Sankhyā : The Indian Journal of Statistics
2022, Volume 84-A, Part 1, pp. 108-126

Bayesian Testing for Exogenous Partition Structures in Stochastic Block Models

Sirio Legramanti
Bocconi University, Milano, Italy
Tommaso Rigon
Duke University, Durham, USA
Daniele Durante
Bocconi University, Milano, Italy

Abstract

Network data often exhibit block structures characterized by clusters of nodes with similar patterns of edge formation. When such relational data are complemented by additional information on exogenous node partitions, these sources of knowledge are typically included in the model to supervise the cluster assignment mechanism or to improve inference on edge probabilities. Although these solutions are routinely implemented, there is a lack of formal approaches to test if a given external node partition is in line with the endogenous clustering structure encoding stochastic equivalence patterns among the nodes in the network. To fill this gap, we develop a formal Bayesian testing procedure which relies on the calculation of the Bayes factor between a stochastic block model with known grouping structure defined by the exogenous node partition and an infinite relational model that allows the endogenous clustering configurations to be unknown, random and fully revealed by the block–connectivity patterns in the network. A simple Markov chain Monte Carlo method for computing the Bayes factor and quantifying uncertainty in the endogenous groups is proposed. This strategy is evaluated in simulations, and in applications studying brain networks of Alzheimer's patients.

AMS (2000) *subject classification.* Primary 62-XX; Secondary 62F15.
Keywords and phrases. Bayes factor, Brain network, Chinese restaurant process, Infinite relational model, Stochastic equivalence

1 Introduction

There is an extensive interest in learning grouping structures among the nodes in a network (see, e.g. Fortunato and Hric, 2016). Classical solutions to this problem focus on detecting community patterns via algorithmic approaches that cluster the nodes into groups characterized by a high number of edges within each community and comparatively few edges between the nodes in different communities (Newman and Girvan, 2004; Blondel et al., 2008; Fortunato, 2010). Despite being routinely implemented, these procedures do not rely on generative probabilistic models and, therefore, face difficulties when the focus is not just on point estimation, but also on hypothesis testing and uncertainty quantification. This issue has motivated several efforts towards developing model–based representations for inference on grouping structures, with the stochastic block model (sbm) (Holland et al., 1983; Nowicki and Snijders, 2001) providing the most notable contribution within this class. Such a statistical model expresses the edge probabilities as a function of the node assignments to groups and of block probabilities among such groups, thus allowing inference on more general block–connectivity patterns beyond classical community structures. The success of sbms in different fields has motivated various extensions (e.g. Kemp et al., 2006; Airoldi et al., 2008; Karrer and Newman, 2011; Geng et al., 2019) and detailed theoretical studies on their asymptotic properties (e.g. Zhao et al., 2012; Gao et al., 2018; van der Pas and van der Vaart, 2018; Ghosh et al., 2020); see Schmidt and Morup (2013), Abbe (2017), & Lee and Wilkinson (2019) and the references therein for a comprehensive overview.

When node–specific attributes are available, the above block models have been generalized in different directions to incorporate such external information in the edge formation mechanism. Common proposals address this goal via the inclusion of nodal attributes within the generative model for the cluster assignments (e.g. Tallberg, 2004; White and Murphy, 2016; Newman and Clauset, 2016; Stanley et al., 2019), or by defining the edge probabilities as a function of block–specific parameters, as in classical sbms, and of pairwise similarity measures among node attributes (e.g. Mariadassou et al. 2010; Choi et al. 2012; Sweet 2015; Roy et al. 2019). Such formulations are powerful approaches to assist the cluster assignment mechanism and, typically, improve the estimation of the edge probabilities. However, when categorical node attributes are available, less attention has been paid to the development of formal Bayesian testing procedures to assess whether the exogenous partitions identified by the labels of the categorical node attributes are in line with the endogenous grouping structure revealed by the block–connectivity behaviors in the network. For example, in structural brain network applications it is often of interest to understand if exogenous anatomical partitions of the brain regions can accurately characterize the endogenous block structures of brain networks (e.g. Sporns, 2013; Faskowitz et al. 2018). This goal could be partially addressed by the aforementioned models via inference on the posterior distribution for the parameters regulating the effect of the node–specific attributes, but these formulations are prone to identifiability and computational issues.

Motivated by the above discussion, we propose a formal and simple Bayesian testing procedure to compare a stochastic block model with known grouping structure, fixed according to a given exogenous node partition, and an infinite relational model (Kemp et al., 2006) where the node assignments are unknown, random and modeled through a Chinese Restaurant Process (CRP) prior (Aldous, 1985), which allows the total number of non–empty clusters H to be inferred. Such a Bayesian nonparametric representation allows flexible learning of the endogenous clustering configurations as revealed by the common connectivity behaviors within the network and, hence, provides a suitable reference model against which to assess the ability of a pre–specified exogenous partition to characterize the block–connectivity structures within the network. In a sense, our goal is related to those of Bianconi et al. (2009) and Peel et al. (2017). However, such contributions compute, under a frequentist perspective, the entropy of a stochastic block model whose groups coincide with the external node partition, and compare it with the distribution of the entropies derived under the same network with grouping structure given by a random permutation of the exogenous node labels. Besides taking a Bayesian approach to inference, our procedure quantifies proximities to endogenous block structures rather than studying departures from a random partition. This allows, as a byproduct, inference on node groupings supported by the data. In fact, leveraging the recent inference methods for Bayesian clustering (Wade and Ghahramani, 2018) brought into the network field by Legramanti et al. (2020), we complement the results of the proposed testing procedures with an analysis of the credible balls for the grouping structure under the infinite relational model.

In Section 2 we describe the proposed testing procedure, based on the calculation of the Bayes factor (e.g. Kass and Raftery, 1995) among the two competing models, and discuss methods for uncertainty quantification on the inferred endogenous clustering. In Section 3, we derive a collapsed Gibbs sampler to obtain samples from the posterior of the endogenous partition, thus allowing Monte Carlo estimation of the marginal likelihood (Newton and Raftery, 1994; Raftery et al., 2007) required to compute the Bayes factor. As illustrated in simulations in Section 4 and in an application to Alzheimer's brain networks in Section 5, the Gibbs sampler is also useful to perform inference on the endogenous groups. Codes to implement the proposed methods can be found at https://github.com/danieledurante/TESTsbm.

2 Model Formulation, Bayesian Testing and Inference

2.1. Endogenous and Exogenous Models Let $\mathbf{Y}$ denote the $n \times n$ symmetric adjacency matrix associated with an undirected binary network without self–loops, so that $y_{vu} = y_{uv} = 1$ if nodes $v = 2, \ldots, n$ and $u = 1, \ldots, v-1$ are connected, and $y_{vu} = y_{uv} = 0$ otherwise. The absence of self–loops implies that the diagonal entries of $\mathbf{Y}$ are not considered for inference. Recalling our discussion in Section 1, we consider a stochastic representation partitioning the nodes into exhaustive and non–overlapping groups, where nodes in the same group are characterized by equal patterns of edge formation. More specifically, let $\mathbf{z} = (z_1, \ldots, z_n)^\intercal \in \mathcal{Z}$ be the vector of cluster membership indicators for the n nodes, with $\mathcal{Z}$ being the space of all the possible group assignments, so that $z_v = h$ if and only if the vth node belongs to the hth cluster. Letting H be the number of non–empty groups in $\mathbf{z}$, we denote with $\boldsymbol{\Theta}$ the $H \times H$ symmetric matrix of block probabilities with generic elements $\theta_{hk} \in (0, 1)$ indexing the distribution of the edges between the nodes in cluster h and those in cluster k. To characterize block–connectivity structures within the network, we assume

$$(y_{vu} \mid z_v = h, z_u = k, \theta_{hk}) \sim \text{Bern}(\theta_{hk}),$$

independently for each $v = 2, \ldots, n$ and $u = 1, \ldots, v-1$, with $\theta_{hk} \sim \text{Beta}(a, b)$, independently for every $h = 1, \ldots, H$ and $k = 1, \ldots, h$. This formulation recalls the classical Bayesian SBM specification (Nowicki and Snijders, 2001) and leverages a stochastic equivalence property that relies on the conditional independence of the edges, whose distribution depends on the cluster membership of the associated nodes. Indeed, by marginalizing out the beta–distributed block probabilities which are typically treated as nuisance parameters in the SBM (e.g. Kemp et al. 2006; Schmidt and Morup 2013), the likelihood for $\mathbf{Y}$ given $\mathbf{z}$ is

$$p(\mathbf{Y} \mid \mathbf{z}) = \prod_{h=1}^{H} \prod_{k=1}^{h} \frac{\mathrm{B}(a + m_{hk}, b + \bar{m}_{hk})}{\mathrm{B}(a, b)}, \tag{2.1}$$

where m_{hk} and $\bar{m}_{hk}$ denote the number of edges and non–edges among nodes in clusters h and k, respectively, whereas $\mathrm{B}(\cdot, \cdot)$ is the beta function. Expression (2.1) is derived by exploiting beta–binomial conjugacy, and, as we will clarify later in the article, is fundamental to compute Bayes factors and to develop a collapsed Gibbs sampler which updates only the endogenous cluster assignments while treating the block probabilities as nuisance parameters. Moreover, as is clear from Eq. 2.1, $p(\mathbf{Y} \mid \mathbf{z})$ is invariant under relabeling of the cluster indicators. Therefore $p(\mathbf{Y} \mid \mathbf{z})$ is equal to $p(\mathbf{Y} \mid \tilde{\mathbf{z}})$ for any relabeling $\tilde{\mathbf{z}}$ of $\mathbf{z}$, meaning that also the Bayes factors computed from these quantities are invariant under relabeling. Hence, in the rest of the paper, $\mathbf{z}$ will denote any element of the equivalence class of its relabelings, whereas $\mathcal{Z}$ will correspond to the space of all the partitions of $\{1, \ldots, n\}$.

Recalling Section 1, our goal is develop a formal Bayesian test to assess whether assuming $\mathbf{z}$ as known and equal to an exogenous assignment vector $\mathbf{z}^*$ produces an effective characterization of all the block structures in $\mathbf{Y}$, relative to what would be obtained by letting $\mathbf{z}$ unknown, random and endogenously determined by the stochastic equivalence relations in $\mathbf{Y}$. The first hypothesized model $\mathcal{M}^*$ can be naturally represented via a SBM as in Eq. 2.1 with a fixed and known exogenous partition $\mathbf{z}^*$, whereas the second model $\mathcal{M}$ requires a flexible prior distribution for the indicators $\mathbf{z}$ in Eq. 2.1 which is able to reveal the endogenous grouping structure induced by the block–connectivity patterns in $\mathbf{Y}$, without imposing strong parametric constraints. A natural option would be to consider a Dirichlet–multinomial prior as in classical SBMs (Nowicki and Snijders, 2001), but such a specification requires the choice of the total number of groups, which is typically unknown. This issue is usually circumvented by relying on BIC metrics that require estimation of multiple SBMs (e.g. Saldana et al., 2017). To avoid these computational costs and increase flexibility, we rely on a Bayesian nonparametric solution that induces a full–support prior on the total number H of non–empty groups in $\mathbf{z}$. This enables learning of H, which is not guaranteed to coincide with the number H^* of non–empty groups in $\mathbf{z}^*$. A widely used prior in the context of SBMs is the CRP (Aldous, 1985), which leads to the so–called infinite relational model (Kemp et al., 2006; Schmidt and Morup, 2013). Under such a prior, each group attracts new nodes in proportion to its size, and the formation of new groups depends only on the size of the network and on a tuning parameter $\alpha > 0$. More specifically, under model $\mathcal{M}$, we assume the following prior over cluster indicators for the vth node, given the memberships $\mathbf{z}_{-v} = (z_1, \ldots, z_{v-1}, z_{v+1}, \ldots, z_n)^\intercal$ of the others

$$\mathrm{pr}(z_v = h \mid \mathbf{z}_{-v}) = \begin{cases} \frac{n_{h,-v}}{n-1+\alpha} & \text{if} \quad h = 1, \ldots, H_{-v}, \\ \frac{\alpha}{n-1+\alpha} & \text{if} \quad h = H_{-v} + 1. \end{cases} \tag{2.2}$$

In Eq. 2.2, H_{-v} is the number of non–empty groups in $\mathbf{z}_{-v}$, the integer $n_{h,-v}$ is the total number of nodes in cluster h, excluding node v, whereas $\alpha > 0$ denotes a concentration parameter controlling the expected number of non–empty clusters. The urn representation in Eq. 2.2 is induced by the joint probability mass function $p(\mathbf{z}) = \alpha^H [\prod_{h=1}^H (n_h - 1)!][\prod_{v=1}^n (v - 1 + \alpha)]^{-1}$, which shows that the CRP is exchangeable. See also Gershman and Blei (2012) for an overview of CRP priors.

2.2. Bayesian Testing To compare the ability of the endogenous ($\mathcal{M}$) and exogenous ($\mathcal{M}^*$) formulations in characterizing the block structures in $\mathbf{Y}$, we define a formal Bayesian test relying on the Bayes factor. More specifically, assuming that the two competing models are equally likely a priori, i.e. $p(\mathcal{M}) = p(\mathcal{M}^*)$, we compare $\mathcal{M}$ against $\mathcal{M}^*$ via

$$\mathcal{B}_{\mathcal{M},\mathcal{M}^*} = \frac{p(\mathbf{Y} \mid \mathcal{M})}{p(\mathbf{Y} \mid \mathcal{M}^*)} = \frac{\sum_{\mathbf{z} \in \mathcal{Z}} p(\mathbf{Y} \mid \mathbf{z}) p(\mathbf{z})}{p(\mathbf{Y} \mid \mathbf{z}^*)}, \tag{2.3}$$

where $\sum_{\mathbf{z} \in \mathcal{Z}} p(\mathbf{Y} \mid \mathbf{z}) p(\mathbf{z})$ and $p(\mathbf{Y} \mid \mathbf{z}^*)$ are the marginal likelihoods of $\mathbf{Y}$ under $\mathcal{M}$ and $\mathcal{M}^*$. Recalling, e.g., Kass and Raftery (1995), Eq. 2.3 defines a formal Bayesian procedure to assess evidence against $\mathcal{M}^*$ relative to $\mathcal{M}$, with high values suggesting that the exogenous assignments in $\mathbf{z}^*$ are not as effective in characterizing the endogenous block structures in $\mathbf{Y}$ as the posterior for $\mathbf{z}$ under $\mathcal{M}$. Under the assumption that $p(\mathcal{M}) = p(\mathcal{M}^*)$, the Bayes factor in Eq. 2.3 coincides with the posterior odds $p(\mathcal{M} \mid \mathbf{Y})/p(\mathcal{M}^* \mid \mathbf{Y})$. When $p(\mathcal{M}) \neq p(\mathcal{M}^*)$, it suffices to rescale $\mathcal{B}_{\mathcal{M},\mathcal{M}^*}$ by $p(\mathcal{M})/p(\mathcal{M}^*)$ to assess posterior evidence against $\mathcal{M}^*$ relative to $\mathcal{M}$.

To evaluate (2.3), note that the quantity $p(\mathbf{Y} \mid \mathbf{z}^*)$ can be computed by evaluating (2.1) at $\mathbf{z} = \mathbf{z}^*$. In contrast, model $\mathcal{M}$ requires the calculation of $p(\mathbf{Y} \mid \mathbf{z})$ and $p(\mathbf{z})$ for every $\mathbf{z} \in \mathcal{Z}$. Although both quantities can be evaluated in closed form as discussed in Section 2.1, this approach is computationally impractical due to the huge cardinality of the set $\mathcal{Z}$, thus requiring alternative strategies relying on Monte Carlo estimation of

$p(\mathbf{Y} \mid \mathcal{M}) = \sum_{\mathbf{z} \in \mathcal{Z}} p(\mathbf{Y} \mid \mathbf{z}) p(\mathbf{z})$. Here, we consider the harmonic mean approach (Newton and Raftery, 1994; Raftery et al., 2007), thus obtaining

$$\hat{p}(\mathbf{Y} \mid \mathcal{M}) = \left[\frac{1}{R} \sum_{r=1}^{R} \frac{1}{p(\mathbf{Y} \mid \mathbf{z}^{(r)})} \right]^{-1}, \tag{2.4}$$

where $\mathbf{z}^{(1)}, \ldots, \mathbf{z}^{(R)}$ are samples from the posterior distribution of $\mathbf{z}$ and $p(\mathbf{Y} \mid \mathbf{z}^{(r)})$ is the likelihood in Eq. 2.1 evaluated at $\mathbf{z} = \mathbf{z}^{(r)}$, for every $r = 1, \ldots, R$. Although recent refinements have been proposed to address some shortcomings of the harmonic estimate (e.g. Lenk, 2009; Pajor, 2017), here we consider the original formula which is computationally more tractable and has proved stable in our simulations and applications; see Figs. 2 and 4.

Leveraging (2.1) and (2.4), our estimate of the Bayes factor in Eq. 2.3 is

$$\hat{\mathcal{B}}_{\mathcal{M},\mathcal{M}^*} = \frac{\hat{p}(\mathbf{Y} \mid \mathcal{M})}{p(\mathbf{Y} \mid \mathcal{M}^*)} = \frac{\left[\frac{1}{R} \sum_{r=1}^{R} \prod_{h=1}^{H^{(r)}} \prod_{k=1}^{h} \frac{\mathrm{B}(a,b)}{\mathrm{B}(a+m_{hk}^{(r)}, b+\bar{m}_{hk}^{(r)})} \right]^{-1}}{\prod_{h=1}^{H^*} \prod_{k=1}^{h} \frac{\mathrm{B}(a+m^*_{hk}, b+\bar{m}^*_{hk})}{\mathrm{B}(a,b)}}, \tag{2.5}$$

where $m_{hk}^{(r)}$ and $\bar{m}_{hk}^{(r)}$ are the counts of edges and non–edges among nodes in groups h and k induced by the rth MCMC sample of $\mathbf{z}$, whereas m^*_{hk} and $\bar{m}^*_{hk}$ denote the number of edges and non–edges among the nodes in clusters h and k induced by the exogenous assignments $\mathbf{z}^*$. Finally, $H^{(r)}$ and H^* are the total numbers of unique labels in $\mathbf{z}^{(r)}$ and $\mathbf{z}^*$. Section 3 describes the collapsed Gibbs algorithm to sample the assignment vectors $\mathbf{z}^{(1)}, \ldots, \mathbf{z}^{(R)}$ from the posterior $p(\mathbf{z} \mid \mathbf{Y})$ under model $\mathcal{M}$. These samples are required to compute (2.5) and, as discussed in Section 2.3, also allow inference on the posterior distribution of the endogenous partitions.

2.3. Inference and Uncertainty Quantification on the Endogenous Partition When the Bayes factor discussed in Section 2.2 provides evidence in favor of model $\mathcal{M}$, it is of interest to study the posterior distribution of $\mathbf{z}$ leveraging the Gibbs samples $\mathbf{z}^{(1)}, \ldots, \mathbf{z}^{(R)}$. Common strategies address this goal by first computing the posterior co–clustering matrix $\mathcal{C}$ with elements $c_{vu} = c_{uv}$ measuring the relative frequency of the Gibbs samples in which nodes $v = 2, \ldots, n$ and $u = 1, \ldots, v-1$ are in the same cluster, and then apply a standard clustering procedure to such a similarity matrix. However, this approach provides only a point estimate of $\mathbf{z}$ and, hence, fails to quantify posterior uncertainty. Legramanti et al. (2020) recently covered this gap by adapting the novel inference methods for Bayesian clustering in Wade and Ghahramani (2018) to the network field. These strategies rely on the variation of information (VI) metric, which quantifies distances between two partitions by comparing their individual and joint entropies.

Under this framework, a point estimate $\hat{\mathbf{z}}$ for $\mathbf{z}$ coincides with that partition having the lowest posterior averaged VI distance from all the other clusterings. Moreover, a $1-\delta$ credible ball around $\hat{\mathbf{z}}$ can be obtained by collecting all those partitions with a VI distance from $\hat{\mathbf{z}}$ below a given threshold, with this threshold chosen to guarantee the smallest–size ball containing at least $1-\delta$ posterior probability. Such inference is useful to complement the results of the test in Section 2.2. Namely, to get further reassurance about the output of the proposed test, we may also study whether the exogenous clustering $\mathbf{z}^*$ is plausible under the posterior distribution for the endogenous partition $\mathbf{z}$ by checking if $\mathbf{z}^*$ lies inside the credible ball around $\hat{\mathbf{z}}$. Refer to Wade and Ghahramani (2018), Legramanti et al. (2020) and to the codes at https://github.com/danieledurante/TESTsbm for more details on the aforementioned inference methods and their implementation.

Finally, although the block probabilities are integrated out, a plug–in estimate for these quantities can be easily obtained. Indeed, by leveraging beta–binomial conjugacy, we have that $(\theta_{hk} \mid \mathbf{Y}, \mathbf{z}) \sim \text{Beta}(a+m_{hk}, b+\bar{m}_{hk})$. Hence, a plug–in estimate of the block probabilities θ_{hk} for $h = 1, \ldots, \hat{H}$ and $k = 1, \ldots, h$ is

$$\hat{\theta}_{hk} = \mathbb{E}[\theta_{hk} \mid \mathbf{Y}, \hat{\mathbf{z}}] = \frac{a + \hat{m}_{hk}}{a + \hat{m}_{hk} + b + \hat{\bar{m}}_{hk}},$$

where $\hat{m}_{hk}$ and $\hat{\bar{m}}_{hk}$ denote the number of edges and non–edges between nodes in groups h and k, respectively, induced by the posterior point estimate $\hat{\mathbf{z}}$ of $\mathbf{z}$.

3 Posterior Computation via Collapsed Gibbs Sampling

The posterior samples of $\mathbf{z}$ under model (2.1) with CRP prior (2.2) can be obtained via a simple collapsed Gibbs sampler which updates the group assignment of each node v conditioned on those of the others by sampling from the full–conditional distribution $p(z_v \mid \mathbf{Y}, \mathbf{z}_{-v})$ (Schmidt and Morup, 2013). By collapsing out the beta priors for the block probabilities, this procedure reduces the computational time in avoiding the updating of θ_{hk} for each $h = 1, \ldots, H$ and $k = 1, \ldots, h$, while improving mixing (Neal, 2000).

Algorithm 1 provides the detailed steps of one cycle of the Gibbs sampler. Note that since (2.1) is the joint probability for a large set of binary edges, manipulating this quantity within Algorithm 1 and in computing the Bayes factor in Eq. 2.5 may lead to practical difficulties due to the need to deal with quantities very close to zero. In these settings, we suggest to work with the logarithm, when possible, and to exploit the log–sum–exp identity $\log[\sum_i \exp(\nu_i)] = \mathrm{d} + \log[\sum_i \exp(\nu_i - \mathrm{d})]$, where d usually coincides with $\max_i \nu_i$.

Algorithm 1 One step of the Gibbs sampler for $\mathbf{z}$ under $\mathcal{M}$.

```
for $v = 1, \ldots, n$ do
    Update each $z_v$ conditionally on $\mathbf{z}_{-v}$ and $\mathbf{Y}$ as follows
        1. Remove node $v$ from the node set.
        2. If no other node belongs to the cluster of $v$, such a cluster is removed.
        3. Reorder the cluster indices so that $1, \ldots, H_{-v}$ are non–empty,
           and sample $z_v$ from the categorical variable with
           full–conditional probabilities
           $$\mathrm{pr}(z_v = h \mid \mathbf{Y}, \mathbf{z}_{-v}) \propto \begin{cases} \frac{n_{h,-v}}{n-1+\alpha} p(\mathbf{Y} \mid z_v = h, \mathbf{z}_{-v}), & \text{if } h = 1, \ldots, H_{-v}, \\ \frac{\alpha}{n-1+\alpha} p(\mathbf{Y} \mid z_v = h, \mathbf{z}_{-v}), & \text{if } h = H_{-v} + 1, \end{cases}$$
           where $p(\mathbf{Y} \mid z_v = h, \mathbf{z}_{-v})$ is computed as in Eq. 2.1
           conditioned on $z_v = h$ and $\mathbf{z}_{-v}$.
return $\mathbf{z} = (z_1, \ldots, z_n)^\intercal$
```

4 Simulation Studies

We consider an illustrative simulation to assess the performance of the new inference procedures presented in Section 2, and to evaluate the ability of model $\mathcal{M}$ to recover underlying endogenous partition structures. Consistent with this goal, we simulate a symmetric binary adjacency matrix $\mathbf{Y}$ from a stochastic block model with $n = 60$ nodes partitioned into $H_0 = 3$ groups of equal size. In particular, we let $\mathbf{z}_0 = (z_{1,0} = 1, \ldots, z_{20,0} = 1, z_{21,0} = 2, \ldots, z_{40,0} = 2, z_{41,0} = 3, \ldots, z_{60,0} = 3)^\intercal$, and simulate each $y_{vu} = y_{uv}$ for

$v = 2, \ldots, n,\ u = 1, \ldots, v-1$ from a Bernoulli with probability 0.8 if nodes v and u are in the same group, and 0.2 otherwise.

In performing posterior inference on the endogenous clustering structure under model $\mathcal{M}$, we set $a = b = 1$ to induce a uniform prior on the block probabilities. This choice is theoretically supported (e.g. Ghosh et al., 2020) and has been widely employed in routine implementations of SBMs (Nowicki and Snijders, 2001; Kemp et al., 2006; Geng et al., 2019). As for α in prior (2.2), we set it equal to 1 following default specifications of the CRP, thus circumventing the need to include a hyper–prior which could affect mixing and convergence of Algorithm 1. Such a default specification has proved effective both in simulations and in applications, and we found the results robust to moderate changes in α. For instance, setting $\alpha = 0.1$ or $\alpha = 10$ did not change the final conclusions of our testing procedures.

Figure 1 shows the simulated adjacency matrix $\mathbf{Y}$ paritioned in blocks according to the estimated $\hat{\mathbf{z}}$ under model $\mathcal{M}$. Such an estimate relies on 15000 MCMC samples produced by Algorithm 1, after a burn–in of 2000. As shown in Fig. 2, such settings are sufficient for good convergence and mixing according to the MCMC diagnostics of key measures for posterior inference, covering the traceplot of the log–likelihood in Eq. 2.1 under model $\mathcal{M}$ and the trajectory of the logarithm of the harmonic mean estimate for the associated marginal likelihood in Eq. 2.4. As is clear from the block partition of $\mathbf{Y}$ in Fig. 1, the posterior for $\mathbf{z}$ under model $\mathcal{M}$ is able to concentrate around the true underlying endogenous partition and allows learning of the correct number of non–empty groups. These results support the use of $\mathcal{M}$ as a benchmark model to test for differences between endogenous and exogenous partitions under the methods presented in Section 2.

To assess the quality of such strategies, we consider four external assignment vectors $\mathbf{z}_0, \mathbf{z}_1, \mathbf{z}_2$ and $\mathbf{z}_3$ evaluated in Table 1. In particular, $\mathbf{z}_0$ denotes the true generative partition, $\mathbf{z}_1$ is obtained by a random permutation of the indices in $\mathbf{z}_0$, while $\mathbf{z}_2$ and $\mathbf{z}_3$ define a refined and a coarsened partitioning of $\mathbf{z}_0$, respectively, in which each cluster is either divided in two additional ones ($\mathbf{z}_2$) or collapsed with others to form a single group ($\mathbf{z}_3$). Due to this, we expect to obtain evidence in favor of the exogenous partition only in the scenario with $\mathbf{z}^* = \mathbf{z}_0$. Table 1 confirms our expectations when compared with the thresholds in Kass and Raftery (1995). Note that, although $\hat{\mathbf{z}} = \mathbf{z}_0$, we obtain a negative Bayes factor in the first scenario, which leads to a strong preference for model $\mathcal{M}^*$ relative to $\mathcal{M}$. Indeed, even if the point estimate $\hat{\mathbf{z}}$ for $\mathbf{z}$ under $\mathcal{M}$ exactly recovers $\mathbf{z}_0$, there is still some amount of posterior uncertainty induced by the CRP prior on $\mathbf{z}$. On the contrary, $\mathcal{M}^*$ is defined in the first scenario by conditioning on the true underlying

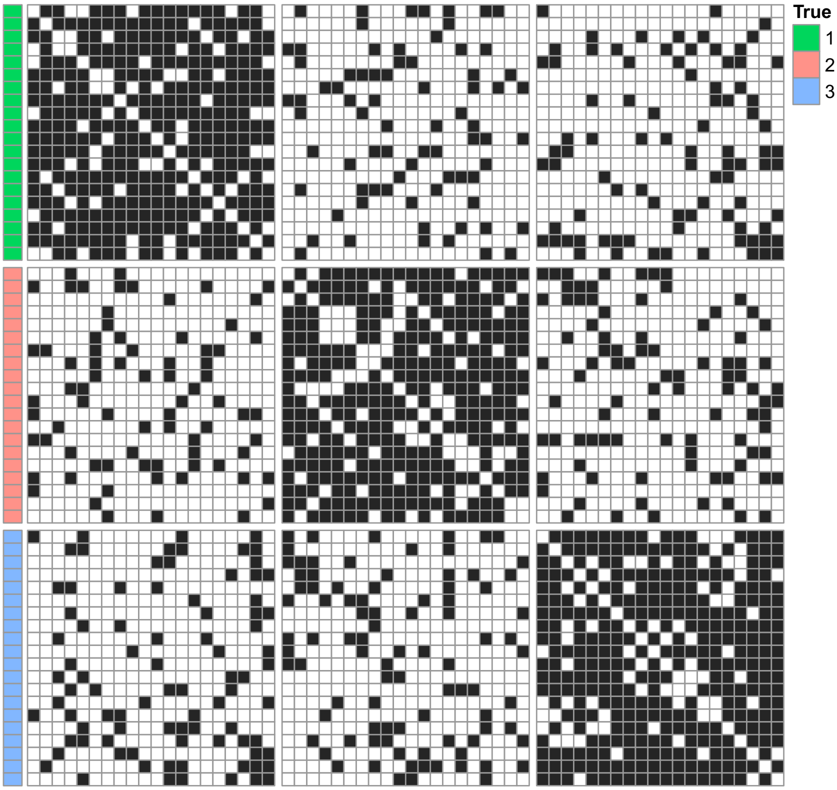

Figure 1: Graphical representation of the simulated adjacency matrix $\mathbf{Y}$ partitioned in blocks according to the estimated endogenous assignments $\hat{\mathbf{z}}$. Black and white cells denote edges and non–edges, respectively, whereas the first colored column represents the true partition $\mathbf{z}_0$. See the online article for the color version of this figure

partition with no uncertainty, thus providing a formulation much closer to the true data–generative mechanism relative to $\mathcal{M}$. All the remaining exogenous partitions $\mathbf{z}_1, \mathbf{z}_2$ and $\mathbf{z}_3$ are, instead, not as effective as model $\mathcal{M}$ in characterizing the endogenous block structures within $\mathbf{Y}$. As expected, this is especially true for the random partition ($\mathbf{z}_1$), but also those obtained from refinements ($\mathbf{z}_2$) or coarsening ($\mathbf{z}_3$) operations on $\mathbf{z}_0$ are not plausible according to the results of the tests. Such results confirm the ability of our procedures to provide accurate conclusions under various configuration of $\mathbf{z}^*$. For instance, although the partition $\mathbf{z}_2$ still leads to homogenous blocks in $\mathbf{Y}$, the additional refinements in $\mathbf{z}_2$ provide an unnecessary addition of further groups which are not required to characterize the block–connectivity

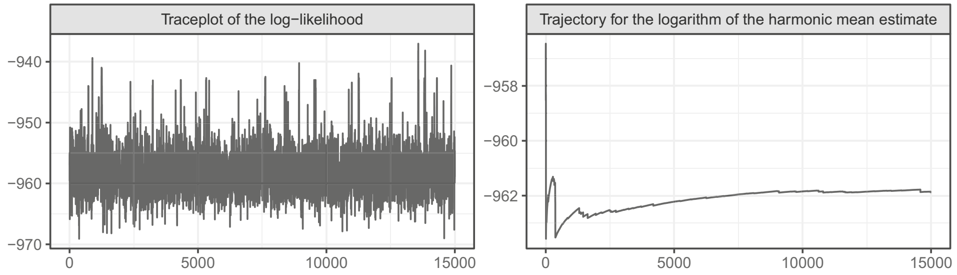

Figure 2: MCMC diagnostics for the simulation study. Left: traceplot for the logarithm of the likelihood in Eq. 2.1 computed at the MCMC samples of $\mathbf{z}$ after burn–in. Right: trajectory of the logarithm of the harmonic mean estimate in Eq. 2.4 for growing R

patterns in $\mathbf{Y}$, thus leading the test to provide evidence in favor of $\mathcal{M}$ rather than $\mathcal{M}^*$ when $\mathbf{z}^* = \mathbf{z}_2$.

The VI distances between the estimated $\hat{\mathbf{z}}$ under model $\mathcal{M}$ and the four exogenous partitions confirm the results of the tests. In particular, the only external assignment vector with a VI distance from $\hat{\mathbf{z}}$ lower than the estimated 0.428 threshold of the 95% credible ball around $\hat{\mathbf{z}}$ is $\mathbf{z}_0$.

5 Application to Brain Networks of Alzheimer's Individuals

There is an intensive research effort aimed at finding the sources of the Alzheimer's disease in human brain networks. Such an increasing interest is motivated by recent developments in brain imaging technologies and by the constant growth of elderly population in the age interval mostly affected by Alzheimer's, thus making such a disease a major concern, both in terms of disability and mortality, especially for countries with longer life expectancy (Ashford et al., 2011a, b; Stam, 2014). Here, we focus on studying structural brain networks encoding the presence or absence of white matter fibers

Table 1: Results of our proposed procedure for testing to what extent four different exogenous partitions $\mathbf{z}^*$ are as effective as the infinite relational model $\mathcal{M}$ in characterizing the endogenous block structures within $\mathbf{Y}$

$\mathbf{z}^*$	$\mathbf{z}_0$ (True)	$\mathbf{z}_1$ (Random)	$\mathbf{z}_2$ (Refined)	$\mathbf{z}_3$ (Coarsened)
$2 \log \hat{\mathcal{B}}_{\mathcal{M},\mathcal{M}^*}$	-5.17	522.27	25.68	260.40
$\text{VI}(\hat{\mathbf{z}}, \mathbf{z}^*)$	0.00	3.16	1.00	0.67

The VI distances between the estimated partition $\hat{\mathbf{z}}$ under the infinite relational model and the exogenous ones are also displayed among anatomical regions in human brains. Such connectivity data have been a source of major interest in several recent studies mostly focused on topological summary measures of Alzheimer's brains and on how these measures change as the disease progresses (Daianu et al., 2013; Sulaimany et al., 2017; John et al., 2017; Mårtensson et al., 2018). Instead, we consider a different perspective by studying the endogenous block structures in a representative Alzheimer's brain network, while assessing whether exogenous region partitions of interest can effectively characterize the block structures within the network.

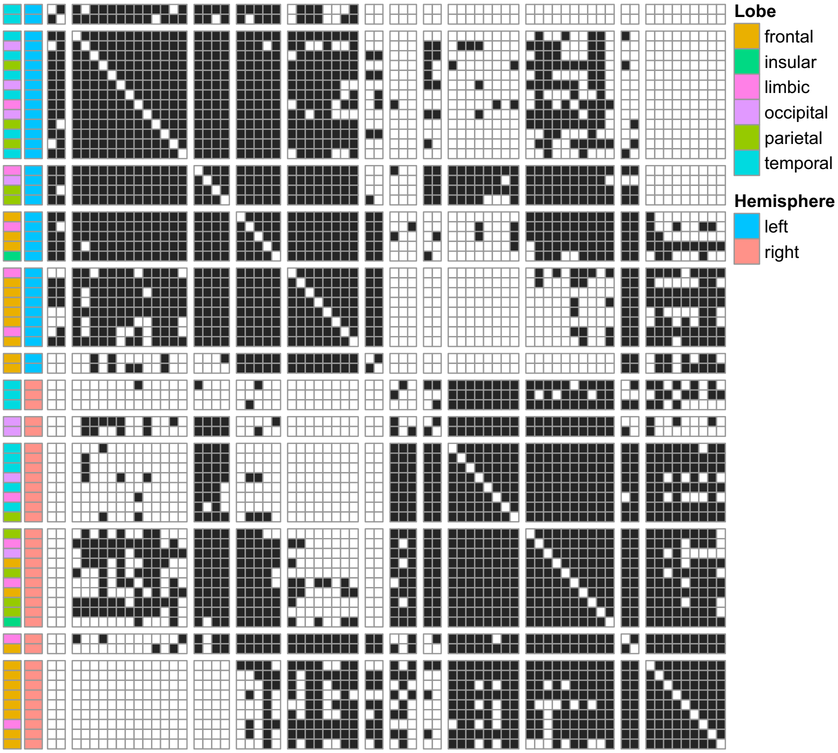

Figure 3: Graphical illustration of a representative brain network $\mathbf{Y}$ for Alzheimer's individuals. Brain regions are re–ordered and partitioned in blocks according to the estimated endogenous assignments $\hat{\mathbf{z}}$. Black and white cells denote edges and non–edges, respectively, whereas the first two colored columns represent the two exogenous anatomical brain partitions into lobes and hemispheres. See the online article for the color version of this figure

Consistent with the above goal, we apply methods in Sections 2–3 to the 68×68 binary adjacency matrix $\mathbf{Y}$ encoding the presence or absence of white matter fibers among anatomical regions in a representative Alzheimer's brain network provided by Sulaimany et al. (2017). In this study, brain regions are defined by the Desikan atlas (Desikan et al., 2006), which provides additional information on hemisphere and lobe memberships (Kang et al., 2012); see Sulaimany et al. (2017) for additional details on the construction of $\mathbf{Y}$. Figure 3 provides a graphical representation of $\mathbf{Y}$ with brain regions suitably reordered and organized in blocks according to the estimated endogenous assignments $\hat{\mathbf{z}}$. The latter are obtained by considering the same MCMC settings and hyper–parameters of the simulation study, which proved effective and robust also in this application; see Fig. 4. As shown in Fig. 3, we learn $\hat{H} = 12$ endogenous groups equally divided between the two hemispheres and showing an overall coherence of the partition structure across left and right regions. As expected, there is an evident block–connectivity within hemispheres, although some groups also display a tendency to connect across hemispheres. For example, brain regions in the frontal lobe tend to create two highly interconnected clusters, one in each hemisphere, with these two blocks showing also a preference to create bridges among the two hemispheres. Despite these anatomical homophily structures, as highlighted in Fig. 3 and in Table 2, hemisphere and lobe partitions are not sufficient to fully characterize the endogenous block structures in Alzheimer's brains. There are, in fact, various sub–blocks within each hemisphere and these clusters typically comprise regions in different lobes.

We conclude by assessing whether the clustering structures inferred from representative brains of individuals in three ordered stages of cognitive

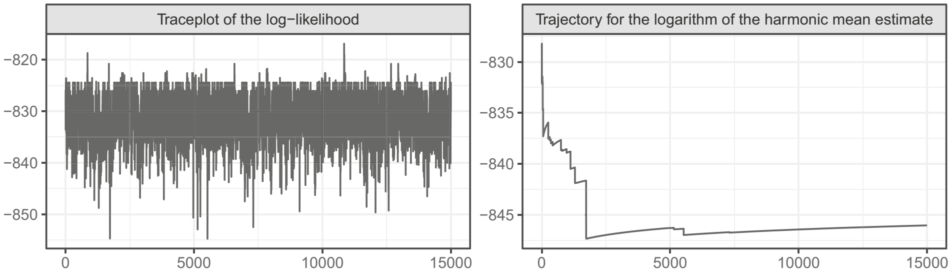

Figure 4: MCMC diagnostics for the application. Left: traceplot for the logarithm of the likelihood in Eq. 2.1 computed at the MCMC samples of $\mathbf{z}$ after burn–in. Right: trajectory of the logarithm of the harmonic mean estimate in Eq. 2.4 for growing R

Table 2: Results of our proposed procedure for testing to what extent exogenous brain partitions $\mathbf{z}^*$ are as effective as model $\mathcal{M}$ in characterizing the endogenous block structure for a representative brain network of Alzheimer's individuals

	Anatomical		Cognitive decline progression		
$\mathbf{z}^*$	Hemispheres	Lobes	Normal Aging	Early Decline	Late Decline
$2 \log \hat{\mathcal{B}}_{\mathcal{M},\mathcal{M}^*}$	713.57	1291.74	156.25	101.45	41.12
$\textsc{vi}(\hat{\mathbf{z}}, \mathbf{z}^*)$	2.29	3.37	1.46	1.33	1.10

Here, we focus on anatomical partitions and on grouping structures identified in representative brains of individuals characterized by three ordered cognitive decline stages. The VI distances between the estimated partition $\hat{\mathbf{z}}$ under the infinite relational model and the exogenous ones are also displayed

decline can effectively explain the endogenous block structures in Alzheimer's brains. To accomplish this goal, we first apply Algorithm 1 to the representative adjacency matrices of individuals characterized by normal aging, early and late cognitive decline (Sulaimany et al., 2017), and then quantify, via the Bayes factors in Table 2, whether these partitions are also effective in modeling the block structures within the Alzheimer's brain. Although $2 \log \hat{\mathcal{B}}_{\mathcal{M},\mathcal{M}^*}$ is above the threshold in Kass and Raftery (1995) suggesting strong evidence against this hypothesis for all the three stages, it is interesting to notice how $2 \log \hat{\mathcal{B}}_{\mathcal{M},\mathcal{M}^*}$ decreases as cognitive decline progresses towards Alzheimers' disease. This means that the inferred partitions could be used, with caution, as a diagnostic strategy to identify the progress of the disease. The VI distances between the estimated $\hat{\mathbf{z}}$ and these external partitions confirm the evidence provided by the Bayes factors.

To further validate the suitability of $\mathcal{M}$ as a flexible model for $\mathbf{Y}$, we also compute the in–sample missclassification error when predicting each y_{vu} with $\hat{\theta}_{\hat{z}_v,\hat{z}_u}$. Such a measure is 0.1, thus confirming that $\mathcal{M}$ can be regarded as a suitable model for this application.

6 Discussion and Future Developments

This article introduces a formal Bayesian testing procedure to assess the ability of a fixed exogenous node partition in characterizing block structures in a network, relative to an infinite relational model. To accomplish this goal, we compare an harmonic mean estimate of the marginal likelihood under this latter representation with the one induced by a stochastic block model conditioned on the external partition of interest. From a computational perspective, we rely on a collapsed Gibbs sampler which additionally allows

Bayesian inference and uncertainty quantification on endogenous partitions. As illustrated in simulations and applications to brain networks, our proposal provides a simple yet effective procedure to obtain further insights on the effects of categorical node attributes on network structures.

There are several directions for future developments. For example, weighted networks comprising counts or continuous measures of strength in the relationship can be easily incorporated within our strategy by simply replacing the likelihood in Eq. 2.1 with a suitable one. This can be obtained by leveraging Poisson–gamma or Gaussian–Gaussian conjugacy, as done for the beta–binomial case. Moreover, while throughout the paper we have considered the problem of testing model $\mathcal{M}$ against model $\mathcal{M}^*$ given a single observed network $\mathbf{Y}$, one may be interested in the same test given a sample of N exchangeable networks. This is feasible under our proposed framework and only requires to substitute $p(\mathbf{Y} \mid \mathbf{z})$ in Eq. 2.1 with $p(\mathbf{Y}_1, \ldots, \mathbf{Y}_N \mid \mathbf{z})$. It is also possible to compare two exogenous partitions, rather than an exogenous and an endogenous one. This task is even simpler than the one analyzed in this article, since the likelihood in Eq. 2.1 can be computed in closed form for both the external partitions under comparison, thus avoiding the need of mcmc methods. For example, one may be interested in comparing an external assignment $\mathbf{z}^*$ with a random permutation of the indices in such a vector to assess whether $\mathbf{z}^*$ offers improvements in modeling network block structures or has no effect. Therefore, the perspective taken by Bianconi et al. (2009) and Peel et al. (2017) can be seen as a special case of our more general solution.

Acknowledgements. The authors are grateful to the Editor, the Associate Editor and the two referees for their useful comments on a first version of this manuscript. Sirio Legramanti and Daniele Durante acknowledge support from miur–prin 2017 grant 20177BRJXS. Tommaso Rigon acknowledges support from grant R01ES027498 of the National Institute of Environmental Health Sciences of the United States National Institutes of Health.

Funding. Open access funding provided by Università Commerciale Luigi Bocconi within the CRUI-CARE Agreement.

References

ABBE, E. (2017). Community detection and stochastic block models: Recent developments. *J. Mach. Learn. Res* **18**, 6446–6531.

AIROLDI, E. M., BLEI, D. M., FIENBERG, S. E. and XING, E. P. (2008). Mixed membership stochastic blockmodels. *J. Mach. Learn. Res.* **9**, 1981–2014.

ALDOUS, D. (1985). Exchangeability and related topics. *École d'Été de Probabilités de Saint-Flour XIII–1983* **1117**, 1–198.

ASHFORD, J. W., ROSEN, A., ADAMSON, M., BAYLEY, P., SABRI, O., FURST, A. and BLACK, S. E. (2011a). *Handbook of Imaging the Alzheimer Brain*: IOS Press.

ASHFORD, J. W., SALEHI, A., FURST, A., BAYLEY, P., FRISONI, G. B., JACK, JR. C.R., SABRI, O., ADAMSON, M. M., COBURN, K. L. and OLICHNEY, J. (2011b). Imaging the Alzheimer brain. *J. Alzheimer's Dis.* **26**, 1–27.

BIANCONI, G., PIN, P. and MARSILI, M. (2009). Assessing the relevance of node features for network structure. *Proc. Natl. Acad. Sci.* **106**, 11433–11438.

BLONDEL, V. D., GUILLAUME, J. L., LAMBIOTTE, R. and LEFEBVRE, E. (2008). Fast unfolding of communities in large networks. *J. Stat. Mech.* **1**, 1–12.

CHOI, D. S., WOLFE, P. J. and AIROLDI, E. M. (2012). Stochastic blockmodels with a growing number of classes. *Biometrika* **99**, 273–284.

DAIANU, M., JAHANSHAD, N., NIR, T. M., TOGA, A. W., JACK, JR. C.R., WEINER, M. W. and THOMPSON, P. M. FOR THE ALZHEIMER'S DISEASE NEUROIMAGING INITIATIVE (2013). Breakdown of brain connectivity between normal aging and Alzheimer's disease: a structural k–core network analysis. *Brain Connect.* **3**, 407–422.

DESIKAN, R. S., SÉGONNE, F., FISCHL, B., QUINN, B. T., DICKERSON, B. C., BLACKER, D., BUCKNER, R. L., DALE, A. M., MAGUIRE, R. P. and HYMAN, B. T. (2006). An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *Neuroimage* **31**, 968–980.

FASKOWITZ, J., YAN, X., ZUO, X. N. and SPORNS, O. (2018). Weighted stochastic block models of the human connectome across the life span. *Sci. Rep.* **8**, 1–16.

FORTUNATO, S. (2010). Community detection in graphs. *Sci. Rep.* **486**, 75–174.

FORTUNATO, S. and HRIC, D. (2016). Community detection in networks: a user guide. *Phys. Rep.* **659**, 1–44.

GAO, C., MA, Z., ZHANG, A. Y. and ZHOU, H. H. (2018). Community detection in degree-corrected block models. *Ann. Stat.* **46**, 2153–2185.

GENG, J., BHATTACHARYA, A. and PATI, D. (2019). Probabilistic community detection with unknown number of communities. *J. Am. Stat. Assoc.* **114**, 893–905.

GERSHMAN, S. J. and BLEI, D. M. (2012). A tutorial on Bayesian nonparametric models. *J. Math. Psychol.* **56**, 1–12.

GHOSH, P., PATI, D. and BHATTACHARYA, A. (2020). Posterior contraction rates for stochastic block models. *Sankhya A* **82**, 448–476.

HOLLAND, P. W., LASKEY, K. B. and LEINHARDT, S. (1983). Stochastic blockmodels: First steps. *Soc. Networks* **5**, 109–137.

JOHN, M., IKUTA, T. and FERBINTEANU, J. (2017). Graph analysis of structural brain networks in Alzheimer's disease: Beyond small world properties. *Brain Struct. Funct.* **222**, 923–942.

KANG, X., HERRON, T. J., CATE, A. D., YUND, E. W. and WOODS, D. L. (2012). Hemispherically–unified surface maps of human cerebral cortex: Reliability and hemispheric asymmetries. *PLoS One* **7**, 1–15.

KARRER, B. and NEWMAN, M. E. (2011). Stochastic blockmodels and community structure in networks. *Phys. Rev. E* **83**, 016107.

KASS, R. E. and RAFTERY, A. E. (1995). Bayes factors. *J. Am. Stat. Assoc.* **90**, 773–795.

KEMP, C., TENENBAUM, J. B., GRIFFITHS, T. L., YAMADA, T. and UEDA, N. (2006). *Learning systems of concepts with an infinite relational model. In Proceedings of the 21st National Conference on Artificial Intelligence - Volume 1*, p. 381-388.

LEE, C. and WILKINSON, D. J. (2019). A review of stochastic block models and extensions for graph clustering. *Appl. Netw. Sci.* **4**, 1–50.

LEGRAMANTI, S., RIGON, T., DURANTE, D. and DUNSON, D. B. (2020). Extended stochastic block models. arXiv:2007.08569.

LENK, P. (2009). Simulation pseudo–bias correction to the harmonic mean estimator of integrated likelihoods. *J. Comput. Graph. Stat.* **18**, 941–960.

MARIADASSOU, M., ROBIN, S. and VACHER, C. (2010). Uncovering latent structure in valued graphs: a variational approach. *Ann. Appl. Stat.* **4**, 715–742.

MÅRTENSSON, G., PEREIRA, J. B., MECOCCI, P., VELLAS, B., TSOLAKI, M., KŁOSZEWSKA, I., SOININEN, H., LOVESTONE, S., SIMMONS, A. and VOLPE, G. (2018). Stability of graph theoretical measures in structural brain networks in Alzheimer's disease. *Sci. Rep.* **8**, 1–15.

NEAL, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *J. Comput. Graph. Stat.* **9**, 249–265.

NEWMAN, M. E. J. and GIRVAN, M. (2004). Finding and evaluating community structure in networks. *Phys. Rev. E* **69**, 1–15.

NEWMAN, M. E. J. and CLAUSET, A. (2016). Structure and inference in annotated networks. *Nat. Commun.* **7**, 1–11.

NEWTON, M. A. and RAFTERY, A. E. (1994). Approximate Bayesian inference with the weighted likelihood bootstrap. *J. R. Stat. Soc. Series B* **56**, 3–26.

NOWICKI, K. and SNIJDERS, T. A. B. (2001). Estimation and prediction for stochastic blockstructures. *J. Am. Stat. Assoc.* **96**, 1077–1087.

PAJOR, A. (2017). Estimating the marginal likelihood using the arithmetic mean identity. *Bayesian Anal.* **12**, 261–287.

PEEL, L., LARREMORE, D. B. and CLAUSET, A. (2017). The ground truth about metadata and community detection in networks. *Sci. Adv.* **3**, 1–8.

RAFTERY, A. E., NEWTON, M. A., SATAGOPAN, J. M. and KRIVITSKY, P. N. (2007). Estimating the integrated likelihood via posterior simulation using the harmonic mean identity. *Bayesian Stat.* **8**, 1–45.

ROY, S., ATCHADÉ, Y. and MICHAILIDIS, G. (2019). Likelihood inference for large scale stochastic blockmodels with covariates based on a divide-and-conquer parallelizable algorithm with communication. *J. Comput. Graph. Stat.* **28**, 609–619.

SALDANA, D. F., YU, Y. and FENG, Y. (2017). How many communities are there? *J. Comput. Graph. Stat.* **26**, 171–181.

SCHMIDT, M. N. and MORUP, M. (2013). Nonparametric Bayesian modeling of complex networks: An introduction. *IEEE Signal Process. Mag.* **30**, 110–128.

SPORNS, O. (2013). Structure and function of complex brain networks. *Dialogues Clin. Neurosci.* **15**, 247–262.

STAM, C. J. (2014). Modern network science of neurological disorders. *Nat. Rev. Neurosci.* **15**, 683–695.

STANLEY, N., BONACCI, T., KWITT, R., NIETHAMMER, M. and MUCHA, P. J. (2019). Stochastic block models with multiple continuous attributes. *Appl. Netw. Sci.* **4**, 1–22.

SULAIMANY, S., KHANSARI, M., ZARRINEH, P., DAIANU, M., JAHANSHAD, N., THOMPSON, P. M. and MASOUDI-NEJAD, A. (2017). Predicting brain network changes in Alzheimer's disease with link prediction algorithms. *Mol. Biosyst.* **13**, 725–735.

SWEET, T. M. (2015). Incorporating covariates into stochastic blockmodels. *J. Educ. Behav. Stat.* **40**, 635–664.

TALLBERG, C. (2004). A Bayesian approach to modeling stochastic blockstructures with covariates. *J. Math. Sociol.* **29**, 1–23.

VAN DER PAS, S. and VAN DER VAART, A. (2018). Bayesian community detection. *Bayesian Anal.* **13**, 767–796.

WADE, S. and GHAHRAMANI, Z. (2018). Bayesian cluster analysis: Point estimation and credible balls. *Bayesian Anal.* **13**, 559–626.

WHITE, A. and MURPHY, T. B. (2016). Mixed–membership of experts stochastic blockmodel. *Netw. Sci.* **4**, 48–80.

ZHAO, Y., LEVINA, E. and ZHU, J. (2012). Consistency of community detection in networks under degree–corrected stochastic block models. *Ann. Stat.* **40**, 2266–2292.

Publisher's Note. Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

SIRIO LEGRAMANTI
DEPARTMENT OF DECISION SCIENCES,
BOCCONI UNIVERSITY, MILANO, ITALY

TOMMASO RIGON
DEPARTMENT OF STATISTICAL SCIENCES,
DUKE UNIVERSITY, DURHAM, USA

DANIELE DURANTE
DEPARTMENT OF DECISION SCIENCES AND
INSTITUTE FOR DATA SCIENCE AND
ANALYTICS, BOCCONI UNIVERSITY,
MILANO, ITALY
E-mail: daniele.durante@unibocconi.it

Paper received: 31 March 2020; accepted 5 November 2020.