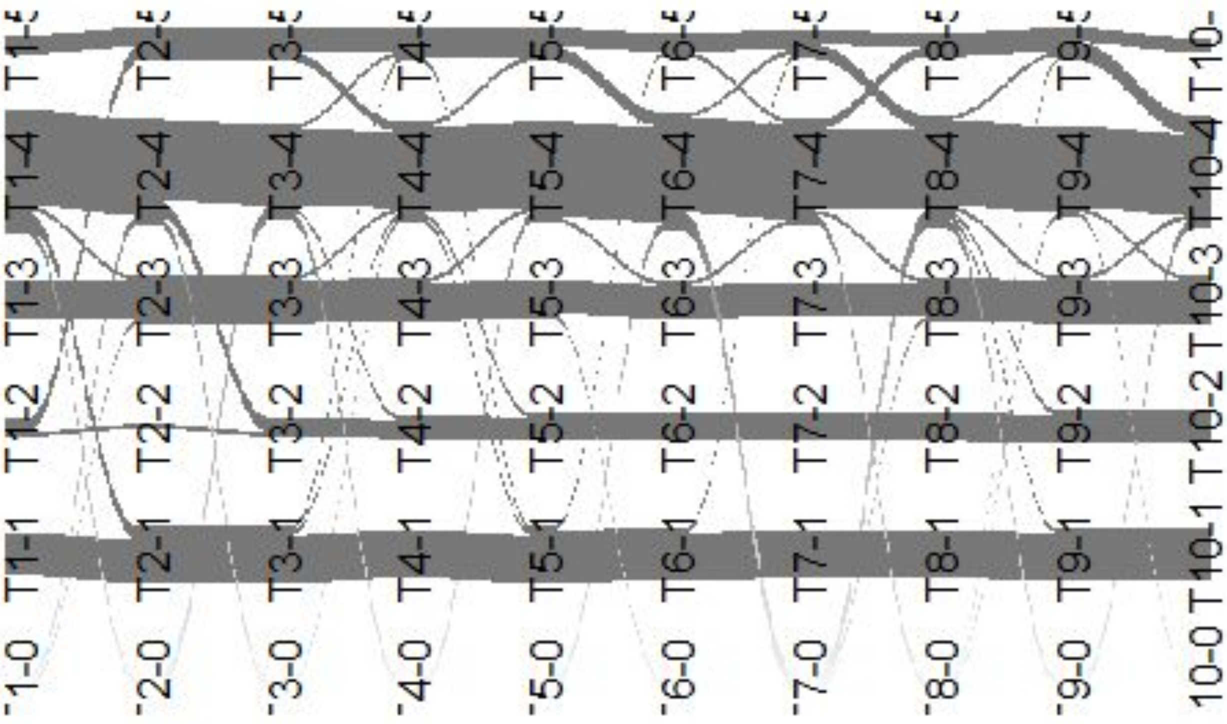


Re-Ordered Adjacency Matrices



# of clusters = 5

“Alluvial Plot” - Group Switching Through Time



Connectivity Through Time