Dealing with reciprocity in dynamic stochastic block models

Analisi della reciprocità in modelli dinamici basati su blocchi stocastici

Francesco Bartolucci, Maria Francesca Marino, Silvia Pandolfi

Abstract For directed relations among a set of nodes with a longitudinal structure, we introduce a dynamic stochastic block model where the blocks are represented by a sequence of latent variables following a Markov chain. Dyads are explicitly modeled conditional on the states occupied by both nodes involved in the relation. We mainly focus on reciprocity and propose three different parameterizations in which: (i) reciprocity is allowed to depend on the blocks of the nodes in the dyad; (ii) reciprocity is assumed to be constant across blocks; and (iii) reciprocity is ruled out. Inference on the model parameters is based on a variational approach. An approximate likelihood ratio test statistic based on the variational approximation is also proposed. This allows us to formally test for both the hypothesis of no reciprocity and that of constant reciprocity with respect to the latent blocks. The proposed approach is illustrated by a simulation study and two applications.

Abstract Si propone un modello dinamico a blocchi stocastici per dati relazionali longitudinali. I blocchi sono identificati da una sequenza di variabili latenti distribuite secondo una catena di Markov. Oggetto dell'analisi è ogni singola diade, la cui distribuzione viene modellata condizionatamente ai blocchi di appartenenza di ciascuno dei nodi coinvolti nella relazione. Particolare enfasi viene posta sullo studio della reciprocità tra i nodi, proponendo tre diverse parametrizzazioni in cui: (i) la reciprocità varia al variare dei blocchi di appartenenza dei nodi, (ii) il livello di reciprocità è constante, (iii) la reciprocità è assente. Per fare inferenze sui parametri del modello si propone l'utilizzo di un approccio variazionale, che rappresenta anche la base per lo sviluppo di un test basato sul rapporto di verosimiglianza ap-

Francesco Bartolucci
Department of Economics, University of Perugia, e-mail: francesco.bartolucci@unipg.it

Maria Francesca Marino
Department of Statistics, Applications, Computer Science, University of Florence, e-mail: mariafrancesca.marino@unifi.it

Silvia Pandolfi
Department of Economics, University of Perugia, e-mail: silvia.pandolfi@unipg.it

prossimato che può essere utilizzato per verificare l'ipotesi di assenza di reciprocità o di reciprocità costante rispetto ai blocchi latenti. L'approccio proposto è illustrato tramite uno studio di simulazione e due applicazioni.

Key words: Dyads, EM algorithm, Hidden Markov models, Likelihood ratio test, Variational inference

1 Introduction

Dynamic Stochastic Block Models (SMBs) [5, 6] represent an important tool of analysis in the dynamic social network literature when the focus is on discovering communities and clustering individuals with respect to their social behavior. According to this specification, nodes in the network can be clustered into k distinct blocks, corresponding to the categories of discrete latent variables which evolve over time according to a first-order Markov chain. The probability of observing a connection between two nodes at a given occasion only depends on their block memberships at the same occasion.

Extending the proposal in [6], we develop an SBM for dynamic directed networks, observed in discrete time, in which the main element of analysis is the *dyad* referred to each pair of nodes. Our main assumption is that of conditional independence between the dyads, given the corresponding latent variables, rather than between univariate responses. This leads to a more flexible specification which does not rely on restrictive assumptions about the reciprocity between nodes. To provide a deeper insight into reciprocity effects, we propose to parametrically specify every dyadic relation by means of a conditional log-linear model. This allows us to effectively distinguish between main and reciprocal effects and improve interpretability of the results. Furthermore, the proposed approach allows us to formulate three different hypotheses: (i) reciprocity may depend on the blocks to which the units involved in the relation belong; (ii) reciprocity is constant across blocks; (iii) reciprocity is absent. Inference on the model parameters is pursued via a variational approach based on a lower bound for the intractable likelihood function. This lower bound also allows us to derive an approximate Likelihood Ratio (LR) test for inferential purposes on the reciprocity parameters.

The reminder of this paper is structured as follows. Section 2 describes the standard dynamic SBM and details the proposed dyadic formulation. In Section 3 we describe the variational approach for model inference. The simulation study and applications are outlined in Sections 4 and 5, respectively. For a detailed description of the proposed approach we remind the reader to [1].

2 Dynamic stochastic block models

Let $Y_{ij}^{(t)}, i, j = 1, \ldots, n, j \neq i$, denote a binary response variable which is equal to 1 if there exists an edge from node i to node j at occasion t, with $t = 1, \ldots, T$, and is equal to 0 otherwise; $y_{ij}^{(t)}$ is used to denote a realization of $Y_{ij}^{(t)}$. We focus on directed networks without self-loops, so that $Y_{ij}^{(t)}$ may differ from $Y_{ji}^{(t)}$ and $Y_{ii}^{(t)}$ is not defined. Moreover, let $\mathbf{Y}^{(t)}$ be the binary adjacency matrix recorded at occasion t, which summarizes the relations between nodes at this occasion and let $\mathscr{Y} = \{\mathbf{Y}^{(1)}, \ldots, \mathbf{Y}^{(T)}\}$ be the set of all network snapshots taken across time occasions.

Standard dynamic SBMs [5, 6] assume that network nodes belong to one of k distinct blocks, which are identified by the node- and time-specific latent variables $U_i^{(t)}$. These are defined on the finite support $\{1, \ldots, k\}$ and are assumed to follow a Markov chain with initial probability vector $\boldsymbol{\lambda} = \{\lambda_u,\, u = 1, \ldots, k\}$ and transition probability matrix $\boldsymbol{\Pi} = \{\pi_{u|v},\, u, v = 1, \ldots, k\}$. A further crucial assumption of dynamic SBMs is that of *local independence*: given the latent variables $U_i^{(t)}$ and $U_j^{(t)}$, the responses $Y_{ij}^{(t)}$ are conditionally independent and follow a Bernoulli distribution with success probability only depending on the blocks of the nodes at occasion t.

We extend the previous formulation by relaxing the local independence assumption and directly accounting for reciprocal effects. For this aim we let $\mathbf{D}_{ij}^{(t)} = (Y_{ij}^{(t)}, Y_{ji}^{(t)})'$ denote the random vector corresponding to the dyad involving nodes i and j at occasion t, with $i = 1, \ldots, n-1$, $j = i+1, \ldots, n$, and $t = 1, \ldots, T$. Conditional on $U_i^{(t)} = u_1$ and $U_j^{(t)} = u_2$, we denote the dyad probabilities by

$$\psi_{y_1 y_2 | u_1 u_2} = p(\mathbf{D}_{ij}^{(t)} = \mathbf{d} \mid U_i^{(t)} = u_1, U_j^{(t)} = u_2),$$

with $u_1, u_2 = 1, \ldots, k$, $y_1, y_2 = 0, 1$, and $\mathbf{d} = (y_1, y_2) \in \{(0,0), (0,1), (1,0), (1,1)\}$. To put emphasis on reciprocity, we use the following log-linear parametrization:

$$\psi_{y_1 y_2 | u_1 u_2} \propto \exp\left[\alpha_{u_1 u_2} y_1 + (\alpha_{u_1 u_2} + \beta_{u_1 u_2}) y_2 + \rho_{u_1 u_2} y_1 y_2\right],$$

where $\beta_{uu} = 0$, for $u = 1, \ldots, k$, $\alpha_{u_1 u_2} = \alpha_{u_2 u_1} + \beta_{u_2 u_1}$, $\beta_{u_1 u_2} = -\beta_{u_2 u_1}$, and $\rho_{u_1 u_2} = \rho_{u_2 u_1}$, for all $u_1 \neq u_2$, to ensure identifiability.

Different versions of the proposed model specification may be obtained by imposing constraints on the $\rho_{u_1 u_2}$ parameters. In particular, under the hypothesis

$$H_I : \rho_{u_1 u_2} = 0, \quad u_1, u_2 = 1, \ldots, k, u_1 \leq u_2,$$

the model directly reduces to the standard dynamic SBM in [6], denoted by M_I, and based on the local independence between the responses $Y_{ij}^{(t)}$. Constant reciprocity effects correspond to the following hypothesis leading to model M_C:

$$H_C : \rho_{u_1 u_2} = \rho, \quad u_1, u_2 = 1, \ldots, k, u_1 \leq u_2.$$

The unconstrained model, with free $\rho_{u_1 u_2}$ parameters, will be denoted by M_U.

3 Variational inference

Let $\mathscr{U} = \{U_i^{(t)}, i = 1,\dots,n, t = 1,\dots,T\}$ denote the overall set of latent variables in the model; based on the assumptions introduced so far, the observed network distribution is obtained by marginalizing out all these latent variables from the joint distribution of $\mathscr{Y}$ and $\mathscr{U}$. This would require the evaluation of a sum over $k^{Tn(n-1)/2}$ terms that, therefore, becomes quickly cumbersome as n, the number of nodes in the network, increases. We then rely on a variational approximation of the intractable likelihood function for making inference on the model parameters.

3.1 Parameter estimation

Let $\boldsymbol{\theta}$ denote the vector of all free model parameters. Following the approach suggested in [5] and [6], we estimate model parameters by a Variational Expectation Maximization (VEM) algorithm [3]. Let $p(\mathscr{U} \mid \mathscr{Y})$ denote the posterior distribution of $\mathscr{U}$ given the observed data $\mathscr{Y}$ and let $Q(\mathscr{U})$ denote its approximation. The VEM algorithm maximizes the following lower bound of the log-likelihood function:

$$\begin{aligned}\mathscr{J}(\boldsymbol{\theta}) &= \log p(\mathscr{Y}) - KL[Q(\mathscr{U}) \,||\, p(\mathscr{U} \mid \mathscr{Y})] \\ &= \sum_{\mathscr{U}} Q(\mathscr{U}) \log p(\mathscr{Y}, \mathscr{U}) - \sum_{\mathscr{U}} Q(\mathscr{U}) \log Q(\mathscr{U}),\end{aligned} \tag{1}$$

where $KL[\cdot \,||\, \cdot]$ stands for the Kullback-Leibler distance. In particular, we use the class of approximate distributions assuming conditional independence between the latent variables in the network given the observed data, namely $Q(\mathscr{U}) = \prod_{i=1}^{n} \prod_{t=1}^{T} q(u_i^{(t)}; \boldsymbol{\tau}_i^{(t)})$, where $q(\cdot; \boldsymbol{\tau}_i^{(t)})$ denotes a multinomial probability distribution with parameters 1 and $\boldsymbol{\tau}_i^{(t)} = \{\tau_{iu}^{(t)}, u = 1,\dots,k\}$. Consequently, function $\mathscr{J}(\boldsymbol{\theta})$ defined in (1) can be rewritten as the sum of the following components:

$$\begin{aligned}\mathscr{J}_1(\boldsymbol{\theta}) &= \sum_{i=1}^{n} \sum_{u=1}^{k} \tau_{iu}^{(1)} \log \lambda_u + \sum_{i=1}^{n} \sum_{t=2}^{T} \sum_{u=1}^{k} \sum_{v=1}^{k} \tau_{iu}^{(t-1)} \tau_{iv}^{(t)} \log \pi_{v|u}, \\ \mathscr{J}_2(\boldsymbol{\theta}) &= \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \sum_{t=1}^{T} \sum_{u=1}^{k} \sum_{v=1}^{k} \tau_{iu}^{(t)} \tau_{jv}^{(t)} \log p(y_{ij}^{(t)}, y_{ji}^{(t)} \mid U_i^{(t)} = u, U_j^{(t)} = v), \\ \mathscr{J}_3(\boldsymbol{\theta}) &= \sum_{i=1}^{n} \sum_{t=1}^{T} \sum_{u=1}^{k} \tau_{iu}^{(t)} \log \tau_{iu}^{(t)}.\end{aligned}$$

To obtain parameter estimates, the VEM algorithm alternates two separate steps until convergence: the E-step and the M-step. At the E-step, we maximize $\mathscr{J}(\boldsymbol{\theta})$ with respect to $\boldsymbol{\tau}_i^{(t)}, i = 1,\dots,n, t = 1,\dots,T$, under the constraint that these quantities are non-negative and $\sum_u \tau_{iu}^{(t)} = 1$. In the M-step, we maximize $\mathscr{J}(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$. Closed form solutions are available for the initial and the transition probabilities of the hidden Markov chain:

$$\lambda_u = \frac{\sum_{i=1}^n \tau_{iu}^{(1)}}{n}, \quad \pi_{v|u} = \frac{\sum_{i=1}^n \sum_{t=2}^T \tau_{iu}^{(t-1)} \tau_{iv}^{(t)}}{\sum_{i=1}^n \sum_{t=2}^T \tau_{iu}^{(t-1)}}.$$

The remaining model parameters are estimated by a standard Netwon-Raphson algorithm for log-linear models.

Two further relevant issues concern the selection of the optimal number of blocks k and the clustering of nodes. Regarding the first aspect, we rely on an Integrated Classification Likelihood (ICL) approach [2]; moreover, nodes may be assigned to one of the k blocks according to a maximum-a-posteriori rule based on the estimated parameters of the multinomial distribution $\hat{\boldsymbol{\tau}}_i^{(t)}$.

3.2 Testing for reciprocity

Reciprocity plays a central role when dealing with directed networks. To test for the absence of reciprocity in the network we propose an approximate LR test based on the lower bound of the likelihood function, $\mathscr{J}(\boldsymbol{\theta})$. Let $\hat{\boldsymbol{\theta}}_I$, $\hat{\boldsymbol{\theta}}_C$, and $\hat{\boldsymbol{\theta}}_U$ denote the vectors of parameters estimated under models M_I, M_C, and M_U, respectively, with the first model incorporating hypothesis H_I and the second incorporating hypothesis H_C. The proposed test is based on the statistic

$$R_I = -2\big[\mathscr{J}(\hat{\boldsymbol{\theta}}_I) - \mathscr{J}(\hat{\boldsymbol{\theta}}_U)\big].$$

We compare the observed value of this test statistic against a χ^2 distribution with a number of degrees of freedom equal to the number of free parameters in $\boldsymbol{\rho}$, that is, $k(k+1)/2$. In fact, we consider R_I as an approximation of the LR statistic $-2\big[\ell(\hat{\boldsymbol{\theta}}_I) - \ell(\hat{\boldsymbol{\theta}}_U)\big]$ that, under suitable regularity conditions, has null asymptotic distribution of this type.

For a more detailed analysis we also consider the decomposition $R_I = R_C + R_{CI}$, where

$$R_C = -2\big[\mathscr{J}(\hat{\boldsymbol{\theta}}_C) - \mathscr{J}(\hat{\boldsymbol{\theta}}_U)\big],$$
$$R_{CI} = -2\big[\mathscr{J}(\hat{\boldsymbol{\theta}}_I) - \mathscr{J}(\hat{\boldsymbol{\theta}}_C)\big],$$

with R_C being the approximate LR test statistic for testing the constant reciprocity assumption H_C and R_{CI} begin the approximate LR test statistic for testing H_I against H_C. To perform the test, the first statistic is compared against a χ^2 distribution with $k(k+1)/2-1$ degrees of freedom and the second against a χ^2 distribution with one degree of freedom only.

4 Simulation study

To assess the properties of the approximate LR test statistics under different scenarios, we performed an intensive simulation study. We randomly drew $1,000$ samples from a two state ($k = 2$) dynamic SBM for $n = 20, 50, 100$ units observed at $T = 10$ different time occasions. The initial probability vector $\boldsymbol{\lambda}$ has elements 0.4 and 0.6 and the transition matrix $\boldsymbol{\Pi}$ has diagonal elements equal to 0.7 and 0.8. For the parameterization of the dyad probabilities, we set $\boldsymbol{\alpha} = (-2, -3, -1)'$, $\beta_{12} = 0$, and different values for the reciprocity parameter ranging from -2.5 to 2.5.

To evaluate the performance of the proposed inferential procedure, for each simulated scenario we considered the distribution of the approximate LR test statistics R_I and R_{CI}, which allow us to compare the independence model (M_I) against the unconstrained model (M_U) and the constant reciprocity model (M_C), respectively. Results are reported in Tables 1 and 2. The tables also report the simulated type I error probability/power of the above test statistics.

Table 1 *Mean $(\bar{R}_I)$, variance $(Var(R_I))$, and simulated type I error probability/power of the test statistic R_I (p) under different scenarios.*

	$n = 20$			$n = 50$			$n = 100$		
ρ	$\bar{R}_I$	Var(R_I)	p	$\bar{R}_I$	Var(R_I)	p	$\bar{R}_I$	Var(R_I)	p
-1.50	35.76	200.22	0.993	229.01	1167.33	1.000	922.82	5912.34	1.000
-1.00	21.15	109.27	0.975	129.17	557.34	1.000	523.86	2605.36	1.000
-0.50	7.53	28.92	0.737	42.13	178.39	1.000	167.06	728.66	1.000
-0.25	2.89	9.25	0.272	12.71	49.36	0.922	47.97	192.24	1.000
0.00	1.00	2.02	0.052	0.93	1.83	0.045	1.02	2.07	0.052
0.25	3.15	13.70	0.297	14.68	57.46	0.957	57.44	260.67	1.000
0.50	10.65	46.61	0.861	62.66	275.15	1.000	251.68	1130.22	1.000
1.00	45.71	220.82	1.000	293.87	1574.97	1.000	1180.30	7741.51	1.000
1.50	114.84	598.56	1.000	736.43	4668.70	1.000	2977.89	29305.76	1.000

Results confirm our conjecture that, when simulating data from model M_I, both approximate test statistics have a distribution reasonably close to a χ^2 distribution, leading to the rejection of H_I in about 5% of the simulated samples. On the other hand, under the homogeneity assumption for the reciprocity effects, we observe that the power increases as much as ρ deviates from 0. Moreover, the power of the test increases as the sample size n increases.

We also explored the performance of the proposed method for clustering units across time. For this aim, we evaluated the agreement between the estimated and the true latent structure in terms of adjusted rand index [4], obtaining rather encouraging results in comparison to alternative approaches.

Table 2 *Mean ($\bar{R}_{CI}$), variance ($Var(R_{CI})$), and simulated type I error probability/power of the test statistic R_{CI} (p) under different scenarios.*

	$n=20$			$n=50$			$n=100$		
ρ	$\bar{R}_{CI}$	Var(R_{CI})	p	$\bar{R}_{CI}$	Var(R_{CI})	p	$\bar{R}_{CI}$	Var(R_{CI})	p
-1.50	38.11	200.09	0.988	231.12	1177.18	1.000	924.85	5930.55	1.000
-1.00	23.67	117.15	0.959	131.21	563.74	1.000	525.90	2617.52	1.000
-0.50	10.21	37.30	0.601	44.10	182.99	1.000	169.01	733.51	1.000
-0.25	5.42	14.42	0.227	14.77	54.87	0.825	49.89	194.19	1.000
0.00	3.51	7.52	0.078	3.00	6.04	0.055	2.94	6.73	0.051
0.25	5.76	19.62	0.244	16.55	60.25	0.883	59.50	268.90	1.000
0.50	13.06	53.14	0.743	64.70	281.35	1.000	253.65	1128.36	1.000
1.00	47.91	224.45	1.000	296.14	1623.78	1.000	1182.33	7749.08	1.000
1.50	116.90	603.89	1.000	738.37	4665.33	1.000	2979.98	29281.81	1.000

5 Empirical applications

5.1 Newcomb Fraternity network

The network at issue consists of 14 network snapshots on preference rankings (coded from 1 to 16) from 17 students. Data were collected longitudinally over 15 weeks between 1953 and 1956 among students living in an off-campus (fraternity) house at the University of Michigan. For the purpose of the analysis, we considered the binary socio-matrices $\mathbf{Y}^{(t)}$ derived from these data that are freely available as part of the R package `networkDynamic`. In each network snapshot, $Y_{ij}^{(t)} = 1$ if student i states a ranking for student j equal to 8 or less at time occasion t.

For these data, we estimated the proposed dynamic SBM with $k = 1,\dots,5$, considering the different model specifications corresponding to different hypotheses on reciprocity. The ICL criterion leads to selecting $k = 3$ latent blocks, regardless the chosen model specification. This criterion also identified M_C as the optimal model specification.

Based on the LR test statistic with $k = 3$, we observe that R_I is statistically significant and, therefore, leads to prefer M_U to M_I. A significant test statistic is also observed when comparing M_I against M_C, again with a p-value smaller than 0.001. On the other hand, we conclude that the assumption of constant reciprocity, H_C, cannot be rejected based on the observed data because $p(\chi^2_5 > R_C) = 0.102$, confirming the result based on the comparison of the ICL values.

The parameter estimates suggest the presence of significant mutual relations between students, irrespective to the cluster they belong to ($\hat{\rho} = 1.044$). Regarding the remaining parameters, we observe that students in block 1 are likely to declare a non-reciprocated friendship with nodes belonging to the same block ($\hat{\alpha}_{11} = 1.203$), while null within-group relations are mainly observed for students belonging to block 2 ($\hat{\alpha}_{22} = -1.069$). A non-significant value is observed for α_{33}. Regarding the estimated initial and transition probabilities of the hidden Markov chain, cluster 2 is the most likely at the beginning of the observation period ($\hat{\lambda}_2 = 0.48$). Moreover, estimated transitions show quite persistent hidden states.

5.2 *Enron email network*

The second example is based on a dynamic network derived from the Enron corpus, consisting of a large set of email messages that was made public during the legal investigation concerning the Enron corporation. The dataset concerns 184 Enron employees; we considered communications recorded between April 2001 and March 2002 and we built an email network for each month, so that the dynamic network has 12 time points. In this application, $Y_{ij}^{(t)} = 1$ if user i sent at least one email message to user j during the t-th month of the analyzed time window, with $i = 1, \ldots, 183$, $j = i+1, \ldots, 184$, and $t = 1, \ldots, 12$.

We estimated a dynamic SBM with a varying number of blocks ($k = 1, \ldots, 7$). ICL values lead to selecting a model with $k = 6$ hidden states for all considered parameterizations. Based on the same index, we selected the unconstrained model M_U, with reciprocity parameters depending on the latent blocks. Even in this case, we may validate the results by comparing the values of the approximate LR statistics. From this comparison, when $k = 6$, we observe that the hypothesis of absence of reciprocity, H_I, is strongly rejected by both tests based on R_I and R_{CI}. Moreover, the observed value of the test statistic R_C allows us to confirm that the unconstrained model has to be preferred to the other model specifications, due to a very low p-value. Accordingly, in this application, we conclude that reciprocal relations are statistically significant, and that they depend on the latent blocks of the nodes.

References

1. Bartolucci, F., Marino, F., Pandolfi, S.: Dealing with reciprocity in dynamic stochastic block models. Comput. Stat. Data An. **123**, 86–100 (2018)
2. Biernacki, C., Celeux, G., Govaert, G.: Assessing a mixture model for clustering with the integrated completed likelihood. IEEE T. Pattern Anal., **22** 719–725 (2000)
3. Daudin, J.-J., Picard, F., Robin, S.: A mixture model for random graphs. Stat. Comput. **18**, 173–183 (2008)
4. Hubert, L., Arabie, P.: Comparing partitions. J. Classif. **2**, 193–218 (1985)
5. Matias, C., Miele, V.: Statistical clustering of temporal networks through a dynamic stochastic block model. J. R. Stat. Soc. B **79**, 1119–1141 (2017)
6. Yang, T., Chi, Y., Zhu, S., Gong, Y., Jin, R.: Detecting communities and their evolutions in dynamic social networks - a Bayesian approach. Mach. Learn. **82**, 157–189 (2011)