A Dynamic Stochastic Block Model with infinite communities

Un modello dinamico con blocchi aleatori e numero infinito di comunità

Roberto Casarin and Ovielt Baltodano López

Abstract This contribution proposes the use of bayesian non–parametric techniques to make inference on the number of communities in a Dynamic Stochastic Block Model which is then applied to real network data on international financial flows.
Abstract *Questo contributo si propone l'uso di metodi bayesiani non–parametrici per fare inferenze sul numero di comunità in un modello dinamico con blocchi aleatori, il quale dopo viene applicato alla rete di flussi finanziari internazionali.*

Key words: Stochastic block models, bayesian non–parametric methods.

1 Introduction

The increase of network data, e.g. online social networks, has shown the importance of clustering and community structures. In this sense, a Dynamic Stochastic Block Model (DSBM) allows to capture heterogeneous relationships between nodes and potential role changes in their interaction. [9, 6] proposed the use of Hidden Markov chains in order to extent the mixture distribution used in a static setting. However, there is no inference and therefore no measure of uncertainty on the number of communities.

On the other hand, in the field of time series analysis, the use of Hidden Markov chains with infinite states has a long tradition. An important extension was proposed by [4]. In their contribution, they introduce state persistence in a Hierarchical Dirichlet process framework used in a hidden Markov chain model. In a nonlinear context, [2] applies the same strategy to a Generalized Auto-Regressive Conditional heteroskedasticity (GARCH) model. In this paper, we combine this persistent

Roberto Casarin
University Ca' Foscari of Venice e-mail: r.casarin@unive.it

Ovielt Baltodano López
University Ca' Foscari of Venice e-mail: ovielt.baltodano@unive.it

Hierarchical Dirichlet process and hidden Markov chain with infinite states to the DSBM. This is in line with [3], but their contribution was centered on a mixed–membership setting, that is, each node can play different roles at the same time, while here we assume each node can be in only one community at each point in time. The description of our model is presented in Section 2. Moreover, in Section 3, we use empirical data on the bilateral financial flows between countries given its relevance for financial stability and interdependence, and to exemplify the use of the DSBM with infinite communities.

2 A DSBM with infinite communities

A weighted graph can be defined as the ordered triplet $\mathscr{G} = (\mathscr{V}, \mathscr{E}, Y)$, where $\mathscr{V} = \{1, \ldots, N\}$ is the set of nodes, $\mathscr{E} \subset \mathscr{V} \times \mathscr{V}$ and Y is a weight matrix, $Y \in \mathbb{R}^N \times \mathbb{R}^N$. The (i,j)–th element of Y is $Y_{ij} = 0$ if $(i,j) \notin \mathscr{E}$ and $Y_{ij} = a \in \mathbb{R}\backslash\{0\}$ if $(i,j) \in \mathscr{E}$. We define the sequence of sets $\mathfrak{V} = \mathscr{V}_1, \ldots, \mathscr{V}_Q$, with $Q \in \{1, 2, \ldots\}$ a partition of $\mathscr{V}$, if each element $\mathscr{V}_j \subset \mathscr{V}$ (called block or community in what follows) satisfies: $\mathscr{V}_i \cap \mathscr{V}_j = \emptyset$ and $\mathscr{V}_1 \cup \ldots \cup \mathscr{V}_Q = \mathscr{V}$.

In this paper we assume a sequence of graphs $\mathscr{G}_{1:T} = \{\mathscr{G}_t, t = 1, \ldots, T\}$ is available and a latent sequence of partitions $\mathfrak{V}_{1:T} = \{\mathfrak{V}_t, t = 1, \ldots, T\}$ drives the topology of the graph. Following [9] and [6], the partition sequence $\mathfrak{V}_{1:T}$ is induced by a set of N hidden Markov chain processes. The membership of $i \in \mathscr{V}$ is captured by $Z_i = \{Z_{it}, t = 1, \ldots, T\}$, which evolves following a Markov Chain process with transition matrix P, where entry $q, r \in \mathscr{Q}$ is given by $P_{qr} \in (0,1)$ and each row $P_{q\cdot}$ sum up to one. At time t, the node i belongs to the block $\mathscr{V}_q$ if $Z_{it} = q$. Although the chains are independent, they share the same transition matrix, thus P gives information on the level of persistence of the communities as a whole.

The node partition induces edge clusters with different existence probabilities and weights. Further, we assume that the contemporaneous network Y_t given $Z_{1:T}$ and $Y_{1:T}$ only depends on $Z_t = \{Z_{1t}, \ldots, Z_{Nt}\}$ and each entry of the adjacency matrix is distributed as

$$Y_{ijt} \left| Z_{it} = q, Z_{jt} = r, \theta_{ijt} \right. \sim (1 - \nu_{qr})\delta(y) + \nu_{qr} f(y|\lambda_{qr}) \tag{1}$$

which is a zero-inflated distribution family, where $\delta(\cdot)$ denotes the Dirac function at zero and $f(\cdot|\lambda_{qr})$ is a probability density function with parameter λ_{qr} and support set $\mathbb{R}\backslash\{0\}$. The community structure is used to allow for partial parameter pooling, that is the edge parameters $\theta_{ijt} = (\nu_{ijt}, \lambda_{ijt})' = \theta^*_{qr}$ if $i \in \mathscr{V}_q$ and $j \in \mathscr{V}_r$ at time t.

Usually, the number of communities is given and the choice depends on some specific criteria. For instance, [6] chooses the model with the highest integrated classification likelihood criterion, after fitting models with different $\mathscr{Q}$ cardinality. In order to infer the number of communities, a Bayesian non–parametric framework can be applied, which to allow for infinite states Markov chains. Since the number of state is infinite, i.e. $\mathscr{Q} = \{1, 2, \ldots\}$, the transition matrix P becomes infinite di-

mensional and a parsimonious model is needed for P, which preserves the labelling of the communities in the different rows. [7] proposed a hierarchical Dirichlet Process (DP) to tie the different rows of P by providing the same centering measure for each row, that is

$$\begin{aligned} Z_{it}|Z_{it-1} = q &\sim G_q,\ q \in \mathscr{Q} \\ G_q|\omega, G_0 &\sim \mathrm{DP}(\omega, G_0) \\ G_0|\eta, H &\sim \mathrm{DP}(\eta, H), \end{aligned} \tag{2}$$

where $\mathrm{DP}(\alpha, H)$ denotes a Dirichlet process with precision parameter α and centering (or base) measure H. Nevertheless, [4] underline the fact that (2) does not differentiate between the main diagonal of P and the transition across different groups, essentially affecting the state persistence. Therefore, using the extension proposed by [4] for the analysis of time–series, (2) can be extended in line with the Chinese restaurant franchise with loyal costumer,

$$\begin{aligned} Z_{it}|P_{q\cdot}, Z_{it-1} = q &\sim \mathrm{Ca}\left(P_{q\cdot}\right) \\ P_{q\cdot}|\omega, \pi &\sim \mathrm{DP}\left(\omega + \kappa, \frac{\omega\pi + \delta(r-q)}{\omega + \kappa}\right) \\ \pi &\sim \mathrm{Stick}(\eta) \\ \theta^*_{qr} &\sim \mathrm{H} \end{aligned} \tag{3}$$

where $\mathrm{Ca}(p)$ denotes a categorical (or multinoulli) distribution with probability parameter p. The parameters $P_{q\cdot}$, π are the weights of the stick-breaking representation of G_q and G_0, and κ is a parameter increasing the self-transition probability $P_{qq},\ q \in \mathscr{Q}$. The $q-th$ element of the infinite vector π is given by $\pi_q = \xi_q \prod_{l=1}^{q-1}(1-\xi_l)$ and $\xi_l \sim \mathrm{Beta}(1, \eta)$. The Fig. 1 summarizes the structure of the DSBM with infinite communities, at each point time the allocation variables $Z_i t$ and $Z_j t$ of the corresponding pair (i, j) determines which parameters θ^*_{qr} applies. Their membership changes on the basis of the infinite dimension P, whose rows have a specific Dirichlet process under the same centering measure.

In the case of a weighted network whose active edges have a support $\mathbb{R}\backslash\{0\}$, a zero-inflated normal can be used in (1) with parameters $\lambda_{qr} = (\beta_{qr}, \sigma^2_{qr})'$. Furthermore, (1) can be rewritten as

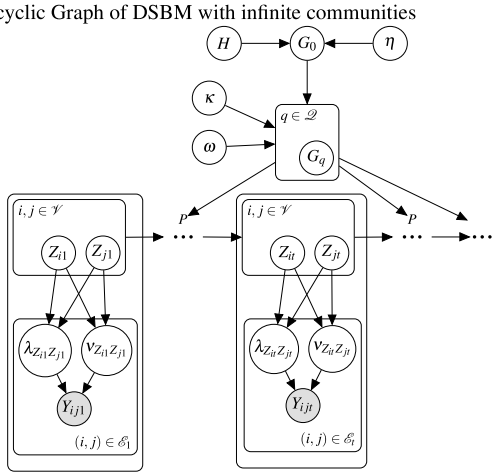
$$Y_{ijt}\left|D_{ijt}, Z_{it} = q, Z_{jt} = r, \theta_{ijt}\right. \sim \begin{cases} \delta(y) & \text{if } D_{ijt} = 0 \\ f\left(y|\lambda_{qr}\right) & \text{if } D_{ijt} = 1 \end{cases} \tag{4}$$

where D_{ijt} is an observable indicator variable such that $D_{ijt} = 1$ if $(i, j) \in \mathscr{E}_t$ and $D_{ijt} = 0$ if $(i, j) \notin \mathscr{E}_t$,

$$D_{ijt}\left|Z_{it} = q, Z_{jt} = r, \theta_{ijt}\right. \sim \mathrm{Bern}\left(\nu_{qr}\right) \tag{5}$$

Under this representation, a full Gibbs sampling procedure can be derived after using a set of slice sampling auxiliary variables u_{it} applied to the stick–breaking

Fig. 1 Directed Acyclic Graph of DSBM with infinite communities



representation in (3). The main full conditional posteriors are presented in Table 1 where the inference also covers the hyperparameters η, κ, ω.[1] The closed form of the parameters of the full conditional posteriors and further details, such as the auxiliary variables $\bar{m}_{\cdot q}$ and g, are standard in the literature [e.g. 4, 2]. Additionally, the allocation variables Z are sampled from Forward filtering backward sampling [5].

Table 1 Gibbs sampling

Prior	Full Conditional Posterior
$\beta_{qr} \sim \mathrm{N}\left(\underline{\beta}_{qr}, \underline{\Sigma}_{qr}\right)$	$\beta_{qr}\|Y,\ldots \sim \mathrm{N}\left(\overline{\beta}_{qr}, \overline{\Sigma}_{qr}\right)$
$\sigma^2_{qr} \sim \mathrm{IG}\left(\underline{d}_{qr}/2, \underline{e}_{qr}/2\right)$	$\sigma^2_{qr}\|Y,\ldots \sim \mathrm{IG}\left(\overline{d}_{qr}/2, \overline{e}_{qr}/2\right)$
$\nu_{qr} \sim \mathrm{Beta}\left(\underline{b}_{qr}, \underline{c}_{qr}\right)$	$\nu_{qr}\|Y,\ldots \sim \mathrm{Beta}\left(\overline{b}_{qr}, \overline{c}_{qr}\right)$
$P_{q\cdot} \sim \mathrm{DP}\left(\omega+\kappa, \frac{\omega\pi+\delta(r-q)}{\omega+\kappa}\right)$	$P_{q\cdot}\|Y,\ldots \sim \mathrm{Dir}\left(\omega\pi_1 + n_{q\cdot,}, \ldots, \omega\pi_q + \kappa + n_{qq}, \ldots, \omega\pi_{Q+1}\right)$
$\pi \sim \mathrm{Stick}(\eta)$	$\pi\|Y,\ldots \sim \mathrm{Dir}\left(\bar{m}_{\cdot 1}, \ldots, \bar{m}_{\cdot Q}, \eta\right)$
$u_{it} \sim \mathrm{Uni}(0,1)$	$u_{it}\|Y,\ldots \sim \mathrm{Uni}\left(0, k^{\mathbb{I}(Z_{it-1}=Z_{it})} P_{Z_{it-1}Z_{it}}\right)$
$\omega+\kappa \sim \mathrm{G}\left(\zeta_1, \zeta_2\right)$	$\omega+\kappa\|Y,\ldots \sim \mathrm{G}\left(\zeta_1 + m - s, \left(1/\zeta_2 - \sum_{q=1}^{Q} \log k_q\right)^{-1}\right)$
$\rho \sim \mathrm{Beta}\left(\chi_1, \chi_2\right)$	$\rho\|Y,\ldots \sim \mathrm{Beta}\left(\chi_1 + g, \chi_2 + m - g\right)$
$\eta \sim \mathrm{G}(\psi_1, \psi_2)$	$\eta\|Y,\ldots \sim \mathrm{G}\left(\psi_1 + \bar{Q} - s, (1/\psi_2 - \log k)^{-1}\right)$

[1] In the case of κ and ω, the prior is set on $\kappa+\omega$ and $\rho = \kappa/(\kappa+\omega)$.

3 Application

The financial flows at international level have experienced significant changes in the last decades including the 2008 crisis [e.g., 8]. The DSBM with infinite communities can identify specific network structures and its evolution, which can have potential consequences in terms of contagion and financial stability. In this sense, the following is an application of the model described in Section 2 to the data collected by Bank for international Settlements (BIS) on bilateral cross–border claims (and liabilities). As in [1], given that the data presented by the BIS is in a bank–country format, that is a banking system reports its position with respect to a country, we transform the data to a country–country format for most of the cases using data triangulation, with this the flows comprise other sectors and the missing data are minimized. The resulting network includes 31 countries for the period 2001–2019.[2]

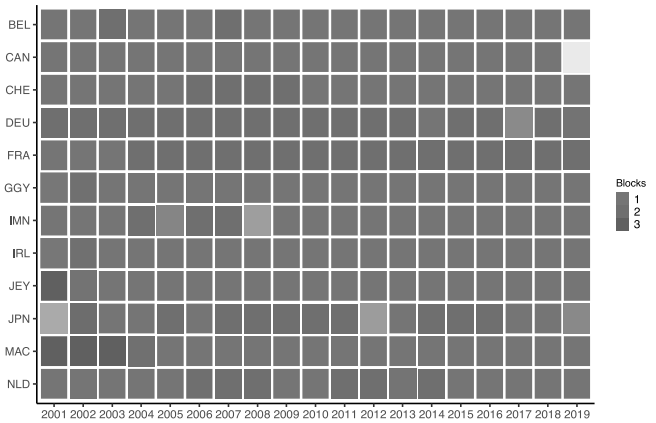
The main results are showed in Fig. 2 and Table 2. Regarding the number of communities, the 53% of draws result in three communities, but still there is some uncertainty given the relative frequency of four communities. Using the former number, 38% of countries have experience at least one change of membership in the period 2001-2019. These countries are presented in Fig.2. Although state persistence is high, there are sudden changes in JPN, NDL, DEU and BEL. Other countries, such as US and GBR are not in the figure because they remain in the same community. In the case of CAN, it seems stable in terms of posterior mode, but in 2019 it starts a transition to another community represented with lower posterior probability.

Table 2 Relative frequency of the number of communities in the network of financial flows

Q	2	3	4	5	6	7	8	9
	0.12	52.70	32.97	9.70	3.28	1.00	0.20	0.01

[2] This sample covers only reporting countries, no destinations, a subset of the countries available. The countries (dependencies or relevant regions) are: Austria, Australia, Belgium, Brazil, Canada, Switzerland, Chile, Germany, Denmark, Spain, Finland, France, United Kingdom, Guernsey, Greece, Hong Kong SAR China, Ireland, Isle of Man, Italy, Jersey, Japan, South Korea, Luxembourg, Macao SAR China, Mexico, Netherlands, Philippines, Sweden, Taiwan, United States and South Africa.

Fig. 2 Countries' membership by year, only the countries which experience a change of membership are included (color intensity is proportional to the posterior probability of the posterior mode)



References

[1] Brei, M., von Peter, G.: The distance effect in banking and trade. Journal of International Money and Finance **81**, 116–137 (2018)

[2] Dufays, A.: Infinite-state Markov-switching for dynamic volatility. Journal of Financial Econometrics **14**(2), 418–460 (2016)

[3] Fan, X., Cao, L., Da Xu, R.Y.: Dynamic infinite mixed-membership stochastic blockmodel. IEEE Transactions on Neural Networks and learning systems **26**(9), 2072–2085 (2014)

[4] Fox, E.B., Sudderth, E.B., Jordan, M.I., Willsky, A.S.: A sticky HDP-HMM with application to speaker diarization. The Annals of Applied Statistics pp. 1020–1056 (2011)

[5] Kim, C., Nelson, C.R.: State-space models with regime switching: classical and Gibbs-sampling approaches with applications, vol. 1. The MIT press (1999)

[6] Matias, C., Miele, V.: Statistical clustering of temporal networks through a dynamic stochastic block model. Journal of the Royal Statistical Society: Series B (Statistical Methodology) **79**(4), 1119–1141 (2017)

[7] Teh, Y.W., Jordan, M.I., Beal, M.J., Blei, D.M.: Hierarchical Dirichlet processes. Journal of the American Statistical Association **101**(476), 1566–1581 (2006)

[8] Tonzer, L.: Cross-border interbank networks, banking risk and contagion. Journal of Financial Stability **18**, 19–32 (2015)

[9] Yang, T., Chi, Y., Zhu, S., Gong, Y., Jin, R.: Detecting communities and their evolutions in dynamic social networks—a bayesian approach. Machine learning **82**(2), 157–189 (2011)