Estimating the number of clusters in a stochastic block model for temporal networks

Riccardo Rastelli

joint work with Pierre Latouche and Nial Friel

CMStatistics 2017

16 - 18 December 2017, University of London, UK
Networks are used to represent the interactions between entities or objects or individuals.

The networks we focus on are:
- binary;
- directed;
- without self-loops;
- dynamically evolving over time.

The time dimension is discrete.

An adjacency cube may be used to represent the network:

\[ x_{ij}^{(t)} = \begin{cases} 
1 & \text{if an edge from } i \text{ to } j \text{ appears at time } t \\ 
0 & \text{otherwise.} 
\end{cases} \]

where \( t \in \{1, \ldots, T\} \) and \( i, j \in \{1, \ldots, N\} \).
- $z_i^{(t)}$ denotes the **cluster membership** of node $i$ at time $t$.

- $z_i^{(t)} \in \{1, \ldots, K\}$ where $K$ is the *overall* number of groups.

- $\forall t: \{z_1^{(t)}, \ldots, z_N^{(t)}\}$ denotes a **partition** of $\{1, \ldots, N\}$.

- The data are **conditionally independent** given the allocations:

  $$
  \mathbb{P}(X|Z, \theta) = \prod_{t=1}^{T} \prod_{\{(i,j): i \neq j\}} \mathbb{P}(x_{ij}^{(t)}|z_i^{(t)}, z_j^{(t)}, \theta)
  $$

- Allocations also determine the **probability of connection**:

  $$
  \mathbb{P}(x_{ij}^{(t)} = 1|z_i^{(t)} = g, z_j^{(t)} = h, \theta) = \theta_{gh}
  $$
- $z_i^{(t)}$ denotes the **cluster membership** of node $i$ at time $t$.

- $z_i^{(t)} \in \{1, \ldots, K\}$ where $K$ is the *overall* number of groups.

- $\forall t : \{z_1^{(t)}, \ldots, z_N^{(t)}\}$ denotes a **partition** of $\{1, \ldots, N\}$.

- The data are **conditionally independent** given the allocations:

\[
P(X | Z, \theta) = \prod_{t=1}^{T} \prod_{\{(i,j) : i \neq j\}} P(x_{ij}^{(t)} | z_i^{(t)}, z_j^{(t)}, \theta)
\]

- Allocations also determine the **probability** of connection:

\[
P(x_{ij}^{(t)} = 1 | z_i^{(t)} = g, z_j^{(t)} = h, \theta) = \theta_{gh}
\]
The allocations evolve according to $N$ independent discrete Markov processes, i.e.:

$$\pi_{gh} = \mathbb{P}\left(z_i^{(t)} = h \mid z_i^{(t-1)} = g\right)$$

Group counts give an approximation of the stationary distribution, which is used to model the initial states:

$$z_i^{(1)} \overset{IID}{\sim} \text{Multinomial} \left(\alpha_1, \ldots, \alpha_K\right)$$

with:

$$\alpha_g \propto \sum_{t=2}^{T} \sum_{i=1}^{N} \mathbb{1}\{z_i^{(t)} = g\}$$
The model parameters are:

\[
Z = \begin{pmatrix}
    z_{1}^{(1)} & \cdots & z_{N}^{(1)} \\
    \vdots & \ddots & \vdots \\
    z_{1}^{(T)} & \cdots & z_{N}^{(T)}
\end{pmatrix}, \quad \pi = \begin{pmatrix}
    \pi_{11} & \cdots & \pi_{1K} \\
    \vdots & \ddots & \vdots \\
    \pi_{K1} & \cdots & \pi_{KK}
\end{pmatrix},
\]

\[
\theta = \begin{pmatrix}
    \theta_{11} & \cdots & \theta_{1K} \\
    \vdots & \ddots & \vdots \\
    \theta_{K1} & \cdots & \theta_{KK}
\end{pmatrix}
\]

- How can we estimate the model parameters for a given \( K \)?
- **Model-choice**: what is the optimal number of groups \( K \)?

We introduce a methodology that provides an answer to both questions within one single algorithmic framework.
Bayesian hierarchical model

- **Connection probabilities:**
  \[ \theta_{gh} \sim \text{Beta}(a_{gh}, b_{gh}) \]
  We assume that \( \phi = a_{gh} = b_{gh} \) for all \( g \) and \( h \).

- **Transition probabilities:**
  \[ \forall g : (\pi_{g1}, \ldots, \pi_{gK}) \sim \text{Dirichlet}(\delta_{g1}, \ldots, \delta_{gK}) \]
  We assume that \( \delta = \delta_{gh} \) for all \( g \) and \( h \).

- In the applications we fix \( \phi = \delta = 1 \), which gives *uninformative priors*. 
Data: $\mathcal{X} = \{X^{(1)}, \ldots, X^{(T)}\}$.

Allocations: $\mathcal{Z} = \{z^{(1)}, \ldots, z^{(T)}\}$.

Model parameters: $\pi$ and $\theta$.

Hyperparameters: $\delta$ and $\phi$. 
Taking advantage of the **conjugacy** of distributions...

- **Analytical formula for the marginal prior** is available:
  \[
  p(\mathcal{Z}|\delta) = \int_{\Pi} p(\mathcal{Z}, \pi|\delta) \, d\pi
  \]

- **Same for the marginal likelihood:**
  \[
  p(\mathcal{X}|\mathcal{Z}, \phi) = \int_{\Theta} p(\mathcal{X}, \theta|\mathcal{Z}, \phi) \, d\theta
  \]

- **Same for the marginal posterior:**
  \[
  p(\mathcal{Z}|\mathcal{X}, \phi, \delta) = \int_{\Pi} \int_{\Theta} p(\mathcal{Z}, \theta, \pi|\mathcal{X}, \phi, \delta) \, d\pi \, d\theta
  \propto p(\mathcal{X}|\mathcal{Z}, \phi) p(\mathcal{Z}|\delta)
  \]
MAP and $\mathcal{ICL}_{ex}$

- $\mathcal{Z}$ are the only unknown variables in $p(\mathcal{Z}|\mathcal{X}, \phi, \delta)$.

- The allocations $\hat{\mathcal{Z}}$ maximising $p(\mathcal{Z}|\mathcal{X}, \phi, \delta)$ are MAP.

- In a finite mixture model context, the same quantity corresponds to the exact Integrated Complete Likelihood:

$$\mathcal{ICL}_{ex} := p(\mathcal{X}, \mathcal{Z}|\phi, \delta) \propto p(\mathcal{Z}|\mathcal{X}, \phi, \delta)$$

- This is the exact quantity that the ICL criterion approximates.

- $\hat{\mathcal{Z}}$ automatically identifies the optimal number of groups, in the ICL sense.
MAP and $\mathcal{ICL}_{ex}$

- $Z$ are the only unknown variables in $p(Z|X, \phi, \delta)$.

- The allocations $\hat{Z}$ maximising $p(Z|X, \phi, \delta)$ are MAP.

- In a finite mixture model context, the same quantity corresponds to the exact Integrated Complete Likelihood:

  \[ \mathcal{ICL}_{ex} := p(X, Z|\phi, \delta) \propto p(Z|X, \phi, \delta) \]

- This is the exact quantity that the ICL criterion approximates.

- $\hat{Z}$ automatically identifies the optimal number of groups, in the ICL sense.
MAP and $\mathcal{ICL}_{ex}$

- $\mathcal{Z}$ are the only unknown variables in $p(\mathcal{Z}|\mathcal{X}, \phi, \delta)$.

- The allocations $\hat{\mathcal{Z}}$ maximising $p(\mathcal{Z}|\mathcal{X}, \phi, \delta)$ are MAP.

- In a finite mixture model context, the same quantity corresponds to the exact Integrated Complete Likelihood:

  $$\mathcal{ICL}_{ex} := p(\mathcal{X}, \mathcal{Z}|\phi, \delta) \propto p(\mathcal{Z}|\mathcal{X}, \phi, \delta)$$

- This is the exact quantity that the ICL criterion approximates.

- $\hat{\mathcal{Z}}$ automatically identifies the optimal number of groups, in the ICL sense.
GreedyIcl algorithm:

1. Initialise $\mathcal{Z}$.

2. Reallocate each node $(t,i)$ in turn, moving it to the group that gives the best increase in $\mathcal{ICL}_{ex}$.

3. Repeat 2 until no further increase in $\mathcal{ICL}_{ex}$ is possible.

4. Perform hierarchical clustering to improve the solution.

Remark: the overall complexity is $O\left(TN(m + K^2)\right)$, where $m$ is the average degree.

Remark: the algorithm is bound to return a local optimum.
Simulations

Average Normalised Mutual Information
Cluster stability = 0.9

Intra–cluster connection probability

Average NMI

●
●
●
●
●
●
●
●
●
●
●
●
●
●
●
●

dynsbm
GreedyICL

Model–choice performance
Cluster stability = 0.9

Proportion of correct estimations of K

● ● ● ●
●
●
●
● ●
● ● ●
●
●
● ● ● ●●
●

dynsbm
GreedyICL
The data consist of all **bike hires** in London from Wednesday 5 June 2013 to Wednesday 12 June 2013.

- $N = 566$ stations are active.

- The time is split in $T = 64$ intervals of **three hours**.

- We create the **adjacency cube** as follows:

$$x_{i,j}^{(t)} = \begin{cases} 
1 & \text{if at least one bike is hired in station } i, \\
& \text{during time frame } t, \text{ and is then returned in } j; \\
0 & \text{otherwise.} 
\end{cases}$$

- Self-edges are discarded.
The data consist of all bike hires in London from Wednesday 5 June 2013 to Wednesday 12 June 2013.

$N = 566$ stations are active.

The time is split in $T = 64$ intervals of three hours.

We create the adjacency cube as follows:

$$x_{ij}^{(t)} = \begin{cases} 
1 & \text{if at least one bike is hired in station } i, \\
& \text{during time frame } t, \text{ and is then returned in } j; \\
0 & \text{otherwise.}
\end{cases}$$

Self-edges are discarded.
The data consist of all bike hires in London from Wednesday 5 June 2013 to Wednesday 12 June 2013.

\[ N = 566 \] stations are active.

The time is split in \[ T = 64 \] intervals of three hours.

We create the adjacency cube as follows:

\[
x_{ij}^{(t)} = \begin{cases} 
1 & \text{if at least one bike is hired in station } i, \\
& \text{during time frame } t, \text{ and is then returned in } j; \\
0 & \text{otherwise.}
\end{cases}
\]

Self-edges are discarded.
Measuring activity level

Number of edges

Wed Thu Fri Sat Sun Mon Tue Wed

Number of edges
The total number of groups is $K = 43$. 

**Measuring heterogeneity**

---

Number of non-empty groups

![Graph showing the number of non-empty groups over a week. The graph displays fluctuations in the number of non-empty groups throughout the week. The total number of groups is stated at $K = 43$.](image-url)
Measuring stability

Stability of nodes

Number of swaps

Frequencies

0 10 20 30 40 50 60
0 50 100 150

Stability of nodes

Number of swaps

Frequencies

0 10 20 30 40 50 60
Instability and positions

Most unstable stations (more than 35 swaps)
Stability and positions

Stations not changing allocation
- Non-binary networks can be handled using **conjugacy**.

- Supervised setting $\Rightarrow$ update only non-fixed allocations.

- This can be exploited to handle a dynamic set of nodes by selecting those “out of the study”.

- The **collapsing** of *transition probabilities* is rather original for **Hidden Markov Models**. This can be exploited in a more general scenario to infer the hidden states and $K$ at the same time.
Non-binary networks can be handled using conjugacy.

**Supervised setting** ⇒ update only non-fixed allocations.

This can be exploited to handle a dynamic set of nodes by selecting those “out of the study”.

The collapsing of transition probabilities is rather original for Hidden Markov Models. This can be exploited in a more general scenario to infer the hidden states and $K$ at the same time.
- Non-binary networks can be handled using **conjugacy**.

- **Supervised setting** ⇒ update only non-fixed allocations.

- This can be exploited to handle a dynamic set of nodes by selecting those “out of the study”.

- The **collapsing** of *transition probabilities* is rather original for **Hidden Markov Models**. This can be exploited in a more general scenario to infer the hidden states and $K$ at the same time.
Conclusions

- The exact ICL is available for a **dynamic stochastic block model**.

- The maximisation of $\mathcal{ICL}_{ex}$ is performed through a heuristic **greedy algorithm**.

- The **number of groups** is obtained **automatically**.

- The method is applied to a **bike sharing** dataset, proposing **model-based measures** of **heterogeneity**, **activity level**, and **stability**.
The exact ICL is available for a dynamic stochastic block model.

The maximisation of $\mathcal{ICL}_{ex}$ is performed through a heuristic greedy algorithm.

The number of groups is obtained automatically.

The method is applied to a bike sharing dataset, proposing model-based measures of heterogeneity, activity level, and stability.
Conclusions

- The exact ICL is available for a dynamic stochastic block model.
- The maximisation of $\mathcal{ICL}_{ex}$ is performed through a heuristic greedy algorithm.
- The number of groups is obtained automatically.
- The method is applied to a bike sharing dataset, proposing model-based measures of heterogeneity, activity level, and stability.
Conclusions

- The exact ICL is available for a **dynamic stochastic block model**.

- The maximisation of $I\!C\!L_{ex}$ is performed through a heuristic **greedy algorithm**.

- The **number of groups** is obtained **automatically**.

- The method is applied to a **bike sharing** dataset, proposing **model-based measures** of **heterogeneity**, **activity level**, and **stability**.