## REVIEW

# A review of stochastic block models and extensions for graph clustering

Clement Lee[1*] [ID] and Darren J. Wilkinson[2,3]

*Correspondence:
clement.lee@lancaster.ac.uk
[1]Department of Mathematics and
Statistics, Lancaster University,
Lancaster, United Kingdom
Full list of author information is
available at the end of the article

## Abstract

There have been rapid developments in model-based clustering of graphs, also known as block modelling, over the last ten years or so. We review different approaches and extensions proposed for different aspects in this area, such as the type of the graph, the clustering approach, the inference approach, and whether the number of groups is selected or estimated. We also review models that combine block modelling with topic modelling and/or longitudinal modelling, regarding how these models deal with multiple types of data. How different approaches cope with various issues will be summarised and compared, to facilitate the demand of practitioners for a concise overview of the current status of these areas of literature.

**Keywords:** Model-based clustering, Stochastic block models, Mixed membership models, Topic modelling, Longitudinal modelling, Statistical inference

## Introduction

Stochastic block models (SBMs) are an increasingly popular class of models in statistical analysis of graphs or networks. They can be used to discover or understand the (latent) structure of a network, as well as for clustering purposes. We introduce them by considering the example in Fig. 1, in which the network consists of 90 nodes and 1192 edges. The nodes are divided into 3 groups, with groups 1, 2 and 3 containing 25, 30 and 35 nodes, respectively. The nodes within the same group are more closely connected to each other, than with nodes in another group. Moreover, the connectivity pattern is rather "uniform". For example, compared to nodes 2 to 25, node 1 does not seem a lot more connected to other nodes, both within the same group or with another group. In fact, this model is generated by taking each pair of nodes at a time, and simulating an (undirected) edge between them. The probability of having such an edge or not is independent of that of any other pair of nodes. For two nodes in the same group, that is, of the same colour, the probability of an edge is 0.8, while for two nodes in different groups, the edge probability is 0.05.

The above example is a *simulation* from a simple SBM, in which there are two essential components, both of which will be explained in "Stochastic block models" section. The first component is the vector of group memberships, given by $(\underbrace{1\ 1 \cdots 1}_{25}\underbrace{2\ 2 \cdots 2}_{30}\underbrace{3\ 3 \cdots 3}_{35})^T$

in this example. The second is the block matrix, each element of which represents the edge probability of two nodes, *conditional on their group memberships.* According the description of the generation of edges above, the block matrix is

**Fig. 1** Network of 90 nodes
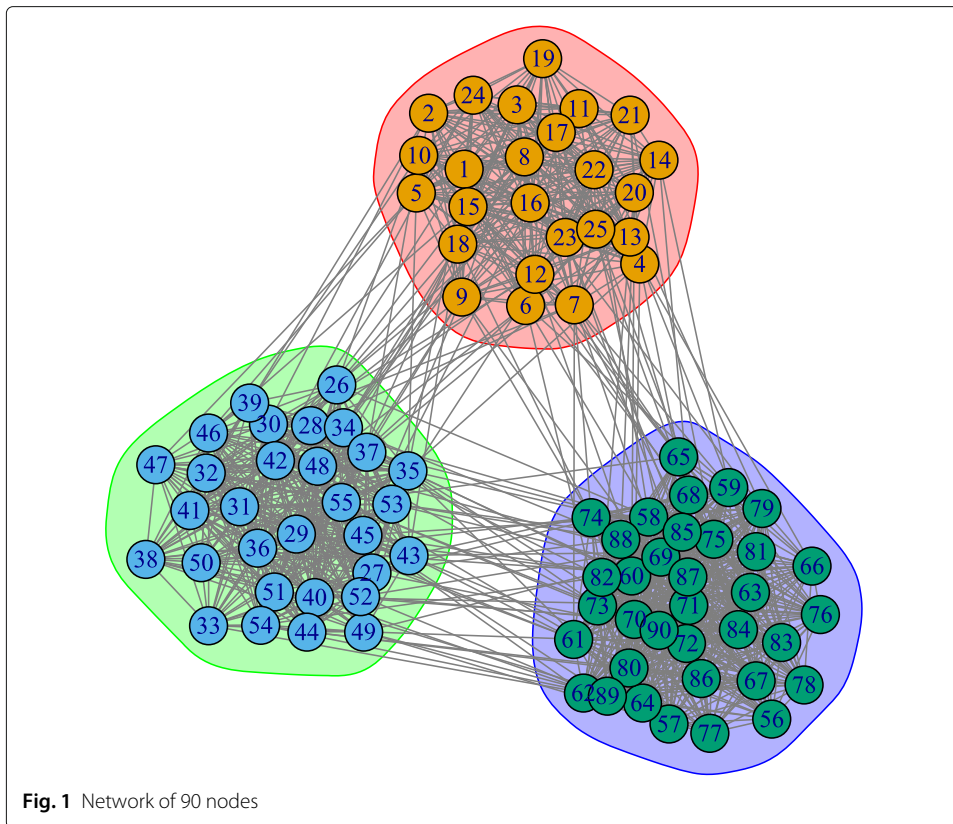
$$\begin{pmatrix} 0.8 & 0.05 & 0.05 \\ 0.05 & 0.8 & 0.05 \\ 0.05 & 0.05 & 0.8 \end{pmatrix}.$$ (1)

Given the group memberships, the block matrix, and the assumptions of an SBM (to be detailed in "Stochastic block models" section), it is straightforward to generate a synthetic network for *simulation* purposes, as has been done in the example. It will also be straightforward to evaluate the likelihood of data observed, for *modelling* purposes. However, in applications to real data, neither the group memberships nor the block matrix is observed or given. Therefore, the goal of fitting an SBM to a graph is to infer these two components simultaneously. Subsequently, the usual statistical challenges arise:

1.  Modelling: How should the SBM be structured or extended to realistically describe real-world networks, with or without additional information on the nodes or the edges?
2.  Inference: Once the likelihood can be computed, how should we infer the group memberships and the block matrix? Are there efficient and scalable inference algorithms?
3.  Selection and diagnostics: Can we compute measures, such as Bayesian information criterion (BIC) and marginal likelihood, to quantify and compare the goodness of fit of different SBMs?

Inferring the group memberships is essentially clustering the nodes into different groups, and the number of groups, denoted by $K$, is quite often unknown prior to modelling and inference. This brings about another challenge:

4. Should we incorporate $K$ as a parameter in the model, and infer it in the inference? Or should we fit an SBM with different fixed $K$'s, and view finding the optimal $K$ as a model selection problem?

In this article, we will review the developments of SBMs in the literature, and compare how different models deal with various issues related to the above questions. Such issues include the type of the graph, the clustering approach, the inference approach, and selecting or estimating the (optimal) number of groups. The issue with the number of groups $K$ is singled out because how differently it is related to questions 1–3 depends largely on the specific model reviewed.

Quite often, other types of data, such as textual and/or temporal, appear alongside network data. One famous example is the Enron Corpus, a large database of over 0.6 million emails by 158 employees of the Enron Corporation before the collapse of the company in 2001. While the employees and the email exchanges represent the nodes and the edges of the email network, respectively, the contents and creation times of the emails provide textual and dynamic information, respectively. It has been studied by numerous articles in the literature, including (Zhou et al. 2006; McCallum et al. 2007; Pathak et al. 2008; Fu et al. 2009; Xing et al. 2010; Gopalan et al. 2012; Sachan et al. 2012; DuBois et al. 2013; Xu and Hero III 2013; Matias et al. 2015; Bouveyron et al. 2016; Corneli et al. 2018). Another example is collections of academic articles, in which the network, temporal and textual data come from the references/citations between the articles, their publication years, and their actual contents, respectively. In these cases, further questions can be asked:

5. Can longitudinal and/or textual modelling, especially for cluster purposes, be incorporated into the SBMs, to utilise all information available in the data?
6. How are inference, model selection/diagnostics, and the issue with $K$ dealt with under the more complex models?

A relevant field to answering question 5 is topic modelling, in which the word frequencies of a collection of texts/articles are analysed, with the goal of clustering the articles into various topics. While topic modelling and SBMs are applicable to different types of information, textual for the former and relational for the latter, their ultimate goals are the same, which is model-based clustering of non-numerical data. Similar issues to the aforementioned ones for SBMs are also dealt with in various works, which will therefore be reviewed in this article.

While incorporating longitudinal and/or topic modelling into SBMs are possible, it should be noted that they are well-developed and large fields on their own. Due to the scope of this article, we will focus on works that are more recent or relatively straightforward extensions of SBMs to these two fields. We hope, from these reviewed interdisciplinary works, to provide future directions to a comprehensive model that handles multiple types of information simultaeously and deals with the questions raised satisfactorily.

While articles on SBMs form a major body of works in the literature, especially over the last decade or so, they should not be completely separated from other methods or

algorithms in statistical network analysis. For example, community detection methods and latent space models are two highly related fields to SBMs, and their developments are interwined with each other. Therefore, several important works, such as reviews of these two topics or articles which make connections with SBMs, will be mentioned as well, for the sake of comprehensiveness.

The rest of this article is organised as follows. A simple version of the SBM is introduced in "Stochastic block models" section. Extensions of the SBM regarding the type of graph are reviewed in "Type of graph and extensions of the SBM" section. Models that relax the usual clustering approach, in which each node is assumed a single group, are introduced in "Clustering approach" section. Related models and methods for graphs to the SBM are discussed in "Related methods for graphs" section. The inference approaches and the related issue of the number of groups are discussed in "Inference approach" and "Number of groups" sections, respectively. Models which incorporate longitudinal modelling are "SBM with longitudinal modelling" section. Topic modelling is briefly introduced in "Topic models" section, and its incorporation in SBMs is reviewed in "SBM with topic modelling" section. A summary and comparison of models are provided in "Comparison" section, and the discussion in "Discussion" section concludes the article.

## Stochastic block models

In this section, we shall first formulate a basic version of the stochastic block model (SBM) and mention the concept of stochastic equivalence, illustrated by continuing with the example in "Introduction" section. This will pave the way for "Type of graph and extensions of the SBM" section, where we consider different extensions to accommodate additional information about the graph, to better describe real-world networks, and to potentially lead to more scalable inference algorithms.

To introduce the terminology, we consider a graph $\mathcal{G} = (\mathcal{N}, \mathcal{E})$, where $\mathcal{N}$ is the node set of size $n := |\mathcal{N}|$, and $\mathcal{E}$ is the edge list of size $M := |\mathcal{E}|$. In the example in Fig. 1, $\mathcal{N} = \{1, 2, \ldots, 90\}$, $n = 90$, and $M = 1192$. We call a pair of nodes a *dyad*, and consider the existence or absence of an edge for the dyad $(p, q)$, through the use of the $n \times n$ adjacency matrix, denoted by $\mathbf{Y}$, which is another useful representation of the graph. If $\mathcal{G}$ is undirected, as is the example, $\mathbf{Y}_{pq} = \mathbf{Y}_{qp} = 1$ if $p$ and $q$ have an edge between them, 0 otherwise, where $\mathbf{M}_{rs}$ represents the $(r, s)$-th element of matrix $\mathbf{M}$. By construction, under an undirected graph, $\mathbf{Y}$ is symmetric along the major diagonal. If $\mathcal{G}$ is directed, $\mathbf{Y}_{pq} = 1$ (0) represents an edge (non-edge) from *p to q*, and is independent of $\mathbf{Y}_{qp}$, for $\mathbf{Y}$ need not be symmetric. We assume $\mathbf{Y}_{pp} = 0$, that is, no node has a *self-edge*, as this is quite often, but not always, assumed in the models reviewed. The adjacency matrix for the example is shown in Fig. 2, where black and white represent 1 and 0, respectively. Graphs with binary adjacency matrices are called binary graphs hereafter.

In the SBM, each node belongs to one of the $K(< n)$ groups, where $K = 3$ in the example. As the groups are unknown before modelling, for node $p = 1, 2, \ldots, n$ also defined is a $K$-vector $\mathbf{Z}_p$, all elements of which are 0, except exactly one that takes the value 1 and represents the group node $p$ belongs to. For instance, as nodes 1, 45 and 90 in the example belong to groups 1, 2 and 3, respectively, we have $\mathbf{Z}_1 = (1\ 0\ 0)^T$, $\mathbf{Z}_{45} = (0\ 1\ 0)^T$, and $\mathbf{Z}_{90} = (0\ 0\ 1)^T$. Also defined is an $n \times K$ matrix $\mathbf{Z} := (\mathbf{Z}_1\ \mathbf{Z}_2\ \cdots\ \mathbf{Z}_n)^T$, such that $\mathbf{Z}_{pi}$ is the $i$-th element of $\mathbf{Z}_p$.
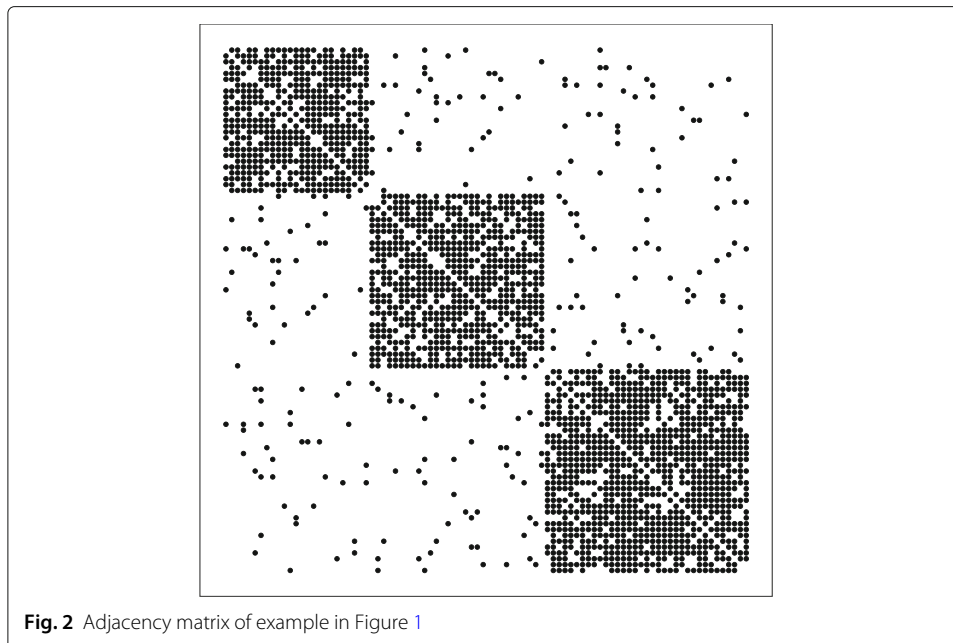
**Fig. 2** Adjacency matrix of example in Figure 1

The group sizes can be derived from $\mathbf{Z}$, and are denoted by $\mathbf{N} = (\mathbf{N}_1 \ \mathbf{N}_2 \ \cdots \ \mathbf{N}_K)^T$. Essentially, $\mathbf{N}_i$ is the sum, or number of non-zero elements, of the $i$-th column of $\mathbf{Z}$. In the example, $\mathbf{N}_1 = 25$, $\mathbf{N}_2 = 30$, $\mathbf{N}_3 = 35$. Finally, the $K \times K$ edge matrix between *groups* can be derived from $\mathbf{Z}$ and $\mathbf{Y}$. It is denoted by $\mathbf{E}$, where $\mathbf{E}_{ij}$ represents the number of edges between groups $i$ and $j$ in undirected graphs, and from group $i$ to group $j$ in directed ones. In the example, $\mathbf{E}$ is symmetric as $\mathcal{G}$ is undirected, and $\mathbf{E}_{11} = 245$, $\mathbf{E}_{22} = 341$, $\mathbf{E}_{33} = 481$, $\mathbf{E}_{12} = \mathbf{E}_{21} = 37$, $\mathbf{E}_{23} = \mathbf{E}_{32} = 52$, and $\mathbf{E}_{31} = \mathbf{E}_{13} = 36$.

In order to describe the generation of the edges of $\mathcal{G}$ according to the groups the nodes belong to, a $K \times K$ block matrix, denoted by $\mathbf{C}$, is introduced. If $\mathcal{G}$ is undirected, for $1 \leq i \leq j \leq K$, $\mathbf{C}_{ij} \in [0, 1]$ and represents the probability of occurrence of an edge between a node in group $i$ and a node in group $j$. Here $\mathbf{C}$ is symmetric, as is (1) in the example. If $\mathcal{G}$ is directed, for $1 \leq i, j \leq K$, $\mathbf{C}_{ij}$ represents the probability of occurrence of a directed edge from a node in group $i$ to a node in group $j$, and $\mathbf{C}$ needs not be symmetric. Note that no rows or columns in $\mathbf{C}$ need to sum to 1.

Whether $\mathcal{G}$ is undirected or directed, the idea of the block matrix $\mathbf{C}$ means that the dyads are conditionally independent given the group memberships $\mathbf{Z}$. In other words, $\mathbf{Y}_{pq}$ follows the Bernoulli distribution with success probability $\mathbf{Z}_p^T \mathbf{C} \mathbf{Z}_q$, and is independent of $\mathbf{Y}_{rs}$ for $(p, q) \neq (r, s)$, *given* $\mathbf{Z}_p$ *and* $\mathbf{Z}_q$. This implies that the total number of edges between any two blocks $i$ and $j$ is a Binomial distributed random variable with mean equal to the product of $\mathbf{C}_{ij}$ and the number of dyads available. For undirected and directed graphs, the latter term is $\mathbf{N}_i \mathbf{N}_j / 2$ and $\mathbf{N}_i \mathbf{N}_j$, respectively. In fact, Fig. 2 can be viewed as a realisation of simulating from the Binomial distribution with the respective means. Conversely, the densities of each pair of blocks in the adjacency matrix, calculated to be

$$\begin{pmatrix} 0.817 & 0.049 & 0.041 \\ 0.049 & 0.784 & 0.05 \\ 0.041 & 0.05 & 0.808 \end{pmatrix},$$

are, as expected, close to (1).

**Stochastic equivalence**

The assumption that the edge probability of a dyad depends solely on their memberships (and **C**) is based on the concept of *stochastic equivalence* (see, for example, Nowicki and Snijders (2001)). In less technical terms, for nodes $p$ and $q$ in the same group, $p$ has the same (and independent) probability of connecting with node $r$, as $q$ does. This echoes the observation in the example that node 1 does not seem more connected to the whole network than other nodes in the same group are. While such probability depends on the group membership of $r$, the equality still holds between $p$ and $q$.

*Block modelling and community detection*

The concept of stochastic equivalence in itself does not require that nodes in the same group are more connected within themselves, than with nodes in another group. Essentially, the elements along the major diagonal of **C** are not necessarily higher than the off-diagonal elements. However, this phenomenon, which is also observed in the example, is often the goal of community detection, a closely related topic to SBMs. Therefore, it is sometimes taken into account in the modelling in the *assortative* or *affinity* SBM by, for example, Gopalan et al. (2012) and Li et al. (2016). Community detection algorithms and assortative SBMs will be further discussed in "Community detection" and "Assortative SBM" sections, respectively.

**Modelling and likelihood**

Given **Z** and **C**, we can write down the likelihood based on the assumption of edges being Bernoulli distributed conditional on the group memberships. If $\mathcal{G}$ is undirected, again assuming no self-edges, the likelihood is

$$
\pi\left(\mathbf{Y}|\mathbf{Z}, \mathbf{C}\right) = \prod_{p<q}^{n} \pi(\mathbf{Y}_{pq}|\mathbf{Z}, \mathbf{C})
$$
$$
= \prod_{p<q}^{n} \left[ \left(\mathbf{Z}_p^T \mathbf{C} \mathbf{Z}_q\right)^{\mathbf{Y}_{pq}} \left(1 - \mathbf{Z}_p^T \mathbf{C} \mathbf{Z}_q\right)^{(1-\mathbf{Y}_{pq})} \right]. \tag{2}
$$

If $\mathcal{G}$ is directed, we replace the index in the product in (2) from $p < q$ to $p \neq q$. The model with this likelihood will be called the Bernoulli SBM hereafter. With a change of index, (2) can be written as

$$
\pi(\mathbf{Y}|\mathbf{Z}, \mathbf{C}) = \prod_{i \leq j}^{K} \mathbf{C}_{ij}^{\mathbf{E}_{ij}} (1 - \mathbf{C}_{ij})^{\mathbf{N}_{ij} - \mathbf{E}_{ij}}, \tag{3}
$$

where $\mathbf{N}_{ij} = \mathbf{N}_i \mathbf{N}_j / 2$ if $i \neq j$, $\mathbf{N}_{ij} = \mathbf{N}_i (\mathbf{N}_i - 1)/2$ if $i = j$. Multiplying over the group indices $i, j$ will become more useful than over the node indices $p, q$ in some cases.

As mentioned in "Introduction" section, when applying SBMs to real-world data, usually neither **Z** nor **C** is known, and has to be inferred. Therefore, assumptions have to be made before modelling and inference. For $p = 1, 2, \ldots, n$, we assume that the latent variable $\mathbf{Z}_p$ is independent of $\mathbf{Z}_q$ *apriori*. Also, we assume that $\Pr(\mathbf{Z}_{pi} = 1) = \theta_i$, where $\theta_i$ is the $i$-th element of the $K$-vector $\boldsymbol{\theta} = (\theta_1\, \theta_2\, \ldots\, \theta_K)^T$ such that $\sum_{i=1}^{K} \theta_i = 1$. Essentially, the latent group $\mathbf{Z}_p$ follows the multinomial distribution with probabilities $\boldsymbol{\theta}$, which means

$$\pi\left(\mathbf{Z}|\boldsymbol{\theta}\right) = \prod_{p=1}^{n} \mathbf{Z}_p^T \boldsymbol{\theta} = \prod_{p=1}^{n} \boldsymbol{\theta}^T \mathbf{Z}_p = \prod_{i=1}^{K} \theta_i^{\mathbf{N}_i}. \tag{4}$$

A further assumption can be made that $\boldsymbol{\theta}$ arises from the Dirichlet$((\alpha \mathbf{1}_K)$ distribution, of which the parameter $\alpha$ comes from a Gamma$(a, b)$ prior. This will be useful in "Poisson and degree-corrected SBMs" section. We shall defer actual inference to "Inference approach" section, and first examine the extensions and variants of SBM in the following section.

### Type of graph and extensions of the SBM

In this section, we briefly revisit the lineage of the SBM, discuss how it is extended for binary graphs, and introduce models for valued graphs, to answer question 1 in "Introduction" section. Special attention will be paid to two increasingly useful variants: the degree-corrected and the microcanonical SBMs.

SBMs originated from their deterministic counterparts. Breiger et al. (1975) illustrated an algorithm to essentially permute the rows and columns of the adjacency matrix $\mathbf{Y}$. The rearranged adjacency matrix contains some submatrices with zeros only, some others with at least some ones. The former and latter kinds of submatrices are summarised by 0 and 1, respectively, in what they called the "blockmodel", which can be viewed as the predecessor of the block matrix $\mathbf{C}$. White and Breiger (1976) followed this line but also calculated the densities of the blocks in some of their examples. The stochastic generalisation of the "blockmodel" was formalised by Holland et al. (1983). While Wang and Wong (1987) applied the SBM to real directed graphs, they assumed that the block structure is known a priori. Snijders and Nowicki (1997) and Nowicki and Snijders (2001) studied *a posteriori* blocking, meaning that the groups are initially unknown and to be inferred via proper statistical modelling, for 2 and an arbitrary number of groups, respectively. These lead to the Bernoulli SBM for binary graphs described in "Stochastic block models" section.

### Binary graphs

Apart from Snijders and Nowicki (1997), binary graphs have been studied by numerous models. In the absence of additional information, such as covariates or attributes associated with the edges or nodes, the focus of any developments has been on proposing alternative models, or optimising the number of groups $K$. For example, the mixed membership models in Airoldi et al. (2008); Fu et al. (2009); Xing et al. (2010); Fan et al. (2015) and Li et al. (2016) allowed each node to belong to multiple groups ("Clustering approach" section). Miller et al. (2009) and Mørup et al. (2011) proposed latent feature models ("Latent feature models" section) that deviate from the SBM, while Zhou (2015) proposed an edge partition model ("Related methods for graphs" section). Mørup et al. (2011) and Fan et al. (2015) also, alongside Kim et al. (2013) and McDaid et al. (2013), *modelled K* in different ways ("Modelling" section), whereas (Latouche et al. 2012; Côme and Latouche 2015) and (Yan 2016) derived various criteria for *selecting K* ("Criteria and model selection" section). Lee and Wilkinson (2018) proposed a model for binary directed *acyclic* graphs (DAGs), in which the possible combinations for $(\mathbf{Y}_{pq}, \mathbf{Y}_{qp})$ are $\{(0, 0), (0, 1), (1, 0)\}$, and no directed cycles of any length are allowed. This means

that if we go from a node to an arbitrary neighbour along the direction of the edges and perform this "random walk" recursively, it is not possible to reach the original node.

Quite often temporal data are avaialble to the networks observed. Therefore, there are numerous models for binary graphs that incoporate longitudinal modelling, including Fu et al. (2009); Xing et al. (2010) Yang et al. (2011); Xu and Hero III (2013); Tang and Yang (2011); Tang and Yang (2014); Fan et al. (2015); Matias et al. (2015), and Ludkin et al. (2018). Therefore, they will be discussed separately in "SBM with longitudinal modelling" section.

While longitudinal SBMs deal with multiple layers of graphs (in a temporal order) directly, there are cases in which the multiple layers of graphs are aggregated, and the observed graph contains one layer only, with information about the specific layers lost. Vallès-Català et al. (2016) proposed a multilayer SBM for such a situation. Using the example of 2 layers, they assumed the adjacency matrix $\mathbf{Y}^l$ of layer $l = 1, 2$ is generated by a Bernoulli SBM according to block matrix $\mathbf{C}^l$ independently, and then considered two versions of the model. In the first version, there is an edge between $p$ and $q$ in the graph to be observed if an edge for the dyad exists in *either* layer, which means $\mathbf{Y}_{pq} = 1 - (1 - \mathbf{Y}_{pq}^1)(1 - \mathbf{Y}_{pq}^2)$. In the second, the existence of the edge in the observed graph requires the existence of the corresponding edge in both layers, which menas $\mathbf{Y}_{pq} = \mathbf{Y}_{pq}^1 \mathbf{Y}_{pq}^2$. Vallès-Català et al. (2016) found that the second version is more predictive (than the usual single-layer SBM) when it comes to link prediction or network reconstruction ("Missingness and errors" section) for real-world networks, which may therefore be better described as an aggregation of multiple layers.

### Valued graphs

The Bernoulli SBM can be easily generalised or modified for valued graphs. Apart from the aforementioned undirected and directed graphs, Nowicki and Snijders (2001) also studied directed signed graphs, where $\mathbf{Y}_{pq}$ can take the value $-1, 0$, or $1$. If we consider the dyad of a node of group $i$ and a node of group $j$, and further define two $K \times K$ matrices $\mathbf{D}$ and $\mathbf{E}$ such that $\mathbf{D}_{ij}$ and $\mathbf{E}_{ij}$ represent the probabilities that the dyad takes the value $0$ and $-1$, respectively, subject to $\mathbf{C}_{ij} + \mathbf{D}_{ij} + \mathbf{E}_{ij} = 1$, the edge probability, with a slight abuse of notation, is

$$\Pr(\mathbf{Y}_{pq} = 1 | \mathbf{Z}, \mathbf{C}, \mathbf{D}, \mathbf{E}) = \mathbf{Z}_p^T \left( \mathbf{C}^{\mathbf{1}\{Y_{pq}=1\}} \mathbf{D}^{\mathbf{1}\{Y_{pq}=0\}} \mathbf{E}^{\mathbf{1}\{Y_{pq}=-1\}} \right) \mathbf{Z}_q,$$

where $\mathbf{1}\{A\}$ is the indicator function of event $A$. They also studied graphs for tournaments, which means the edge between nodes $p$ and $q$ can represent a match between the two nodes. The possible combinations for $(\mathbf{Y}_{pq}, \mathbf{Y}_{qp})$ are $\{(-1, 1), (1, -1), (0, 0)\}$, corresponding to a loss, win and tie for $p$, respectively. The specification of the edge probability is omitted here because of the higher *notational* complexity. There is limited adoption of models in the literature for this kind of graphs.

Kurihara et al. (2006) considered a graph in which multiple edges between two nodes can be accounted for. Instead of taking the product over all dyads, they assumed each edge arises from the usual Bernoulli($\mathbf{Z}_p^T \mathbf{C} \mathbf{Z}_q$) distribution *after* drawing the two nodes $p$ and $q$ independently from a multinomial distribution with weights $\boldsymbol{\phi} = (\phi_1 \, \phi_2 \cdots \phi_n)^T$. If we assume the $m$-th edge corresponds to the dyad $(p_m, q_m)$, the likelihood is

$$\pi(\mathcal{E} | \boldsymbol{\phi}, \mathbf{Z}, \mathbf{C}) = \prod_{m=1}^{M} \left[ \phi_{p_m} \phi_{q_m} \left( \mathbf{Z}_{p_m}^T \mathbf{C} \mathbf{Z}_{q_m} \right)^{\mathbf{Y}_{p_m q_m}} \left( 1 - \mathbf{Z}_{p_m}^T \mathbf{C} \mathbf{Z}_{q_m} \right)^{1 - \mathbf{Y}_{p_m q_m}} \right].$$

While Kurihara et al. (2006) mainly considered bipartite graphs, we consider the two types of nodes as essentially the same set of nodes, to align with the notation in other models.

Yang et al. (2011) mainly worked with binary undirected graphs in their dynamic SBM (see "SBM with longitudinal modelling" section), but briefly extended to the valued version, where $\mathbf{Y}_{pq}$, now the discrete number of interactions for dyad $(p, q)$, is being modelled by a geometric distribution:

$$\pi\left(\mathbf{Y}_{pq} | \mathbf{Z}, \mathbf{C}\right) = \left(\mathbf{Z}_p^T \mathbf{C} \mathbf{Z}_q\right)^{\mathbf{Y}_{pq}} \left(1 - \mathbf{Z}_p^T \mathbf{C} \mathbf{Z}_q\right). \quad (5)$$

DuBois et al. (2013) also proposed a model based on the SBMs for network dynamics. In its simplest version, for the dyad $(p, q)$, the interactions are being modelled by a Poisson process with intensity $\exp\left(\mathbf{Z}_p^T \mathbf{C} \mathbf{Z}_q\right)$, where the elements of $\mathbf{C}$ are not bounded by 0 and 1 here. Further modifications of the model are discussed in "SBM with longitudinal modelling" section.

### Poisson and degree-corrected SBMs

Karrer and Newman (2011) also worked with (undirected) valued graphs, but arguably in a more natural way. They first redefined $\mathbf{Y}_{pq}$ to be the number of edges for the dyad $(p, q)$ following a Poisson distribution, and $\mathbf{C}_{ij}$ the *expected* number of edges from a node in group $i$ to a node in group $j$. The density of $\mathbf{Y}_{pq}$ is now

$$\pi\left(\mathbf{Y}_{pq} | \mathbf{Z}, \mathbf{C}\right) = \left(\mathbf{Y}_{pq}!\right)^{-1} \exp\left(-\mathbf{Z}_p^T \mathbf{C} \mathbf{Z}_q\right) \left(\mathbf{Z}_p^T \mathbf{C} \mathbf{Z}_q\right)^{\mathbf{Y}_{pq}}. \quad (6)$$

They argued that, in the limit of a large sparse graph where the edge probability equals the expected number of edges, this version of the SBM, called the Poisson SBM, is asymptotically equivalent to the Bernoulli counterpart in (2). To further modify the model, a parameter $\phi_p$ is introduced for each node, subject to a constraint $\sum_{p=1}^n \phi_p \mathbf{1}\{\mathbf{Z}_{pi} = 1\} = 1$ for every group $i$, so that the expected number of edges for the dyad $(p, q)$ is now $\phi_p \phi_q \mathbf{Z}_p^T \mathbf{C} \mathbf{Z}_q$. The density of $\mathbf{Y}_{pq}$ becomes

$$\pi\left(\mathbf{Y}_{pq} | \mathbf{Z}, \mathbf{C}, \boldsymbol{\phi}\right) = \left(\mathbf{Y}_{pq}!\right)^{-1} \exp\left(-\phi_p \phi_q \mathbf{Z}_p^T \mathbf{C} \mathbf{Z}_q\right) \left(\phi_p \phi_q \mathbf{Z}_p^T \mathbf{C} \mathbf{Z}_q\right)^{\mathbf{Y}_{pq}}, \quad (7)$$

where $\boldsymbol{\phi} = (\phi_1 \ \phi_2 \ \cdots \ \phi_n)^T$. This is termed the degree-corrected (DC) SBM. The parameters $\phi_p$ and $\mathbf{C}_{ij}$ have natural interpretations as their maximum likelihood estimates (MLEs) are the ratio of $p$'s degree to the sum of degrees in $p$'s group, and the total number of edges between groups $i$ and $j$, respectively. While Karrer and Newman (2011) have also considered self-edges in their model, such treatment is omitted here for easier notational alignment.

Of importance is the reason behind the DC-SBM. Karrer and Newman (2011) argued that the then existing SBMs usually ignores the variation in the node degrees in real-world networks. This is quite evident as, under the original SBM, the expected degree is the same for all nodes in each group, given $\mathbf{Z}$ and $\mathbf{C}$. Karrer and Newman (2011) illustrated that the DC-SBM managed to discover the known factions in a karate club network, while the original counterpart failed to do so. Quite a few recent works built upon the DC-SBM in various ways. For instance, Yan et al. (2014) provided an approach to model selection ("Criteria and model selection" section) between the two versions of SBMs. Also, Lu and Szymanski (2019b) introduced additional parameters to $\boldsymbol{\phi}$ as a means of taking the *assortativeness* into account; see "Assortative SBM" section.

McDaid et al. (2013) managed to integrate out the model parameters of the Poisson SBM in (6), to form a *collapsed* SBM. Apart from (2), (4) and the Dirichlet distribution for $\boldsymbol{\theta}$, they also assumed *apriori* each $\mathbf{C}_{ij}$ is independent and identically distributed (i.i.d.) according to a Gamma$(\gamma, \beta)$ distribution (with mean $\gamma/\beta$). With a change of indices similar to (3), the likelihood can be written as

$$
\pi(\mathbf{Y}|\mathbf{Z}, \mathbf{C}) = \prod_{p<q}^{n} \left(\mathbf{Y}_{pq}!\right)^{-1} \exp\left(-\mathbf{Z}_p^T \mathbf{C} \mathbf{Z}_q\right) \left(\mathbf{Z}_p^T \mathbf{C} \mathbf{Z}_q\right)^{\mathbf{Y}_{pq}}
$$

$$
= \prod_{i \leq j}^{K} \exp\left(-\mathbf{C}_{ij} \mathbf{N}_i \mathbf{N}_j\right) \mathbf{C}_{ij}^{\mathbf{E}_{ij}} \times \prod_{p<q}^{n} (\mathbf{Y}_{pq}!)^{-1} \tag{8}
$$

Now, the joint density of $\mathbf{Y}$ and $\mathbf{Z}$, with $\mathbf{C}$ and $\boldsymbol{\theta}$ integrated out, can be derived:

$$
\pi(\mathbf{Y}, \mathbf{Z}|\alpha, \beta, \gamma) = \pi(\mathbf{Y}|\mathbf{Z}, \beta, \gamma) \times \pi(\mathbf{Z}|\alpha)
$$

$$
= \int \pi(\mathbf{Y}|\mathbf{Z}, \mathbf{C})\pi(\mathbf{C}|\beta, \gamma)d\mathbf{C} \times \int \pi(\mathbf{Z}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\alpha)d\boldsymbol{\theta}
$$

$$
= \prod_{p<q}^{n}(\mathbf{Y}_{pq}!)^{-1} \times \int \left[\prod_{i \leq j}^{K} e^{-\mathbf{C}_{ij}\mathbf{N}_i\mathbf{N}_j}\mathbf{C}_{ij}^{\mathbf{E}_{ij}} \times \prod_{i \leq j}^{K} e^{-\beta\mathbf{C}_{ij}}\mathbf{C}_{ij}^{\gamma-1}\beta^\gamma / \Gamma(\gamma)\right]d\mathbf{C}
$$

$$
\times \int \prod_{i=1}^{K} \theta_i^{\mathbf{N}_i} \times \left[\Gamma(K\alpha)\, \mathbf{1}\{\sum_{i=1}^{K}\theta_i = 1\}\prod_{i=1}^{K}\frac{\theta_i^{\alpha-1}}{\Gamma(\alpha)}\right]d\boldsymbol{\theta}
$$

$$
= \prod_{p<q}^{n}(\mathbf{Y}_{pq}!)^{-1} \times \prod_{i \leq j}^{K}\frac{\Gamma(\gamma+\mathbf{E}_{ij})\beta^\gamma}{\Gamma(\gamma)\left(\beta+\mathbf{N}_i\mathbf{N}_j\right)^{\gamma+\mathbf{E}_{ij}}} \times \frac{\Gamma(K\alpha)\prod_{i=1}^{K}\Gamma(\alpha+\mathbf{N}_i)}{\Gamma(\alpha)^K\Gamma(K\alpha+n)} \tag{9}
$$

Now (9) is a function of $\mathbf{Y}$, $\mathbf{Z}$ (through $\mathbf{E}$ and $\mathbf{N}$), and the three scalar (hyper-)parameters only. This becomes particularly useful in inference ("Collapsed SBMs and efficient moves" section). Also, (9) can be compared with (11) in the microcanonical SBM below. Furthermore, if $\alpha$, $\beta$ and $\gamma$ are fixed or integrated out, $\pi(\mathbf{Y}, \mathbf{Z})$ is the exponential of the *integrated complete data log-likelihood* (ICL). It is a useful quantity when it comes to inference ("Variational methods" section) and model selection ("Criteria and model selection" section).

Aicher et al. (2015) introduced a unifying framework for modelling binary and valued graphs, by observing that both the Bernoulli and Poisson distributions belong to the exponential family of distributions. For example, (2) can be written as

$$
\pi(\mathbf{Y}|\mathbf{Z}, \mathbf{C}) = \prod_{p<q}^{n}\left[\left(\mathbf{Z}_p^T\mathbf{C}\mathbf{Z}_q\right)^{\mathbf{Y}_{pq}}\left(1 - \mathbf{Z}_p^T\mathbf{C}\mathbf{Z}_q\right)^{(1-\mathbf{Y}_{pq})}\right]
$$

$$
= \exp\left\{\sum_{p<q}^{n}\left[\mathbf{Y}_{pq}\log\left(\frac{\mathbf{Z}_p^T\mathbf{C}\mathbf{Z}_q}{1-\mathbf{Z}_p^T\mathbf{C}\mathbf{Z}_q}\right) + \log\left(1 - \mathbf{Z}_p^T\mathbf{C}\mathbf{Z}_q\right)\right]\right\}
$$

$$
= \exp\left\{\sum_{p<q}^{n}\boldsymbol{S}(\mathbf{Y}_{pq})^T\boldsymbol{\eta}(\mathbf{Z}_p^T\mathbf{C}\mathbf{Z}_q)\right\}, \tag{10}
$$

where $\boldsymbol{S}(x) = (x, 1)^T$ and $\boldsymbol{\eta} = \left(\log\left(x/(1-x)\right), \log\left(1-x\right)\right)^T$ are both vector-valued functions, of the sufficient statistics and the natural parameters of the Bernoulli distribution,

respectively. Now, with a different type of edges observed, $S(x)$ and $\eta(x)$ can be specified according to the appropriate distribution in the exponential family. In the case of the Poisson SBM (6), $S(x) = (x, -\log(x!), -1)^T$ and $\eta(x) = (\log x, 1, x)^T$.

Aicher et al. (2015) also used this *weighted* SBM to clarify the meaning of zeros in valued graphs, as they could mean a non-edge, an edge with weight zero, or missing data. To overcome this ambiguity, they extended (10), so that the (log-)likelihood is a mixture of distributions:

$$\pi(\mathbf{Y}|\mathbf{Z}, \mathbf{C}) = \exp\left\{ \psi \sum_{p<q}^{n} S_1(\mathbf{Y}_{pq})^T \eta_1(\mathbf{Z}_p^T \mathbf{C}\mathbf{Z}_q) + (1-\psi) \sum_{p<q}^{n} S_2(\mathbf{Y}_{pq})^T \eta_2(\mathbf{Z}_p^T \mathbf{C}\mathbf{Z}_q) \right\},$$

where $\psi$ is the weight assigned to the two components. They can correspond to, for example, the Bernoulli and the Poisson SBM, for modelling the edge *existence* and edge *weight*, respectively.

#### Microcanonical SBM

Arguably the most important recent developments is the microcanonical SBM and its nested version (Peixoto 2017a; Peixoto 2017b), as these works are a culmination of applying the principle of minimum description length (MDL) (Grünwald 2007; Rosvall and Bergstrom 2007), a fundamental result regarding $K$ that is related to community detection (Peixoto 2013) ("Community detection" section), an efficient MCMC inference algorithm (Peixoto 2014a), a hierarchical structure that models $K$ simultaneously (Peixoto 2014b) ("Modelling" section), and an approach that models $\mathbf{Z}$ and $\mathbf{C}$ differently from (4), leading to efficient inference algorithm ("Monte Carlo methods" section). The microcanonical SBM is mentioned here because it can be derived from modifying the Poisson SBM. The case for undirected graphs is illustrated here, with (8) utilised again. Next, for $1 \le i \le j \le K$ and *conditional on* $\mathbf{Z}$ *and an extra parameter* $\lambda$, $\mathbf{C}_{ij}$ is assumed to follow the Exponential distribution with rate $\mathbf{N}_i\mathbf{N}_j/\lambda$. This assumption replaces that according to (4). By doing so, $\mathbf{C}$ can be integrated out from the product of (8) and the exponential density of $\mathbf{C}$:

$$\pi(\mathbf{Y}|\mathbf{Z}, \lambda) = \int \pi(\mathbf{Y}|\mathbf{Z}, \mathbf{C})\pi(\mathbf{C}|\mathbf{Z}, \lambda)d\mathbf{C}$$

$$= \prod_{p<q}^{n}(\mathbf{Y}_{pq}!)^{-1} \times \int \left[ \prod_{i\le j}^{K} e^{-\mathbf{C}_{ij}\mathbf{N}_i\mathbf{N}_j}\mathbf{C}_{ij}^{\mathbf{E}_{ij}} \times \prod_{i\le j}^{K} e^{-\mathbf{C}_{ij}\mathbf{N}_i\mathbf{N}_j/\lambda}\left(\mathbf{N}_i\mathbf{N}_j/\lambda\right) \right] d\mathbf{C}$$

$$= \prod_{p<q}^{n}(\mathbf{Y}_{pq}!)^{-1} \times \prod_{i\le j}^{K}\left(\mathbf{N}_i\mathbf{N}_j/\lambda\right) \times \prod_{i\le j}^{K}\frac{\Gamma(\mathbf{E}_{ij}+1)}{\left[\mathbf{N}_i\mathbf{N}_j(1+1/\lambda)\right]^{\mathbf{E}_{ij}+1}}$$

$$= \prod_{p<q}^{n}(\mathbf{Y}_{pq}!)^{-1} \times \prod_{i\le j}^{K}\frac{\lambda^{\mathbf{E}_{ij}}}{(1+\lambda)^{\mathbf{E}_{ij}+1}} \times \prod_{i\le j}^{K}\frac{\mathbf{E}_{ij}!}{(\mathbf{N}_i\mathbf{N}_j)^{\mathbf{E}_{ij}}}$$

$$= \underbrace{\frac{\lambda^M}{(1+\lambda)^{M+K(K+1)/2}}}_{\pi(\mathbf{E}|\mathbf{Z},\lambda)} \times \underbrace{\prod_{p<q}^{n}(\mathbf{Y}_{pq}!)^{-1} \times \prod_{i\le j}^{K}\frac{\mathbf{E}_{ij}!}{(\mathbf{N}_i\mathbf{N}_j)^{\mathbf{E}_{ij}}}}_{\pi(\mathbf{Y}|\mathbf{E},\mathbf{Z},\lambda)}, \tag{11}$$

where $M$, as defined in "Stochastic block models" section, is the total number of edges. Now, $\mathbf{Z}$ influences (11), or what we call the *integrated* likelihood, through $\mathbf{E}$ and $\mathbf{N}$ only. (We refrain from calling it the *marginal* likelihood, which we refer to the likelihood with

$\mathbf{Z}$ also integrated out.) Furthermore, this $\pi(\mathbf{Y}|\mathbf{Z}, \lambda)$ can be split into the product of the two underbraced terms, which is the joint likelihood of a microcanonical model (Peixoto 2017a). It is termed "microcanonical" because of the hard constraints imposed, as $\mathbf{Y}$ and $\mathbf{Z}$ together fix the value of $\mathbf{E}$, and therefore

$$\pi(\mathbf{Y}|\mathbf{Z}, \lambda) = \pi(\mathbf{Y}, \mathbf{E}|\mathbf{Z}, \lambda) = \pi(\mathbf{Y}|\mathbf{E}, \mathbf{Z}, \lambda) \times \pi(\mathbf{E}|\mathbf{Z}, \lambda).$$

Further marginalisation of $\lambda$, which is straightforward with one-dimensional integration of $\pi(\mathbf{E}|\mathbf{Z}, \lambda)$ in (11), results in

$$\pi(\mathbf{Y}|\mathbf{Z}) = \pi(\mathbf{Y}|\mathbf{E}, \mathbf{Z}) \times \pi(\mathbf{E}|\mathbf{Z}), \tag{12}$$

where no model parameters are involved, which however, as argued by Peixoto (2017a), is not compulsory in the microcanonical formulation. It is also called a nonparametric SBM, not because of having no parameters but because that $K$ is being modelled ("Modelling" section). Finally, the same marginalisation of $\mathbf{C}$ (and $\lambda$) can be applied to (7) to arrive at the microcanonical DC-SBM.

### Graphs with covariates or attributes

Tallberg (2005) proposed a model in which the group memberships $\mathbf{Z}$ depend on the covariates $\boldsymbol{x}$, through what they called a random utility model. Specifically, assume that $\boldsymbol{x}$ is $d$-dimensional, and associated with each group $i$ is a $d$-vector $\boldsymbol{\beta}_i$. The group membership of node $p$ is determind by $\mathbf{Z}_p = \underset{i}{\operatorname{argmax}} \left( \boldsymbol{x}_p^T \boldsymbol{\beta}_i + \epsilon_{pi} \right)$, where $\epsilon_{pi}$ is an i.i.d. Gaussian-distruted error. In this way, the covariates $\boldsymbol{x}_p$ determine the memberships $\mathbf{Z}_p$ through the group-specific vectors $\{\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \cdots, \boldsymbol{\beta}_K\}$ (and the error terms).

Vu et al. (2013) also studied binary graphs as well as directed signed graphs. Furthermore, they proposed a model that connected the SBMs with exponential random graph models (ERGMs), another prominent class of social network analysis models which can be traced back to Holland (1981). In one example of the model by Vu et al. (2013), the edge probability is

$$\pi\left(\mathbf{Y}_{pq}|\mathbf{Z}, \mathbf{C}, \boldsymbol{x}\right) \propto \exp\left[\psi f(\boldsymbol{x}, \mathbf{Y}_{pq}) + \left(\mathbf{Z}_p^T \mathbf{C} \mathbf{Z}_q\right) g(\boldsymbol{x}, \mathbf{Y}_{pq})\right], \tag{13}$$

where $\boldsymbol{x}$ are the covariates, and $f$ and $g$ are functions that may depend on $\boldsymbol{x}$. We do not specify the index of $\boldsymbol{x}$ because the covariates may depend on the nodes or the dyads. The parameter $\psi$ is constant to both $\boldsymbol{x}$ and the groups $p$ and $q$ belong to, while the use of $\mathbf{C}$, each element of which is possibly vector-valued and not bounded by 0 and 1, is to illustrate the dependence on the blocks and alignment with other models.

Peixoto (2018a) extended the microcanonical SBM by incorporating attributes $\boldsymbol{x}$. Specifically, the joint density of $\mathbf{Y}$ and $\boldsymbol{x}$ is

$$\pi\left(\mathbf{Y}, \boldsymbol{x}|\mathbf{Z}\right) = \pi\left(\boldsymbol{x}|\mathbf{Y}, \mathbf{Z}\right) \pi\left(\mathbf{Y}|\mathbf{Z}\right), \tag{14}$$

where $\pi(\mathbf{Y}|\mathbf{Z})$ is given by (12), and the model is termed the nonparametric weighted SBM. Again, a hierarchical or nested structure can be incorporated, so that the *groups* are modelled by another SBM, and so on, if necessary. Please see "Number of groups" section for further details.

Stanley et al. (2019) proposed an *attribute* SBM in which the group memberships $\mathbf{Z}$ of the nodes determine both the graph $\mathbf{Y}$ and the non-relational attributes $\boldsymbol{x}$, which are assumed conditionally independent given $\mathbf{Z}$. In the model formulation, in addition to a

Bernoulli SBM for an undirected graph, they assumed that the attribute $x_p$ for node $p$ comes from a mixture of Gaussian distributions with weights $\theta$, which are the probabilities for generating $Z_p$ as defined in "Modelling and likelihood" section. By utilising both $x$ and $Y$ together in inferring $Z$, Stanley et al. (2019) found that the two types of information can complement each other, and therefore their *attribute* SBM is useful for link prediction ("Missingness and errors" section).

## Clustering approach

Most of the models introduced so far adopt a hard clustering approach, that is, each node belongs to one group. However, for real-world networks, it is not unreasonable to allow a node to belong to multiple groups. In this section, we will look at models that do so, by incoporating a *soft clustering* approach in an SBM. Care has to be taken regarding how nodes that can belong to more than one group interact to form edges.

### Mixed membership SBM

In the mixed membership stochastic block model (MMSBM) by Airoldi et al. (2008), for each node $p$, the latent variable $Z_p$, which contains exactly one 1, is replaced by a membership vector, also of length $K$, denoted by $\theta_p$. The elements of $\theta_p$, which represent weights or probabilities in the groups, have to be non-negative and sum to 1. Using the example in Fig. 1, $Z_1 = (1\ 0\ 0)^T$ could be replaced by $\theta_1 = (0.7\ 0.2\ 0.1)^T$, which roughly means that, on average, node 1 spends 70%, 20% and 10% of the time in groups 1, 2 and 3, respectively. Next, each node can belong to different groups when interacting with different nodes. In order to do so, still assuming that $\mathcal{G}$ is undirected, for each dyad $(p, q)$, a latent variable $Z_{pq}$ is drawn from the multinomial distribution with probabilities $\theta_p$. As a $K$-vector containing exactly one 1, $Z_{pq}$ now represents the group node $p$ is in *when interacting with $q$*. (Also drawn is $Z_{qp}$ from $\theta_q$ to represent the group node $q$ is in when interacting with $p$.) Going back to the example, if $Z_{12} = (0\ 0\ 1)^T$ and $Z_{13} = (1\ 0\ 0)^T$, which are drawn independently from $\theta_1$, node 1 belongs to groups 3 and 1, respectively, when $Y_{12}$ and $Y_{13}$ are concerned. As each $Z_{pq}$ is a $K$-vector, the collection of latent variables $Z$ is now an $n \times n \times K$ array. The *apriori* density of $Z$ now becomes

$$\pi (Z|\Theta) = \prod_{p \neq q}^{n} \left( Z_{pq}^T \theta_p \times Z_{qp}^T \theta_q \right),$$

where $\Theta := (\theta_1\ \theta_2\ \cdots\ \theta_n)^T$ is the $n \times K$ matrix of membership probabilities such that $\Theta_{pi}$ is the $i$-th element of $\theta_p$. Comparing with (2), the likelihood is

$$\pi (Y|Z, C) = \prod_{p \leq q}^{n} \left[ \left( Z_{pq}^T C Z_{qp} \right)^{Y_{pq}} \left( 1 - Z_{pq}^T C Z_{qp} \right)^{(1-Y_{pq})} \right]. \tag{15}$$

We can carry out (Bayesian) inference once we specify the prior distributions. However, we shall defer this to "Inference approach" section. For the derivations of the model for directed graphs, please see Airoldi et al. (2008). What should be noted here is that the main goal of inference is not for the pairwise latent variables $Z$, but the mixed memberships $\Theta$.

Several articles built upon the MMSBM introduced. Li et al. (2016) proposed a scalable algorithm ("Inference approach" section), Fu et al. (2009) and Xing et al. (2010) incorporated longitudinal modelling ("SBM with longitudinal modelling" section), and Kim et al.

(2013) and Fan et al. (2015) focused on modelling $K$, the number of groups ("Modelling" section). Fan et al. (2016) observed that the assumption that $\mathbf{Z}_{pq}$ and $\mathbf{Z}_{qp}$ are independent is not quite realistic in real-world networks, as nodes may have higher correlated interactions towards the ones within the same groups. Therefore, they proposed a copula MMSBM for modelling these intra-group correlations.

Godoy-Lorite et al. (2016) modified the MMSBM for recommender systems, in which the observed data is the *ratings* some users give to some items (such as books or movies), and the goal of modelling and inference is to predicting user preferences. The ratings are therefore treated as the observed edges in a bipartite graph, but it is not the *existence* of the these edges that is being modelled. Rather, it is the value of the ratings, that is, the edge weight, that depends on the depends on the respective mixed memberships of the users and items. By inferring these memberships as well as the block matrix, predictions on user preferences can be made *for unobserved combinations of users and items.*

The MMSBM is related to, or has been compared with other models for graphs. For example, the latent feature model (Miller et al. 2009) deviated from the hard clustering SBM in a different way than MMSBM did, and therefore made comparison with the latter in terms of performance. For the class of latent feature models, and their practical difference with MMSBM, please see "Latent feature models" section. A close connection with the latent space models (Hoff et al. 2002; Handcock et al. 2007) have been drawn by Airoldi et al. (2008); please see "Latent space models" section.

### Overlapping SBM

When applying MMSBMs, while some nodes might have genuine mixed memberships, some other nodes might have single memberships, that is, $\boldsymbol{\theta}_p$ being a vector with all 0's but one 1. Such phenomenon is being addressed in another group of models, called *overlapping* SBMs, in a more direct way. For example, if there are $K = 2$ groups in the network, there will be nodes belonging to group 1 only, some others belonging to group 2 only, and the rest belonging to both groups simultaneously. When $K > 2$, overlapping between more than two groups is allowed. Unlike the hard clustering SBMs, the groups are not disjoint anymore in these overlapping models, which therefore are an alternative to the MMSBMs, as far as soft clustering is concerned. Theoretically, there are $2^K - 1$ choices of membership combinations for each node, as, for each node, there is a binary choice for belonging to each of the $K$ groups, with the only constraint that it has to have at least one grou membership.

The difference with the MMSBM is noted by Peixoto (2015b), who used the MDL approach (Peixoto 2017a) ("Microcanonical SBM" section) for overlapping models. They first considered a variation of the Bernoulli (or Poisson) SBM, in which the memberships are relaxed in the way described in the paragraph above. They then observed that, for sparse graphs, such an overlapping model can be approximated by a non-overlapping model for an augmented graph. Each distinct membership of a single node in the original graph can be considered as a different node with a single membership in the augmented graph. Modelling and inference ("Inference approach" section) are then straightforward. The expected degree for a node $p$ (in the original graph) with membership in multiple groups will be larger than that for nodes in either group, as $p$ received edges associated with each of the groups independently. Contrastly, in MMSBM, such quantity will be the weighted average of the corresponding quantities of the groups $p$ belongs to. A

degree-corrected version which incorporates the DC-SBM (Karrer and Newman 2011) is also derived, in which the soft clustering is achieved through hard clustering of the half-edges, rather than the hard clustering of the augmented graph in the non-degree-corrected version.

The increased number of membership choices (from $K$ to $2^K - 1$) naturally brings about the increased complexity of the overlapping SBM. However, such complexity is usually not favoured in applications to real-world networks. By comparing the MDL of the non-overlapping and overlapping SBMs, Peixoto (2015b) managed to carry out model selection, and found that the latter is more likely to overfit, and is selected as the better model only in a few cases. This finding is echoed by Xie et al. (2013) in the context of *community detection algorithms* ("Community detection" section). They observed through a comparative study that, in real-world networks, each overlapping node typically belongs to 2 or 3 groups. Furthermore, the proportion of overlapping nodes is relatively small for real-world networks, usually less than 30%.

Ranciati et al. (2017) proposed an overlapping model that is similar to but not an overlapping SBM, because the connections between the nodes, which they called actors, are unknown in the data. Instead, available in the data are whether the actors have attended certain events. Equivalently, the data can be viewed as a bipartite network between the actors and the events. In the proposed model, clustering is applied to the actor nodes, which can belong to one or more groups, hence the overlapping nature of the model. Subsequently, memberships were inferred without the direct knowledge of edges *between* the actor nodes.

## Related methods for graphs

In this section, two classes of models for graphs and one class of models for *hypergraphs*, are reviewed, with a focus on how they work with network or graph data in different ways to SBMs. Community detection algorithms will also be mentioned, which are a class of *methods* with a similar (but not identical) goal to SBMs.

### Latent feature models

A class of models closely related to SBMs is the latent feature models (Miller et al. 2009; Mørup et al. 2011), in which there are no longer $K$ groups but $K$ *features*. For example, if Fig. 1 represents a social network where nodes and edges correspond to people and personal connections, respectively, then the $K = 3$ features could be gender (0 for female and 1 for male), whether they wear glasses, and whether they are left-handed (0) or right-handed (1). Each element of $\mathbf{Z}_p$ is a binary latent variable without constraint, representing the absence or presence of a latent *feature*, meaning that the sample space of $\mathbf{Z}_p$ is the $2^K$ combinations of 0's and 1's (note the similarity with the number of combinations in an overlapping SBM in "Overlapping SBM" section). Continuing with the example, if node 1 is a female who wears glasses and is right-handed, $\mathbf{Z}_1 = (0\ 1\ 1)^T$.

The element $\mathbf{C}_{ij}$ ($1 \leq i, j \leq K$) in the matrix $\mathbf{C}$ represents the probability of an edge from a node with feature $i$ to a node with feature $j$. In their infinite multiple relational model (IMRM), Mørup et al. (2011) assumed that the feature combinations are independent, which means that the probability of an edge for the dyad $(p, q)$ is

$$
\begin{aligned}
\Pr\left(\mathbf{Y}_{pq}=1\middle|\mathbf{Z},\mathbf{C}\right) &= 1 - \Pr\left(\mathbf{Y}_{pq}=0\middle|\mathbf{Z},\mathbf{C}\right) \\
&= 1 - \prod_{i,j}^{K} \Pr\left(\text{No edge from } p \text{ with feature } i \text{ to } q \text{ with feature } j|\mathbf{Z},\mathbf{C}\right) \\
&= 1 - \prod_{i,j}^{K} \left(1-\mathbf{C}_{ij}\right)^{\mathbf{Z}_{pi}\mathbf{Z}_{qj}} \\
&= 1 - \exp\left(\sum_{i,j}^{K} \mathbf{Z}_{pi}\log(1-\mathbf{C}_{ij})\mathbf{Z}_{qj}\right) = 1 - \exp\left(\mathbf{Z}_p^T\mathbf{P}\mathbf{Z}_q\right),
\end{aligned}
\tag{16}
$$

where $\mathbf{P}$ is a matrix such that $\mathbf{P}_{ij}=\log(1-\mathbf{C}_{ij})$. Miller et al. (2009) specified their latent feature relational model (LFRM) in a slightly different way, by using a weight matrix $\mathbf{W}$ in place of $\mathbf{C}$ such that $\mathbf{Z}_p^T\mathbf{W}\mathbf{Z}_q$ can take any real value, and a function $\sigma(\cdot)$ that maps $(-\infty,\infty)$ to $(0,1)$ such that

$$
\Pr(\mathbf{Y}_{pq}=1|\mathbf{Z},\mathbf{W}) = \sigma\left(\mathbf{Z}_p^T\mathbf{W}\mathbf{Z}_q\right).
\tag{17}
$$

Not only do (16) and (17) look similar to the (conditional) edge probability $\Pr(\mathbf{Y}_{pq}=1|\mathbf{Z},\mathbf{C})=\mathbf{Z}_p^T\mathbf{C}\mathbf{Z}_q$ in the aforementioned version of SBM, the latent feature models can also be reduced to the SBM when only one feature is allowed, by imposing the constraint $\mathbf{Z}_p^T\mathbf{1}_K=1$, where $\mathbf{1}_K$ is a $K$-vector of 1's.

The latent feature models should not be confused with the mixed membership models ("Clustering approach" section), where a node can belong to multiple groups with weights. The practical difference is, while the SBM so far and the latent feature model allow one and multiple 1's in each $\mathbf{Z}_p$, respectively, the MMSBMs allow non-binary and non-negative weights in $\boldsymbol{\theta}_p$, subject to the constraint that these weights sum to 1 for each $p$.

Zhou (2015) proposed a similar model, called the edge partition model (EPM), in which each element of $\mathbf{Z}$ and $\mathbf{W}$ is assumed to come from the Gamma distribution, resulting in a non-negative value for $\mathbf{Z}_p^T\mathbf{W}\mathbf{Z}_q$, which is assumed to be the mean rate of *interaction*, for dyad $(p,q)$. Assuming that the number of interactions is Poisson distributed and that $p$ is connected to $q$ if they have interacted once, we have

$$
\Pr\left(\mathbf{Y}_{pq}=1|\mathbf{Z},\mathbf{W}\right) = 1 - \exp\left(-\mathbf{Z}_p^T\mathbf{W}\mathbf{Z}_q\right).
\tag{18}
$$

Palla et al. (2012) extended the LFRM by Miller et al. (2009) by introducing subclusters for the latent features in their infinite latent attribute (ILA) model. Universally, there are still $K$ features and an $n \times K$ binary matrix $\mathbf{Z}$ representing the presence or absence of latent features for the nodes. Additionally, for feature $m$, there are $K^{(m)}$ subclusters, a $K^{(m)} \times K^{(m)}$ weight matrix $W^{(m)}$, and an $n$-vector denoted by $\mathbf{V}^{(m)}$ such that $\mathbf{V}_p^{(m)}$ represents the subcluster that node $p$ belongs to if it has feature $m$. If we denote the collections of subcluster vectors and weight matrices by $\mathbf{V}$ and $\mathbf{W}$, respectively, we have

$$
\Pr\left(\mathbf{Y}_{pq}=1|\mathbf{Z},\mathbf{V},\mathbf{W}\right) = \sigma\left(\sum_{m=1}^{K} \mathbf{Z}_{pm}\mathbf{W}_{\mathbf{V}_p^{(m)}\mathbf{V}_q^{(m)}}^{(m)}\mathbf{Z}_{qm}\right),
\tag{19}
$$

where $\sigma(\cdot)$ is the same as in Miller et al. (2009) that maps $(-\infty,\infty)$ to $(0,1)$, such as the sigmoid function $\sigma(x)=(1+\exp(-x))^{-1}$ or the probit function $\sigma(x)=\Phi(x)$.

Comparing the models introduced in this section, the EPM (Zhou 2015) is found to outperform the ILA model (Palla et al. 2012), which in turn outperforms the LFRM (Miller

et al. 2009), which in turn outperforms the MMSBM (Airoldi et al. 2008) introduced in "Clustering approach" section. However, (Mørup et al. 2011) did not compare their IMRM (Mørup et al. 2011) with the models here, not was there a single comparison between all these latent feature models.

### Hypergraph models

Coauthorship or collaboration networks are a popular kind of data that statistical network methods have been applied to (Newman 2001a; 2001b; 2004; Newman and Girvan 2004; Ji and Jin 2016). However, the graphs are usually constructed by assigning an edge, possibly valued, to two authors if they have coauthored one or more articles. Such representation, however, does not preserve all the information (Ng and Murphy 2018) and may not be very realistic. For example, pairwise edges between nodes (authors) 1, 2 and 3 could mean that each pair have collaborated separately, or that all three of them have written one or more articles as a whole, or a combination of both. Furthermore, when an article is written by, say, more than 20 authors, it is unrealistic to assume that each pair of authors know each other with equal strengths.

A more natural representation of such data is through the use of hypergraph. Specifically, a *hyperedge* is an unordered subset of the node set $\mathcal{N}$, and when all hyperedges are node pairs, the hypergraph is reduced to a graph. In the example with the three authors, each pair having collaborated separately corresponds to 3 hyperedges: $\{1, 2\}$, $\{2, 3\}$ and $\{3, 1\}$, whereas all three of them collaborating together corresponds to 1 hyperedge: $\{1, 2, 3\}$.

Hypergraph data can also be modelled with the same goal of clustering the nodes. However, it is not quite direct to extending from SBMs to "connect a random number of two or more nodes", making it more difficult to work with hyperedges. Ng and Murphy (2018) resorted to and extended the latent class analysis (LCA), in which the *hyperedges* are clustered into the latent groups, and the memberships of the nodes can be seen as a mixture of the memberships of the hyperedges they are in.

Lunagómez et al. (2017) considered a geometric representation of the nodes in an Euclidean space to construct a hypergraph. For ease of explanation, we assume the nodes lie on a 2-dimensional plane, and, for each node, a circle of the same radius is drawn. Then for each set of nodes that have their circles overlapped, a hyperedge is assigned. Essentially, instead of clustering the nodes into groups, this model *projects* them onto an Euclidean space and infers their latent *positions*.

### Latent space models

Projecting the nodes of a *graph* $\mathcal{G}$ to an Euclidean space and discovering their latent positions has also been explored in the literature. Hoff et al. (2002) proposed the latent space model, in which the latent variable associated with node $p$, still denoted by $\mathbf{Z}_p$ here, does not correspond to the group membership, but a position represented by, for example, the vector of coordinates in the Euclidean space. Then the probability of nodes $p$ and $q$ having an edge in $\mathcal{G}$, assuming it is undirected, depends on the distance between $\mathbf{Z}_p$ and $\mathbf{Z}_q$:

$$\Pr(\mathbf{Y}_{pq} = 1 | \mathbf{Z}) = \frac{\exp\left[-d(\mathbf{Z}_p, \mathbf{Z}_q)\right]}{1 + \exp\left[-d(\mathbf{Z}_p, \mathbf{Z}_q)\right]},$$

where $d(\cdot, \cdot)$ is a distance measure, possibly with some parameters, satisfying the triangular inequality. So, probability of an edge between $p$ and $q$ decreases with the distance between $\mathbf{Z}_p$ and $\mathbf{Z}_q$. If covariates $\boldsymbol{x}_{pq}$ about the dyad $(p, q)$ are available, they can also be incorporated into the model:

$$\Pr(\mathbf{Y}_{pq} = 1 | \mathbf{Z}, \alpha, \boldsymbol{\beta}) = \frac{\exp\left[\alpha + \boldsymbol{\beta}^T \boldsymbol{x}_{pq} - d(\mathbf{Z}_p, \mathbf{Z}_q)\right]}{1 + \exp\left[\alpha + \boldsymbol{\beta}^T \boldsymbol{x}_{pq} - d(\mathbf{Z}_p, \mathbf{Z}_q)\right]},$$

where $\alpha$ and $\boldsymbol{\beta}$ are extra parameters. They noted that this model formulation is useful for handling undirected graphs because of the symmetry between $p$ and $q$. For directed graphs, they proposed

$$\Pr(\mathbf{Y}_{pq} = 1 | \mathbf{Z}, \alpha, \boldsymbol{\beta}) = \frac{\exp\left[\alpha + \beta^T \boldsymbol{x}_{pq} + \mathbf{Z}_p^T \mathbf{Z}_q / |\mathbf{Z}_q|\right]}{1 + \exp\left[\alpha + \beta^T \boldsymbol{x}_{pq} + \mathbf{Z}_p^T \mathbf{Z}_q / |\mathbf{Z}_q|\right]},$$

where the asymmetric term $\mathbf{Z}_p^T \mathbf{Z}_q / |\mathbf{Z}_q|$ is the signed magnitude of $\mathbf{Z}_p$ in the direction of $\mathbf{Z}_q$.

Real-world networks usually exhibit transitivity, which means that, if both nodes A and B are connected to node C, then A and B are likely to be connected. Another common phenonmenon is homophily, which means that nodes with similar attributes are more likely to be connected. They are accounted for by the above model through the use of latent space and dyad-specific covariates $\boldsymbol{x}_{pq}$, respectively. Handcock et al. (2007) proposed an extension in the form of a latent space *cluster* model, by considering the *apriori* distribution of the latent positions. Specifically, for each node $p$, $\mathbf{Z}_p$ is assumed to be drawn from a mixture of $K$ Gaussian distributions, each of which has a different mean and covariance matrix to represent a different group/cluster. In this way, the clustering of the nodes are accounted for explicitly.

Airoldi et al. (2008) noted the similarity between their MMSBM and the latent space models. For generating an edge between nodes $p$ and $q$, the terms $\mathbf{Z}_{pq}^T \mathbf{C} \mathbf{Z}_{qp}$ and $\mathbf{Z}_p^T \mathbf{I} \mathbf{Z}_q$ are involved in the former and the latter, respectively, where $\mathbf{I}$ is an identity matrix. They have also compared their performances when being applied to the same set of data.

A recent development with latent space models is by Sanna Passino and Heard (2019), who first used spectral clustering (von Luxburg 2007) to project, or embed, the graph to a $d$-dimensional Euclidean space. Then, a Gaussian mixture model, with $K$ components, is fit to these spectral embeddings. Their novelty is the estimation of *d and K* simultaneously ("Modelling" section) in their inference algorithm. For more details on spectral clustering and its relation to SBM, please see Rohe et al. (2011).

## Community detection

Without the pretext of statistical or probabilistic modelling, community detection can be the goal of analysing a network, which is to cluster nodes so that the edge density is high within a group and low between groups. This concept is also called assortativeness. In the context of SBMs, this means $\mathbf{C}_{ii}$ $(i = 1, 2, \ldots, K)$ is high while $\mathbf{C}_{ij}$ is low for $j \neq i$. As mentioned in "Stochastic equivalence" section, this is not guaranteed by the concept of stochastic equivalence alone. While SBMs can find communities with high within-group edge densities, they are in fact a more general method that allow other types of structure in the network to be found (Guimerà and Sales-Pardo 2009; McDaid et al. 2013).

The above effect is illustrated by, for example, the difference between the DC-SBM by Karrer and Newman (2011) and the original version, when applied to a real-world network with $K = 2$. While the former accounted for the variation in the degree and managed to discover the two communities, the latter put the highly connected nodes together in one group, the rest in another. In bigger networks with nodes on the *periphery* of the network, that is, they are only connected to one or a few nodes which are more central to the network, these peripheral nodes will be put together in a "miscellaneous" group with a low edge density, under the original SBM, instead of the same groups as the more central nodes they are connected to.

### Assortative SBM

One way of achieving community detection is to modify the SBMs to align with this goal. In the assortative (or affinity) SBM (Gopalan et al. 2012; Li et al. 2016), a constraint is imposed that $\mathbf{C}_{ij} = \delta$ for $i \neq j$, where $\delta$ is a parameter presumed to be smaller than $\mathbf{C}_{ii}$. While reducing the number of parameters in $\mathbf{C}$ from $K(K + 1)/2$ (in the case of undirected graphs, $K^2$ in the case of directed ones) to $K + 1$ may not significantly reduce the computational cost unless $K$ is large, it implies assortativeness. However, it should be noted that incorporating assortativeness in SBMs is not a universal solution. For example, it is not sensible when bipartite networks are modelled, in which connectivity is high between groups but zero within groups. Therefore, caution should be taken whenever an assortative SBM is used, although the stochastic gradient method by Li et al. (2016) should be easily generalisable to a non-assortative model.

Lu and Szymanski (2019b) proposed a regularised SBM which extends the DC-SBM to control the desired level of assortativeness. The expected number of edges for the dyad $(p, q)$ is now $\phi_p \phi_q \mathbf{Z}_p^T \mathbf{C} \mathbf{Z}_q$ if $\mathbf{Z}_p = \mathbf{Z}_q$, $(k_p - \phi_p)(k_q - \phi_q)\mathbf{Z}_p^T \mathbf{C} \mathbf{Z}_q$ otherwise, where $k_p$ is the degree of node $p$. While a different expected number of edges is allowed according to the group memberships, the parameter $\phi_p$ is regulated by a parameter $h$, which is a number between 0 and 1, according to $\phi_p = \max(hk_p, 1)$. The tuning parameter $h$ is not estimated but varied, to give differnt clustering results corresonding to different levels of assortativeness. A high value of $h$ leads to a more assortative partition and, in the application, recovers the same known factions in the karate club network as in Karrer and Newman (2011).

The assortative models introduced so far are actual SBMs and not merely related methods for graphs. It should be noted that they are introduced here because of the proximity to the goal of community detection.

### Non-probabilistic and modularity methods

Another way of achieving community detection is to step away from SBMs, and apply methods which are not based on statistical or probabilistic *modelling* but mainly on heuristics, and are usually iterative in nature. For example, in the label propogation algorithm by Raghavan et al. (2007), initially each node is randomly assigned to one of the $K$ groups. Then each node takes turn to join the group to which the maximum number of its neighbours belong (with ties broken uniformly randomly). The iterative process continues until no node changes its group membership anymore. Other methods, for example, are based on the edge betweenness centrality measure Girvan and Newman (2002), random walks on the graph (Pons and Latapy 2006), and network flows and information-theoretic principles (Rosvall et al. 2009).

One very popular framework of community detection is the use, and optimisation, of the modularity of a network, by the highly-cited Newman and Girvan (2004). Assuming an undirected (but possibly valued) graph, the formulae of the modularity, denoted by $Q$, is

$$Q = \frac{1}{2M} \sum_{p,q}^{n} \left[ \mathbf{Y}_{pq} - \frac{k_p k_q}{2M} \right] \mathbf{1}\{\mathbf{Z}_p = \mathbf{Z}_q\}, \tag{20}$$

where $k_p = \sum_{q=1}^{n} \mathbf{Y}_{pq}$ is the *degree* of node $p$. As it can be interpreted as the number of edges within groups minus the expected number of such edges, the modularity is a very useful measure for how "good" the clustering is, and numerous algorithms have been proposed for its optimisation (Clauset et al. 2004; Newman 2006a; Newman 2006b; Wakita and Tsurumi 2007; Blondel et al. 2008). Moreover, it can be computed for results not based on modularity optimisation, and compared between all methods. Clauset et al. (2004) suggested that, in practice, a modularity above 0.3 is a good indicator of significant community structure.

Newman (2016) established a connection between modularity optimisation and the DC-SBM ("Poisson and degree-corrected SBMs" section). Specifically, maximising the likelihood of a special case of the DC-SBM is equivalent to maximising a generalised version of the modularity. This is also noted by, for example, (Lu and Szymanski 2019a), when they tackled the resolution limit problem, which is reviewed below.

### Resolution limit

Community detection algorithms are in general easy to implement, and likely to be faster than applying an SBM. However, its disadvantages includes that different initial configurations may lead to different results even under the same algorithm, and that different results give vastly different results. More importantly, modularity optimisation methods suffer from a resolution limit (Fortunato and Barthélemy 2007; Good et al. 2010; Lancichinetti and Fortunato 2011). This means that by maximising the modularity, smaller groups or clusters cannot be detected, especially for large graphs. For example, performing community detection on a network of 1 million nodes may result in the smallest group having 500 nodes, but the algorithm performed on the whole network cannot go deeper to further cluster these 500 nodes (or that modularity will decrease by doing so). The need of a possible additional round of community detection is not ideal as it should have been a systematic part of the initial community detection.

The above issue of resolution limit is tackled by recent works in various ways. Chen et al. (2014) proposed an alternative to modularity, called modularity density, which theoretically resolves the resolution limit problem, and emprically improves results significantly when being applied to community detection algorithms. Lu and Szymanski (2019a) proposed an agglomerative community detection algorithm to detect communities at multiple scales, thus avoiding the resolution limit issue. Peixoto (2013) noted the connection with the resolution limit for SBMs, and went on to resolve the issue in the hierarchical model (Peixoto 2014b) ("Microcanonical SBM" section).

### Miscellaneous

Community detection is, in itself, a largely studied topic in the literature, and a few important reviews should be referred to for further exploration of the topic. Fortunato

(2010) provided an introduction and comprehensive review of community detection, while Abbe (2018) surveyed the developments that establish the fundamental limits for community detection in SBMs. Schaub et al. (2014) reviewed a spectrum of algorithms according to their different motivations that underpin community detection, with the aim of providing guildlines for selecting appropriate algorithms for the given purposes. In addition to providing an overview of algorithms based on modularity optimisation, Cherifi et al. (2019) reviewed recent advances on two specific topics, namely community detection for time evolving networks, and immunisation strategies in networks with overlapping and non-overlapping community structure. The former is, in the context of SBMs, reviewed in "SBM with longitudinal modelling" section, while the latter is useful for, for example, targeting a small group of nodes to prevent the spread of epidemics in networks.

### Inference approach

In this section, the general framework of inference for SBMs is reviewed, in which there are two main classes of methods: Monte Carlo ("Monte Carlo methods" section) and variational ("Variational methods" section), to answer question 2 in "Introduction" section. Greedy methods will be mentioned in "Greedy methods and others" section, while methods for predicting or correcting the edges or non-edges will be presented in "Missingness and errors" section. While the general approach is discussed here, the more specific models or algorithms for estimating or selecting the number of groups $K$ will be mentioned in "Number of groups" section. Nevertheless, inference and the issue of $K$ are, most of the time, intertwined with each other, and are presented in separate sections here for clarity.

We first observe that, by combining (2) and (4), inference is possible by the frequentist approach. This can be achieved via direct maximisation of likelihood or the expectation-maximisation (EM) algorithm (Dempster et al. 1977), both of which are illustrated in Snijders and Nowicki (1997) for a simple case where $K = 2$. However, we will focus on the more popular and arguably more powerful Bayesian approach here, as did Nowicki and Snijders (2001). What remains is assigning priors to $\mathbf{C}$ and $\boldsymbol{\theta}$ before inference can be carried out. We assume each element of $\mathbf{C}$ has an independent Beta prior, that is $\mathbf{C}_{ij} \sim \text{Beta}\left(\mathbf{A}_{ij}, \mathbf{B}_{ij}\right)$, where $\mathbf{A}$ and $\mathbf{B}$ are $K \times K$ matrices with all positive hyperparameters. We also assume $\boldsymbol{\theta}$ arises from the Dirichlet($\alpha \mathbf{1}_K$) distribution, of which the parameter $\alpha$ comes from a Gamma($a, b$) prior.

The joint posterior of $\mathbf{Z}, \boldsymbol{\theta}, \mathbf{C}$ and $\alpha$, up to a proportionality constant, is

$$
\begin{aligned}
\pi\left(\mathbf{Z}, \boldsymbol{\theta}, \mathbf{C}, \alpha | \mathbf{Y}\right) &\propto \pi\left(\mathbf{Y}, \mathbf{Z}, \boldsymbol{\theta}, \mathbf{C}, \alpha\right) \\
&= \pi\left(\mathbf{Y}|\mathbf{Z}, \boldsymbol{\theta}, \mathbf{C}, \alpha\right) \times \pi\left(\mathbf{Z}|\boldsymbol{\theta}, \mathbf{C}, \alpha\right) \times \pi\left(\boldsymbol{\theta}|\mathbf{C}, \alpha\right) \times \pi\left(\mathbf{C}, \alpha\right) \\
&= \pi\left(\mathbf{Y}|\mathbf{Z}, \mathbf{C}\right) \times \pi\left(\mathbf{Z}|\boldsymbol{\theta}\right) \times \pi\left(\boldsymbol{\theta}|\alpha\right) \times \pi\left(\mathbf{C}\right) \times \pi(\alpha) \\
&\propto \prod_{p<q}^{n}\left[\left(\mathbf{Z}_p^T \mathbf{C} \mathbf{Z}_q\right)^{\mathbf{Y}_{pq}}\left(1-\mathbf{Z}_p^T \mathbf{C} \mathbf{Z}_q\right)^{\left(1-\mathbf{Y}_{pq}\right)}\right] \times \prod_{p=1}^{n} \mathbf{Z}_p^T \boldsymbol{\theta} \\
&\quad \times\left[\Gamma\left(K\alpha\right) \mathbf{1}\{\sum_{i=1}^{K} \theta_i = 1\} \prod_{i=1}^{K} \frac{\theta_i^{\alpha-1}}{\Gamma(\alpha)}\right] \\
&\quad \times \prod_{i \leq j}^{K}\left[\mathbf{C}_{ij}^{\mathbf{A}_{ij}-1}\left(1-\mathbf{C}_{ij}\right)^{\mathbf{B}_{ij}-1}\right] \times \alpha^{a-1} e^{-b\alpha}.
\end{aligned}
\tag{21}
$$

Under the soft clustering approach in "Clustering approach" section, the weight vectors $\{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_K\}$ are assumed to be i.i.d. according to the Dirichlet($\alpha \mathbf{1}_K$) distribution. This yields the joint posterior of $\mathbf{Z}$, $\Theta$, $\mathbf{C}$ and $\alpha$:

$$
\begin{aligned}
\pi\left(\mathbf{Z}, \Theta, \mathbf{C}, \alpha | \mathbf{Y}\right) &\propto \pi\left(\mathbf{Y}|\mathbf{Z}, \mathbf{C}\right) \times \pi\left(\mathbf{Z}|\Theta\right) \times \pi\left(\Theta|\alpha\right) \times \pi\left(\mathbf{C}\right) \times \pi(\alpha) \\
&\propto \prod_{p<q}^{n} \left[ \left(\mathbf{Z}_{pq}^T \mathbf{C} \mathbf{Z}_{qp}\right)^{\mathbf{Y}_{pq}} \left(1 - \mathbf{Z}_{pq}^T \mathbf{C} \mathbf{Z}_{qp}\right)^{(1-\mathbf{Y}_{pq})} \right] \\
&\quad \times \prod_{p<q}^{n} \left(\mathbf{Z}_{pq}^T \boldsymbol{\theta}_p \mathbf{Z}_{qp}^T \boldsymbol{\theta}_q\right) \times \prod_{p=1}^{n} \left[ \Gamma\left(K\alpha\right) \mathbf{1}\{\boldsymbol{\theta}_p^T \mathbf{1}_K = 1\} \prod_{i=1}^{K} \frac{\Theta_{pi}^{\alpha-1}}{\Gamma(\alpha)} \right] \\
&\quad \times \prod_{i \leq j}^{K} \left[ \mathbf{C}_{ij}^{\mathbf{A}_{ij}-1} \left(1 - \mathbf{C}_{ij}\right)^{\mathbf{B}_{ij}-1} \right] \times \alpha^{a-1} e^{-b\alpha}.
\end{aligned}
\tag{22}
$$

Comparing this with (21) illustrates why the hard clustering approach is preferred in the literature of SBMs. As $K$ is a lot smaller than $n$ usually (hence the purpose of clustering), the computational cost mainly depends on the number of latent variables $\mathbf{Z}$. In the hard and soft clustering approaches, this amounts to $O(n)$ and $O(n^2)$ iterations, respectively. The quadratic computation cost means that a simple Gibbs sampler is not very scalable in soft clustering (Mørup et al. 2011).

### Monte Carlo methods

If algorithmic simplicity is preferred to computational efficiency, inference can be carried out in a straightforward way via Markov chain Monte Carlo (MCMC). More specifically, a simple regular Gibbs sampler can be used, where all the parameters and latent variables (except $\alpha$) can be updated via individual Gibbs steps; see, for example, Nowicki and Snijders (2001) and Lee and Wilkinson (2018). In quite a few cases, a Gibbs sampler is natural when $K$ is being modelled ("Modelling" section), and this is used by both SBMs (Tang and Yang 2011; 2014; Palla et al. 2012; Fan et al. 2015; Zhou 2015) and a latent feature model (Miller et al. 2009). Other articles on SBMs that use MCMC include Yang et al. (2011), DuBois et al. (2013), McDaid et al. (2013), Li et al. (2016), Ludkin et al. (2018), Peixoto (2017a), Peixoto (2017b) and Peixoto (2018a). Mørup et al. (2011) proposed using Hamiltonian Monte Carlo (HMC) for (a transformation of) each element in $\mathbf{C}$, in which the gradient of the log-posterior is utilised. In their dynamic SBM, Xu and Hero III (2013) incorporated a sequential Monte Carlo (SMC) step with a label-switching algorithm by Karrer and Newman (2011) (explained below).

If computational efficiency is the focus of the MCMC algorithm, careful considersations are required so as not to waste computational time on naive moves. Mørup et al. (2011) noted that, in a latent feature model, the possible combinations of latent features is $2^{Kn}$, compared to the $K^n$ possible combinations of group memberships in an SBM, and so standard Gibbs samplers are unlikely to be scalable. Therefore, in their algorithm, certain types of moves are employed so that the computational complexity increases linearly with the number of *edges*, not of *node combinations*. Li et al. (2016) made a similar claim for their stochastic gradient MCMC algorithm, in which only a small mini-batch of the nodes is required in each iteration, to greatly reduce the computational overhead. These algorithms benefit from the fact that large graphs are usually sparse in reality.

### Collapsed SBMs and efficient moves

While both McDaid et al. (2013) and Peixoto (2017a) and Peixoto (2017b) integrated **C** out in their respective SBMs, they proposed different efficient moves in their MCMC algorithms. In McDaid et al. (2013), one of the following four moves is selected randomly uniformly and performed in each iteration. The first one is a Gibbs move for a randomly selected node. The second is a move that selects two groups at random and proposes to reassign all the nodes in these two groups, in a scheme described by Nobile and Fearnside (2007). The last two are moves that affect the number of groups alongside the memberships, as $K$ is treated as a parameters and estimated. McDaid et al. (2013) also proposed a method to deal with the issue of label-switching, which is circumvented in the spectral clustering model by Sanna Passino and Heard (2019).

Peixoto (2017a); Peixoto (2017b) applied a single-node move proposed by Peixoto (2014a) that works as follows. For node $p$, whose membership $\mathbf{Z}_p$ is to be updated, a neighbour $q$ is selected randomly uniformly, whose membership is, say, $\mathbf{Z}_q = i$. Then node $p$ is proposed to move to a group $j$ (which could be the same as $i$) with probability proportional to $\mathbf{E}_{ij}$, that is, the number of edges between groups $i$ and $j$. Note that this is different from proposing to move $p$ to a group $j$ with probability proportional to the number of neighbours of $p$ in group $j$. Another move is also proposed for merging groups, as $K$ is being modelled ("Modelling" section).

### Variational methods

An alternative to the Monte Carlo methods is the class of variational expectation-maximisation (VEM) methods. The principle of these algorithms is to first provide a lower bound to the marginal log-likelihood, in which the latent variables ($\mathbf{Z}$ here) are integrated out, by the use of an approximate *variational* distribution. This lower bound is then tightened or maximised, with respect to the model parameters ($\boldsymbol{\theta}$, $\mathbf{C}$ and $\alpha$ here) as well as those of the variational distribution. We illustrate this using the Bernoulli SBM. By writing $\boldsymbol{\eta} = \{\boldsymbol{\theta}, \mathbf{C}, \alpha\}$, for any distribution $Q$ of the latent variables $\mathbf{Z}$, we have

$$
\begin{aligned}
\log \pi(\mathbf{Y}|\boldsymbol{\eta}) &= \log \int \frac{\pi(\mathbf{Y}, \mathbf{Z}|\boldsymbol{\eta})}{Q(\mathbf{Z})} Q(\mathbf{Z}) d\mathbf{Z} \\
&\geq \int \log \frac{\pi(\mathbf{Y}, \mathbf{Z}|\boldsymbol{\eta})}{Q(\mathbf{Z})} Q(\mathbf{Z}) d\mathbf{Z} \\
&= \mathbb{E}_Q \left[ \log \pi(\mathbf{Y}, \mathbf{Z}|\boldsymbol{\eta}) - \log Q(\mathbf{Z}) \right].
\end{aligned}
\tag{23}
$$

The second line is due to Jensen's inequality as the logarithm function is concave. As $\log \pi(\mathbf{Y}|\boldsymbol{\eta}) = \int \log \pi(\mathbf{Y}|\boldsymbol{\eta}) Q(\mathbf{Z}) d\mathbf{Z}$, the difference in the inequality is

$$
\begin{aligned}
&\log \pi(\mathbf{Y}|\boldsymbol{\eta}) - \int \log \frac{\pi(\mathbf{Y}, \mathbf{Z}|\boldsymbol{\eta})}{Q(\mathbf{Z})} Q(\mathbf{Z}) d\mathbf{Z} \\
&= \int \left[ \log \pi(\mathbf{Y}|\boldsymbol{\eta}) - \log \frac{\pi(\mathbf{Y}, \mathbf{Z}|\boldsymbol{\eta})}{Q(\mathbf{Z})} \right] Q(\mathbf{Z}) d\mathbf{Z} \\
&= \mathbb{E}_Q \left[ \log \frac{Q(\mathbf{Z})}{\pi(\mathbf{Z}|\mathbf{Y}, \boldsymbol{\eta})} \right],
\end{aligned}
$$

which is the Kullback-Leibler (K-L) divergence of $\pi(\mathbf{Z}|\mathbf{Y}, \boldsymbol{\eta})$ from $Q(\mathbf{Z})$, denoted by $D_{KL}(Q(\mathbf{Z}) \,||\, \pi(\mathbf{Z}|\mathbf{Y}, \boldsymbol{\eta}))$. Therefore,

$$
\log \pi(\mathbf{Y}|\boldsymbol{\eta}) = \mathbb{E}_Q \left[ \log \pi(\mathbf{Y}, \mathbf{Z}|\boldsymbol{\eta}) - \log Q(\mathbf{Z}) \right] + D_{KL}(Q(\mathbf{Z}) \,||\, \pi(\mathbf{Z}|\mathbf{Y}, \boldsymbol{\eta})).
\tag{24}
$$

As the marginal likelihood on the left-hand side is constant to the choice of $Q$, maximising the first term with respect to $Q$ and $\boldsymbol{\eta}$ is equivalent to minimising the K-L divergence, thus improving the approximation. While the ideal choice of $Q(\mathbf{Z})$ is the conditional distribution $\pi(\mathbf{Z}|\mathbf{Y}, \boldsymbol{\eta})$ such that the K-L divergence is 0, the latter is usually intractable in the models discussed here, and so the best tractable choices of $Q(\mathbf{Z})$ are being sought. Usual choices are such that $Q(\mathbf{Z})$ is factorisable, making analytical calculations of the lower bound possible. Finally, the lower bound is iteratively maximised with respect to $\boldsymbol{\eta}$ in the M step, and (the parameters of) $Q$ in the E step, in an EM algorithm.

The factorisable variational distributions and the iterative steps have been illustrated by, for example, Kurihara et al. (2006) and the following references in this section. Gopalan et al. (2012) and Kim et al. (2013) considered stochastic optimisation in place of EM algorithms, while Matias et al. (2015) considered variants of the M step. Hayashi et al. (2016) opted for belief propagation as the alternative approach to obtaining the variational distribution $Q$.

That the variational approach is adopted is due to several reasons in different articles. Airoldi et al. (2008) argued that in the MMSBMs (Airoldi et al. 2008; Fu et al. 2009; Xing et al. 2010) it outperforms (with computational cost $O(nK + 2K)$) the corresponding MCMC methods (with computation cost $O(n^2)$). Similarly, (Vu et al. 2013) argued that their algorithm for SBM is more scalable ($O(n)$) compared to latent space models ($O(n^2)$) ("Latent space models" section). In some other cases, it facilitates the computation of a criterion, such as the ICL $\pi(\mathbf{Y}, \mathbf{Z})$ (Latouche et al. 2012; Matias et al. 2015; Hayashi et al. 2016; Matias and Miele 2017), or the observed data log-likelihood $\pi(\mathbf{Y}|\boldsymbol{\eta})$ (Yan et al. 2014). This criterion can be directly used for model selection, or equivalently selecting $K$ ("Criteria and model selection" section).

### Greedy methods and others

While most articles in the literature have used either Monte Carlo methods or variational methods, there are a few exceptions. In the DC-SBM (Karrer and Newman 2011), the log-likelihood can be obtained by summing the logarithm of (7) over all possible dyads. Instead of integrating out $\mathbf{C}$ and $\boldsymbol{\phi}$, they adopted a frequentist approach and substituted their respective MLEs to obtain the objective function $\pi(\mathbf{Y}|\mathbf{Z})$, which is the basis of their label-switching algorithm. In each step, each node is proposed to move from one group to another, and selected is the move that will most increase or least decrease $\pi(\mathbf{Y}|\mathbf{Z})$. Once all nodes have been moved and the group memberships $\mathbf{Z}$ are according updated, the objective function is calculated for all the steps this greedy algorithm passed through, and the one with the highest objective is selected as the initial state of another run of the algorithm. The algorithm is stopped when there is no further increase in the objective function. This label-switching algorithm is also incorporated by Xu and Hero III (2013) in the inference algorithm for their dynamic SBM, which will be discussed in "SBM with longitudinal modelling" section.

Côme and Latouche (2015) also proposed a greedy step in their inference algorithm. Before doing so, required is the ICL, $\log \pi(\mathbf{Y}, \mathbf{Z})$, by integrating out the parameters $\boldsymbol{\eta} = \{\boldsymbol{\theta}, \mathbf{C}, \alpha\}$. They worked out an asymptotic version via certain approximations, as well as an exact version under certain priors. The ICL then becomes the objective function in their greedy optimisation algorithm. Similar to Karrer and Newman (2011), in each step, each node is proposed to move from one group to another, and selected is the move

that will most increase the ICL. However, if no proposed move results in an increase, $\mathbf{Z}_p$ remains unchanged. The algorithm terminates again if there is no further increase in the objective.

Yan (2016) combined the work by Karrer and Newman (2011) and Côme and Latouche (2015), by approximating the ICL for the DC-SBM. As the focus is selecting the number of groups by comparing the objective under different values of $K$, Yan (2016) only provided the calculations for the ICL in the absence of an inference algorithm for the parameters and/or the latent variables.

In the work by van der Pas and van der Vaart (2018), they did not consider the inference algorithm and were mainly concerned with the posterior mode, that is, the value of $\mathbf{Z}$ that maximises the posterior density in (21), or the collapsed version thereof according to McDaid et al. (2013). They found that this mode converges to the true value of $\mathbf{Z}$ (as $n \to \infty$), under the frequentist setup that there *is* such a true value, and with the condition that the expected degree is at least of order $(\log n)^2$. The reason that the priors of $\mathbf{Z}$ (and of other parameters) are used even in a frequentist setup is that such priors actually play a part in establishing the convergenece. Furthermore, van der Pas and van der Vaart (2018) established a connection of such maximised posterior density, termed Bayesian modularity, with the likelihood modularity defined by Bickel and Chen (2009) ("Criteria and model selection" section).

### Missingness and errors

In some cases, the goal of applying an SBM is not (only) the inference of the group memberships $\mathbf{Z}$, but on dealing with partially or errorfully observed graphs. For example, there is no information regarding the interactions between some dyads. Inferring these edges or non-edges, with some (un)certainty, then becomes the goal of inference, and is called the *link prediction* problem in the literature. In the context of Sbms, this has been looked into by, for example, Guimerà and Sales-Pardo (2009), Zhou (2015), Vallès-Català et al. (2016), Zhao et al. (2017), Vallès-Català et al. (2018), Stanley et al. (2019) and Tarrés-Deulofeu et al. (2019), and also by Miller et al. (2009) for latent feature models. Closely related is the issue with errors in the observed graphs, where the spurious edges may lead to wrong conclusions. Here, *network reconstruction* becomes another goal of inference. This has been studied by, for example, Guimerà and Sales-Pardo (2009), Priebe et al. (2015) and Peixoto (2018b), in the context of SBMs.

### Number of groups

As the main goal of SBMs is to cluster nodes into groups or communities, without a given number of groups $K$, it is difficult to evaluate the likelihood and infer the group memberships $\mathbf{Z}$. Unless there is prior information on $K$, it might not be objective to carry out inference for one fixed value of $K$. Solutions usually come in two main directions. The first is to fit an SBM to multiple values of $K$, with a measure computed to quantify the goodness of fit, followed by selecting the optimal $K$. Works in this direction will be reviewed in "Criteria and model selection" section. The second is to model $K$ as yet another parameter and estimate it in the inference, most likely by transdimensional inference algorithms. Works in this direction will be reviewed in "Modelling" section. Together, these methods can be seen as an answer to questions 3 and 4 in "Introduction" section simultaneously.

In some articles, the number of groups is fixed, rather than being estimated. Snijders and Nowicki (1997) set $K = 2$, enabling the MLE calculations and working out the associated EM algorithm. Tallberg (2005) and Karrer and Newman (2011) used $K = 2$ and $K = 3$, respectively, for their data but acknowledged the issue of assuming that $K$ is given, which is usually not the case in practice. Yang et al. (2011) focussed on the dynamic structure of social networks, and fixed $K$ to 2 for two of their datasets, and to 3 for a third dataset, by using prior knowledge on the three sets of data. Similarly, Vu et al. (2013) set $K = 5$ by following external relevant practices of using five groups for their data. Xu and Hero III (2013) fixed $K = 7$ for their Enron email network data, also using prior knowledge on the classes the nodes (employees) belonged to. In their model for recommender systems, Godoy-Lorite et al. (2016) reported the results for 10 groups of users and 10 groups of items, and found no differences in performance when either number of groups is increased. The data analysed by Zhang et al. (2017) contained individual attributes that allowed them to divide the nodes into $K = 3$ groups, and their focus was on showing that their dynamic SBM obtained results closer to the ground truth than the static counterpart did. When van der Pas and van der Vaart (2018) illustrated the connection of their Bayesian modularity with the likelihood modularity, through the proximity of the empirical results to those by Bickel and Chen (2009), they used both $K = 2$ and $K = 4$ in the application to the same set of data.

### Criteria and model selection

Nowicki and Snijders (2001) made separate fits to their data using $K = 2, 3, 4, 5$, and compared the *information* as well as a parameter representing the "clearness" of the block structure, to aid their decision on $K = 3$. This set the precedent to using some kind of criterion to select the optimal $K$.

A measure previously mentioned that requires no modelling is the modularity in (20), which has been compared with the profile log-likelihood by Bickel and Chen (2009), who called it a likelihood modularity. To obtain this criterion, first observe that, in (3), the MLE of $\mathbf{C}_{ij}$ is simply $\hat{\mathbf{C}}_{ij} = \mathbf{E}_{ij}/\mathbf{N}_{ij}$, the edge density between groups $i$ and $j$. This estimate can then be plugged in (3) to obtain

$$\log \pi(\mathbf{Y}|\mathbf{Z}, \hat{\mathbf{C}}) = \sum_{i,j}^{K} \mathbf{N}_{ij} \left[ \frac{\mathbf{E}_{ij}}{\mathbf{N}_{ij}} \log \frac{\mathbf{E}_{ij}}{\mathbf{N}_{ij}} + \left( 1 - \frac{\mathbf{E}_{ij}}{\mathbf{N}_{ij}} \right) \log \left( 1 - \frac{\mathbf{E}_{ij}}{\mathbf{N}_{ij}} \right) \right].$$

While Bickel and Chen (2009) did not use the profile log-likelihood mainly for model selection, it can be seen as a precedent in moving from the non model-based modularity to a criterion based on SBM.

Another criterion used in earlier works is the Bayesian information criterion (BIC) (Airoldi et al. 2008; Fu et al. 2009; Xing et al. 2010). However, according to Latouche et al. (2012) and Côme and Latouche (2015), it is not tractable in most cases as it depends on the *observed data* log-likelihood $\log \pi(\mathbf{Y}|\boldsymbol{\eta})$, where $\boldsymbol{\eta}$ represents the collection of parameters, and is known to misestimate $K$ in certain situations (Yan et al. 2014). Yan (2016), however, drew connection with the minimum description length (MDL) (Peixoto 2013) and dervied a more principled BIC.

Instead of evaluating the BIC, Yan et al. (2014) focussed on approximating the observed data log-likelihood $\log \pi(\mathbf{Y}|\boldsymbol{\eta})$. They applied what they called belief propagation in their EM algorithm, in order to approximate $\log \pi(\mathbf{Y}|\boldsymbol{\eta})$ from $\log \pi(\mathbf{Y}|\mathbf{Z}, \boldsymbol{\eta})$, which is derived

for both DC-SBM and its original counterpart. By doing so, a (log-)likelihood ratio can be computed, denoted by $\log \pi(\mathbf{Y}|\boldsymbol{\eta}_1) - \log \pi(\mathbf{Y}|\boldsymbol{\eta}_2)$, where $\boldsymbol{\eta}_1$ and $\boldsymbol{\eta}_2$ correspond to the parameters under the DC-SBM and the original version, respectively. With its asymptotic behaviour derived, a likelihood ratio test can be performed to answer the question of whether degree correction should be applied to the data in hand.

While $\log \pi(\mathbf{Y}|\boldsymbol{\eta})$ or its approximation is also being used by, for example, Decelle et al. (2011) (who termed it free-energy density in statistical physics terms), more works were based on the ICL $\log \pi(\mathbf{Y}, \mathbf{Z})$, which is equivalent to the MDL (Peixoto 2014b; Newman and Reinert 2016) under compatiable assumptions. In order to obtain it, we start with the quantity $\log \pi(\mathbf{Y}, \mathbf{Z}|\boldsymbol{\eta})$ in (24), and attempt to integrate $\boldsymbol{\eta}$ out:

$$\log \pi(\mathbf{Y}, \mathbf{Z}) = \log \int \pi(\mathbf{Y}, \mathbf{Z}|\boldsymbol{\eta})\pi(\boldsymbol{\eta})d\boldsymbol{\eta},$$

However, this is usually not tractable (Côme and Latouche 2015). Daudin et al. (2008) provided an approximation for directed graphs:

$$\log \pi(\mathbf{Y}, \mathbf{Z}) \approx \max_{\boldsymbol{\eta}} \log \pi(\mathbf{Y}, \mathbf{Z}|\boldsymbol{\eta}) - \frac{K^2}{2}\log(n(n-1)) - \frac{K-1}{2}\log(n). \tag{25}$$

Daudin et al. (2008) also provided an algorithm for the optimisation of the first term on the right hand side. This approximate ICL has been adopted by, for example, Matias et al. (2015), Matias and Miele (2017), and Stanley et al. (2019).

The above approximation, however, can be improved by noting that some parameters in $\boldsymbol{\eta}$, such as $\mathbf{C}$ and $\boldsymbol{\theta}$, can be integrated out by considering conjugate priors. This is exactly the case in (9), which in fact is equivalent to $\log \pi(\mathbf{Y}, \mathbf{Z}|\boldsymbol{\eta})$ in (24) but with $\boldsymbol{\eta} = (\alpha, \beta, \gamma)$. Noting such potential, Latouche et al. (2012) and Côme and Latouche (2015) integrated out parameters whose dimensions depend on $K$ or $n$, and fixed the values of the remaining scalar parameters in $\boldsymbol{\eta}$, but parted ways in their subsequent novelty. While Côme and Latouche (2015) derived an exact ICL, Latouche et al. (2012) proposed to approximate the marginal log-likelihood $\log \pi(\mathbf{Y})$, using the version of (24) without $\boldsymbol{\eta}$. After the convergence of the variational algorithm ("Variational methods" section), the first term on the right hand side of (24) can be computed by plugging in the estimated $\mathbf{Z}$. This maximised lower bound $\mathbb{E}_Q\left[\log \pi(\mathbf{Y}, \mathbf{Z}) - \log Q(\mathbf{Z})\right]$ is then used by Latouche et al. (2012) as the approximation of the marginal log-likelihood, under the assumption that the K-L divergence is close to zero and does not depend on $K$. This is also being adopted by Aicher et al. (2015), who calculated the Bayes factor of one model to another, according to the difference of their respective approximate marginal log-likelihoods. Similarly, Hayashi et al. (2016) provided an approximation to the marginal log-likelihood, this time an asymptotic one, called the fully factorised information criterion ($F^2IC$).

Wang and Bickel (2017) also investigated log-likelihood ratio similar to Yan et al. (2014), but this time for an underfitting/overfitting $K$ against the true $K$. Using its asymptotic results, they derived a penalty term $\lambda \frac{K(K+1)}{2} n \log n$, where $\lambda$ is a tuning parameter, and subsequently a penalised (log-)likelihood criterion. However, Hu et al. (2019) argued that this penalty tends to underestimate $K$, and therefore proposed a lighter penalty $\lambda n \log K + \frac{K(K+1)}{2} \log n$ in their *corrected* BIC.

Other criteria used include the mean log-likelihood of held-out test data (Gopalan et al. 2012; DuBois et al. 2013; Li et al. 2016), termed perplexity by the latter, as the criterion for selecting $K$. Finally, Ludkin et al. (2018) considered the ratio of the sum of the squared

distances in the *k*-means clustering for all nodes in different groups, to the sum of squared distances between all node pairs, in what they called the Elbow plot, to determine the number of groups.

Vallès-Català et al. (2018) investigated an empirical approach to model selection, which is to maximise the performance in link prediction ("Missingness and errors" section). They found that the results are sometimes inconsistent with those by a criterion formulated in a similar way to those described above, which is the posterior likelihood $\pi(\mathbf{Z}|\mathbf{Y})$ in their case.

### Modelling

The formulation of the SBM and the inference algorithm in "Inference approach" section relies on $K$ being specified beforehand. To increase the flexibility of the model, $\boldsymbol{\theta}$ can be assumed to arise not from a Dirichlet distribution but from a (Hierarchical) Dirichlet process, which will be introduced in "Hierarchical Dirichlet process" section, as it was first used in topic modelling. In this way, $K$ becomes a random quantity generated by the process, and potentially an infinite number of groups is allowed. By incorporating the Dirichlet process in the SBM, as did Kurihara et al. (2006), Mørup et al. (2011), Tang and Yang (2011), Tang and Yang (2014), Kim et al. (2013) and Fan et al. (2015), $K$ can be estimated along other parameters and latent variables.

A similar structure for models which are not SBMs can be incorporated, so that $K$ can be modelled and inferred. In the latent feature model by Miller et al. (2009), an Indian buffet process is used, while a hierarchical Gamma process is used in the edge partition model by Zhou (2015).

One major issue with modelling and inferring $K$ is that, in the SBMs with $\mathbf{C}$ and/or $\boldsymbol{\theta}$ *not* integrated out, the number of parameters changes with $K$. As we have seen in (9) and (11), as well as in "Criteria and model selection" section, in some cases they can be marginalised, usually with conjugate priors. Having a likelihood in which the number of parameters is constant to $K$ greatly facilitates the associated inference algorithm, regardless of whether $K$ is modelled or selected by a criterion. Therefore, in recent years, these models have been preferred to the aforementioned models with a nonparametric *process* for $K$. In the case where $K$ is being modelled, estimation is very often carried out using MCMC, and care has to be taken when $K$ grows or shrinks in the algorithm. Both McDaid et al. (2013) and Sanna Passino and Heard (2019) included two moves that allow so. The first move proposes to add or remove an empty group, while the second proposes to merge two groups into one *or* split one into two.

Peixoto (2014a) also allowed $K$ to be modelled, and proposed a progressive way of merging groups. To explain such way, first consider each group $i$ in the original graph, denoted by $\mathcal{G}^{(0)}$, as a *node* in a graph one level above, denoted by $\mathcal{G}^{(1)}$. Similarly, the total number of edges $\mathbf{E}_{ij}^{(0)}$ between groups $i$ and $j$ in $\mathcal{G}^{(0)}$ is now the valued edge $\mathbf{Y}_{ij}^{(1)}$ between nodes $i$ and $j$ in $\mathcal{G}^{(1)}$. Then proposing to merge groups in $\mathcal{G}^{(0)}$ can be seen as proposing to move the membership of a node in $\mathcal{G}^{(1)}$, and this can be carried out in a similar fashion to that described in "Collapsed SBMs and efficient moves" section.

Similar to McDaid et al. (2013), Newman and Reinert (2016) also considered the possibility of empty groups, through modelling $K$ explicitly as a parameter. Specifically, $K$ represents the number of groups that the nodes can potentially occupy, not the number they actually do according to $\mathbf{Z}$. Such possibility of empty groups ensure the joint

posterior of **Z** and $K$, with other parameters integrated out in their Poisson SBM, to be computed correctly:

$$\pi(\mathbf{Z}, K | \mathbf{Y}) = \frac{\pi(K)\pi(\mathbf{Z}|k)\pi(\mathbf{Y}|\mathbf{Z})}{\pi(\mathbf{Y})}.$$

Now, applying an MCMC algorithm with $\pi(\mathbf{Z}, K|\mathbf{Y})$ as the target density will give a (marginal) posterior of $K$, $\pi(K|\mathbf{Y})$.

### Nested model

Peixoto (2017a) presented an example of a synthetic network containing 64 *cliques* of 10 nodes. The nodes are connected with each other within a clique, while there is no edges between nodes in any two different cliques. The results of fitting the microcanonical SBM ("Microcanonical SBM" section) suggested an optimal of 32 groups, each containing two cliques, mean that the model suffers from *underfitting* rather than *overfitting*, due to the fact that the maximum detectable $K$ is (proportional to) $\sqrt{n}$ (Peixoto 2013). Therefore, Peixoto (2014b) proposed a nested SBM, also adopted by Peixoto (2017a), Peixoto (2017b), which can be described using the aforementioned level-one graph $\mathcal{G}^{(1)}$. While $\mathcal{G}^{(0)}$ is characterised by the original SBM, the resulting $\mathcal{G}^{(1)}$ is characterised by another SBM, which results in $\mathcal{G}^{(2)}$, and so on, until there is only one group at the top level. Accompanied by the moves described above for modelling and inferring $K$, they found that this nested SBM managed to overcome the underfitting issue, while discovering a hierarchical structure.

## Comparison

In this section, an overall comparison in two aspects is provided between the works reviewed. The first aspect concerns the performances of the models that all applied to two real-world examples. The second aspect concerns the approaches reviewed in "Clustering approach", "Inference approach" and "Number of groups" section.

### Real-world examples and performances

A classic undirected network studied in the literature is the karate club reported by Zachary (1977). Due to the nature of the study, it was known that the club has been split into two main factions over a disagreement, eventually forming two smaller clubs. Zachary (1977) managed to observe the friendship among $n = 34$ of the members, and compiled the undirected network of $M = 78$ edges, which is plotted in Fig. 3 (there were more members who however did not interact with this biggest component of the network). It has been studied for both community detection methods and SBMs (Karrer and Newman 2011; Yan et al. 2014; Yan 2016; van der Pas and van der Vaart 2018; Lu and Szymanski 2019b). Within the community detection algorithms, Girvan and Newman (2002) and Newman and Girvan (2004) managed to reveal the two true groups, Blondel et al. (2008) stopped short at four groups. But this should not be taken as an indication of superiority automatically, as, in fact, there is also no unanimous agreement within the SBMs. On one hand, the models mostly agreed on the optimal $K$, as Decelle et al. (2011) found $K = 2$ using their belief propagation algorithm and the observed data log-likelihood criterion, and Newman and Reinert (2016) found that the posterior $\pi(K|\mathbf{Y})$ in their DC-SBM is indeed maximised at $K = 2$. On the other hand, while the DC-SBM (Karrer and

**Fig. 3** The karate club network by Zachary (1977)
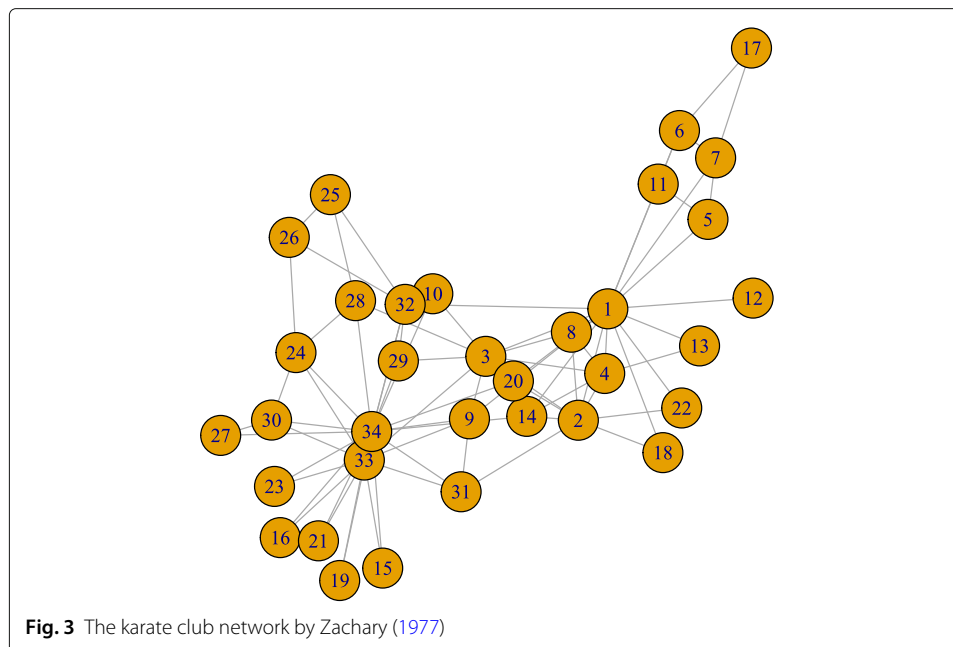
Newman 2011) revealed the two known factions that the original SBM failed, the likelihood ratio test by Yan et al. (2014) suggested that there were not enough evidence to suggest that the network was generated by DC-SBM (against the original SBM), even when the inhomogeneous degree distribution prompted the degree correction in the first place. Furthermore, Yan (2016) found that using the ICL criterion selected $K = 1$ under the DC-SBM, meaning that any blocking leads to overfitting, while $K = 4$ was selected under the original SBM, with different groups corresponding to different levels of node degrees.

Another example is the directed network of political blogs by Adamic and Glance (2005) during the 2004 U.S. Presidential Election. While it contains 1490 blogs in total, usually only the giant component of 1222 blogs (nodes) with 19089 is analysed. Furthermore, the blogs were labeled as either "liberal" or "conservative" by Adamic and Glance (2005) according to their political leanings, providing some sort of ground truth. By forcing $K = 2$, the DC-SBM by Karrer and Newman (2011) again showed a high agreement with the ground truth, while the original SBM did not. The evidence for the former is supported by the likelihood ratio test by Yan et al. (2014). Yan (2016) presented a mixed picture in which the DC-SBM always dominated the original SBM, even more so at smaller $K$, but the ICL criterion actually increased with $K$, thus potentially requiring some penalty. While Wang and Bickel (2017) selected $K = 4$ using their penalised log-likelihood, closer investigation revealed that one ground-truth group matched well with one inferred group, while the other ground-truth group is split into three smaller inferred groups. However, Hu et al. (2019) found that, at a certain value of the tuning parameter $\lambda$ ("Criteria and model selection" section) the penalised log-likelihood selected $K = 1$, therefore arguing for their corrected BIC that has a heavier penalty. Finally, the nested SBM (Peixoto 2017a) managed to cluster the nodes beneath the two main political factions, and discovered 20 groups in the lowest-level SBM according to the DC-SBM.

While there is an agreement over the big picture, the models usually differ in the fine details.

### Approaches

Several aspects of SBMs and related models have been considered so far, namely the type of graph, whether the model is an SBM or not, the inference approach, the clustering approach, the number of groups, and whether there is longitudinal modelling. However, these aspects are not completely independent of each other. For example, opting for a latent feature model (Miller et al. 2009; Mørup et al. 2011) not only deviates from an SBM, but also increases the computation complexity because the number of latent variable combination increases from $K^n$ to $2^{Kn}$, thus in turn prompting for an efficient and scalable inference algorithm. Another example is that a soft clustering approach influences partly how $K$ is being modelled or selected. The models considered here either opted for modelling by a Dirichlet process (Kurihara et al. 2006; Kim et al. 2013; Fan et al. 2015) or selecting by a criterion (Airoldi et al. 2008; Fu et al. 2009; Xing et al. 2010; Gopalan et al. 2012; Li et al. 2016). Also, the added computational complexity prompted (Li et al. 2016) to propose an inference algorithm which exploits the sparity of graphs. Finally, using a variational approach more often results in $K$ being selected by a criterion (Airoldi et al. 2008; Fu et al. 2009; Xing et al. 2010; Gopalan et al. 2012; Latouche et al. 2012; Aicher et al. 2015; Matias et al. 2015; Hayashi et al. 2016; Matias and Miele 2017) than not (Kurihara et al. 2006; Kim et al. 2013; Vu et al. 2013). Note that these are not intended as definitive arguments but to highlight that the modelling, inference, and clustering approaches, and the issue with the number of groups, are quite interconnected. For a comprehensive comparison, please see Tables 1 and 2 for models with and without longitudinal modelling, respectively.

### SBM with longitudinal modelling

The SBMs and related models introduced so far assume that the graph is observed at one instant, which can be regarded as a cross-section of a graph that is evolving over time. If temporal information of the interactions is available, or the graph is observed at multiple instants, longitudinal or dynamic models for graphs can be applied. Therefore, articles that incoporate longitudinal modelling will be reviewed in this section, providing a partial answer to question 3 in "Introduction" section.

In the dynamic model by Fu et al. (2009) and Xing et al. (2010), which are direct extensions of the MMSBM by Airoldi et al. (2008), the membership probabilities of node $p$ is now indexed by time $t$, denoted by $\boldsymbol{\theta}_p^{(t)}$, and is dependent on $\boldsymbol{\theta}_p^{(t-1)}$ via a state space model. Similarly, the block matrix now becomes $\mathbf{C}^{(t)}$, and evolves over time according to a separate and independent state space model. The latent pairwise group memberships at time $t$, $\mathbf{Z}^{(t)}$, is then generated by $\Theta^{(t)} = \left( \boldsymbol{\theta}_1^{(t)} \ \boldsymbol{\theta}_2^{(t)} \ \cdots \ \boldsymbol{\theta}_n^{(t)} \right)^T$. Finally, the observed graph $\mathbf{Y}^{(t)}$ is assumed to arise from an MMSBM (Airoldi et al. 2008) with parameters $\mathbf{C}^{(t)}$ and latent variables $\mathbf{Z}^{(t)}$.

Fan et al. (2015) proposed two dynamic models, by combining the Dirichlet process and the MMSBM. In their mixture time variant (MTV) model, at each instant the group memberships $\mathbf{Z}^{(t)}$ are drawn according to the membership probability vector $\Theta^{(t)}$, elements of which arises from a Dirichlet process. The process parameters in turn depend

**Table 1** Models for graphs without longitudinal or topic modelling

| Article | Type of graph | | SBM | Inference | Clustering | Number of groups |
|---|---|---|---|---|---|---|
| Snijders and Nowicki (1997) | Binary | Undirected | Yes | MLE/EM/MCMC | Hard | Fixed (2) |
| Nowicki and Snijders (2001) | Either | Either | Yes | MCMC | Hard | Criterion (Information) |
| Tallberg (2005) | Discrete | Directed | Yes | MCMC | Hard | Fixed (3) |
| Kurihara et al. (2006) | Valued | Directed | Yes | Variational | Soft | Dirichlet process |
| Airoldi et al. (2008) | Binary | Directed | Yes | Variational | Soft | Criterion (BIC) |
| Daudin et al. (2008) | Binary | Undirected | Yes | Variational | Hard | Criterion (ICL) |
| Miller et al. (2009) | Binary | Directed | No | MCMC | Hard | Indian buffet process |
| Decelle et al. (2011) | Binary | Directed | Yes | Belief propagation | Hard | Criterion (Observed data log-like.) |
| Karrer and Newman (2011) | Valued | Undirected | Yes | Greedy algo.+MLE | Hard | Fixed (2) |
| Mørup et al. (2011) | Binary | Undirected | No | MCMC+HMC | Hard | Dirichlet process |
| Gopalan et al. (2012) | Binary | Undirected | Yes | Variational | Soft | Criterion (Perplexity) |
| Latouche et al. (2012) | Binary | Undirected | Yes | Variational | Hard | Criterion (Marginal log-like.) |
| Palla et al. (2012) | Binary | Undirected | No | MCMC | Hard | Dirichlet process |
| Kim et al. (2013) | Binary | Undirected | Yes | Variational | Soft | Dirichlet process |
| McDaid et al. (2013) | Binary | Either | Yes | MCMC | Hard | Estimated |
| Vu et al. (2013) | Discrete | Either | Yes | Variational | Hard | Fixed (5, 20) |
| Yan et al. (2014) | Valued | Undirected | Yes | EM | Hard | Criterion (Observed data log-like.) |
| Aicher et al. (2015) | Valued | Directed | Yes | Variational | Hard | Criterion (Marginal log-like.) |
| Côme and Latouche (2015) | Binary | Directed | Yes | Greedy algo. | Hard | Criterion (ICL) |
| Zhou (2015) | Valued | Undirected | No | MCMC | Hard | Hier. Gamma process |
| Godoy-Lorite et al. (2016) | Valued | Bipartite | Yes | Variational | Soft | Fixed (10) |
| Hayashi et al. (2016) | Binary | Directed | Yes | Variational | Hard | Criterion ($F^2$IC) |
| Li et al. (2016) | Binary | Undirected | Yes | MCMC | Soft | Criterion (Perplexity) |
| Newman and Reinert (2016) | Binary | Undirected | Yes | MCMC | Hard | Estimated |
| Vallès-Català et al. (2016) | Binary | Undirected | Yes | Unknown | Hard | Unknown |
| Yan (2016) | Binary | Undirected | Yes | Unknown | Hard | Criterion (BIC) |
| Peixoto (2015b); Peixoto (2017a); Peixoto (2017b); Peixoto (2018a) | Either | Either | Yes | MCMC | Either | Estimated |
| Ranciati et al. (2017) | Binary | Bipartite | No | MCMC | Soft | Criterion (DIC) |
| Wang and Bickel (2017) | Binary | Undirected | Yes | Variational | Hard | Criterion (Penalised log-like.) |
| Gerlach et al. (2018) | Valued | Bipartite | Yes | MCMC | Hard | Estimated |
| van der Pas and van der Vaart (2018) | Binary | Undirected | Yes | Unknown | Hard | Fixed |
| Hu et al. (2019) | Binary | Undirected | Yes | Unknown | Hard | Criterion (BIC, corrected) |
| Lu and Szymanski (2019b) | Valued | Undirected | Yes | MCMC | Hard | Fixed (2) |
| Sanna Passino and Heard (2019) | Binary | Undirected | Yes | MCMC | Hard | Estimated |
| Stanley et al. (2019) | Binary | Undirected | Yes | Variational | Hard | Criterion (ICL) |

**Table 2** Stochastic block models with longitudinal modelling

| Article | Type of graph | | Inference | Clustering | Number of groups |
|---|---|---|---|---|---|
| Fu et al. (2009) | Binary | Directed | Variational | Soft | Criterion (BIC) |
| Xing et al. (2010) | Binary | Directed | Variational | Soft | Criterion (BIC) |
| Tang and Yang (2011); Tang and Yang (2014) | Binary | Directed | MCMC | Hard | Dirichlet process |
| Yang et al. (2011) | Either | Undirected | MCMC | Hard | Fixed (2,2,3) |
| DuBois et al. (2013) | Valued | Directed | MCMC | Hard | Criterion (Perplexity) |
| Xu and Hero III (2013) | Binary | Directed | Greedy algo.+SMC | Hard | Fixed (7) |
| Fan et al. (2015) | Binary | Directed | MCMC | Soft | Dirichlet process |
| Matias et al. (2015) | Binary | Undirected | Variational | Hard | Criterion (ICL) |
| Peixoto (2015a) | Either | Either | MCMC | Either | Estimated |
| Matias and Miele (2017) | Either | Either | Variational | Hard | Criterion (ICL) |
| Zhang et al. (2017) | Either | Undirected | Greedy algo. | Hard | Fixed |
| Ludkin et al. (2018) | Binary | Undirected | MCMC | Hard | Criterion (Elbow plot) |

on the history of group memberships $\{\mathbf{Z}^{(1)}, \mathbf{Z}^{(2)}, \ldots, \mathbf{Z}^{(t-1)}\}$. In the mixture time invariant (MTI) model, at each instant the group memberships $\mathbf{Z}_p^{(t)}$ depends on $K$ time-invariant $\boldsymbol{\theta}_p^{(i)}(i = 1, 2, \ldots, K)$, also from a Dirichlet process, *and* the previous group memberships $\mathbf{Z}^{(t-1)}$. The graph $\mathbf{Y}^{(t)}$ is then generated from an MMSBM with $\mathbf{Z}^{(t)}$, and a universal block matrix $\mathbf{C}$. This means that the historical group memberships influence the distribution of the current ones, in the MTV and MTI models, through the Dirichlet process parameters and the group-specific (and time-invariant) membership probabilities, respectively.

Tang and Yang (2011); Tang and Yang (2014) adopted a similar idea in their dynamic SBM with temporal Dirichlet process. As hard clustering is used here, there is one membership vector $\boldsymbol{\theta}^{(t)}$ for all nodes at each instant. However, this $\boldsymbol{\theta}^{(t)}$ is dependent on the Dirichlet process parameter *and* the group memberships $\mathbf{Z}^{(t-1)}$ at the previous instant, so that the collection $\{\boldsymbol{\theta}^{(t)}\}$ arises from what they called the recurrent Chinese restaurant process. The ingredients for generating the graph $\mathbf{Y}^{(t)}$ include the block matrix $\mathbf{C}$, which is constant over time, and the group memberships $\mathbf{Z}^{(t)}$, which are generated by $\boldsymbol{\theta}^{(t)}$ in turn.

Yang et al. (2011) considered a simpler evolution mechanism of the group memberships in their dynamic SBM. A Markov chain is assumed for $\mathbf{Z}^{(t)}$, which means, for node $p$, $\mathbf{Z}_p^{(t-1)}$ moves to $\mathbf{Z}_p^{(t)}$ according to a transition matrix, which remains unchanged over time. Also assumed constant over time are the model parameters $\boldsymbol{\theta}$ and $\mathbf{C}$. Matias and Miele (2017) proposed a similar model, except that the graph is allowed to be valued, in which case each element in $\mathbf{C}$ does not necessarily mean the edge probability. They noted that allowing the group memberships *and* the block matrix to vary over time will lead to identifiability and label-switching issues, therefore fixing $\mathbf{C}$ to be constant over time.

Xu and Hero III (2013) used a dynamic SBM quite different to the ones introduced so far. They considered the observed density of edges between two groups to be a noisy observation of a dynamic system, in which the corresponding element in the block matrix is the state. While $\mathbf{C}^{(t)}$ is modelled by a state space model, there is none specified for the time-specific group memberships $\mathbf{Z}^{(t)}$.

Another distinctive model is the autoregressive SBM (Ludkin et al. 2018), in which the group membership $\mathbf{Z}_p^{(t)}$ follows a continuous-time Markov chain (CTMC), meaning that node $p$ spends an exponentially distributed time in a group, before moving to another group chosen uniformly at random from the remaining groups. If dyad

$(p, q)$ *belongs completely* to one group, that is, both $p$ and $q$ belong to the same group, the presence or absence of an edge over time is modelled by a separate CTMC, the transition rates of which are universal to all dyads that belong completely to the same group as $(p, q)$. For all the remaining dyads where the group memberships do not coincide, there is one extra set of transition rates governing the independent edge process for each dyad. Essentially, there are $n(n-1)/2$ CTMCs modelling the dyads, the parameters of which are determined, as always in SBMs, by the group memberships.

Whilst also using a CTMC in their dynamic model, Zhang et al. (2017) placed a it on the edge existence/absence $\mathbf{Y}_{pq}^{(t)}$, instead of the group membership $\mathbf{Z}_p$. Using the DC-SBM (7), the rate that edges are removed depends on the group memberships of $p$ and $q$, while the rate that edges are added is the product of the edge-removal rate and $\phi_p \phi_q \mathbf{Z}_p^T \mathbf{C} \mathbf{Z}_q$, which is the same term used in the static model (7). Using the Markov property of a CTMC, they worked out the probabilities of the graph observed at one time point, *given* the graph at the previous time point, thus the whole likelihood. For inference, *conditional on* $\mathbf{Z}$, the MLEs of the other parameters can be obtained analytically (in terms of $\mathbf{Z}$). These estimates are then plugged in to the original log-likelihood to arrive at the profile log-likelihood, which is then maximised using a greedy algorithm, to obtain the MLEs of $\mathbf{Z}$.

Peixoto (2015a) proposed two versions of modelling longitudinal or temporal networks, in which group memberships are the same across layers or time, based on the MDL approach (Peixoto 2017a). In the first, the total numbers of edges between the groups across *all* layers are viewed as a *collapsed* graph that arises from one SBM (according to, for example, (3) or (8)). The edges between groups $i$ and $j$ are then assumed to be uniformly distributed, conditioned only on the number of edges between $i$ and $j$ *in each layer*. In the second version, each layer is generated from an independent SBM, but nodes are allowed to belong only to a subset of the layers. A node not belonging to a layer means that the node is forbidden to have edges in that layer. Equivalence can be established between special cases of the two versions for non-degree-corrected SBMs, but their degree-corrected counterparts are in general not equivalent. This is because degree variability across layers is allowed for the second version but not the first. For temporal modelling, unlike the models introduced above, there are no temporal dynamics explicitly specified. However, Peixoto (2015a) suggested binning the times so that each bin can be viewed as a layer, within which the large-scale network structure does not change significantly. Depending on the version chosen and whether degree correction is incorporated, how the edges can be assigned to the nodes at different layers is flexibly regulated. Finally, selecting the most appropriate time binning, alongside the choice between the two versions and that of degree correction or not, can be aided by model selection through the use of the MDL.

The models introduced so far model the evolution of a graph, usually a binary one, and observed at different instants. They are however not suitable for recurrent interaction events in a network, of which the times of interaction are random in nature. In the aforementioned DuBois et al. (2013), the interactions for the dyad $(p, q)$ arise from a Poisson process with intensity $\exp(\mathbf{Z}_p^T \mathbf{C} \mathbf{Z}_q)$. The intensity can be extended to depend on not only the group memberships only but also on the historical interactions. Specifically, it is the dot product of a vector summarising the

interactions of the dyad and $\mathbf{Z}_p^T\mathbf{C}\mathbf{Z}_q$, which is now allowed to be vector-valued (essentially making $\mathbf{C}$ a 3-dimensional array). However, due to such a generalisation, it is misleading to continue calling such model a Poisson process, because the piecewise constant intensity depends on the history of the process, thus making it self-exciting in nature.

Matias et al. (2015) formulated a similar model using conditional Poisson processes. The block matrix $\mathbf{C}(t)$ is now a $K \times K$ matrix of intensity *functions*, not probabilities or scalar parameters. *Conditional on $\mathbf{Z}_p$ and $\mathbf{Z}_q$*, which are assumed to remain unchanged over the course of time, the interactions of the dyad $(p,q)$ follows a nonhomogeneous Poisson process with intensity $\mathbf{Z}_p^T\mathbf{C}(t)\mathbf{Z}_q$. Similar to Matias and Miele (2017), identifiability and label-switching issues were also discussed.

It should be noted that having both the temporal information in the data and a longitudinal/dynamic SBM does not guarantee that the actual groups will be discovered. Ghasemian et al. (2016) studied a dynamic SBM, and derived a threshold as a function of the rate of change and the strength of the groups/communities. They found that, below that threshold and empirically, no efficient inference algorithms can identify the groups better than chance.

### Topic models

In this section, we briefly introduce the general form of topic models, in particular latent Dirichlet allocation, to prepare for "SBM with topic modelling" section, where they are incoporated in SBMs. While topic models are not the main focus of this review, we will discuss aspects including the inference approach, dealing with the number of topics, and longitudinal modelling, which have been covered in "Inference approach, "Number of groups", "SBM with longitudinal modelling" section for SBMs, respectively.

The main goal of topic modelling is to cluster a collection of *documents* into different *topics*. Each latent topic is represented by its distribution over the words that appear in the documents, with such distribution usually being visualised by wordclouds. The roles of the documents, the topics, and the words are then analogous to those of the nodes, the groups, and the edges, respectively. One major difference is that a document is usually assumed to be independent of other documents *apriori*, whereas a node in a graph is defined by its interactions, or the lack thereof, with others. Another difference is that soft clustering is the dominating approach in topic modelling, meaning that each document usually has non-binary weights over multiple topics. For example, an article in genetics may be about biology, chemistry, statistics, and medicine, for 40%, 30%, 20% and 10%, respectively. This is in contrast to the (usual) hard clustering approach in SBMs.

We first introduce some terminology and notation, which may look different to that in the literature, but is intended to align with what we have introduced for SBMs. Assume that the there are $m$ documents, and $n$ distinct words in total, denoted by $\mathbf{V} := (\mathbf{V}_1\,\mathbf{V}_2\,\cdots\,\mathbf{V}_n)^T$. In document $p = 1, 2, \ldots, m$, there are $N_p$ words. The $q$-th word is represented by an $n$-vector $\mathbf{W}_{pq} := \left(\mathbf{W}_{pq1}\,\mathbf{W}_{pq2}\,\ldots\,\mathbf{W}_{pqn}\right)^T$, in which one element is 1, the rest 0. If $\mathbf{W}_{pqk} = 1$, this means the $k$-th word in $\mathbf{V}$ is the one used as the $q$-th word in document $p$. We also define the $N_p \times n$ matrix $\mathbf{W}_p := \left(\mathbf{W}_{p1}\,\mathbf{W}_{p2}\,\cdots\,\mathbf{W}_{pN_p}\right)^T$, and $\mathbf{W} := \{\mathbf{W}_1\,\mathbf{W}_2\,\cdots\,\mathbf{W}_m\}$, which essentially is the whole of the data. Finally, an $m \times n$ matrix $\mathbf{M}$, called the document-word frequency matrix, is defined. Each element

$\mathbf{M}_{pk}$ ($p = 1, 2, \ldots, m; k = 1, 2, \ldots, n$) represents the frequency of the word $\mathbf{V}_k$ in the $p$-th document.

A common assumption in topic modelling is that there are $K$ latent *topics*. Associated with the $i$-th topic is an $n$-vector $\boldsymbol{\phi}_i := \left( \boldsymbol{\phi}_{i1} \, \boldsymbol{\phi}_{i2} \, \cdots \, \boldsymbol{\phi}_{in} \right)^T$, subject to the constraint $\boldsymbol{\phi}_i^T \mathbf{1}_n = 1$, which represents the distribution of the vocabulary in this topic. The sequence $\{\boldsymbol{\phi}_i\}$ ($i = 1, 2, \ldots, K$) is assumed to be independent and identically distributed (i.i.d.) according to the Dirichlet($\beta \mathbf{1}_K$) distribution. Collectively, we write $\Phi := \left( \boldsymbol{\phi}_1 \, \boldsymbol{\phi}_2 \, \cdots \, \boldsymbol{\phi}_K \right)^T$, which is a $K \times n$ matrix.

In a similar fashion, associated with the $p$-th document is a $K$-vector $\boldsymbol{\theta}_p$, subject to the constraint $\boldsymbol{\theta}_p^T \mathbf{1}_K = 1$, which represents the distribution of the $K$ topics for this document, or its mixed membership in the topics. The sequence $\{\boldsymbol{\theta}_p\}$ ($p = 1, 2, \ldots, m$) is assumed to be i.i.d. according to the Dirichlet($\alpha \mathbf{1}_K$) distribution. Collectively, we write $\Theta := \left( \boldsymbol{\theta}_1 \, \boldsymbol{\theta}_2 \, \cdots \, \boldsymbol{\theta}_m \right)^T$, which is an $m \times K$ matrix. This is similar to the definitions in "Clustering approach" section.

The main difference between various topic models is the generating mechanism of the words in the documents. The principle in latent Dirichlet allocation (LDA) by Blei et al. (2003) is that a document can belong to different topics when generating each word. Specifically, for document $p$, associated with the $q$-th word ($q = 1, 2, \ldots, N_p$) is a $K$-vector latent variable, denoted by $\mathbf{Z}_{pq}$. Only one element of $\mathbf{Z}_{pq}$ is 1, representing the topic the document belongs to *for this particular word*, the rest of which 0. (This is similar to the latent group membership of a node $p$ for its particular interaction with $q$ in MMSBM (Airoldi et al. 2008).) Collectively, we write $\mathbf{Z}_p := \left( \mathbf{Z}_{p1} \, \mathbf{Z}_{p2} \, \cdots \, \mathbf{Z}_{pN_p} \right)^T$, which is an $N_p \times K$ matrix, and $\mathbf{Z} := \{ \mathbf{Z}_1, \mathbf{Z}_2, \ldots, \mathbf{Z}_n \}$, which is a sequence of matrices as well as a collection of all latent variables.

Now, with the data $\mathbf{W}$, the latent variables $\mathbf{Z}$, the parameters $\Theta$ and $\Phi$, the likelihood can be computed in the usual manner, and inference carried out. As with MMSBM ("Clustering approach" section), the main goal of inference is for $\Theta$ (and $\Phi$) but not $\mathbf{Z}$. Instead of going through the derivations of the posterior in detail, we shall mention works which have worked on various aspects in "Inference approach", "Number of groups" and "SBM with longitudinal modelling" sections.

### Inference approach

In the usual approach which is Bayesian inference, in addition to the Dirichlet distribution assumptions made for $\Phi$ and $\Theta$, once the priors are assigned to $\alpha$ and $\beta$, the joint posterior of $\mathbf{Z}$, $\Phi$, $\Theta$, $\alpha$ and $\beta$ can be derived. Unlike in SBMs, even with the soft clustering approach, the computational complexity grows linearly, but not quadratically, with the number of documents $m$.

Similar to what has been described in the beginning of "Inference approach" section, because of the use of conjugate priors for $\Phi$, $\Theta$, $\alpha$ and $\beta$, for algorithmic simplicity a Gibbs sampler can be derived, in which the parameters and the latent variables can be updated via individual Gibbs steps. However, Griffiths and Steyvers (2004) were not interested in $\Phi$ and $\Theta$ and therefore integrated them out, to obtain a *collapsed* Gibbs sampler. Furthermore, as they fixed the values for $\alpha$ and $\beta$, only the latent groups $\mathbf{Z}$ are to be inferred. This is similar to collapsing the hard clustering SBMs in, for example, (McDaid et al. 2013), (Côme and Latouche 2015), Peixoto (2017a) and Peixoto (2017b), although the parameters $\mathbf{C}$ in the soft clustering MMSBM (Airoldi et al. 2008) are not

being integrated out. Other uses of MCMC algorithms include Teh et al. (2005), Ren et al. (2008), Papaspiliopoulos and Roberts (2008), Ahmed and Xing (2010) and (Griffiths and Ghahramani 2011). However, such algorithms are more useful when the number of topics $K$ is being modelled ("Number of topics" section), rather than selected by a criterion, by processes such as the hierarchical Dirichlet process and the Indian buffet process.

Blei et al. (2003), along with Wang et al. (2011) used variational inference, the general formulation of which is described in "Variational methods" section. Blei and Lafferty (2006) combined variational inference and a Kalman filter for their dynamic topic model, in which a state space model is incorporated, to be explained in "Longitudinal modelling" section. Arun et al. (2010) focussed on finding the optimal number of topics and did not explicitly state their inference approach.

### Number of topics

Using the same notation $K$ as in "Stochastic block models" section is because the number of topics in a topic model is analogous to the number of groups in an SBM. Therefore, there also exists the issue of whether $K$ should be fixed, selected by a criterion, or modelled. Of the topic models reviewed here, only Blei and Lafferty (2006) used a fixed number of topics ($K = 20$). Blei et al. (2003) used the perplexity, which is the likelihood of held-out test data, to determine $K$. This is (almost) the same as the definition of perplexity in SBMs (Gopalan et al. 2012; DuBois et al. 2013; Li et al. 2016). Griffiths and Steyvers (2004) selected $K$ by the marginal log-likelihood, which is similar to Latouche et al. (2012) for SBMs, although they have not provided the derivations of this quantity $\log \pi (\mathbf{W})$, or approximations thereof.

Arun et al. (2010) proposed an empirical measure to find the optimal $K$. They started with viewing that LDA essentially "factorises" the document-word frequency matrix $\mathbf{M}$ into the two stochastic matrices $\mathbf{\Phi}$ and $\Theta$. (Of course, this does not mean that $\mathbf{M} = \Theta\mathbf{\Phi}$, however.) By defining the vector of document lengths $\boldsymbol{N} := (N_1 \ N_2 \ \cdots \ N_m)^T$, they gave the expression of the criterion for selecting $K$, as the symmetric K-L divergence between the distribution of the singular values of $\mathbf{\Phi}$, and the distribution obtained by normalising $\boldsymbol{N}^T\Theta$.

#### *Hierarchical Dirichlet process*

The Dirichlet process (DP) and its hierarchical version have to be introduced before incorporating into topic models. As it is only used in some of the models reviewed, we focus on the ideas behind these *nonparametric Bayesian* processes, rather than their technical details.

The topic model introduced in the beginning of this section assumes that, for each document, there is a same number ($K$) of topics to choose from, and the membership probabilities $\boldsymbol{\theta}_p$ ($p = 1, 2, \ldots, m$) form a $K$-vector which comes from a Dirichlet($\alpha\mathbf{1}_K$) *distribution*. Dirichlet process takes this one step further by not requiring $K$ to be prespecified. While the elements of $\boldsymbol{\theta}_p$ drawn from a DP still sum to 1, $K$ varies between i.i.d. samples and is implied by the length of $\boldsymbol{\theta}_p$. This means, for example, $\boldsymbol{\theta}_p = (0.2 \ 0.5 \ 0.3)^T$ and $\boldsymbol{\theta}_p = (0.15 \ 0.45 \ 0.35 \ 0.05)^T$ can be two independent samples from the same DP, implying $K = 3$ and $K = 4$, respectively.

To take the DP one step further, Teh et al. (2005) assumed that $\boldsymbol{\theta}_p$ comes from a DP, the parameters of which come from another DP. Introducing this structure in their hierarchical Dirichlet process (HDP) leads to a clustering effect. In the context of topic models, the topics generated for each document are uncorrelated in a DP, while the same set of topics is shared across all documents in an HDP.

This description of the DP and HDP is an informal one, and has omitted the technical details of actual sampling from a DP; see, for example, Teh et al. (2005), Ren et al. (2008), and Ahmed and Xing (2010). What is to be highlighted here is that $K$ arises naturally when sampling from a DP or HDP, and therefore does not need to be pre-specified. As these nonparametric Bayesian processes concerns the generation of the mixing vectors $\boldsymbol{\theta}_p$ and is independent of the topic models, it can be incorporated in models that require these membership probabilities, hence its use in the SBMs discussed in "Modelling" section.

It has been argued in "Inference approach" section and shown by, for example, Griffiths and Steyvers (2004) that a Gibbs sampler can be derived for the parameters and the latent variables **Z** of the topic model when $K$ is fixed. Teh et al. (2005) have shown that a Gibbs sampler is also possible under the HDP formulation, hence the preference to using MCMC as the inference algorithm too, in Ren et al. (2008) and Ahmed and Xing (2010). While Papaspiliopoulos and Roberts (2008) proposed two MCMC algorithms to improve the then existing ones for HDP models in general, Wang et al. (2011) pointed out the limitation that such algorithms require multiple passes through all the data and are therefore not very scalable. They proposed a variational infernce algorithm as an alternative.

So far, for a topic model or SBM, incorporating an HDP means that the quantity $K$ corresponds to is potentially infinite and to be modelled. The equivalent for latent feature models exists and is termed the Indian buffet process, and again a Gibbs sampler for such models is possible (Miller et al. 2009). For a detailed introduction and an extensive review, see Griffiths and Ghahramani (2011).

### Longitudinal modelling

As temporal information is sometimes available for text data, such as academic articles, longitudinal or dynamic models can be incorporated in a similar way as they are for SBMs or other models for graphs. Blei and Lafferty (2006) proposed a dynamic topic model, in which the distribution of vocabulary for the $i$-th topic is now indexed by time $t$, denoted by $\boldsymbol{\phi}_i^{(t)}$, and is dependent on $\boldsymbol{\phi}_i^{(t-1)}$ via a state space model. The incorporation of a state space model is similar to that in Fu et al. (2009); Xing et al. (2010) and Xu and Hero III (2013), mentioned in "SBM with longitudinal modelling" section. The main difference is the generation the membership probabilities, as once the $p$-th document is generated from $\boldsymbol{\theta}_p$ there is no need to evolve its distribution over the topics. Instead, there is a "universal" vector of membership probabilities, denoted by $\boldsymbol{\psi}^{(t)}$, and depends on $\boldsymbol{\psi}^{(t-1)}$ via a separate state space model. Assuming that, without loss of generality, the $p$-th document is to occur at time $t$, $\boldsymbol{\theta}_p$ is generated from $\boldsymbol{\psi}^{(t)}$ with random noise. Words of the document are then generated in the usual way according to $\boldsymbol{\theta}_p$ and $\boldsymbol{\phi}^{(t)}$.

As mentioned before, in a non-dynamic model that models $K$, the memberships $\boldsymbol{\theta}_p$ arises from an HDP. Ren et al. (2008) introduced temporal dependency in their dynamic model, by assuming that the documents occurred sequentially, and that $\boldsymbol{\theta}_p$ depends on both $\boldsymbol{\theta}_{p-1}$ and an HDP which represents innovation. Ahmed and Xing (2010) modified the HDP in a slightly different way. The membership probabilities now come from,

loosely speaking, an evolving DP, which depends on the process at the previous time point through a state space model. They argued that their dynamic model allows topics to be born and die at any instant.

## SBM with topic modelling

In this section articles that combine a model for graphs and a topic model are reviewed. Similar to "Topic models" section, different aspects related to modelling and inference approaches are discussed, thus providing answers to questions 5 and 6 in Introduction section.

Before looking at SBMs with topic modelling for graph *and* textual data, it is useful to look at a recently proposed SBM *for* textual data by Gerlach et al. (2018). Instead of probabilitistically factorising the document-word frequency matrix $M$ ("Topic models" section), they treated the relations between the words and the documents as a bipartite graph, in which both the words and the documents are the nodes. By doing so, the SBMs reviewed before can be direct applied to the graph. In particular, they fit a nested SBM (Peixoto 2014b), thus constructing a hierarchical structure and allowing the number of groups to be inferred. The proposed model was found to perform better than the traditional approach of LDA. It can be compared with the model by Godoy-Lorite et al. (2016) for recommender systems, who also applied an SBM for seemingly non-relational data that can however be represented as graphs.

### Type of data and modelling approach

In the articles reviewed in this section, the modelling approach is largely influenced by the type of data available, in particular whether the textual information is for the edges or nodes in the graph. The availability of such textual information is in turn determined by the nature of the data set itself. Therefore, we will review these two aspects together.

#### *Textual edges*

One famous example of data that contains both network and text information is the Enron Corpus, a large database of over 0.6 million emails by 158 employees of the Enron Corporation before the collapse of the company in 2001. The nature of an email network leads to a directed and valued graph, in which the nodes and edges are the employees and the email exchanges, respectively, with the latter of which the texts are associated. In terms of the models reviewed in "Stochastic block models" section, it has been studied by Fu et al. (2009); Xing et al. (2010); Gopalan et al. (2012); DuBois et al. (2013); Xu and Hero III (2013) and Matias et al. (2015). Models in which both the graph and the texts are studied include Zhou et al. (2006); McCallum et al. (2007); Pathak et al. (2008); Sachan et al. (2012); Bouveyron et al. (2016) and Corneli et al. (2018), which will be discussed individually.

Zhou et al. (2006) proposed two models in which the graph is not modelled explicitly. Instead, each email, or communication document in general, is generated by an extension of LDA. In the first model, the membership probabilities $\boldsymbol{\theta}$ is not associated with the document, but with the users involved with this document (for example, sender and receiver of an email). Then, the latent topic variable for word $q$ in document $p$, denoted by $\mathbf{Z}_{pq}$ as in "Topic models" section, is generated according to the membership probabilities of the users. There is an additional layer in which users come from different groups

(communities), but it is not that each user has a mixed membership over the groups, but that for each group there is a distribution of users representing their participation. In the second model, associated with each group is the membership probabilities $\theta$ of the topics, and associated with each topic is a distribution of the *users*. This time, it is not made clear how a word in the document is generated from the user, who is in turn generated by the topic.

Similar to those by Zhou et al. (2006) is the author-recipient-topic model by McCallum et al. (2007), in which the membership probabilities $\theta$ are now specific to each author-recipient pair. For document $p$, the author and the set of recipients are treated as given. For word $q$ in this document, a recipient is selected at random uniformly, and the latent topic $\mathbf{Z}_{pq}$ follows a multinomial distribution according to the aforementioned pair-specific membership probabilities. Finally, the word is generated, as usual, according to the word distribution $\phi$ over the selected latent topic. Pathak et al. (2008) extended the author-recipient-topic model, by adding the group element of the authors and recipients. In particular, chosen at random uniformly is not the recipient but the group. Then, similar to Zhou et al. (2006), the authors and recipients are selected according to a group-specific distribution of users.

Sachan et al. (2012) introduced a topic user community model, which is different to those by Zhou et al. (2006) despite the similarities in the model names. For each sender, there is a group distribution, and for each sender-group pair, there is a distribution over the topics. There is one latent group for each document by the sender, and the latent topic variables for individual words are generated according to the aforementioned sender-group-specific probabilities. To complete the model specification, the recipients are chosen by a group-specific distribution of the latent group.

Bouveyron et al. (2016) combined the SBM and LDA to form the stochastic topic block model (STBM). As in the SBM introduced in "Stochastic block models" section, the groups of nodes $p$ and $q$ arise from a multinomial distribution, and the edge variable $\mathbf{Y}_{pq}$ follows a Bernoulli distribution with parameter $\mathbf{Z}_p^T\mathbf{C}\mathbf{Z}_q$, where $\mathbf{Z}_p$ and $\mathbf{Z}_q$ represent the memberships in the *groups*, not the topics. Now, for each pair of groups $i$ and $j$, there is a specific vector $\theta_{ij}$ representing the memberships in the *topics*, the collection of which, still denoted by $\Theta$, is a 3-dimensional array. Then, the latent topic variables for individual words of a document from $p$ to $q$ follow a multinomial distribution with parameters $\mathbf{Z}_p^T\Theta\mathbf{Z}_q$, *conditional on* $\mathbf{Y}_{pq} = 1$. This means that the latent groups influence both the dyad and potentially the words in the document, if an edge exists.

Corneli et al. (2018) extended the STBM by incorporating a dynamic component, and is the only work reviewed in this section with longitudinal modelling, possibly due to complexity of combining a model for a graph and a topic model. Instead of conditioning on $\mathbf{Y}_{pq} = 1$, the occurrences of the documents from $p$ to $q$ arise from a nonhomogeneous Poisson process, according to the collection of group-pair-specific intensity functions and the latent groups $\mathbf{Z}_p$ and $\mathbf{Z}_q$. Furthermore, these intensity functions are piecewise constant, that is, constant within each of the time clusters, which are also latent variables and have to be inferred. Within each time cluster and conditioned on the existence of a document, the latent topic variables and individual words are again generated in the same way as in Bouveyron et al. (2016). Essentially, Corneli et al. (2018) proposed a model for simultaneously clustering three aspects, namely the nodes into groups, the documents into

topics, and the occurrences into time clusters. It can also been seen as a generalisation, or even a direct combination, of both the STBM and the dynamic extension of the LDA.

### Textual nodes

In another set of models, the entities are usually documents with texts and links between them. Note that a document is defined in its broad sense, as it can be a Wikipedia page, a blog post, or an academic article, the links in the last of which are citations or references. The data structure is then a graph, usually directed, with texts in the *nodes*. We shall discuss a few of them.

Liu et al. (2009) argued that links between documents are not only determined by content similarity, but also by the connections between authors, because authors are naturally more aware of documents in their community and might not be aware of the possibly more relevant documents outside it. They introduced a topic-link LDA model, in which the edge probability between document $p$ and $q$ depends on a linear combination of document similarity and author similarity. The former similarity is the dot product of the topic memberships $\theta_p$ and $\theta_q$ of the documents, generated as described in "Topic models" section, while the latter is the dot product of the memberships of the authors, drawn in a similar way from a separate Dirichlet distribution. The coauthorship network of the authors, however, is not incorporated to enrich the information on the author memberships.

Chang and Blei (2010) proposed a relational topic model, in which the graph of the documents depends on their content similarity. First, each document is generated according to LDA. Next, for documents $p$ and $q$, their edge probability depends on the similarity between the latent variables $\mathbf{Z}_p$ and $\mathbf{Z}_q$ as defined in "Topic models" section, quantified by a link probability function. Different version of this function are considered.

Ho et al. (2012) introduced a model called TopicBlock, in which a latent hierarchical or tree structure is assumed to generate the documents, which are the leaf nodes. Also defined is a hierarchical node $h$, that is, the root or anything between the root and the leaves. Associated with $h$ is a word distribution, denoted by $\phi_h$ for alignment with notation in "Topic models" section, as well as a parameter for the edge probability between any two documents that share $h$ as their deepest common ancestor. Three things are then generated for each document. First, the path from the root to the document arises from a nested Chinese Restaurant process, which is related to the DP. Second, the words are generated according to a mixture of the distributions $\phi_h$ of all the nodes $h$ along the path. Finally, the presence or absence of edge of this document with another document follows a Bernoulli distribution according to the parameter associated with their deepest common ancestor. The use of a hierarchical latent structure to infer the number of groups or topics and for scalability is similar to Peixoto (2017a).

### Inference approach

As it is illustrated in "Stochastic block models" and "Topic models" sectiona that a naive MCMC algorithm can be derived for a block model and a topic model, respectively, it is natural and possible to derive a similar algorithm for a model that combines both, if scalability is less of a concern compared to algorithmic simplicity. Each of Zhou et al. (2006); McCallum et al. (2007); Pathak et al. (2008); Ho et al. (2012) and Sachan et al. (2012) have derived a Gibbs sampler for their corresponding model. In particular, Ho et

**Table 3** Models combining block models and topic models

| Article | Type of graph | | Textual | Inference | Clustering | Number of groups/topics | Longitudinal |
|---|---|---|---|---|---|---|---|
| Zhou et al. (2006) | Valued | directed | Edges | MCMC | Soft | Fixed (6 groups; 20 topics) | No |
| McCallum et al. (2007) | Valued | directed | Edges | MCMC | Soft | Criterion (Perplexity) | No |
| Pathak et al. (2008) | Valued | directed | Edges | MCMC | Soft | Fixed (8 groups; 25 topics) | No |
| Liu et al. (2009) | Binary | directed | Nodes | Variational | Soft | Criterion (Perplexity) | No |
| Chang and Blei (2010) | Binary | undirected | Nodes | Variational | Soft | Fixed (5,10,15,20,25 topics) | No |
| Ho et al. (2012) | Binary | directed | Nodes | MCMC | Soft | Dirichlet process | No |
| Sachan et al. (2012) | Valued | directed | Edges | MCMC | Soft; Soft | Criterion (Perplexity) | No |
| Bouveyron et al. (2016) | Valued | directed | Edges | Variational | Soft; Hard | Criterion (BIC-like) | No |
| Corneli et al. (2018) | Valued | directed | Edges | Variational | Soft; Hard | Criterion (ICL) | Yes |

Under the column "clustering", the first approach refers to the topics. If there is a second approach, it refers to the groups

al. (2012) followed Griffiths and Steyvers (2004) and integrated out the parameters other than the latent variables in their collapsed Gibbs sampler.

On the other hand, the VEM algorithm described in "Variational methods" section is very general and equally feasible as the alternative. Therefore, it has been used by Liu et al. (2009); Chang and Blei (2010); Bouveyron et al. (2016) and Corneli et al. (2018) as the inference algorithm. Bouveyron et al. (2016) and Corneli et al. (2018) observed that, in their STBM and dynamic counterpart, the equivalent of the lower bound in (23) can be split into two components, one depending on the variational distribution of the latent topic vectors $\mathbf{Z}$ and the other not. Therefore they applied an extension of the VEM algorithm, called the classification VEM algorithm, in which the lower bound is maximised alternatively in two steps. In the first step, the lower bound is maximised with respect to the variational distribution and the collection of the word distributions, $\mathbf{\Phi}$, in the usual way of a VEM algorithm. In the second step, the maximisation is carried out in a greedy fashion with respect to the remaining latent variables, namely the latent groups in Bouveyron et al. (2016), and the latent groups and latent time clusters in Corneli et al. (2018).

### Clustering approach

For data with textual edges, while soft clustering still applies to discovering the topics of these documents, it is less clear for the nodes (users). For example, in Zhou et al. (2006) and Pathak et al. (2008), a distribution of users is associated with each group, rather than the other way round. In such cases, we only report the clustering approach for the documents in Table 3. Sachan et al. (2012) are the only ones who soft clustered both the documents and users, the latter of which can belong to multiple documents *and* multiple groups. Bouveyron et al. (2016) and Corneli et al. (2018) hard clustered the nodes and

soft clustered the documents, which are the predominant approaches in SBMs and topic models, respectively.

For data with textual nodes, the models essentially assume that the same set of latent variables, that is their group memberships influence both the generation of the words and the connection between the documents. The soft clustering or mixed membership approach adopted by LDA is possible to be carried over just to a model which combines LDA with a block model, overriding the usual hard clustering approach for the latter. This is the case for all of Liu et al. (2009); Chang and Blei (2010), and Ho et al. (2012).

### Number of groups/topics

The number of groups or topics is usually fixed or determined by certain criterion, possibly because of the complexity of the model. For data with textual edges, both clustering the nodes into groups and the documents into topics are required. Zhou et al. (2006) used 6 groups and 20 topics for the Enron data set, while Pathak et al. (2008) used 8 groups and 25 topics, and perplexity is used by both McCallum et al. (2007) and Sachan et al. (2012) as the criterion, the latter of which selected 10 groups and 20 topics. On the other hand, Bouveyron et al. (2016) used a BIC-like criterion to find 10 groups and 5 topics, while Corneli et al. (2018) used the ICL to find 6 groups and 9 topics. It should be noted that, such discrepancy is due to not only the choice of the criterion but also the model itself.

For data with textual nodes, clustering the nodes into groups is equivalent to clustering the documents in topics. Chang and Blei (2010) used various numbers (5,10,15,20,25) of topics for their data, while Liu et al. (2009) used perplexity as the criterion. Ho et al. (2012) are the only ones who used incorporated a DP in their hierarchical model, although the number of groups found is not reported. Furthermore, they fixed the hierarchical level to 2 or 3 for their data sets.

### Comparison

A comparison is provided in this section regarding works that incorporate longitudinal and/or topic model in SBMs and other models for graphs. Specifically, we will look at their performances on one widely studied dataset.

### Real-world example

The Enron email network dataset (Klimt and Yang 2004), which contains the email communications between its employees from 1999 to 2002 before its bankrupty in 2001, has been widely studied because of the availability of the temporal information (the time points of the email exchanges), the textual information (the actual contents of the emails), and the subsequent graph (the network of the employees). While the works reviewed have a different size of network, due to the different time periods studied, in most cases the number of nodes $n$ is around 150, except Xu and Hero III (2013) and Sanna Passino and Heard (2019) with $n = 184$. A few events happened during that period:

1. 2001-08-14: Then CEO Jeff Skilling resigned.
2. 2001-08-15 to 2001-08-22: The fininical flaws of the company were disclosed.
3. 2001-09-11: September 11 attacks, which had no direct relationship with Enron though.
4. 2001-10-31: The opening of fraud investigation by the Securities and Exchange Comission.

5.    2001-12-02: The filing for bankruptcy of the company.

These events likely affected the dynamics of the employees, which could be reflected from the email exchanges, and were looked into by different works below.

### Longitudinal modelling

The dynamic SBMs reviewed in "SBM with longitudinal modelling" section provided various insights into the dynamics of the network. In their dynamic MMSBM, Fu et al. (2009) and Xing et al. (2010) examined the temporal evolution of the mixed memberships of each employee, which can be interpreted as different (latent) roles. They found that, perhaps not surprisingly, that employees with strong connections with multiple groups and/or important positions tended to have multiple active roles most of the time. They also discovered the major changes in the mixed memberships aligned with real-life events, such as 2 and 5 above. Related observations were made by Xu and Hero III (2013), who examined the evolution of the block matrix $\mathbf{C}^{(t)}$ instead, and fixed the $K = 7$ groups according to the known roles of the employees. They discovered that the elements of $\mathbf{C}^{(t)}$ spiked at event 1.

Other applications include Matias et al. (2015), who found out that their ICL criterion did not manage to select a reasonably small $K$ under their SBM with conditional Poisson processes. Fixing $K = 3$, they discovered that two groups had intra-group communications that peaked at different periods, while the remaining group had very little activity (within itself and with other groups). While Fan et al. (2015) also applied their MTI model to the Enron dataset, they only reported that it outperformed the then existing models such as the MMSBM (Airoldi et al. 2008) and the LFRM (Miller et al. 2009). Similar claims of better performance over SBM were made by DuBois et al. (2013) in their dynamic model.

### Topic modelling

Next, we look at the topics in the emails exchanged, discovered by models reviewed in "SBM with topic modelling" section. McCallum et al. (2007) found the $K = 50$ topics according to their author-recipient-topic model matched well with specific issues in the operations of a company, and such observation is echoed by the respective models of Zhou et al. (2006) and Pathak et al. (2008). Upon comparison with simply fitting an SBM to the network, they found that two employees in the same group inferred by the SBM might actually have vastly different roles in the company. This means that, while the two employees wrote to roughly the same set of employees (hence the stochastic equivalence), the difference in their roles might not be revealed without taking into account the texts of the emails. Using their community-author-recipient-topic model, Pathak et al. (2008) showed that the groups found were topically meaningful, which means that different groups wrote emails on different topics, implying a successful incorporation of the textual information. However, Zhou et al. (2006); McCallum et al. (2007), and (Pathak et al. 2008) did not compare the groups they found with the actual roles of the employees.

Apart from making a similar observation to Pathak et al. (2008); Sachan et al. (2012) managed to compute the modularity 20, and found that not only did their topic user community model outperformed the other models mentioned in this section across different values of $K$, it also achieved a maximimum modularity (at $K = 10$) greater than 0.3, indicating significant community structure (Clauset et al. 2004). As modularity is not a formal

criterion for model selection, Sachan et al. (2012) used perplexity to select an optimal number of groups, which was also around $K = 10$.

In their STBM, Bouveyron et al. (2016) optimised the BIC-likelihood criterion with respect to the number of groups $K$ and the number of topics *simultaneously*. The optimal $K = 10$ (and 5 topics) is larger than that found by an SBM alone ($K = 8$). Closer investigation revealed that some employees in the same group found by SBM talked about different topics than the rest of the group, hence the splitting of a group into smaller ones found by STBM. This is in agreement with the observation by McCallum et al. (2007).

### Longitudinal and topic modelling

In the dynamic STBM by Corneli et al. (2018), model selection was done by maximising the ICL criterion with respect to the number of groups $K$, the number of topics, *and* the number of time clusters. Compared to 10 groups and 5 topics in the STBM, 6 groups, 9 topics and 4 time clusters were selected here. The 4 time clusters corresponded well (with a slight lag) to the four periods separated by events 3, 4 and 5 described at the beginning of this section. Furthermore, they found different interactions between groups at different time clusters, with different topics discussed, thus showing the need of incorporating both temporal and textual information.

### Approaches

Similar to the comparison for SBMs in "Comparison" section, the modelling approach in topic models influences, and is influenced by other aspects discussed. For example, the use of a hierarchical process for the number of topics usually leads naturally to an MCMC algorithm, in particular a Gibbs sampler. While topic models are not the focus of this review, we related some of them with their counterparts for graphs, as similar issues regarding inference and dealing with $K$ arise in both sets of models. For a comprehensive comparison between the topic models, please see Table 4.

### Discussion

In this review we have seen a spectrum of statistical models for graphs, in particular the SBMs, with or without the longitudinal and/or topic modelling. They have been compared in different aspects, including the clustering approach, the inference approach, and how the number of groups $K$ is being coped with. Instead of looking at each model as a unit, we investigated them in a systematic and cross-sectional way, hopefullying

**Table 4** Topic models and related clustering models

| Article | Inference | Clustering | Number of topics | Longitudinal |
| --- | --- | --- | --- | --- |
| Blei et al. (2003) | Variational | Soft | Criterion (Perplexity) | No |
| Griffiths and Steyvers (2004) | MCMC | Soft | Criterion (Marginal log-like.) | No |
| Teh et al. (2005) | MCMC | Soft | Dirichlet process | No |
| Blei and Lafferty (2006) | Variational+SMC | Soft | Fixed (20) | Yes |
| Ren et al. (2008) | MCMC | Soft | Dirichlet process | Yes |
| Papaspiliopoulos and Roberts (2008) | MCMC | Soft | Dirichlet process | No |
| Ahmed and Xing (2010) | MCMC | Soft | Dirichlet procees | Yes |
| Arun et al. (2010) | Unknown | Soft | Criterion (K-L divergence) | No |
| Griffiths and Ghahramani (2011) | MCMC | Soft | Indian buffet process | No |
| Wang et al. (2011) | Variational | Soft | Dirichlet process | No |

presenting the landscape of the literature in a straightforward manner, so that the models most relevant to the reader can be compared with relative ease.

It should be noted that, however, these aspects are not completely independent of each other, nor are they independent of the model specification. In particular, by looking at Table 1, a shift in the approach to $K$, which is highly related to the developments in the models themselves, can be seen over the years. While previous SBMs have relied on nonparametric Bayesian processes used in topic models, more recent SBMs are usually collapsed. With model parameters being integrated out, this allows model selection criteria to be computed (Côme and Latouche 2015), or tractable posterior (McDaid et al. 2013; Newman and Reinert 2016; Peixoto 2017a) to be used with efficient MCMC algorithms (Peixoto 2014b), therefore enabling $K$ to be selected or estimated without trans-dimensional methods.

Two of the recent developments are worth singling out. The first is the nested micro-canonical SBM (Peixoto 2017a), which establishes the MDL (Peixoto 2013) and its equivalence with the usual Bayesian inference approach, incorporates an efficient MCMC algorithm (Peixoto 2014a) and a hierarchical structure (Peixoto 2014b) to model $K$ and circumvent the issue with potential underfitting due to the maximum $K$ detectable in an SBM (Peixoto 2013). One possible direction is to apply these methods and techniques to hypergraphs as well, as they are still a growing field to our knowledge.

The second is the increasing adopting of the DC-SBM (Karrer and Newman 2011), not only in modifying the model, but also in aiding model selection in different ways (Yan et al. 2014; Yan 2016; Wang and Bickel 2017; Hu et al. 2019). While it is particularly useful for dealing with the degree heterogeneity, the results are more mixed when it is applied to real-world networks, as seen in "Comparison" section. Moreover, the usefulness of the DC-SBM should not be seen as an indication of the inferiority of the original SBM, as it captures different underlying structures that follow stochastic equivalence. One possible direction is to combine the regularised SBM (Lu and Szymanski 2019b) and the weighted SBM (Aicher et al. 2015), and have both versions of the SBM in an unifying model. Rather than aimlessly increasing the model complexity, one should aim at modifying or extending an SBM to realistically capture the properties of real-world networks.

Future directions can also be made for extending models that combine SBMs and topic models, and can be split into the two broad categories as in "Type of data and modelling approach" section. For graphs with textual edges, while Corneli et al. (2018) currently represents the state-of-the-art among the models reviewed, a mixed membership version for the nodes could be incorporated. The number of groups/topics/time clusters could also be modelled, possibly with the associated parameters, whose dimensions grow with these numbers integrated out, potentially leading to more accurate computations of criterion and/or more efficient inference algorithms.

For graphs with textual nodes, there are a few possible directions. One way is to develop a truly hierarchical model, potentially by incorporating a nested SBM, to infer the number of groups *and* the depth of the hierarchical structure simultaneously. Another way is to combine Chang and Blei (2010) with an MMSBM to soft cluster the nodes, accompanied by a criterion to infer the number of groups/topics. Methods by, for example, Li et al. (2016) can also be incorporated to obtain a scalable and efficient inference algorithm. The model by Gerlach et al. (2018) can also be applied to account for the graph between words and documents *and* the graph between documents simultaneously, in an unifying

framework. Finally, the SBM by Stanley et al. (2019) that allows continuous attributes on the nodes can potentially be modified to model *textual* attributes instead.

## Author details
[1]Department of Mathematics and Statistics, Lancaster University, Lancaster, United Kingdom. [2]School of Mathematics, Statistics and Physics, Newcastle University, Newcastle, United Kingdom. [3]The Alan Turing Institute, London, United Kingdom.

## References
Abbe E (2018) Community detection and stochastic block models: recent developments. J Mach Learn Res 18:1–86

Adamic L, Glance N (2005) The political blogosphere and the 2004 u.s. election: divided they blog. In: Proceedings of the WWW-2005 Workshop on the Weblogging Ecosystem. https://doi.org/10.1145/1134271.1134277

Ahmed A, Xing EP (2010) Timeline: a dynamic hierarchical dirichlet process model for recovering birth/death and evolution of topics in text stream. In: Grunwald P, Spirtes P (eds). Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence (UAI'10). AUAI Press, Arlington. pp 20–29

Aicher C, Jacobs AZ, Clauset A (2015) Learning latent block structure in weighted networks. J Complex Netw 3:221–248

Airoldi EM, Blei DM, Fienberg SE, Xing EP (2008) Mixed membership stochastic blockmodels. J Mach Learn Res 9:1981–2014. http://www.jmlr.org/papers/v9/airoldi08a.html

Arun R, Suresh V, Madhavan CEV, Murty MN (2010) On finding the natural number of topics with Latent Dirichlet Allocation: some observations. In: PAKDD'10 Proceedings of the 14th Pacific-Asia conference on Advances in Knowledge Discovery and Data Mining. https://doi.org/10.1007/978-3-642-13657-3_43

Bickel P. J, Chen A (2009) A nonparametric view of network models and Newman-Girvan and other modularities. Proc Natl Acad Sci 106(50):21068–21073

Blei DM, Lafferty JD (2006) Dynamic topic models. In: Proceedings of the 23rd International Conference on Machine Learning. https://doi.org/10.1145/1143844.1143859

Blei D. M, Ng A. Y, Jordan M. I (2003) Latent Dirichlet allocation. J Mach Learn Res 3:993–1022. http://www.jmlr.org/papers/v3/blei03a.html

Blondel V. D, Guillaume J-L, Lambiotte R, Lefebvre E (2008) Fast unfolding of communities in large networks. J Stat Mech Theory Exp. https://doi.org/10.1088/1742-5468/2008/10/p10008

Bouveyron C, Latouche P, Zreik R (2016) The stochastic topic block model for the clustering of networks with textual edges. Working paper or preprint. https://hal.archives-ouvertes.fr/hal-01299161

Breiger RL, Boorman SA, Arabie P (1975) An algorithm for clustering relational data with applications to social network analysis and comparison with multidimensional scaling. J Math Psychol 12(3):328–383

Chang J, Blei D. M (2010) Hierarchical relational models for document networks. Ann Appl Stat 4(1):124–150

Chen M, Kuzmin K, Szymanski B. K (2014) Community detection via maximization of modularity and its variants. IEEE Trans Comput Soc Syst 1(1):46–65

Cherifi H, Palla G, Szymanski BK, Lu X (2019) On community structure in complex networks: challenges and opportunities. ArXiv e-prints. arXiv:1908.04901

Clauset A, Newman M. E. J, Moore C (2004) Finding community structure in very large networks. Phys Rev E 70:066111

Côme E, Latouche P (2015) Model selection and clustering in stochastic block models based on the exact integrated complete data likelihood. Stat Model 15(6):564–589

Corneli M, Bouveyron C, Latouche P, Rossi F (2018) The dynamic stochastic topic block model for dynamic networks with textual edges. Stat Comput. https://doi.org/10.1007/s11222-018-9832-4

Daudin J. J, Picard F, Robin S (2008) A mixture model for random graphs. Stat Comput 18(2):173–183

Decelle A, Krzakala F, Moore C, Zdeborová L (2011) Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications. Phys Rev E 84:066106

Dempster A. P, Laird NM, Rubin D (1977) Maximum likelihood from incomplete data via the EM algorithm. J R Stat Soc Ser B (Methodol) 39(1):1–38. http://www.jstor.org/stable/2984875

DuBois C, Butts C, Smyth P (2013) Stochastic blockmodeling of relational event dynamics. In: Proceedings of the 16th International Conference on Artificial Intelligence and Statistics (AISTATS) 2013. Scottsdale, AZ, USA. Volume 31 of JMLR: W&CP 31

Fan X, Cao L, Xu RYD (2015) Dynamic infinite mixed-membership stochastic blockmodel. IEEE Trans Neural Netw Learn Syst 26(9):2072–2085

Fan X, Xu RYD, Cao L (2016) Copula mixed-membership stochastic blockmodel. In: Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI'16. AAAI Press. pp 1462–1468. http://dl.acm.org/citation.cfm?id=3060621.3060824. Accessed 1 Jan 2016

Fortunato S (2010) Community detection in graphs. Phys Rep 486:75–174

Fortunato S, Barthélemy M (2007) Resolution limit in community detection. Proc Natl Acad Sci 104(1):36–41

Fu W, Song L, Xing EP (2009) Dynamic mixed membership blockmodel for evolving networks. In: Proceedings of the 26th International Conference on Machine Learning. pp 329–336. https://doi.org/10.1145/1553374.1553416

Gerlach M, Peixoto TP, Altmann EG (2018) A network approach to topic models. Sci Adv 4:eaaq1360

Ghasemian A, Zhang P, Clauset A, Moore C, Peel L (2016) Detectability threshold and optimal algorithms for community structure dynamic networks. Phys Rev X 6:031005

Girvan M, Newman MEJ (2002) Community structure in social and biological networks. Proc Natl Acad Sci 99(12):7821–7826

Godoy-Lorite A, Guimerà R, Moore C, Sales-Pardo M (2016) Accurate and scalable social recommendation using mixed-membership stochastic block models. Proc Natl Acad Sci 113(50):7–14212

Good BH, de Montjoye Y-A, Clauset A (2010) Performance of modularity maximization in practical contexts. Phys Rev E 81:046106

Gopalan PK, Mimno DM, Gerrish S, Freedman M, Blei DM (2012) Scalable inference of overlapping communities. In: Pereira F, Burges CJC, Bottou L, Weinberger KQ (eds). Advances in, Neural Information Processing Systems 25 Curran Associates, Inc. pp 2249–2257. http://papers.nips.cc/paper/4573-scalable-inference-o%f-overlapping-communities.pdf

Griffiths TL, Ghahramani Z (2011) The Indian buffet process: An introduction and review. J Mach Learn Res 12:1185–1224

Griffiths TL, Steyvers M (2004) Finding scientific topics. Proc Natl Acad Sci 101:5228–5235

Grünwald PD (2007) The Minimum Description Length Principle. The MIT Press. https://doi.org/10.1007/springerreference_179252

Guimerà R, Sales-Pardo M (2009) Missing and spurious interactions and the reconstruction of complex networks. Proc Natl Acad Sci 106(52):22073–22078

Handcock MS, Raftery AE, Tantrum JM (2007) Model-based clustering for social networks. J R Stat Soc Ser A (Stat Soc) 170:301–354

Hayashi K, Knishi T, Kawamoto T (2016) A tractable fully Bayesian method for the stochastic block model. ArXiv e-prints. arXiv:1602.02256

Ho Q, Eisenstein J, Xing EP (2012) Document hierarchies from text and links. In: Proceedings of the 21st International Conference on World Wide Web. pp 739–748. https://doi.org/10.1145/2187836.2187936

Hoff PD, Raftery AE, Handcock MS (2002) Latent space approaches to social network analysis. J Am Stat Assoc 57(460):1090–1098

Holland PW, Laskey KB, Leinhardt S (1983) Stochastic blockmodels: First steps. Soc Netw 5(2):109–137

Holland PW (1981) An exponential family of probability distributions for directed graphs. J Am Stat Assoc 76(373):33–50

Hu J, Qin H, Yan T, Zhao Y (2019) Corrected Bayesian information criterion for stochastic block models. J Am Stat Assoc. https://doi.org/10.1080/01621459.2019.1637744

Ji P, Jin J (2016) Coauthorship and citation networks for statisticians. Ann Appl Stat 10(4):1779–1812

Karrer B, Newman MEJ (2011) Stochastic blockmodels and community structure in networks. Phys Rev E 83:016107

Kim DI, Gopalan PK, Blei D, Sudderth E (2013) Efficient online inference for Bayesian nonparametric relational models. In: Burges CJC, Bottou L, Welling M, Ghahramani Z, Weinberger KQ (eds). Advances in Neural Information Processing Systems 26. Curran Associates, Inc.

Klimt B, Yang Y (2004) The Enron Corpus: A new dataset for email classification research. In: Boulicaut J-F, Esposito F, Giannotti F, Pedreschi D (eds). Machine Learning: ECML 2004', Vol. 3201 of Lecture Notes in Computer Science. Springer, Berlin Heidelberg. pp 217–226

Kurihara K, Kameya Y, Sato T (2006) A frequency-based stochastic blockmodel. In: Workshop on Information-Based Induction Sciences, Osaka

Lancichinetti A, Fortunato S (2011) Limits of modularity maximization in community detection. Phys Rev E 84:066122

Latouche P, Birmelé E, Ambroise C (2012) Variational Bayesian inference and complexity control for stochastic block models. Stat Model 12(1):93–115

Lee C, Wilkinson D. J (2018) A social network analysis of articles on social network analysis. ArXiv e-prints. arXiv:1810.09781

Li W, Ahn S, Welling M (2016) Scalable MCMC for mixed membership stochastic blockmodels. In: Gretton A, Robert CC (eds). Proceedings of the 19th, International Conference on Artificial Intelligence and Statistics, Vol. 51 of Proceedings of Machine Learning Research. PMLR, Cadiz. pp 723–731. http://proceedings.mlr.press/v51/li16d.html

Liu Y, Niculescu-Mizil A, Gryc W (2009) Topic-link LDA: Joint models of topic and author community. In: Proceedings of the 26th International Conference on Machine Learning. pp 665–672. https://doi.org/10.1145/1553374.1553460

Lu X, Szymanski BK (2019) Asymptotic resolution bounds of generalized modularity and statistically significant community detection. ArXiv e-prints. arXiv:1902.04243

Lu X, Szymanski BK (2019) A regularized stochastic block model for the robust communitydetection in complex networks. Sci Rep 9. https://doi.org/10.1038/s41598-019-49580-5

Ludkin M, Eckley I, Neal P (2018) Dynamic stochastic block models: parameter estimation and detection of changes in community structure. Stat Comput 28:1201–1213

Lunagómez S, Mukherjee S, Wolpert RL, Airoldi EM (2017) Geometric representations of random hypergraphs. J Am Stat Assoc 112(517):363–383

Matias C, Miele V (2017) Statistical clustering of temporal networks through a dynamic stochastic block model. J R Stat Soc Ser B (Stat Methodol) 79(4):1119–1141

Matias C, Rebafka T, Villers F (2015) Estimation and clustering in a semiparametric poisson process stochastic block model for longitudinal networks. ArXiv e-prints. arXiv:1512.07075v1

McCallum A, Wang X, Corrada-Emmanuel A (2007) Topic and role discovery in social networks with experiments on Enron and academic email. J Artif Intell Res 30:249–272

McDaid AF, Murphy TB, Friel N, Hurley NJ (2013) Improved Bayesian inference for the stochastic block model with application to large networks. Comput Stat Data Anal 60:12–31

Miller K, Jordan MI, Griffiths TL (2009) Nonparametric latent feature models for link prediction. In: Bengio Y, Schuurmans D, Lafferty JD, Williams CKI, Culotta A (eds). Advances in Neural Information Processing Systems 22. Curran Associates, Inc. pp 1276–1284

Mørup M, Schmidt MN, Hansen LK (2011) Infinite multiple membership relational modeling for complex networks. In: Proceedings of the 2011 IEEE International Workshop on Machine Learning for Signal Processing. https://doi.org/10.1109/mlsp.2011.6064546

Newman MEJ (2001) Scientific collaboration networks. I. network construction and fundamental results. Phys Rev E 64:016131. https://doi.org/10.1103/physreve.64.016131

Newman MEJ (2001) The structure of scientific collaboration networks. Proc Natl Acad Sci 98(2):404–409

Newman MEJ (2004) Coauthorship networks and patterns of scientific collaboration. Proc Natl Acad Sci 101:5200–5205

Newman MEJ (2006) Finding community structure in networks using the eigenvectors of matrices. Phys Rev E 74:036104

Newman MEJ (2006) Modularity and community structure in networks. Proc Natl Acad Sci 103(23):8577–8582

Newman MEJ (2016) Equivalence between modularity optimization and maximum likelihood methods for community detection. Phys Rev E 94:052315

Newman MEJ, Girvan M (2004) Finding and evaluating community structure in networks. Phys Rev E 69:026113

Newman MEJ, Reinert G (2016) Estimating the number of communities in a network. Phys Rev Lett 117:078301

Ng TJ, Murphy TB (2018) Model-based clustering for random hypergraphs. ArXiv e-prints. arXiv:1808.05185

Nobile A, Fearnside A (2007) Bayesian finite mixtures with an unknown number of components: the allocation sampler. Stat Comput 17:147–162

Nowicki K, Snijders TAB (2001) Estimation and prediction for stochastic blockstructures. J Am Stat Assoc 96(455):1077–1087

Palla K, Knowles DA, Ghahramani Z (2012) An infinite latent attribute model for network data. In: Proceedings of the 29th International Conference on Machine Learning, Edinburgh

Papaspiliopoulos O, Roberts GO (2008) Retrospective Markov chain Monte Carlo methods for Dirichlet process hierarchical models. Biometrika 95(1):169–186

Pathak N, DeLong C, Banerjee A, Erickson K (2008) Social topic models for community extraction. In: The 2nd SNA-KDD Workshop '08'

Peixoto TP (2013) Parsimonious module inference in large networks. Phys Rev Lett 110:148701

Peixoto TP (2014) Efficient Monte Carlo and greedy heuristic for the inference of stochastic block models. Phys Rev E 89:012804

Peixoto TP (2014) Hierarchical block structures and high-resolution model selection in large networks. Phys Rev X 4:011047

Peixoto, TP (2015) Inferring the mesoscale structure of layered, edge-valued, and time-varying networks. Phys Rev E 92:042807

Peixoto TP (2015) Model selection and hypothesis testing for large-scale network models with overlapping groups. Phys Rev X 5:011033

Peixoto TP (2017) Bayesian stochatic blockmodeling. ArXiv e-prints. arXiv:1705.10225

Peixoto TP (2017) Nonparametric Bayesian inference of the microcanonical stochastic block model. Phys Rev E 95:012317

Peixoto TP (2018) Nonparametric weighted stochastic block models. Phys Rev E 97:012306

Peixoto TP (2018) Reconstructing networks with unknown and heterogeneous errors. Phys Rev X 8:041011

Pons P, Latapy M (2006) Computing communities in large networks using random walks. J Graph Algoritm Appl 10(2):191–218

Priebe CE, Sussman DL, Tang M, Vogelstein JT (2015) Statistical inference on errorfully observed graphs. J Comput Graph Stat 24(4):930–953

Raghavan UN, Albert R, Kumara S (2007) Near linear time algorithm to detect community structures in large-scale networks. Phys Rev E 76:036106

Ranciati S, Vinciotti V, Wit EC (2017) Identifying overlapping terrorist cells from the noordin top actor-event network. ArXiv e-prints. arXiv:1710.10319

Ren L, Dunson DB, Carin L (2008) The dynamic hierarchical Dirichlet process. In: Proceedings of the 25th International Conference on Machine Learning. pp 824–831. https://doi.org/10.1145/1390156.1390260

Rohe K, Chatterjee S, Yu B (2011) Spectral clustering and the high-dimensional stochastic blockmodel. Ann Stat 39(4):1878–1915

Rosvall M, Axelsson D, Bergstrom CT (2009) The map equation. Eur Phys J 178:13–23

Rosvall M, Bergstrom CT (2007) An information-theoretic framework for resolving community structure in complex networks. Proc Natl Acad Sci 104(18):7327–7331

Sachan M, Contractor D, Faruquie TA, Subramaniam LV (2012) Using content and interactions for discovering communities in social networks. In: Proceedings of the 21st International Conference on World Wide Web. pp 331–340. https://doi.org/10.1145/2187836.2187882

Sanna Passino F, Heard NA (2019) Bayesian estimation of the latent dimension and communities in stochastic blockmodels. ArXiv e-prints. arXiv:1904.05333

Schaub MT, Delvenne J, Rosvall M, Lambiotte R (2014) The many facets of community detection in complex networks. Appl Netw Sci 2:4

Snijders TAB, Nowicki K (1997) Estimation and prediction for stochastic blockmodels for graphs with latent block structure. J Classif 14(1):75–100

Stanley N, Bonacci T, Kwitt R, Niethammer M, Mucha PJ (2019) Stochastic block models with multiple continuous attributes. Appl Netw Sci 4(54). https://doi.org/10.1007/s41109-019-0170-z

Tallberg C (2005) A Bayesian approach to modeling stochastic blockstructures with covariates. J Math Sociol 29:1–23

Tang X, Yang CC (2011) Dynamic community detection with temporal Dirichlet process. In: 2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing. pp 603–608. https://doi.org/10.1109/passat/socialcom.2011.37

Tang X, Yang CC (2014) Detecting social media hidden communities using dynamic stochastic blockmodel with temporal Dirichlet process. ACM Trans Intell Syst Technol 5(2). https://doi.org/10.1145/2517085

Tarrés-Deulofeu M, Godoy-Lorite A, Guimerà R, Sales-Pardo M (2019) Tensorial and bipartite block models for link prediction in layered networks and temporal networks. Phys Rev E 99:032307

Teh YW, Jordan MI, Beal MJ, Blei DM (2005) Sharing clusters among related groups: Hierarchical Dirichlet processes. In: Saul LK, Weiss Y, Bottou L (eds). Advances in Neural Information Processing Systems 17 (NIPS 2004). MIT Press. pp 1385–1392

Vallè T, Massucci FA, Guimerà R, Sales-Pardo M (2016) Multilayer stochastic block models reveal the multilayer structure of complex networks. Phys Rev X 6:011036

Vallè T, Peixoto TP, Sales-Pardo M, Guimerà R (2018) Consistencies and inconsistencies between model selection and link prediction in networks. Phys Rev E 97:062316

van der Pas SL, van der Vaart AW (2018) Bayesian community detection. Bayesian Anal 13(3):767–796

von Luxburg U (2007) A tutorial on spectral clustering. Stat Comput 17(4):395–416

Vu DQ, Hunter DR, Schweinberger M (2013) Model-based clustering of large networks. Ann Appl Stat 7(2):1010–1039

Wakita K, Tsurumi T (2007) Finding community structure in mega-scale social networks. In: Proceedings of the 16th International Conference on World Wide Web. pp 1275–1276. https://doi.org/10.1145/1242572.1242805

Wang C, Paisley J, Blei DM (2011) Online variational inference for the hierarchical Dirichlet process. In: Proceedings of the 14th International Conference on Artificial Intelligence and Statistics (AISTATS)Š 2011. Volume 15 of JMLR: W&CP 15, Fort Lauderdale. pp 752–760

Wang YJ, Wong GY (1987) Stochastic blockmodels for directed graphs. J Am Stat Assoc 82:8–19

Wang YXR, Bickel PJ (2017) Likelihood-based model selection for stochastic block models. Ann Stat 45(2):500–528

White HC, Boorman, Breiger RL (1976) Social structure from multiple networks. i. blockmodels of roles and positions. Am J Sociol 81(4):730–780

Xie J, Kelley S, Szymanski BK (2013) Overlapping community detection in networks: The state-of-the-art and comparative study. ACM Comput Surv 45(4):43:1–43:35. http://doi.acm.org/10.1145/2501654.2501657

Xing EP, Fu W, Song L (2010) A state-space mixed membership blockmodel for dynamic network tomography. Ann Appl Stat 4(2):535–566

Xu KS, Hero III AO (2013) Dynamic stochastic blockmodels: Statistical models for time-evolving networks. In: Social Computing, Behavioral-Cultural Modeling and Prediction. pp 201–210. https://doi.org/10.1007/978-3-642-37210-0_22

Yan X (2016) Bayesian model selection of stochastic block models. In: 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM). pp 323–328. https://doi.org/10.1109/asonam.2016.7752253

Yan X, Shalizi C, Jensen JE, Krzakala F, Moore C, Zdeborová L, Zhang P, Zhu Y (2014) Model selection for degree-corrected block models. J Stat Mech Theory Exp 5:P05007

Yang T, Chi Y, Zhu S, Gong Y, Jin R (2011) Detecting communities and their evolutions in dynamic social networks - a Bayesian approach. Mach Learn 82:157–189

Zachary WW (1977) An information flow model for conflict and fission in small groups. J Anthropol Res 33(4):452–473. https://www.jstor.org/stable/3629752

Zhang X, Moore C, Newman MEJ (2017) Random graph models for dynamics networks. Eur Phys J B 90:200

Zhao Y, Wu Y-J, Levina E, Zhu J (2017) Link prediction for partially observed networks. J Comput Graph Stat 26(3):725–733

Zhou D, Manavoglu E, Li J, Giles CL, Zha H (2006) Probabilistic models for discovering e-communities. In: Proceedings of the 15th International Conference on World Wide Web. https://doi.org/10.1145/1135777.1135807

Zhou M (2015) Infinite edge partition models for overlapping community detection and link prediction. In: Proceedings of the 18th International Conference on Artificial Intelligence and Statistics (AISTATS) 2015. JMLR: W&CP volume 38, San Diego. pp 1135–1143

## Publisher's Note