Statistics and Its Interface Volume 10 (2017) 369–378

LCN: a random graph mixture model for community detection in functional brain networks*

Christopher Bryant, Hongtu Zhu†,
Mihye Ahn, and Joseph Ibrahim

The aim of this article is to develop a Bayesian random graph mixture model (RGMM) to detect the latent class network (LCN) structure of brain connectivity networks and estimate the parameters governing this structure. The use of conjugate priors for unknown parameters leads to efficient estimation, and a well-known nonidentifiability issue is avoided by a particular parameterization of the stochastic block model (SBM). Posterior computation proceeds via an efficient Markov Chain Monte Carlo algorithm. Simulations demonstrate that LCN outperforms several other competing methods for community detection in weighted networks, and we apply our RGMM to estimate the latent community structures in the functional resting brain networks of 185 subjects from the ADHD-200 sample. We find overlap in the estimated community structure across subjects, but also heterogeneity even within a given diagnosis group.

AMS 2000 subject classifications: Primary 62P10, 62-09; secondary 62-07.

1. INTRODUCTION

A problem of particular interest in the field of neuroscience is to understand the structure and organization of functional and structural brain networks and their relationships with predictors such as disease status and behavior [16, 3]. The existing literature has largely focused on various topological measures, such as degree distribution, clustering coefficient and network diameter, and their clinical implications [3, 26, 14, 9]. These types of global and local network characteristics are convenient in their ability to reduce large networks to a small set of statistics that describe their large-scale organization. The community network structure, in which there exist groups of nodes (sometimes called "modules") that have dense connections within each group and sparse connections between different groups, has been observed in numerous real life networks [34, 10, 7, 15, 33], including functional brain networks [26, 16]. See [7] for a comprehensive review of various approaches to the community detection problem.

Our motivating data are resting-state functional magnetic resonance images (fMRI) from the ADHD-200 sample, which is downloadable from http://fcon_1000.projects.nitrc.org/indi/adhd200. In this study, we have used the dataset with 215 subjects collected at New York University. After removing the ADHD hyperactive/impulsive subtype due to a small sample size, our analysis dataset consists of 185 subjects: 91 typically developing controls, 62 of the ADHD combined subtype, and 32 of the ADHD inattentive subtype. For each subject, we calculated a 116×116 Fisher-transformed correlation matrix based on the 116 predefined regions of interest (ROI) defined by the automated anatomical labeling (AAL) template [30] and used it as a resting-state fMRI connectivity network. See Figure 1 for the networks of two randomly selected subjects, which have been visualized with the BrainNet Viewer (http://www.nitrc.org/projects/bnv/) [32]. Thus, our networks consist of 116 nodes (brain regions) and the weighted edges between them (Fisher-transformed correlations between time courses). We are interested in understanding the modular structure of these functional brain networks, which we address by formulating a Bayesian random graph mixture model to detect the latent community structure in each network and estimate the modularity parameters governing the edge weights.

Two major classes of community detection methods include optimization algorithms and model-based methods. The typical approach to model-based community detection is via the stochastic block model (SBM), which summarizes the network characteristics through a low dimensional latent space, while partitioning the network into blocks of nodes with similar connectivity characteristics [20, 17, 18]. The SBM can be seen as an extension of the well-known Erdős-Rényi random graph model for binary graphs [6]. While much of the focus has been on binary graphs (e.g. Nowicki and Snijders [20], Choi, Wolfe and Airoldi [4], Vu, Hunter and Schweinberger [31], Schweinberger and Handcock [24]), versions of the SBM have been proposed to estimate the

*Research reported in this article was partially supported by the National Cancer Institute of the National Institutes of Health through the training grant *Biostatistics for Research in Genomics and Cancer*, NCI grant 5T32CA106209-07 (T32 the research of Dr. Ibrahim was supported by NIH grants GM70335 and POICA142538 and the research of Dr. Zhu was supported by NSF grants SES-1357666 and DMS-1407655, NIH grant MH 086633, and a grant from Cancer Prevention Research Institute of Texas). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.
†Corresponding author.

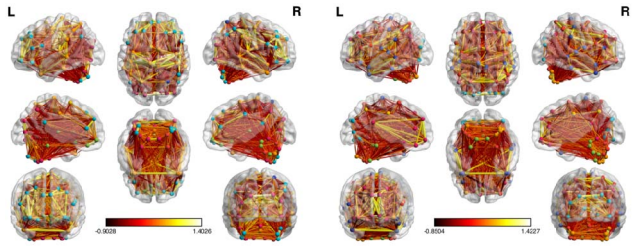

Figure 1. Functional brain networks for subject 1 (L) and subject 2 (R). There are 116 brain regions in each image. Different colored nodes indicate different estimated latent classes, but colors are not comparable between subjects.

community structure of random *weighted* graphs [17]. However, calculating unknown parameters in these SBMs represents major computational challenges. For instance, maximum likelihood estimation is only possible for small graphs due to the intractable summation in the EM algorithm for the SBM likelihood [27]. Alternatively, Bayesian methods based on Markov chain Monte Carlo (MCMC) sampling and variational algorithms have been developed for the calculation of posterior estimates for the SBM [17, 12, 1]. Moreover, other approximating methods, such as the use of a composite likelihood and moment estimators, have been proposed to compute parameter estimates for some versions of the SBM [2].

In this paper, we develop a fully Bayesian framework for the weighted SBM as a hierarchical random graph mixture model (RGMM), in order to estimate the latent class network (LCN) structure in functional brain networks. We propose conjugate priors for the unknown parameters in order to achieve efficient estimation and use the more parsimonious affiliation version of the SBM to avoid a well-known nonidentifiability issue. We develop an efficient Markov chain Monte Carlo (MCMC) algorithm to draw random samples from the desired posterior distribution. Our MCMC algorithm can handle graphs with thousands of nodes or relatively few nodes without having to rely on any asymptotic assumptions or approximations. Our simulations demonstrate that our estimation approach outperforms several existing methods for the weighted SBM in terms of both classification accuracy and accuracy in estimating the modularity parameters, and we apply our method to the sample of functional brain networks and examine the patterns of estimated latent community structure across children from different ADHD diagnosis groups.

The rest of the article is organized as follows. Section 2 introduces the formulation of our random graph mixture model and its associated MCMC sampling algorithm. In Section 3, we compare our method to several competing methods using simulated data. Then in Section 4, we apply our method to the functional brain network dataset discussed above. In Section 5, we present some concluding remarks.

2. METHODOLOGY

2.1 Random graph mixture model

Let $\mathbf{Y} = (Y_{ij})$ denote an observed undirected graph with n nodes, where Y_{ij} denotes the weighted edge value between node i and node j. We assume that the n nodes each fall into one of Q latent classes, with the unobserved class label of node i given by the random vector $Z_i = (Z_{i1}, \dots Z_{iQ})$, where $Z_{iq} = 1$ indicates that node i is in the $q-$th group. Following the version of the SBM in [2], our RGMM consists of:

- (i) A latent class model for characterizing the class label Z_i for each node $i = 1, \dots n$.
- (ii) A measurement model for characterizing the conditional distribution of Y_{ij} given $\{Z_i, Z_j\}$.

We assume that the latent classes $\{Z_i\}$ are independently and identically distributed as Multinomial random variables with the probability vector $\pi = (\pi_1, \dots \pi_Q)$ such that $0 \le \pi_q \le 1$ and $\sum_q \pi_q = 1$. The measurement model is a two-component mixture model: we assume that Y_{ij} conditional on $\{Z_i\}_{1 \le i \le n}$ are independent and the conditional distribution of Y_{ij} given $Z_{ic} \cdot Z_{jd} = 1$ is given by

$$p_{cd} f(\cdot; \theta_{cd}) + (1 - p_{cd}) \delta_0(\cdot) \quad \text{for} \quad i, j = 1, \dots n,$$

where $f(\cdot; \theta_{cd})$ is a prefixed probability distribution with an unknown parameter vector θ_{cd} and $\delta_0(\cdot)$ denotes the Dirac measure at zero accounting for non-present edges. By assuming that the edge values are conditionally independent given the latent classes of the nodes, the (marginal) dependencies of the graph are fully determined by the latent community structure.

Furthermore, we impose the affiliation SBM by reducing the $Q \cdot (Q + 1)$ parameters in $\{p_{cd}\}_{1 \le c \le d \le Q}$ and $\{\theta_{cd}\}_{1 \le c \le d \le Q}$ to:

$$p_{cd} = \begin{cases} p_{in} & \text{if } c = d, \\ p_{out} & \text{if } c \ne d, \end{cases} \quad \text{and} \quad \theta_{cd} = \begin{cases} \theta_{in} & \text{if } c = d, \\ \theta_{out} & \text{if } c \ne d. \end{cases}$$

Use of this parameterization allows us to avoid the typical problem of label switching/swapping in Bayesian mixture modeling. When non-symmetric priors are used for the group proportions, the nonidentifiability of the order of the latent classes of nodes can lead to the class labels changing between successive MCMC samples and make posterior inference difficult. The affiliation SBM does not have class-specific parameters, so the sampler arbitrarily sets the order in the initialization step and then sampling proceeds without label swapping.

This framework is flexible and can model directed graphs by utilizing a bivariate distribution for $f(Y_{ij}, Y_{ji})$ and allowing $p_{cd} \ne p_{dc}$ and $\theta_{cd} \ne \theta_{dc}$. Here we focus on Gaussian-weighted edges, such that $\theta = (\theta_{in}, \theta_{out}) = (\mu_{in}, \tau_{in}, \mu_{out}, \tau_{out})$, but we can easily incorporate different distributions for the edge distribution $f(\cdot)$. We can also

adapt the model for more complex latent structures such as the overlapping SBM and correlated latent groups by alterations within this hierarchical formulation.

2.2 Prior distributions

Priors are chosen to preserve conjugacy to allow for efficient MCMC estimation as follows:

$$\pi|Q \sim \text{Dirichlet}(a_1, \dots a_Q), \quad p_{in}, p_{out} \sim \text{Uniform}(0,1),$$
$$\mu_{in}|\tau_{in} \sim N(\mu_{0,in}, \tfrac{\sigma^2_{0,in}}{\tau_{in}}), \quad \mu_{out}|\tau_{out} \sim N(\mu_{0,out}, \frac{\sigma^2_{0,out}}{\tau_{out}}),$$
$$\tau_{in} \sim \text{Ga}(\alpha_{0,in}, \beta_{0,in}), \quad \tau_{out} \sim \text{Ga}(\alpha_{0,out}, \beta_{0,out}),$$

where $\text{Ga}(a,b)$ is a gamma distribution with shape a and rate b. To achieve relatively flat priors, we set the hyperparameters to be: $a_1, \dots a_Q = 1$, $\mu_{0,in} = \mu_{0,out} = 0$, $\sigma^2_{0,in} = \sigma^2_{0,out} = 10$, and $\alpha_{0,in} = \alpha_{0,out} = \beta_{0,in} = \beta_{0,out} = 0.01$.

2.3 Estimation

We utilize a Gibbs sampler for posterior computation, with all full conditional posterior distributions given in the Appendix. The Gibbs sampler involves sampling from a series of conditional distributions while each of the components is updated in turn. Our Gibbs sampler starts as follows:

- Initialize $\pi_q^{(0)} = 1/Q$ for $q = 1, \dots Q$.
- Sample $Z_i^{(0)}$ from Dirichlet$(\pi^{(0)})$ for $i = 1, \dots n$.
- Initialize $p_{in}^{(0)} = p_{out}^{(0)} = \frac{1}{2}, \mu_{in}^{(0)} = \mu_{out}^{(0)} = 0, \tau_{in}^{(0)} = \tau_{out}^{(0)} = 1$.

Then for $t = 1, \dots N$, we sequentially update all parameters as follows:

- Sample $\pi^{(t)}$ from $P(\pi|Q, Y, Z^{(t-1)})$.
- For $i = 1, \dots n$, sample $Z_i^{(t)}$ from

$$P(Z_i|Q, Y, Z^{(t)}_{1:(i-1)}, Z^{(t-1)}_{(i+1):n}, \pi^{(t)}, p_{in}^{(t-1)}, p_{out}^{(t-1)}, \mu_{in}^{(t-1)}, \mu_{out}^{(t-1)}, \tau_{in}^{(t-1)}, \tau_{out}^{(t-1)}).$$

- Sample $\mu_{in}^{(t)}$ from $P(\mu_{in}|Q, Y, Z^{(t)}, \tau_{in}^{(t-1)})$ and $\mu_{out}^{(t)}$ from $P(\mu_{out}|Q, Y, Z^{(t)}, \tau_{out}^{(t-1)})$.
- Sample $\tau_{in}^{(t)}$ from $P(\tau_{in}|Y, Z^{(t)}, \mu_{in}^{(t)})$ and $\tau_{out}^{(t)}$ from $P(\tau_{out}|Y, Z^{(t)}, \mu_{out}^{(t)})$.
- Sample $p_{in}^{(t)}$ from $P(p_{in}|Y, Z^{(t)})$ and $p_{out}^{(t)}$ from $P(p_{out}|Y, Z^{(t)})$.

To improve sampling performance, we run multiple MCMC chains and use the Integrated Completed Likelihood (ICL) criterion to automatically select the chain that maximizes ICL [17, 5]. For a graph with n nodes, the ICL criterion is given by:

$$(1) \quad \text{ICL}_Q = \max_\theta \log \mathbb{P}(Y, \tilde{Z}|Q, \theta) - \frac{1}{2}[P_Q \cdot \log\left(\frac{n(n-1)}{2}\right) + (Q-1)\cdot\log(n)],$$

where $\tilde{Z}$ denotes the predictions for the latent Z and P_Q denotes the number of independent parameters. In this case, we have $\theta = (p_{in}, p_{out}, \mu_{in}, \mu_{out}, \tau_{in}, \tau_{out})$ and $P_Q = 6$. Moreover, we plug in the univariate mode of each parameter into ICL_Q. This amounts to maximizing the observed data likelihood when comparing two MCMC chains with Q and n fixed.

To achieve better sampling performance for large graphs, we propose using spectral clustering to estimate the initial value of the latent structure $Z^{(0)}$; we can use the k-means clustering algorithm [11] to cluster all n nodes into Q groups according to the first Q eigenvectors of a graph. Moreover, the diagnostic tools in the *coda* R package [22] can be used to assess posterior convergence.

3. SIMULATIONS

We carried out simulations to examine the finite sample performance of the LCN RGMM in detecting the community structure of simulated networks and quantify their network modularity.

3.1 Setup

We simulated networks as follows: for a given $Q*$, π was randomly generated from Dirichlet$(a_1, \dots a_{Q*})$, and then each Z_i for $i = 1, \dots n$ was independently generated from Multinomial$(\pi_1, \dots \pi_{Q^*})$. The data Y_{ij} were generated from a mixture of zero-valued edges, randomly drawn from either Bernoulli$(1-p_{in})$ or Bernoulli$(1-p_{out})$ distributions and either Normal$(\mu_{in}, \tau_{in}^{-1})$ or Normal$(\mu_{out}, \tau_{out}^{-1})$, depending on whether nodes i and j are in the same latent class. We set hyper-parameters $\sigma^2_{0,in}$ and $\sigma^2_{0,out}$ to one. The parameters $p_{in}, p_{out}, \mu_{in}, \mu_{out}, \tau_{in}$, and τ_{out} were fixed at various values in order to examine the finite sample performance of LCN and the associated MCMC algorithm as modularity measures change.

We considered six schemes and simulated 200 independent graphs for each scheme. Simulation schemes are listed in Table 1. Scheme 1 is an example of a relatively easy community detection problem with $p_{in} >> p_{out}$ and $\mu_{in} >>$

Table 1. 200 datasets were simulated from each of these schemes, then analyzed using 2 MCMC chains, and the chain with the greatest ICL was selected

Sim	n	$Q*$	Q (est)	p_{in}	p_{out}	μ_{in}	μ_{out}	τ_{in}	τ_{out}
1	50	3	3	0.8	0.3	1	-1	1	1
2	50	3	3	1	1	0.5	-0.3	0.2	0.4
3	50	3	5	0.8	0.3	1	-1	1	1
4	500	3	3	0.8	0.3	0.5	-0.5	1	1
5	100	5	10	0.8	0.3	0.5	-0.5	1	1
6	50	10	10	0.8	0.3	0.5	-0.5	1	1

μ_{out}. Scheme 2 is a much harder problem with decreased distance between mixture distributions and fully dense graphs (no zero edges). Schemes 3 and 5 were designed to test performance when the number of latent groups is misspecified. Scheme 4 represents a scenario with a large number of nodes. Scheme 6 is a scenario with a relatively large number of smaller latent groups.

For each graph, we ran two independent chains of the Gibbs sampler and then used ICL to choose the best chain as described previously. We also compared our method with several methods for community detection in weighted graphs: the approximating method of Ambroise and Matias based on a composite likelihood [2] (AM), the Bayesian implementation of the original SBM of Nowicki and Snijders in the *hergm* R package [20, 28, 25] (HERGM), the spin-glass model of [23] (SPIN) [19, 29], and a simple spectral clustering algorithm, using k-means [11] on the eigenvectors of the adjacency matrix (SPEC). To deal with the label switching phenomenon seen in the *hergm* output, MCMC samples were relabeled with the use of the loss function from Carvalho (2013) [21], which is included in the R function *hergm.postprocess*.

3.2 Results

Classification is typically accurate under all of the simulation schemes, as shown via box plots of the misclassification rates in Figure 2, though expectedly less so with more similar mixture distributions. The most probable classes were estimated from the 10,000 MCMC samples for each simulation, and the misclassification rate was estimated as the sum of false positives (nodes estimated to be in the same community when they are not) and false negatives (nodes estimated to be in different communities when they are in the same) divided by the total number of possible latent connections $(n \cdot (n-1)/2)$. Most misclassification occurred in MCMC chains that did not converge to the true posterior distribution, which is seen in the tails of the box plots – many of these incorrectly estimated a single latent class containing all the nodes.

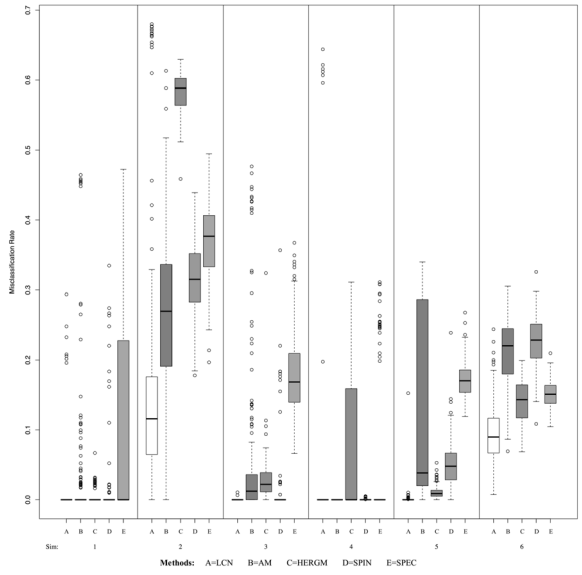

Figure 2. Boxplots of misclassification rates by simulation scheme. The 6 schemes, each with 200 simulated datasets, are listed in Table 1. Misclassification rate is defined as the sum of the false positives and false negatives divided by the total number of possible node pairs.

The methods mostly do well for the "easy" community detection problem (Scheme 1). Our method outperformed the other methods for the selection of the true number of groups when more groups were specified (Schemes 3 and 5). The approximating method of Ambroise and Matias (AM) fares well with a large number of nodes (Scheme 4), but it is not as accurate for smaller graphs (Schemes 2 and 6). The Bayesian method (HERGM) [20, 28, 25] is approximately exact, but it involves a computationally intensive algorithm for solving the label switching problem, which adds another level of error in estimating the latent structure, especially in the difficult Scheme 2. The spin-glass method (SPIN) [19, 29], which is based on extension of modularity maximization to networks with positive and negative weights, tends to be accurate in classification but less so in Scheme 6, where there are a greater number of smaller groups; it has been shown previously that modularity optimization methods can fail to detect communities that are smaller than a value which depends on the total network size and the connectedness of separate communities [8]. The simple spectral clustering algorithm (SPEC) only performs well in Scheme 4, where the simulated networks are larger than in the other schemes.

In our estimation method, when the MCMC chain converges to the true distribution, estimation of the other parameters is accurate. Figure 3 shows the absolute deviation from between the posterior median and the true parameter value, scaled by the magnitude of the parameter. For the edge parameters $(\mu_{in}, \mu_{out}, \tau_{in}, \tau_{out})$ estimated in both our formulation and the parameterization used in [2], our approach typically has less estimation error. Figure 4 gives coverage of the 95% highest posterior density (HPD) regions for the edge parameters; coverage is near 95% for most edge parameters, except for p_{in} and p_{out} in Scheme 2 – in which the parameters are on the boundary of the parameter space. For large graphs such as in Scheme 4, decreasing HPD widths indicate efficient estimation of the edge parameters.

4. ADHD-200 RESTING-STATE FMRI NETWORKS

The resting state fMRI scans were acquired using a Siemens Allegra 3T scanner for six minutes (voxel size = $3 \times 3 \times 4$ mm, slice thickness = 4 mm, number of slices=33, TR=2 s, TE=15 ms, flip angle=90°, field of view=240 mm). The Athena pipeline was applied for data preprocessing and the images were band-pass filtered within a frequency range

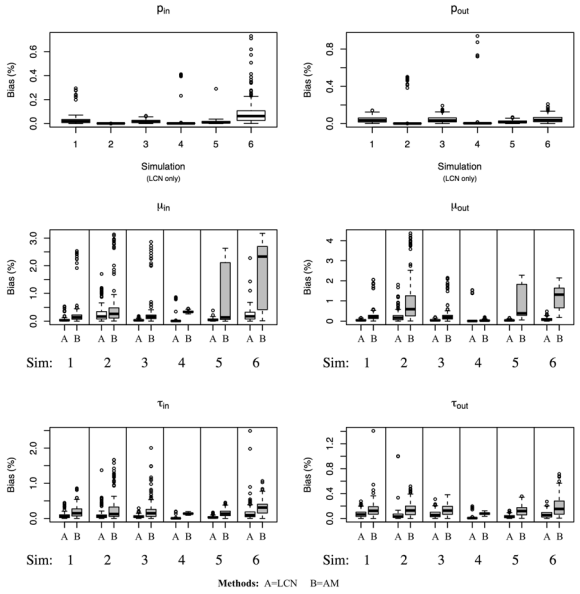

*Figure 3. Absolute deviation between the posterior median of each parameter and the true value, scaled by the true value, from each of the 6 schemes listed in Table 1. For the edge parameters $\mu_{in}, \mu_{out}, \tau_{in}, \tau_{out}$, results from our Bayesian random graph model (**A** – on the left of each panel) are compared to the method of Ambroise and Matias (**B** – on right).*

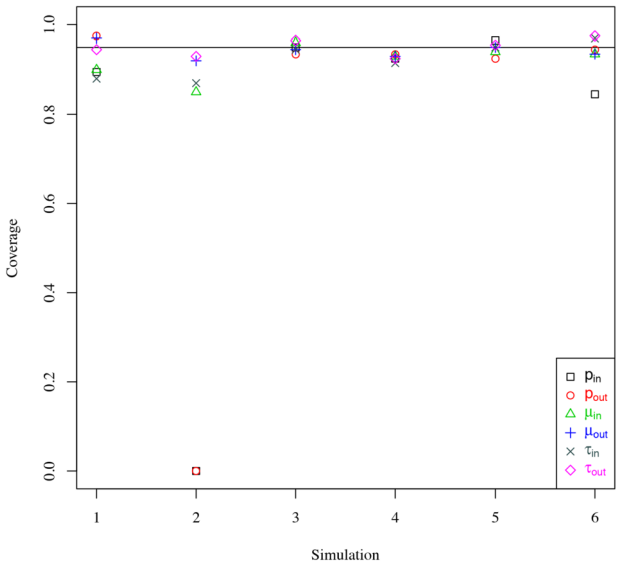

Figure 4. Percent of the 95% HPD intervals containing the true value, across 200 simulations in each scheme. Simulation schemes are listed in Table 1. The horizontal line indicates 95%.

of $(0.009, 0.08)$ hz. We deleted the scans showing movement artifacts or other problems based on the quality control information given in the phenotypic dataset and then, for the

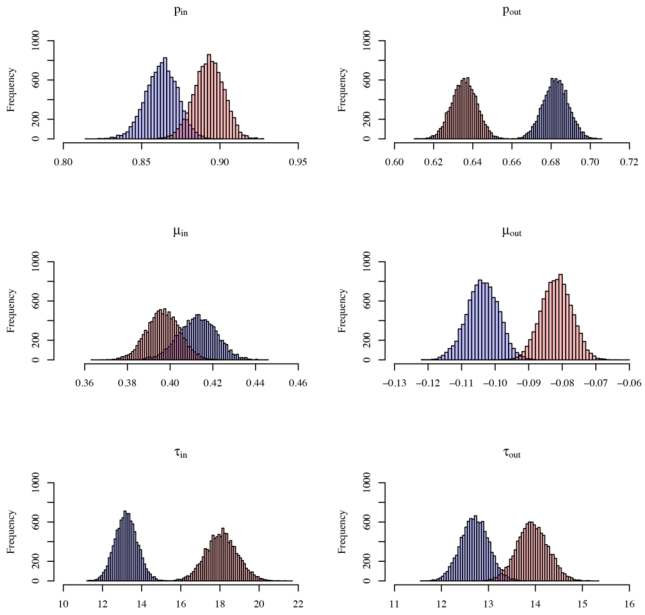

Figure 5. Posterior estimates of modularity parameters (Sparsity: p_{in} and p_{out}, edge weights: $\mu_{in}, \mu_{out}, \tau_{in}, \tau_{out}$) for subject 1 (Blue) and subject 2 (Red). First 100 samples were dropped, 9900 samples of each parameter shown.

subjects with at least one scan passing quality control, we selected a single scan for calculation of that subject's connectivity network.

The automated anatomical labeling (AAL) template [30] was used to split patients' brains into 116 non-overlapping regions of interest (ROIs); blood-oxygen-level dependent (BOLD) contrast signals were averaged within each region for each of 172 time points, and a Pearson correlation matrix was estimated for each subject's 116 ROI $\times$ 172 time point matrix. Subsequently, the elements in each 116 x 116 matrix were transformed to approximate normality via the Fisher transformation, $z = 0.5 \times \ln(\frac{1+r}{1-r})$. Additionally, the Fisher-transformed correlation matrices were thresholded at $\pm$ 0.1 (which corresponds to $r \sim \pm 0.1$) to allow for some level of sparsity.

We applied our RGMM to each subject's weighted network as follows: two parallel MCMC chains of our Gibbs sampling algorithm were run for each of Q = 3, 6, 9, and 12, and then ICL was used to choose the best of the 8 chains, which allowed for anywhere between 1 and 12 latent classes for each subject. Figure 1 shows the estimated latent classes of the 116 ROIs for two randomly selected subjects as the color of nodes in the networks; subject 1 (L) has 7 latent classes of regions and subject 2 (R) has 8 latent classes of regions. To assess the overlap of the community structures of the two subjects, the adjusted Rand Index [13] between the two clusterings was estimated to be 0.182, which is significantly different from zero (which would indicate no overlap at all) via permutation testing (10,000 permutations of

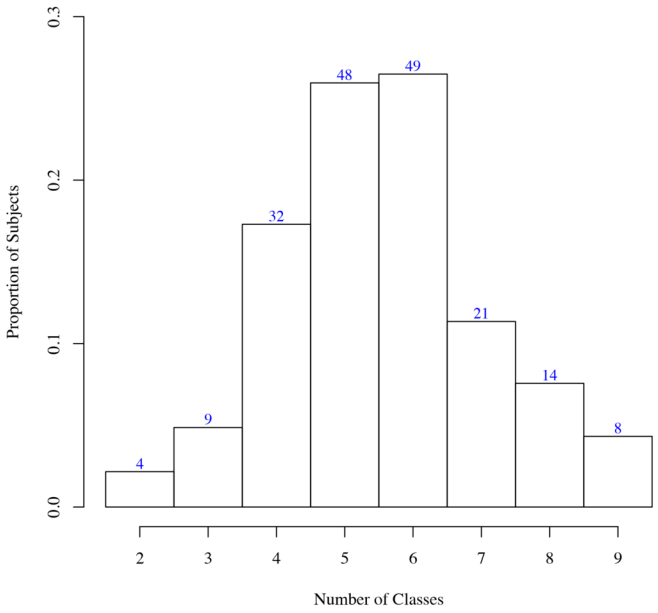

Figure 6. Number of latent classes of brain regions selected across 185 subjects from the ADHD-200 sample.

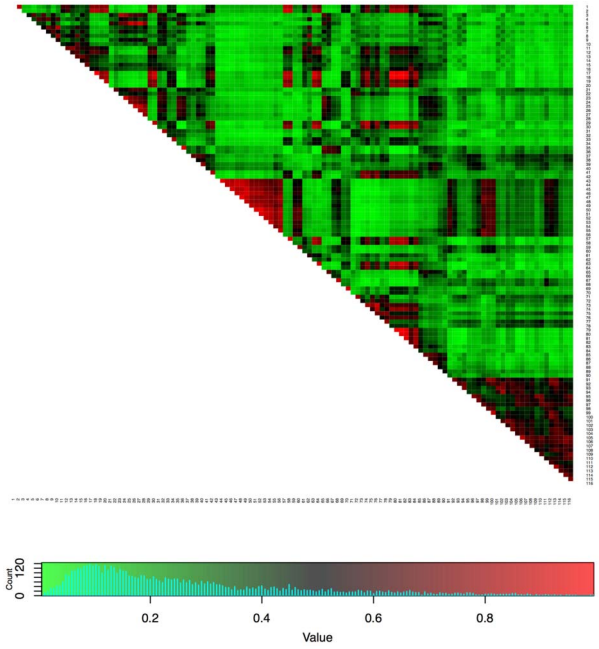

Figure 7. Overlap of the latent class structure across 185 subjects from the ADHD-200 sample. Each element of the matrix is the proportion of all 185 subjects in whom the corresponding two nodes fall in the same estimated latent class. The list of ROIs is given in Table 2 in the Appendix.

the class labels, $p < 0.001$). So, while the latent community structure for these two subjects is different, there is significant overlap between them, which suggests that there may be a shared latent structure and individual deviations from this structure.

Figure 6 shows the estimated number of latent classes across the 185 subjects, with values ranging from 2 to 9; more than half of the subjects have either 4 or 5 estimated latent classes of ROIs. In Figure 7, the overlap of the latent structures of all 185 subjects is shown; the node pairs in red are those that are in the same latent class in most networks, while the node pairs in green are in different latent classes in most networks. The functional overlap between these node pairs could be considered as the shared latent functional brain structure, while other groups of node pairs are in the same latent classes in only a subset of the subjects (see the node pairs in black, which have been estimated to be in the same latent class in approximately 50% of subjects). Additionally, the posterior distributions of the modularity parameters appear to vary across many of the subjects, indicating heterogeneity in latent community structure even beyond the latent class membership of the 116 ROIs. See Figure 5 for posterior samples of the modularity parameters of the two subjects from Figure 1.

5. DISCUSSION

We have developed the weighted affiliation SBM as a Bayesian RGMM. Our RGMM utilizes an intuitive hierarchical parametric framework that accurately captures the affiliation community structure in simulated data. The benefits of using this fully Bayesian framework include incorporation of prior data, the ability to characterize the entirety of the posterior distribution, as well as the validity of estimates and accurate classification with smaller graphs. Additionally, this approach yields estimates of the modularity of the network as parameters in the model. For highly modular graphs, in which nodes in one latent class have considerably more connections and different weights as compared to nodes in different classes, our estimation method performs well with minimal misclassification and accurate estimates of the parameters.

Within the 185 functional brain networks from the ADHD-200 sample, subjects were estimated to have between 2 and 9 latent classes of brain regions, but considerable overlap in the latent structure is seen between some subjects. The commonalities between subjects appear to include some level of symmetry in the latent classes across the left and right hemispheres, as well as the functional overlap in the regions of the occipital lobe (see the red region near the center of the diagnoal of Figure 7) and several other groups of ROIs. While the SBM is unrealistic as a model of the true data-generating process in fMRI studies, it is nonetheless a useful and principled statistical tool for uncovering the large-scale structure in correlation between ROI time courses, which can help inform future studies in functional brain connectivity.

This framework allows the flexibility to utilize different distributions for the edge weights, detect overlapping com-

munities, and estimate the community structure in directed graphs, all by straightforward alterations to the model. Current work is focused on extending this model to allow for groups of subjects that share a common structure, which appear plausible based on our analyses of the resting-state fMRI networks from the ADHD-200 sample. Additionally, we are working to incorporate regression and hypothesis testing to assess the differences in functional brain structure associated with changes in covariates.

APPENDIX A. ROI LABELS

Table 2. ROI labels for Figure 7 (From Left (1) to Right (116) and Top (1) to Bottom (116))

Label	ROI	Label	ROI
1	Precentral_L	59	Parietal_Sup_L
2	Precentral_R	60	Parietal_Sup_R
3	Frontal_Sup_L	61	Parietal_Inf_L
4	Frontal_Sup_R	62	Parietal_Inf_R
5	Frontal_Sup_Orb_L	63	SupraMarginal_L
6	Frontal_Sup_Orb_R	64	SupraMarginal_R
7	Frontal_Mid_L	65	Angular_L
8	Frontal_Mid_R	66	Angular_R
9	Frontal_Mid_Orb_L	67	Precuneus_L
10	Frontal_Mid_Orb_R	68	Precuneus_R
11	Frontal_Inf_Oper_L	69	Paracentral_Lobule_L
12	Frontal_Inf_Oper_R	70	Paracentral_Lobule_R
13	Frontal_Inf_Tri_L	71	Caudate_L
14	Frontal_Inf_Tri_R	72	Caudate_R
15	Frontal_Inf_Orb_L	73	Putamen_L
16	Frontal_Inf_Orb_R	74	Putamen_R
17	Rolandic_Oper_L	75	Pallidum_L
18	Rolandic_Oper_R	76	Pallidum_R
19	Supp_Motor_Area_L	77	Thalamus_L
20	Supp_Motor_Area_R	78	Thalamus_R
21	Olfactory_L	79	Heschl_L
22	Olfactory_R	80	Heschl_R
23	Frontal_Sup_Medial_L	81	Temporal_Sup_L
24	Frontal_Sup_Medial_R	82	Temporal_Sup_R
25	Frontal_Med_Orb_L	83	Temporal_Pole_Sup_L
26	Frontal_Med_Orb_R	84	Temporal_Pole_Sup_R
27	Rectus_L	85	Temporal_Mid_L
28	Rectus_R	86	Temporal_Mid_R
29	Insula_L	87	Temporal_Pole_Mid_L
30	Insula_R	88	Temporal_Pole_Mid_R
31	Cingulum_Ant_L	89	Temporal_Inf_L
32	Cingulum_Ant_R	90	Temporal_Inf_R
33	Cingulum_Mid_L	91	Cerebelum_Crus1_L
34	Cingulum_Mid_R	92	Cerebelum_Crus1_R
35	Cingulum_Post_L	93	Cerebelum_Crus2_L
36	Cingulum_Post_R	94	Cerebelum_Crus2_R
37	Hippocampus_L	95	Cerebelum_3_L
38	Hippocampus_R	96	Cerebelum_3_R
39	ParaHippocampal_L	97	Cerebelum_4_5_L
40	ParaHippocampal_R	98	Cerebelum_4_5_R
41	Amygdala_L	99	Cerebelum_6_L
42	Amygdala_R	100	Cerebelum_6_R
43	Calcarine_L	101	Cerebelum_7b_L
44	Calcarine_R	102	Cerebelum_7b_R
45	Cuneus_L	103	Cerebelum_8_L
46	Cuneus_R	104	Cerebelum_8_R
47	Lingual_L	105	Cerebelum_9_L
48	Lingual_R	106	Cerebelum_9_R
49	Occipital_Sup_L	107	Cerebelum_10_L
50	Occipital_Sup_R	108	Cerebelum_10_R
51	Occipital_Mid_L	109	Vermis_1_2
52	Occipital_Mid_R	110	Vermis_3
53	Occipital_Inf_L	111	Vermis_4_5
54	Occipital_Inf_R	112	Vermis_6
55	Fusiform_L	113	Vermis_7
56	Fusiform_R	114	Vermis_8
57	Postcentral_L	115	Vermis_9
58	Postcentral_R	116	Vermis_10

APPENDIX B. DISTRIBUTIONS

Prior and sampling distributions are listed as follows: The latent class for each node i is distributed as

$$P(Z_i|\pi,Q)=\pi_1^{Z_{i1}}\cdots\pi_Q^{Z_{iQ}},\quad P(\pi|Q)=\prod_{q=1}^{Q}\pi_q^{a_q-1}.$$

$P(Y|Z,\pi,Q,\theta,p)$ is given by

$$\begin{aligned}
&\prod_{1<i<j<n}P(Y_{ij}|\theta_{in},\theta_{out},p_{in},p_{out},Z_i,Z_j,Q)\\
=&\prod_{1<i<j<n}\prod_{q,l}P(Y_{ij}|\theta_{in},\theta_{out},p_{in},p_{out},Z_{iq}Z_{jl}=1,Q)\\
=&\prod_{1<i<j<n}\prod_{q=l}\left[D_{in}(i,j)\right]^{Z_{iq}Z_{jl}}\prod_{q\neq l}\left[D_{out}(i,j)\right]^{Z_{iq}Z_{jl}}\\
=&\prod_{A}\left(p_{in}(\frac{\tau_{in}}{2\pi})^{1/2}\exp\{-\frac{\tau_{in}}{2}(Y_{ij}-\mu_{in})^2\}\right)\\
&\prod_{B}\left(p_{out}(\frac{\tau_{out}}{2\pi})^{1/2}\exp\{-\frac{\tau_{out}}{2}(Y_{ij}-\mu_{out})^2\}\right)\\
&\prod_{C}(1-p_{in})\prod_{D}(1-p_{out}),
\end{aligned}$$

where $D_{in}(i,j)=$

$$\left(p_{in}\left(\frac{\tau_{in}}{2\pi}\right)^{1/2}\exp\{-\frac{\tau_{in}}{2}\left(Y_{ij}-\mu_{in}\right)^2\}\right)^{1(Y_{ij}\neq 0)}\times\left(1-p_{in}\right)^{1(Y_{ij}=0)},$$

$D_{out}(i,j)=$

$$\left(p_{out}\left(\frac{\tau_{out}}{2\pi}\right)^{1/2}\exp\{-\frac{\tau_{out}}{2}\left(Y_{ij}-\mu_{out}\right)^2\}\right)^{1(Y_{ij}\neq 0)}$$

$$\times (1-p_{out})^{1(Y_{ij}=0)},$$

and A, B, C, and D satisfy

$$\begin{aligned}
A &= \{i<j, q : Z_{iq}Z_{jq}=1, Y_{ij}\neq 0\},\\
B &= \{i<j, q\neq l : Z_{iq}Z_{jl}=1, Y_{ij}\neq 0\},\\
C &= \{i<j, q : Z_{iq}Z_{jq}=1, Y_{ij}= 0\},\\
D &= \{i<j, q\neq l : Z_{iq}Z_{jl}=1, Y_{ij}= 0\}.
\end{aligned}$$

Moreover, we set

$$P(p_{in}) = 1(0<p_{in}<1), \quad P(p_{out}) = 1(0<p_{out}<1),$$

$$P(\mu_{in}|\tau_{in}) = \left(\frac{\tau_{in}}{2\pi\sigma_{0,in}^2}\right)^{1/2} \exp\{\frac{\tau_{in}}{2\sigma_{0,in}^2}(\mu_{in}-\mu_{0,in})^2\},$$

$$P(\tau_{in}) = \beta_{0,in}^{\alpha_{0,in}} \frac{1}{\Gamma(\alpha_{0,in})} \tau_{in}^{\alpha_{0,in}-1} \exp\{-\beta_{0,in}\tau_{in}\},$$

$$P(\mu_{out}|\tau_{out}) = \left(\frac{\tau_{out}}{2\pi\sigma_{0,out}^2}\right)^{1/2} \exp\{\frac{\tau_{out}}{2\sigma_{0,out}^2}(\mu_{out}-\mu_{0,out})^2\},$$

$$P(\tau_{out}) = \beta_{0,out}^{\alpha_{0,out}} \frac{1}{\Gamma(\alpha_{0,out})} \tau_{out}^{\alpha_{0,out}-1} \exp\{-\beta_{0,out}\tau_{out}\}.$$

Then, the full conditional distributions are derived as follows:

First, we have

$$\begin{aligned}
&P(Z_i|\pi, Q, p_{in}, p_{out}, \theta, Y) \propto P(Y|\theta, p, \pi, Z, Q)P(Z_i|\pi, Q)\\
&\propto \prod_{1\leq i<j\leq n} \prod_{q=l} \left[D_{in}(i,j)\right]^{Z_{iq}Z_{jl}} \prod_{q\neq l} \left[D_{out}(i,j)\right]^{Z_{iq}Z_{jl}}\\
&\times \left(\pi_1^{Z_{i1}} \cdots \pi_Q^{Z_{iQ}}\right).
\end{aligned}$$

Therefore, the full conditional distribution of Z_i given all others is proportional to

$$\begin{aligned}
&\prod_{q=1}^{Q} \left[\prod_{j\neq i} D_{in}(i,j)^{Z_{jq}} D_{out}(i,j)^{\sum_{r\neq q} Z_{jr}}\right]\\
&\times \left(\pi_1^{Z_{i1}} \cdots \pi_Q^{Z_{iQ}}\right),
\end{aligned}$$

Thus, we have $Z_i|\ldots \sim \text{Multinomial}(\tilde{\pi}_{i1}, \ldots \tilde{\pi}_{iQ})$, where $\tilde{\pi}_{iq}$ is given by

$$\tilde{\pi}_{iq} = \frac{\pi_q \prod_{j\neq i} D_{in}(i,j)^{Z_{jq}} D_{out}(i,j)^{\sum_{k\neq q} Z_{jk}}}{\sum_{q=1}^{Q} \pi_q \prod_{j\neq i} D_{in}(i,j)^{Z_{jq}} D_{out}(i,j)^{\sum_{k\neq q} Z_{jk}}} \text{ for } q=1,\ldots Q.$$

The full conditional distribution of π is given by

$$\begin{aligned}
P(\pi|Q, Z, Y) &\propto P(Z|\pi, Q)P(\pi|Q)\\
&\propto \left(\prod_{i=1}^{n} \pi_1^{Z_{i1}} \cdots \pi_Q^{Z_{iQ}}\right) \left(\prod_{q=1}^{Q} \pi_q^{a_q-1}\right)\\
&\propto \pi_1^{\sum_{i=1}^{n} Z_{i1}+a_1-1} \cdots \pi_Q^{\sum_{i=1}^{n} Z_{iQ}+a_Q-1},
\end{aligned}$$

which implies that

$$\pi|\ldots \sim \text{Dirichlet}\left(\sum_{i=1}^{n} Z_{i1}+a_1, \ldots \sum_{i=1}^{n} Z_{iQ}+a_Q\right).$$

The full conditional distribution of p_{in} is given by

$$P(Y|\theta, p, \pi, Z, Q)P(p_{in}) \propto p_{in}^{n_A}(1-p_{in})^{n_C} \cdot 1(0<p_{in}<1),$$

so $p_{in}|\ldots \sim \text{Beta}(n_A+1, n_C+1)$, where $n_A = |A|$ and $n_B = |B|$. Similarly, we have

$$P(p_{out}|\cdots) \sim \text{Beta}(n_B+1, n_D+1),$$

where $n_C = |C|$ and $n_D = |D|$.

The full conditional distribution of μ_{in} is given by

$$\begin{aligned}
&P(\mu_{in}|\tau, \pi, Z, Q, p, Y) \propto P(Y|\theta, p, \pi, Z, Q)P(\mu_{in}|\tau_{in}^{-1})\\
&\propto \exp\{\sum_{A} -\frac{\tau_{in}}{2}(Y_{ij}-\mu_{in})^2\}\\
&\times \exp\{-\frac{\tau_{in}}{2\sigma_{0,in}^2}(\mu_{in}-\mu_{0,in})^2\},
\end{aligned}$$

which implies that

$$\mu_{in}|\cdots \sim \text{N}\left(\frac{\sigma_{0,in}^2 \sum_{A} Y_{ij} + \mu_{0,in}}{n_A\sigma_{0,in}^2+1}, \frac{\sigma_{0,in}^2\tau_{in}^{-1}}{n_A\sigma_{0,in}^2+1}\right).$$

Similarly, we have

$$\mu_{out}|\cdots \sim \text{N}\left(\frac{\sigma_{0,out}^2 \sum_{B} Y_{ij} + \mu_{0,out}}{n_A\sigma_{0,out}^2+1}, \frac{\sigma_{0,out}^2\tau_{out}^{-1}}{n_B\sigma_{0,out}^2+1}\right).$$

The full conditional distribution of τ_{in} is given by

$$\begin{aligned}
&P(\tau_{in}|\mu, \pi, Z, Q, p, Y) \propto P(Y|\theta, p, \pi, Z, Q)\\
&\times P(\mu_{in}|\tau_{in})P(\tau_{in})\\
\propto\ &\left[\prod_{A} \left(p_{in}(\frac{\tau_{in}}{2\pi})^{1/2} \exp\{-\frac{\tau_{in}}{2}(Y_{ij}-\mu_{in})^2\}\right)\right]\\
&\times \left(\frac{\tau_{in}}{2\pi\sigma_{0,in}^2}\right)^{1/2} \exp\{\frac{\tau_{in}}{2\sigma_{0,in}^2}(\mu_{in}-\mu_{0,in})^2\}\\
&\times \beta_{0,in}^{\alpha_{0,in}} \frac{1}{\Gamma(\alpha_{0,in})} \tau_{in}^{\alpha_{0,in}-1} \exp\{-\beta_{0,in}\tau_{in}\}\\
\propto\ &\tau_{in}^{\left(\frac{n_A+1}{2}+\alpha_{0,in}-1\right)} \exp\{\frac{-\tau_{in}}{2}\sum_{A}(Y_{ij}-\mu_{in})^2\}
\end{aligned}$$

$$\times \exp\{-\frac{\tau_{in}}{2\sigma^2_{0,in}}\left(\mu_{in}-\mu_{0,in}\right)^2-\beta_{0,in}\tau_{in}\},$$

which implies that

$$\tau_{in}|\cdots \sim \text{Gamma}\left(\alpha^*_{in},\beta^*_{in}\right),$$

where

$$\alpha^*_{in}=\frac{n_A+1}{2}+\alpha_{0,in}$$

and

$$\beta^*_{in}=\frac{1}{2}\sum_A\left(Y_{ij}-\mu_{in}\right)^2+\frac{1}{2\sigma^2_{0,in}}\left(\mu_{in}-\mu_{0,in}\right)^2+\beta_{0,in}\,.$$

Similarly, we have

$$\tau_{out}|\cdots \sim \text{Gamma}\left(\alpha^*_{out},\right),$$

with

$$\alpha^*_{out}=\frac{n_B+1}{2}+\alpha_{0,out},$$

and

$$\beta^*_{out}=\frac{1}{2}\sum_B\left(Y_{ij}-\mu_{out}\right)^2+\frac{1}{2\sigma^2_{0,out}}\left(\mu_{out}-\mu_{0,out}\right)^2+\beta_{0,out}\,.$$

Received 12 May 2016

REFERENCES

[1] Aicher, C., Jacobs, A. Z. and Clauset, A. (2015). Learning latent block structure in weighted networks. *Journal of Complex Networks* **3** 221–248. MR3365464

[2] Ambroise, C. and Matias, C. (2012). New consistent and asymptotically normal parameter estimates for random-graph mixture models. *Journal of the Royal Statistical Society. Series B* **74** 3–35. MR2885838

[3] Bullmore, E. and Sporns, O. (2009). Complex brain networks: graph theoretical analysis of structural and functional systems. *Nature Reviews. Neuroscience* **10** 186–98.

[4] Choi, D. S., Wolfe, P. J. and Airoldi, E. M. (2012). Stochastic blockmodels with a growing number of classes. *Biometrika* **99** 273–284. MR2931253

[5] Daudin, J. J., Picard, F. and Robin, S. (2008). A mixture model for random graphs. *Statistics and Computing* **18** 173–183. MR2390817

[6] Erdös, P. and Rényi, A. (1959). On random graphs I. *Publ. Math. Debrecen* **6** 290–297. MR0120167

[7] Fortunato, S. (2010). Community detection in graphs. *Physics Reports* **486** 75–174. MR2580414

[8] Fortunato, S. and Barthélemy, M. (2007). Resolution limit in community detection. *Proceedings of the National Academy of Sciences of the United States of America* **104** 36–41.

[9] Gao, W., Zhu, H., Giovanello, K. S., Smith, J. K., Shen, D., Gilmore, J. H. and Lin, W. (2009). Evidence on the emergence of the brain's default network from 2-week-old to 2-year-old healthy pediatric subjects. *Proceedings of the National Academy of Sciences* **106** 6790–6795.

[10] Girvan, M. and Newman, M. E. J. (2002). Community structure in social and biological networks. *Proceedings of the National Academy of Sciences of the United States of America* **99** 7821–7826. MR1908073

[11] Hartigan, J. A. and Wong, M. A. (1979). Algorithm AS 136: a k-means clustering algorithm. *Journal of the Royal Statistical Society C* **28** 100–108.

[12] Hofman, J. M. and Wiggins, C. H. (2008). A Bayesian approach to network modularity. *Physical Review Letters* **100**.

[13] Hubert, L. and Arabie, P. (1985). Comparing partitions. *Journal of Classification* **2** 193–218.

[14] Kolaczyk, E. D. (2009). *Statistical Analysis of Network Data: Methods and Models*. Springer Science+Business Media. MR2724362

[15] Lancichinetti, A., Kivelä, M., Saramäki, J. and Fortunato, S. (2010). Characterizing the community structure of complex networks. *PloS One* **5** e11976.

[16] Lichtman, J. W., Pfister, H. and Shavit, N. (2014). The big data challenges of connectomics. *Nature Neuroscience* **17** 1448–1454.

[17] Mariadassou, M., Robin, S. and Vacher, C. (2010). Uncovering latent structure in valued graphs: a variational approach. *Annals of Applied Statistics* **4** 715–742. MR2758646

[18] Matias, C. and Robin, S. (2014). Modeling heterogeneity in random graphs: a selective review. *arXiv preprint arXiv:1402.4296*. MR3419385

[19] Newman, M. E. J. and Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical Review E* **69** 16.

[20] Nowicki, K. and Snijders, T. A. B. (2001). Estimation and prediction for stochastic blockstructures. *Journal of the American Statistical Association* **96** 1077–1087. MR1947255

[21] Peng, L. and Carvalho, L. (2013). Bayesian degree-corrected stochastic blockmodels for community detection. *arXiv preprint arXiv:1309.4796* 1–23.

[22] Plummer, M., Best, N., Cowles, K. and Vines, K. (2006). CODA: convergence diagnosis and output analysis for MCMC. *R News* **6** 7–11.

[23] Reichardt, J. and Bornholdt, S. (2006). Statistical mechanics of community detection. *Physical Review E* **74** 16110. MR2276596

[24] Schweinberger, M. and Handcock, M. S. (2015). Local dependence in random graph models: characterisation, properties, and statistical inference. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **77** 647–676. MR3351449

[25] Schweinberger, M. and Luna, P. (2015). HERGM: hierarchical exponential-family random graph models. *Journal of Statistical Software*.

[26] Simpson, S. L., Bowman, F. D. and Laurienti, P. J. (2013). Analyzing complex functional brain networks: fusing statistics and network science to understand the brain. *Statistics Surveys* **7** 1–36. MR3161730

[27] Snijders, T. A. (2011). Statistical models for social networks. *Annual Review of Sociology* **37** 131–153.

[28] R Core Team (2015). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

[29] Traag, V. A. and Bruggeman, J. (2009). Community detection in networks with positive and negative links. *Physical Review E – Statistical, Nonlinear, and Soft Matter Physics* **80** 7.

[30] Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N., Mazoyer, B. and Joliot, M. (2002). Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. *NeuroImage* **15** 273–289.

[31] Vu, D. Q., Hunter, D. R. and Schweinberger, M. (2013). Model-based clustering of large networks. *Annals of Applied Statistics* **7** 1010–1039. MR3113499

[32] Xia, M., Wang, J. and He, Y. (2013). BrainNet Viewer: a network visualization tool for human brain connectomics. *PloS One* **8** e68910.

[33] Yang, J. and Leskovec, J. (2014). Structure and overlaps of ground-truth communities in networks. *ACM Transactions on Intelligent Systems and Technology* **5** 26.

[34] Zhang, Y., Friend, A. J., Traud, A. L., Porter, M. A., Fowler, J. H. and Mucha, P. J. (2008). Community structure in congressional cosponsorship networks. *Physica A: Statistical Mechanics and its Applications* **387** 1705–1712.

Christopher Bryant
135 Dauer Drive
3101 McGavran-Greenberg Hall
CB #7420
Chapel Hill, NC 27599-7420
USA
E-mail address: cmbrya@unc.edu

Hongtu Zhu
1515 Holcombe Blvd.
Unit 1411
Houston, TX 77030
USA
E-mail address: hzhu5@mdanderson.org

Mihye Ahn
1664 N. Virginia Street
Reno, NV 89557
USA
E-mail address: mahn@unr.edu

Joseph Ibrahim
135 Dauer Drive
3101 McGavran-Greenberg Hall
CB #7420
Chapel Hill, NC 27599-7420
USA
E-mail address: ibrahim@bios.unc.edu