NeuroImage 220 (2020) 116611

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/neuroimage

Multi-subject Stochastic Blockmodels for adaptive analysis of individual differences in human brain network cluster structure

Dragana M. Pavlović [a,b,*], Bryan R.L. Guillaume [a,c], Emma K. Towlson [d,e], Nicole M.Y. Kuek [b], Soroosh Afyouni [a], Petra E. Vértes [f], B.T. Thomas Yeo [b], Edward T. Bullmore [f,g,h], Thomas E. Nichols [a,i]

[a] Big Data Institute, Nuffield Department of Population Health, University of Oxford, Oxford, United Kingdom
[b] Department of Electrical and Computer Engineering, Clinical Imaging Research Centre, N.1 Institute for Health and Memory Networks Programme, National University of Singapore, Singapore
[c] Department of Biomedical Engineering, National University of Singapore, Singapore
[d] Center for Complex Network Research and Department of Physics, Northeastern University, Boston, MA, United States
[e] Media Laboratory, Massachusetts Institute of Technology, Cambridge, MA, United States
[f] Behavioural and Clinical Neuroscience Institute, Department of Psychiatry, University of Cambridge, Cambridge, United Kingdom
[g] Cambridgeshire and Peterborough National Health Service Foundation Trust, Cambridge, United Kingdom
[h] GlaxoSmithKline, Clinical Unit Cambridge, Addenbrooke's Hospital, Cambridge, United Kingdom
[i] Warwick Manufacturing Group, University of Warwick, Coventry, United Kingdom

ARTICLE INFO

Keywords:
Mixture models
Stochastic blockmodel
Stochastic block model
Community detection
Modularity
Variational approximation
Firth estimation
Wald test
Likelihood ratio
Permutation test
Network analysis
Multi-subject network analysis
Integrated classification likelihood criterion

ABSTRACT

There is considerable interest in elucidating the cluster structure of brain networks in terms of modules, blocks or clusters of similar nodes. However, it is currently challenging to handle data on multiple subjects since most of the existing methods are applicable only on a subject-by-subject basis or for analysis of an average group network. The main limitation of per-subject models is that there is no obvious way to combine the results for group comparisons, and of group-averaged models that they do not reflect the variability between subjects. Here, we propose two new extensions of the classical Stochastic Blockmodel (SBM) that use a mixture model to estimate blocks or clusters of connected nodes, combined with a regression model to capture the effects of subject-level covariates on individual differences in cluster structure. The proposed Multi-Subject Stochastic Blockmodels (MS-SBMs) can flexibly account for between-subject variability in terms of homogeneous or heterogeneous covariate effects on connectivity using subject demographics such as age or diagnostic status. Using synthetic data, representing a range of block sizes and cluster structures, we investigate the accuracy of the estimated MS-SBM parameters as well as the validity of inference procedures based on the Wald, likelihood ratio and permutation tests. We show that the proposed multi-subject SBMs recover the true cluster structure of synthetic networks more accurately and adaptively than standard methods for modular decomposition (i.e. the Fast Louvain and Newman Spectral algorithms). Permutation tests of MS-SBM parameters were more robustly valid for statistical inference and Type I error control than tests based on standard asymptotic assumptions. Applied to analysis of multi-subject resting-state fMRI networks (13 healthy volunteers; 12 people with schizophrenia; $n = 268$ brain regions), we show that Heterogeneous Stochastic Blockmodel (Het-SBM) identifies a range of network topologies simultaneously, including modular and core structures.

1. Introduction

Network-like representations of brain connectivity (e.g., functional, structural or causal associations) allow us to explore the link between the architecture of the brain and the way it facilitates (i) specialised segregative processes and (ii) complex integrative processes. Two network markers, 'modular structure' and 'rich-club' have shaped our understanding of brain networks. The 'modular structure' represents a decomposition of the brain network into clusters of densely connected nodes, whose ties to the other clusters in the network are much sparser,

* Corresponding author. Big Data Institute, Nuffield Department of Population Health, University of Oxford, Oxford, United Kingdom.
E-mail address: pavlovic.mile.dragana@gmail.com (D.M. Pavlović).

https://doi.org/10.1016/j.neuroimage.2020.116611
Received 7 June 2019; Received in revised form 31 January 2020; Accepted 4 February 2020
Available online 10 February 2020

and, as suggested by Rubinov and Sporns (2010), this feature is informative of specialised segregative processes. Traditionally, a modular structure is detected with the Newman Spectral (Newman, 2006) or Fast Louvain (Blondel et al., 2008) algorithms. On the other hand, rich-club is a concept such that the high degree nodes tend to be more densely connected to each other than to the lower degree nodes and it is believed to be informative of complex integrative processes (see, e.g., van den Heuvel and Sporns, 2013, for a review). Nevertheless, as demonstrated in the brain network of roundworm *Caenorhabditis elegans* (*C. elegans*) by Pavlovic et al. (2014), both aspects of this organisation can simultaneously be captured by a more general concept of 'cluster structure', which is defined as the average proportion of edges; within each cluster, and between each cluster pair. This fitted cluster structure typically provides the individual cluster sizes, the total number of clusters Q and a $Q \times Q$ matrix of within-cluster and between-cluster connectivity averages $\boldsymbol{\pi}$. The variations in mean connectivity values among the cell elements of $\boldsymbol{\pi}$, traditionally called 'blocks', allow us to characterise the overall network organisation and quantify the portions that are modular and those that are hub-like. In the case of *C. elegans*, the cluster structure turned out to be a combination of two rich club clusters, comprising sensory neurons which regulate forward and backward locomotions while the remainder of the clusters were found to be modular. Thus, from a purely methodological view, the cluster structure can be thought of as a rich and elegant tool which marries many isolated graph-theoretical measures, yet remains general enough to capture a catalogue of network organisations, including 'core-periphery', 'disassortative', 'star-patterned' structures and more (see, e.g., Daudin et al., 2008; Picard et al., 2009; Matias and Robin, 2014; Betzel et al., 2018b, for a review), whose occurrence and functional relevances in brain networks are open for discussion.

Beyond *C. elegans*, the analysis of human brain connectivity data is a rapidly growing field of study with a current focus on between-group comparisons between brains of patients and healthy controls in multi-subject datasets (Van Den Heuvel and Pol, 2010). Despite the need for comparative network studies, there has not been much progress in extending single network data analyses of cluster structure to multi-subject data analyses and between-group comparisons. The reasons for this can be found in a set of challenges brought about by the multi-subject nature of the data, including:

(i) The need to estimate a common network decomposition over subjects while accounting for between-subject variability in connectivity rates.
(ii) How to use such a network decomposition to infer differences between populations (e.g., Cases vs Controls) or effects of covariates.

Furthermore, ground truth about cluster structure in human neuroimaging data is generally unknown, and assumptions on its form may not be prudent. Hence, there is a pressing need for a statistical framework that provides estimates of cluster labels and general cluster structure, independent from prior assumptions about its form (e.g., assuming that the cluster structure can only be various shades of modular). In line with the multi-subject nature of research on human brain networks, this framework should ideally be rich enough to accommodate for the above mention set of challenges. To address this gap, we follow the class of probabilistic network clustering models originated in the work of Snijders and Nowicki (Nowicki and Snijders, 2001; Snijders and Nowicki, 1997), called Stochastic Blockmodel (SBM). The SBM uses a framework of mixture models to describe heterogeneity in the distribution of network edges. Principally, one type of probability density is selected for the entire network, such as the Bernoulli density, and each cluster and cluster pair has its own Bernoulli parameter. These parameters represent the mean values in the cluster structure $\boldsymbol{\pi}$ or, in this specific example, they represent the probability that there is an edge between any two of the nodes which can either be in (i) the same cluster or (ii) two different clusters. However, as the clusters can be of different sizes, we, therefore, need a parameter $\boldsymbol{\alpha}$ to indicate the proportion of each cluster in the overall mixture of Bernoulli densities. Hence, the larger the cluster, the more likely it is that a randomly selected node falls into it and, thus, is expected to contribute more to the mixture.

To obtain the estimates of $\boldsymbol{\pi}, \boldsymbol{\alpha}$ and the latent cluster labels, Snijders and Nowicki (1997) considered maximum likelihood estimates (ML) (based on Gibbs sampling and Expectation Maximisation (EM) algorithm) and showcased some computational challenges in the optimisation that limited these techniques to small networks (e.g., < 100 nodes). A trivial change of network probability density and reparametrisation of the SBM of Snijders and Nowicki (1997) was also considered in the work of Newman and Leicht (2007), in which the authors referred to it as the 'Newman and Leicht model'. More recently, Daudin et al. (2008) introduced frequentist variational approximation as an optimisation strategy of Snijders and Nowicki's SBM and applied a criterion of Biernacki et al. (1998) to measure the amount of variation in the data explained by a particular network partition. This criterion allowed not only to estimate the optimal number of clusters but also to quantify the overall goodness of fit of each candidate clustering in the dataset. Since then, the SBMs have been adapted for various biological datasets, including the Overlapping SBM (Latouche et al., 2011, 2014), the closely related Mixed-membership SBM (Airoldi et al., 2009; Mørup et al., 2011), the bayesian SBM (Latouche et al., 2012; Mørup and Schmidt, 2012; Schmidt and Mørup, 2013; Ambrosen et al., 2013, 2014; Andersen et al., 2014; Côme and Latouche, 2015), the online SBM (Zanghi et al., 2008), the SBM with nodal features (Zanghi et al., 2010) and the SBM with edge features by Mariadassou et al. (2010). SBMs have been applied in the analysis of macaque anatomical cortex (Picard et al., 2009), *C. elegans* brain network (Pavlovic et al., 2014) and human brain networks (Hinne et al., 2015; Moyer et al., 2015; Pavlovic, 2015). Of note, SBM hybrids have also been considered in the analysis of dynamic connectivity (Matias and Miele, 2017) and in neuroimaging (Robinson et al., 2015). Theoretical properties of the SBMs have additionally been studied in the work of Bickel et al. (2013), Choi et al. (2012), Ambroise and Matias (2012), Wolfe and Olhede (2013), Olhede and Wolfe (2014), Gao et al. (2015) and Mossel et al. (2016) to mention a few.

From this literature, we especially highlight the work of Mariadassou et al. (2010) who developed the generalised SBM for the analysis of a single network. We refer to this model as generalised SBM for two reasons. First, the variational fitting procedure of SBMs holds for any distribution from the exponential family, which makes it applicable to a wide range of datasets. Second, the authors used generalised linear models to statistically link edge-based features (or covariates) and network cluster structure $\boldsymbol{\pi}$. The elegance of this model is that it allows us to study edge-covariate effects, which affect connectivity within each element of the cluster structure either with the same global intensity (homogeneous effects), or with differing intensities (heterogeneous effects). It is also interesting to note that, in both their real data analyses and simulation experiments, the authors only investigated the 'homogeneous effect' version of their model and, while the 'heterogeneous effect' model was defined, the behaviour of this model and its domain of application was not explored.

Inspired by the work of Mariadassou et al. (2010), we consider three Multi-Subject SBMs (MS-SBMs), Binomial SBM (Bin-SBM), Homogeneous SBM (Hom-SBM) and Heterogeneous SBM (Het-SBM). In order to understand the conceptual differences between these multi-subject SBMs, we have simulated three simple examples that could arise from these three models. For this, we have considered three subjects, aged 20, 40 and 90 years. As shown in Fig. 1 (a), Bin-SBM poses the same block probabilities for all subjects, regardless of their age; hence, there is almost no variation in the connectivity between them. In contrast, Hom-SBM poses a negative effect of age so that with increasing age, there is a decrease in connectivity over the entire cluster structure for all subjects (see Fig. 1 (b)). Consequently, the variability of the cluster structure is tuned to tolerate only minor variations across subjects, as the model only accounts for

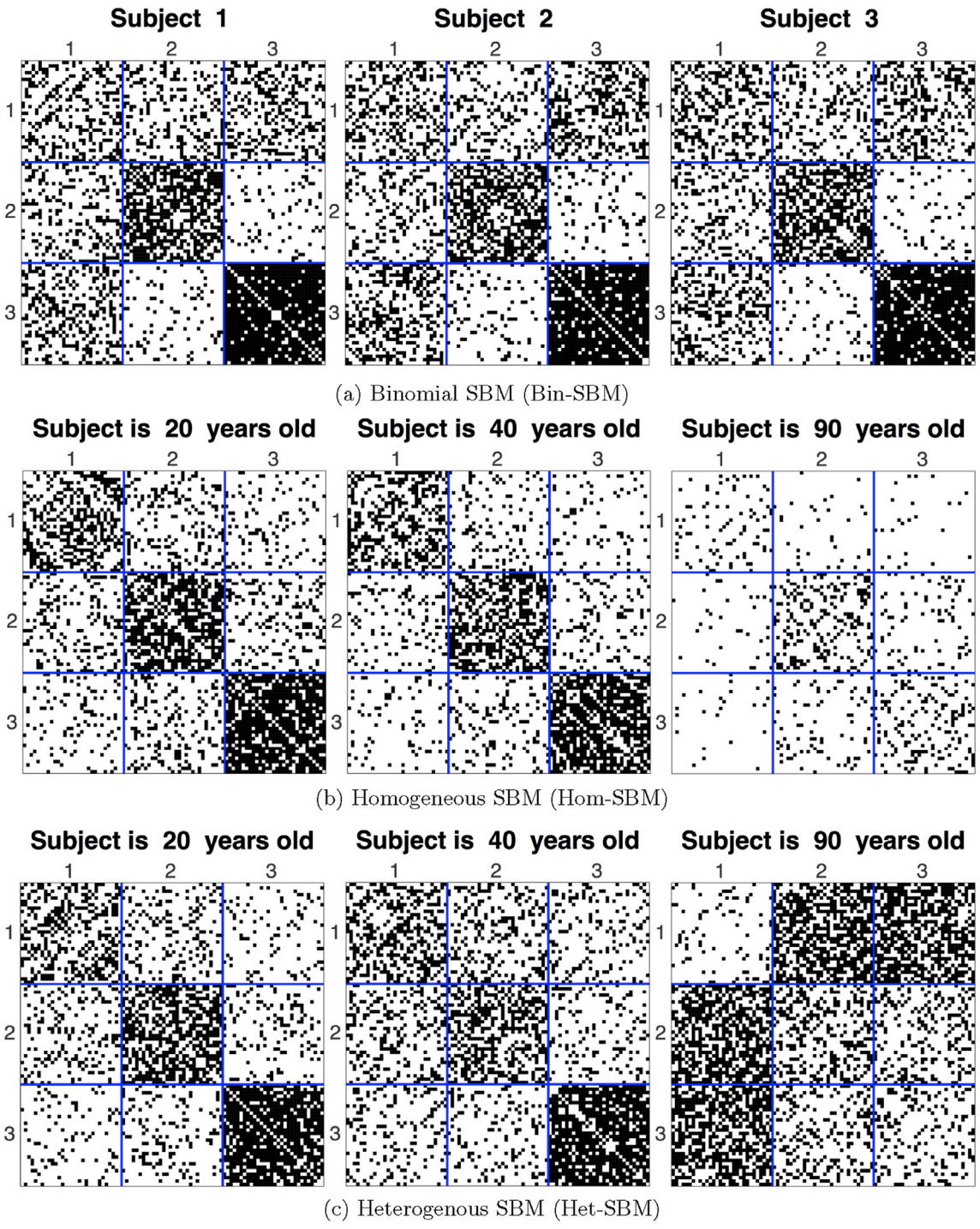

Fig. 1. Conceptual differences between three multi-subject SBM models. Simulated data are provided for three subjects aged 20, 40 and 90 years, and each subject's network is represented as a reorganised adjacency matrix with three blocks, labelled numerically (1–3). (a) Binomial SBM (Bin-SBM) does not use the subjects' age to model the variability between the three subjects. The three individual networks thus conform very closely to each other. (b) In Homogeneous SBM (Hom-SBM), subject level covariates can affect connectivity probabilities globally by increasing or decreasing the strength of connectivity across all blocks. In these data, age is modelled as a subject-level covariate, and increasing age is associated with globally decreasing connectivity across all blocks and all subjects. (c) In Heterogeneous SBM (Het-SBM), the subject-level covariates can affect connectivity probabilities specifically within each block or between each block pair. The effect of age is seen as a locally heterogeneous increase or decrease of connectivity in each block. For example, the intra-block connectivity in Block (1, 1) decreases as a function of age whereas the inter-block connectivity in Block (1, 2) increases with age.

unidirectional covariate effects. In that regard, Het-SBM is much more flexible as it allows the effect of age in each block to decrease or increase independently and, thus, different types of cluster structures can be observed across subjects (see Fig. 1 (c)). For example, the first subject has a modular structure while the third subject has an overall loss of modular structure in Block 1 (or Block (1, 1)) as there is an increase of connectivity patterns in Blocks (1, 2) and (1, 3) which is 'disassortative' (Hu and Wang, 2009) (inter-block connections > intra-block connections). Although it is implausible that such an extreme variation between subjects is present in real brain data, it is noteworthy that the model is rich and flexible enough

to handle this degree of structural variation. For these reasons, we mainly focus on Het-SBM in this work.

As shown in Fig. 1, these models assume that the cluster labels are fixed across all subjects, but there is an inter-subject variation through subject covariates in a logistic regression model. To our knowledge, this combination of a constrained block structure and an inter-subject regression model is novel. Unlike the model in Mariadassou et al. (2010) which utilises edge-based covariates, the new model class we propose utilises covariates which vary by subject and not by nodes or by edges. This modelling decision is guided by the high explanatory power of primary subject covariates, like age and gender (Dosenbach et al., 2010). Furthermore, the inclusion of subject variables allows for inferences between subject populations, such as testing for differences between patient and control groups. In that regard, not much literature about hypothesis testing for multi-subject networks exists. One such work, Alexander-Bloch et al. (2012), discussed how statistical inference could be utilised to compare some global network statistics or some nodal statistics between two groups of networks. However, a significant group difference in a global network statistic does not allow us to locate where the group difference occurs within the multi-subject network. While this shortcoming can be alleviated using nodal statistics, these statistics have a high degree of noise and may yield a large number of tests (at least, one per node). In this paper, instead, we utilise our proposed MS-SBMs to make statistical inference on the cluster-wise regression parameters. Such test statistics (i) have a lower level of noise than nodal statistics (due to the pooling effect of the nodes in each cluster), (ii) are localised within the network (i.e. by noting the nodes in each involved cluster), and (iii) may potentially yield less tests than nodal statistics. Although we do not dismiss the potential value of including edge/node-based covariates (e.g., one could consider a distance-based edge covariate) in our proposed MS-SBMs, we view these to be secondary to the value of an inter-subject regression framework and beyond the scope of the present work.

The remainder of this paper will be organised as follows. First, we will use Bin-SBM to review the frequentist variational approximation and the model selection criterion (ICL). Second, we will define Het-SBM for which we discuss an estimation procedure based on a variational approximation and Firth regularisation. Similar derivations for Hom-SBM can be found in Supplemental Information (**SI**) **D**. Third, hypothesis testing (or equivalently inference) procedures based on parametric and non-parametric tests like the Wald, likelihood ratio and permutation tests will be described. Fourth, Monte Carlo simulations will be used to evaluate the accuracy of the cluster structure estimated by Bin-SBM and Het-SBM, compared to two prior methods for modular decomposition: Newman's Spectral algorithm and the Fast Louvain algorithm. As an additional comparison, we will also consider a block-to-node parametrisation of the SBM (BL-SBM), given in the work of Newman and Leicht (2007). This SBM will be assessed only on a slice of our simulation because it is a single-network model and, as the authors did not implement a strategy to estimate Q, we will benchmark this model only on the ground truth Q. Fifth, advantages of using Het-SBM over Bin-SBM will be shown, and an example in which the between-subject variation strongly impacts the estimated cluster structure and cluster labels will be raised. Sixth, using simulations, we will assess the performance of parametric (the Wald and likelihood ratio tests) and non-parametric tests (permutation test) in terms of their control of false positives (Type I error). Seventh and lastly, we will apply Het-SBM to a resting-state fMRI study with 25 subjects split into two groups: healthy subjects (Controls) and subjects diagnosed with schizophrenia (Patients).

2. Methods

In this section, we first setup the preliminary notation, and then we describe the MS-SBMs followed by their associated hypothesis testing procedures based on the Wald, likelihood ratio and permutation tests.

2.1. Notation

We will employ the usual statistical convention of Roman capital letters to denote random variables and lower case letters to denote their observed realisations. Scalar and non-scalar values are denoted by lightface and boldface fonts, respectively. For example, a non-scalar random variable will be denoted as $\boldsymbol{X}$ and its realisation as $\boldsymbol{x}$. If $\boldsymbol{X}$ is a discrete random variable, $f_{\boldsymbol{X}}(\boldsymbol{x}) \equiv f(\boldsymbol{x}) \equiv \mathrm{P}(\boldsymbol{X}=\boldsymbol{x})$ are three equivalent notations for its probability mass function or discrete density. Densities conditional on a random variable are written using a vertical bar, as in $f(\boldsymbol{x}|\boldsymbol{z})$. Furthermore, we will be using the notation $\mathbb{E}_{f^*}[g(X)]$ with f^* indicating the density of X. That is, $\mathbb{E}_{f^*}[g(X)] = \sum_{x\in S} g(x) f^*(x)$, where S is the sample space of X. As shown in Eq. (5), the notation is utilised to establish the variational nature of the underlying distribution.

In both models, the random variable X_{ijk} represents the possibility of an edge between the nodes V_i and V_j in the k-th subject, and x_{ijk} denotes a binary realisation of X_{ijk} with 1 being an edge and 0 no edge. Therefore, for the k-th subject, $\boldsymbol{X}_k = ((X_{ijk}))_{1\le i\neq j\le n}$ denotes an $n \times n$ random, symmetric adjacency matrix with elements X_{ijk}, and, for a total of K subjects, $\boldsymbol{X}$ denotes the set of independently distributed random matrices $\boldsymbol{X} = \{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_K\}$. The individual matrices $\boldsymbol{x}_k = ((x_{ijk}))_{1\le i\neq j\le n}$ are assumed to be undirected, without self-connected nodes and without multiple-edges between the nodes. Hence, they are binary and symmetric matrices with 0s on their principal diagonal and a total of $n(n-1)/2$ data points.

The goal of each model is to estimate a common cluster structure among K subjects. Thus, both multi-subject models assume that the set of nodes, labelled as $\{V_1, \ldots, V_n\}$, is divided into Q unknown (latent) blocks or clusters. Block membership of a particular node V_i, will be indicated by a random vector $\boldsymbol{Z}_i = (Z_{i1}, \ldots, Z_{iQ})$ whose elements Z_{iq} take the value 1 if $V_i \in q$-th group and 0 otherwise. Pooling this information across nodes, the $n \times Q$ random matrix $\boldsymbol{Z}$ can be defined such that the vectors $\boldsymbol{Z}_i$ are mutually independent and follow a categorical distribution with Q possible outcomes

$$\boldsymbol{Z}_i \;\sim\; \mathrm{Categorical}(Q, \boldsymbol{\alpha}), \tag{1}$$

where $\boldsymbol{\alpha}$ is a $1 \times Q$ dimensional vector of success probabilities $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_Q)$ and $\sum_q^Q \alpha_q = 1$. Specifically, if we assume that the cluster labels for each node are given, we interpret an individual α_q as the probability that a randomly selected node falls into the q-th block. Here, it is important to note that the choice of categorical (or equivalently single trial multinomial) distribution implies that fitted blocks form a partition of all nodes, in which each node belongs to only one block (i.e. disjoint blocks). This is formally noted as $\sum_{q=1}^Q z_{iq} = 1$.

2.2. Binomial Stochastic Blockmodel (Bin-SBM)

In this model, all subjects are assumed to have the same connectivity profile, such that the distribution of edges is not influenced by subject-specific information (covariates). Consequently, for the k-th subject, the edges are assumed to follow a Bernoulli distribution

$$X_{ijk} \big| Z_{iq} = 1, Z_{jl} = 1 \;\sim\; \mathrm{Bernoulli}(\pi_{ql}), \tag{2}$$

where π_{ql} is the connectivity rate that expresses the probability that nodes are connected between (q,l)-th blocks. For all blocks, connectivity rates are compiled into a $Q \times Q$ matrix $\boldsymbol{\pi}$ wherein the within-block rates are located on the diagonal, and between-block rates are located on the off-diagonal of this matrix. As the edges are undirected, the $\boldsymbol{\pi}$ matrix is symmetrical (i.e. $\pi_{ql} = \pi_{lq}$). In particular, each block-specific connectivity rate π_{ql} is the expected mean of its edges (i.e. $\pi_{ql} = \mathbb{E}(X_{ij} | Z_{iq} = 1, Z_{jl} = 1)$). Thus, by parameterising each block component separately, Bin-SBM can represent a host of different network topologies, including

various degrees of modular organisation and core-modular structures.

Block-specific connectivity rates are assumed to be constant across all subjects, suggesting that random variables $X_{ij} := \sum_{k=1}^{K} X_{ijk}$ follow a binomial distribution,

$$X_{ij}\big|Z_{iq}=1, Z_{jl}=1 \;\sim\; \text{Binomial}\big(K, \pi_{ql}\big), \tag{3}$$

where realisations across subjects can be compiled into a symmetric connectivity matrix, denoted by the random variable $\boldsymbol{X} = ((X_{ij}))_{1\le i\neq j\le n}$. Notably, this trivial model matches the case of binomially distributed edges in a single network in Mariadassou et al. (2010), where edges between node pairs represent the total number of observed edges across K subjects.

Estimation. Closely following Daudin et al. (2008) and Mariadassou et al. (2010), the framework of variational approximation is utilised to estimate the model parameters. The standard approach formulates the statistical model in terms of *complete data*, represented by the joint discrete density $f(\boldsymbol{x}, z; \boldsymbol{\pi}, \boldsymbol{\alpha})$ whose likelihood is given as

$$\log f(\boldsymbol{x}, z; \boldsymbol{\pi}, \boldsymbol{\alpha}) = \log f(\boldsymbol{x}|z; \boldsymbol{\pi}) + \log f(z; \boldsymbol{\alpha}). \tag{4}$$

The marginal of the complete data likelihood is noted as an *incomplete data* likelihood, $\log f(\boldsymbol{x}; \boldsymbol{\pi}, \boldsymbol{\alpha})$ and has the same parameters as the complete data likelihood. Ideally, the estimation of the model parameters should be based on $f(\boldsymbol{x}; \boldsymbol{\pi}, \boldsymbol{\alpha})$. However, as the explicit calculation of this density is computationally challenging, we use a variational approach. In the variational approach, the model parameters are estimated by optimising the variational bound $\mathcal{J}(f^*(z;\boldsymbol{\tau}); \boldsymbol{\alpha}, \boldsymbol{\pi})$, defined as

$$\mathcal{J}(f^*(z;\boldsymbol{\tau}); \boldsymbol{\alpha}, \boldsymbol{\pi}) = \mathbb{E}_{f^*}[\log f(\boldsymbol{x}, \boldsymbol{Z}; \boldsymbol{\pi}, \boldsymbol{\alpha})] - \mathbb{E}_{f^*}[\log f^*(\boldsymbol{Z}; \boldsymbol{\tau})], \tag{5}$$

where $f^*(z;\boldsymbol{\tau})$ depends on the variational parameter $\boldsymbol{\tau}$ ($n\times$ Q matrix of posterior probabilities), and it denotes the parametric family that is closest, in the Kullback-Leibler sense, to $f(z|\boldsymbol{x}; \boldsymbol{\pi}, \boldsymbol{\alpha})$. The complete derivation of this bound and its relationship to $f(\boldsymbol{x}; \boldsymbol{\pi}, \boldsymbol{\alpha})$ can be found in **SI** A. Based on Eq. (5), computation of the variational bound requires taking expectations of $\boldsymbol{Z}$ with respect to their variational density $f^*(z;\boldsymbol{\tau})$. To make this computationally feasible, $f^*(z;\boldsymbol{\tau})$ is taken to be a product of individual densities of $\boldsymbol{Z}_i$. Each density is categorical with block-specific probabilities, that are independent in each node

$$f^*(z;\boldsymbol{\tau}) = \prod_{i=1}^{n}\prod_{q=1}^{Q} \tau_{iq}^{z_{iq}}, \tag{6}$$

where $\sum_{q=1}\tau_{iq} = 1$, and $\mathbb{E}_{f^*}(Z_{iq}) = \tau_{iq}$, $\mathbb{E}_{f^*}(Z_{iq}Z_{jl}) = \tau_{iq}\tau_{jl}$. In particular, τ_{iq} is the strength of evidence that a node V_i is a member of block q having observed the data. For example, the node V_i in a three cluster structure has a vector $\widehat{\boldsymbol{\tau}}_i = (0.1, 0.01, 0.89)$. Based on the 'maximum a posteriori' (MAP) estimate shown in Eq. (11), this node has the strongest affiliation to the third block as its posterior probability is the highest (i.e. 0.89).

Taking the expectations stated in Eq. (5), the variational bound can more concisely be written as

$$\begin{aligned}\mathcal{J}(f^*(z;\boldsymbol{\tau}); \boldsymbol{\alpha}, \boldsymbol{\pi}) = &\frac{1}{2}\sum_{i=1}^{n}\sum_{j\neq i}^{n}\sum_{q,l}^{Q}\tau_{iq}\tau_{jl}\log f\big(x_{ij}|z_{iq}, z_{jl}; \pi_{ql}\big)\\ &+\sum_{i=1}^{n}\sum_{q=1}^{Q}\tau_{iq}\log\alpha_q - \sum_{i=1}^{n}\sum_{q=1}^{Q}\tau_{iq}\log\tau_{iq},\end{aligned} \tag{7}$$

where $f(x_{ij}|z_{iq}, z_{jl}; \pi_{ql}) = \binom{K}{x_{ij}}\pi_{ql}^{x_{ij}}(1-\pi_{ql})^{K-x_{ij}}$. Optimising Eq. (7) with respect to the variational parameter $\boldsymbol{\tau}$ and the model's parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\pi}$, according to the constraints $\sum_{q=1}^{Q}\tau_{iq} = 1$ and $\sum_{q=1}^{Q}\alpha_q = 1$, we find the following fixed point relations

$$\widehat{\alpha}_q = \frac{1}{n}\sum_{i=1}^{n}\widehat{\tau}_{iq}, \tag{8}$$

$$\widehat{\pi}_{ql} = \frac{\sum_{i=1}^{n}\sum_{j\neq i}^{n}\widehat{\tau}_{iq}\widehat{\tau}_{jl}x_{ij}}{K\sum_{i=1}^{n}\sum_{j\neq i}^{n}\widehat{\tau}_{iq}\widehat{\tau}_{jl}} \tag{9}$$

$$\widehat{\tau}_{iq} \propto \widehat{\alpha}_q \prod_{j\neq i}^{n}\prod_{l=1}^{Q}\big[f\big(x_{ij}|z_{iq}, z_{jl}; \widehat{\pi}_{ql}\big)\big]^{\widehat{\tau}_{jl}}. \tag{10}$$

An estimate of the probability that a randomly selected node is assigned to the q-th block ($\widehat{\alpha}_q$) in Eq. (8) is calculated as the average of posterior probabilities of all nodes in block q. An estimate of within- or between-block connection probability ($\widehat{\pi}_{ql}$) in Eq. (9) is calculated as the expected number of edges between block q and l, taking into account the posterior probability of nodes belonging to block q and l. The estimate of $\widehat{\tau}_{iq}$ (Eq. (10)) is proportional to (i) the estimate of the node belonging to the q-th block ($\widehat{\alpha}_q$) and (ii) the probability of observing connectivity of node V_i ($f(x_{ij}|z_{iq}, z_{jl}; \widehat{\pi}_{ql})$), taking into account posterior block assignment probabilities of other nodes ($\widehat{\tau}_{jl}$).

Finally, estimates of the classification vector $\widehat{z}_i$ are obtained from $\widehat{\boldsymbol{\tau}}_i$ such that

$$\widehat{z}_{iq} = \begin{cases} 1 & \text{if } \operatorname{argmax}_{q'}\big(\widehat{\tau}_{iq'}\big) = q \\ 0 & \text{otherwise.}\end{cases} \tag{11}$$

Model Selection. To estimate an optimal model with Q clusters, denoted by m_Q, we use the Integrated Classification Likelihood (ICL) criterion proposed by Biernacki et al. (1998) and its adaptation presented in Daudin et al. (2008), (see **SI** B for details). The notation m_Q refers to all the parameter estimates of Bin-SBM ($\widehat{\boldsymbol{\tau}}$, $\widehat{\boldsymbol{\pi}}$ and $\widehat{\boldsymbol{\alpha}}$) for Q clusters. The ICL criterion is defined on a complete data log-likelihood (see Eq. (4)) such that

$$\begin{aligned}\mathrm{ICL}(\mathrm{m}_Q) = &\log f(\boldsymbol{x}, \hat{z}|\mathrm{m}_Q, \widehat{\boldsymbol{\pi}}, \widehat{\boldsymbol{\alpha}}) - \frac{1}{2}\left(\frac{Q(Q+1)}{2}\right)\log\left[\frac{n(n-1)}{2}\right]\\ &-\frac{Q-1}{2}\log[n].\end{aligned} \tag{12}$$

Thus, ICL serves as a goodness of fit model statistic by mediating the trade-off between the fitted complete data likelihood (first term in Eq. (12)) and the complexity of the model (second and third terms in Eq. (12)). As Q increases, the second and the third terms in Eq. (12) progressively penalise the fitted complete data likelihood with increasing severity. For example, in a network with 100 nodes, a model with three clusters has a penalisation term of 30, but, if a model has 30 clusters, the penalisation term increases to 2045. Thus, selecting a model with 30 clusters over a model with three clusters can only be justified if the difference between their complete data likelihoods is significant enough to outweigh the heavy penalisation term.

Implementation Details. Beginning with the initial values for $\boldsymbol{\tau}^{(0)}$ (see **SI** C.2 for details), the model parameters are updated in two iterative steps:

1. $(\boldsymbol{\alpha}^{(h+1)}, \boldsymbol{\pi}^{(h+1)}) = \arg\max_{(\boldsymbol{\alpha},\boldsymbol{\pi})}[\mathcal{J}(f^*(z;\boldsymbol{\tau}^{(h)}); \boldsymbol{\alpha}, \boldsymbol{\pi})]$,
2. $\boldsymbol{\tau}^{(h+1)} = \arg\max_{(\boldsymbol{\tau})}[\mathcal{J}(f^*(z;\boldsymbol{\tau}); \boldsymbol{\alpha}^{(h+1)}, \boldsymbol{\pi}^{(h+1)})]$.

First, in the Maximisation Step or (M-Step), we use Eq. (8) and Eq. (9) to update $\boldsymbol{\alpha}$ and $\boldsymbol{\pi}$, respectively. Second, in the Expectation Step (or E-Step), we use Eq. (10) to update $\boldsymbol{\tau}$. Both steps are iterated until parameter estimates reach convergence. Technical details of the algorithm and a pseudocode can be found in **SI** C.2.

As shown in Daudin et al. (2008), this fitting procedure ensures that the algorithm iteratively climbs the variational bound (Eq. (7)), which in turn maximises the incomplete data log-likelihood $\log f(\boldsymbol{x}; \boldsymbol{\alpha}, \boldsymbol{\pi})$ without the need to calculate it directly. This guarantees that the algorithm will converge at a local maximum typically in the neighbourhood of the initial estimate $\boldsymbol{\tau}^{(0)}$. However, it is important to note that it does not guarantee

convergence at the global maximum. Identical limitations have been found in the classical EM-algorithm (Wu, 1983). Nevertheless, selecting the fit with the highest ICL score among many candidate fits could provide some practical evidence of attaining the global maximum. For a given network dataset, the likelihood surface can be heavily spiked, and, therefore, an informative starting point will yield a better likelihood fit than a non-informative one. The quality of parameter estimates depends on how informative $\boldsymbol{\tau}$ is. Therefore, instances in which many nodes can be equally assigned to all clusters will tend to produce poor estimates of $\boldsymbol{\pi}$, and consequently, a very small ICL score. However, selecting an informative starting point is often very difficult in practice (Scrucca and Raftery, 2015) as a large number of possible choices (Q^n) make it impossible to explore every possible starting point. In our application, hierarchical clustering (Murtagh and Legendre, 2014) was utilised to generate an initial estimate of $\boldsymbol{\tau}^0$ as it has been found to perform relatively well in practice and had been used successfully by Mariadassou et al. (2010). Nevertheless, other approaches may also be suitable. Depending on time constraints and dimensionality of the data (i.e. the number of nodes in the studied network), it is also possible to consider a random initialisation whereby the node labels are sampled from Q number of blocks. In this alternative, one would also need to consider multiple initialisations so that the sample space of all possible node classification is searched more effectively. Finally, it is worth mentioning that similar issues are found in many graphical models using variational algorithms (e.g., Mixed Membership SBM, Overlapping SBM, Latent Dirichlet Association model, Hidden Markov models and more). While selecting a local maximum solution may not be ideal, the algorithm can yield better results by improving upon the starting point obtained through other clustering algorithms (unless, of course, the starting point is already a local maximum). Hence, the procedure may be expected to provide an improvement upon the fit of any other clustering algorithms.

2.3. Homogeneous Stochastic Blockmodel (Hom-SBM)

In this model, variability between subjects is assumed to be a global feature of the multi-subject networks. Thus, conditional on its node assignments, each edge is assumed to follow a Bernoulli distribution with a probability of connection as a function of the subject covariates via a logistic regression model

$$\begin{aligned} &X_{ijk}\big|Z_{iq}=1,Z_{jl}=1 \;\sim\; \text{Bernoulli}\big(\pi_{qlk}\big),\\ &\log\left(\frac{\pi_{qlk}}{1-\pi_{qlk}}\right)=\theta_{ql}+\boldsymbol{d}_k^{\top}\boldsymbol{\beta} \quad \text{or} \quad \text{equivalently} \quad \pi_{qlk}=\frac{1}{1+e^{-\left(\theta_{ql}+\boldsymbol{d}_k^{\top}\boldsymbol{\beta}\right)}} \end{aligned} \tag{13}$$

where π_{qlk} is the k-th subject's connectivity rate in block (q,l), and θ_{ql} is the common intercept in block (q,l) across all subjects (i.e. baseline block probability on the logit scale). In particular, $\boldsymbol{d}_k^{\top}$ is the k-th subject's $1\times P$ dimensional vector of covariates and $\boldsymbol{\beta}$ is a $P\times 1$ vector of regression coefficients. This model has a block specific intercept for all subjects (θ_{ql}), and covariates that relate globally to the overall block structure but not individual blocks.

In contrast to Het-SBM, where each covariate was allowed to interact with the cluster structure in a block-wise manner (see Fig. 1 (c)), the covariates in Hom-SBM only engage with the entire cluster structure in a global fashion (see Fig. 1 (b)). Such modelling constraint suggests a smaller degree of variation in the cluster structure across subjects in Hom-SBM compared to Het-SBM. Nevertheless, Hom-SBM may be useful in neuroimaging applications (e.g. accounting for a global nuisance effect during preprocessing), and we have provided a detailed derivation of the estimation and inference procedures of in **SI** D. However, Het-SBM is preferable for testing varying covariate effects which may increase the number of connections between some brain areas, but decrease it between other brain areas. Given that such modelling principles may capture the observed plasticity in the brain caused, for example, by ageing or disease effects, we have chosen in this paper to primarily focus on Het-SBM.

2.4. Heterogeneous Stochastic Blockmodel (Het-SBM)

In this model, variability between subjects is assumed to influence within- and between-block connectivity. Thus, conditional on its node assignments, each edge is assumed to follow a Bernoulli distribution, whose rates of connection depend on the subject covariates via a logistic regression model. Formally,

$$\begin{aligned} &X_{ijk}\big|Z_{iq}=1,Z_{jl}=1 \;\sim\; \text{Bernoulli}\big(\pi_{qlk}\big)\\ &\log\left(\frac{\pi_{qlk}}{1-\pi_{qlk}}\right)=\boldsymbol{d}_k^{\top}\boldsymbol{\beta}_{ql} \quad \text{or} \quad \text{equivalently} \quad \pi_{qlk}=\frac{1}{1+e^{-\boldsymbol{d}_k^{\top}\boldsymbol{\beta}_{ql}}}, \end{aligned} \tag{14}$$

where π_{qlk} is the connectivity of block (q,l) associated with subject k, $\boldsymbol{d}_k^{\top}$ is the $1\times P$ vector of covariates associated with subject k (typically the first element will be 1, representing the intercept), and $\boldsymbol{\beta}_{ql}$ is a $P\times 1$ vector of regression parameters. The s-th component of $\boldsymbol{d}_k$ and $\boldsymbol{\beta}_{ql}$ are denoted as $\boldsymbol{d}_{ks}$ and $\boldsymbol{\beta}_{qls}$, respectively, while the total of $Q(Q+1)/2$ individual regression vectors $\boldsymbol{\beta}_{ql}$ are collectively denoted as $\boldsymbol{\beta}$.

Estimation. In this model, the variational optimisation strategy and Fisher's scoring algorithm are used to estimate the parameters $\boldsymbol{\tau},\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. When estimating $\boldsymbol{\beta}$, circumstances may arise where the inferences about $\boldsymbol{\beta}$ may be biased due to small or sparse samples (i.e. rare events[1]) or due to a perfect matching between the covariate and binary scores (i.e. complete separation[2]). To prevent such biases, we use a Firth type estimation (Firth, 1993; Kosmidis, 2014).

Similar to the estimation strategy used for Bin-SBM in Section 2.2, we begin by defining the lower bound. Since $\boldsymbol{\beta}_{ql}=\boldsymbol{\beta}_{lq}$, we use the notation γ_{ijql} to indicate the following cases

$$\gamma_{ijql}=\begin{cases}\tau_{iq}\tau_{jq} & \text{if} \quad q=l\\ \tau_{iq}\tau_{jl}+\tau_{il}\tau_{jq} & \text{if} \quad q<l.\end{cases} \tag{15}$$

Thus, the variational bound can be stated as

$$\begin{aligned} \mathscr{J}(f^*(z;\boldsymbol{\tau});\boldsymbol{\alpha},\boldsymbol{\beta})=&\frac{1}{2}\sum_{k=1}^{K}\sum_{i=1}^{n}\sum_{j\neq i}^{n}\sum_{q\leq l}^{Q}\gamma_{ijql}\log f\big(x_{ijk}\big|z_{iq},z_{jl};\boldsymbol{\beta}_{ql}\big)\\ &+\sum_{i=1}^{n}\sum_{q}^{Q}\tau_{iq}\log\alpha_q-\sum_{i=1}^{n}\sum_{q=1}^{Q}\tau_{iq}\log\tau_{iq}, \end{aligned} \tag{16}$$

where $f(x_{ijk}|z_{iq},z_{jl};\boldsymbol{\beta}_{ql})=\pi_{qlk}^{x_{ijk}}(1-\pi_{qlk})^{1-x_{ijk}}$ and π_{qlk} depends on $\boldsymbol{\beta}_{ql}$ (Eq. (14)). Optimising Eq. (16) for $\boldsymbol{\tau}$ and $\boldsymbol{\alpha}$ yields the following point estimating equations

$$\widehat{\alpha}_q=\frac{1}{n}\sum_{i=1}^{n}\widehat{\tau}_{iq}, \tag{17}$$

$$\widehat{\tau}_{iq}\propto\widehat{\alpha}_q\prod_{k=1}^{K}\prod_{j\neq i}^{n}\prod_{l=1}^{Q}\left[f\big(x_{ijk}\big|z_{iq},z_{jl};\widehat{\boldsymbol{\beta}}_{ql}\big)\right]^{\widehat{\tau}_{jl}}. \tag{18}$$

The Firth type estimates for $\boldsymbol{\beta}_{ql}$ are generated from optimising the modified variational bound $\mathscr{J}^*(f^*(z;\boldsymbol{\tau});\boldsymbol{\alpha},\boldsymbol{\beta})$ such that

[1] Rare events refer to an unusual distribution of 0s and 1s in a sample where there is enormous disparity between their counts.

[2] Complete separation is a special case in the logistic regression, where it is possible to establish a perfect Correspondence between all possible values of a covariate and $\{0,1\}$ outcomes in data. For example, complete separation has occurred when fitting a logistic regression curve to the block (q,l), if all smokers have an edge in this block while all non-smokers do not have an edge in this block.

$$\mathscr{J}^*(f^*(z;\boldsymbol{\tau});\boldsymbol{\alpha},\boldsymbol{\beta}) = \mathscr{J}(f^*(z;\boldsymbol{\tau});\boldsymbol{\alpha},\boldsymbol{\beta}) + \frac{1}{2}\sum_{q\le l}^{Q}\log\left[\mathrm{Det}\left(\mathscr{I}_{ql}(\boldsymbol{\beta}_{ql})\right)\right], \tag{19}$$

where $\mathscr{J}(f^*(z;\boldsymbol{\tau});\boldsymbol{\alpha},\boldsymbol{\beta})$ is given in Eq. (16) and the term $\frac{1}{2}\sum_{q\le l}^{Q}\log[\mathrm{Det}(\mathscr{I}_{ql}(\boldsymbol{\beta}_{ql}))]$ is a penalisation term. The first order partial derivatives of $\mathscr{J}(f^*(z;\boldsymbol{\tau});\boldsymbol{\alpha},\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}_{ql}$ can be written as $\mathbf{U}_{ql}(\boldsymbol{\beta}_{ql})$ such that

$$\mathbf{U}_{ql}(\boldsymbol{\beta}_{ql}) = \frac{1}{2}\sum_{k=1}^{K}\sum_{i=1}^{n}\sum_{j\neq i}^{n}\gamma_{ijql}\left(x_{ijk}-\pi_{qlk}\right)\boldsymbol{d}_k. \tag{20}$$

Similarly, negative second order partial derivatives with respect to $\boldsymbol{\beta}_{ql}$ yield a Fisher Information matrix $\mathscr{I}(\boldsymbol{\beta})$, expressed as a $Q(Q+1)/2P \times Q(Q+1)/2P$ block diagonal matrix of individual sub-matrices $\mathscr{I}_{ql}(\boldsymbol{\beta}_{ql})$. This can be defined as

$$\mathscr{I}_{ql}(\boldsymbol{\beta}_{ql}) = \frac{1}{2}\sum_{k=1}^{K}\sum_{i=1}^{n}\sum_{j\neq i}^{n}\gamma_{ijql}\pi_{qlk}\left(1-\pi_{qlk}\right)\boldsymbol{d}_k\boldsymbol{d}_k^{\top}. \tag{21}$$

To find estimates of $\boldsymbol{\beta}_{ql}$, the modified bound given in Eq. (19) is optimised according to Fisher's scoring formula

$$\boldsymbol{\beta}_{ql}^{(r)} = \boldsymbol{\beta}_{ql}^{(r-1)} + \mathscr{I}_{ql}^{-1}\left(\boldsymbol{\beta}_{ql}^{(r-1)}\right)\mathbf{U}_{ql}^{*}\left(\boldsymbol{\beta}_{ql}^{(r-1)}\right), \tag{22}$$

where (r) is the r-th iteration and $\mathbf{U}_{ql}^{*}(\boldsymbol{\beta}_{ql})$ is the modified score vector. The s-th element of $\mathbf{U}_{ql}^{*}(\boldsymbol{\beta}_{ql})$ is

$$\mathbf{U}_{qls}^{*}(\boldsymbol{\beta}_{ql}) = \mathbf{U}_{qls}(\boldsymbol{\beta}_{ql}) + \frac{1}{2}\mathrm{Tr}\left[\mathscr{I}_{ql}^{-1}(\boldsymbol{\beta}_{ql})\frac{\partial}{\partial\boldsymbol{\beta}_{qls}}\mathscr{I}_{ql}(\boldsymbol{\beta}_{ql})\right], \tag{23}$$

where

$$\frac{\partial\mathscr{I}_{ql}(\boldsymbol{\beta}_{ql})}{\partial\boldsymbol{\beta}_{qls}} = \frac{1}{2}\sum_{k=1}^{K}\sum_{i=1}^{n}\sum_{j\neq i}^{n}\gamma_{ijql}\pi_{qlk}\left(1-\pi_{qlk}\right)\left(1-2\pi_{qlk}\right)\boldsymbol{d}_k\boldsymbol{d}_k^{\top}\boldsymbol{d}_{ks}. \tag{24}$$

Model Selection. Similar to Bin-SBM, the ICL criterion for Het-SBM is given as

$$\mathrm{ICL}(\mathrm{m}_Q) = \log f(\boldsymbol{x},\widehat{\boldsymbol{z}}|\mathrm{m}_Q,\widehat{\boldsymbol{\alpha}},\widehat{\boldsymbol{\beta}}) - \frac{1}{2}\left(\frac{Q(Q+1)}{2}P\right)\log\left[\frac{n(n-1)}{2}K\right] - \frac{Q-1}{2}\log[n]. \tag{25}$$

Implementation Details. Beginning with some initial values for $\boldsymbol{\tau}^0$, the model parameters are iteratively updated according to a two-step procedure.

1. $(\boldsymbol{\alpha}^{(h+1)}\boldsymbol{\beta}^{(h+1)}) = \mathrm{argmax}_{(\boldsymbol{\alpha},\boldsymbol{\beta})}\left[\mathscr{J}(f^*(z;\boldsymbol{\tau}^{(h)});\boldsymbol{\alpha},\boldsymbol{\beta}) + \frac{1}{2}\sum_{q\le l}^{Q}\log[\mathrm{Det}(\mathscr{I}_{ql}(\boldsymbol{\beta}_{ql}))]\right]$,

2. $\boldsymbol{\tau}^{(h+1)} = \arg\max_{\boldsymbol{\tau}}[\mathscr{J}(f^*(z;\boldsymbol{\tau});\boldsymbol{\alpha}^{(h+1)},\boldsymbol{\beta}^{(h+1)})]$.

In the first step, Eqs. (17) and (22) are used to update $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, respectively. The part of the variational bound related to $\boldsymbol{\beta}$ is calculated by the sum of products between the non-negative weights $\boldsymbol{\gamma}$ and the conditional log-likelihoods of the logistic regression model (see the first term in Eq. (16)). As the conditional log-likelihood of the logistic regression is globally concave, and the individual components of $\boldsymbol{\gamma}$ are non-negative, then the sum of their product is also globally concave. This implies that there is at most one global maximum and, if the global maximum exists, the algorithm will converge on it. However, in extreme cases of complete or quasi separation, the maximum likelihood solution will not exist. In this situation, the algorithm will typically return high values for the regression coefficients but their variances would diverge to infinity, rendering inferences inaccurate. However, an optimisation with Firth penalisation typically ensures the existence of a global maximum, high quality estimates of regression coefficients and a numerically stable algorithm even in samples with rare events and extreme circumstances of separation. The latter property was confirmed by Heinze and Schemper (2002) who wrote the package logistf in R software which implements Firth regression for binary data. In the well-behaved datasets (i.e. without any of the above mentioned peculiarities), there is a perfect agreement between the maximum likelihood and Firth estimates. Fisher's scoring algorithm is initialised with zero as the starting values for the parameters ($\boldsymbol{\beta}$). For each iteration of Fisher's scoring algorithm, a maximal absolute change of value 5 is set for all the parameters (e.g., if the change for a parameter is estimated to be -7.2, it is forced to be -5). This ensures that changes in the parameters are not too large, which might cause the algorithm to overshoot the maximum. When the weighted log-likelihoods (i.e. block-wise log-likelihoods) are smaller than values obtained with previous parameter estimates, a step-halving procedure is executed. Parameter updates are reduced by half until an improvement is observed or until a maximal number of halving steps is reached, wherein the previous parameter estimates are retained. This strategy is consistent with the implementation in logistf.

In the second step, $\boldsymbol{\tau}$ is updated according to Eq. (18). These two steps are iterated until the convergence is obtained. The convergence is measured in terms of the relative changes of the parameter estimates. The pseudocode for these algorithms can be found in **SI** E and the computational time analysis for $n = 200, Q = 10$ and $K \in \{25, 50, 75, \ldots, 200\}$ can be found in **SI** F.

2.5. Inference in multi-subject Stochastic Blockmodels

Multi-subject models estimate a common cluster structure across subjects, serving as the common ground for making comparisons between subjects. Since Het-SBM imposes a logistic regression model on each element of the cluster structure, the methodological framework of logistic regression can be used to estimate differences between groups of subjects or the effects of covariates on the connectivity rate at each block. While these could simply be used as network summary metrics, the logistic regression model offers the additional possibility to determine if linear combinations of these quantities are statistically different from a specified constant (e.g., 0). In this context, several strategies can be used for parametric inference (i.e. the Wald and likelihood ratio tests), or non-parametric inference (i.e. permutation test) and each such test is assumed to be conditional on $\boldsymbol{\tau}$. Due to the joint optimisation in the variational algorithm of Het-SBM, cluster labels estimates depend on the covariates in the logistic regression model. Once the model with the most substantial clustering evidence is fitted, inferences are carried out without accounting for the variance of $\boldsymbol{\tau}$. This setup notably ensures the interpretability of results for non-parametric tests as it prevents scenarios where the permuted clusters do not overlap with the original cluster.

Wald test. The Wald test has commonly been used in logistic regression analysis to make inferences on the estimates of regression coefficients. In the context of Het-SBM, due to the special block diagonal structure of the Fisher Information matrix, the null hypothesis can be written as $\mathscr{H}_0 : \boldsymbol{L}_{ql}\boldsymbol{\beta}_{ql} = \boldsymbol{b}_{ql0}$. The Wald statistic takes the form

$$W_{ql} = \left(\boldsymbol{L}_{ql}\widehat{\boldsymbol{\beta}}_{ql} - \boldsymbol{b}_{ql0}\right)^{\top}\left(\boldsymbol{L}_{ql}\mathscr{I}_{ql}^{-1}(\widehat{\boldsymbol{\beta}}_{ql})\boldsymbol{L}_{ql}^{\top}\right)^{-1}\left(\boldsymbol{L}_{ql}\widehat{\boldsymbol{\beta}}_{ql} - \boldsymbol{b}_{ql0}\right) \Big/ c_{ql} \tag{26}$$

where $\boldsymbol{L}_{ql}$ is a matrix (or a vector) that defines the combination of the parameters (or contrast) tested and c_{ql} denotes the rank of $\boldsymbol{L}_{ql}$. Asymptotically, W_{ql} follows a $\chi^2_{c_{ql}}$ distribution. If $\boldsymbol{L}_{ql}$ is a vector, then

$$W_{ql}^{*} = \frac{\boldsymbol{L}_{ql}\widehat{\boldsymbol{\beta}}_{ql} - \boldsymbol{b}_{ql0}}{\sqrt{\boldsymbol{L}_{ql}\mathscr{I}_{ql}^{-1}(\widehat{\boldsymbol{\beta}}_{ql})\boldsymbol{L}_{ql}^{\top}}}, \tag{27}$$

asymptotically follows a Standard Normal distribution. In both Ordinary and Firth MLE approaches (Firth, 1993; Heinze and Schemper, 2002), standard errors of model parameters are estimated with the Fisher Information matrix (see Eq. (21)).

Likelihood ratio (LR) test. An alternative to the Wald test for the inference on the combination of parameters $\mathscr{H}_0 : \boldsymbol{L}_{ql}\boldsymbol{\beta}_{ql} = \boldsymbol{b}_{ql0}$ is the likelihood ratio (LR) test. Conceptually, the likelihood ratio test compares the full (or alternative) model against the restricted (or null) model. Thus, the likelihood ratio statistic is formulated as

$$\Lambda_{ql} = 2[\mathscr{J}^*(f^*(z;\widehat{\boldsymbol{\tau}});\widehat{\boldsymbol{\alpha}},\widehat{\boldsymbol{\beta}}) - \mathscr{J}^*(f^*(z;\widehat{\boldsymbol{\tau}});\widehat{\boldsymbol{\alpha}},\tilde{\boldsymbol{\beta}})], \tag{28}$$

where $\tilde{\boldsymbol{\beta}}$ is the set of all parameters estimated under the null hypothesis. More precisely, $\tilde{\boldsymbol{\beta}}$ is estimated from the modified variational bound of the restricted model, which has a different penalisation term from the one of the full model. While $\tilde{\boldsymbol{\beta}}$ is obtained using the modified variational bound of the null model, it is substituted into the modified variational bound of the full model to finally get the likelihood ratio statistic.

Similarly to the Wald test, under the null hypothesis, Λ_{ql} is assumed to follow a $\chi^2_{c_{ql}}$ distribution, where c_{ql} is the rank of $\boldsymbol{L}_{ql}$. Note that for Ordinary MLEs, the likelihood ratio is based on the non-modified variational bound given by Eq. (16).

Multiple testing. Inference procedures for the block level parameters effectively make $Q(Q+1)/2$ individual tests, and this presents multiple comparison problem. To control the family-wise error rate (FWE), defined as the probability of making at least one Type I error, the Bonferroni correction (Holm, 1979) is utilised. This correction is valid for any dependence structure and is easy to apply: Instead of using a nominal α_0 significance value (e.g., 0.05), α_0/n_T is used instead, where n_T is the number of tests (here, $n_T = Q(Q+1)/2$).

Permutation test. In addition to the Wald and likelihood ratio (LR) tests, that depend on asymptotic sampling distributions, permutation tests are also considered (Good, 2000). Permutation tests are based on the premise that, under the null hypothesis, the data can be exchanged without altering its distribution. The exchangeability implies that the distribution of any test statistic can be found empirically through repeated evaluations of randomly rearranged (or permuted) data. In the context of Het-SBM, permutation tests are used to make inferences on the parameter vector $\boldsymbol{\beta}_{ql}$ under the null hypothesis, in which there is no association between edge occurrence and the covariate tested. In our simulations, only tests of the entire parameter vector $\boldsymbol{\beta}_{ql}$ are considered. However, it is essential to note that such an approach would not be possible in real fMRI data as the model depends on more than one covariate. To account for this, we use a simple permutation strategy proposed by Potter (2005). In this approach, the covariate of interest is regressed on the remaining nuisance covariates (using a linear regression model), and the residuals from this model are used in place of the original covariate. Following the permutation of these residuals, logistic regressions are fitted and *P*-values are computed by comparing the resulting test statistics against their original scores.

The *P*-value of the observed Wald test statistic w_{ql0} is computed by a Monte Carlo sampling scheme such that for a sequence of random permutations indexed by t ($t = 1,\ldots,M$), there is M Wald statistics labelled as $w_{ql1}\ldots, w_{qlM}$. Including the observed Wald statistic w_{ql0} in the permuted scores, the Monte Carlo *P*-value is computed as

$$\mathrm{P}\left(W_{ql} \geq w_{ql0} \middle| H_0\right) = \frac{\sum_{t=0}^{M} I\left(w_{qlt} \geq w_{ql0}\right)}{M+1}, \tag{29}$$

where $I(\cdot)$ is the indicator function. The same approach can be used with the likelihood ratio test to obtain *P*-values without assuming an asymptotic χ^2 distribution.

Improved Multiple Testing Procedures with Permutation. The Bonferroni method for controlling the FWE is conservative in the presence of dependence. Thus, there is a need to consider an alternative permutation based procedure (Westfall and Young, 1993). For the Wald statistic w_{ql0}, the FWE corrected *P*-value is given as

$$\mathrm{P}\left(\max\left[\left(\left(W_{ql}\right)\right)_{1\leq q,l\leq Q}\right] \geq w_{ql0} \middle| \mathscr{H}_0\right) = \frac{\sum_{t=0}^{M} I\left(\max\left[\left(\left(w_{qlt}\right)\right)_{1\leq q,l\leq Q}\right] \geq w_{ql0}\right)}{M+1}. \tag{30}$$

The FWE corrected *P*-values for the likelihood ratio (LR) statistics are computed in the same way.

3. Methodology of simulations and results

This section details the methodology and results of three sets of simulations labelled Simulation I-III. To make the cross-referencing between the initial simulation parameters and the results more convenient for the reader, this section first details the goals and setups of each simulation, and then it gives their corresponding results. The birds-eye view of each simulation pipeline is given in Fig. 2.

3.1. Simulation I goals and setup

Goals of Simulation I. The goal of Simulation I is to assess the overall performance of Bin-SBM and Het-SBM in a range of different cluster structures which exhibit substantial deviations from a classically modular organisation[3]. The between-subject variations in the cluster structures are assumed to be independent of subject covariates. Since this setup does not consider covariate effects, it allows for the comparison between multi-subject SBMs and modular decomposition methods such as the Fast Louvain (FL) and Newman Spectral (NS) algorithms. The modular algorithms are utilised in terms of their stand-alone implementations as well as in combination with consensus clustering.

Setup of Simulation I. In this simulation, the total number of blocks, subjects and nodes has been fixed to 10, 30 and 200, respectively (i.e. $Q = 10$, $K = 30$ and $n = 200$). The range of block sizes was determined according to three designs, which we will refer to as Balanced, Mildly Unbalanced (M. Unbalanced) and Unbalanced designs. Each design is characterised by varying degrees of heterogeneity of block sizes, detailed in Table 1. Given the network sizes, a Q of 10 was selected in order to yield similar block sizes to those already reported in applied literature (Picard et al., 2009). While one could also vary n and Q, and perhaps even study entirely different experimental questions like small sample behaviours (Pavlovic, 2015), our prior experience with different n and Q have been found to yield comparable outcomes to what was obtained.

In this simulation, we have chosen three types of clusters structures for $\boldsymbol{\pi}$. As noted in Fig. 3 these are *Homogeneous-Modular* (Hom-Modular), *Heterogeneous-Modular* (Het-Modular) and *Core-Modular*. The first two are classic representations of modular network structure while the Core-Modular structure is a combination between the heterogeneous modular structure (Blocks 1–8) and a densely integrated core (Blocks 9 & 10). The Core-Modular structure is closely related to the 'core-on-modules' structure found in *C. elegans* connectome using SBM (Pavlovic et al., 2014), which deviates from a purely modular organisation due to the inclusion of densely inter- and intra-connected core blocks.

In order to simulate a network, we generate Bernoulli realisations according to the connectivity rates in Fig. 3 and the three block designs in Table 1. For example, to generate a network with a Hom-Modular structure where Block 1 has 20 nodes and a within-block connectivity rate of 0.9, we perform $20 \times 19/2$ (maximum number of edges in Block 1) Bernoulli trials where the probability of seeing an edge (i.e. outcome 1) is 0.9. Likewise, to generate connections between 20 nodes in Block 1

[3] The clusters in a modular organisation are characterised by maximised within-cluster connectivity and minimised between-cluster connectivity. The probabilities of connections within each cluster are assumed to be equal, and the probabilities of connections between clusters are also assumed to be equal.

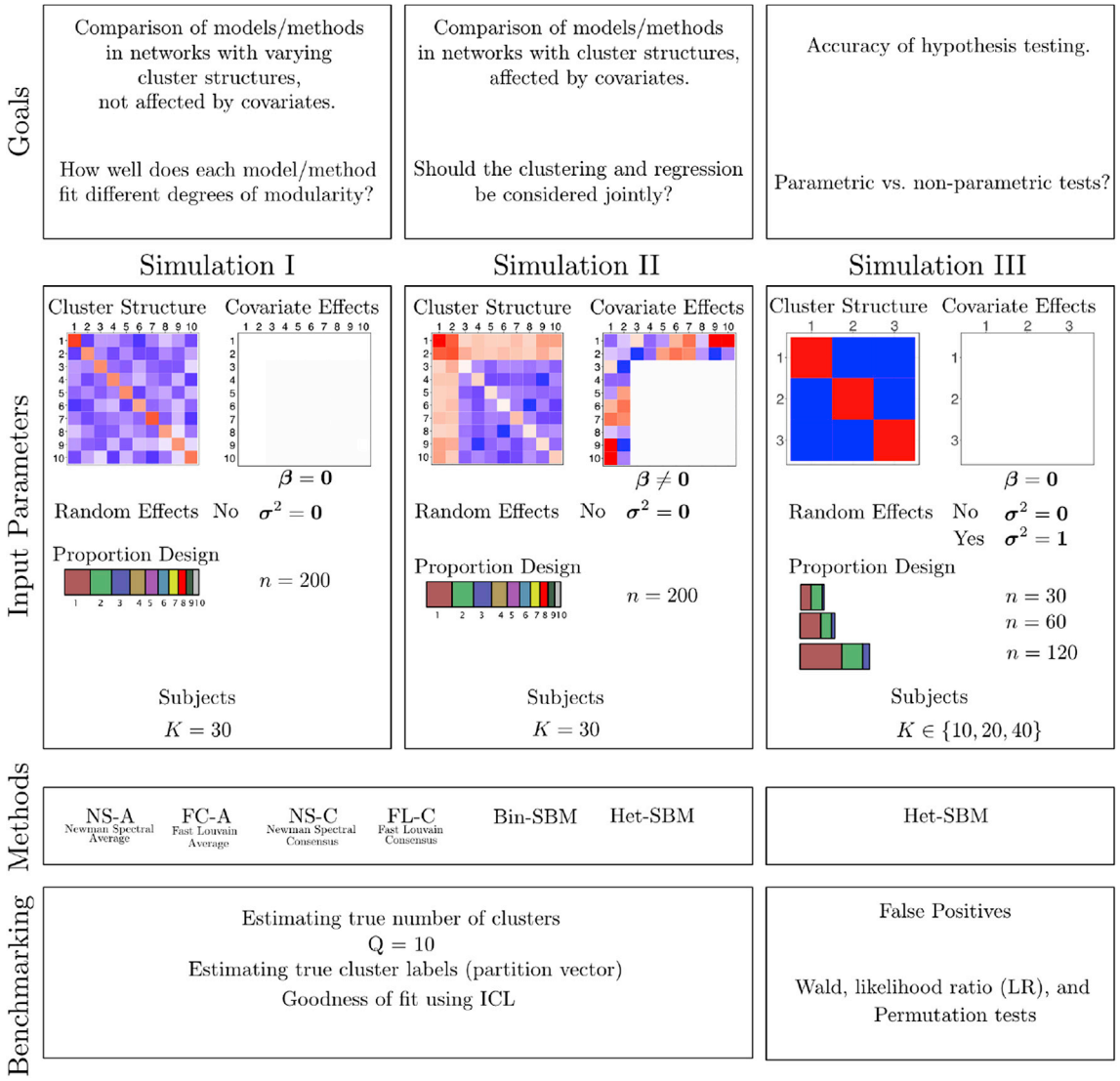

Fig. 2. Pipeline of each simulation organised in terms of its goals, input parameters, fitting methods and benchmarking strategies. The input parameters consist of a cluster structure (capturing the underlying network topology), covariate effects (how a cluster structure varies across subjects as a function of their demographics), random effects (introducing dependencies within blocks & within-subjects), a network size (i.e. the total number of nodes *n*), a proportion design (i.e. the individual cluster sizes) and the total number of subjects (*K*). The input parameters show only one instance of each parameter and its full range of values can be found in each simulation section (see Sections 3.1, 3.3 and 3.5 for further information).

Table 1
Block sizes for three block size designs. In the Balanced design, all blocks have the same size while, in the Mildly Unbalanced (M. Unbalanced) and Unbalanced designs, there is a systematic decrease in the overall number of nodes such that the last blocks are the smallest. In the M. Unbalanced design, the block sizes are mildly varying while, in the Unbalanced case, the block sizes are changing more rapidly.

	Block Labels									
Design Label	1	2	3	4	5	6	7	8	9	10
Balanced	20	20	20	20	20	20	20	20	20	20
M. Unbalanced	38	32	27	23	19	16	14	12	10	9
Unbalanced	60	41	29	20	15	11	8	6	5	5

and 20 nodes in Block 2, we perform 20×20 (maximum number of edges between Blocks 1 & 2) Bernoulli trials using the probability specified by the connection rate between these two blocks (e.g., 0.05). In this block-wise manner, we systematically built an entire network, and we generate 100 individual networks for each of these combinations of parameters (i.e. block design and cluster structure). It is important to note that, firstly, the block assignment of each node is already known and, therefore, the ground truth can be used to benchmark the ability of a model to estimate true cluster assignments correctly. Secondly, our data generating process does not take into account the between-subject variation.

Each of the simulated multi-subject networks is fitted with four approaches: Bin-SBM, Het-SBM, the Newman Spectral (NS) and Fast Louvain (FL) algorithms. This setup allows us to study the ability of each approach to retrieve the correct clustering. Specific details about each fitting procedure are given below.

Fitting procedure for MS-SBMs. For a given network and each value of $\widehat{Q} \in \{2, 3, \ldots, 18\}$ blocks, Bin-SBM and Het-SBM are fitted 30 times with different initialisations obtained from k-means, random label sampling and hierarchical clustering (hclust). For k-means, 10 initialisations are generated using the function Kmeans (Lucas, 2014, R package amap) on the average adjacency matrix, which allowed the restarts to vary due to the stochastic nature of k-means. For random label sampling, 10 starting points are generated through a uniform sampling of labels between one and $\widehat{Q}$. For hclust, which is deterministic, 10 initialisations are generated

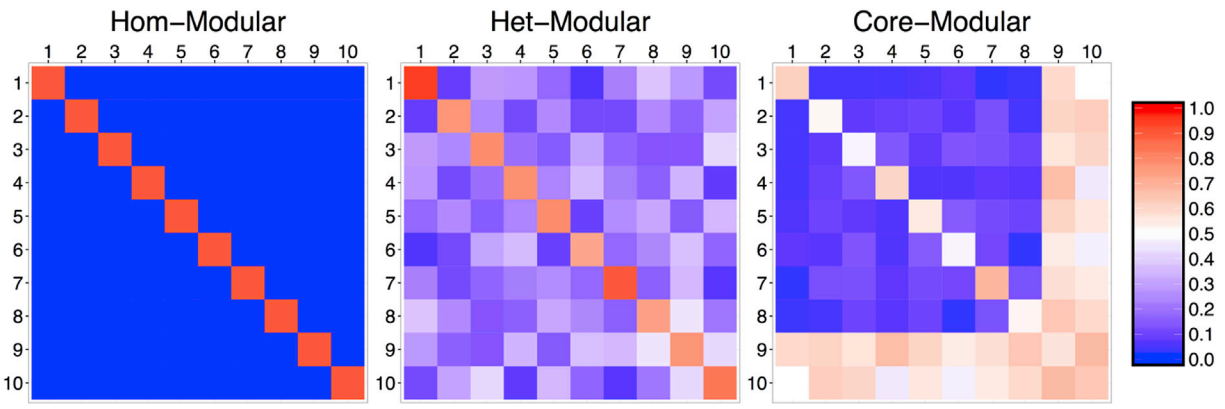

Fig. 3. Connectivity matrices (π) for three different cluster structures. Each cluster structure is characterised by differing degrees of modular organisation, defined as a relatively higher within-blocks connectivity rates compared to the between-blocks connectivity rates. The homogeneous modular (Hom-Modular) case represents perfect, classical modularity characterised by a high within-block connectivity rate that is constant for all blocks and a low between-block connectivity rate that is constant across all block pairs. The heterogeneous modular (Het-Modular) case retains the modular structure, but it allows for varying rates of connectivity within and between different blocks. The Core-Modular structure is a hybrid organisation characterised by (i) a heterogeneous modular structure (Blocks 1–8) and (ii) two core blocks with dense within- and between-block connections (Blocks 9 & 10).

by applying the hclust function (R Core Team, 2017, R package stats) on the average adjacency matrix, and nine randomly selected subject-specific adjacency matrices. The best solution is selected as the one with the highest ICL score both with respect to different initialisations and values of Q.

Fitting of Newman Spectral (NS) and Fast Louvain (FL) algorithms. We fit the modular algorithms in two ways. Firstly, we apply them on the average adjacency matrices, the result of which we refer to as NS-A and FL-A. Secondly, we apply them on the individual subject matrices using the consensus clustering algorithm, the results of which we refer to as NS-C and FL-C. In both cases, the first two letters indicate the algorithm in question and the last letter stands for either the averaged data (-A) or consensus (-C) clustering. Each algorithm utilises 11 different values for their resolution parameter γ ranging from 0.5 to 1.5 with a step size of 0.1 and, whenever possible, uses the same initialisations as the MS-SBMs. The functions for the NS and FL algorithms can be found respectively as modularity_und and community_louvain in the Matlab Brain Connectivity Toolbox (Rubinov and Sporns, 2010, http://www.brain-connectivity-toolbox.net/, last accessed May 01, 2018). A complete description of implementation details can be found in **SI** G.

Benchmarking Methodology. In this simulation, we evaluate each approach by the ability to (i) estimate the optimal number of clusters (i.e. $Q = 10$) and (ii) achieve the true clustering. The latter is assessed by the Adjusted Rand Index (ARI) (Handl et al., 2005; Hubert and Arabie, 1985, see **SI** H, for details), which indicates the overall agreement between the selected best fits and the true clustering, and the ICL scores evaluated given the estimated clustering vector via Eq. (12).

3.2. Simulation I results

Table 2 reports, for Het-SBM in the Core-Modular scenarios, the ranges over the 100 Monte Carlo samples of the best adjusted ICL scores, each defined as the difference between the maximum ICL score for a given $\widehat{Q}$ and the maximum ICL score over all $\widehat{Q}$. The range $[0, 0]$ indicates that, for all 100 Monte Carlo samples, the best ICL score systematically points towards the correct number of clusters (i.e. $\widehat{Q} = 10$). The ICL scores tend to decrease when $\widehat{Q}$ is taking values which are going further away from the ground truth ($\widehat{Q} = 10$) with a higher rate for values below 10. The same behaviour has been observed for all the other scenarios as well as for Bin-SBM (see **SI** I; Tables S2–S3 for Het-SBM and S4–S6 for Bin-SBM), indicating that both models are reliable in retrieving the correct number of clusters.

Fig. 4 shows the distribution of the estimated $\widehat{Q}$ for all approaches.

Table 2
Ranges of adjusted ICL scores for Het-SBM in the scenarios with Core-Modular cluster structure. For each simulated dataset and a given $\widehat{Q}$, the adjusted ICL scores are given as the difference between the maximum score across the restarts (local maximum) and all values of $\widehat{Q}$ (global maximum). For each $\widehat{Q}$, the range of ICL scores is shown across all 100 Monte Carlo samples. The range $[0, 0]$ indicates that, for all 100 Monte Carlo samples, the best ICL score systematically points towards the correct number of clusters (i.e. $\widehat{Q} = 10$). The ICL scores tend to decrease when $\widehat{Q}$ utilises values that increasingly diverge from the ground truth $\widehat{Q} = 10$ with a higher rate for values below 10.

$\widehat{Q}$	Core-Modular		
	Balanced	M. Unbalanced	Unbalanced
2	[−31252, −30193]	[−44884, −43403]	[−51276, −49308]
3	[−23621, −22803]	[−24265, −23200]	[−22118, −21019]
4	[−18689, −17173]	[−16489, −15498]	[−10824, −10095]
5	[−14118, −12602]	[−10843, −10036]	[−6742, −5866]
6	[−9936, −8991]	[−6912, −6318]	[−3058, −2704]
7	[−6296, −5649]	[−4536, −3566]	[−1827, −1547]
8	[−3365, −2986]	[−2331, −1992]	[−843, −675]
9	[−974, −763]	[−520, −335]	[−628, −140]
10	[0, 0]	[0, 0]	[0, 0]
11	[−76, −63]	[−77, −63]	[−77, −62]
12	[−164, −137]	[−156, −134]	[−157, −139]
13	[−243, −215]	[−238, −215]	[−244, −221]
14	[−327, −299]	[−328, −300]	[−334, −315]
15	[−429, −389]	[−427, −395]	[−432, −407]
16	[−521, −479]	[−531, −496]	[−540, −513]
17	[−633, −594]	[−641, −602]	[−656, −607]
18	[−739, −703]	[−758, −704]	[−770, −728]

Each modular fit is given in terms of three γ values (i.e. $\gamma \in \{0.5, 1, 1.5\}$). The full set of results can be found in **SI** I; see Figs. S1, S3 and S5. Consistent with the results mentioned above, Bin-SBM and Het-SBM accurately estimate the optimal number of clusters in all simulation scenarios. In contrast, the modular algorithms show a variation in their estimates and tend to strongly depend on the block designs and cluster structures. In particular, both algorithms are accurate in the Hom-Modular case with the Balanced and M. Unbalanced designs, where there is minimal influence of the tuning parameter (γ) and minimal difference between average and consensus clustering. However, in the case of Hom-Modular with Unbalanced design, there is a strong influence of γ in average clustering but not in consensus clustering. This result suggests that consensus clustering is robust in cases with small clusters and perfectly modular networks. Unlike Hom-Modular, all Het-Modular and

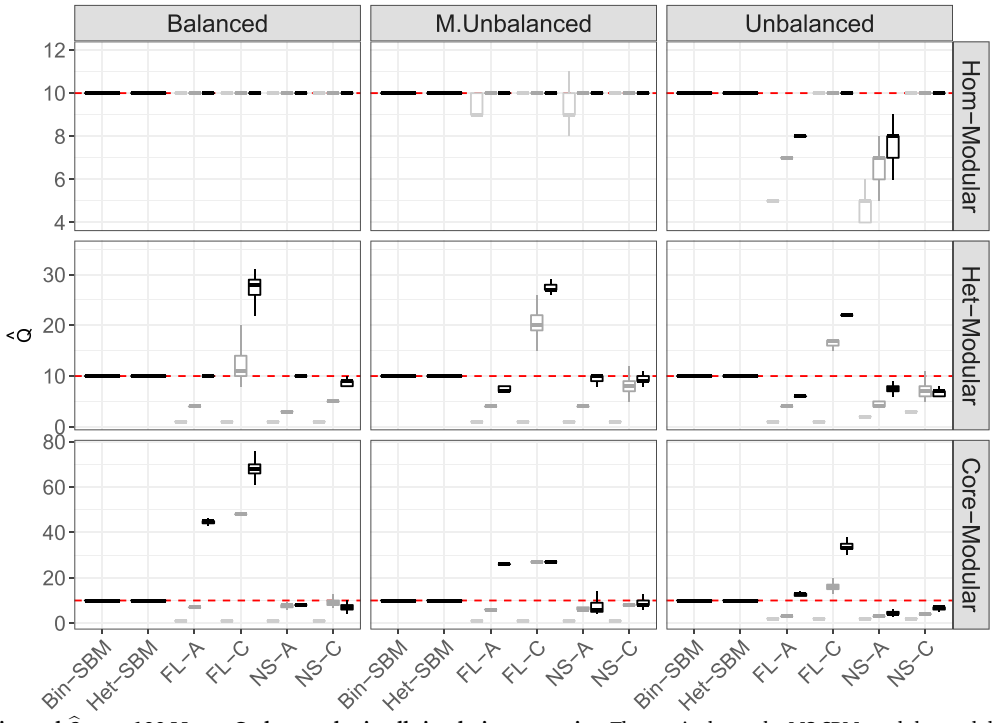

Fig. 4. Boxplots of estimated $\widehat{Q}$ over 100 Monte Carlo samples in all simulation scenarios. The *x*-axis shows the MS-SBMs and the modular algorithms such that:

Core-Modular cases exhibit significant dependence on γ. For the Fast Louvain algorithm, there is a tendency to underestimate Q for small values of γ and to overestimate Q for large values of γ. In contrast, the Newman Spectral algorithm tends to underestimate Q for all the range of γ values investigated. It is interesting to note that the values of γ that yield the most accurate results are strongly varied across block designs and cluster structures, indicating a potential difficulty in setting its value in real data applications. Another noteworthy point is that, although the Fast Louvain algorithm had the same initialisations as the MS-SBMs, it obtained very different estimates of the optimal number of clusters.

In Fig. 5, the boxplots represent the Adjusted Rand Index (ARI) scores (ARI $\in [0, 1]$; see **SI** H) over 100 Monte Carlo samples, with 1 denoting a perfect agreement and 0 a complete disagreement. For space considerations, only three different values of the tuning parameter (i.e. $\gamma \in \{0.5, 1, 1.5\}$) are shown for each of the modular fits (see **SI** I Fig. S2, S4 and S6 for the complete results). Bin-SBM and Het-SBM exhibit high accuracy in all simulation scenarios. In contrast, the modular algorithms show difficulties in retrieving the actual cluster labels, especially in the Het-Modular and Core-Modular cases. In particular, the ARI scores appear to be strongly influenced by the tuning parameter γ with the tendency to

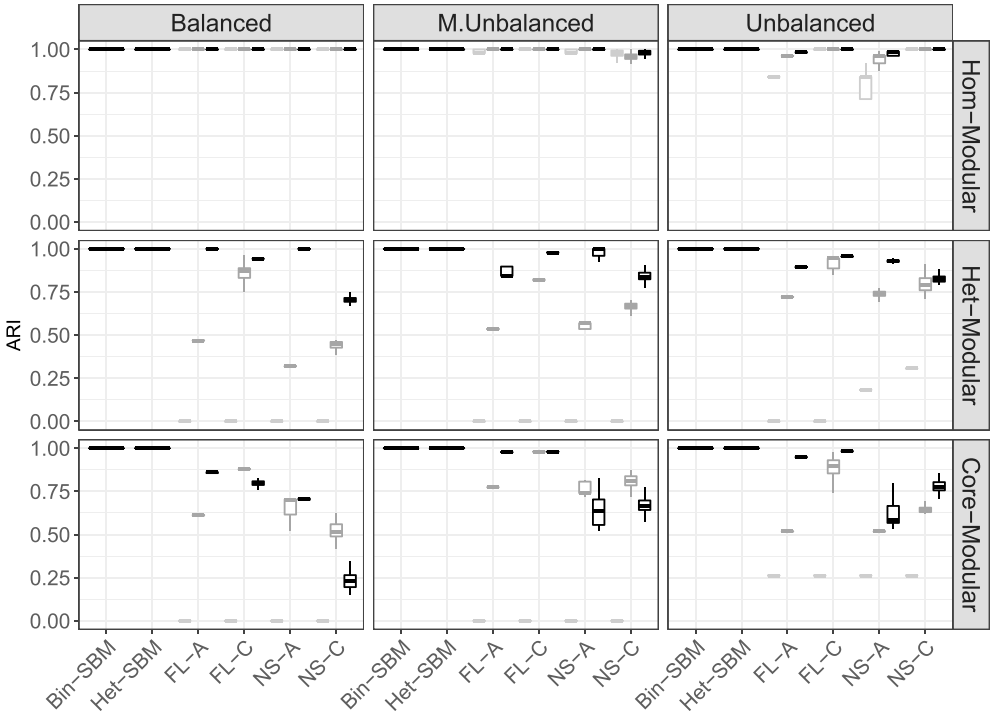

Fig. 5. Boxplots showing the recovery of the true cluster assignments over 100 Monte Carlo samples in terms of ARI scores. FL and NS stand for Fast Louvain

be at their highest values when γ is large. Surprisingly, some modular fits with very high ARI scores (close to 1) tend to overestimate the total number of clusters strongly. The most notable example of this can be seen for FL-A and FL-C ($\gamma = 1.5$) in the Core-Modular and M. Unbalanced design (see Figs. 4 and 5). This unexpected result can be explained by the fact that some portions of the ground-truth cluster labels are correctly retrieved while the few remaining ground-truth clusters are entirely split into single-node clusters in the candidate modular solutions. The contributions to the ARI scores are therefore driven by the sizes of the true clusters and the way these are split into single-node clusters in the candidate solutions. For example, if the last two clusters in the true clustering are very tiny and the modular solution splits them into single-node clusters, the ARI scores will still stay very high as the larger clusters are correctly matched (similar issues are discussed in Gates et al., 2019). A potential remedy to this misclassification bias is to consider the ICL criterion (see Eq. (25)) given the fitted partition ($\widehat{z}$). In this simulation, each fitted partition was assigned an ICL score, and the results across 100 Monte Carlo datasets have been summarised in Table 3. To make a comparison across all approaches, we consider a range of adjusted ICL scores for the Core-Modular scenarios which were computed as the difference between (i) the ICL scores of each approach, and (ii) the highest ICL score across all approaches. The fits of Bin-SBM and Het-SBM, as shown in Fig. 5, perfectly recover the true partitions in all designs, yielding a $[0, 0]$ range. However, the modular algorithms were not found to yield optimal partitions. For example, despite ARI values being close to 1 in the M. Unbalanced case (see Fig. 5), the fits of FL-A and FL-C for $\gamma = 1.5$ resulted in ranges of $[-6454, -5275]$, and $[-2822, -2040]$, respectively. Such extreme range values suggest very poor fits as the candidate solutions are not parsimonious. Similar results were obtained in the Het-Modular cases while, for the Hom-Modular cases, the fits were reasonably accurate (see **SI** I Figs. S7 and S8).

3.3. Simulation II goals and setup

Goals of Simulation II. The goal of Simulation II is to explore situations in which covariate effects pose a strong influence on the estimation of cluster labels. Although our data has been simulated according to fixed cluster labels for all subjects, it is still possible to encounter situations in which the joint optimisation of cluster labels and logistic regression covariates strongly informs the final clustering fit. In this simulation, the fit based on Bin-SBM portrays an example in which covariates are discounted from the estimation of cluster labels. For example, this would be a case in which a researcher initially fits cluster labels independently of the logistic regression model, and then subsequently uses the clusters to fit a logistic regression model. At first glance, this may appear to be a viable approximation that reduces the computational burden from logistic regression optimisation. However, there are situations in which a collection of nodes with somewhat similar connectivities show evidence supporting further clustering when fitted with informative covariates.

Setup of Simulation II. In this simulation, the parameters related to the total number of blocks, subjects and nodes are set to the same values as in Simulation I (i.e. Q = 10, K = 30 and n = 200). Likewise, the block sizes vary according to the Balanced, M. Unbalanced and Unbalanced designs (see Table 1). The subjects are divided into two groups of 15 subjects. The subjects in the first group have individual $\boldsymbol{d}_k^\top$ set to $(+1, +1)$, while the remaining subjects have individual $\boldsymbol{d}_k^\top$ set to $(+1, -1)$. The first entry corresponds to a common intercept representing the average effect of both groups, while the second entry corresponds to a differential effect between the first and second groups. Panels (A) and (B) in Fig. 6 show the block-specific common intercept and group effect coefficients, respectively. All 30 subjects share a common Core-Modular cluster structure with a densely integrated core comprising Blocks 1 & 2, and a heterogeneous set of modules comprising Blocks 3–10. The group effects largely interact with this cluster structure via the core blocks, and this is accomplished by setting the regression coefficients of Blocks 1 & 2 to a range of values on a scale between +2.50 and − 2.50. The group effects on the modular structure are almost negligible (0.05). Therefore, the differences between the two groups will only be apparent in the core blocks and not in the modular components of the cluster structure. This scenario can be expected in the clinical setting in which neurological disorders preferentially impact high degree hubs that constitute the core or rich club of the cluster structure (Crossley et al., 2014).

Fitting and Benchmarking. As noted earlier in Eq. (14), the intercept and group coefficients that are given in panels (A) and (B) of Fig. 6 are utilised to simulate 100 Monte Carlo realisations of every subject connectivity matrix. For a given multi-subject network realisation and each value of candidate block numbers, $\widehat{Q} \in \{2, 3, \ldots, 18\}$, Bin-SBM and Het-SBM are both fitted 10 times using some initialisations based on hierarchical clustering (hclust). More precisely, the initialisations are generated by applying the hclust function (R Core Team, 2017, R package stats) once on the average adjacency matrix and nine times on a randomly selected subject-specific adjacency matrix. The best solution is selected as the one with the highest ICL score across all the different initialisations and values of $\widehat{Q}$. The modular algorithms utilise the same initialisations as for Bin-SBM and Het-SBM and are fitted according to the procedures described in Simulation I (see Section 3.1). The ARI and ICL scores are used to evaluate the accuracy of the estimated cluster structure.

Table 3
Ranges of adjusted ICL score of the best fits for several approaches in the scenarios with Core-Modular cluster structure. For each simulated dataset and approach, the adjusted ICL score of the best fit is the difference between its ICL score and the maximum ICL score across all approaches. For each approach, the range of ICL scores is shown across all 100 Monte Carlo samples. In brief, Bin-SBM and Het-SBM systematically yielded the highest ICL scores.

	Core-Modular		
Approach	Balanced	M. Unbalanced	Unbalanced
Bin-SBM	[0, 0]	[0, 0]	[0, 0]
Het-SBM	[0, 0]	[0, 0]	[0, 0]
FL-A (0.5)	[−94759, −93042]	[−86412, −84361]	[−51276, −49308]
FL-A (1)	[−80967, −78202]	[−59720, −55804]	[−42216, −40272]
FL-A (1.5)	[−21556, −10203]	[−6454, −5275]	[−17463, −8801]
FL-C (0.5)	[−94759, −93042]	[−86412, −84361]	[−51276, −49308]
FL-C (1)	[−8391, −7072]	[−3675, −2040]	[−12814, −2025]
FL-C (1.5)	[−18591, −9751]	[−2822, −2040]	[−4648, −1912]
NS-A (0.5)	[−94759, −93042]	[−86412, −84361]	[−51276, −49308]
NS-A (1)	[−86307, −76052]	[−61199, −55297]	[−42216, −40272]
NS-A (1.5)	[−82353, −67263]	[−65910, −42283]	[−44189, −19932]
NS-C (0.5)	[−94759, −93042]	[−86412, −84361]	[−51276, −49308]
NS-C (1)	[−85879, −68256]	[−62754, −48389]	[−44352, −32619]
NS-C (1.5)	[−92240, −80997]	[−74128, −56713]	[−51527, −37443]

3.4. Simulation II results

Tables 4 and 5 report for Bin-SBM and Het-SBM the ranges over the 100 Monte Carlo samples of the best adjusted ICL scores. Each such score is defined as the difference between (i) the maximum ICL score for a given $\widehat{Q}$ and (ii) the maximum ICL score over all $\widehat{Q}$. For a particular number of clusters, the range $[0, 0]$ indicates that this total number of clusters is consistently selected in all 100 Monte Carlo samples. In Table 4, Bin-SBM systematically underestimates the actual number of clusters across all three block designs, suggesting 9 clusters instead of 10. In contrast to this, the results in Table 5 show that Het-SBM accurately estimates the correct number of clusters across the Balanced and M. Unbalanced designs in all the simulations. In the Unbalanced design, however, Het-SBM estimates the actual number of clusters in 90% of the simulations but overestimates it by one cluster in the remaining 10% of simulations. This behaviour may be explained by the fact that, in some simulations, the algorithm may not have converged to the global maximum as we have used a minimal number of initialisations of which all were based on hierarchical clustering. A potential remedy to this

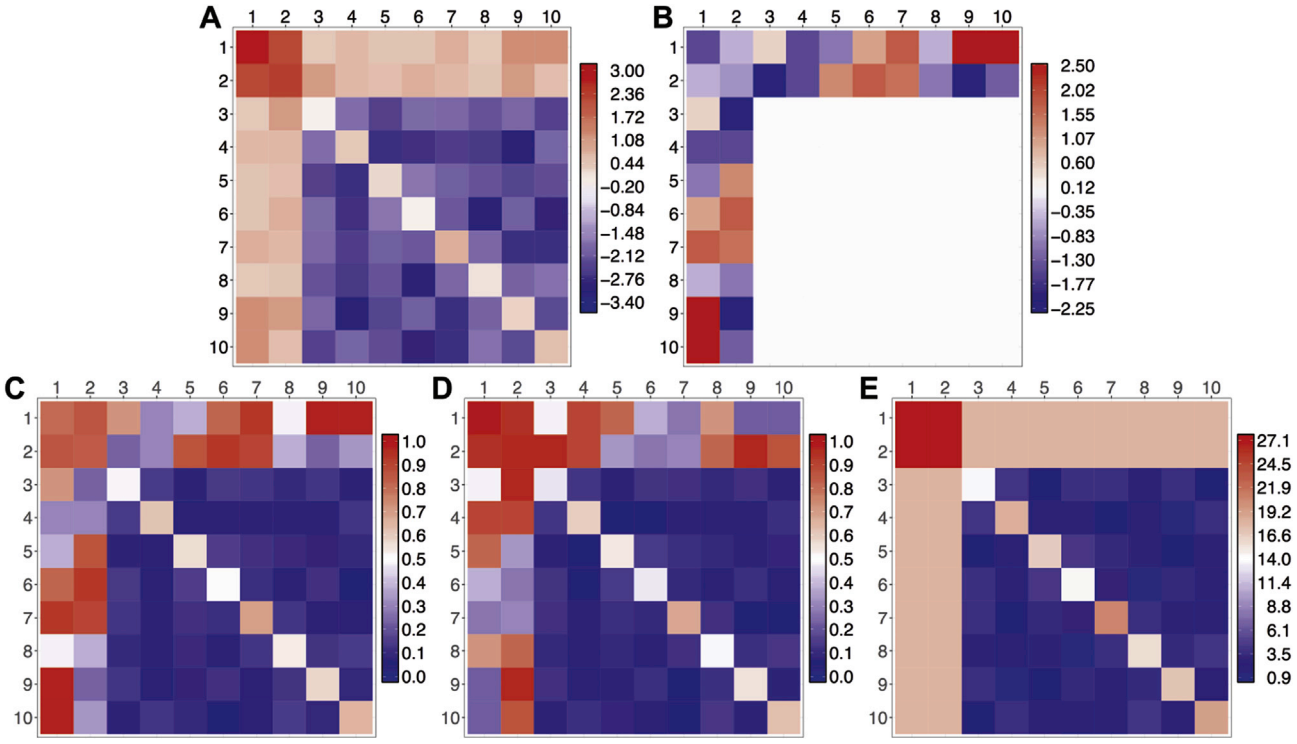

Fig. 6. Simulation II setup information. The block-specific intercept effects given as a $Q \times Q$ matrix of values $((\beta_{ql1}))_{q,l\in\{1,\ldots,Q\}}$ in **A** create a common Core-Modular cluster structure for all subjects as the block-specific intercept scores sharply delineate a tightly integrated core comprised of Blocks 1 & 2 and a heterogeneous modular structure comprised of Blocks 3–10. The block-specific differential effects between the first and second groups are given as a $Q \times Q$ matrix of values $((\beta_{ql2}))_{q,l\in\{1,\ldots,Q\}}$ in **B** and were sampled on a scale between +2.50 and −2.50 for the two core blocks and set to 0.05 otherwise. Thus, the Control-Case differences in the cluster structure are represented specifically as deviations from the intra-block and inter-block connectivity on a common Core-Modular cluster structure. The cluster structure of the subjects in the first and second group is given in **C** and **D**, respectively. The sum of all the subject connectivity matrices is given in **E** and shows that, while each subject has clearly delineated cluster structure with 10 blocks (see **C** & **D**), the overall average of connectivity matrices shows only the evidence of 9 blocks, as Block 1 & Block 2 are merged.

Table 4
Ranges of adjusted ICL scores for Bin-SBM. For each simulated dataset and a given $\widehat{Q}$, the adjusted ICL scores are given as the difference between (i) the maximum score across the restarts (local maximum) and (ii) all values of $\widehat{Q}$ (global maximum). For each $\widehat{Q}$, the range of ICL scores is shown across all 100 Monte Carlo samples. The range $[0, 0]$ indicates that, for all 100 Monte Carlo samples, the best ICL score systematically underestimates the optimal number of clusters (i.e. $\widehat{Q} = 9$).

$\widehat{Q}$	Balanced	M. Unbalanced	Unbalanced
2	[−28480, −27138]	[−18727, −17732]	[−11519, −10847]
3	[−21206, −20078]	[−12998, −12049]	[−7678, −6280]
4	[−16318, −15356]	[−8938, −8318]	[−3482, −3152]
5	[−12402, −11101]	[−6500, −5443]	[−2080, −1800]
6	[−8487, −7671]	[−4303, −3379]	[−1269, −1058]
7	[−5618, −4785]	[−2039, −1778]	[−552, −390]
8	[−2371, −2121]	[−1074, −685]	[−330, −139]
9	[0, 0]	[0, 0]	[0, 0]
10	[−77, −44]	[−99, −46]	[−122, −47]
11	[−141, −92]	[−184, −107]	[−218, −105]
12	[−198, −148]	[−261, −158]	[−304, −183]
13	[−254, −212]	[−342, −212]	[−387, −269]
14	[−311, −278]	[−418, −284]	[−464, −334]
15	[−379, −342]	[−469, −354]	[−551, −400]
16	[−455, −423]	[−540, −429]	[−652, −464]
17	[−536, −496]	[−637, −510]	[−747, −536]
18	[−612, −578]	[−670, −592]	[−784, −605]

situation would be to introduce more initialisation strategies as in Simulation I.

Table 5
Ranges of adjusted ICL scores for Het-SBM. For each simulated dataset and a given $\widehat{Q}$, the adjusted ICL scores are presented as the difference between (i) the maximum score across the restarts (local maximum) and (ii) all values of $\widehat{Q}$ (global maximum). For each $\widehat{Q}$, the range of ICL scores is shown across all 100 Monte Carlo samples. The range $[0, 0]$ indicates that, for all 100 Monte Carlo samples in the Balanced and M. Unbalanced designs, the best ICL score systematically points towards the correct number of clusters (i.e. $\widehat{Q} = 10$). In the Unbalanced designs, however, the model accurately estimates the optimal number of clusters in 90% of simulations, but in 10% it overestimates the optimal number of clusters by one. This result can be explained by the limited number of initialisations used in this simulation.

$\widehat{Q}$	Balanced	M. Unbalanced	Unbalanced
2	[−66095, −64206]	[−68990, −67233]	[−63637, −62204]
3	[−49671, −48018]	[−49121, −46608]	[−45079, −43575]
4	[−37345, −35761]	[−36239, −34848]	[−32554, −31715]
5	[−25638, −23925]	[−22773, −21361]	[−17800, −17117]
6	[−18602, −17267]	[−13319, −12652]	[−8661, −8114]
7	[−13371, −12119]	[−9187, −7252]	[−3931, −3497]
8	[−8358, −7784]	[−3732, −3298]	[−1423, −1106]
9	[−4275, −3843]	[−1326, −1129]	[−1463, −228]
10	[0, 0]	[0, 0]	[−421, 0]
11	[−137, −126]	[−149, −127]	[−149, 0]
12	[−292, −266]	[−311, −274]	[−313, −147]
13	[−451, −424]	[−487, −426]	[−491, −313]
14	[−629, −598]	[−676, −603]	[−680, −512]
15	[−820, −779]	[−878, −783]	[−883, −681]
16	[−1021, −985]	[−1075, −974]	[−1093, −908]
17	[−1244, −1197]	[−1319, −1185]	[−1322, −1126]
18	[−1470, −1424]	[−1543, −1412]	[−1564, −1360]

The differences in results between Bin-SBM and Het-SBM are not surprising since the simulated data contains an apparent group effect (see panels (C) and (D) in Fig. 6, in which each group shows strong evidence for 10 clusters). However, by assuming that the estimation of cluster labels is independent of the group effect, information about the 10 cluster structure is lost and smoothed out into a cluster structure with 9 connectivity profiles (see panel (D) in Fig. 6). As a consequence of this, Bin-SBM fails to correctly estimate the optimal number of clusters by one cluster in all simulations and design cases. This example illustrates a potential issue that can arise when separating the clustering procedure from the logistic regression model and shows the benefit of using a joint optimisation of the cluster labels and regression model parameters. As we can see in the case of the Bin-SBM fit, this may lead to an inaccurate clustering, which then poses further problems for post hoc analysis, like fitting a regression model or making inferences.

The distribution of the total number of clusters across all approaches is shown in Fig. 7. The estimates of $\widehat{Q}$ for Bin-SBM and Het-SBM are consistent with the results shown in Tables 4 and 5. The respective pairwise median total number of clusters for Bin-SBM and Het-SBM is [9, 10] across all three designs. Their corresponding ARI median values (see Fig. 8) are [0.89,1], [0.76,1] and [0.66,1] across all three designs. The decrease of ARI scores observed for Bin-SBM across the three designs is linked to the increasing size of the first two clusters which tend to be merged by Bin-SBM. In contrast to this, depending on the tuning parameter values, the modular algorithm estimates of $\widehat{Q}$ show a wide scope of fitted values ranging from 1 to less than 200 clusters. The default value 1 for the tuning parameter seems to yield less extreme estimates across all three designs with pairwise median total number of clusters for FL-A and NS-A [7,7], [5,4] and [4,3], and FL-C and NS-C [8,7], [17,6] and [14,5]. The ARI scores for such cases seem to show a decreasing accuracy across the three designs for FL-A and NS-A with their respective median values of [0.59,0.59], [0.32,0.27] and [0.19,0.15]. The median values for FL-C and NS-C suggest some improvements across the three designs [0.73,0.66], [0.85,0.72] and [0.91,0.83].

3.5. Simulation III goals and setup

Goals of Simulation III. The goal of Simulation III is to assess the accuracy of the proposed parametric and non-parametric inference procedures in order to give recommendations for real data analyses. A particular focus is given to the inferential validity of parametric inferences (the Wald and likelihood ratio tests) and non-parametric inferences (permutation test).

Setup of Simulation III. The total number of subjects is set to 10, 20 and 40 ($K \in \{10, 20, 40\}$), which corresponds to the average numbers of subjects in a neuroimaging study. In the interest of computational feasibility and simplicity of presentation, the total number of clusters is set to three ($Q = 3$), and networks with 30, 60 and 120 nodes ($n \in \{30, 60, 120\}$) are utilised. The cluster sizes vary according to three designs Balanced, M. Unbalanced and Unbalanced (see **SI** N Table S9 for exact reference values). By controlling the total number of nodes, the total number of subjects and the cluster sizes, this simulation introduces a range of small sample scenarios that potentially could be encountered in real data analyses as the model is not restricted to fit clusters of specific sizes. Another type of small sample effects may occur in the data generated by probability values that are close to one or zero. In the literature, such values are known to introduce a significant bias and pose different problems to the validity of the parametric tests (see, for example, King and Zeng, 2001) while very little bias is expected for the values close to 0.5 for which the logit function tends to be linear. To explore such cases, we simplify the cluster structure by setting them to four different versions of intra-block and inter-block connectivity labelled as $\boldsymbol{\pi}_1, \ldots, \boldsymbol{\pi}_4$ with values 0.99/0.01, 0.95/0.05, 0.90/0.10 and 0.85/0.15 (see panel (A) in Fig. 9).

Network data is generated according to a logistic regression model which includes an intercept and age covariate (generated as a random sample of values between 20 and 60 with replacement). The intercept effects are set to the logit transformations of the four cluster structures ($\boldsymbol{\pi}_1, \ldots, \boldsymbol{\pi}_4$) shown in Fig. 9 panel (A) and the age effects are set to zero. In addition to this, we also consider a data generating process with random effects which introduces dependencies between edges within each block and violates the independence assumption of the model. More precisely, the data is generated according to subject-specific connectivity rates obtained on the logit scale from $\boldsymbol{d}_k^\top \widehat{\boldsymbol{\beta}}_{ql} + u_{qlk}$, where u_{qlk} is a block- and subject-specific random intercept obtained from the Standard Normal distribution and $\boldsymbol{d}_k^\top \widehat{\boldsymbol{\beta}}_{ql}$ is the fixed effect data. The use of a random effect allows us to introduce dependence between the edges within each subject-specific block. This scenario is relevant as dependencies between the edges of fitted blocks are quite plausible in real neuroimaging data, and this allows us to assess the likely behaviour of the proposed

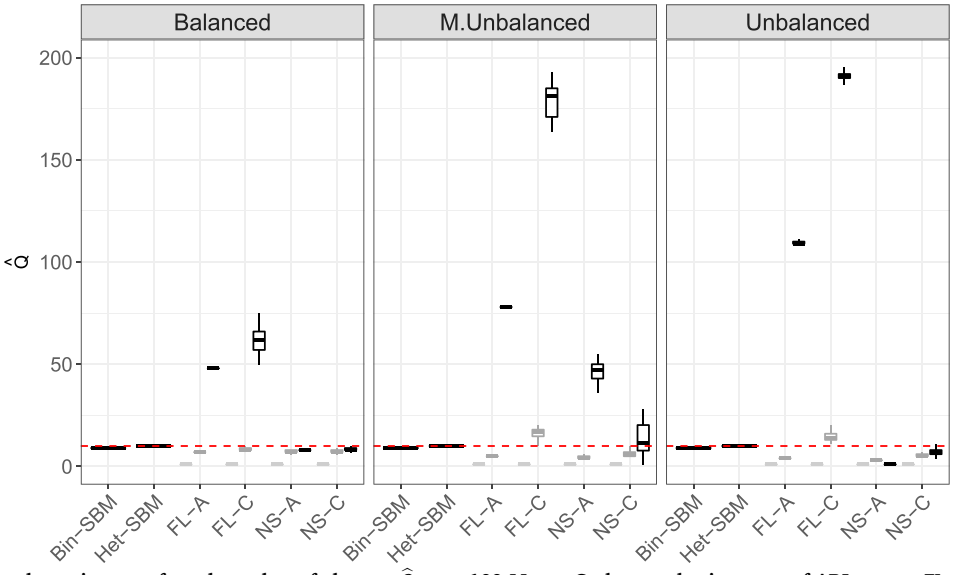

Fig. 7. Boxplots showing the estimates of total number of clusters $\widehat{Q}$ over 100 Monte Carlo samples in terms of ARI scores. FL and NS stand for the Fast

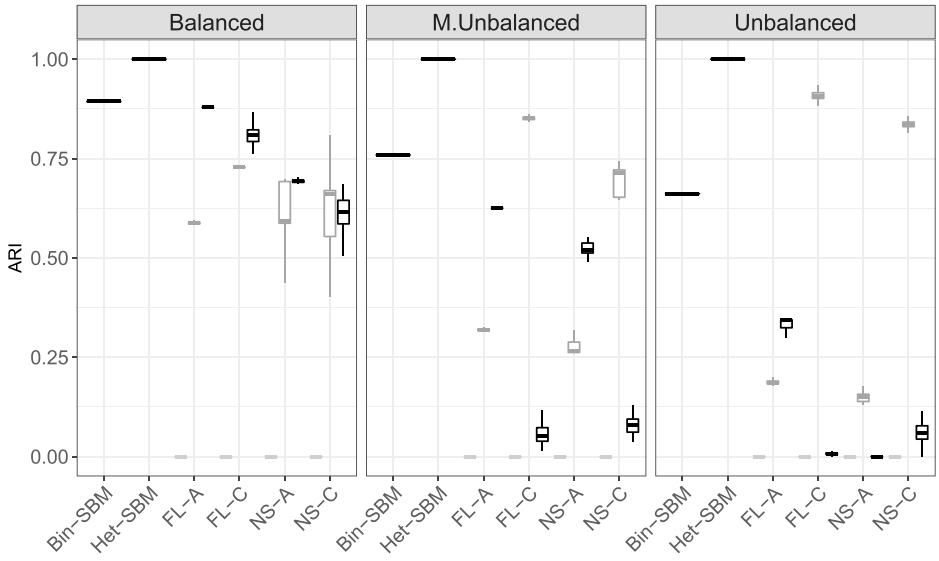

Fig. 8. Boxplots showing the recovery of the true cluster assignments over 100 Monte Carlo samples in terms of ARI scores. FL and NS stand for the Fast

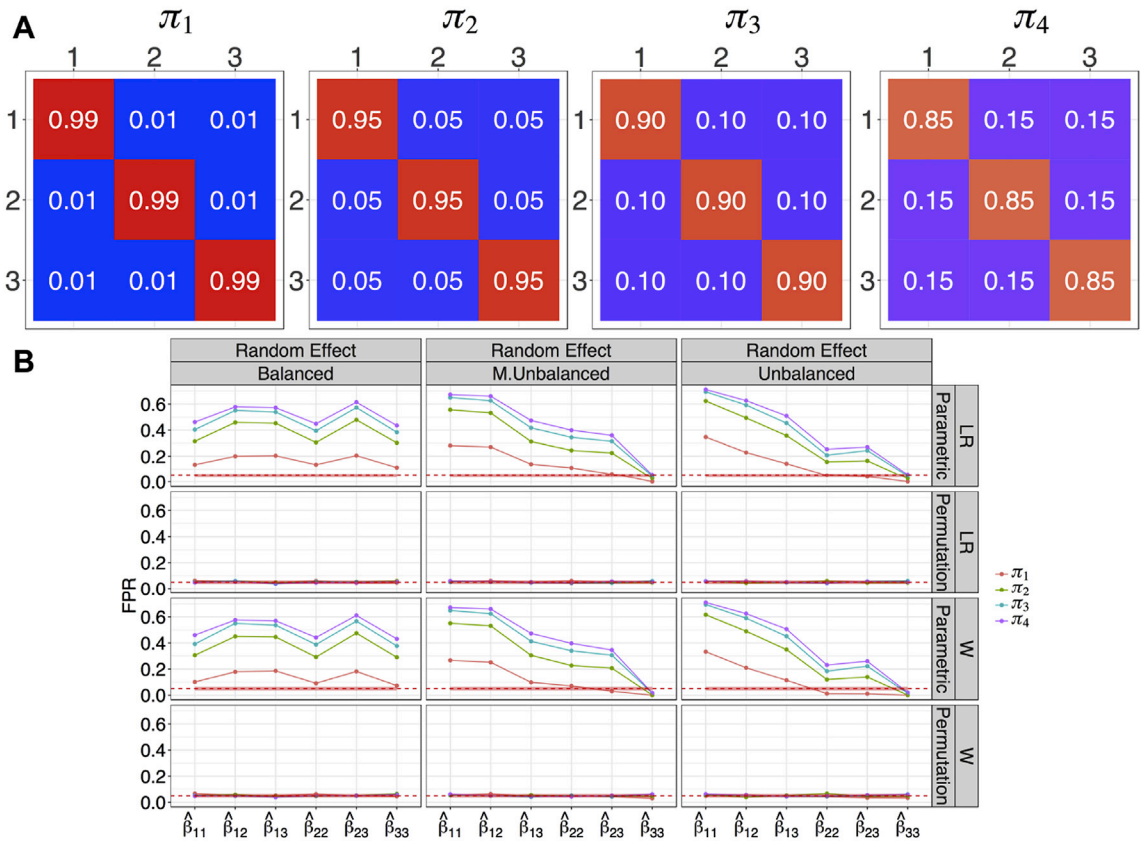

Fig. 9. A: Four different connectivity rates ($\pi_1, \ldots, \pi_4$) are used to simulate various levels of intra-block and inter-block connectivity in a homogeneous modular cluster structure. B: Observed FPR for the simulated networks with 30 nodes, 10 subjects and random effect. The *x*-axis represents the block specific fitted regression coefficients. The observed FPRs are colour coded according to 4 different connectivity matrices. The columns represent the 3 different block size designs, and the rows denote the likelihood ratio (LR) and Wald (W) scores based on the parametric and permutation tests. The red shaded strip represents the 95% Monte Carlo binomial proportion confidence intervals. Overall the parametric tests show highly inflated FPRs while the permutation test is accurate.

parametric and non-parametric tests in a more realistic setting.

In total, we generate 1000 matrices for each condition of the simulations, and we evaluate how well each of the parametric (the Wald and LR tests) and non-parametric (permutation test) inference procedures control the false positive rates (FPR). The *P*-values of the permutation tests are obtained by computing 1000 permutations of the age covariate in each of the simulations.

Table 6
Average observed FPR at 5% level significance in several scenarios. The average is taken across the 6 blocks, 4 connectivity matrices and 3 block size designs considered in Simulation III.

		No random effect				Random effect			
		Parametric		Permutation		Parametric		Permutation	
		LR	W	LR	W	LR	W	LR	W
30 nodes	10 subjects	0.044	0.035	0.053	0.050	0.338	0.324	0.052	0.051
	20 subjects	0.045	0.042	0.049	0.048	0.353	0.347	0.050	0.049
	40 subjects	0.050	0.046	0.051	0.050	0.367	0.363	0.052	0.052
60 nodes	10 subjects	0.049	0.045	0.051	0.050	0.554	0.546	0.053	0.052
	20 subjects	0.049	0.048	0.050	0.050	0.557	0.555	0.049	0.049
	40 subjects	0.050	0.049	0.051	0.051	0.581	0.579	0.053	0.053
120 nodes	10 subjects	0.051	0.049	0.053	0.053	0.735	0.733	0.052	0.051
	20 subjects	0.050	0.050	0.052	0.052	0.742	0.741	0.051	0.051
	40 subjects	0.051	0.050	0.052	0.052	0.755	0.755	0.053	0.054

3.6. Simulation III results

In Table 6, the average observed false positive rates (FPR) are generally well controlled by all inferential procedures when there was no random effect. Although the table gives raw averages across all simulation cases, there is still some evidence of conservative behaviour from the Wald test, which is more notable in the individual plots in **SI** N (see Figs. S25 – S33), and especially in the data sampled according to $\boldsymbol{\pi}_1$. In this instance, the blocks need to contain at least 12 nodes, and there should be at least 20 subjects in a study (see Fig. S30) to yield accurate inferences. While the likelihood ratio test also tends to be conservative in small samples, it appears to be less frequently sensitive to the small sample effects compared to the Wald test (e.g., see Fig. S25 M. Unbalanced design). In contrast to the parametric tests, the permutation tests are shown to be more accurate in the small sample scenarios and, for this reason, they seem to be more suited for real data applications than the parametric procedures.

In the more realistic scenario of a per block random effect, the parametric tests of both the Wald and likelihood ratio statistics had severely uncontrolled FPRs, whereas the permutation tests of both statistics maintained FPR ~ 0.05 in all the conditions (see Table 6 and panel (B) in Fig. 9). The parametric tests assume that the edges within a block are independent and because the simulation conditions violate these assumptions, the tests fail to control the FPRs. In contrast, the assumption of exchangeability, necessary for the permutation tests, is still valid because the presence of dependencies between edges within a subject does not impact the exchangeability between subjects. The full set of results for all subject numbers and cases can be found in **SI** N.

4. Multi-subject resting state fMRI data and analysis setup

Functional magnetic resonance imaging (fMRI) data were collected on 13 healthy volunteers (Controls) and 12 patients with schizophrenia (Cases), who have been scanned at rest after being given placebo medication (see Lynall et al., 2010, for details). Subjects were instructed to quietly lay in the scanner with their eyes closed for 17 min and 12 s. In each session, a total of 512 scans were acquired with a repetition time of 2 s. Each such data set was corrected for motion artefacts and then registered to the MNI standard space. A Gaussian kernel of 6 mm was used to smooth the registered images spatially and the time series were high-pass filtered with a cutoff frequency of ≈ 0.008 Hz. Each image was parcellated into 325 anatomically defined regions (ROIs) using the AFNI TT_N27_EZ_ML atlas (in the Talairach coordinate system) and 28 regions were discarded due to missing data for some individuals. As the majority of these were from the cerebellum, the remaining 29 cerebellar nodes were also removed (see **SI** O Fig. S42). The averaged voxel time series in each region was further decomposed into four frequency scales by the discrete wavelet transform (Percival and Walden, 2000). Our subsequent analysis considers correlations between regional fMRI time series in the frequency interval of $0.06 - 0.125$ Hz, which was previously shown to be the low-frequency range most strongly associated with Case-Control differences in pairwise wavelet correlations in these data (Lynall et al., 2010; Bullmore et al., 2001). These pairwise inter-regional estimates of band-passed correlations between low frequency fMRI oscillations were compiled to form a (268×268) functional connectivity matrix $((r_{ijk}))_{1 \le i,j \le n,k}$ for each subject k and then transformed by Fisher's formula (Fisher, 1915)

$$\frac{1}{2}\log\left[\frac{1+r_{ijk}}{1-r_{ijk}}\right]. \tag{31}$$

Under the null hypothesis that the population correlation is zero, the Fisher transformation statistic follows a Normal distribution with zero mean and variance $1/(T-3)$, where T is the number of discrete wavelet coefficients within a given frequency band; in this study, $T = 128$. On this basis, adjacency matrices can be constructed by testing if the Fisher transformation statistics are significantly higher than zero. Noting that this entails a large number of multiple, non-independent tests (35,778), the false discovery rate (FDR) was controlled at the 5% level (Benjamini and Yekutieli, 2001). If the results were significant, the value 1 was assigned and 0 otherwise. In this manner, each subject k is represented by an adjacency matrix $((x_{ijk}))_{1 \le i,j \le n}$ that captures a binary and undirected graph (or network).

To this set of adjacency matrices, we fit Het-SBM, described in Section 2.4, including as covariates in the regression model a common intercept for both groups (Cases & Controls), differential effect (Cases-Controls), and mean centred: age, premorbid intelligence, and head movement parameters. We use a similar initialisation strategy as in Section 3.1 so that we consider random label sampling, k-means and hierarchical clustering on each subject and their group average. For the first two procedures, we have used 1000 initialisations for each subject, their group average and a range of values for $\widehat{Q} \in \{2, 3, \ldots, 35\}$. The best fit is selected according to the ICL score.

5. Multi-subject resting state fMRI data analysis results

Het-SBM estimated a common cluster structure across both groups (Cases & Controls) comprising 21 blocks, whose anatomical labels can be found in **SI** P Figs. S44 and S45. The fit is based on the initialisation obtained by the hierarchical clustering on the group average data.

Fig. 10 panel (A) shows the average network data reorganised in terms of the 21-block fit found by Het-SBM. As shown in Fig. 10 panel (B), the network structure does not appear to be modular or core-modular. Instead, it seems to exhibit a structure with a varying degree of connectivity pattern per cluster with some clusters acting strongly as a core (e.g., Blocks 1, 2, 4, 12 & 17), some being mildly connected to other clusters (e.g., Blocks 3, 7, 10, 13 & 20) and finally some clusters acting

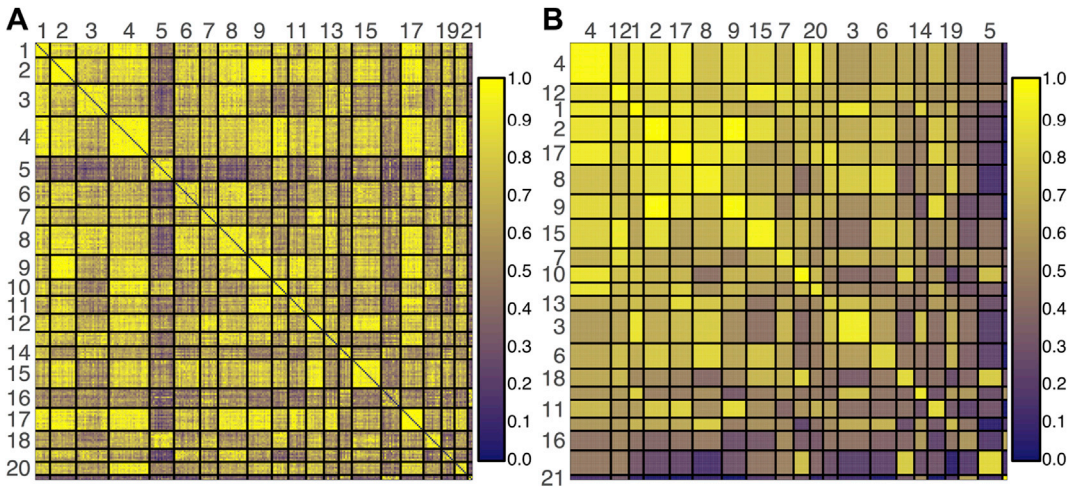

Fig. 10. A: Grand average network reorganised in terms of the 21-block obtained from the Het-SBM fit. B: Fitted connectivity matrix from the Het-SBM fit, ordered by decreasing average block probabilities. The network structure seems to present a non-modular structure.

more like modular cluster being almost not connected to any other clusters (e.g., Blocks 5 & 21). We can also observe that Het-SBM separate well some clusters that may have some high degree of connection between them, but a somewhat different profile of connection with the other clusters. For example, Blocks 5 & 10 exhibit a high within- and between-block connectivity, but Het-SBM separates them because their connectivity profiles to the other clusters are different.

In Fig. 11, we show the family-wise error rate corrected *P*-values for the test of difference between the average block connectivity levels of Patient and Control groups based on the LR parametric test (see panel (A)) and the permutation test on LR scores (see panel (B)). The non-significant block *P*-values, thresholded at a 5% level significance, are given in white while the significant block *P*-values are plotted on a $-\log_{10}(\cdot)$ scale so that, for example, the value of 305.3 represents the *P*-value of $10^{-305.3}$.

There is a striking difference between the two tests in terms of the number of significantly declared blocks. In particular, the LR test seems to declare a total of 215 significant blocks which is about 107 times more than the number of significant blocks declared by the permutation test. This particular scenario is mainly consistent with the results of Simulation III and is suggestive of an invalid LR test. Indeed, in Simulation III, we have already shown that when there is some sort of dependence between the edges of a block (i.e. per block random effect) the asymptotic test tend to have profoundly inflated false positive rates. To investigate this on the real multi-subject data, we conducted a parametric bootstrap test based on LR statistics to find evidence of a random effect in each of the blocks (the full details on the fitting and hypothesis testing procedures are given in SI O.1). Except for Block (1, 1) and Block (17, 17), all the remaining fitted blocks in Fig. S43 show significant random effects. Given the nature of functional connectivity data and strongly coupled correlation patterns, it is not surprising that the dependencies between the nodes continue to persist in the data, even after the clustering model is fitted. Interestingly, such dependences in Simulation III showed almost no particular impact on the clustering results but proved to be very disastrous for the control of false positive rates in the asymptotic tests. In contrast, the overall performance of the permutation test is unaffected by

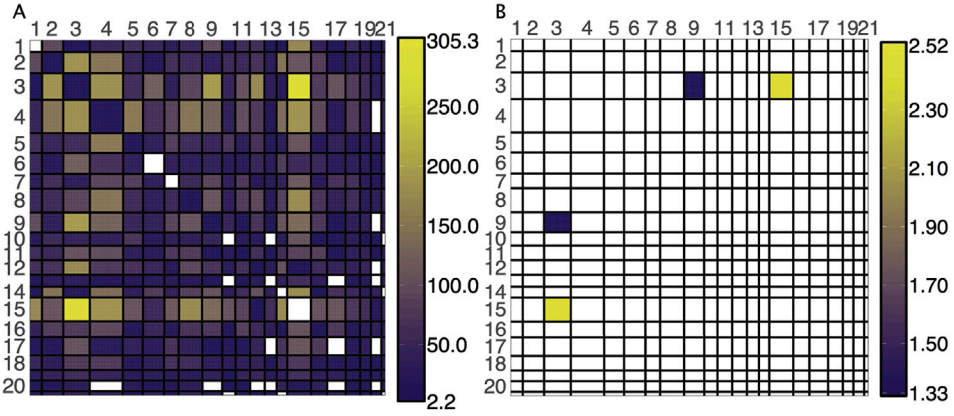

Fig. 11. Family-wise error rate corrected significant *P*-values for the null hypothesis of no average difference in connectivity between the Patient and Control groups with panel (A) showing the results of the LR test and panel (B) showing the results of the permutation test based on the LR statistic. The *P*-values were computed according to the strategies described in Section 2.5 and are stated in terms of $-\log_{10}(P-\text{values})$ (e.g., the legend score of 305.3 corresponds to the *P*-value of $10^{-305.3}$) and the non-significant *P*-values at 5% significance are given in white.

this condition, and the false positive rates are controlled reasonably well. This specific observation is echoed in Fig. 11 panel (B), where we only see two significant blocks, namely the connectivities between Blocks 3 & 9 and Blocks 3 & 15.

In order to see the direction of differential effect (Patients vs Controls) and the anatomical locations of nodes in Blocks 3, 9 and 15, we look at the observed differences in their connections (Δx_{ij}) at each edge. This statistic is expressed as the difference between the two groups relative to their total number of subjects (i.e. $\Delta x_{ij} = \frac{1}{K_1}\sum_{k=1}^{K_1} x_{ijk} - \frac{1}{K_2}\sum_{k=1}^{K_2} x_{ijk}$) where K_1 and K_2 are the total number of subjects in the Patient and Control groups, respectively. As shown in Fig. 12, the connectivity rates between the blocks tend to be much weaker in Patients than Controls, and this is consistent with the general literature which classifies schizophrenia as 'dysconnectivity disorder' (Friston and Frith, 1995). This profile of attenuated functional connectivity between areas of the temporal, frontal and cingulate cortex is compatible with other graph-theoretical and functional connectivity studies of schizophrenia that have emphasised fronto-temporal profiles of dysconnectivity (Fornito et al., 2012). The observation that functional dysconnectivity in schizophrenia is restricted to a subset of clusters is more novel but consistent with previous studies demonstrating that abnormal developmental trajectories of cortical shrinkage in schizophrenia were restricted to a specific module of the structural network community structure comprising areas of frontal, temporal and cingulate cortex (Alexander-Bloch et al., 2013, 2014). It seems that the cluster structure of brain networks may constrain the expression of dysconnectivity phenotypes in schizophrenia. This, in turn, may be related to the emerging idea that the modular organisation of brain networks reflects co-expression of genes and that genes associated with a risk for schizophrenia have a non-uniform expression across the cortex with high levels of schizophrenia risk gene expression associated with developmental changes in structure of frontal cortex (Whitaker et al., 2016).

6. Discussion

This work proposed a multi-subject framework based on three extensions of the classical SBM, referred to as Bin-SBM, Hom-SBM and Het-SBM. The last two models are non-trivial and use subject-specific covariates to explain variations in cluster structure between subjects. Focusing on Het-SBM, this work benchmarked it against classical modular algorithms, explored the validity of its inference procedures based on parametric and non-parametric tests and how these could be used in a real data analysis to test for a group effect between healthy controls and patients diagnosed with schizophrenia.

6.1. Benchmarking the models

To investigate the modelling strengths and weaknesses of Het-SBM and Bin-SBM, we conducted three separate simulations (Simulations I-III).

Simulation I. In Simulation I, we compared our models to popular clustering methods such as the Newman Spectral (NS) and Fast Louvain (FL) algorithms, based on (i) average (-A) and (ii) consensus clustering (-C), with different values for the resolution parameter γ. Overall, the MS-SBMs outperformed the modular algorithms and were accurate in all scenarios. The modular algorithms showed accuracy in the Hom-Modular structure and reasonably balanced cluster sizes with an almost negligible effect of γ. In the Hom-Modular structure with the Unbalanced cluster size design, consensus clustering was found to perform better than average clustering. However, it seems possible to improve the results of average clustering by using the larger values of γ. For Het-Modular and Core-Modular structures, the modular solutions were found to be less accurate, with a higher dependence on γ. These results suggest that, in practice, special care must be taken when choosing the value of γ, especially when the cluster structure deviates from a pure Hom-Modular structure. Another problematic aspect is that the resolution parameter can lead to a very strong overestimation of the total number of clusters and introduce partitions which are not parsimonious. In Het-SBM, this problem is solved with the ICL criterion, which penalises the log-likelihood for the complexity of the model and preserves parsimony. Unlike modularity scores, which cannot be compared over different values of γ (controlling the total number of clusters), ICL scores are comparable across different values of Q. Furthermore, ICL scores can be used to evaluate the partitions obtained through any clustering method and obtain their goodness of fits.

In our simulations, the NS and FL algorithms were found to overestimate the correct number of blocks by introducing many single-node blocks. This selection bias may be a potential concern, especially because the ground truth in our simulations did not contain such blocks. However, it is possible that these nodes may be clustered alone because (i) they deviate from perfectly modular connectivity profile and the algorithms struggle to cluster them due to the limited choices of the clustering patterns, or because (ii) these nodes display unique patterns of connectivity which supports their clustering into single block nodes. Given that the simulations used ground-truth partitions with at least five nodes per-block, and with distinct block connectivity patterns, we believe the latter to be unlikely. Nevertheless, if the true clustering does involve single-node blocks, the clustering method in question would be expected to estimate them. Specific to the case of MS-SBMs, single-node blocks are indeed possible, but in our experience, they tend to be very rare. Observing such extreme patterns might indicate artefacts in the data, but if such possibilities have been ruled out, it is plausible to have them as the real feature of the data. In these instances, the data is expected to show strong evidence for their existence by attaining the winning ICL score, thus showing that other types of cluster structures are less likely.

BN-SBM in Simulation I. In addition to the classical modular algorithms, we also benchmarked Block to Node SBM (or BN-SBM) (Newman and Leicht, 2007), which can be regarded as a particular case (or special re-parametrisation) of the classical SBM of Snijders and Nowicki (1997) (see **SI** K for details). There are, however, two main reasons as to why this model could not be benchmarked in the full simulation setting but only on a subset of cases. First, this model is intended for the analysis of a single, binary network, and our simulations are multi-subject. Second, the model and its `C` based implementation use estimated likelihood scores to compare different fits, which are only valid for partitions with the same Q. Unlike MS-SBMs which use ICL scores to compare fits based on different values of Q, this model can only be benchmarked when the ground-truth cluster number is supplied (i.e. Q was set to 10). As shown in Fig. S10, the model performs very well in the case of the Hom-Modular structure, but it is less accurate in the Het-Modular and Core-Modular cases. One possible explanation is that the combination of block-to-node parametrisation and the parametric constraints of Categorical distribution (i.e. that the parameters must add up to 1 across the nodal categories) pose limitations on the types of cluster structures that the model can estimate. It appears that the modelling of nodal connectivity profiles comes at a substantial cost to the model richness and ability to estimate different cluster structures, a trade-off that we have found to be unsatisfactory.

Simulation II. In Simulation II, we showcased an example in which the between-subject variation induced by a covariate strongly influenced the overall cluster structure. As a result, Bin-SBM erroneously merged the two clusters while Het-SBM correctly treated them as two distinct clusters. This example demonstrates that it can be essential to account for between-subject variations in cluster structure when estimating the cluster labels. Therefore, it is not advisable to perform clustering separately from the regression analysis.

Simulation III. in Simulation III, we tested the accuracy of Het-SBM inference procedures based on parametric and non-parametric tests. Overall, non-parametric (permutation) tests were found to be more accurate than the Wald and LR tests. When the assumptions of independence between the edges in a block are satisfied, parametric and non-parametric

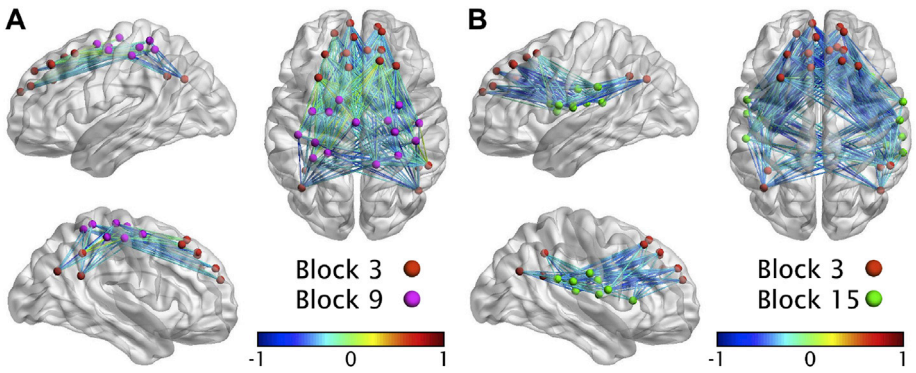

Fig. 12. Difference between the observed average connectivities of Patients vs Controls (Δx_{ij}) at each edge. The difference in edges is computed between Block 3 & Block 9 panel (A) and Block 9 & Block 15 panel (B). Overall, the connectivity strengths between blocks tend to be much weaker in Patients than Controls.

tests were both found to be valid. However, in the instances of dependence between edges in a block, the parametric Wald and LR tests were found to be invalid due to liberal control of FPRs while the permutation tests remained accurate. These results lend strong support for permutation testing as the most reasonable inference method for situations where independence between block edges cannot be assumed. Nevertheless, it is important to note that permutation testing relies on the assumption that data is exchangeable under the null hypothesis. This assumption may not hold if the data exhibits some form of heteroskedasticity. For example, if we assume that data is generated through a mixed-effects model with a different random intercept per group, data will not be exchangeable between groups, and the permutation test may not be valid.

Inference with Het-SBM in Real Data. In this paper, we have illustrated Het-SBM in an application to a resting-state fMRI study with healthy subjects and subjects diagnosed with schizophrenia, which have yielded a fit with 21 clusters. Interestingly, this data contained the same type of edge dependencies in the blocks as in Simulation III, which made the parametric LR test liberal. While more time consuming (approximately 25 s per permutation), a much more reliable inference was achieved by the permutation test on LR scores which found significant differences between the two groups in Block $(3, 9)$ and Block $(3, 15)$, containing nodes in temporal, frontal and cingulate cortex.

Sensitivity Analysis of Het-SBM. While the settings in Simulations I-III were useful to benchmark Het-SBM, they did not provide much information about the sensitivity of the model to separate clusters with very similar connectivity profiles. For this reason, we have performed an additional set of simulations in **SI** J that studies the sensitivity of Het-SBM to separate such a kind of clusters. As can be seen in Fig. S8, Het-SBM may fail to separate such clusters if the difference in connectivity is weak and if the number of subjects is low. However, as shown in Fig. S8 this weakness of the model can be overcome by increasing the total number of subjects.

6.2. Generalised MS-SBMs

Non-binary Edges. Using the general approach outlined in this paper, it is also relatively simple to derive other variants of Het-SBM or Hom-SBM for non-binary edges. As any distribution in an exponential family of distributions can be utilised to describe the edges in a dataset, the estimation strategies described in this work can be utilised in the same way (as mentioned above in this paper), to derive a variant of Het-SBM that suits the edges of the dataset. For example, if researchers are interested in clustering observed correlations between ROI time series (without a thresholding step), one could utilise Fisher Z to transform the correlations and subsequently assume that edges are sampled from Normal distribution. This situation dramatically simplifies the estimation procedure as a Normal linear model utilises closed-form solutions for maximum likelihood estimates. Likewise, one can imagine the uses for this type of modelling in networks obtained from diffusion tensor imaging. For example, if the edges represent white matter fibre counts between the ROIs, a Poisson distribution might be a reasonable fit for such networks. Subsequently, one would use a Poisson regression model which would invoke the Fisher scoring step as in this work. In general, depending on the type of density used in the modelling of the edge data, researchers could choose an appropriate generalised linear model while still applying the methodology discussed in this work in order to develop a Het-SBM model tailored to their dataset. As part of our future work, we will implement these two new versions of the model.

Edge-wise Covariates. While, in this work, we mainly focused on multi-subject models where the distribution of edges is dependent on subject-specific covariates like age or gender, it is also possible to adapt Het-SBM and Hom-SBM for a distribution of edges that is dependent on edge-level or node-level covariates. For example, (i) between-regional Euclidean distances, (ii) total numbers of tractography streamlines, and (iii) correlations based on cortical thickness, can be seen as edge-level covariates. The Euclidean distance can be a powerfully informative factor in explaining the cluster structure, especially as the long-distance and short-distance connections are suggestive of integration and segregation in the brain. Thus, assuming heterogeneous effects (i.e. Het-SBM), we can write this model as $X_{ijk}|Z_{iq} = 1, Z_{jl} = 1 \sim \text{Bernoulli}(\pi_{ijk})$, where $\log\left[\frac{\pi_{ijk}}{1-\pi_{ijk}}\right] = \boldsymbol{d}_{ijk}^{\top}\boldsymbol{\beta}_{ql}$ such that $\boldsymbol{d}_{ijk}$ is a $1 \times P$ vector of edge features associated with the subject k and edge x_{ij}. In this particular example, $P = 4$, as the regression model consists of the intercept, Euclidean distance, total number of tractography streamlines and cortical thickness correlations. In general, it can be easily shown that by setting $\boldsymbol{d}_{ijk}$ to be equal to $\boldsymbol{d}_k$, this model becomes exactly the Het-SBM discussed in this work. In this version of the model, edge-level covariates could have a much stronger influence on the estimation of cluster assignments than subject-level covariates, so it would be interesting to see how much of this influence is tied to the homogeneous and heterogeneous versions of the model. In parallel to this, it would also be interesting to detect a collection of edge-covariates which can explain the clustering of network data that give rise to known resting-state networks. By doing this type of analysis, the researchers could obtain a host of potential network-based biomarkers, which could potentially be more closely tied to specific brain disorders. However, a potential limitation to these edge-covariate models is the possible violation of the exchangeability assumption that renders the permutation testing inappropriate. In such a situation, alternative resampling strategies will be needed, and the accuracy of selected inference procedures must be investigated.

Nodal Covariates. It is also possible to consider the use of node-based covariates for modelling distribution of edges. One way to account for them is to transform those node-based covariates into edge-based covariates using a set of meaningful symmetric transformation functions $f_p(a, b)$ such that the element of the edge-based covariate p corresponding

to the nodes V_i and V_j is given by $f_p(d_{ikp}, d_{jkp})$. The choice these functions would depend on the nature of the node-based covariates and could be, for example, the mean function $1/2(a+b)$ which would assume an addictive effect of the node-based covariate values on the edge connectivities.

6.3. Limitations and future work

It is important to note that the current setup of our proposed approach does not allow for per-subject varying cluster labels. This is a limitation of our model that will be important to overcome as some studies have identified that those cluster assignments of nodes substantially change as a function of a task (Hearne et al., 2017) and diagnostic status (Alexander-Bloch et al., 2012). More recently, Gordon et al. (2017) and Kong et al. (2018) suggested that the cluster structure is individualised and exhibits subject-specific nuances. As part of our future work, we are now actively working on an extension of Het-SBM, which will accommodate per-subject varying cluster labels and compare it to the modular multi-subject alternatives proposed by Betzel et al. (2018a).

The current model makes inferences on covariate effects which are dependent on cluster label fit. The accuracy of the fit is paramount in the interpretation of the results, and it is closely related to the functionally meaningful block structures. If blocks cannot be associated with a particular brain function, it may results in statistically significant blocks without any meaningful interpretation. This may be circumvented by conducting an edge-specific inference. However, this strategy goes against the principle of parsimony, currently enforced by making a block-specific inference. Moreover, edge-specific inferences would drastically increase the number of tests, worsening the issue of multiple comparisons. Therefore, while it is indeed possible to conduct edge-wise inferences, block-specific inferences might be preferable, particularly if the blocks can be associated with meaningful functions. Also, it is worth noting that the issue of selecting optimal clustering is not uniquely specific to our model, but to all clustering methods. In general, given the same dataset, all clustering methods experience similar limitations as they explore the same space of candidate models (i.e. potential cluster labels) and this space for a particular Q has a finite size of Q^n. Given that this is the extremely high number of fits to consider, it is tough to ensure that the final model is indeed a global solution and not merely a reasonable local maximum. In our analysis of real data, Het-SBM yielded a fit with 21 clusters and significant differences between the two groups in two blocks. Ideally, a repeated sample would be needed to confirm these findings, especially as the sample size of the dataset is rather small, which may have prevented to discover other significant associations. Thus, it would be interesting to repeat this analysis on a larger cohort and to see how robust and reproducible these results are.

CRediT authorship contribution statement

Dragana M. Pavlović: Conceptualization, Methodology, Software, Formal analysis, Investigation, Writing - original draft, Writing - review & editing, Visualization, Project administration. **Bryan R.L. Guillaume:** Software, Formal analysis, Writing - review & editing, Visualization. **Emma K. Towlson:** Data curation, Writing - review & editing. **Nicole M.Y. Kuek:** Writing - review & editing. **Soroosh Afyouni:** Visualization. **Petra E. Vértes:** Data curation, Writing - review & editing. **B.T. Thomas Yeo:** Writing - review & editing, Supervision, Funding acquisition. **Edward T. Bullmore:** Writing - review & editing, Funding acquisition. **Thomas E. Nichols:** Writing - review & editing, Supervision, Funding acquisition, Project administration.

Acknowledgements

We would like to thank Eugene Demidenko, David Firth, Iréne Gannaz, Georg Heinze and Stéphane Robin for valuable discussions at various stages of this project. We would also like to thank Andrew Zalesky and a cohort of anonymous reviewers for providing great suggestions which improved the quality of the paper.

D.M.P was supported by the MRC Industrial CASE award with the GlaxoSmithKlines Clinical Unit Cambridge (UK) PhD studentship. B.R.L.G was supported by the EU within the PEOPLE Programme (FP7): Initial Training Networks (FP7-PEOPLE-ITN-2008), Grant Agreement No. 238593 NEUROPHYSICS. E.K.T was funded by the Medical Research Council. P.E.V was supported by the Medical Research Council (grant number MR/K020706/1) and is a Fellow of MQ: Transforming Mental Health (MQF17_24) and of the Alan Turing Institute funded by EPSRC grant EP/N510129/1. Y.B.T.T is supported by NUS Tier 1, Singapore MOE Tier 2 (MOE2014-T2-2-016), NUS Strategic Research (DPRT/944/09/14), NUS SOM Aspiration Fund (R185000271720), Singapore NMRC (CBRG14nov007, NMRC/CG/013/2013) and NUS YIA. E.T.B is employed half-time by the University of Cambridge and half-time by GlaxoSmithKline; and holds stock in GSK. The Behavioural and Clinical Neuroscience Institute is supported by the Medical Research Council (UK) and Wellcome Trust. T.E.N is supported by NIH U54MH091657-03 and the Wellcome Trust. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.neuroimage.2020.116611.

References

Airoldi, E.M., Blei, D.M., Fienberg, S.E., Xing, E.P., 2009. Mixed membership stochastic blockmodels. In: In: Advances in Neural Information Processing Systems, pp. 33–40.
Alexander-Bloch, A., Lambiotte, R., Roberts, B., Giedd, J., Gogtay, N., Bullmore, E., 2012. The discovery of population differences in network community structure: new methods and applications to brain functional networks in schizophrenia. Neuroimage 59 (4), 3889–3900.
Alexander-Bloch, A., Raznahan, A., Bullmore, E., Giedd, J., 2013. The convergence of maturational change and structural covariance in human cortical networks. J. Neurosci. 33 (7), 2889–2899.
Alexander-Bloch, A.F., Reiss, P.T., Rapoport, J., McAdams, H., Giedd, J.N., Bullmore, E.T., Gogtay, N., 2014. Abnormal cortical growth in schizophrenia targets normative modules of synchronized development. Biol. Psychiatr. 76 (6), 438–446.
Ambroise, C., Matias, C., 2012. New consistent and asymptotically normal parameter estimates for random-graph mixture models. J. Roy. Stat. Soc. B 74 (1), 3–35.
Ambrosen, K.S., Herlau, T., Dyrby, T., Schmidt, M.N., Mørup, M., 2013. Comparing structural brain connectivity by the infinite relational model. In: 2013 InternationalWorkshop on Pattern Recognition in Neuroimaging. IEEE, pp. 50–53.
Ambrosen, K.S., Albers, K.J., Dyrby, T.B., Schmidt, M.N., Mørup, M., 2014. Nonparametric bayesian clustering of structural whole brain connectivity in full image resolution. In: 2014 InternationalWorkshop on Pattern Recognition in Neuroimaging. IEEE, pp. 1–4.
Andersen, K.W., Madsen, K.H., Siebner, H.R., Schmidt, M.N., Mørup, M., Hansen, L.K., 2014. Non-parametric bayesian graph models reveal community structure in resting state fmri. NeuroImage 100, 301–315.
Benjamini, Y., Yekutieli, D., 2001. The control of the false discovery rate in multiple testing under dependency. Ann. Stat. 1165–1188.
Betzel, R.F., Bertolero, M.A., Gordon, E.M., Gratton, C., Dosenbach, N.U., Bassett, D.S., 2018a. The Community Structure of Functional Brain Networks Exhibits Scale-specific Patterns of Variability across Individuals and Time. bioRxiv, p. 413278.
Betzel, R.F., Medaglia, J.D., Bassett, D.S., 2018b. Diversity of meso-scale architecture in human and non-human connectomes. Nat. Commun. 9 (1), 346.
Bickel, P., Choi, D., Chang, X., Zhang, H., 2013. Asymptotic normality of maximum likelihood and its variational approximation for stochastic blockmodels. Ann. Stat. 1922–1943.
Biernacki, C., Celeux, G., Govaert, G., et al., 1998. Assessing a Mixture Model for Clustering with the Integrated Classification Likelihood.
Blondel, V., Guillaume, J., Lambiotte, R., Lefebvre, E., 2008. Fast unfolding of communities in large networks. J. Stat. Mech. Theor. Exp. 2008 (10), P10008.
Bullmore, E., Long, C., Suckling, J., Fadili, J., Calvert, G., Zelaya, F., Carpenter, T.A., Brammer, M., 2001. Colored noise and computational inference in neurophysiological (fmri) time series analysis: resampling methods in time and wavelet domains. Hum. Brain Mapp. 12 (2), 61–78.
Choi, D.S., Wolfe, P.J., Airoldi, E.M., 2012. Stochastic Blockmodels with a Growing Number of Classes. Biometrika, asr053.
Côme, E., Latouche, P., 2015. Model selection and clustering in stochastic block models based on the exact integrated complete data likelihood. Stat. Model. Int. J. 15 (6), 564–589.

Crossley, N.A., Mechelli, A., Scott, J., Carletti, F., Fox, P.T., McGuire, P., Bullmore, E.T., 2014. The hubs of the human connectome are generally implicated in the anatomy of brain disorders. Brain 137 (8), 2382–2395.
Daudin, J., Picard, F., Robin, S., 2008. A mixture model for random graphs. Stat. Comput. 18 (2), 173–183.
Dosenbach, N.U., Nardos, B., Cohen, A.L., Fair, D.A., Power, J.D., Church, J.A., Nelson, S.M., Wig, G.S., Vogel, A.C., Lessov-Schlaggar, C.N., et al., 2010. Prediction of individual brain maturity using fmri. Science 329 (5997), 1358–1361.
Firth, D., 1993. Bias reduction of maximum likelihood estimates. Biometrika 80 (1), 27–38.
Fisher, R.A., 1915. Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. Biometrika 507–521.
Fornito, A., Zalesky, A., Pantelis, C., Bullmore, E.T., 2012. Schizophrenia, neuroimaging and connectomics. Neuroimage 62 (4), 2296–2314.
Friston, K.J., Frith, C.D., 1995. Schizophrenia: a disconnection syndrome. Clin. Neurosci. 3 (2), 89–97.
Gao, C., Lu, Y., Zhou, H.H., et al., 2015. Rate-optimal graphon estimation. Ann. Stat. 43 (6), 2624–2652.
Gates, A.J., Wood, I.B., Hetrick, W.P., Ahn, Y.-Y., 2019. Element-centric clustering comparison unifies overlaps and hierarchy. Sci. Rep. 9 (1), 8574.
Good, P., 2000. Permutation Tests. Springer.
Gordon, E.M., Laumann, T.O., Gilmore, A.W., Newbold, D.J., Greene, D.J., Berg, J.J., Ortega, M., Hoyt-Drazen, C., Gratton, C., Sun, H., et al., 2017. Precision functional mapping of individual human brains. Neuron 95 (4), 791–807.
Handl, J., Knowles, J., Kell, D., 2005. Computational cluster validation in post-genomic data analysis. Bioinformatics 21 (15), 3201–3212.
Hearne, L.J., Cocchi, L., Zalesky, A., Mattingley, J.B., 2017. Reconfiguration of brain network architectures between resting-state and complexity-dependent cognitive reasoning. J. Neurosci. 37 (35), 8399–8411.
Heinze, G., Schemper, M., 2002. A solution to the problem of separation in logistic regression. Stat. Med. 21 (16), 2409–2419.
Hinne, M., Ekman, M., Janssen, R.J., Heskes, T., van Gerven, M.A., 2015. Probabilistic clustering of the human connectome identifies communities and hubs. PloS One 10 (1), e0117179.
Holm, S., 1979. A simple sequentially rejective multiple test procedure. Scand. J. Stat. 65–70.
Hu, H.-B., Wang, X.-F., 2009. Disassortative mixing in online social networks. EPL (Europhysics Letters) 86 (1), 18003.
Hubert, L., Arabie, P., 1985. Comparing partitions. J. Classif. 2 (1), 193–218.
King, G., Zeng, L., 2001. Logistic regression in rare events data. Polit. Anal. 9 (2), 137–163.
Kong, R., Li, J., Orban, C., Sabuncu, M.R., Liu, H., Schaefer, A., Sun, N., Zuo, X.-N., Holmes, A.J., Eickhoff, S.B., et al., 2018. Spatial topography of individual-specific cortical networks predicts human cognition, personality, and emotion. Cerebr. Cortex 29 (6), 2533–2551.
Kosmidis, I., 2014. Bias in parametric estimation: reduction and useful side-effects. Wiley Interdisciplinary Reviews: Comput. Stat. 6 (3), 185–196. https://doi.org/10.1002/wics.1296.
Latouche, P., Birmele, E., Ambroise, C., 2012. Variational bayesian inference and complexity control for stochastic block models. Stat. Model. Int. J. 12 (1), 93–115.
Latouche, P., Birmelé, E., Ambroise, C., 2014. Overlapping Clustering Methods for Networks.
Latouche, P., Birmelé, E., Ambroise, C., et al., 2011. Overlapping stochastic block models with application to the French political blogosphere. Ann. Appl. Stat. 5 (1), 309–336.
Lucas, A., 2014. Amap: Another Multidimensional Analysis Package. R Package Version 0.8-14. URL. https://CRAN.R-project.org/package=amap.
Lynall, M.-E., Bassett, D.S., Kerwin, R., McKenna, P.J., Kitzbichler, M., Muller, U., Bullmore, E., 2010. Functional connectivity and brain networks in schizophrenia. J. Neurosci. 30 (28), 9477–9487.
Mariadassou, M., Robin, S., Vacher, C., et al., 2010. Uncovering latent structure in valued graphs: a variational approach. Ann. Appl. Stat. 4 (2), 715–742.
Matias, C., Miele, V., 2017. Statistical clustering of temporal networks through a dynamic stochastic block model. J. Roy. Stat. Soc. B 79 (4), 1119–1141.
Matias, C., Robin, S., 2014. Modeling heterogeneity in random graphs through latent space models: a selective review. ESAIM: Proceedings and Surveys 47, 55–74.
Mørup, M., Schmidt, M.N., 2012. Bayesian community detection. Neural Comput. 24 (9), 2434–2456.
Mørup, M., Schmidt, M.N., Hansen, L.K., 2011. Infinite multiple membership relational modeling for complex networks. In: 2011 IEEE International Workshop on Machine Learning for Signal Processing. IEEE, pp. 1–6.
Mossel, E., Neeman, J., Sly, A., et al., 2016. Belief propagation, robust reconstruction and optimal recovery of block models. Ann. Appl. Probab. 26 (4), 2211–2256.
Moyer, D., Gutman, B., Prasad, G., Faskowitz, J., Ver Steeg, G., Thompson, P., 2015. Blockmodels for connectome analysis. In: 11th International Symposium on Medical Information Processing and Analysis, vol. 9681. International Society for Optics and Photonics, p. 96810A.
Murtagh, F., Legendre, P., 2014. Ward's hierarchical agglomerative clustering method: which algorithms implement ward's criterion? J. Classif. 31 (3), 274–295.
Newman, M., 2006. Modularity and community structure in networks. Proc. Natl. Acad. Sci. Unit. States Am. 103 (23), 8577–8582.
Newman, M.E., Leicht, E.A., 2007. Mixture models and exploratory analysis in networks. Proc. Natl. Acad. Sci. Unit. States Am. 104 (23), 9564–9569.
Nowicki, K., Snijders, T., 2001. Estimation and prediction for stochastic blockstructures. J. Am. Stat. Assoc. 96 (455), 1077–1087.
Olhede, S.C., Wolfe, P.J., 2014. Network histograms and universality of blockmodel approximation. Proc. Natl. Acad. Sci. Unit. States Am. 111 (41), 14722–14727.
Pavlovic, D.M., 2015. Generalised Stochastic Blockmodels and Their Applications in the Analysis of Brain Networks. Ph.D. thesis. University of Warwick.
Pavlovic, D.M., Vértes, P.E., Bullmore, E.T., Schafer, W.R., Nichols, T.E., 2014. Stochastic blockmodeling of the modules and core of the caenorhabditis elegans connectome. PloS One 9 (7), e97584.
Percival, D.B., Walden, A.T., 2000. Wavelet Methods for Time Series Analysis (Cambridge Series in Statistical and Probabilistic Mathematics).
Picard, F., Miele, V., Daudin, J.-J., Cottret, L., Robin, S., 2009. Deciphering the connectivity structure of biological networks using mixnet. In: BMC Bioinformatics, vol. 10. BioMed Central, p. S17.
Potter, D.M., 2005. A permutation test for inference in logistic regression with small-and moderate-sized data sets. Stat. Med. 24 (5), 693–708.
R Core Team, 2017. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. URL. https://www.R-project.org/.
Robinson, L.F., Atlas, L.Y., Wager, T.D., 2015. Dynamic functional connectivity using state-based dynamic community structure: method and application to opioid analgesia. Neuroimage 108, 274–291.
Rubinov, M., Sporns, O., 2010. Complex network measures of brain connectivity: uses and interpretations. Neuroimage 52 (3), 1059–1069.
Schmidt, M.N., Morup, M., 2013. Nonparametric bayesian modeling of complex networks: An introduction. IEEE Signal Process. Mag. 30 (3), 110–128.
Scrucca, L., Raftery, A.E., 2015. Improved initialisation of model-based clustering using Gaussian hierarchical partitions. Advances in data analysis and classification 9 (4), 447–460.
Snijders, T., Nowicki, K., 1997. Estimation and prediction for stochastic blockmodels for graphs with latent block structure. J. Classif. 14 (1), 75–100.
Van Den Heuvel, M.P., Pol, H.E.H., 2010. Exploring the brain network: a review on resting-state fmri functional connectivity. Eur. Neuropsychopharmacol 20 (8), 519–534.
van den Heuvel, M.P., Sporns, O., 2013. Network hubs in the human brain. Trends Cognit. Sci. 17 (12), 683–696.
Westfall, P.H., Young, S.S., 1993. Resampling-based Multiple Testing: Examples and Methods for P-Value Adjustment, vol. 279. John Wiley & Sons.
Whitaker, K.J., Vértes, P.E., Romero-Garcia, R., Váša, F., Moutoussis, M., Prabhu, G., Weiskopf, N., Callaghan, M.F., Wagstyl, K., Rittman, T., et al., 2016. Adolescence is associated with genomically patterned consolidation of the hubs of the human brain connectome. Proc. Natl. Acad. Sci. Unit. States Am. 113 (32), 9105–9110.
Wolfe, P.J., Olhede, S.C., 2013. Nonparametric Graphon Estimation arXiv preprint arXiv: 1309.5936.
Wu, C.J., 1983. On the convergence properties of the em algorithm. Ann. Stat. 95–103.
Zanghi, H., Ambroise, C., Miele, V., 2008. Fast online graph clustering via erdős–rényi mixture. Pattern Recogn. 41 (12), 3592–3599.
Zanghi, H., Volant, S., Ambroise, C., 2010. Clustering based on random graph model embedding vertex features. Pattern Recogn. Lett. 31 (9), 830–836.