Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani
PIANO NAZIONALE DI RIPRESA E RESILIENZA
UNIVERSITÀ DEGLI STUDI DI MILANO BICOCCA

Link prediction in temporal networks: A dynamic stochastic block model approach

Luca Brusa[1]
(luca.brusa@unimib.it)

Silvia Pandolfi[2]
Fulvia Pennoni[1]
Francesco Bartolucci[2]

[1]University of Milano-Bicocca, Department of Statistics and Quantitative Methods (Milan, Italy)

[2]University of Perugia, Department of Economics (Perugia, Italy)

Outline

Longitudinal network data

- **Multiple snapshots** of a network involving n nodes are available at T different time occasions
- We focus on **undirected graphs without self-loops:** the existence of an edge from node i to node j implies an edge from j to i
- **Main notation** $(i, j = 1, \ldots, n,\ i \neq j,\ t = 1, \ldots, T)$:
  - $Y_{ij}^{(t)}$ **binary random variable** equal to 1 if there is an edge between nodes i and j at occasion t, and 0 otherwise
  - $y_{ij}^{(t)}$: realization of $Y_{ij}^{(t)}$
  - $\boldsymbol{Y}^{(t)}$: **symmetric adjacency matrix** recorded at occasion t
  - $\mathcal{Y} = \{\boldsymbol{Y}^{(t)}, t = 1, \ldots, T\}$: set of **network snapshots** taken across time

Dynamic stochastic block model (DSBM)

- Aim: identify homogeneous (latent) blocks of nodes having a **similar social behavior** and their **dynamic across time**
- Block membership depends on **individual- and time-specific** discrete latent variables $U_i^{(t)} \in \{1, \ldots, k\}$
- The vectors $\boldsymbol{U}_i = (U_i^{(1)}, \ldots, U_i^{(T)})'$ are **mutually independent** and **identically distributed** according to a **hidden Markov chain**:
  - **initial probabilities:** $\pi_u = p(U_i^{(1)} = u), \quad u = 1, \ldots, k$
  - **transition probabilities:** $\pi_{v|u} = p(U_i^{(t)} = v \mid U_i^{(t-1)} = u)$, $t = 2, \ldots, T,\ u, v = 1, \ldots, k$
- $\mathcal{U} = \{\boldsymbol{U}_i, i = 1, \ldots, n\}$ overall set of **latent variables**

Likelihood function

- Conditional on $U_i^{(t)} = u$ and $U_j^{(t)} = v$, variables $Y_{ij}^{(t)}$ are independent (**local independence assumption**) and identically distributed:

$$Y_{ij}^{(t)} \mid U_i^{(t)} = u, U_j^{(t)} = v \overset{i.i.d.}{\sim} \mathcal{B}(\beta_{uv})$$

  - **connection probabilities** $\beta_{uv} = p(Y_{ij}^{(t)} = 1 \mid U_i^{(t)} = u, U_j^{(t)} = v)$

- **Likelihood function** (observed network distribution):

$$p(\mathcal{Y}) = \sum_{\mathcal{U}} p(\mathcal{Y} \mid \mathcal{U}) p(\mathcal{U})$$

- **Full maximum likelihood** estimates are not achievable, apart from networks of a very limited size

Variational maximum likelihood inference

- A **variational approximation** of the Expectation-Maximization algorithm (**VEM**) is frequently considered as an alternative (Yang et al., 2011; Matias & Miele, 2017; Bartolucci & Pandolfi, 2020)
- The approximation is based on maximizing a **lower bound** $\mathcal{J}(\boldsymbol{\theta})$ of the log-likelihood function:

$$\mathcal{J}(\boldsymbol{\theta}) = \log p(\mathcal{Y}) - KL\left[\mathbb{Q}(\mathcal{U}) \,||\, p(\mathcal{U} \mid \mathcal{Y})\right]$$

  - $\boldsymbol{\theta}$ denotes the vector of all free model parameters
  - KL denotes the **Kullback-Leibler divergence** between the true **intractable** posterior distribution of the latent variables $p(\mathcal{U} \mid \mathcal{Y})$ and a suitable approximation $\mathbb{Q}(\mathcal{U})$

Variational expectation-maximization algorithm

- We assume **a posteriori** independence across individuals and time occasions:

$$\mathbb{Q}(\mathcal{U}) = \prod_t \prod_i \mathbb{Q}(U_i^{(t)}) = \prod_t \prod_i \prod_u \tau(t,i,u)^{I(U_i^{(t)}=u)}$$

  - $\tau(t,i,u)$: approximation of $p(U_i^{(t)} = u \mid \mathcal{Y})$

- To maximize function $\mathcal{J}(\boldsymbol{\theta})$ the VEM algorithm alternates **two steps** until convergence:
  - the **E-step** maximizes $\mathcal{J}(\boldsymbol{\theta})$ with respect to $\tau(t,i,u)$, with $\boldsymbol{\theta}$ fixed at the values obtained from the previous iteration
  - the **M-step** maximizes $\mathcal{J}(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$ with the $\tau(t,i,u)$ fixed at the value obtained from the E-step

Outline

Out-of-sample forecasting

- Aim: predict the **one-step-ahead probability** of a potential edge between a given couple of nodes (i, j) at time occasion $T + 1$:

  1. **conditional on the predicted latent memberships**:

$$\hat{p}_{ij}^{(T+1)} = p(Y_{ij}^{(T+1)} = 1) = \hat{\beta}_{\hat{u}_i^{(T+1)}, \hat{u}_j^{(T+1)}}$$

  2. **unconditional on the predicted latent memberships**:

$$\tilde{p}_{ij}^{(T+1)} = p(Y_{ij}^{(T+1)} = 1) = \sum_u \sum_v \hat{\tau}(T+1, i, u)\hat{\tau}(T+1, j, v)\hat{\beta}_{uv}$$

  based on the observed data and the estimated model parameters:

  - $\hat{\tau}(T+1, i, u) = \sum_{\bar{u}} \hat{\tau}(T, i, \bar{u})\hat{\pi}_{u|\bar{u}}$ posterior block membership probabilities at time $T + 1$
  - $\hat{u}_i^{(T+1)} = \underset{u}{\text{argmax}}\, \hat{\tau}(T+1, i, u)$: predicted block (MAP rule)
  - $\hat{\pi}_{u|\bar{u}}$, $\hat{\beta}_{uv}$: estimated transition and connection probabilities

Conditional vs unconditional

- The choice between the **conditional** and **unconditional** approaches may depend on the specific task
- When the goal is to predict the **link probability for a specific node pair**, or to produce a **ranking of the most likely future links**, both methods are suitable
- When the goal is to predict the **joint probability of future connections** over a subgraph or the entire graph the **conditional** approach is more appropriate
  - The DSBM assumes that edge existence is conditionally independent across node pairs given the block assignments
  - The **unconditional** approach may **overlook the underlying dependencies** captured by the block structure

Evaluation metrics for temporal link prediction

- In order to assess the performance of the proposed link prediction, we rely on standard evaluation metrics:
  - **Area under the ROC curve** (AUC)
  - **F1-score**
- To calculate the F1-score it is necessary to set a **threshold** $c \in [0, 1]$: we predict a future link between nodes i and j if

$$\hat{p}_{ij}^{(T+1)} > c \quad \text{or} \quad \tilde{p}_{ij}^{(T+1)} > c$$

- To calibrate c, we rely on the **(in-sample) link predictions** $\hat{p}_{ij}^{(t)}$ or $\tilde{p}_{ij}^{(t)}$, for $t = 1, \ldots, T$
- The **optimal value of the threshold**, c^*, is the one that maximizes the F1-score based on the comparison between the (in-sample) predictions and the observed links

Outline

Simulation design

- Different scenarios to evaluate the **performance of the link prediction** approaches:
  - number of nodes ($n = 100, 200$)
  - number of time occasions ($T = 10, 20$)
  - number of blocks ($k = 3, 4$)
  - levels of group stability (highly and weakly persistent latent blocks)
  - **community structures**:
    - M_1: strong intra-group connectivity and weak inter-group connectivity
    - M_2: high inter-group connectivity and weak intra-group connectivity
    - M_3: high intra-group connection probabilities and varying levels of inter-group separation (moderate connectivity, weak connectivity, and near disconnection) between different pairs of blocks
- For each scenario 100 networks are generated and for each network snapshots are partitioned into a **training set** (first $T - 1$ time occasions) and a **test set** (last observed time T)

Results

Table 1: Performance metrics for link prediction, on the basis of predicted probabilities computed conditionally ($\hat{p}_{ij}^{T+1}$) and unconditionally ($\tilde{p}_{ij}^{T+1}$) on the predicted block memberships, under different simulated scenarios

	Benchmark[1]	
	$\hat{p}_{ij}^{T+1}$	$\tilde{p}_{ij}^{T+1}$
F1-score (in-sample)	0.850	0.850
F1-score (out-of-sample)	0.674	0.669
c^* (average)	0.500	0.500
AUC	0.780	0.803
Sensitivity/Recall	0.708	0.696
Specificity	0.829	0.833
Precision	0.644	0.647
Accuracy	0.792	0.791

[1] Benchmark scenario: $n = 100$ nodes, $T = 10$ time snapshots, $k = 3$ latent blocks, high persistence probabilities, connectivity parameters corresponding to model M_1.

Table 1: cont'd

	$n = 200$		$k = 4$		$T = 20$	
	$\hat{p}_{ij}^{T+1}$	$\tilde{p}_{ij}^{T+1}$	$\hat{p}_{ij}^{T+1}$	$\tilde{p}_{ij}^{T+1}$	$\hat{p}_{ij}^{T+1}$	$\tilde{p}_{ij}^{T+1}$
F1-score (in-sample)	0.850	0.850	0.834	0.834	0.847	0.847
F1-score (out-of-sample)	0.679	0.679	0.651	0.641	0.676	0.673
c^* (average)	0.500	0.500	0.500	0.500	0.496	0.496
AUC	0.787	0.807	0.783	0.821	0.783	0.803
Sensitivity/Recall	0.711	0.711	0.661	0.639	0.712	0.698
Specificity	0.832	0.832	0.881	0.888	0.826	0.824
Precision	0.649	0.649	0.641	0.647	0.645	0.642
Accuracy	0.795	0.795	0.828	0.827	0.791	0.785
	low persistence		M_2		M_3	
	$\hat{p}_{ij}^{T+1}$	$\tilde{p}_{ij}^{T+1}$	$\hat{p}_{ij}^{T+1}$	$\tilde{p}_{ij}^{T+1}$	$\hat{p}_{ij}^{T+1}$	$\tilde{p}_{ij}^{T+1}$
F1-score (in-sample)	0.825	0.825	0.773	0.773	0.883	0.883
F1-score (out-of-sample)	0.473	0.275	0.695	0.695	0.805	0.805
c^* (average)	0.451	0.451	0.500	0.500	0.500	0.500
AUC	0.622	0.642	0.672	0.692	0.794	0.799
Sensitivity/Recall	0.523	0.154	0.856	0.856	0.782	0.782
Specificity	0.699	0.744	0.494	0.494	0.760	0.760
Precision	0.438	0.440	0.585	0.585	0.829	0.829
Accuracy	0.646	0.564	0.659	0.659	0.774	0.774

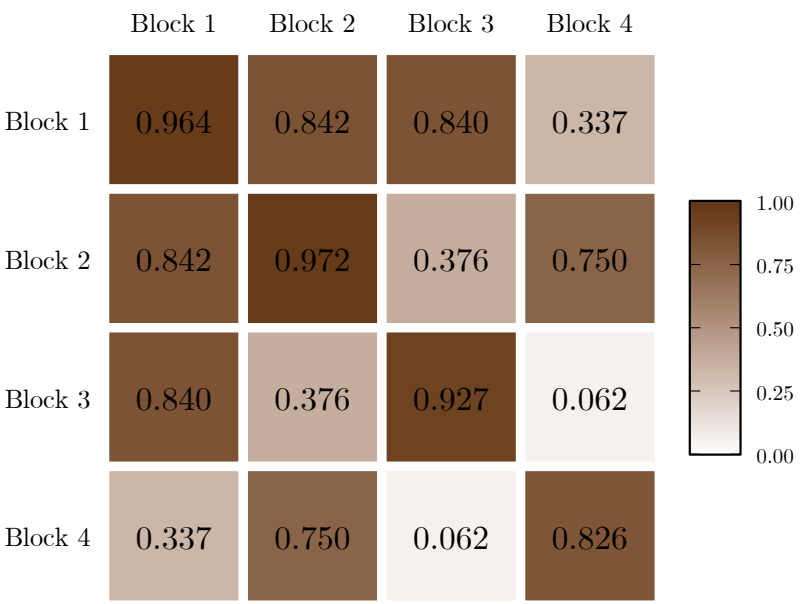
Outline

Application

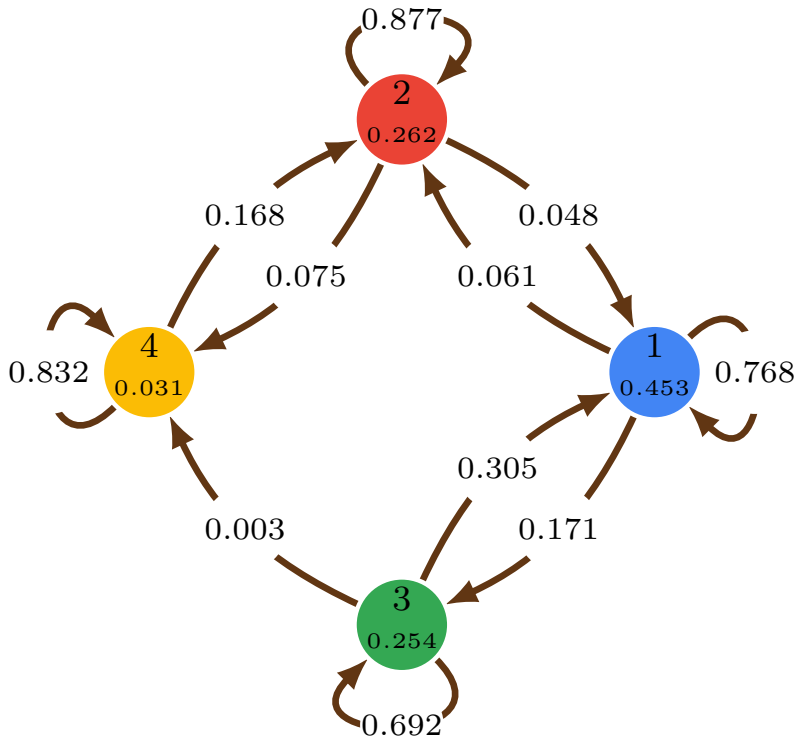
- Data are referred to **daily interactions** within a colony of $n = 164$ ants over a period of 11 days
- An **interaction** is defined as occurring whenever the front end of one ant enters the proximity region of another
- The binary network $\boldsymbol{Y}^{(t)}$ is built such that an edge between nodes i and j is included if **at least one contact** was recorded between the corresponding ants on day t
- **Additional observed information** is also available, including ant age and size, number of visits to the brood, nest entrance, and rubbish piles, as well as number of foraging events
- We estimate a DSBM with $k = 4$ **latent blocks** on the data from the first 10 days ($T = 10$) and use the estimated model parameters to **forecast connections** for day $T + 1 = 11$

Application: estimation results

Figure 1: Summary of the estimated model parameters for the ant colony data under the DSBM with $k = 4$ latent blocks



(a) Estimated connection probabilities ($\hat{\beta}_{uv}$): darker shades indicate higher probabilities



(b) Estimated initial ($\hat{\pi}_u$, within the nodes) and transition ($\hat{\pi}_{v|u}$, along the edges) probabilities

Application: estimation results

- Characterization of the estimated blocks:
  - *Block 1* - **"Cleaners"**: highest number of visits to rubbish piles, hence mainly engaged in waste removal; average age of around 93 days
  - *Block 2* - **"Near-nest foragers"**: highest number of visits to the nest entrance and of foraging events; average age of around 153 days
  - *Block 3* - **"Nurses"**: highest number of visits to the brood; youngest ants, with an average age of around 89 days
  - *Block 4* - **"Distant foragers"**: high number of foraging events, but fewer visits to the nest entrance compared to ants in block 2, likely performing longer trips and bringing back more resources; oldest ants with an average age of around 221 days

Application: link prediction results

Table 2: **Main performance metrics** for prediction of connections among ants during the 11^{th} day for the conditional and unconditional case

	$\hat{p}_{ij}^{(T+1)}$	$\tilde{p}_{ij}^{(T+1)}$
Sensitivity/Recall	0.930	0.998
Specificity	0.439	0.056
Precision	0.794	0.711
Accuracy	0.782	0.715
F1-score	0.856	0.831
AUC	0.761	0.825

Outline

Conclusions

- We consider the dynamic stochastic block model and we focus on **link prediction** to estimate the probability of future interactions in longitudinal networks
- This capability is particularly valuable in **early warning scenarios**, as it enables the detection of high-risk or anomalous connections
- We show a **good predictive accuracy** of the proposed approaches through simulations on synthetic data under different scenarios and an application
- **Future research**:
  - incorporate network-specific features (e.g., node similarities or node-level attributes) into the model formulation
  - adapt the method to address network sparsity
  - improve the choice of the optimal cutoff, considering more advanced approaches such as cross-validation and cost-sensitive methods

Outline

Main references

Bartolucci F, Marino M, Pandolfi S. 2018. Dealing with reciprocity in dynamic stochastic block models. *Computational Statistics & Data Analysis* 123:86–100

Bartolucci F, Pandolfi S. 2020. An exact algorithm for time-dependent variational inference for the dynamic stochastic block model. *Pattern Recognition Letters* 138:362–369

Matias C, Miele V. 2017. Statistical clustering of temporal networks through a dynamic stochastic block model. *Journal of the Royal Statistical Society, Series B* 79:1119–1141

Xu, K. S, Hero A.O. 2014. Dynamic stochastic blockmodels for time-evolving social networks. *IEEE Journal of Selected Topics in Signal Processing* 8:552–562

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

Acknowledgments

Acknowledgment: The authors acknowledge the financial support from the grant "Hidden Markov Models for Early Warning Systems" of Ministero dell'Università e della Ricerca (PRIN 2022TZEXKF) funded by European Union - Next Generation EU, Mission 4, Component 2, CUP J53D23004990006.