Bocconi

Bachelor of Science in BEMACS

Advanced statistical methods in network analysis: the role of Stochastic Block Modeling

Advisor:
Prof. ANTONIO LIJOI

Bachelor of Science thesis by:
TOMMASO VEZZOLI
student ID no. 3144207

Academic Year 2023-2024

Bocconi
Università Commerciale
Luigi Bocconi

Ringraziamenti

Ai miei genitori Alberto e Raffaella, che hanno sempre fatto molti sacrifici affinché potessi seguire il percorso che volevo, e mi hanno sopportato in tutti i miei momenti di difficoltà in questi anni.
Ai miei fratelli Lorenzo e Filippo, che mi hanno regalato diversi attimi di svago e di stacco dallo studio a casa, rendendolo più piacevole e alleggerendo la pressione per gli esami.
A tutti i miei zii e cugini, che mi hanno sempre visto come un piccolo genietto.
A mia nonna Anna, una vera forza della natura, che porta sempre tante risate quando ci viene a trovare e crede in me più di quanto non lo faccia io.
Alla mia ragazza Anastasia, che mi ha reso una versione migliore di me stesso; in questi anni ha sopportato molte volte le mie ansie ma è sempre stata convinta che avrei ottenuto ottimi risultati.
Al mio gruppo di amici dell'università: Giulio, Riccardo, Tomas, Vittorio, e Giovanni, con cui ho condiviso tanti momenti divertenti, sessioni di studio intense in biblioteca, pause pranzo in posti poco raccomandabili, e vacanze esotiche.
Al mio supervisore Antonio Lijoi, per esser stato una guida preziosa durante lo svolgimento della tesi, mostrando sempre molta disponibilità.

Ringrazio tutti immensamente.

Contents

1 THE IMPORTANCE OF RELATIONAL DATA TODAY

Nowadays, data have emerged as one of the most valuable assets in the world economy and its availability is continuously increasing due to the technological advancements. In particular, the role of relational data has evolved from a mere academic interest to a fundamental mean to support operational and strategic decisions of organizations worldwide. Understanding the multifaceted relationships and interactions represented by these data becomes critical for unlocking insights that drive innovation, efficiency, and growth. The burgeoning significance of relational data in our modern era can be explored through the lens of advanced statistical methodologies like stochastic block modelling (SBM).

Relational data, characterized by their focus on the connections and relationships between entities rather than on the entities themselves, provides a unique approach to analyse complex systems. This type of data is intrinsic to networks, which can represent various aspects of reality: from the social interactions that shape our societies to the biological networks that govern life processes. For instance, Facebook, a leading social media platform, reported over 2.8 billion monthly active users in 2021, illustrating the scale and impact of social networks in contemporary life. Similarly, the Internet of Things (IoT) has connected billions of devices, generating a vast network of relational data that encompasses everything from home appliances to industrial machinery.

Statistical models provide a systematic and rigorous framework for understanding the intrinsic processes in interactions, estimating parameters, making predictions, and testing hypotheses. Techniques such as stochastic block modelling allow us to uncover latent patterns and groupings within networks, revealing the underlying organizational principles that govern the behaviour of entities within the network. This methodological approach is not merely academic; it has practical applications across a wide array of disciplines.

In social sciences, for instance, relational data analysis can uncover the dynamics of social networks, identifying influential individuals, understanding the spread of information or misinformation, and mapping community structures. This can have profound

implications for fields such as sociology, political science, and marketing, where understanding the flow of information and the structure of social groups is crucial. An illustrative case is the analysis of Twitter data during political campaigns, which can reveal how information flows between different community clusters and identify the most influential nodes in the network.

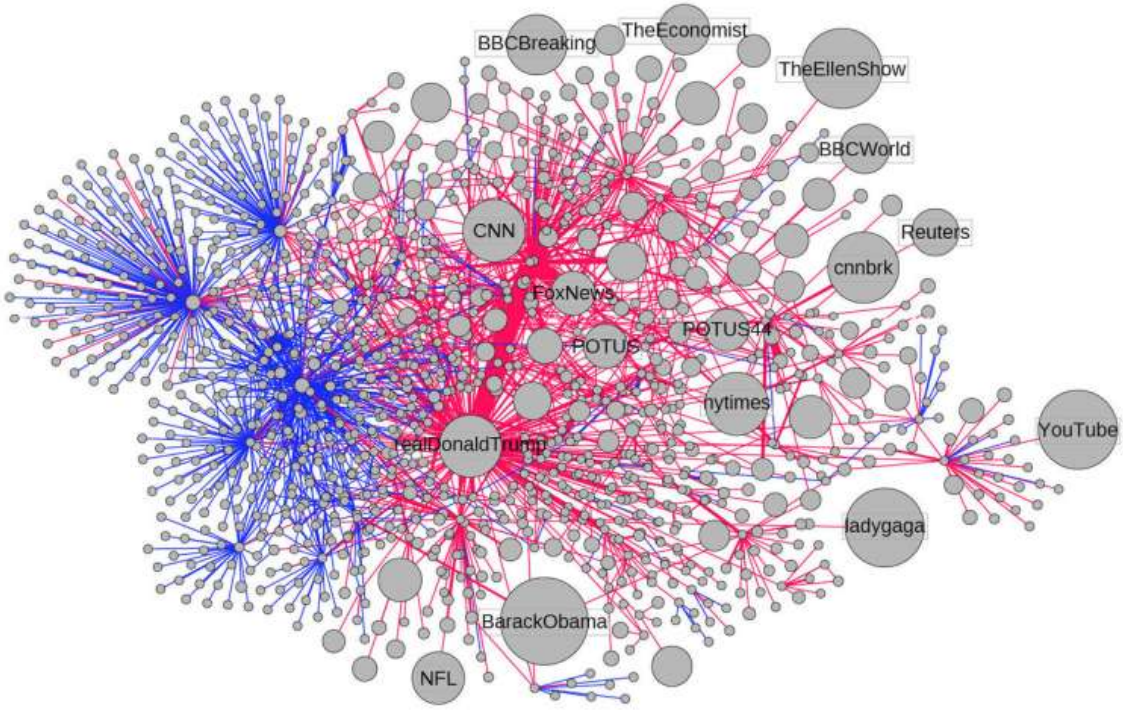

Figure 1: Analysis of the role of automated accounts in the spreading of misinformation during 2016 US elections (Indiana University, 2018) (Chengcheng Shao et al., 2018). Despite the falseness of the news in question, some of the most influential accounts on the platform believed it because they have been cited by many accounts (actually, bots), increasing the credibility of the information.

In the realm of biology and medicine, network analysis helps in mapping genetic interactions, understanding the spread of diseases within populations, and identifying potential pathways for therapeutic interventions. The complexity of biological systems makes them ideal candidates for analysis through stochastic block modelling, providing insights that can lead to breakthroughs in treatment and prevention strategies. For example, the Human Protein Interaction Network has been extensively studied to understand the relationships between different proteins and their impact on various diseases, leading to new insights into cancer and neuro-degenerative disorders.

In technology and business, the analysis of relational data is at the heart of network optimization, algorithm development, and customer relationship management. Companies leverage network analysis to improve supply chain operations, develop targeted marketing campaigns, and enhance customer service. Amazon, for example, utilizes relational data to refine its recommendation algorithms, thereby enhancing the customer shopping experience and driving sales.

Despite its vast potential, the analysis of relational data poses significant challenges, from the collection and storage of large datasets to the development of algorithms capable of handling the complexity and dynamism of real-world networks. Privacy and ethical considerations also come to the forefront, particularly as data becomes more personal and its analysis more penetrating. The balance between leveraging relational data for insights and respecting individual privacy rights is a critical issue that requires careful consideration and regulatory compliance.

This thesis will explore the mathematical and statistical fundamentals of stochastic block modelling (Section 2), focusing on the definition and analysis of well-known models in the literature (Section 3). Additionally, a practical application on real-world data will be used to show the potential of these models and the central role of relational data for understanding complex systems (Section 4).

The ultimate goal of the paper is to explain complex relational data through the application of stochastic block modeling, providing a robust methodological framework. By clearly explaining the mathematical fundamentals, critically evaluating well-established models, and applying these to tangible data, this work aims to establish a simple and practical connection between theoretical statistical concepts and practical data analysis challenges. Through this exploration, the thesis will offer a comprehensive resource for effectively harnessing the potential of SBM in a variety of contexts.

2 MATHEMATICAL AND STATISTICAL FUNDAMENTALS

2.1 Introduction to SBM and Notation

One of the primary challenges in network analysis involves identifying communities within networks that exhibit analogous patterns of connectivity. The Stochastic Block Model (SBM) (Holland et al., 1983), a rigorous statistical approach for community detection, addresses this challenge. SBM analyzes connections in a population represented as a graph of interlinked nodes and identifies distinct groups based on connectivity patterns. Its principal objective is to determine the latent block memberships of nodes, which are considered the fundamental factor determining inter-node connections. The term 'stochastic' underscores the inherent randomness in the model, where each pair of clusters is linked to a specific probability of an edge existing between them. This probabilistic approach not only allows for the simulation of varied network outcomes, thereby more accurately representing the unpredictability of real-world networks but also introduces a level of adaptability in modeling network complexities. Furthermore, SBM assigns prior probabilities to each node's community membership, usually based on a known distribution, enhancing the model's ability to capture the nuanced variability and uncertainty within the data.

All stochastic block models share some common elements on which they are based. The network is in the form of a graph with N nodes with edges connecting them, and an $N \times N$ adjacency matrix Y of 1s or 0s indicating the presence or absence of an edge between two nodes. The graph can be either undirected or directed; in the latter case the adjacency matrix will not be symmetric. An important underlying assumption is that each node belongs to one of the $K < N$ communities, and such membership is represented by a vector **z**.The number of communities K can be either fixed or learnt from the data through different techniques (Section 2.2). In the first case, it implies to have prior information about the data or to empirically test the model with a set of values for K and select the most performing one; the second case is represented by more complex models which simultaneously estimate the number of clusters and the block membership vector. The last elements are two $K \times K$ matrices, namely E, which denotes the number of edges between each pair of groups, and B, the probability matrix

representing the probability of observing an edge between nodes in different groups. The model assumes the so called stochastic equivalence, that is, the presence of an edge is conditionally independent given the group membership. As a consequence, all the entries in the adjacency matrix follow a Bernoulli distribution and are conditionally independent from each other given their block membership.

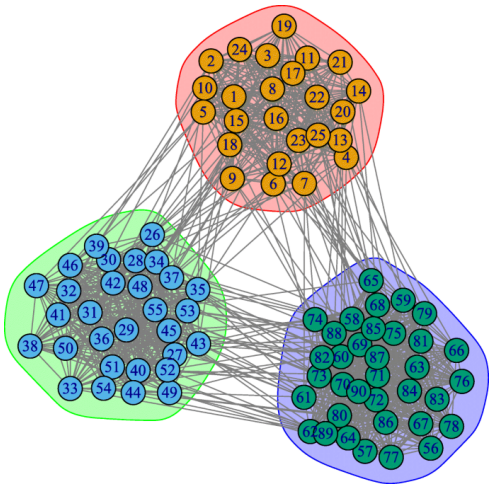

Figure 2: Example of SBM applied on a network; nodes are grouped in 3 clusters, and we can observe more dense interactions within the same cluster than between different groups (Lee et al, 2019)

When applying SBM to real-world data, the block membership **z** and the probability matrix B are not observable, and hence they are the variables of interest to infer. An important assumption to do is that the probability of belonging to each class is independent of the others, making **z** a multinomial variable with a probability vector $\boldsymbol{\pi}$, which usually follows a Dirichlet distribution.

2.2 Advanced Techniques for Mixture Models

The most crucial aspect of SBMs is related to the number of communities K in the network. Many models assume prior knowledge of K or estimate it via cross validation, a technique which tests the performance of a model at different assignments of the parameters, allowing to choose the best value. However, these methods can be computationally expensive, not always feasible, or they don't consider the intrinsic uncertainty behind this parameter. An important step forward involves the definition of models which simultaneously estimates the number of communities and the block memberships by using a probabilistic approach instead of empirical methods (Geng et al., 2019). These types of models are based on advanced concepts from Bayesian nonparametric statistics, which encapsulates methods that do not restrict models to a

fixed set of parameters; instead, it allows for an infinite dimensional parameter space, enabling the model to adapt and grow in complexity with the amount of available data. The techniques that will be used later in the paper are described below in terms of urn models, random partition models and random probability measures.

2.2.1 Pólya Urn

The Pólya urn scheme (Pólya et al., 1923, Blackwell et al., 1973) refers to a general framework of a stochastic process which can be explained as follows: suppose there is an urn with N balls of K different colours, and let x_i for $i \in \{1, ..., K\}$ indicate the number of balls of color i contained in the urn ($\sum_{i=1}^{K} x_i = N$). At each iteration, a ball is drawn at random, hence color i will be observed with probability $\frac{x_i}{N}$; if color i is observed from the draw, we update $x_i = x_i + 1$, that is, we return the ball to the urn along with an additional ball of the same colour. The main idea behind it is the opposite of the sampling without replacement, where every time a colour is observed, it becomes less and less probable to sample it again, until all the balls of that colour have been taken out of the urn. In the Pólya urn scheme, instead, both the balls of a specific colour and the total number of balls in the urn increase at each iteration; therefore, the act of drawing the same colour over time will affect continuously less the future samplings. The main limitation of this scheme, however, is that it allows only for fixed prior number of colors K.

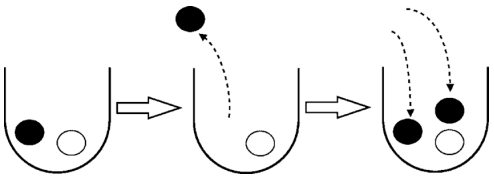

Figure 3: Graphical representation of the Pólya urn scheme, where the draw of a black ball is followed the replacement of 2 black balls.

2.2.2 Chinese Restaurant Process

The Chinese Restaurant Process (Pitman et al., 1995) is a modification of the Pólya urn scheme in which, taking the previous example of the urn, the new ball introduced will be of the same colour with a certain probability, otherwise a new colour will be introduced. The process is explained using the analogy of a Chinese restaurant with potentially infinite tables, each of which has potentially infinite seats. In this process, a new customer entering the restaurant faces a choice: join an existing table or start a

new one. The likelihood of joining an already occupied table is directly proportional to the number of diners already seated there, effectively making it more probable that the new arrival will choose to sit with others rather than alone. This probability is conditional on the table assignments of previous customers, demonstrating how each decision influences subsequent ones. Despite this tendency to join already occupied tables, due to the inherent randomness of the process, the probability of generating a new tiny extraneous cluster is not negligible and, depending on the concentration parameter of the CRP in question, we would observe fewer large clusters or average clusters with some small, isolated groups. It is important to notice that the CRP cannot be used to consistently estimate the number of clusters in an SBM since it diverges when the sample grows infinitely.

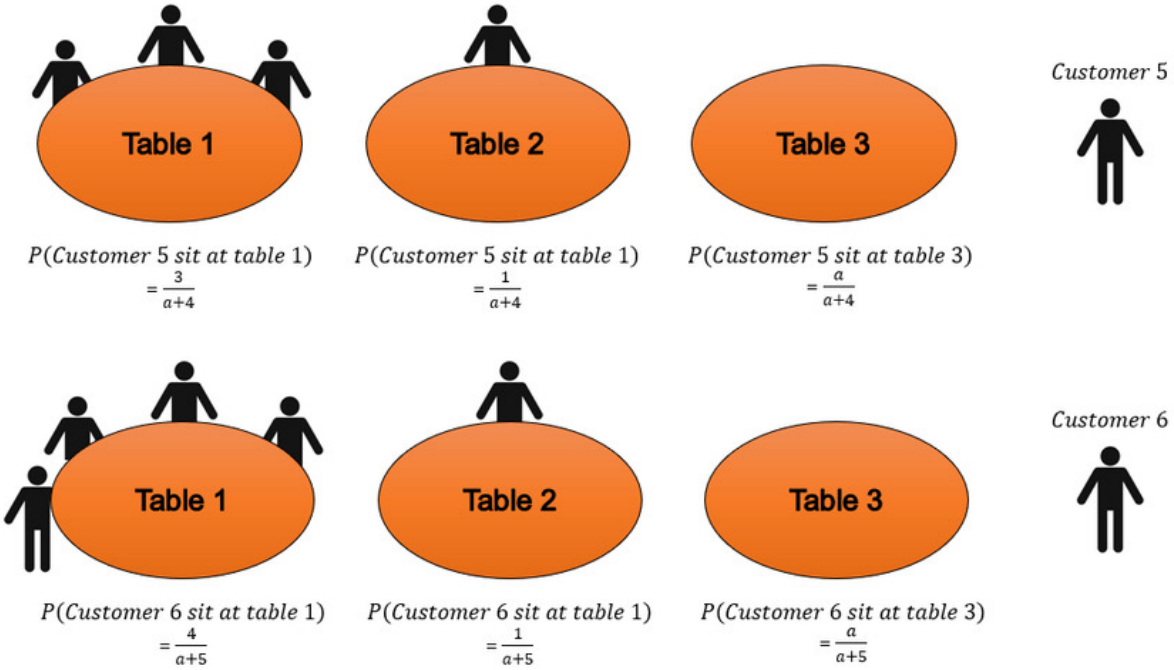

Figure 4: CRP scheme example, with a being the concentration parameter (Chan & Chin, (2019))

Let **z**$= (z_i), \forall i \in \{1, ..., N\}$ be the vector indicating at which table customer i is sit. Then

$$Pr(z_i = c|z_1, ..., z_{i-1}) \propto \begin{cases} |c| & \text{if } c \text{ is an existing table with customers} \\ \alpha & \text{if } c \text{ is a new table} \end{cases}$$

where $\boldsymbol{\alpha}$ is the concentration parameter.

2.2.3 Dirichlet Process

The Dirichlet process (Ferguson, 1973) is a fundamental concept which provides a framework for modelling non-parametric probability distributions. The DP can be considered as the underlying probability distribution of the clustering configuration in the CRP, hence enabling the sampling of cluster assignments from a distribution.

Formally, the Dirichlet Process is defined by two parameters: a base distribution G_0 and a concentration parameter α. The base distribution G_0 is a probability distribution that characterizes the expected value of the process, essentially dictating the underlying structure of the data or "clusters" in the absence of any observed data.

Mathematically, if $G \sim \text{DP}(\alpha, G_0)$, then for any measurable partition $(A_1, \ldots, A_k)$ of the sample space, the vector $(G(A_1), \ldots, G(A_k))$ follows a Dirichlet distribution with parameters $(\alpha G_0(A_1), \ldots, \alpha G_0(A_k))$. The concentration parameter α, a positive scalar, influences the variability of the process. It controls the number of distinct clusters (or tables in the CRP metaphor) that are likely to be formed: a larger α leads to a higher probability of new cluster formation, whereas a smaller α tends to favor the concentration of observations within existing clusters. The probability of a new customer joining a new table, distinct from the existing ones, is given by $\frac{\alpha}{\alpha+N}$, where N is the number of previous customers. This is derived from the stick-breaking construction of the DP, where the weights π_i of the clusters are generated as follows:

- Let $V_i \sim \text{Beta}(1, \alpha)$ for each i.
- Define $\pi_i = V_i \prod_{j=1}^{i-1}(1 - V_j)$ for each i.

This construction shows how the weights of the clusters are dependent on α and the sequence of Beta-distributed random variables V_i. The weights π_i represent the proportions of the total population that belong to each cluster, and the infinite-dimensional nature of the DP allows for an unlimited number of potential clusters, reflecting the non-parametric flexibility of the model.

In essence, the Dirichlet Process offers a probabilistic mechanism for modeling an infinite mixture model, where the number of mixture components (clusters) does not need to be specified a priori. This property makes the DP particularly useful for applications where the structure of the data is unknown and potentially complex, allowing for model adaptability as more data becomes available.

2.2.4 Mixture of Finite Mixtures

The Mixture of Finite Mixtures (Gnedin & Pitman, 2006) is a modification of the traditional mixture model described by the DP, adding an additional layer of complexity to the CRP.

The innovative aspect of MFM, as indicated by its name, lies in its assumption that each component within the mixture model is, in itself, a finite mixture distribution, allowing for more control over the distribution dynamics of cluster assignments. In practice, the MFM assumes that the true number of components K is a random variable following a probability mass function (pmf) on $\mathbb{N}^+$, the set of positive natural numbers. It can be proven that the number of clusters deriving from this pmf, out of N observations, converges almost surely to the true value K as N grows, hence it is consistent. This convergence contrasts with the DP, in which the number of clusters tends to diverge at a $\log N$ rate, hence leading to inconsistent number of clusters. As a result, the MFM defines a more balanced cluster assignment, typically resulting in groups with more uniform size, while in the DP there would be many tiny extraneous clusters.

3 LITERATURE REVIEW

Before analysing the more advanced models based on Bayesian nonparametric methods, it is appropriate to present a simpler model, which enables a better understanding of the structure of SBMs.

3.1 Bayesian Inference in SBMs with Value-Directed Graphs

The model proposed is an a posteriori blockmodel with a fixed number of communities K (Nowicki et al., 2001), which can be adapted to undirected, directed, and value-directed graph. The distinctive characteristic of SBMs is the inference method proposed to estimate the variables of interest: the block membership vector **z**, and the edge probability matrix B. In this case the inference method is a generalized Bayesian approach combined with an MCMC algorithm, the Gibbs sampler. This model ensures, under the correct assumptions and conditions, to reach convergence and consistent results.

3.1.1 Model Definition

The block structure of the SBM is defined as the joint distribution of the observed relations and of the block membership vector, based on the following distributions:

- Unconditional distribution of block membership:

$$Pr(z_i = j,\ \forall i \in \{1, ..., N\}, \text{ with } j \in \{1, ..., K\}) = \prod_{j=1}^{K} \pi_j^{m_j}$$

  where $z_i \in \{1, ..., K\}$ is the block membership of node i, π_j is the probability of being assigned to a specific community j, and $m_j = \sum_{i=1}^{N} \mathbb{1}\{z_i = j\}$ indicates the number of community members in group j;

- Conditional distribution of relationships:

$$Pr(Y \mid \mathbf{z}, \boldsymbol{\pi}, B) = \left(\prod_{\mathbf{a} \in A} \prod_{1 \leq k < h \leq K} (B_{\mathbf{a}}(k,h))^{E_{\mathbf{a}}(k,h)} \right) \times \left(\prod_{\mathbf{a} \in A'} \prod_{k=1}^{K} (B_{\mathbf{a}}(k,k))^{E_{\mathbf{a}}(k,k)} \right)$$

  where Y is the $N \times N$ adjacency matrix of observed relations between the each pair of nodes, **a** is a 2-tuple containing the directed relations between nodes in two

clusters in both directions, A is the set of all possible **a**, A' is a subset of A which contains the symmetric relations (those which are equal in the two directions) and the half of the set of asymmetric relations, B is a $K \times K \times |A|$ array containing the probabilities of observing each possible relation **a** $\in A$ between any pair of clusters (k, h); since we are considering the case of a value-directed graph, the adjacency matrix will not contain only 0s and 1s but the actual values of the relationships, in accordance to the given alphabet A. Therefore, instead of being the standard $K \times K$ edge probability matrix, B here is a multi-dimensional array. Analogously, the edge-counting matrix E takes the form of a $K \times K \times |A|$ array; its entries, $E_{\mathbf{a}}(k, h)$, counts the number of times relation **a** is observed from a node in cluster k and one in cluster h

- Stochastic block model distribution:

$$Pr(Y, \mathbf{z}|\boldsymbol{\pi}, B) = \left(\prod_{k=1}^{K} \pi_k^{m_k}\right) \times \left(\prod_{\mathbf{a}\in A} \prod_{1\leq k<h\leq K} (B_{\mathbf{a}}(k,h))^{E_{\mathbf{a}}(k,h)}\right) \times \left(\prod_{\mathbf{a}\in A'} \prod_{k=1}^{K} (B_{\mathbf{a}}(k,k))^{E_{\mathbf{a}}(k,k)}\right)$$

Given the distributions defined above, it is possible to recover the block structure defined by the latent membership vector **z** by applying the Bayes rule to get $Pr(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\pi}, B)$. In words, this distribution represent the posterior distribution of memberships given the observed relations of the adjacency matrix Y. The last step for achieving the recovery of the structure is probably the most important, that is, to estimate the posterior distributions of the parameters $(\boldsymbol{\pi}, B)$ given **z** and **y**. The prior distributions assigned to the parameters are two independent uniform Dirichlet distributions (with all the K parameters equal), since both parameters describes probabilities, hence values between 0 and 1. An approximate evaluation of posterior estimates of such parameters is determined through a Gibbs Sampling algorithm.

3.1.2 Gibbs Sampling

Gibbs sampling is an MCMC simulation method (Geman et al., 1984) used to generate approximate samples from the posterior distribution through an iterative process which alternates a simulation from the conditional distribution of a subgroup of the parameters of interest, given the other subgroup, and a simulation of the latter, given the new instances elaborated in the previous step. In the context of the proposed model, the

first group is given by π and B; the second-step parameter will be the vector **z**. Below is the pseudo-code of the algorithm:

---

initialize: $\boldsymbol{\pi}^{(p)}, B^{(p)}, \mathbf{z}^{(p)}$

Draw $\boldsymbol{\pi}^{(p+1)}, B^{(p+1)}$ from posterior distribution $(\boldsymbol{\pi}, B)$ given $(\mathbf{z}^{(p)}, Y)$

repeat

  for $i = 1$ **to** N **do**

    Draw $z_i^{(p+1)}$ from the conditional distribution of z_i given

    $\boldsymbol{\pi}^{(p+1)}, B^{(p+1)}, Y, z_h^{(p+1)}$ for $h = 1, ..., i-1$, and $z_l^{(p)}$ for $l = i+1, ..., N$

  until convergence

Tests on this algorithm show that for networks of moderate size (below 100 nodes) convergence happens after approximately $2M_0$ iterations (Nowicki et al., 2001), where M_0 is a preset number of iterations in which the parameters of the Dirichlet prior distribution of $\boldsymbol{\pi}$ are decreased linearly from $T_1 = 10N$ to $T_{M_0} = 100K$ and the parameters of the Dirichlet prior distribution of B are multiplied by a linearly increasing factor w from $w_1 = 1/N$ to $w_{M_0} = 1$. The reason for these M_0 iterations is to improve the probability of convergence by starting with over-dispersed prior distributions in order to avoid ending stuck in a local optimum. The Gibbs sampler allows to obtain all the elements necessary to recover the predictive posterior distribution of **z** given Y, and hence the block membership of all the nodes in the network.

3.2 Mixed Membership Stochastic Block Model

One of the main limitations of several SBMs is the strict assumption that each element of the network can belong exclusively to one community.

In the real world the scheme is much more complex, with the group of belonging of each member possibly changing when it interacts with a specific individual rather than another. In such a complex framework, the Mixed Membership Model (MMB) (Airoldi et al., 2008) enables a more complete and realistic analysis of reality, by relaxing the stricter assumption of SBM.

3.2.1 Model Definition

The model presents some differences from the standard set of variables in SBMs. The graphical representation of the data remains unchanged, with a directed graph with N nodes, K communities, and an adjacency matrix Y, where Y $(p, q) \in \{0, 1\}$ indicates the presence of an interaction (edge) from node p to node q. Each node p is associated with a random K-dimensional vector $\boldsymbol{\pi}_p$, the elements of which indicates the probability of belonging to each possible cluster. The $K \times K$ matrix B is defined as the probability matrix of observing an interaction between nodes in each possible pair clusters. The main difference consists in the definition of the block membership vector $\mathbf{z}$: each node p has an indicator vector $\mathbf{z}_{p\to q}$ for each other node q in the network: it is a K-dimensional vector the entries of which are all 0 except the entry indicating the block membership of node p when interacting with node q, which will have value 1. The distributions of the different elements described are the following:

$$
\begin{aligned}
&\boldsymbol{\pi}_p \sim \text{Dirichlet}(\boldsymbol{\alpha}), \qquad \forall\, p \in \{1, ..., N\} \\
&\mathbf{z}_{p\to q} \sim \text{Multinomial}(\pi_p), \qquad \forall\, (p, q) \in \{1, ..., N\} \times \{1, ..., N\} \\
&\mathbf{z}_{p\leftarrow q} \sim \text{Multinomial}(\pi_q), \qquad \forall\, (p, q) \in \{1, ..., N\} \times \{1, ..., N\} \\
&Y\,(p, q) \sim \text{Bernoulli}\,(\mathbf{z}_{p\to q}^{\top} B\, \mathbf{z}_{p\leftarrow q}), \qquad \forall\, (p, q) \in \{1, ..., N\} \times \{1, ..., N\}
\end{aligned}
$$

The MMB model can be seen as a generalization of the SBM, where the same latent components can generate different networks among the same individuals. In fact, the group membership of each node is context dependent, that is, it varies when the node interacts with different nodes.

3.2.2 Variational Bayes Inference Algorithm

The Variational Bayes Method is an alternative to MCMC sampling methods like the Gibbs sampling (Attias, 1999). The main difference with the MC techniques consists in the fact that instead of sampling a numerical approximation of the true posterior distribution of the parameter, variational methods usually consider an auxiliary distribution of free latent parameters over the unknown parameters of interest, which can be seen as an approximation of the true posterior and find an optimal solution for this approx-

imation. In this sense, the variational methods can be seen as a form of Expectation – Maximization algorithm: an iterative method which develops in an E step, where the expectation of the log-likelihood is evaluated at the current estimated value of the parameters, and a step M, which maximizes the expected log-likelihood in the previous step with respect to the parameters and updates such values. By doing so we continuously adjust our approximation until it gets very close to the true posterior.
At the core of the variational methods there is the idea of minimizing the Kullback-Leibler divergence between the true posterior distribution and the approximate one, that is, a measure of the information lost when using the approximate distribution instead of the true. The KL divergence is minimized through a coordinate ascent algorithm which can be seen as a generalization of the EM algorithm.

The main advantage of using variational methods when compared with the Gibbs sampler and other MCMC algorithms comes from the fact that, when the parameter space to estimate is very large, the variational methods converge to a solution at a greater speed. On the other hand, the variational methods may be limited in the accuracy of their approximation because of the dependence on the chosen parametric family, while the Gibbs sampler simply generates sample from the true distributions, hence handling more complex posterior distributions. Therefore, the price to pay for a faster convergence is the risk of introducing biases. Finally, the variational methods make some specific assumptions about the approximate distribution, that is, it must be fully factorized, which may be impractical in some cases.

3.2.3 Parameters Estimation

To estimate the parameters $\boldsymbol{\alpha}$, B, π and the arrays containing $\mathbf{z}_{p\to q}$ and $\mathbf{z}_{p\leftarrow q}$ ($\forall p, q$), namely $Z_{\rightarrow}$ and $Z_{\leftarrow}$, a nested variational inference algorithm is implemented, based on the concept of variational Bayes method, using a set of free parameters, namely Γ and Φ. The prior distributions for $\boldsymbol{\pi}_{1:N}$, $Z_{\rightarrow}$, and $Z_{\leftarrow}$ are the following, based on these

parameters:

$\boldsymbol{\pi}_p \overset{\text{ind}}{\sim} \text{Dirichlet}(\boldsymbol{\gamma}_p)$, $\quad \forall p \in \{1, ..., N\}, \boldsymbol{\gamma}_p$ is a K-dimensional vector

$\mathbf{z}_{p \to q} \overset{\text{ind}}{\sim} \text{Multinomial}(\boldsymbol{\phi}_{p \to q})$, $\quad \forall p, q \in \{1, ..., N\}, \boldsymbol{\phi}_{p \to q}$ is a K-dimensional vector

$\mathbf{z}_{p \leftarrow q} \overset{\text{ind}}{\sim} \text{Multinomial}(\boldsymbol{\phi}_{p \leftarrow q})$, $\quad \forall p, q \in \{1, ..., N\}, \boldsymbol{\phi}_{p \leftarrow q}$ is a K-dimensional vector

Instead of alternating the two update steps (one for the free parameters of Z and one for the free parameters of π) as it is usually done in variational algorithms, they use a nested "for" loop to repeatedly update $\boldsymbol{\phi}_{p \to q}$ and $\boldsymbol{\phi}_{p \leftarrow q}$ and then update $\boldsymbol{\gamma}_p$. In doing so, the algorithm allocates less memory at each iteration and hence it converges faster to a solution. The pseudocode of the algorithm (Airoldi et al., 2008) is the following

---

initialize: $\gamma^0_{p,k} = \frac{2N}{K}$ for all p, k
repeat
  for $p = 1$ **to** N **do**
    for $q = 1$ **to** N **do**
      get variational $\phi^{t+1}_{p \to q}$ and $\phi^{t+1}_{p \leftarrow q} = f(Y(p, q), \gamma^t_p, \gamma^t_q, B^t)$;
      partially update γ^{t+1}_p, γ^{t+1}_q, and B^{t+1};
  until convergence

The step of the inner for loop to get the variational for ϕ is a nested iterative algorithm:

---

initialize: $\phi^0_{p \to q,g} = \phi^0_{p \leftarrow q,h} = \frac{1}{K}$ for all $g, h \in K$
repeat
  for $g = 1$ **to** K **do**
    update $\boldsymbol{\phi}^{s+1}_{p \to q} \propto f_1(\boldsymbol{\phi}^s_{p \leftarrow q}, \boldsymbol{\gamma}_p, B)$;
    normalize $\boldsymbol{\phi}^{s+1}_{p \to q}$ to sum to 1;
  for $h = 1$ **to** K **do**
    update $\boldsymbol{\phi}^{s+1}_{p \leftarrow q} \propto f_2(\boldsymbol{\phi}^s_{p \to q}, \boldsymbol{\gamma}_q, B)$;
    normalize $\boldsymbol{\phi}^{s+1}_{p \leftarrow q}$ to sum to 1;
  until convergence

---

where f_1 and f_2 are specific updating formulas and $\phi^0_{p \to q,g}$ and $\phi^0_{p \leftarrow q,h}$ are respectively the probability for node p to belong to group g when interacting with node q and the probability for node q to be in group h when interacting with node p (since $\phi^0_{p \to q}$ and $\phi^0_{p \leftarrow q}$ are vectors with the group membership probabilities of each cluster). Finally, to

derive the parameters of interest (B and $\boldsymbol{\alpha}$), an MLE approximation and a linear-time Newton-Rapson method can be respectively used, based on the estimation of the free parameters obtained in the algorithm. The approximate MLE of B is

$$\hat{B}(g,h) = \frac{\sum_{p,q} Y(p,q) \cdot \phi_{p \to q,g} \cdot \phi_{p \leftarrow q,h}}{(1-\rho) \cdot \sum_{p,q} \phi_{p \to q,g} \cdot \phi_{p \leftarrow q,h}}$$

where ρ is a sparsity parameter (Section 3.2.4) that modifies the probability parameter of the Bernoulli distribution over $Y(p,q)$.

The closed form solution for the approximate MLE of $\boldsymbol{\alpha}$ does not exist (Blei et al., 2003), but the Newton-Rapson method applied to it uses the following gradient and Hessian

$$\frac{\delta \mathcal{L}_\alpha}{\delta \alpha_k} = N \left(\psi \left(\sum_{k \in K} \alpha_k \right) - \psi(\alpha_k) \right) + \sum_{p \in N} \left(\psi(\gamma_{p,k}) - \psi \left(\sum_{k \in K} \gamma_{p,k} \right) \right),$$

$$\frac{\delta \mathcal{L}_\alpha}{\delta \alpha_{k_1} \alpha_{k_2}} = N \left(\mathbb{1}_{(k_1 = k_2)} \cdot \psi'(\alpha_{k_1}) - \psi'(\sum_{k \in K} \alpha_k) \right)$$

where N is the number of nodes and $\psi(x)$ is the derivative of the log-gamma function.

3.2.4 Modeling Sparsity

Adjacency matrices are often sparse in networks, that is, they present many zeros (non-interactions), but not all of them are significant for the model. Therefore, it is important to make a distinction between observed non-interactions caused by some limits which are not informative for the model and those which are due to the block structure itself. A possible solution is to include a sparsity parameter ρ in the model (Airoldi et al., 2008) as previously mentioned. Defining ρ as the portion of non-interactions which are not informative, the new probability of observing a successful interaction becomes $[(1-\rho) \cdot \mathbf{z}_{p \to q}^\top B \, \mathbf{z}_{p \leftarrow q}]$. By doing so, the whole parameter estimation changes: since the probability of observing an interaction is pre-multiplied by a factor between 0 and 1, the interactions will be more informative now, and hence will influence more the parameter estimation. In fact, by looking at the previous formula for the MLE of B, it is possible to notice that having $(1-\rho)$ at the denominator increases the value of the ratio, the numerator of which is a product of the total interactions and the free parameters. To estimate the best value of the sparsity parameter, an approximate MLE is proposed in

the following formula, using the estimated free parameters obtained in the variational algorithm

$$\hat{\rho} = \frac{\sum_{p,q}(1 - Y(p,q)) \cdot (\sum_{g,h} \phi_{p\to q,g}\phi_{p\leftarrow q,h})}{\sum_{p,q}\sum_{g,h} \phi_{p\to q,g}\phi_{p\leftarrow q,h}}$$

3.2.5 BIC Approach for Selection of Hyperparameter K

The BIC (Bayesian Information Criterion) is a technique base on the principle of model selection of trade-off between model complexity and goodness-of-fit to the data. In the context of block modeling, it can be used to estimate the number of communities in the network, namely K. The process generally consists in considering a set of possible values of K and fitting the model for each value, that is, finding the estimated parameters of the block structure. The BIC score is then computed as the log-likelihood of observing the data given the estimated parameters plus a penalty term for the number of parameters in the model. Applying the BIC criterion to the model discussed above, we obtain

$$BIC = 2 \cdot \log[Pr(Y|\hat{\boldsymbol{\pi}}, \hat{Z}, \hat{\boldsymbol{\alpha}}, \hat{B})] - (|\boldsymbol{\alpha}| + |B|) \cdot \log|Y|$$

where $|\boldsymbol{\alpha}| + |B|$ indicates the number of parameters in the model (in this case, $|\boldsymbol{\alpha}| = K$, $|B| = K^2$), and $|Y|$ is the number of observed interactions different from 0. The value of K which minimizes the BIC score is considered the best fitting value. The main idea behind the BIC approach is to penalize models with higher complexity, discouraging overfitting. On the other hand, the BIC approximation may not be accurate when the true underlying structure in the network is complex and it may cause underfitting. For this reason, we should take the value of K estimated by the BIC method as a starting point for further investigation on the true number of communities.

3.3 The SBM-MFM Model

As already discussed, one of the most evident limitations of the traditional SBMs is the uncertainty about the true number of communities in a network. The SBM-MFM model (Geng et al., 2019) makes a step forward and its contribution is two-folds: it allows to simultaneously infer the number of clusters K and the community membership by combining the SBM techniques with the Mixture of Finite Mixtures (MFM) approach, and it establishes a framework to consistently detect convergence to the correct configura-

tion by deriving non asymptotic bounds, according to which the posterior distribution on the community assignments concentrates on the true configuration as the size of the network N increases. The model proposed in the paper is the following:

$$
\begin{aligned}
&K \sim p(\cdot), \text{ where } p(\cdot) \text{ is a p.m.f. on } \{1, 2, ...\} \\
&B_{rs} = B_{sr} \overset{\text{ind}}{\sim} \text{Beta}(a,\, b), \qquad \forall r, s \in \{1, ..., K\} \\
&Pr(z_i = j \,|\, \boldsymbol{\pi}, K) = \pi_j, \qquad \forall j \in \{1, ..., K\},\, i \in \{1, ..., N\} \\
&\boldsymbol{\pi} \,|\, K \sim \text{Dirichlet}(\boldsymbol{\alpha}) \\
&Y_{ij} \,|\, \mathbf{z},\, B,\, K \overset{\text{ind}}{\sim} \text{Bernoulli}(B_{z_i z_j}), \qquad 1 \leq i < j \leq n
\end{aligned}
$$

where K is the true number of clusters, B is the probability matrix of observing an edge between a node in cluster r and one in cluster s, since the SBM assumption that such probability only depends on the community membership is valid, **z** is the vector of block membership, $\boldsymbol{\pi}$ is the probability vector containing the probabilities associated with each cluster, $\boldsymbol{\alpha}$ is a $K \times 1$ constant vector (all entries are equal), and Y is the adjacency matrix indicating the presence of an edge between each pair of nodes. The most suitable p.m.f. to choose for K is a Poisson (1) truncated to be positive (Geng et al., 2019). In fact, the Poisson distribution is often widely used in Bayesian nonparametric settings for modeling count data, including the number of clusters K in MFM; the reason behind that is the potentially unbounded number of clusters this distribution can support. Moreover, the parameter used in the Poisson distribution is clearly interpretable as the expected value and variance of the number of clusters, and setting such parameter to 1 ensures a moderate model complexity (number of communities), which reflects many realistic scenarios.

Another important innovation is the exploitation of the Pólya urn scheme properties in the context of MFM to develop a Gibbs sampler with K marginalized out. In doing so, it is possible to avoid using the reversible jump MCMC algorithm (Green 1995), which is more difficult to implement. The ultimate result is an efficient Gibbs sampler to obtain the posterior distributions of interest, namely, those of **z** and B.
It is worth noticing that the exploitation of the Pólya urn scheme properties is not strictly dependent on the choice of the pmf of K, but specific properties of this prior distribution

can positively affect the effectiveness of the Pólya urn scheme in the Gibbs sampler. In general, the scheme aligns with distributions that allow for unbounded number of categories, but it can work with any pmf by adjusting the probability of forming new clusters based on the probability mass not yet allocated to existing clusters.

The advantage of using the SBM-MFM model and the Gibbs Sampler above comes from the fact that it enables the possibility to consistently estimate K even when the sample is very large (as if it is infinite), while other models are able to estimate with a certain degree of uncertainty the number of communities mainly for a finite sample.
This capability stems from a nuanced understanding and utilization of the relationship between the posterior probabilities of the cluster assignments, **z**, and the number of clusters, K. In the SBM-MFM model, the asymptotic behavior of the posterior probability of **z**, is leveraged to inform the estimation of K. Specifically, this aspect is expressed by observing that both the posterior distributions of **z** and K tends to concentrate around the respective true values as N diverges. For **z** this implies that $Pr(\mathbf{z} = \mathbf{z}_0|Y) \to 1$ as $N \to \infty$, where $\mathbf{z}_0$ is the true underlying configuration. Similarly, $Pr(K = K_0|Y) \to 1$ as $N \to \infty$, where K_0 is the true number of clusters. The the Gibbs sampler ensures a mutual reinforcement in the estimation process of the two parameters:

- As the estimation of **z** becomes more accurate, the model improves its ability to identify distinct clusters based on the connectivity patterns. This improved clarity aids in more accurately estimates of K, as it becomes clearer how many distinct clusters the nodes are truthfully divided into;

- An improved estimate of K provides the correct range of values that each cluster assignment can take, guiding the estimation of **z**.

3.3.1 Consistency with Fixed and Known K Communities

First, it is important to clarify what ”consistency” means in the context of SBM, especially within a Bayesian framework. In Bayesian statistics, consistency refers to the property that as the size of the network N goes to ∞, the posterior distribution of the estimated parameters (in this case, the block structure of the network) converges to the true underlying parameters. This is a frequentist property of Bayesian procedures, implying that the Bayesian posterior concentrates around the true parameter value as more data is observed.
On the other hand, convergence in the context of an algorithm, such as MCMC (Markov Chain Monte Carlo), refers to the algorithm's ability to reach a stable solution as the number of iterations increases. This is about the algorithm's performance and stability, not the accuracy of the solution in estimating the true underlying structure. When K is assumed to be fixed and known, there are different approaches to detect and prove the consistency of the results.

As presented in Nowicki et al.(2001), it is possible to compute some measure of adequacy from the output of the Gibbs sampler. This process involves launching multiple independent runs of the sampler, each with a different initialization, and compare the values of the measures of interest. Convergence is assumed to be reached when a clear block structure is indicated across multiple runs by these measures of adequacy. When the sampler converges, the results are consistent. The measures of adequacy in question are indeed the information contained in the observed relations, I_y, expressed by the negative log-likelihood of the block structure and of the probability parameter of observing the relations, and H_x: an index measuring to which extent the distribution of **z** defines a clear partition of nodes into classes. It is computed as a function of $\theta_{ij} = Pr(z_i = z_j|Y)$, that is, the probability that two vertices i and j belong to the same class:

$$H_x = \frac{4}{N(N-1)} \sum_{i,j=1}^{N} \pi_{ij}(1 - \pi_{ij})$$

It takes value between 0 and 1, and values close to 0 indicates that the partition is estimated by the model with high confidence, while a value close to 1 shows a higher

level of uncertainty as it is not clear whether i and j belongs to the same cluster or not. However, this approach mainly ensures the convergence of the algorithm.

A more rigorous Bayesian approach is discussed in Geng et al. (2019), and it is based on the assumption that there exist an underlying true data-generating mechanism responsible for the true cluster assignment vector $\mathbf{z}_0$.
To deal with typical issues of SBMs like the identifiability problem (Section 4) the model employs a permutation-invariant loss function, specifically a modification of the Hamming distance, which counts discrepancies between two vectors of the same length. To simplify the discussion, it is possible focus on the case of homogeneous SBMs, characterized by a matrix B with all equal diagonal entries (p) and all off-diagonal entries equal (q). As discussed in the mentioned paper, as long as the prior probability of $\mathbf{z}$ given K is labelling invariant, so will be the posterior probability given Y, computed in accordance with the Bayes theorem. For this reason, a logical choice for the prior distribution given K is a Dirichlet-multinomial, with the probability vector following a symmetric Dirichlet distribution.

By leveraging common properties of homogeneous SBMs, it is possible to define an upper bound for the Bayes risk of their model, defined as the expected value of the permutation-invariant Hamming distance between the posterior cluster assignments z and the true partition z_0:

$$\mathbb{E}[d(\mathbf{z}, \mathbf{z}_0)|Y] \leq \exp\left\{-\frac{CN\overline{D}(p_0, q_0)}{K}\right\}$$

where $d(\mathbf{z}, \mathbf{z}_0)$ is the permutation-invariant Hamming distance between the true configuration $\mathbf{z}_0$ and the configuration $\mathbf{z}$ sampled from the model, C is a constant which does not depend on the other factors, $\overline{D}(p_0, q_0)$ is a relation between the true edge probabilities p_0 and q_0 (respectively, the within and between groups edge probabilities) such that $\frac{N\overline{D}(p_0,q_0)}{K} \to \infty$ as $N \to \infty$. The bound formulated above decreases exponentially to 0 as N diverges, implying that the marginal posterior distribution of the cluster membership almost surely concentrates on the true distribution at an exponential rate as

number of nodes increases:

$$\Pi[\langle \mathbf{z} \rangle = \langle \mathbf{z}_0 \rangle | Y] \geq 1 - \exp\left\{ -\frac{CN\overline{D}(p_0, q_0)}{K} \right\} \text{ almost surely as } N \to \infty$$

where $\Pi[\mathbf{z}|Y]$ is the posterior probability of $\mathbf{z}$, and $\langle \mathbf{z} \rangle$ indicates all possible permutations of a given $\mathbf{z}$. This result shows that the estimated block membership $\mathbf{z}$ is consistent by the definition of consistency in a Bayesian framework.

Another metric for evaluating clustering accuracy is the Rand Index (Rand, 1971), which computes the ratio of the node pairs which are consistently partitioned in the true configuration and the modelled one to the total number of possible node pairs in a graph with N nodes. As a result, the Rand Index is a number between 0 and 1, and the closer it is to 1, the more the two partitions are similar.

The consistency results defined in Geng et al. (2019) are more robust than those presented in Nowicki et al. (2001), as the latter bases the consistency of the model's results exclusively on the convergence of the MCMC algorithm, contrasting with the more comprehensive Bayesian approach that incorporates advanced probabilistic concepts proposed in the other paper.

3.3.2 Consistency with Unknown K Communities

When extending the discussion to scenarios where K is unknown and subject to a prior distribution, the challenge intensifies significantly. This issue forms a central debate in the realm of mixture models and stochastic block modeling, focusing on the feasibility of accurately identifying the true values of K and $\mathbf{z}_0$.

A simplified scenario where K adheres to a probability mass function (p.m.f.) defined solely over $\{2, 3\}$ allows for a more tangible exploration of these challenges. Within the framework of the SBM-MFM model, it has been demonstrated that it is feasible to adjust an initial estimate of K towards its true value, thereby showcasing an instance of model-selection consistency. This concept pertains to the capability of a statistical approach to accurately select the optimal model from a spectrum of potential candidates,

even when the initial predictions deviate from reality.

In contrast, the broader scenario, where K is distributed according to a p.m.f. over the set $\{1, 2, \ldots\}$, necessitates imposing an upper limit on the possible number of communities. This requirement emerges from the computational demand associated with reconciling the true number of communities K against every possible estimate. Such a process, involving comparisons across various configurations of estimated and actual values of K, imposes a significant computational burden. The SBM-MFM model's approach to this challenge involves sophisticated statistical techniques that aim to balance the accuracy of cluster recovery with the practicalities of computational resource limitations.

This nuanced exploration reveals the inherent complexities in achieving model consistency within SBMs, especially under conditions of uncertain cluster numbers. Geng et al. (2019) contribute to this ongoing discussion by proposing methodologies that, while computationally intensive, offer pathways to more accurate and consistent cluster identification in the face of uncertainty about K.

In summary, consistency in SBM under both fixed and unknown K scenarios requires rigorous methods to ensure that the posterior distributions of cluster memberships and the number of clusters accurately reflect the true underlying structure as the network size grows. This approach contrasts with merely achieving convergence of an algorithm, which, while necessary, does not alone guarantee the recovery of the true block structure.

4 PROBLEMS AND LIMITATIONS OF SBMs

4.1 Identifiability Issue

A recurrent challenge in SBMs is the so-called identifiability, that is, to uniquely determine the labels associated to the blocks estimated given the observed network. In fact, SBMs often present the so called label-switching problem, where different permutations of the labels associated to the clusters result in the same likelihood given the observed data and estimated parameters.

A direct implication of this problem is that there exist multiple distinct configurations that the observed data can generate without any loss in fit quality. The main consequence of the non-identifiability issue is the lack of interpretability of the results: the model does not provide a rigorous and distinct description of the communities detected, leading to uncertainty or potentially erroneous conclusions about what the output actually represent. This issue is also connected to the estimation via algorithms: the same network may lead to different community structures depending on the random seed of the algorithm, since most of the estimation processes uses stochastic procedures.

The identifiability problem becomes even more evident when the community structure is not too evident, or the boundaries between communities is not clear. For this reason, the MMB model explained in Section 3.2 is a clear example of this scenario: allowing for a more flexible model where each node may belong to different communities depending on who it is interacting with trades off the more realistic representation of how networks are in real world and an increasing vulnerability to the identifiability problem.

Although it is impossible to completely avoid this issue, there are several approaches that can help mitigate it in SBMs. First of all, the inclusion of metadata and prior knowledge about the possible community labels can help distinguish otherwise symmetric solutions. Another possibility is to use the estimated block membership after fitting the model as input in relabeling algorithms to align labels across different model fits. This approach is called ”post-hoc analysis” and is particularly useful when consistency of the cluster assignments is the final goal of the study.

4.2 Other Known Problems

4.2.1 Computational Complexity

The estimation processes used in SBMs, as seen in the previous sections, often involve the evaluation of likelihood functions and numerous iterations using `for` loops. When the input size, that is, the size of the network to analyse, increases, the computational complexity increases, leading to possible obstacles in the estimation process if the computational power is not enough. This issue is very common since the majority of the real-world networks usually involve large datasets.
A possible approach to solve the problem is to put significant effort in the improvement of the scalability of the process, that is, to make the algorithm more efficient so that it can deal with larger inputs. A stark example of this approach is the variational method described in Section 3.2.2, although it often trades off accuracy for speed and may not fully capture the complexity of the network structure.

4.2.2 Sensitivity to Model Specification

As already discussed in sections 3.3.1 and 3.3.2, the choice of the number of blocks K is crucial in SBMs for determining consistency, and it dramatically influences the outcome of the analysis. Choosing a too small K leads to underfitting, that is, the model is not able of capturing significant structural details in the network; on the other hand, picking a large value for K will increase the probability of overfitting, a phenomenon in which the model captures the intrinsic noise in data as a significant structural element. When a prior distribution for K is specified, as seen in the SBM-MFM model, the estimation is still very sensitive to the range of values of K determined by the distribution. In fact, it has been proved that it is computationally unachievable nowadays to consistently estimate the true value of K following an unbounded prior distribution, and the best results are obtained on priors over a very small set of values.

Some possible selection methods for K had already been mentioned, like BIC and cross-validation. The main disadvantage of these methods is their computational intensity, since they usually involve to valuate the model multiple times for each possible value of K in a given interval and find the most suitable one according to some mea-

sure of goodness-of-fit. Moreover, these techniques are not applicable to the so called ”dynamic networks”, datasets in which the structure and the interactions change over time. Making the choice of K dynamic and adaptive over time series data (Rastelli et al. 2018), meaning that both the true and estimated K are not fixed, but change over time along with the interactions in the network, is one of the biggest challenges in the field due to the time-consuming experimentation it requires. In this case, K takes value in a set and can be estimated using the model selection techniques already presented. Treating the case where K follows a prior distribution is too computationally expensive for the reasons showed in section 3.3.2.

5 REAL-DATA APPLICATION

To demonstrate the potential of Stochastic Block Models (SBMs), we implement a frequentist version of the model in `R` and apply it to a real dataset. Our goal is to assess the model's goodness-of-fit and consistency in capturing the underlying network structures.

5.1 The `war` Dataset

The chosen dataset is part of the `R` library `sbm` and contains two networks extracted from the Correlates of War website. The first network, known as "belligerent" (Sarkees et al., 2010), comprises 83 nodes representing countries. The edges between these nodes indicate that there has been at least one war between the connected countries during the years from 1816 to 2007. In SBM terms, this implies that $Y_{ij} = 1$ if and only if country i and j have been at war with each other.
The second network is called "alliance" (Gibler et al., 2009) and features 171 nodes. Each node again represents a country, and edges denote the existence of at least one formal alliance between the countries within the timeframe from 1816 to 2012.
It is noteworthy that the "belligerent" network includes fewer nodes than the "alliance" network because countries that never entered a war within the specified period are excluded from the network. Moreover, both datasets contain states which may not exist anymore: for instance, Bavaria, Baden and Wuerttemburg had been independent for years in the past 2 centuries before being annexed to Germany.

The dataset includes additional attributes for each country. The "power" attribute is a measure of increasing military capability, while "trade" reflects the intensity of trade relationships between pairs of countries. However, for the purposes of this thesis, these variables will not be included in the analysis. The focus will remain on the structural properties of the networks, keeping the analysis straightforward and manageable.

Both networks are structured as `igraph` objects from the homonymous library, which is a widely used method to represent and analyze networks in `R`. This format facilitates the manipulation and visualization of complex network data and will be instrumental in implementing and evaluating the SBM.

5.2 Model Implementation

The implementation of the SBM in `R` will involve fitting the model to both the ”belligerent” and ”alliance” networks separately to identify potential blocks or clusters of countries that exhibit similar interaction patterns. The goal of the analysis is to understand the geopolitical relationships and alliances by dividing the countries in groups which are historically consistent, in the sense that the results obtained align with the actual historical events and frameworks.

In particular, the results on the ”belligerent” network will focus on grouping the countries in classes of belligerence: in each class we expect to observe countries which actually took part to a similar number of wars, identifying those having a higher tendency to start a conflict. This result can be a good indicator for understanding the main actors in future wars. The analysis on the ”alliance” network aims at clustering countries in blocks that may have a similar cultural, or geographical background, or that for historical reasons ended up collaborating with each other for a significant amount of time.

5.2.1 Model Estimation & Evaluation

The estimation algorithm, based on the function `estimateSimpleSBM` of the `sbm` library, is a variational EM algorithm similar to the one explained in Section 3.2.2, while the model selection method for the most fitting number of clusters (the hyperparameter K) is the Integrated Classification Likelihood (ICL). This method is used to evaluate the fit of a model at different values for K while penalizing for complexity, effectively balancing the trade-off between model generalization and fitting to the data. This technique is very similar to the Bayesian Information Criterion of Section 3.2.5, but with some adjustments to address network models like SBMs. The likelihood L of observing a clustering configuration **z** given the parameters, can be written as

$$L(\mathbf{z}|Y, \boldsymbol{\pi}, B) = \prod_{i<j} B_{z_i,z_j}^{Y_{ij}} (1 - B_{z_i,z_j})^{Y_{ij}}$$

The ICL is defined as

$$ICL = \log L(\hat{\mathbf{z}}|Y, \hat{\boldsymbol{\pi}}, \hat{B}) - \frac{\lambda}{2} \log N$$

where $\log L(\hat{\mathbf{z}}|Y, \hat{\boldsymbol{\pi}}, \hat{B})$ is the log-likelihood of observing the configuration $\hat{\mathbf{z}}$ given the parameters estimated by the model, N is the number of nodes, λ is the total number of parameters of the model, hence being the sum of the number of elements present in B and $\boldsymbol{\pi}$. $\frac{\lambda}{2} \log N$ is a penatly term subtracted to the estimated log-likelihood to penalize too complex models, avoiding overfitting. The ICL is computed for different values of K and the one maximizing it is selected as the most suitable number of clusters.

5.3 Results on the Alliance Network

The model identified 11 blocks, reflecting a nuanced subdivision of alliances based on historical, cultural, or possibly economic similarities among countries. Each block can potentially represent a unique pattern of alliance formation that correlates with historical events, such as wars, treaties, or economic agreements.

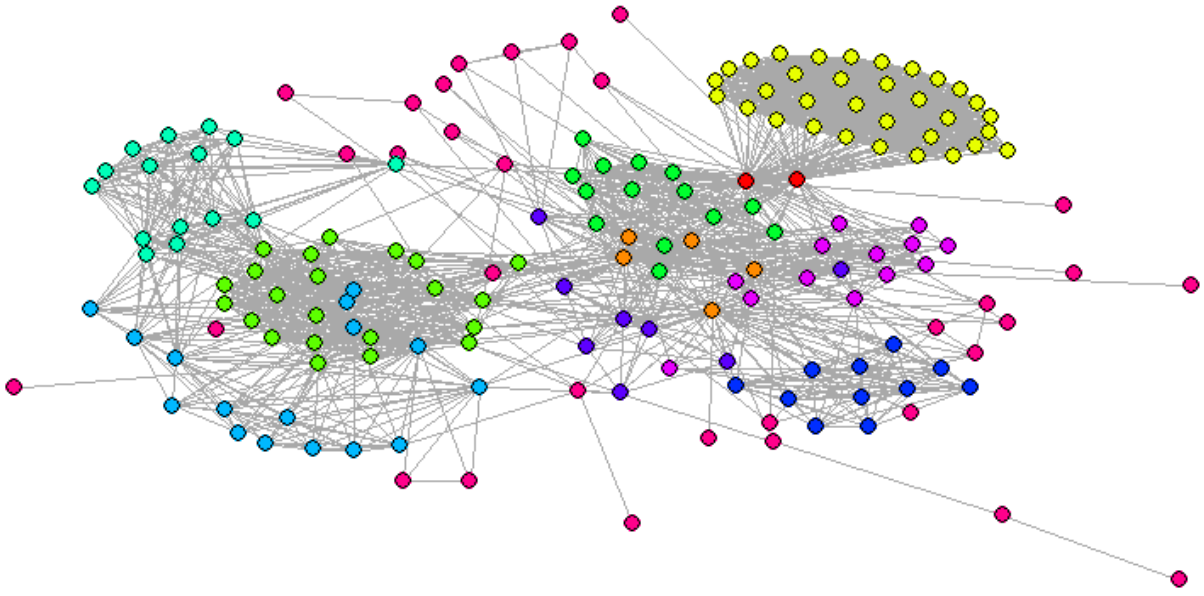

Figure 5: Visualization of the ”alliance” network with nodes colored by membership

Below is a table containing the countries in each block with a consistent explanation of the memberships.

Block	Members	Descriptions
1	United States, Canada	These countries are part of North America, sharing similar economic policies and defense strategies, often aligned in international policies.
2	UK, France, Poland, Russia, Turkey	Members have significant historical influence, playing crucial roles in European and global geopolitics, often involved in major wars and alliances.
3	Countries in Caribbean and Latin America	These nations share regional proximity and have similar post-colonial development paths, with many being part of trade blocs like CARICOM or OAS.
4	Middle Eastern countries	These countries share cultural, religious, and historical links, often involved in similar geopolitical conflicts and economic organizations like OPEC.
5	Western European countries	Highly integrated economically and politically, many are EU members with shared values on democracy and economic policies.
6	West African countries	Many of these countries share colonial histories, economic challenges, and are members of the Economic Community of West African States.
7	Central and East African countries	These nations often face similar developmental challenges and are part of regional bodies like the East African Community.
8	Former Soviet republics	Sharing a common post-Soviet transition history, these countries have political, cultural, and economic ties.
9	Eastern European and some Asian countries	This group includes countries that experienced communist rule and have transitioned to market economies, sharing a history of political upheaval and economic transformation.
10	Diverse Asia-Pacific countries	This group includes major economic players with robust development trajectories and significant influence in regional politics.
11	Diverse group from multiple continents	This eclectic group might share unique bilateral relations, historical ties, or common interests in global forums like the UN.

Table 1: Block Memberships and Descriptions

The plot below displays the estimation process for values of K ranging from 1 to 17. Notably, the points colored in red highlight the run for each each value of K which gave the highest ICL value. The concave curve delineated by these points indicates that the ICL value increases until $K = 11$, and then starts decreasing, indicating 11 as the most suitable number of clusters.

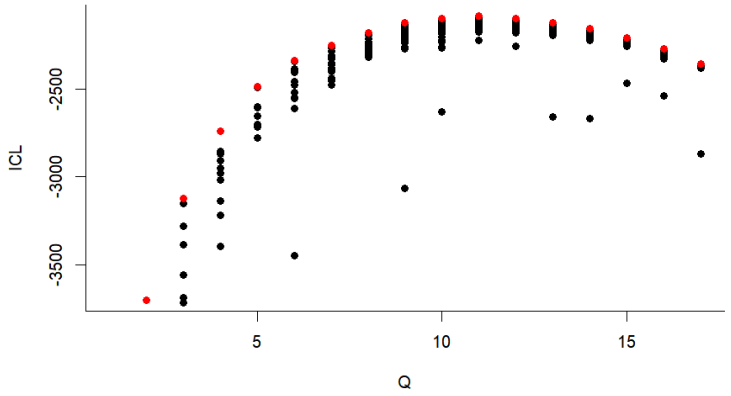

Figure 6: Integrated Completed Likelihood (ICL) values for different numbers of clusters (Q) in the model selection and estimation process for the ”alliance” network

To better visualize the results obtained, the data frame of the network is joint with an already built dataset storing the coordinates of a vast number of cities for each country of the world. This dataset allows to plot a world map with the borders between countries marked, and is combined with the block membership configuration.

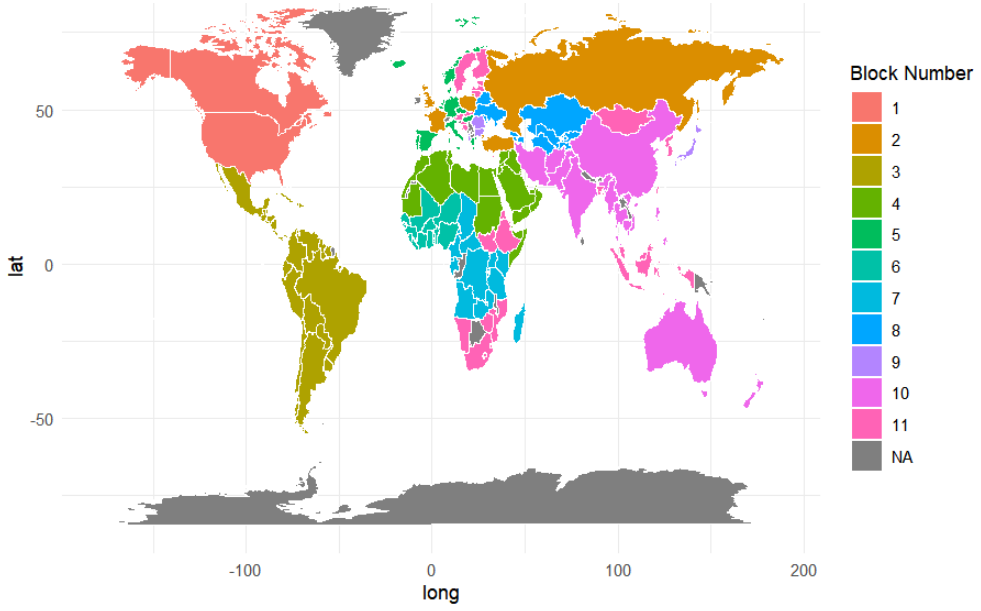

Figure 7: World map divided in blocks using the estimated memberships and the `world` dataset included in the library `ggplot2`. The countries in grey are not present in the ”alliance” network

5.4 Results on the Belligerent Network

The number of blocks estimated by the algorithm on the ”belligerent” network is smaller than on the ”alliance” one, with only 3 blocks identified. The main reason is that the network in question has less than half the number of countries present in the ”alliance” network, due to the fact that not all those countries were in war during the considered time frame. A natural consequence of a smaller network is a more simplified model, and probably less significant results. However, it is worth analysing the estimated block configuration to confirm that the results obtained are still consistent from an historical point of view.

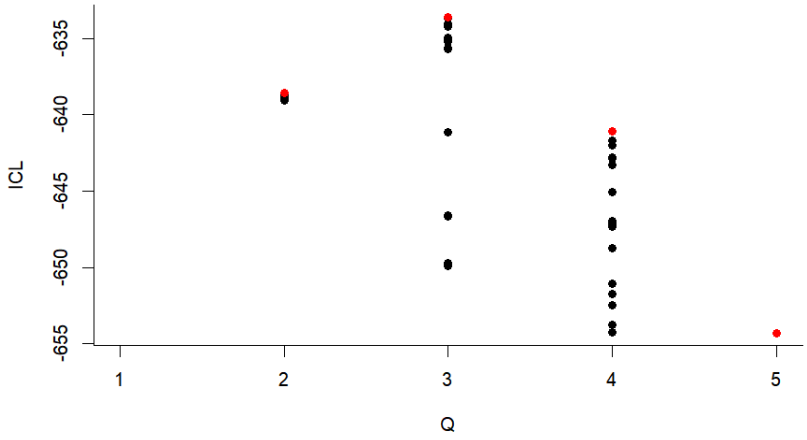

Figure 8: ICL values for number of clusters ranging from 1 to 5

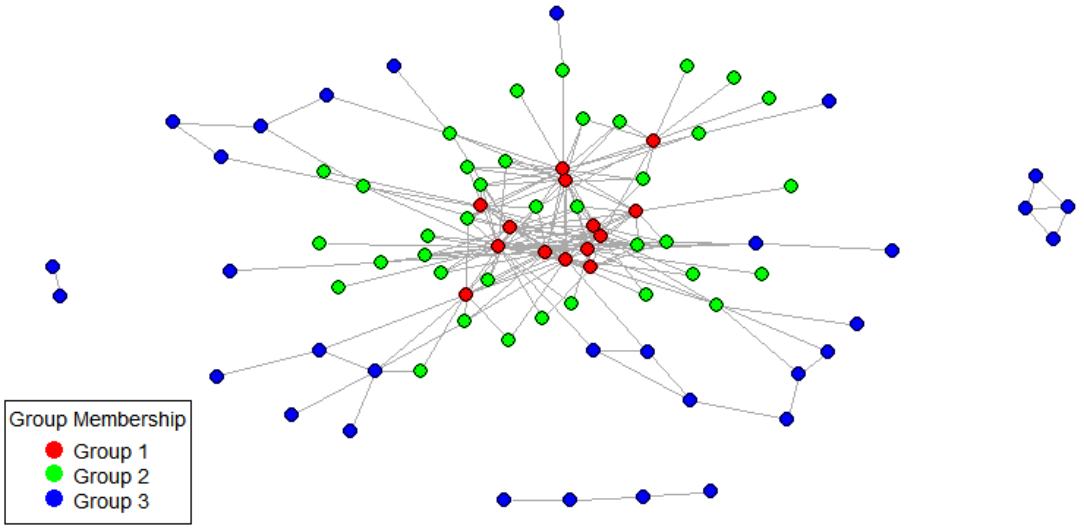

Figure 9: Network representation with colored block membership

Block	Members
1	United States of America, United Kingdom, France, Germany, Austria-Hungary, Italy, Bulgaria, Russia, Turkey, Iraq, China, North Korea, Japan, Vietnam
2	Canada, Mexico, Colombia, Netherlands, Belgium, Spain, Portugal, Bavaria, Baden, Wuerttemburg, Poland, Yugoslavia, Greece, Cyprus, Romania, Estonia, Latvia, Finland, Norway, Denmark, Ethiopia, South Africa, Iran, Egypt, Syria, Afghanistan, Mongolia, Taiwan, South Korea, India, Thailand, Cambodia, Laos, Republic of Vietnam, Philippines, Australia, New Zealand
3	Cuba, Guatemala, Honduras, El Salvador, Nicaragua, Ecuador, Peru, Brazil, Bolivia, Paraguay, Chile, Argentina, Hungary, Czechoslovakia, Lithuania, Armenia, Azerbaijan, Chad, Democratic Republic of the Congo, Uganda, Tanzania, Somalia, Eritrea, Angola, Morocco, Libya, Lebanon, Jordan, Israel, Saudi Arabia, Yemen Arab Republic, Pakistan

Table 2: Block membership of the countries in the ”belligerent” network

- *Group 1*: it consists of major global powers and countries involved in the largest conflicts in history. These countries are characterized by more dense connections with each other and with countries in other blocks, underlying the high bellicosity of this group. The average number of countries with which a member in group 1 has engaged war with is approximately 14.57.

- *Group 2*: it comprises Western and European democracies, many of which are military and economic alliances such as NATO and the EU. These countries adopted less belligerent politics and the estimated average number of countries engaged is 3.11.

- *Group 3*: this group mostly includes African, Latin American, and Middle East countries. These countries were characterized by a similar pattern of internal conflicts, revolutions, and were influenced by the Cold War politics. Despite the internal conflicts, this group present the lowest level of bellicosity since the wars were mainly local and did not involve many countries. The average number of countries engaged in war is estimated to be 2.03.

6 CONCLUSION

This thesis explores Stochastic Block Modeling, a powerful tool which can be used to analyse and understand the complex structure of relational data and networks, which have become an extremely important source of pivotal information. By deeply examining the mathematical and statistical fundamentals of SBMs, the purpose of this work is to show the models' ability to recover the latent block structures in large networks, providing useful insights about the relationships between nodes.

The introduction of basic notation in SBMs and the general review of the statistical concepts paved the way to more advanced techniques of mixture models, allowing for enhanced flexibility and robustness, ending up in the realm of Bayesian nonparametric statistics.

The core of the work is the literature review, which focuses on significant models that led to a turning point in the application of SBMs. Nowicki and Snijders were the first to build a model which produced consistent results using a Gibbs Sampling algorithm. The Mixed Membership Block model (MMB) provided a more realistic analysis of network structures by relaxing the strict assumption of a unique membership; moreover, the introduction of advanced inference techniques, such as the Variational Bayes algorithm, offers a computationally efficient alternative to traditional MCMC methods.
The MFM-SBM model introduces an additional layer of complexity by combining the foundations of SBM with innovative Bayesian nonparametric techniques and by providing pivotal insights into consistency with both known and unknown number of clusters K, setting the bases for a more rigorous framework to evaluate the performance of a model.

Despite the profound capabilities of SBMs, the thesis recognizes their limitations and presents the main challenges. The identifiability issue stands as the most significant problem, but it is worth mentioning the significant efforts that have been made in improving computational efficiency and sensitivity of the models. These challenges underscore the need for continuous research in network analysis in order to exploit the full potential of relational data.

Finally, the thesis presented an original contribution to the realm of SBM by applying the theoretical concepts presented to real-world data using a frequentist approach. The results were analysed both from a statistical and an historical perspective, extensively explaining how the results obtained from an estimation algorithm actually reflects real-world facts. These findings highlight the potential of SBMs to provide useful insights across various fields, from social sciences to biology.

Research is expected to continue in directions that include refining inference algorithms, exploring models that integrate several node attributes in addition to network interactions, and examining dynamic networks evolving over time.

References

[1] Airoldi, E. M., Blei, D., Fienberg, S., & Xing, E. (2008). *Mixed membership stochastic blockmodels*. Advances in Neural Information Processing Systems, 21.

[2] Attias, H. (1999). *A variational Bayesian framework for graphical models*. Advances in Neural Information Processing Systems, 12.

[3] Blackwell, D., & MacQueen, J. B. (1973). *Ferguson distributions via Pólya urn schemes*. The Annals of Statistics, 1(2), 353-355.

[4] Blei, D. M., Kucukelbir, A., & McAuliffe, J. D. (2017). *Variational inference: A review for statisticians*. Journal of the American Statistical Association, 112(518), 859-877.

[5] Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). *Latent dirichlet allocation*. Journal of Machine Learning Research, 3(Jan), 993-1022.

[6] Chan, T. K., & Chin, C. S. (2019). *Unsupervised Bayesian Nonparametric Approach with Incremental Similarity Tracking of Unlabeled Water Demand Time Series for Anomaly Detection*. Water, 11(10), 2066.

[7] Chiquet, J., Barbillon, P., & Donnet, S. (2020). *sbm: Stochastic blockmodels*.

[8] Ferguson, T. S. (1973). *A Bayesian analysis of some nonparametric problems*. The Annals of Statistics, 209-230.

[9] Geman, S., & Geman, D. (1984). *Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images*. IEEE Transactions on Pattern Analysis and Machine Intelligence, (6), 721-741.

[10] Geng, J., Bhattacharya, A., & Pati, D. (2019). *Probabilistic community detection with unknown number of communities*. Journal of the American Statistical Association, 114(526), 893-905.

[11] Gibler, D. M. (2008). *International military alliances, 1648-2008*. CQ Press.

[12] Gnedin, A., & Pitman, J. (2006). *Exchangeable Gibbs partitions and Stirling triangles*. Journal of Mathematical Sciences, 138, 5674-5685.

[13] Green, P. J., & Hastie, D. I. (2009). *Reversible jump MCMC*. Genetics, 155(3), 1391-1403.

[14] Holland, P. W., Laskey, K. B., & Leinhardt, S. (1983). *Stochastic blockmodels: First steps*. Social Networks, 5(2), 109-137.

[15] Karrer, B., & Newman, M. E. (2011). *Stochastic blockmodels and community structure in networks*. Physical Review E, 83(1), 016107.

[16] Lee, C., & Wilkinson, D. J. (2019). *A review of stochastic block models and extensions for graph clustering*. Applied Network Science, 4(1), 1-50.

[17] Leger, J. B. (2016). *Blockmodels: A R-package for estimating in Latent Block Model and Stochastic Block Model, with various probability functions, with or without covariates*. arXiv preprint arXiv:1602.07587.

[18] Nowicki, K., & Snijders, T. A. B. (2001). *Estimation and prediction for stochastic blockstructures*. Journal of the American Statistical Association, 96(455), 1077-1087.

[19] Pitman, J. (1995). *Exchangeable and partially exchangeable random partitions*. Probability Theory and Related Fields, 102(2), 145-158.

[20] Pólya, G. (1923). *On some problems of probability and statistics relating to the theory of irreversibility*.

[21] Rand, W. M. (1971). *Objective criteria for the evaluation of clustering methods*. Journal of the American Statistical Association, 66(336), 846-850.

[22] Rastelli, R., Latouche, P., & Friel, N. (2018). *Choosing the number of groups in a latent stochastic blockmodel for dynamic networks*. Network Science, 6(4), 469-493.

[23] Sarkees, M. R., & Wayman, F. W. (2010). *Resort to War: A Data Guide To Inter-State, Extra-State, Intra-State, And Non-State Wars*.

[24] Shao, C., Ciampaglia, G. L., Varol, O., Yang, K. C., Flammini, A., & Menczer, F. (2018). *The spread of low-credibility content by social bots*. Nature Communications, 9(1), 1-9.

[25] Tabouy, T., Barbillon, P., & Chiquet, J. (2019). *Variational inference for stochastic block models from sampled data*. Journal of the American Statistical Association.

[26] Wickham, H., Chang, W., & Wickham, M. H. (2016). *Package 'ggplot2': Create elegant data visualisations using the grammar of graphics. Version, 2*(1), 1-189.