Consistency and Quality of fit of Stochastic Block Models in Realistic Settings

By
Felipe Vaca-Ramírez

Submitted to
Central European University
Department of Network and Data Science

In partial fulfillment of the requirements for the degree of Doctor of Philosophy in Network Science

Supervisor: Prof. Tiago P. Peixoto

Vienna, Austria
2024

Researcher declaration

I, Felipe Vaca-Ramírez certify that I am the author of the work *Consistency and Quality of fit of Stochastic Block Models in Realistic Settings*. I certify that this is solely my own original work, other than where I have clearly indicated, in this declaration and in the thesis, the contributions of others. The thesis contains no materials accepted for any other degrees in any other institutions. I declare that no unidentified and illegitimate use was made of the work of others, and no part of the thesis infringes on any person's or institution's copyright.

Statement of inclusion of joint work

I confirm that Chapter 3 is based on a paper which was written in collaboration with Dr. Tiago P. Peixoto [1]. Dr. Peixoto and I conceived the idea of studying the quality of fit of SBMs in empirical networks using posterior predictive checks. He contributed with his expertise, suggesting experiments and analyses. I fitted the models, conducted the simulation experiments, computed indices, and processed the data. Both authors conducted the data analysis and contributed to the writing of the paper on which the chapter is based and gave final approval for publication. Dr. Peixoto endorses this statement with his signature below.

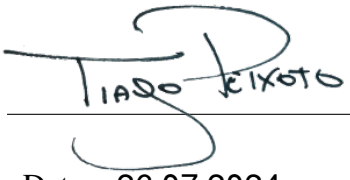
I confirm that Chapter 4 is based on a research project in collaboration with Dr. Tiago P. Peixoto. Dr. Peixoto conceived the idea of studying the reconstruction performance of SBMs in empirical networks. He contributed with his expertise, suggesting experiments and analyses. I fitted the models, conducted the simulation experiments, computed indices, processed and analyzed the data. Dr. Peixoto endorses this statement with his signature below.

I confirm that Chapter 5 is based on a research project in collaboration with Dr. Tiago P. Peixoto, Dr. Marta Sales-Pardo, and Dr. Roger Guimerà. Dr. Peixoto proposed the idea of revisiting and expanding a previous work of the aforementioned researchers [2], incorporating recent advances in statistical inference of networks. They contributed with their expertise, suggesting experiments and analyses. I fitted the models, conducted the simulation experiments, computed indices, processed and analyzed the data. Dr. Peixoto, Dr. Sales-Pardo, and Dr. Guimerà endorse this statement with their signatures below.

Signature of PhD Candidate:

Date: 26.07.2024

Signature of Dr. Tiago P. Peixoto, endorsing statement of joint work:



Date: 26.07.2024

Signature of Dr. Marta Sales-Pardo, endorsing statement of joint work:

MARTA SALES PARDO - DNI 36523875J

Digitally signed by MARTA SALES PARDO - DNI 36523875J
Date: 2024.07.22 10:11:40 +02'00'

Date: July 23rd 2024

Signature of Dr. Roger Guimerà, endorsing statement of joint work:

ROGER GUIMERA MANRIQUE - DNI 46355046H

Signed by: ROGER GUIMERA MANRIQUE - DNI 46355046H
Date: 2024-07-22 18:02:49 GMT+2

Date:

Abstract

The structure of real-world networked systems is crucial for understanding their origin, evolution, and behavior. Network structure can be summarized by decomposing the network into subsets of elements and assuming that the rate of interactions between individual elements is driven by such groupings. These groups, commonly referred to as "communities", play an important role in the network formation process and may significantly shape the behavior of the underlying system.

Generative network models are flexible and robust approaches to detect communities in network data. The family of Stochastic Block Models (SBMs), along with Bayesian inference tools, has proven useful for community detection and link prediction tasks. SBMs yield a coarse-grained description of the network data in a statistically principled way, which prevents drawing misleading conclusions due to spurious patterns, and simultaneously, allows the discovery of existing patterns in the data.

However informative, SBMs are approximations of real-world networks and rely on several simplifying assumptions that are unlikely to be valid in various empirical settings. Currently, neither the extent of these potential discrepancies in empirical network data nor the consequences that SBM modeling inconsistencies can introduce are well understood. This dissertation aims to address this issue by conducting large-scale studies of SBM fits to hundreds of empirical networks to uncover systematic patterns in SBMs performance. We consider two complementary approaches to assess the quality of the model, namely model checking and model selection.

In model checking, the goal is to understand how the model fails in describing the data, as a path towards model comprehension, revision, and improvement. To this end, we first use posterior predictive checks, which involves comparing networks generated by the inferred model with the empirical network, according to a set of network descriptors. Additionally, we conduct another study in a scenario with noisy network measurements, where we use a network reconstruction framework to test the accuracy of SBM estimates of underlying patterns of empirical networks. In both analyses, we observe that while the SBM provides accurate descriptions or estimates for most networks in the corpus, it does not fulfill all modeling requirements, particularly for transportation networks.

Finally, we study model selection approaches, considering several variants of the SBM. We evaluate the models based on their compression ability and predictive power, and examine the agreements and disagreements between these model selection criteria. Overall, we find consistency between such criteria, i.e., the most compressive model is also the most predictive. Nevertheless, compression criteria tend to be more reliable for model selection, as predictive criteria cannot always determine which SBM variant is better. Thus, this dissertation aims to provide a better understanding of the behavior of SBMs, their capabilities, and limitations as approximations of true underlying models of real-world networks.

Acknowledgements

This has been a long journey, and as I approach the end, I feel fortunate to have met people who supported me every step of the way. Without their help, I would not have been able to complete this work.

First, I thank my supervisor, Tiago P. Peixoto, for giving me the opportunity to delve into the field of statistical inference of networks and for teaching me invaluable lessons. His nearly infinite patience, guidance, commitment, and attention to detail have been crucial to my research.

I am grateful to my collaborators, Santo Fortunato, Roger Guimerà, and Marta Sales-Pardo, for sharing their expertise and providing valuable guidance on our research projects. I also thank Santo for hosting me during my visit to Indiana University, Bloomington, and for the opportunity to learn from other communities.

I want to thank my colleagues of the Department of Network and Data Science at Central European University for fostering a friendly and stimulating work environment. Special thanks go to my colleagues in the Inferential Network Science group — Martina Contisciani, Bukyoung Jhun, Sebastian Kusch, Sebastian Morel-Balbi, Thomas Robiglio, and Sina Sajjadi — for sharing their knowledge, helping to improve my dissertation, and, most importantly, for bringing more joy to both my life and research. Another special thanks to Abdullah Alrhmoun, Elsa Andres, Lisette Espín (+Reinhard), Martí Medina Hernández, Onkar Sadekar, Gergely Ódor (+Dori), and Manran Zhu for making me feel at home.

I would also like to thank the School of Informatics, Computing, and Engineering of Indiana University, especially Fan Huang, Zoher Kachwala, Sadamori Kojaku, Filipi Silva, and Attila Varga, for lifting my spirit with food, beer, and conversations during my visit to Bloomington.

I am deeply grateful to my parents, Silvia and Nelson, my grandfather Bienvenido, and my brother Eduardo, for their love, wholehearted support, and understanding from a corner of the Andes. I also thank Nacha, Pepe, Javier, Julieta, and Angie, for embracing me as one of their own.

Finally, I wholeheartedly thank my wife, Gabriela, for her immense love and patience, for believing in me, and for encouraging me to keep going. Words cannot express my deep gratitude, love, and admiration for you.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In this Chapter, we explain the motivation of this work (Sec. 1.1), the rationale of the approaches that we take to study SBMs (Sec. 1.2 and 1.3), and the summary of results and main contributions of this dissertation (Sec. 1.4).

1.1 Motivation

The structure of real-world networked systems, i.e., the connection patterns between the elements of the system, is of crucial importance to understand their origin, evolution, and behavior. One of most active directions of research to learn about network structure is based on the assumption that the network can be meaningfully divided into subsets of elements, commonly referred as "communities" or "blocks". These groups can play a significant role in network formation processes and have implications for the behavior of dynamic phenomena occurring on such networks.

In the past two decades, the task of finding such groups in networks, known as *community detection*, has received increasing attention. Various scientific domains have benefited from the application of community detection methods. Relevant examples include studying groups in social networks [3, 4], biological functions in metabolic networks [5, 6], fraud in telecommunications networks [7], and homology in genetic similarity networks [8]. Alongside the growing number of applications, many competing approaches have been proposed in the literature [4, 9], which differ not only on the motivations and goals they pursue, but also on how they define "community" [10]. One useful way to navigate through such diversity of approaches and gain intuition on how appropriate they are to describe the structure of real-world networks, is by using a statistical taxonomy. More specifically, we can distinguish between "descriptive" and "inferential" approaches to community detection [11].

Descriptive methods rely on heuristics to find a partition that fulfills some definition of com-

munity structure, focusing on a feature of the data rather than its generation. Often, these approaches have varied origins and goals, leading to disagreements on the meaning of community structure. Although the resulting partition can be used to describe the network, it remains unclear what is the role of the groups in forming connections between the elements of the system. Among these approaches, *modularity maximization* [12] is arguably the most popular one. Its goal is to find the partition that maximizes the modularity function, i.e., the number of connections within-groups minus the expected fraction of such quantity in a randomized version of the network. Despite its wide adoption, modularity maximization method suffers from several limitations. The most prominent problem from which this, and other descriptive methods, suffer is *overfitting*, which means that the algorithm finds spurious communities because it conflates structure with randomness in the data. In particular, this method finds partitions with high modularity in fully random graphs [13] and in graphs with non-assortative structures, such as lattices, trees, and tree-like networks [14, 15]. Additionally, this method suffers from a *resolution limit* [16, 17], i.e., it finds a number of groups no larger than $\sqrt{2E}$ in a connected network, being E its number of edges. This occurs because optimizing modularity in sufficiently large networks entails merging the small clusters. This behavior would, in turn, prevent the method from finding small communities in large networks, even if there is sufficient statistical evidence to support them. In other words, modularity maximization not only overfits, but it is also prone to *underfitting* the data.

In contrast, inferential approaches rely on generative models of network structure, explicitly incorporating modeling assumptions such as network formation mechanisms, prior information about model parameters, and data collection processes. These methods aim to infer the most likely *latent* groups of nodes which would have been responsible for the placement of edges in the observed network. The most prominent of inferential approaches is the family of Stochastic Block Models (SBMs) [18–22]. In its simplest version, the SBM divides the nodes in an undirected network into B groups, with the probability of having an edge between two nodes depending only on their group memberships. If we denote by b_i the group to which node i belongs, then we can define a $B \times B$ matrix $\mathbf{p}$, such that the matrix element $p_{b_i b_j}$ is the independent probability of having an edge between nodes i and j. In this way, SBMs can describe structures with arbitrary mixing patterns, such as assortative, bipartite, and core-periphery structures.[1]

The combination of SBMs with Bayesian inference [22] has proven powerful for analyzing network data. This framework allows us to be agnostic about what kind of structure is to be inferred and overcomes the limitations of descriptive approaches. Overfitting is addressed by incorporating regularization in the inferential framework [22], *via* the Minimum Description Length (MDL) principle [24], which prefers simpler hypotheses unless evidence in data supports more complex ones. This means that the description length also serves as a model selection criterion. Underfitting is tackled by incorporating suitable prior knowledge about the

[1] We refer the reader to Ref. [22, 23] for a description on different variants of SBMs.

parameters of the model, e.g., through a hierarchy of priors and hyperprior distributions [25]. This ensures that statistically significant patterns are uncovered, avoiding overly simplistic explanations. Thus, Bayesian SBMs protect against both overfitting and underfitting, preventing misleading conclusions when analyzing real-world networks.

However expressive and reliable our models are, it is important to evaluate whether SBMs provide an accurate description of real-world networks. Once an SBM is fitted, it is essential to understand the behavior of the model and assess its quality of fit to the data [26]. SBMs rely on several simplifying assumptions that may not hold in various empirical settings. Currently, neither the extent of these potential discrepancies in empirical network data nor the consequences that SBM modeling inconsistencies can introduce are well understood. This dissertation aims to address this issue by conducting large scale studies of SBM fits to hundreds of empirical networks to uncover systematic patterns in SBM performance. We consider two complementary approaches to assess the quality of the model, namely model *checking* and *model selection*. In the following sections, we provide further details on how we used these approaches to evaluate the quality of fit of SBMs.

1.2 Evaluating SBMs with Model Checking

Model checking consists on comparing the data to replicated data under the model. In our Bayesian framework, a useful and direct way of assessing the fit of the model to various aspects of the data is through *posterior predictive checking*. In Chapter 3, we use this approach to explore which aspects of empirical networks, according to a set of network descriptors, are not well described by the SBMs expectations. The goal is not to test whether the model's assumptions are "true", because all models are approximate. Instead, the goal is to assess exactly how the model fails in describing the data, as a path towards model comprehension, revision, and improvement.

While informative, posterior predictive checking might be overly simplistic as it does not fully reflect real-world situations. Many empirical studies rely on indirect measurements that yield noisy, incomplete, or unreliable network data. For example, measuring technological networks can involve incomplete sampling and technical limitations [27–29]; measurements of social networks might be affected by subjectivity, accuracy, and reliability of both participants and experimenters [30–32]; natural variation and inconsistent lab measurements might introduce significant variability and discrepancies in the measurement of biological networks [33–35]. Despite the pervasiveness of measurement errors in empirical studies, many practitioners neither acknowledge nor incorporate these aspects into their modeling frameworks. Instead, they assume that the measured network is the "true" underlying network, conduct the analysis, and draw conclusions, which might be erroneous or misleading [36–38].

In the network science literature, several attempts have been made to address this issue, with link prediction [39–42] being one of the most common approaches. The drawback of this method is that it does not explicitly incorporate a mechanism for measurement error. A more robust approach for obtaining the best possible estimates of network structure, given unreliable data, is *network reconstruction* [43, 44]. In this framework, it is possible to combine a model of measurement with a model of network structure, where the SBM is a suitable candidate for the latter [43] due to its high expressiveness. Most evaluations of network reconstruction or link-prediction methods are confined to relative comparisons between competing algorithms. Although the SBM has been shown to consistently outperform alternative methods for link prediction [42], evaluations of the reconstruction performance of the SBM in absolute terms are lacking. Specifically, we lack understanding on how accurate SBM estimates of underlying network patterns of empirical networks are. We address this issue in Chapter 4.

1.3 Evaluating SBMs with Model Selection

When evaluating SBMs with model checking, we considered one model class of the SBM. In practice, even when working with a single network, we often fit several models. Although these models might disagree with the network data in various ways, it might be still valuable to compare them. Thus, we can also evaluate the quality of a model by testing its performance against alternative models, and consequently, doing model selection. Two principled approaches for model selection are compression ability — where the best model is the one that compresses the data most effectively — and predictive power — where the best model is the one that is able to generalize from data and predict missing observations accurately.[2] In the community detection literature, some examples of using compression criteria can be found in Ref. [11, 49], while for predictive criteria in Ref. [42, 50, 51]. Both approaches aim to prevent overfitting by favoring the most parsimonious model that yields the best performance. Consequently, one might expect that the most predictive model among a set of alternative models is also the most compressive one. However, Vallès-Català *et. al* [2] showed that although compression and prediction are consistent in most networks in consideration, there are also notable instances where they do not agree. While their work offered valuable insights, it was constrained by the available tools at the moment, such as mostly using point estimates for compression indices and considering a few dozen empirical networks. Since then there have been significant advances in area of statistical inference of network structure. First, a robust network reconstruction framework for

[2]We discard other approaches to compare models due to their unrealistic assumptions. One such method involves measuring the agreement between the obtained partition and node metadata, which is assumed to represent "ground truth" communities. We refer the reader to Ref. [45, 46] for a discussion on how such comparison can be misleading. Additionally, other approaches use synthetic graphs or artificial benchmarks for comparisons [47], which may not be representative of real-world networks. For example, some of these benchmarks assume that the degrees are broadly distributed following a power law, yet one can find networks whose degree distribution significantly deviates from such assumption, as in the friendship networks of Ref. [48].

link prediction has been developed [43], which is summarized and used in Chapter 4. Second, a method has been introduced to characterize the posterior distribution of network partitions [52], which, in turn, can be used to approximate a measure of compression known as *model evidence*. Third, more efficient MCMC algorithms to sample from the posterior distribution of network partitions are now available [53]. Finally, access to hundreds of network datasets in network repositories has become possible [54]. In Chapter 5, we harness these innovations to revisit and extend upon the work of Vallès-Català *et. al* [2] in a more systematic way.

1.4 Outline and Contributions of the Dissertation

We start this dissertation by describing the methodological background of SBMs that is used throughout this work in Chapter 2. In particular, we refer to the generation and inference of SBMs, some relevant model variants, a model selection approach based on the Minimum Description Length principle, and the realism of the underlying modelling assumptions.

In Chapter 3, we perform a systematic analysis of the quality of fit of the SBM for 275 empirical networks spanning a wide range of domains and orders of magnitude in size. We employ posterior predictive model checking as a criterion to assess the quality of fit, which involves comparing networks generated by the inferred model with the empirical network, according to a set of network descriptors. We observe that the SBM is capable of providing an accurate description for the majority of networks considered, but falls short of saturating all modeling requirements. In particular, networks possessing a large diameter and slow-mixing random walks tend to be badly described by the SBM. However, contrary to what is often assumed, networks with a high abundance of triangles can be well described by the SBM in many cases. We demonstrate that simple network descriptors can be used to evaluate whether or not the SBM can provide a sufficiently accurate representation, potentially pointing to possible model extensions that can systematically improve the expressiveness of this class of models. The results of Chapter 3 have been published in the following article:

> *Systematic assessment of the quality of fit of the stochastic block model for empirical networks*. Felipe Vaca-Ramírez & Tiago P. Peixoto. Physical Review E. 2022. [1].

In Chapter 4, we assess the performance of the stochastic block model (SBM) in reconstructing 248 empirical networks spanning several domains and orders of size magnitude. We simulate a noisy measurement process and evaluate the model's ability at recovering various descriptors of the network structure. We observe that the SBM yields accurate estimates for most networks in the corpus, but this behavior is not ubiquitous. In particular, we mostly observe large reconstruction errors in networks having large diameter and slow-mixing random walks — corresponding typically to networks embedded in space. Contrary to what is often assumed, the

SBM is able to provide accurate estimates on networks with a high abundance of triangles. We also demonstrate that incorporating a more detailed error assessment while doing measurement tends to improve the quality of the reconstruction.

In Chapter 5, we study the agreements and disagreements between compression criteria and predictive criteria for selecting a model variant of the SBM while performing community detection in networks. We consider a corpus containing 392 empirical and synthetic networks, and fit two SBM variants to them. Then we obtain compression and predictive indices, and select the best model according to them. This allows us to determine whether the most compressive model is the same as the most predictive one or not, when disagreements occur, and in which magnitude. For synthetic networks, we find consistency between model selection criteria, i.e., the most compressive model is also the most predictive one, while for empirical networks, consistency is often the case, with few exceptions. Although agreements between model selection approaches are quite frequent, we observe that predictive criteria cannot always tell which model is better, since there are many cases in which the AUC of competing models is statistically equivalent. On the contrary, both the description length or evidence always tells us which model compresses more the data, and provides a degree of confidence for ruling out the alternative model. In that sense, compression criteria would be more a reliable approach for model selection in the context of community detection.

In writing this dissertation, we have aimed to provide a better understanding of the behavior of SBMs in empirical settings. In particular, we focused on the capabilities and limitations of SBMs as approximations of true underlying models of real-world networks. In turn, this should have provided insights on which improvements may be necessary to be introduced in the models. Additionally, we payed attention to model selection approaches of SBMs, which can be viewed as another way to understand and compare multiple model variants fitted to the same data. Since both model selection and model checking should go hand-in-hand when analyzing the structure of networks, we hope that, by reading these pages, the reader also feel motivated to think about the assessment of network models, and more broadly, about the whole process of analysis of network data. Nevertheless, because of the focus of the dissertation, there have inevitably been extensions and topics that are beyond its scope. Therefore, in Chapter 6, we provide conclusions of this work and comments on future directions of research.

Chapter 2

Methodological Background

In this chapter, we present relevant terminology and technical background used in this dissertation. Specifically, in Sec. 2.1, we refer to networks and network data. In Sec. 2.2, we refer to network models, focusing on two relevant examples. Finally, in Sec. 2.3, we refer to Stochastic Block Models, which are central to this dissertation. We note that the purpose of this chapter is to provide a common framework rather than an exhaustive treatment of the aforementioned topics. For a more detailed treatment of the first two sections, we refer the reader to Ref. [55, 56], and for the latter section, to Ref. [22, 23, 57].

2.1 Elementary Network Theory

A *network* or *graph* is a mathematical object $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ formed by a set of vertices (or nodes) $\mathcal{V}$ and a set of edges (or links) $\mathcal{E}$, where an edge is an unordered pair of vertices (i, j), such that $i \neq j$. The number of nodes N and the number of edges E are sometimes referred as the *size* of the network.

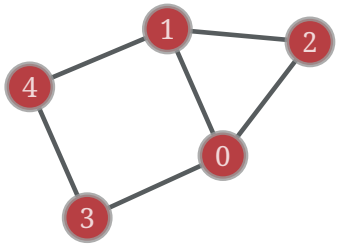
There is an important connection between graph theory and matrix algebra that offers tools to characterize graphs, and in general, treat them rigorously. In fact, a graph can be fully defined by its *adjacency matrix* $\boldsymbol{A} = \{A_{ij}\}$ of dimension $N \times N$, where

$$A_{ij} = \begin{cases} 1, & \text{if there is an edge between nodes } i \text{ and } j. \\ 0, & \text{otherwise.} \end{cases}$$

Furthermore, a graph can be represented graphically, as shown in Fig. 2.1.

This is the simplest type of graph, being called *simple* graph, and there exist several extensions of the concepts mentioned above to incorporate other features. If the graph contains self-loops and multi-edges, it is called *multigraph*. Furthermore, if edges contain directions, i.e., the pair

(a)



(b)

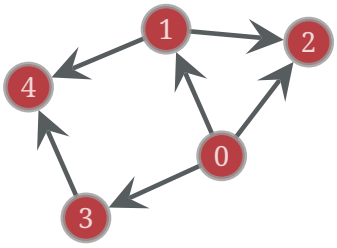
$$\boldsymbol{A} = \begin{bmatrix} 0 & 1 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 \end{bmatrix}$$

Figure 2.1: (a) An example of a graph having $N = 5$ nodes and $E = 6$ edges, in which nodes are labelled by an index. (b) The adjacency matrix of the graph shown in (a).

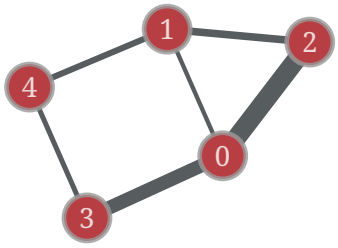
(i, j) is ordered and indicates that the edge goes from i to j, then we have a *directed* graph. If edges contain weights, then we have a *weighted* graph. Note that these features are not exclusive, but can be combined.

Since a graph is connected to an adjacency matrix, the latter is also modified accordingly. In the simplest case, the adjacency matrix is symmetric, but for directed graphs it is not. Additionally, entries are binary for simple graphs, while in weighted graphs they can be integers or real numbers. In the simplest case, the diagonal contained zeros, but in multigraphs, this is not necessarily the case.[1] Some examples are shown in Fig. 2.2.

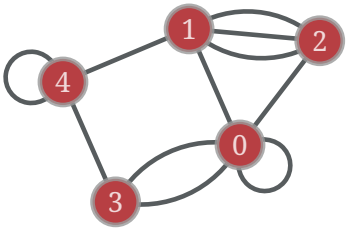
(a) directed



(b) weighted



(c) multigraph



(d)

$$\boldsymbol{A} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 \end{bmatrix}$$

(e)

$$\boldsymbol{A} = \begin{bmatrix} 0.0 & 1.2 & 5.3 & 3.8 & 0.0 \\ 1.2 & 0.0 & 2.1 & 0.0 & 1.7 \\ 5.3 & 2.1 & 0.0 & 0.0 & 0.0 \\ 3.8 & 0.0 & 0.0 & 0.0 & 1.4 \\ 0.0 & 1.7 & 0.0 & 1.4 & 0.0 \end{bmatrix}$$

(f)

$$\boldsymbol{A} = \begin{bmatrix} 2 & 1 & 1 & 2 & 0 \\ 1 & 0 & 3 & 0 & 1 \\ 1 & 3 & 0 & 0 & 0 \\ 2 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 & 2 \end{bmatrix}$$

Figure 2.2: (a-c) Examples of directed graph, weighted graph, and multigraph, respectively. (d-f) The adjacency matrices of the graphs shown in (a), (b), and (c).

An important feature of a vertex is its *degree*. The degree of a vertex i in a simple graph is defined as the number of connections it has, i.e.,

$$k_i = \sum_{j=1}^{N} A_{ij}.$$

[1] For multigraphs, self-loops appear in the adjacency matrix with a value of 2, since for every self-loop, there are 2 half-edges incident to the node.

The sum of all degrees is equal to twice the number of edges, i.e., $\sum_{i=1}^{N} k_i = 2E$. Additionally, the *average degree* summarizes the *density*, or equivalently the *sparsity*, of the graph,

$$\begin{aligned}\langle k \rangle &= \frac{1}{N}\sum_i k_i \\ &= \frac{2E}{N}.\end{aligned}$$

In the case of other types of graphs, such as directed and weighted graphs, the definition of degree is modified accordingly. In this dissertation, we only deal with simple graphs.

A more informative picture of the graph connectivity can be obtained by looking at the distribution of vertex degrees, or *degree distribution*. This distribution indicates the probability that a randomly selected node has degree k. Importantly, its shape and broadness may also give insights about the process of formation or the robustness of a network. We will refer to some cases of interest in Sec. 2.2.

2.1.1 Network Data

The term *network* has several uses in the scientific literature, depending on the context in which it appears. In the field of network science, a network is a representation of a complex system. Specifically, nodes represent elements of the system and edges represent interactions or relations between these elements. Therefore, how we define nodes and edges in the network becomes an important choice, as it influences the way the network will be analyzed and, ultimately, the conclusions drawn about the system. The term *network data* corresponds to the measurements of a system conceptualized as a network, or the behavior coming from it. In other words, a network is constructed from network data. Although we acknowledge the distinction between these terms, and recognize that we almost always deal with data rather than the underlying network, we use them interchangeably for ease of exposition.

The collection and analysis of network data dates back to at least to 1930s, with the works of Helen Jennings and Jacob Moreno [58, 59]. Since the beginning of the 21st century, there has been a surge in the applications that involve networks, so one can find examples of them in a variety of contexts.

In biology, molecular biologists study protein-protein interaction networks, in which nodes are proteins, and a link represents an interaction between proteins, whose measurements come from experiments [60]. Neuroscientists study brain networks, in which nodes represent neurons or brain regions, and an edge represents an anatomical or functional connection. The data sometimes is derived from functional MRI or magnetoencephalography [61]. Ecologists study

food webs, in which nodes represent species, and edges represent feeding of one species on another [62].

In the social sciences, sociologists study interactions or relations (e.g., friendship) between people or groups of people. Surveys are often conducted to collect data [48]. Political scientists are interested in political discussion and collective demonstrations. Examples include the study of a network of political blogs, in which nodes are blogs, and edges are references between blogs [3], and the study of online social media networks during public protests, in which accounts are nodes, and there is a directed edge if one account follows another [63]. Data collection methods for such studies include web crawling and APIs usage.

In engineering, researchers study technological networks or networks of physical infrastructure. One relevant example is the Internet, where nodes represent computers or related devices, and edges represent physical connections between them [64]. Another example corresponds to transportation networks, such as urban street networks. Nodes can represent street junctions and edges can represent street segments. They are often obtained by processing maps [65–67].

Besides the diversity of sources and data collection techniques, network data have other characteristics that introduce intricacies in their analysis [56]. The first is high dimensionality. Network data not only contains nodes and edges, but also attributes on each of them (e.g., demographic characteristics of people or strength of relations), on parts of the graph, or even on the entire graph. A combination of such characteristics is sometimes represented as a multilayer (attributed) network [51, 68–70].

The second aspect that brings challenges to analyzing network data is dependency. The creation of a link might be influenced by other already existing links (e.g., via triadic closure [49,71,72]), node attributes (e.g., via homophily [73–76]), or a combination of both. Furthermore, networks can change in time [77–79], which might introduce a temporal dependency.

The third aspect is the size of datasets, which are currently much larger than in the past, with some datasets containing millions of nodes[2]. Some approaches to analyzing network data involve taking parts of the graph and aggregating the results to draw conclusions about the whole system. These approaches can be potentially misleading, as taking subsets of data might distort the structure of the network and the behavior of dynamical processes occurring on it in unexpected ways. Therefore, designing efficient algorithms to analyze network data is still a non-trivial task and remains an active area of research.

Finally, network data might contain errors (data is noisy), or there might be some parts of the graph that have not been measured (missing data). Sometimes, measurements of a network might come indirectly, e.g., by thresholding matrices whose entries indicate some relation between pairs of elements. The structure of the resulting graph might be sensible to the imposed threshold, which in turn, may impact the robustness of conclusions. Developing modeling

[2]For instance, see Ref. [80, 81]

frameworks that incorporate these aspects into the analysis is another area that offers avenues of research.

2.2 Network Models

An important part of the foundations of network science corresponds to network models. Here, we are especially interested in models of network structure, i.e., models whose goal is to generate patterns of connection similar to those observed in real-world networked systems, and subsequently, understand the implications of such patterns in the behavior of the system under study. Of particular interest are the so-called *random graph models*, i.e., probabilistic models that, given a pre-specified set of parameters or network features, randomly generate a collection (or *ensemble*) of networks.[3]

These models serve various purposes in the study of networks. Due to their relative simplicity, structural and dynamical properties can be derived analytically, offering insights about the interplay between structure and dynamics in networks. Consequently, they have also been used as a starting point for understanding a variety of dynamical processes on networks, before considering other networks generated with more complicated models or even real-world networks. Sometimes, these models are also used as a reference or "null" model to test the "significance" of a network property. In such cases, the property measured in an empirical network is compared with the property computed across the ensemble of networks. Examples of such applications include the detection of motifs [92, 93] and, as mentioned earlier, community detection based on modularity maximization [12]. Finally, they are also incorporated in studies either as benchmarks for testing the accuracy or performance of other models or as parts of the sampling design or estimation strategy, e.g., in the inference of population graph parameters [94].

In the remaining of this section, we refer to two of the most widely studied and simplest random graph models, namely the Erdős-Rényi Model and the Configuration Model. In the next section, we refer to the Stochastic Block Model (SBM), which can be seen as a generalization of these models.

[3]There are several relevant models of network structure that are beyond the focus of this thesis. Relevant examples include the preferential attachment models [82–84], whose goal is to generate networks having a power law degree distribution; the Watts and Strogatz model [85], which attempts to reproduce high clustering (or abundance of triangles) observed in real-world networks; Latent Space Models [86–88], which embed nodes in a lower-dimensional latent space to capture network structure; and the family of Exponential Random Graph Models [89–91], which are statistical models of network structure relying on local edge-based structures.

2.2.1 The Erdős-Rényi Model

Given a certain number of nodes N and edges E, one of the simplest ways to generate networks is by choosing E node pairs uniformly at random from all possible pairs, and creating an edge for each of them. This model is known as the $G(N,E)$ model and each graph $\boldsymbol{G}$ in the ensemble has a probability of being drawn with probability $P(\boldsymbol{G}) = 1/\Omega$, where

$$\Omega = \binom{\binom{N}{2}}{E}, \tag{2.1}$$

$\binom{N}{2}$ is the binomial coefficient that indicates the total number of node pairs, and E is the number of edges in the graph.

Another version of this model is the so-called $G(N,p)$ model, which sometimes is preferred to the $G(N,E)$ model because some calculations are easier to handle. In the $G(N,p)$ model, the number of edges is not fixed. Instead, for a given number of nodes N, each edge is placed with a probability p. Thus, the probability of a simple graph $\boldsymbol{G}$ is given by

$$P(\boldsymbol{G}) = p^E (1-p)^{\binom{N}{2}-E}. \tag{2.2}$$

The study of this model can be traced back to at least the works of Solomonoff and Rapoport (1960) [95] and Gilbert (1959) [96]. However, in the literature, it is commonly referred as the *Erdős-Rényi Model* [97], due the seminal contributions of Paul Erdős and Alfréd Rényi to the model [97–99].

Considering Eq. (2.2), the probability of observing a simple graph having N nodes and E vertices is given by the following binomial distribution,

$$P(E) = \binom{\binom{N}{2}}{E} p^E (1-p)^{\binom{N}{2}-E}. \tag{2.3}$$

Thus, the expected number of edges in the $G(N,p)$ model is given by

$$\langle E \rangle = \sum_{E=0}^{\binom{N}{2}} E\, P(E) = \binom{N}{2} p. \tag{2.4}$$

In the previous section, we mentioned that the average degree in a graph can be computed as $\langle k \rangle = 2E/N$. Using this relation, the mean degree of a node in this model is given by

$$\langle k \rangle = \frac{2\langle E \rangle}{N} = \frac{2\binom{N}{2} p}{N} = (N-1)p. \tag{2.5}$$

Finally, we focus on the degree distribution of the $G(N,p)$ model. Since each node can connect to another node with probability p, the probability that a given node is connected to a specific set of k other nodes is given by

$$p^k(1-p)^{N-1-k}. \tag{2.6}$$

Since the number of ways in which k nodes can be chosen from the $N-1$ total candidates is $\binom{N-1}{k}$, the total probability of a node being connected to k others is

$$P(k_i = k) = \binom{N-1}{k} p^k (1-p)^{N-1-k}, \tag{2.7}$$

which means that the degree distribution of the $G(N,p)$ model is binomial.

If we assume that the number of nodes N is large, i.e., $N \to \infty$, so that the probability of connection $p = \langle k \rangle / (N-1)$ becomes vanishingly small; and $\langle k \rangle$ is fixed, it can be shown that the degree distribution of Eq. (2.7) becomes a Poisson distribution. It is for this reason that the $G(N,p)$ model is sometimes referred in the literature as the Poisson random graph. A reference to other properties of this model can be found in Ref. [55].

2.2.2 The Configuration Model

The Configuration Model [100, 101] is a generalization of the Erdős-Rényi Model, in the sense that, it is not restricted to a Poissonian degree distribution, but admits arbitrary degree distributions. More precisely, this model generates networks with a fixed *degree sequence*, i.e., the list of degrees of each node in the network, instead of a pre-established degree distribution.[4] The procedure used to generate a network having a degree sequence relies on "stub matching", which we describe in the following.

Assume that, for every node i, we fix its degree k_i. This assumption, in turn, fixes the number of edges E, since $2E = \sum_i k_i$. Another way to interpret this assumption is that each node i has k_i labelled half-edges or "stubs", existing a total of $2E$ stubs in the network. We form an edge by choosing two stubs uniformly at random and connecting them. Then we repeat this procedure with the $2E-2$ remaining stubs, and continue until all stubs have been matched or paired. In this way, the configuration model generates multigraphs, i.e., networks that may have self-loops and multiedges, since there is no restriction for its formation. This might turn the configuration model into an unrealistic model for real-world networks. However, this might

[4] It is possible to adapt the model to the case when only the degree distribution is known. The idea is to draw a degree sequence from the specified distribution, and use such sequence along with the configuration model to generate the network.

not represent a problem, because in sufficiently large and sparse networks, the probability of finding a self-loop or a multiedge between any two specific nodes tends to zero.

It should be noted that, the configuration model defines an ensemble of pairings, in which, each possible pairing has the same probability of being drawn. Nevertheless, since stubs are labelled, different matchings can create the same network, and consequently the networks in the ensemble do not have the same probability of being drawn. The reason is that a permutation of the stubs at each node creates the same graph. Thus, the probability of drawing a graph $\boldsymbol{G}$ under the configuration model is given by the ratio between the number of matchings ν corresponding to $\boldsymbol{G}$, and the total number of matchings Ω in the ensemble, i.e., $P(\boldsymbol{G}) = \nu/\Omega$. In principle, there are $\prod_i k_i!$ matchings for a given network. However, for each multiedge, there are $A_{ij}!$ permutations of stubs at one end, which do not generate new matchings. Furthermore, for each self-loop, there is a further factor of two since permutations on both ends do not generate new matchings either. Therefore, the number of matchings corresponding to a network is given by

$$\nu = \frac{\prod_i k_i!}{\prod_{i<j} A_{ij}! \prod_i A_{ii}!!}, \tag{2.8}$$

where $n!! = n(n-2)(n-4)...2$, with n even. The total number of matchings Ω is given by

$$\Omega = \frac{(2E)!}{E! 2^E \prod_k (k!)^{N_k}}, \tag{2.9}$$

where N_k is the number of nodes having degree k. The edge probability in the configuration model can be also derived by looking at stubs. Consider a pair of nodes i and j, having k_i and k_j stubs, respectively. If we take one stub of i, there are in total $2E-1$ other stubs to form a connection, from which only k_j belong to node j. Since there are k_i possibilities in which we could have chosen the initial stub of node i, the probability of a connection between nodes i and j is given by

$$p_{ij} = \frac{k_i k_j}{2E-1}. \tag{2.10}$$

It should be noted that Eq. (2.10) corresponds to the expected number of edges between nodes i and j rather than the probability of having an edge between those nodes. However, they coincide when $E \to \infty$ and k_i and $p_{ij} < 1$ for given k_i and k_j. This equation does not hold for self-loops; for further details on this and other aspects of the model, we refer the reader to Ref. [55].

We conclude this section by noting that the configuration model provides useful tools to study structural properties of networks, such as degree distributions, clustering coefficients, and connected components. Many researchers have used this model, often in combination with power

law degree distributions to model real-world networks having a heterogeneous distribution of degrees. However, despite its usefulness, the configuration model still contains assumptions that are unrealistic for real-world networks. For instance, it generates networks with vanishingly small clustering as the number of nodes increase, and it does not incorporate correlations in the network generation process. Therefore, more complex models are needed to learn the structure of real-world networks. We devote the next section to one of such models, namely the Stochastic Block Models. They are not only generalizations of the Erdős-Rényi Model and Configuration Model, but also the focus of this dissertation.

2.3 Stochastic Block Models

Stochastic Block Models are probabilistic models of network structure. They allow us not only to generate networks with arbitrary group structure, but also to infer such structure when we provide a network as input. In the following, we present foundations of SBMs that are relevant for this dissertation: an overview of the model from the perspective of generation and inference (Sec. 2.3.1 and 2.3.2), microcanonical formulations of the model (Sec. 2.3.3 and 2.3.4), a connection between inference and information theory (Sec. 2.3.5), MCMC methods for inferring network partitions (Sec. 2.3.6), and a brief discussion on how realistic SBM assumptions are (Sec. 2.3.7). We refer the reader to Ref. [22, 23, 57] for further details of these models.

2.3.1 Generative Model

The stochastic block model is a generative model for blocks, groups, or communities in networks. In its simplest version, it takes as parameters the partition of the nodes into B groups, denoted by a vector $\boldsymbol{b}$, with $b_i \in \{1,\dots,B\}$, and a $B \times B$ matrix of probabilities $\boldsymbol{p}$, where p_{rs} is the independent probability of having an edge between a node from group r and a node from group s. This means that the probability of having an edge between nodes i and j only depends on their group membership. Thus, a network having adjacency matrix $\boldsymbol{A}$ is generated according to the following likelihood

$$P(\boldsymbol{A}|\boldsymbol{p},\boldsymbol{b}) = \prod_{i<j} p_{b_i,b_j}^{A_{ij}} (1-p_{b_i,b_j})^{(1-A_{ij})}, \tag{2.11}$$

By changing the parametrization of the matrix $\boldsymbol{p}$, it is possible to generate networks having different kind of structures, such as assortative, core-periphery, bipartite, or even a combination of them (see Fig. 2.3).

The model of Eq. (2.11) has its origins in the social sciences [18, 102, 103] and appears under different names [104–109], being one of them *Bernoulli* SBM since it relies on Bernoulli

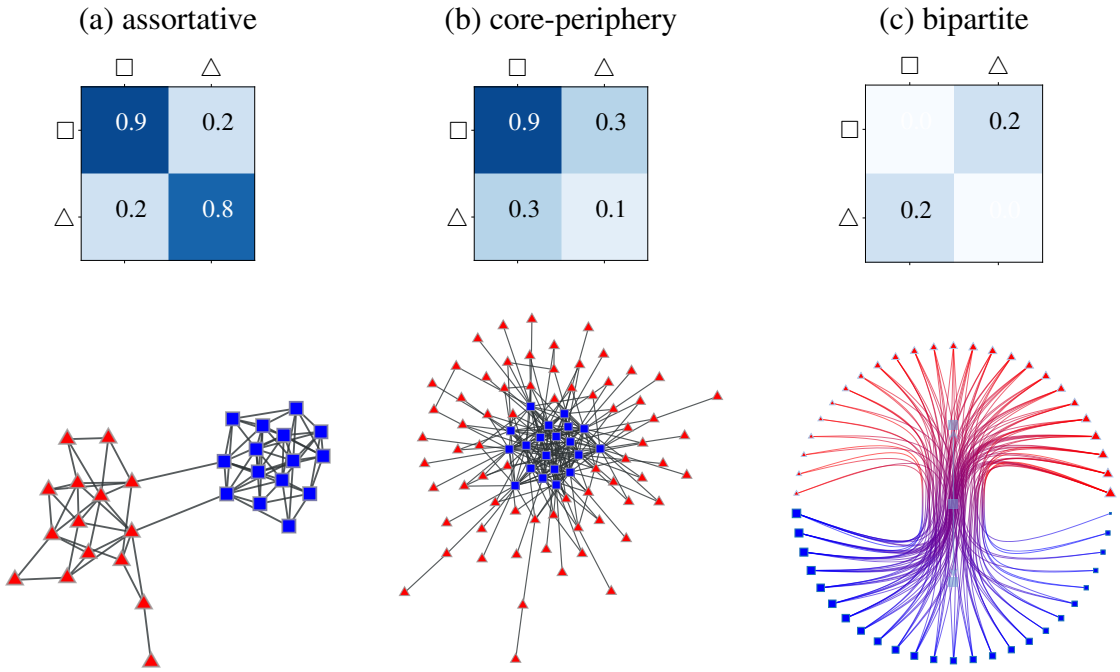


Figure 2.3: Examples of network structures that can be generated by the SBM. On top, we show the connectivity matrix between groups, and on the bottom, an instance of the model. (a) *assortative* structure, where there are more links inside groups than between groups. (b) *core-periphery* structure, where there is a dense group called *core*, and a sparse group that mostly connects to the core, called *periphery*. (c) *bipartite* structure, where nodes of one group only have connections with the other group. The labels of the connectivity matrices correspond to the node shapes in the generated graphs.

random variables to sample edges. It generates simple undirected networks, but it can be very easily modified to generate directed networks instead, by making $\boldsymbol{p}$ an asymmetric matrix, and adjusting the model likelihood accordingly.

This model generates networks where nodes belonging to the same group tend to have very similar degrees. Specifically, the Bernoulli SBM implicitly assumes that the expected degree of nodes within the same community is identical, with the expected degree of any node i approximately following a Poisson distribution if the communities are large. This is a major drawback of the model, as many empirical networks exhibit degree heterogeneity [55,84], often spanning several orders of magnitude. Consequently, when applying this version of the SBM to such networks, the model would tend to group nodes according to their degree, resulting in groups that would not exhibit heterogeneous degree distributions found in real-world networks.

An improved model which can generate networks with arbitrary mixing patterns and accommodate degree heterogeneity is the so called *degree-corrected* SBM (DC-SBM) [110]. This model uses a Poisson distribution to sample edges and introduces an extra parameter $\boldsymbol{\theta} = \{\theta_i\}$, one θ_i per node, which allows us to control the number of edges connecting to each node. This model generates multigraphs with probability

$$P(\boldsymbol{A}|\boldsymbol{\theta},\boldsymbol{\lambda},\boldsymbol{b}) = \prod_{i<j} \frac{e^{-\lambda_{b_i b_j}\theta_i\theta_j}(\lambda_{b_i b_j}\theta_i\theta_j)^{A_{ij}}}{A_{ij}!} \times \prod_i \frac{e^{-\lambda_{b_i b_i}\theta_i^2/2}(\lambda_{b_i b_i}\theta_i^2/2)^{A_{ij}/2}}{(A_{ij}/2)!}, \tag{2.12}$$

where λ_{rs} controls the expected number of edges between groups r and s, θ_i is the propensity of node i to receive edges, which is proportional to its expected degree. It can be assumed that the node propensities are normalized for each group,

$$\sum_i \theta_i \delta_{b_i,r} = 1,$$

such that the value λ_{rs} will correspond to the average number of edges between groups r and s (or twice that if $r = s$). The *non-degree-corrected* SBM, also known as *Poisson* SBM, is recovered from the model in Eq. (2.12) by setting $\theta_i = 1/n_{b_i}$, where n_{b_i} is the number of nodes in group r. For conciseness, we will describe other relevant variants of the SBM in Sec. 2.3.3. Now that we have defined how networks with prescribed modular structure are generated, we will consider the reverse procedure, i.e., how to infer the modular structure from data.

2.3.2 Nonparametric statistical inference

The inference task consists on determining which partition $\boldsymbol{b}$ generated an observed network $\boldsymbol{A}$, assuming the generative model is a variant of the SBM (e.g., see Fig 2.4). Specifically, we can express our uncertainty about the network partition $\boldsymbol{b}$, conditioned on the network data $\boldsymbol{A}$, according to the Bayesian posterior probability

$$P(\boldsymbol{b}|\boldsymbol{A}) = \frac{P(\boldsymbol{A}|\boldsymbol{b})P(\boldsymbol{b})}{P(\boldsymbol{A})}, \tag{2.13}$$

where

$$P(\boldsymbol{A}|\boldsymbol{b}) = \int_{\boldsymbol{\Theta}} P(\boldsymbol{A}|\boldsymbol{\Theta},\boldsymbol{b})P(\boldsymbol{\Theta}|\boldsymbol{b})d\boldsymbol{\Theta} \tag{2.14}$$

is the *marginal likelihood* integrated over the remaining model parameters $\boldsymbol{\Theta}$, $P(\boldsymbol{b})$ and $P(\boldsymbol{\Theta}|\boldsymbol{b})$ are the prior probabilities of the model parameters, which encode our prior beliefs about the model, and

$$P(\boldsymbol{A}) = \sum_{\boldsymbol{b}} P(\boldsymbol{A}|\boldsymbol{b})P(\boldsymbol{b}) \tag{2.15}$$

is a normalizing constant called the *model evidence*, and corresponds to the total probability of the data summed over all model parameters. The computation of $P(\boldsymbol{A})$ is intractable, but fortunately, the inference procedure only requires to evaluate $P(\boldsymbol{b}|\boldsymbol{A})$ up to a normalizing constant.

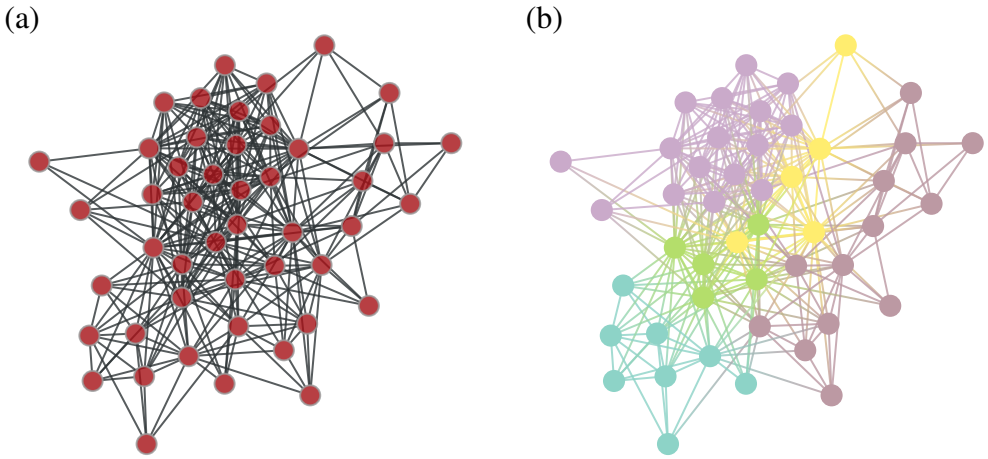


Figure 2.4: (a) A network of cortical regions in the Macaque cortex [112]. For visualization purposes, multiple edges are deleted. (b) An SBM fit to network data of (a). Node colors indicate the inferred groups.

Besides the marginal likelihood, the prior probabilities are important elements of the inference procedure, since they will affect the shape of the posterior distribution, and ultimately, our inference results. We postpone the choice of priors to Sec. 2.3.3, where we introduce a more convenient formulation of the SBM.

The inference procedure considered here will consist in either finding a network partition that maximizes Eq. (2.13), or sampling different partitions according to its posterior probability. In both cases, we rely on efficient Markov Chain Monte Carlo (MCMC) methods [53, 111], which are described in Sec. 2.3.6. Furthermore, this inference approach is non-parametric, to the extent that, the number of groups B will be an outcome of the inference procedure, rather than an input. Thus, the posterior of Eq. (2.13) will put low probabilities on partitions that are not backed by sufficient statistical evidence in the network structure, i.e., it will prevent *overfitting*. The reason why this approach prevents overfitting is based on a connection between Bayesian inference and information theory, which we refer to in Sec. 2.3.5.

To conclude this section we note that, when fitting an SBM to a network, we infer $P(\boldsymbol{b}|\boldsymbol{A})$, i.e., the posterior probability of the partition $\boldsymbol{b}$. Since SBMs are generative models, the estimated parameters can be used not only to simulate new predictions but also for model criticism and revision. We will turn our attention to this task in Chapter 3.

2.3.3 Microcanonical versions of the SBM

The models presented in Sec. 2.3.1 are canonical versions of the SBM. The term "canonical" comes from the field of statistical physics, and in such context, means that the model parameters correspond to "soft" constraints imposed on the ensemble of generated networks, i.e., constraints (e.g., the total number of edges) are only fulfilled on average. These models can be

reinterpreted and reformulated in a "microcanonical" way, i.e., model parameters (such as the total number of edges) correspond to "hard" constraints, so that they are fulfilled without any variation.

The microcanonical formulation of the degree-corrected SBM (DC-SBM) [57] combines arbitrary mixing patterns between groups together with arbitrary degree sequences. The parameters of this model are the partition of the nodes into B groups, $\boldsymbol{b} = \{b_i\}$, with $b_i \in [1,B]$ being the group membership of node i; the degree sequence $\boldsymbol{k} = \{k_i\}$, where k_i is the degree of node i; and the edge counts between groups $\boldsymbol{e} = \{e_{rs}\}$ (or twice that number for $r = s$), given by $e_{rs} = \sum_{ij} A_{ij}\delta_{b_i,r}\delta_{b_i,s}$. Given these constraints, the network is generated like in the configuration model [100, 101], with probability [57]

$$P(\boldsymbol{A}|\boldsymbol{k},\boldsymbol{e},\boldsymbol{b}) = \frac{\prod_{r<s} e_{rs}! \prod_r e_{rr}!! \prod_i k_i!}{\prod_{i<j} A_{ij}! \prod_i A_{ii}!! \prod_r e_r!}, \tag{2.16}$$

where $\boldsymbol{A} = \{A_{ij}\}$ is the adjacency matrix of an undirected multigraph with potential self-loops, and $e_r = \sum_s e_{rs}$. In this case, all the samples of the model have the same edge count matrix $\boldsymbol{e}$ and the same node degree sequence $\boldsymbol{k}$. This differs from the parameters $\boldsymbol{\lambda}$ and $\boldsymbol{\theta}$ in Eq. (2.12), which determine only the average number of edges between groups and the average node degrees. The actual values of these parameters fluctuate between samples.

One advantage of the microcanonical formulation over its canonical counterpart is that the former does not require any actual computation of the marginal likelihood.[5] In particular, the marginal likelihood of the microcanonical DC-SBM is given by

$$P(\boldsymbol{A}|\boldsymbol{b}) = \sum_{\boldsymbol{k},\boldsymbol{e}} P(\boldsymbol{A}|\boldsymbol{k},\boldsymbol{e},\boldsymbol{b}) P(\boldsymbol{k}|\boldsymbol{e},\boldsymbol{b}) P(\boldsymbol{e}|\boldsymbol{b}) \tag{2.17}$$

$$= P(\boldsymbol{A}|\boldsymbol{k}(\boldsymbol{A}),\boldsymbol{e}(\boldsymbol{A}),\boldsymbol{b}) P(\boldsymbol{k}(\boldsymbol{A})|\boldsymbol{e}(\boldsymbol{A}),\boldsymbol{b}) P(\boldsymbol{e}(\boldsymbol{A})|\boldsymbol{b}). \tag{2.18}$$

Notably, the summation over $\boldsymbol{k}$ and $\boldsymbol{e}$ of Eq. 2.17 reduces to a single term because only one term in the summation is compatible with the observed network. Given a network partition $\boldsymbol{b}$, there is only one pair of $(\boldsymbol{k},\boldsymbol{e})$ that matches the observed network data. All other parameter values are inconsistent and have zero probability.

Since $P(\boldsymbol{A}|\boldsymbol{k},\boldsymbol{e},\boldsymbol{b})$ was already defined in Eq. (2.16), we still need to chose the priors $P(\boldsymbol{k}|\boldsymbol{e},\boldsymbol{b})$ and $P(\boldsymbol{e}|\boldsymbol{b})$. One approach is to select these priors such that the microcanonical and canonical versions of the DC-SBM are equivalent, i.e., their marginal likelihoods are the same. As we will

[5] There are canonical formulations of the SBM for which the marginal likelihood can be computed exactly, e.g. see Ref. [25, 41, 113–115]. However, these models only include simple non-informative or conjugate priors, which potentially prevents the identification of all groups in large networks [116].

use the microcanonical formulation of the SBM in the following chapters, we omit the derivation of such equivalence and refer the reader to Ref. [22]. Instead, we focus on microcanonical priors, which, without requiring to compute the marginal likelihood, can be incorporated in Bayesian hierarchies of priors and hyperpriors. This approach leads to fewer assumptions on the data generating process and improves the quality of fit to data [57].

Priors of the SBM Parameters

Prior for the node partition

At the very first, we refer to the prior for the node partition. The simplest choice consists on being completely agnostic about the partitions, and choose among all of them with equal probability,

$$P(\boldsymbol{b}|B) = B^{-N}. \tag{2.19}$$

However, this uniform prior is not suitable for modeling real-world networks. The main reason is that most partitions into B groups have similar groups sizes N/B. Consequently, assuming a uniform prior becomes unrealistic and limits the potential of the inferential framework to achieve a better compression of the network data. A better prior relies on a parametric distribution, which is conditioned on the group sizes $\boldsymbol{n} = \{n_r\}$, where n_r is the number of nodes in group r,

$$P(\boldsymbol{b}|\boldsymbol{n}) = \frac{\prod_r n_r!}{N!}. \tag{2.20}$$

This is a maximum entropy distribution (all allowed configurations are equally likely), constrained on the fixed group sizes. In order to be agnostic about the size of communities, we can use a noninformative *hyperprior* on the node counts,

$$P(\boldsymbol{n}|B) = \left(\!\!\binom{B}{N}\!\!\right)^{-1}, \tag{2.21}$$

where $\left(\!\!\binom{n}{m}\!\!\right) = \binom{n+m-1}{m}$ counts the number of m-combinations from a set of size n, or equivalently, the number of possible histograms with n bins with counts that sum to m. It should be noted that, this prior also generates groups with size zero, which implies that we can also find partitions containing empty groups in the posterior distribution. This would force us to treat the number of groups as a free variable, since the nominal number of groups is not necessarily equal to the actual (nonempty) number of groups [115]. In order to avoid dealing

with such empty groups, we simply exclude them from our prior distribution, by using instead

$$P(\boldsymbol{n}|B) = \binom{N-1}{B-1}^{-1}, \tag{2.22}$$

which is a uniform distribution over all histograms with B nonempty bins and counts that sum to N. With this modification, the number of groups becomes a hard constraint as well, being always tied to the partition.

Lastly, the number of nonempty groups becomes a *hyperparameter*, for which, we can choose a uniform *hyperprior*, i.e., $P(B) = 1/N$, for $B \in [1, N]$. Therefore, the nonparametric prior for the node partition is given by

$$P(\boldsymbol{b}) = P(\boldsymbol{b}|\boldsymbol{n})P(\boldsymbol{n}|B)P(B) = \frac{\prod_r n_r!}{N!}\binom{N-1}{B-1}^{-1} N^{-1}. \tag{2.23}$$

Before specifying the priors for the remaining parameters, it should be noted that, we could have increased the depth of the Bayesian hierarchy by introducing a *hyperhyperprior* on other higher-order aspect of the group sizes $\boldsymbol{n}$. We do not proceed in that direction, and thus, remain with Eq. (2.23). As shown in Peixoto (2017) [22, 57], the reason is that we would gain at most a fairly marginal improvement proportional to $\ln N$ in the log-probability of the data generating process $\ln P(\boldsymbol{b})$. Consequently, for most cases, this would make little practical difference in the inference outcome.

Prior for the degrees

For the microcanonical degree-corrected SBM, the simplest choice we can make for the prior of degrees is to sample the degrees inside each group from a uniform distribution,

$$P(\boldsymbol{k}|\boldsymbol{e},\boldsymbol{b}) = \prod_r \left(\!\!\binom{n_r}{e_r}\!\!\right)^{-1}, \tag{2.24}$$

where $\left(\!\!\binom{n_r}{e_r}\!\!\right)$ counts the number of possible degree sequences on n_r nodes, constrained such that their total sum equals e_r. This uniform prior may not be suitable for modeling real-world networks, as sampling from it will result in degree sequences where most nodes have very similar degrees. Specifically, if the number of nodes is sufficiently large, this prior will lead to exponential degree distributions within each group [57]. These distributions have a much smaller variance than what is observed in empirical networks [117].

A better prior for $\boldsymbol{k}$ should be conditioned on an arbitrary degree distribution $\boldsymbol{\eta} = \{\eta_k^r\}$, with η_k^r being the number of nodes with degree k that belong to group r,

$$P(\boldsymbol{k}|\boldsymbol{e},\boldsymbol{b}) = P(\boldsymbol{k}|\boldsymbol{\eta})P(\boldsymbol{\eta}|\boldsymbol{e},\boldsymbol{b}). \tag{2.25}$$

The first term of the right hand side is a uniform distribution of degree sequences constrained by the overall degree counts, i.e.,

$$P(\boldsymbol{k}|\boldsymbol{\eta}) = \prod_r \frac{\prod_k \eta_k^r!}{n_r!}. \tag{2.26}$$

The second term is the distribution of the overall degree counts, i.e.,

$$P(\boldsymbol{\eta}|\boldsymbol{e},\boldsymbol{b}) = \prod_r q(e_r,n_r)^{-1}, \tag{2.27}$$

where $q(m,n)$ is the number of different degree counts, such that the sum of degrees is exactly m and the number of non-zero counts is at most n. This is also known as the number of *restricted partitions* of the integer m into at most n parts [118]. The function $q(m,n)$ can be computed recursively using the following expression:

$$q(m,n) = q(m,n-1) + q(m-n,n), \tag{2.28}$$

with boundary conditions $q(m,1) = 1$ for $m > 0$, and $q(m,n) = 0$ for $m \leq 0$ or $n \leq 0$.

Prior for the edge counts

As a starting point, we can assume again a uniform prior for the edge counts between groups, i.e.,

$$P(\boldsymbol{e}) = \left(\left(\!\!\binom{B}{2}\!\!\right)\right)^{-1}, \tag{2.29}$$

where $\left(\left(\!\!\binom{B}{2}\!\!\right)\right)$ counts the number of symmetric e_{rs} matrices with a constrained sum $\sum_{rs} e_{rs} = 2E$.

As before, the uniform prior is not a good choice either. In this case, this assumption introduces a "resolution limit", where the largest number of groups that can be inferred scales as $B_{\max} \sim \sqrt{N}$ [116], similarly to what is observed with the modularity maximization approach [16]. Roughly speaking, this means that smaller groups are typically merged together with neighboring blocks, so the methods cannot retrieve the true community structure. We

describe a solution to this limitation in the next section.

2.3.4 Nested DC-SBM

In practice, the "resolution limit" affects a community detection approach by limiting its ability to find small groups in very large networks. Although the Bayesian approach is robust against overfitting, it is still susceptible of *underfitting* due to the "resolution limit". Underfitting occurs when the model yields an overly simplistic partition compared with the actual pattern in the network data, i.e., when it mistakes statistically significant structure for randomness. Peixoto (2014) [25] proposed a solution to this problem by deepening the Bayesian hierarchy, i.e., the noninformative priors are replaced by a hierarchy of priors and hyperpriors. This new version of the model is called *nested* SBM or *hierarchical* SBM.

The main idea consists on viewing the matrix $\boldsymbol{e}$ as the adjacency matrix of a multigraph with B (meta)nodes and E edges. A reasonable assumption is that this multigraph is generated by another SBM, such that each group r belongs to one of another set of (meta)groups. The SBM at one level above the original model serves as a prior for the edge count matrix at the bottom level. This procedure can be repeated recursively L times until we end up with one node at the highest level, i.e.,

$$P(\{\boldsymbol{e}_l\}|\{\boldsymbol{b}_l\}) = \prod_{l=1}^{L} P(\boldsymbol{e}_l|\boldsymbol{e}_{l+1},\boldsymbol{b}_l), \tag{2.30}$$

where $\boldsymbol{b}_l$ is the partition of the groups in level l, $\boldsymbol{e}_l$ is the (weighted) adjacency matrix at level l, and we enforce always that $B_L = 1$.

The probability of sampling a multigraph from the microcanonical SBM at each level l is given by [119]

$$P(\boldsymbol{e}_l|\boldsymbol{e}_{l+1},\boldsymbol{b}_l) = \prod_{r<s}\left(\!\!\binom{n_r^l n_s^l}{e_{rs}^{l+1}}\!\!\right)^{-1} \prod_{r}\left(\!\!\binom{n_r^l(n_r^l+1)/2}{e_{rr}^{l+1}/2}\!\!\right)^{-1}. \tag{2.31}$$

In this regard, the uniform prior of Eq. (2.29) is a special case of Eq. (2.30) when $L = 1$. Since we have a partition per level, we also need to choose a prior for them, i.e.,

$$P(\{\boldsymbol{b}_l\}) = \prod_{l=1}^{L} P(\boldsymbol{b}_l), \tag{2.32}$$

where Eq. (2.23) is used accordingly at each level, replacing B for B_l, and N for B_{l-1}, with $B_0 = N$.

By putting together the model likelihood with all the priors, we obtain the joint distribution for

the hierarchical microcanonical DC-SBM,

$$\begin{aligned} P(\boldsymbol{A},\boldsymbol{k},\{\boldsymbol{e}_l\},\{\boldsymbol{b}_l\}) &= P(\boldsymbol{A}|\boldsymbol{k},\boldsymbol{e},\boldsymbol{b}_1)\times P(\boldsymbol{k}|\boldsymbol{e}_1,\boldsymbol{b}_1)\times P(\{\boldsymbol{e}_l\})\times P(\{\boldsymbol{b}_l\}) \\ &= \frac{\prod_i k_i!\prod_{r<s}e_{rs}!\prod_r e_{rr}!!}{\prod_r e_r!\prod_{i<j}A_{ij}!\prod_i A_{ii}!!}\times\prod_r\frac{\prod_k \eta_k^r!}{n_r!}q(e_r,n_r)^{-1}\times \\ &\quad \prod_{l=1}^{L}\prod_{r<s}\left(\!\!\binom{n_r^l n_s^l}{e_{rs}^{l+1}}\!\!\right)^{-1}\prod_r\left(\!\!\binom{n_r^l(n_r^l+1)/2}{e_{rr}^{l+1}/2}\!\!\right)^{-1}\times \\ &\quad \frac{\prod_r n_r^l!}{B_{l-1}!}\binom{B_{l-1}-1}{B_l-1}^{-1}\frac{1}{B_{l-1}}. \end{aligned} \tag{2.33}$$

Thus, the nested SBM accounts for a nested hierarchy of partitions, which besides improving the resolution limit, allows us to describe the data at multiple scales, having the possibility of uncovering potentially different mixing patterns at each level.

2.3.5 The Minimum Description Length Principle

We can interpret the Bayesian approach outlined above in an information-theoretic way, through the so-called *minimum description length* (MDL) principle [24]. This equivalence does not depend on a model variant, but holds in general. Nevertheless, it becomes clearer and more direct to appreciate through the microcanonical formulation.

We start by recalling that the inference procedure consists on finding the most likely partitions of the network supported by the data, either by sampling from or maximizing the posterior distribution of Eq. (2.13). The numerator of this equation can be rewritten as

$$P(\boldsymbol{A}|\boldsymbol{b})P(\boldsymbol{b}) = 2^{-\Sigma(\boldsymbol{A},\boldsymbol{b})}, \tag{2.34}$$

where $\Sigma(\boldsymbol{A},\boldsymbol{b})$ is called the *description length* of the network $\boldsymbol{A}$. [24, 120]. Considering the DC-SBM, whose marginal likelihood was given in Eq. (2.17), $\Sigma(\boldsymbol{A},\boldsymbol{b})$ is computed as

$$\Sigma(\boldsymbol{A},\boldsymbol{b}) = \underbrace{-\log_2 P(\boldsymbol{A}|\boldsymbol{k},\boldsymbol{e},\boldsymbol{b})}_{\mathcal{D}(\boldsymbol{A}|\boldsymbol{k},\boldsymbol{e},\boldsymbol{b})}\underbrace{-\log_2 P(\boldsymbol{k}|\boldsymbol{e},\boldsymbol{b}) - \log_2 P(\boldsymbol{e}|\boldsymbol{b}) - \log_2 P(\boldsymbol{b})}_{\mathcal{M}(\boldsymbol{k},\boldsymbol{e},\boldsymbol{b})}, \tag{2.35}$$

where the sum has been dropped since only one term is non-zero given a fixed network $\boldsymbol{A}$. In Eq. (2.35), the second set of terms $\mathcal{M}(\boldsymbol{k},\boldsymbol{e},\boldsymbol{b})$ quantifies the amount of information in bits necessary to encode the parameters of the model, while the first term $\mathcal{D}(\boldsymbol{A}|\boldsymbol{k},\boldsymbol{e},\boldsymbol{b})$ measures the number of bits needed to encode the network, once the model parameters are known. This connection between Bayesian inference and information theory allows us to draw an equivalence

between inferring a partition in a network and compressing it. Thus, inferring the partition that maximizes the posterior distribution of Eq. (2.13) is equivalent to finding the partition that most compresses the data, according to the description length of Eq. (2.35).

Importantly, the MDL approach to inference implements a principle of parsimony (or *Occam's razor*), as it penalizes overly complex models that are not supported by the data. For instance, consider two examples of the DC-SBM used to describe a network: a simpler one with few groups and a more complex one with many groups. For the complex example, the model term of the description length $\mathcal{M}(\boldsymbol{k},\boldsymbol{e},\boldsymbol{b})$ will be large because there are more parameters to encode compared to the simpler model. However, an increase in model complexity results in a reduction of the first term, $\mathcal{D}(\boldsymbol{A}|\boldsymbol{k},\boldsymbol{e},\boldsymbol{b})$, since fewer networks are compatible with the complex partition, and consequently, less information is needed to encode the network once the parameters are known. In this sense, the term $\mathcal{M}(\boldsymbol{k},\boldsymbol{e},\boldsymbol{b})$ functions as a *penalty* term. Thus, compressing a network involves finding a balance between model fit and model complexity, where simpler models are preferred unless increasing the model complexity is justified by a significant reduction in the total description length.

The reason why the compression approach, and consequently inference, prevents overfitting lies in a fundamental theorem from information theory, known as Shannon's source coding theorem [121]. This theorem establishes the statistical limits to possible data compression. Specifically, it states that the optimal compression of a sufficiently large sample of data $\boldsymbol{x}$, drawn from a probability distribution $P(x)$, can only be achieved using the code associated with the true probability distribution. In our framework, where data compression and inference are equivalent, Shannon's theorem implies that the best compression of an instance of the Erdős-Rényi model is achieved with an SBM that has only one group. Consequently, dividing the network in more groups does not yield additional compression or explanatory power. This result holds exactly when the size of the networks tend to infinite, while for finite-sized networks, the probability of better compression becomes vanishingly small. In Fig. 5.4, the reader will find the results of an experiment that are consistent with these statements.

Since the MDL approach prevents overfitting, the description length Σ can also be used as a criterion to select between models of different classes, i.e., having a different internal structure and set of parameters (such as the degree-corrected and non-degree corrected SBMs). The result of the model comparison should be the simplest model that is able to explain the data according to its statistical significance. As we will see in Chapter 5, besides the MDL principle, there are other approaches to do model selection.

2.3.6 Inference using Markov Chain Monte Carlo (MCMC)

Considering the Nested DC-SBM, the inference task consists on sampling from (or maximizing) the posterior distribution of the hierarchical partition,

$$P(\{\boldsymbol{b}_l\}|\boldsymbol{A}) = \frac{P(\boldsymbol{A},\{\boldsymbol{b}_l\})}{P(\boldsymbol{A})}. \tag{2.36}$$

Our approach is based on an efficient Markov chain Monte Carlo (MCMC) algorithm [111], which attempts to move the membership of nodes in different hierarchical levels at random, so that by accepting or rejecting such proposed moves after a sufficiently long time, the hierarchical partitions are sampled according to the posterior distribution of Eq. (2.36). The procedure is summarized in the following, and we start by noting that such posterior can be factorized as

$$\begin{aligned} P(\{\boldsymbol{b}_l\}|\boldsymbol{A}) &= \frac{\prod_l P(\boldsymbol{e}_{l-1},\boldsymbol{b}_l|\boldsymbol{e}_l)}{P(\boldsymbol{A})} \\ &= \prod_l P(\boldsymbol{b}_l|\boldsymbol{e}_{l-1},\boldsymbol{e}_l) \end{aligned} \tag{2.37}$$

with per-level posteriors

$$P(\boldsymbol{b}_l|\boldsymbol{e}_l,\boldsymbol{e}_{l+1}) = \frac{P(\boldsymbol{e}_l|\boldsymbol{e}_{l+1},\boldsymbol{b}_l)P(\boldsymbol{b}_l)}{P(\boldsymbol{e}_l|\boldsymbol{e}_{l+1})}, \tag{2.38}$$

where we assume $\boldsymbol{e}_0 = A$, and $P(\boldsymbol{e}_l|\boldsymbol{e}_{l+1})$ is a normalization constant. Thus, we can sample partitions at each level separately, according to its individual posterior conditioned on the remaining levels, which are fixed temporarily. This approach ensures *ergodicity*, i.e., every state is eventually visited. Furthermore, if the moves at each individual level are reversible, the overall distribution will correspond to the desired full posterior of Eq. (2.36). It is important to note that this procedure should only allow node membership moves at level l that do not invalidate the partition at level $l+1$.

At each individual level l, we select a node i and propose moving it from its current group r to a new group s. This move is made with probability $P(b_i^{(l)} = r \to s)$, whose definition will be provided shortly. Then, we compute the corresponding difference in the log-likelihood $\Delta \ln P_l$, and employ the Metropolis-Hastings criterion [122, 123], which states that we should accept the move with probability

$$a = \min\left\{1, e^{\Delta \ln P_l}\frac{P(b_i^{(l)} = s \to r)}{P(b_i^{(l)} = r \to s)}\right\}, \tag{2.39}$$

where $P(b_i^{(l)} = s \to r)$ is the probability of the reverse move being proposed. The log-likelihood difference at level l is computed as

$$\Delta \ln P_l = \ln \frac{P(b_i^{(l)} = s, \boldsymbol{b}_l \setminus b_i^{(l)} | \boldsymbol{e}_l, \boldsymbol{e}_{l+1})}{P(b_i^{(l)} = r, \boldsymbol{b}_l \setminus b_i^{(l)} | \boldsymbol{e}_l, \boldsymbol{e}_{l+1})}, \tag{2.40}$$

where $\boldsymbol{b}_l \setminus b_i^{(l)}$ means the partition of the remaining nodes excluding node i. Computing Eq. (2.40) does not require the normalization constant of Eq. (2.38); it only needs a subset of terms in the joint distribution in Eq. (2.33). The number of these terms is proportional to the degree k_i of node i. Furthermore, the number of groups in the bottom level ($l = 0$) is typically much larger than in the upper levels, showing an exponential decrease. Consequently, an entire "sweep" of the algorithm — attempting one move per node in the network — can be completed in $O(E)$ time, independent of the total number of groups, which makes the algorithm suitable for inference in very large networks.

Importantly, the use of the Metropolis-Hastings criterion enforces the property of *detailed balance*, which in combination with ergodicity, provides theoretical guarantees that the hierarchical partitions are eventually sampled from the correct posterior distribution $P(\{\boldsymbol{b}_l\}|\boldsymbol{A})$. However, in practice, the equilibration time might be prohibitively large for some choices of move proposals we make unless they are close to the actual posterior. According to Ref. [57], a move proposal that can significantly improve mixing times is given by Ref. [111]:

$$P(b_i^{(l)} = r \to s) = \sum_t P(t|i,l) \frac{e_{ts}^l + \varepsilon}{e_t^l + \varepsilon(B_l + 1)}, \tag{2.41}$$

where $P(t|i,l) = \sum_j A_{ij}^{(l)} \delta(b_j^{(l)}, t)/k_i^{(l)}$ is the fraction of neighbors of node i in level l that belong to group t, and $\varepsilon > 0$ is an arbitrary parameter that enforces ergodicity, without other significant impact in the algorithm, provided it is sufficiently small. Importantly, these move proposals do not affect the computation time, which remains $O(E)$, and they eliminate the dependency on the number of groups B.

Additionally, the efficiency of the algorithm might also be significantly impacted by the starting state. For instance, starting from a random partition can lead to metastable states, from which the chain takes a long time to escape. To reduce the tendency to get trapped in a metastable state and to have an initialization protocol that reduces the mixing time of the MCMC, Peixoto (2014) [111] proposed an agglomerative approach. This approach initially places each node in their own group and then progressively chooses between merging groups or making individual node moves. This procedure is iteratively run for each hierarchical level as described in Ref. [25].

The approach outlined above serves for sampling from the posterior distribution of Eq. (2.36).

However, if the goal of inference is maximizing the posterior, it can be easily adapted by introducing an "inverse temperature" parameter β in Eq. 2.39 and replacing $\Delta \ln P_l$ with $\beta \Delta \ln P_l$. By making $\beta \to \infty$ the algorithm becomes a greedy heuristic that provides a reliable estimate of the maximum, as long as the procedure is repeated many times. Alternatively, one can start with $\beta = 1$, and gradually increase β at each step until $\beta \to \infty$. This method is also known as *simulated annealing* [124].

Additionally, we note that the MCMC algorithm from Ref. [25], which we described, considers single node moves at each step. Peixoto (2020) [53] proposed a refined version of the algorithm which in addition to single node moves, implements merges and splits of groups. Since this version tends to produce faster mixing times, we consider it in this work. Specifically, we use the implementation of these methods available in the `graph-tool` library [125].

2.3.7 Model Realism

In the previous sections, we have highlighted the flexibility of the SBMs to model arbitrary mixing patterns and the robustness of the Bayesian inferential framework to prevent overfitting and underfitting when modelling empirical networks. We have seen that the degree-corrected SBM is able to accommodate arbitrary degree distributions, and we note that the models presented above can be easily extended to consider networks with directed edges. Despite these features, SBMs remain as approximations of real-world networks, so it is legitimate to ask to which extent their modelling assumptions are accurate representations of these networks. In this section we discuss some relevant cases where the previously described models might need extension or revision.[6]

Nodes might belong to more than one group

The versions that we have considered yield "hard" partitions of the set of nodes, meaning that each node can only belong to one group. Nevertheless, there are situations in which it is intuitively appealing to assume that nodes belong to more than one group simultaneously. For example, in social support networks, individuals share attributes such as kinship, religious beliefs, caste [130]. Similarly, proteins can belong to several protein complexes at the same time [131]. In these scenarios, the connection patterns of the nodes are assumed to be a mixture of the "pure" groups, leading to a richer type of model [132].

A relevant family of statistical models that relaxes the single-membership assumption and allows nodes to belong to multiple clusters is known as *mixed membership models*. These models have been applied in various domains for non relational data, including the analysis of survey

[6] Regarding extensions of the model for dynamical networks, we refer the reader to Ref. [126–128], and for multilayer networks to Ref. [127, 129].

data [133], text analysis [134], and image processing [135]. For a comprehensive review of the topic, we refer the reader to Ref. [136]. In the context of network data, the first approaches were mostly concerned with mixed membership models for single layer networks [132, 137–141], while more recent works have aimed to account for richer scenarios, such as multilayer networks [51, 142], noisy multiple reported network data [143], and hypergraphs [144].

In the context of microcanonical versions of SBMs, an overlapping variant was introduced in Ref. [145]. This model also incorporates the MDL principle in the inference process and allows for extensions to nested variants, as in the case of its non-overlapping counterparts. A detailed description of these extensions, along with a comparison of different SBM variants (both overlapping and non-overlapping) fitted to 42 empirical networks can be found in Ref. [145]. In this comparison, the non-overlapping degree-corrected SBM systematically yielded smaller description lengths and better fits than other model variants, implying that the overlapping models tend to overfit, especially in larger networks. In turn, this suggests that degree heterogeneity might be more pervasive than overlapping groups in real-world networks, at least within the current modelling framework.

Many real-world networks are simple graphs, not multigraphs

The DC-SBM generates multigraphs with potential self-loops according to Eq. (2.16). However, an important fraction of real-world networks are simple graphs, for which the above model can give only an approximation. Peixoto (2020) [146] demonstrated that the use of multigraph models based on the Poisson distribution (or equivalently, microcanonical models based on the pairing of half-edges, as above) cannot ascribe probabilities to simple edges (i.e. $A_{ij} = 1$) that are larger than $1/\mathrm{e} \approx 0.37$. This limits the applicability of such models on networks with heterogeneous density, either due to broad degree distributions or sufficiently dense communities, which are common properties of empirical networks.

To address this limitation, Peixoto (2020) [146] proposed a *Latent Poisson Multigraph* model, which assumes that an underlying unobserved multigraph $\boldsymbol{A}$ is in fact responsible for the observed simple graph $\boldsymbol{G}$ simply via the removal of the edge multiplicities and self-loops, i.e.

$$P(\boldsymbol{G}|\boldsymbol{A}) = \prod_{i<j} \left(1 - \delta_{A_{ij},0}\right)^{G_{ij}} \delta_{A_{ij},0}{}^{1-G_{ij}}. \tag{2.42}$$

Note that $P(\boldsymbol{G}|\boldsymbol{A})$ can only take a value of 0 or 1, depending on whether $\boldsymbol{G}$ and $\boldsymbol{A}$ are compatible. Via this mathematical construction, the final model

$$P(\boldsymbol{G}|\boldsymbol{k},\boldsymbol{e},\boldsymbol{b}) = \sum_{\boldsymbol{A}} P(\boldsymbol{G}|\boldsymbol{A}) P(\boldsymbol{A}|\boldsymbol{k},\boldsymbol{e},\boldsymbol{b}) \tag{2.43}$$

can express both arbitrary mixing patterns between groups as well as degree correction, without the limitations of the multigraph model for networks with large local densities [146]. The inference of this model is performed by sampling from the posterior distribution

$$P(\boldsymbol{A},\boldsymbol{k},\boldsymbol{e},\boldsymbol{b}|\boldsymbol{G}) = \frac{P(\boldsymbol{G}|\boldsymbol{A})P(\boldsymbol{A}|\boldsymbol{k},\boldsymbol{e},\boldsymbol{b})P(\boldsymbol{k},\boldsymbol{e},\boldsymbol{b})}{P(\boldsymbol{G})}, \tag{2.44}$$

which remains tractable. A suitable choice for $P(\boldsymbol{k},\boldsymbol{e},\boldsymbol{b})$ is the hierarchy of priors and hyperpriors described in Refs. [25, 57], which attempt to prevent underfitting, as mentioned previously. Furthermore, experiments on empirical networks from Ref. [146] suggest that the Latent Poisson Multigraph DC-SBM outperforms the Nested Poisson DC-SBM in the task of community detection. The former model reveals a larger number of groups due to its increased ability of identifying heterogeneous densities. Additionally, within a small corpus of empirical networks, Peixoto (2021) [52] showed that this model yields smaller description lengths than other SBM variants, suggesting it might be a better fit for these networks. In Chapter 3, we will extensively use this model to fit it on hundreds of empirical networks and evaluate its quality of fit.

The network formation process might not only depend on the groups

The basic versions of the SBM assume that the creation of edges depends solely on the group membership of the nodes. However, other mechanisms might be simultaneously responsible for the network formation. Neglecting this possibility and directly inferring a partition of the network, might result in spurious patterns and misleading conclusions due to the conflation of such mechanisms.

A relevant example is the conflation between homophily — the tendency of similar nodes, based on their attributes, to be connected — and transitivity — the tendency of two nodes to be connected if they share a common neighbor.[7] On one hand, homophily may induce the formation of triangles between similar nodes. On the other, triadic closure may induce locally dense regions. Both cases might result in a similar pattern, i.e., abundance of triangles, making it difficult to interpret which process is responsible for what we observe.

Some recent studies have integrated both concepts into their modeling framework and provided insights into the interplay between triangles and homophily [148–152]. However, they have not directly addressed the process of formation of triangles, the presence of large-scale homophily, or the specific contributions of each mechanism. A widely used approach in the social sciences to model the occurrence of triangles and homophily is the family of Exponential Random Graph Models (ERGMs) [89–91]. Although ERGMs are conceptually appealing because they rely on local edge-based structures that could explain the network formation, they suffer from *de-*

[7] For another interesting example involving the conflation of ranking and preference of connections (via degree imbalance), we refer the reader to Ref. [147].

generacy [89, 153–155]. This means that the large majority of the probability distribution is concentrated on either an empty or a full graph, which makes these models implausible for real-world networks.

In the context of Bayesian SBMs, Peixoto (2022) [49] proposed a modified version of the SBM that incorporates triadic closure (SBM/TC). This approach, combined with Bayesian inference can determine the most plausible mechanism responsible for the existence of every edge in the network, in addition to the underlying community structure.

The main assumption of the model is that the observed network $\boldsymbol{G}$ is composed by a seminal or substrate network $\boldsymbol{A}$, which is generated by the DC-SBM conditioned on a partition $\boldsymbol{b}$ of the nodes, and a set of ego graphs $\boldsymbol{g}$ potentially containing triadic closure edges, i.e., edges that connect two nodes if they share a common neighbor in the substrate network $\boldsymbol{A}$.

The inference procedure consists on sampling from the posterior distribution

$$P(\boldsymbol{g},\boldsymbol{A},\boldsymbol{b}|\boldsymbol{G}) = \frac{P(\boldsymbol{G},\boldsymbol{g},\boldsymbol{A}|\boldsymbol{b})}{P(\boldsymbol{G})}, \tag{2.45}$$

which encompasses all possible divisions into seminal and triadic closure edges, weighted according to their plausibility. For further details, we refer the reader to Ref. [49].

Experiments conducted in empirical networks show that, in most cases, the observed structure can be better explained by a non-trivial combination of underlying mixing patterns (community structure) with a tendency of forming triangles (triadic closure). The relevance of each component depends on the network and must be inspected individually. Additionally, the SBM/TC systematically performs better in reproducing the clustering coefficient of such networks compared with the DC-SBM. As before, this behavior cannot be solely explained by a single mechanism.

The model might be misspecified

Finally, we should note that inferential methods are not one-size-fits-all approaches, and there might be situations where they are unrealistic. If our model poorly represents relevant aspects of the true data-generating process, i.e., the model is *misspecified*, then our inferences might be inaccurate and our conclusions misleading. However, even though the model is misspecified, we may still want to use it in the hope that our inferences reveal some structure of the underlying generating process. Below, we illustrate these ideas with an example.

In Fig. 2.5(a), we show an urban street network along with the node partition obtained by fitting the Latent Poisson Multigraph Model from Ref. [146] to such network[8]. The inference procedure yielded groups that primarily correspond to contiguous spatial regions. This partition is

[8]We also considered the DC-SBM and hierarchical priors [25, 57] as part of the whole model.

interpretable and captures some patterns that might appear using a more suitable model (e.g., a latent space model [86]). This suggests that we can achieve some degree of data compression and predictive accuracy with an SBM, although they may not be optimal. However, the SBM may not be a suitable choice for describing this empirical network, as the model would assign it a very low probability. By inspecting a sampled network from the fitted model (see Fig. 2.5(b)) we note that edges are sampled in a way that violate spatial constraints, creating longer connections than those observed in the data, consequently distorting the distances between nodes and potentially other network properties. Models better suited to capture relevant properties of spatial networks may include geometric graphs [156, 157], spatial growth models [158–164], and optimal network models [159, 165, 166]. For a comprehensive review on spatial network models, see Ref. [67].

Importantly, the inferential framework described above provides tools for detecting signs of poor fit or model misspecification. One of them consists on checking how close summaries of the data are to those computed in networks sampled from the model. Large deviations might indicate poor fit (see Chapter 3). Another approach consists on inspecting the posterior distribution of network partitions [52] (see Chapter 5). If the posterior is too broad, i.e., there are many alternative hypotheses for the same data being equally plausible, then the model structure might be unable to capture the structure in the data.

(a) Empirical network

(b) Sampled network

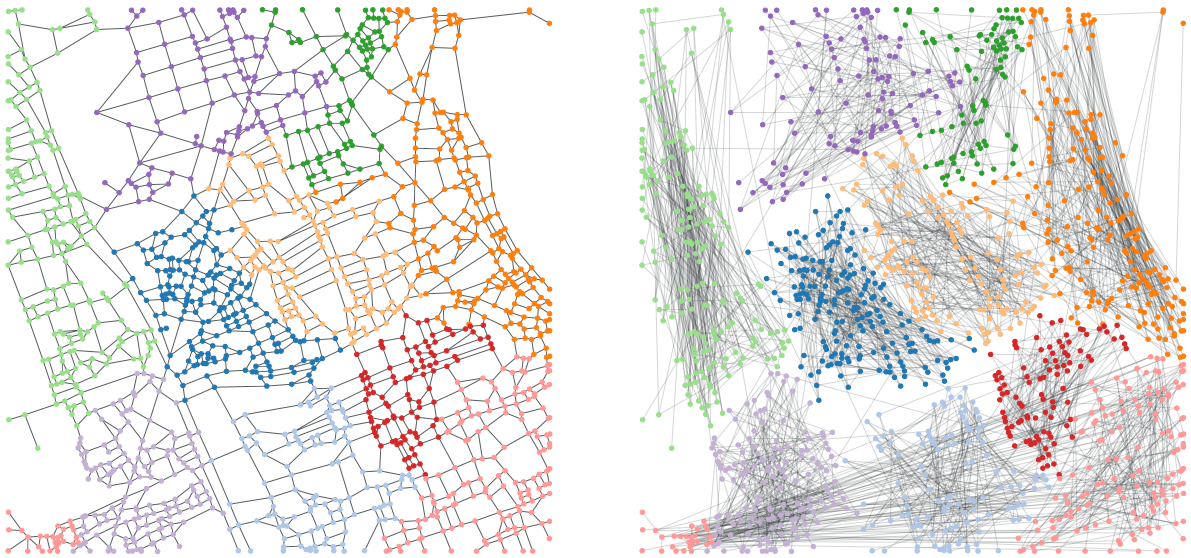


Figure 2.5: (a) Cairo street network [167, 168] where nodes represent street junctions and edges represent street segments. The node colors correspond to the groups inferred by the Latent Poisson Multigraph Model from Ref. [146], along with the DC-SBM and hierarchical priors [25, 57], by minimizing the description length. (b) A sampled network from the SBM fit described in (a).

Chapter 3

Quality of fit of the SBM for empirical networks

The SBM is a widely used model for community detection and link prediction tasks [22]. The parameters of this model, i.e., the partition of the nodes and the affinities between groups, are *latent* information that need to be obtained via inference algorithms. The family of SBMs also serve as generalizations of more fundamental random network models. For instance, the Erdős-Rényi model [97] can be recovered with the traditional non-degree corrected SBM with one group, while the configuration model [169] is analogously recovered with the degree-corrected version of the model. In fact, it has been shown that the SBM is able to approximate a broad class of generative models that are different from it [170]. In this regard, approximating the generative mechanism of a network with an SBM would be analogous to approximating the underlying probability distribution of some quantity. In the first case, we infer the partition of the observed network and affinities between groups, while in the second, we infer the bins of a histogram by fitting it to the data.

The level of complexity of the SBM can be controlled by the number of *latent* groups forming the partition. By increasing the number of groups, we can express increasingly elaborate types of network structures, being the rate of connections between nodes determined by their group membership. However, the expressiveness of the SBM is not absolute, especially when the networks are *sparse*, i.e. when their average degree is much smaller than the total number of nodes. In such a situation, there is no guarantee that the SBM is capable of arbitrarily approximating the true underlying model, regardless of how we infer it. By increasing the model complexity we move from a situation where we are *underfitting*, i.e. extracting patterns that do not sufficiently capture all the features of the true model, to a situation where we are *overfitting*, i.e. incorporating randomness into the model description. In both cases, we significantly deviate from the true model. When we find the most adequate inference that balances statistical evidence against model complexity to prevent overfitting, we might still be missing important

features of the true model, simply because it cannot be sufficiently well captured under the SBM parametrization.

In this work, we are not interested in evaluating the SBM as a plausible generative process of networks across all domains, since it does not represent an ultimately credible mechanism for any of them. Instead, our objective is to assess how capable it is of providing a general *effective* description of empirical networks, and in which aspects and to what extent (and not *whether*) it tends to be misspecified (see Sec. 2.3.7). Understanding the limits of the SBM representation in empirical settings is therefore a nuanced undertaking that is likely to be affected by a variety of possible sources of deviations (e.g., network size, structure, and domain). Since the SBM tends to yield very good comparative performance in link prediction tasks [42, 50], it is therefore known that it tends to outperform alternative models in capturing the structure of networks, but we still lack a more accurate assessment of its qualities and shortcomings in absolute terms.

In this chapter, we evaluate the quality of fit of the SBM in empirical contexts by performing *model checking* on Bayesian inferences. Based on a diverse collection of 275 networks spanning various domains and several orders of size magnitude, we compare the values of many network descriptors computed on the observed network with what would be typically obtained with networks sampled from the inferred SBM. In this way, any significant discrepancy can be interpreted as a form of "residual" that points to a shortcoming of the SBM in capturing that particular network property.

Overall we find that the SBM is capable of encapsulating the network structure to a significant degree for a large fraction of the networks studied, but falls short of completely exhausting the modelling requirements in many cases. We find that for networks with very large *diameter* or a very slow *mixing random walk* [171] the SBM tends to provide a poor description.[1] This includes, for example, many transportation networks — which are typically embedded in a low dimensional space — as well as some economic networks.[2] However, for other types of networks the quality of fit tends to be good overall.

In the remainder of this chapter, we describe in detail the model and inference procedure (Sec. 3.1), our criteria to evaluate the quality of fit (Sec. 3.2), the network corpus in which we assessed the model and the results of our analysis (Sec. 3.3). We finalize in Sec. 3.4 with a conclusion.

[1] These network properties are defined formally in App. A.2. For now, it is enough to consider the diameter of a network as the longest of all shortest paths in a network, and the mixing time of a random walk as the time it takes for a random walker to converge to a "stable" distribution over the nodes.

[2] See Ref. [55] for a qualitative overview of the different network classifications we consider.

3.1 Model and inference

For our analysis we will use the microcanonical degree-corrected SBM (DC-SBM) [57, 110], which combines arbitrary mixing patterns between groups together with arbitrary degree sequences. This was introduced in Eq. (2.16). Since all the networks we will be studying are undirected simple graphs, we will use the Latent Poisson Multigraph Model from Ref. [146], which was described in Sec. 2.3.7. For $P(\boldsymbol{k},\boldsymbol{e},\boldsymbol{b})$, we assume the nonparametric microcanonical hierarchical priors and hyperpriors described in Refs. [25,57]. For the inference procedure, we use the merge-split Markov chain Monte Carlo (MCMC) algorithm described in Ref. [53] to efficiently sample from the posterior distribution of Eq (2.44).

Note that for $P(\boldsymbol{k},\boldsymbol{e},\boldsymbol{b})$ we use the nonparametric microcanonical hierarchical priors and hyperpriors described in Refs. [25,57]. Importantly, this kind of approach determines the appropriate model complexity (via the number of groups) according to the statistical evidence available in the data. As has been shown in these previous works, this choice guarantees that only compressive inferences are made in a manner that prevents overfitting (finding a number of groups B that is too large), but also with a substantial protection against underfitting (finding a number that is too small), which tends to happen when noninformative priors are used instead.

In addition to the DC-SBM we will also use the configuration model as a comparison, obtained by reshuffling the edges of the input network while preserving its degree sequence (here we use the edge-switching MCMC algorithm [169]). We note that the configuration model is an approximate special case of the DC-SBM considered above when there is only a single group.[3] Therefore, whenever the Bayesian approach above identifies more than one group with a large probability, this automatically implies a selection of the DC-SBM in lieu of the configuration model. This happens for every network that we consider in this work, meaning that the DC-SBM is the favored model for all of them. Nevertheless, the configuration model serves as a good baseline to determine to what extent the quality of fit obtained with the DC-SBM can be ascribed to the degree sequence alone or to the group-based mixing patterns uncovered.

3.2 Assessing quality of fit

Posterior Predictive Distribution

The approach we use to assess the quality of fit of the DC-SBM consists on obtaining the *posterior predictive distribution* [172,173] of certain network descriptors and checking whether the model is able to capture such aspects of the network or not. More precisely, for a scalar

[3] This is only approximately true since the configuration model and the latent Poisson models are not identical, but sufficiently similar for the purposes of this work [146].

network descriptor $f(\boldsymbol{G})$, its posterior predictive distribution is given by

$$P(y|\boldsymbol{G}) = \sum_{\substack{\boldsymbol{G}',\boldsymbol{A}',\boldsymbol{A} \\ \boldsymbol{k},\boldsymbol{e},\boldsymbol{b}}} \delta(y - f(\boldsymbol{G}'))P(\boldsymbol{G}'|\boldsymbol{A}') \times P(\boldsymbol{A}'|\boldsymbol{k},\boldsymbol{e},\boldsymbol{b})P(\boldsymbol{A},\boldsymbol{k},\boldsymbol{e},\boldsymbol{b}|\boldsymbol{G}), \tag{3.1}$$

where $\delta(x)$ is the Dirac delta function. In other words, for each inferred parameter set $(\boldsymbol{k},\boldsymbol{e},\boldsymbol{b})$, weighted according to its posterior probability, we sample a new network $\boldsymbol{G}'$ from the model defined above (which can be done in time $O(E+N)$ where E and N are the total number of edges and nodes, respectively, as we show in Appendix A.1), and obtain the descriptor value $y = f(\boldsymbol{G}')$.[4]

Note that we could consider only one point estimate of the parameters (e.g., the one that maximizes the posterior distribution), sample networks, and compute descriptors to check the model. However, this not desirable, because we may loose all the information about uncertainty in the entire posterior distribution. This loss of information may lead us to overconfidence, to the extent that, the posterior would produce a narrower distribution than the posterior predictive distribution. Therefore, we may be tempted to believe that the model is more consistent with the data than it really is, if we find that our replicated data is similar to our observations. Instead, we would like to propagate the parameter uncertainty, i.e., carry it forward, as we evaluate the implied predictions. In this sense, the posterior predictive distribution of Eq. (3.1) can be seen as averaging across parameters and implied distributions of outcomes.

Measures of discrepancy

We know how to obtain predictive posterior distributions of network descriptors, but we still need to define how to quantify the magnitude of discrepancies between the simulations and the network data. We can say that a model captures well the value of a descriptor if its predictive posterior distribution ascribes high probability to values that are close to what was observed in the original network. We can obtain a compact summary of the level of agreement in two different ways. The first measures the statistical significance of the deviation, e.g. via the z-score [174]

$$z = \frac{f(\boldsymbol{G}) - \langle y \rangle}{\sigma_y}, \tag{3.2}$$

where $\langle y \rangle$ and σ_y are the mean and standard deviation of $P(y|\boldsymbol{G})$. The second criterion is the relative deviation, which here we compute in two different ways,

$$\Delta_1 = \frac{f(\boldsymbol{G}) - \langle y \rangle}{f(\boldsymbol{G})}, \quad \Delta_2 = \frac{f(\boldsymbol{G}) - \langle y \rangle}{f_{\max} - f_{\min}}, \tag{3.3}$$

[4]The posterior predictive distribution for the configuration model is analogous, i.e. $P(y|\boldsymbol{G}) = \sum_{G'} \delta(y - f(\boldsymbol{G}'))P(\boldsymbol{G}'|\boldsymbol{k})$, where $\boldsymbol{k}$ are the observed degrees, and $P(\boldsymbol{G}|\boldsymbol{k})$ is the likelihood of the configuration model.

Symbol	Descriptor	Range	Δ
r	Degree assortativity	$[-1,1]$	Δ_2
$\langle c \rangle$	Mean k-core value	$[0,\infty]$	Δ_1
C_l	Mean local clustering coefficient	$[0,1]$	Δ_2
C_g	Global clustering coefficient	$[0,1]$	Δ_2
λ_1^A	Leading eigenvalue of the adjacency matrix	$[0,\infty]$	Δ_1
λ_1^H	Leading eigenvalue of the Hashimoto matrix	$[0,\infty]$	Δ_1
τ	Characteristic time of a random walk	$[0,\infty]$	Δ_1
$\varnothing$	Pseudo-diameter	$[1,\infty]$	Δ_1
R_r	Node percolation profile (random removal)	$[0,1/2]$	Δ_2
R_t	Node percolation profile (degree-targeted removal)	$[0,1/2]$	Δ_2
S	Fraction of nodes in the largest component	$[0,1]$	Δ_2

Table 3.1: List of network descriptors used in this work, with their respective symbol, range of values, and how the relative deviation was computed. More details on how the descriptors are computed are given in Appendix A.2.

depending on whether the descriptor values are bounded in a well defined interval $[f_{\min}, f_{\max}]$ (Δ_2) or not (Δ_1).

The z-score and relative deviation measure complementary aspects of the agreement between data and model, and represent different criteria which should be used together. While a high value of the z-score can be used to reject the inferred model as a plausible explanation for the data, by itself it tells us nothing about how good an approximation it is. Conversely, the relative deviation tells us how well the descriptor is being reproduced by the model, but nothing about the statistical significance of the comparison.

In Fig. 3.1 we show examples that illustrate how the different criteria operate. In Fig. 3.1(a) and (b) we see examples that show good and bad agreements between model and data, respectively, according to both criteria simultaneously. In these cases, the conclusion is unambiguous: we either see no reason whatsoever to condemn the model, or we see a definitive reason to do so. However, in Fig. 3.1(c) and (d) we reach mixed conclusions. Fig. 3.1(c) the model typically yields different values than observed in the data, but it still ascribes a large probability to it. We cannot condemn the model as an implausible explanation for the data, but it is conceivable that the true generative model would be more concentrated on the observed value. Conversely, in Fig. 3.1 (d) we see a situation where the model ascribes close to zero probability to the actual descriptor value seen in the data, but, in absolute terms, the discrepancy is quite small. Although we find evidence to condemn the plausibility of the model, we could still claim that it is a good approximation.

Overall, since we know that a model like the DC-SBM cannot possibly correspond to the true

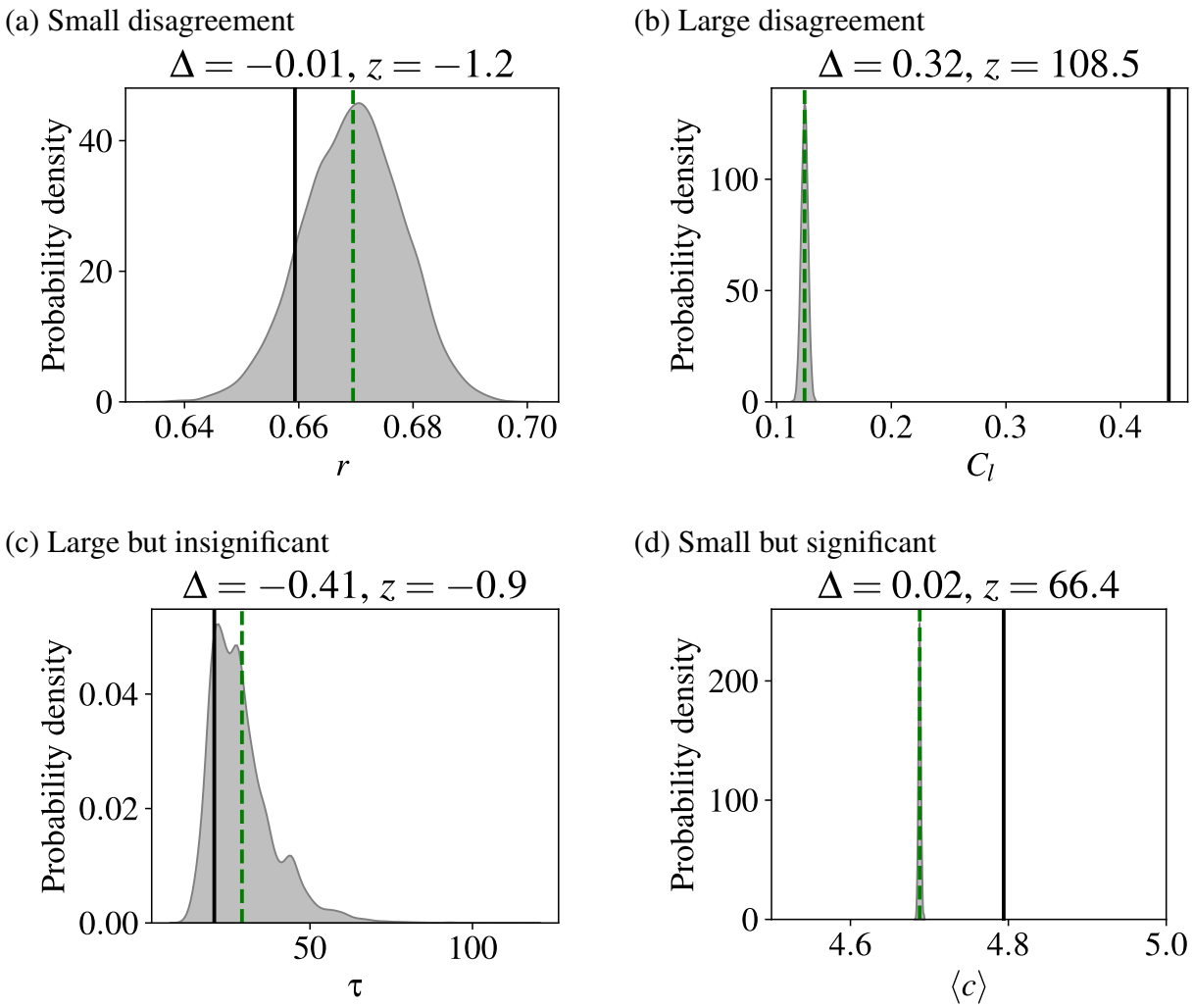


Figure 3.1: Examples of posterior predictive distributions for some descriptors (see Table 3.1 for definitions) using the DC-SBM, together with z-score and relative deviation. The solid black line shows the empirical value of the descriptor $f(\boldsymbol{G})$, and the dashed green line the mean of the predictive posterior distribution. In (a) and (b) we see examples where employing both criteria reveal unambiguously good and bad agreements, respectively, between data and model. However, in (c) we see a situation where despite a substantial disagreement with respect to the relative deviation, the z-score indicates that the model cannot be discarded as a plausible explanation for the data. In (d) we see a situation where the z-score points to decisive rejection of the model, but the small relative deviation allows us to accept it as an accurate approximation.

generative model of empirical networks, we should expect that in situations where the network is sufficiently large, and hence there is more abundant data, the values of the z-score will tend to be high. Here we argue that since the objective of a model like the DC-SBM is to obtain a good approximation of the underlying model, not an exact representation, the ultimate criterion is a combination of the two, where we may deem the model compatible with the data when *either* the z-score *or* the relative deviation has a sufficiently low magnitude. For the purpose of clarity and simplicity of our analysis, we will consider the thresholds $|z| = 3$ and $|\Delta| = 0.05$ as reasonable choices to deem the model compatible with data, although our results will not depend on these particular choices, and we will always report the full range of values.

Before continuing, some important considerations regarding model checking should be made. While an excellent model should fulfill both of the above criteria simultaneously, we need to observe that a model that maximally overfits, i.e. ascribes to the observed network a probability of one, and to any other a probability of zero, will achieve the best possible performance according to both relative deviation and statistical significance. This occurs because we are using the same data to perform both the model inference and evaluate its quality, which is an invalid approach for *model selection*. Therefore, it is important to recognize the crucial difference

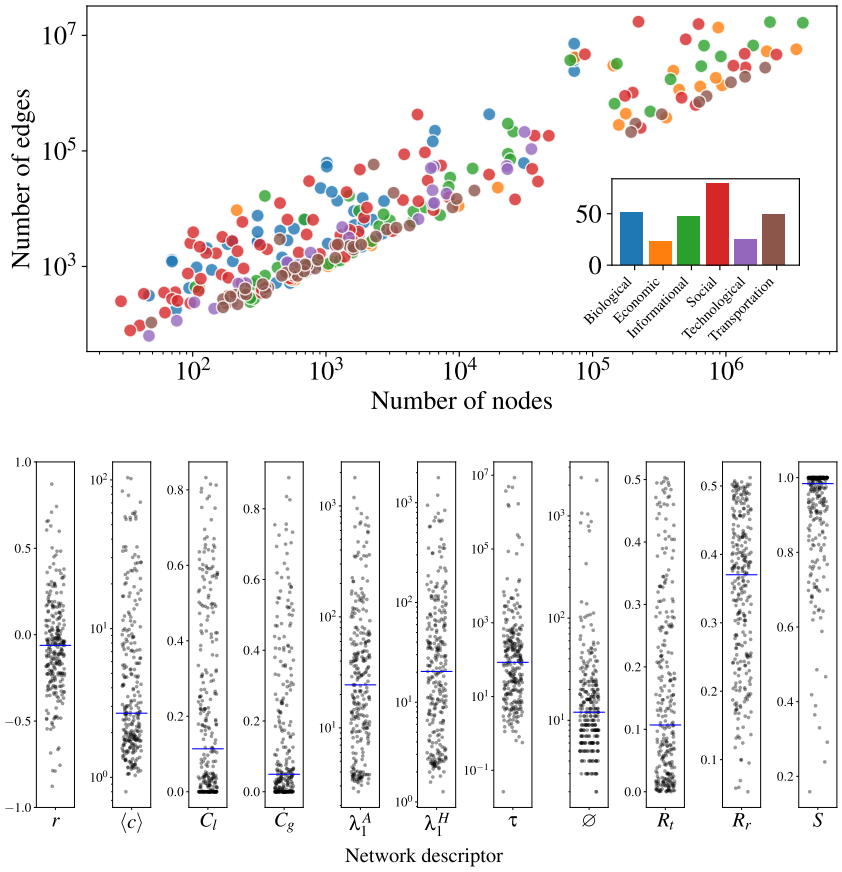


Figure 3.2: (Top) Number of nodes and edges for the networks in the corpus used in this work, and their domain composition (inset). (Bottom) Distribution of descriptor values for the networks in the corpus. The horizontal line marks the median values.

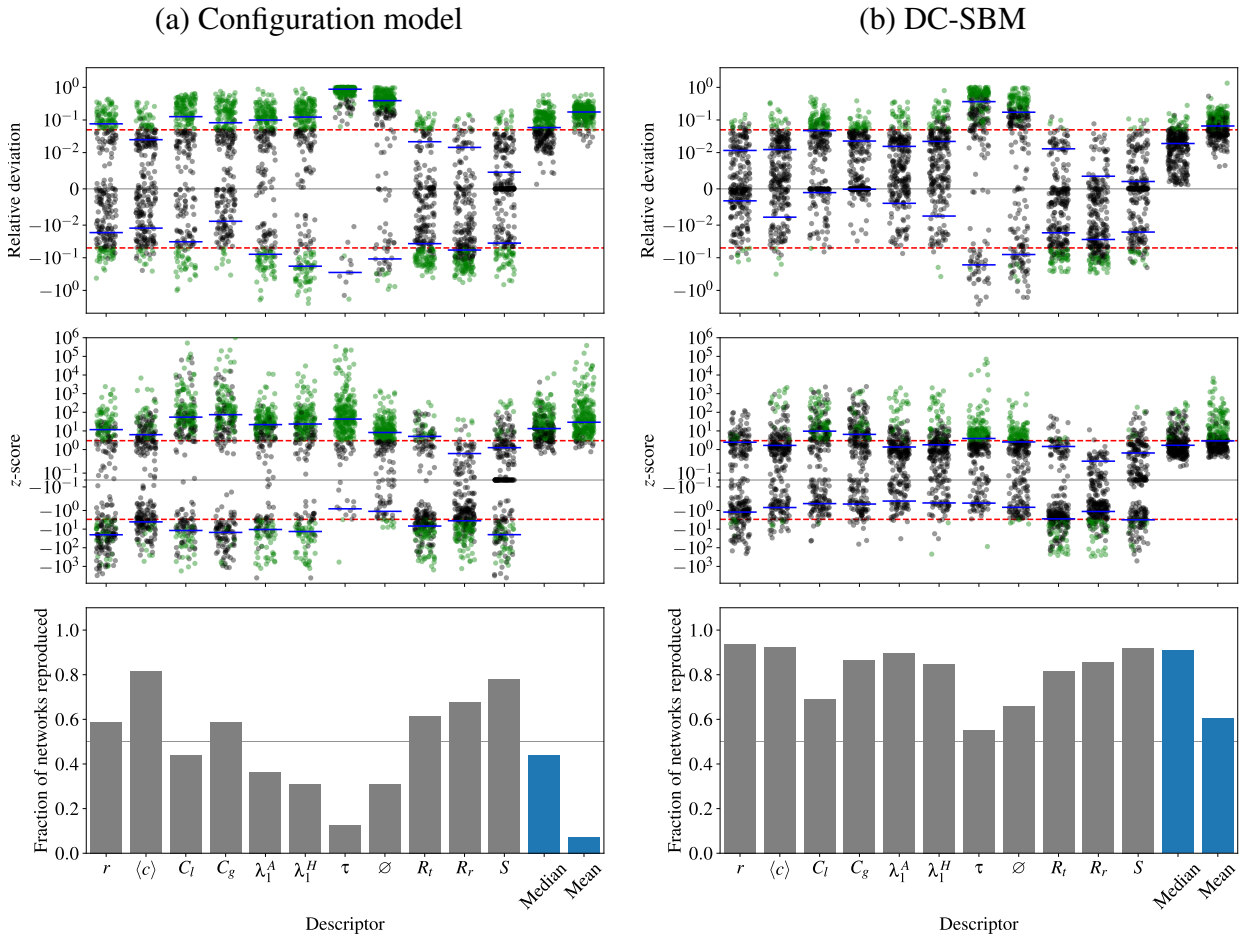


Figure 3.3: Distribution of relative deviation (top), z-score (middle), and fraction of networks reproduced (bottom) for (a) the configuration model and (b) the DC-SBM, according to their respective predictive posterior distributions for each descriptor. We also show the median and mean of the absolute values for all descriptors for each network. The solid blue lines mark the negative and positive median values, and the dashed red line marks the values of $|\Delta| = 0.05$ and $|z| = 3$. The fraction of networks reproduced correspond to those that have the absolute value of either Δ or z below these thresholds. The points in green color correspond to the networks that are not reproduced according to this combined criterion.

between model checking and model selection: the latter attempts to find the model alternative that is better justified according to statistical evidence, while the former simply finds systematic discrepancies between the inferred model and data. In our analysis, protection against overfitting is obtained via Bayesian inference, and we use model checking only to evaluate the discrepancies (indeed, the fact we find discrepancies to begin with shows that we cannot be massively overfitting). Another observation is that when performing multiple comparison over many networks and descriptors, some amount of "statistically significant" deviations are always expected, even if the models inferred correspond to the true ones, unless we incorporate the fact that we are doing multiple comparisons in our criterion of statistical significance, which would be the methodologically correct approach. We will not perform such a correction in our analysis, because we do not seek to demonstrate the absolute quality of DC-SBM as a

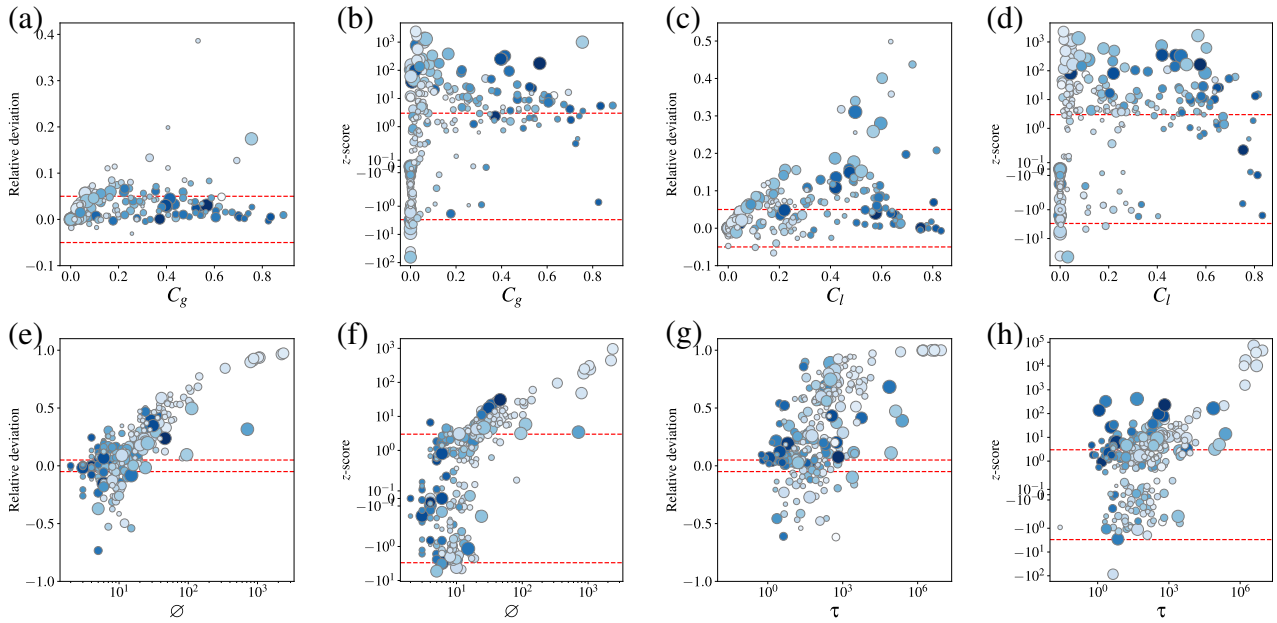


Figure 3.4: Relative deviation and z-score values for the global and mean local clustering coefficients, C_g and C_l, as well as diameter and characteristic time of a random walk, $\varnothing$ and τ, as a function of their empirical values, for every network in the corpus, when using the DC-SBM. The dashed red line marks the values of $|\Delta| = 0.05$ and $|z| = 3$. The size of the symbol corresponds to the logarithm of the number of edges in the network, and the darkness to the mean degree.

ultimately plausible hypothesis for network formation. As we will see from our results, such a correction would gain us very little.

Finally, in Table 3.1 we list the network descriptors that are used in this work. Our approach requires scalar values, so we constrained ourselves to this category, and furthermore we chose quantities that can be computed quickly, so that robust statistics from the predictive posterior distributions can be obtained. Given these restrictions, we then chose descriptors that measure different aspects of the network structure, both at a local and global levels. Further details on the network descriptors are given in Appendix A.2.

3.3 Quality of fit of the SBM in empirical networks

We carry out our analysis on a corpus containing 275 networks spanning various domains and several orders of size magnitude, as shown in Fig. 3.2. We have not collected every network at our disposal, but instead chosen networks that are as diverse as possible, both in size and domain, and avoided many networks that are closely related by belonging to the same subset. The networks in our corpus can be downloaded from the Netzschleuder repository [54]. In this work, every network is a simple graph, i.e., we considered symmetrized versions of directed networks removing parallel edges and self-loops.

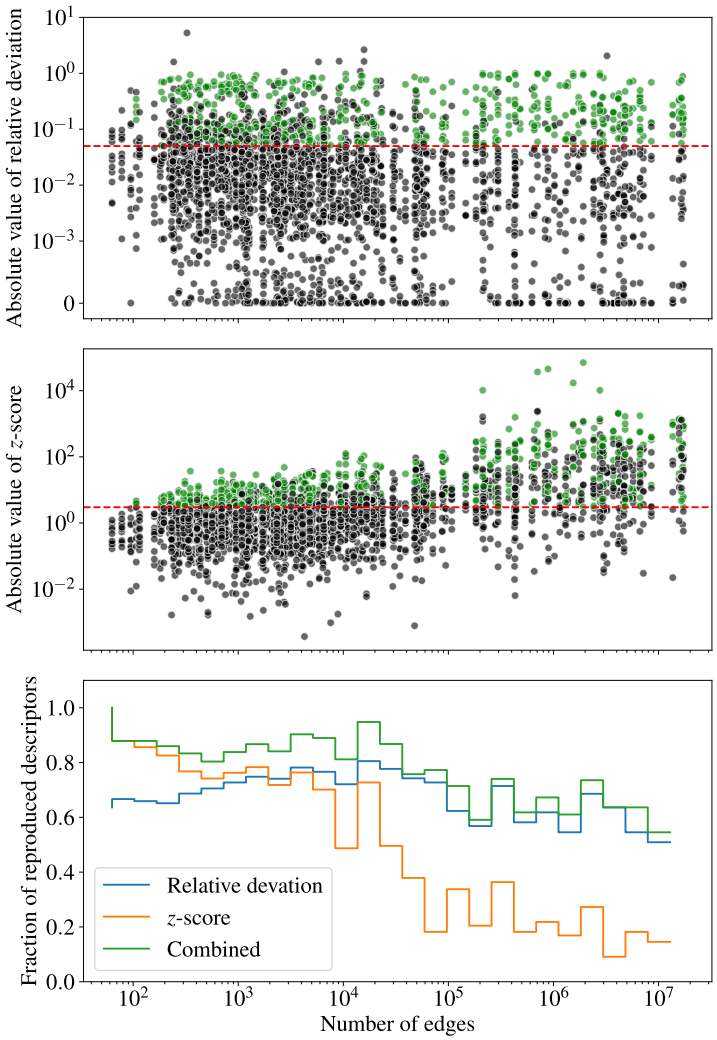


Figure 3.5: Absolute value of the relative deviation (top), z-score (middle) and fraction of reproduced descriptors (bottom), as a function of the number of edges, for every network in the corpus. The dashed red line marks the values of $|\Delta| = 0.05$ and $|z| = 3$. The fraction of descriptors reproduced correspond to those that have the value of either Δ or z below these thresholds. The points in green color correspond to the descriptors that are not reproduced according to this combined criterion.

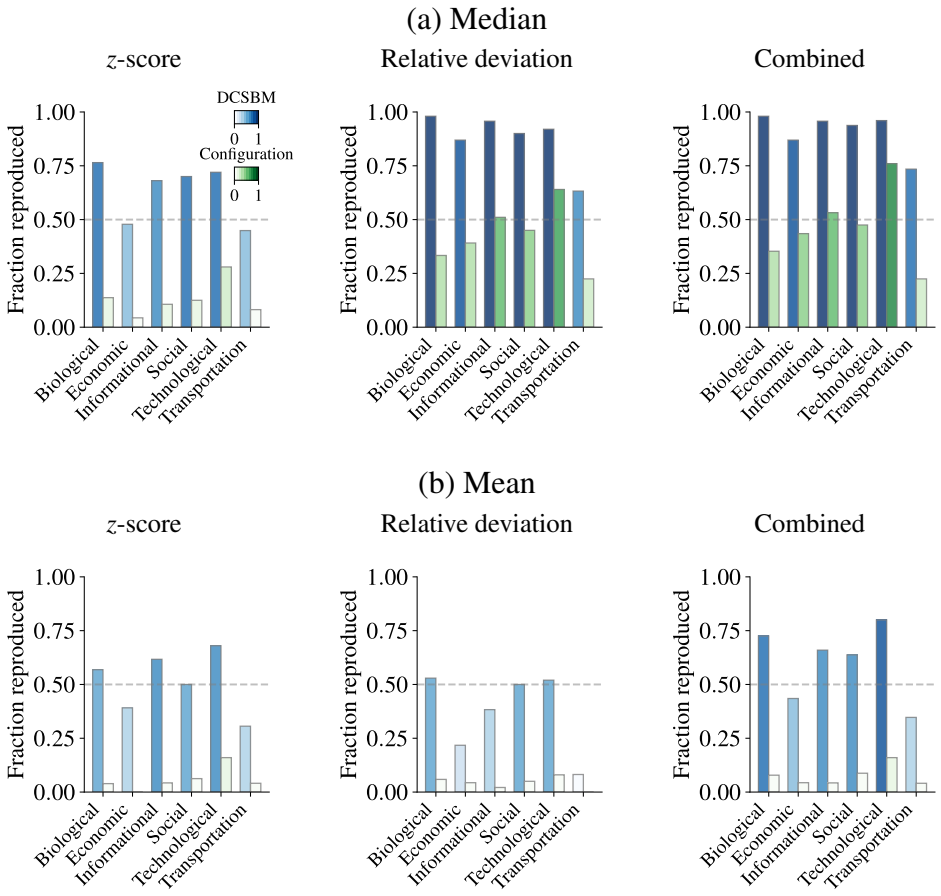


Figure 3.6: Fraction of reproduced networks according to their domain, considering the (a) median and (b) mean values of either the z-score, the relative deviations, or their combined values, for both models (as shown in the legend). When the combined values are used, this means that a model is deemed compatible with a network when we obtain either $|\Delta| < 0.05$ or $|z| < 3$.

In Fig. 3.3, we show the summaries of the posterior predictive checks for each descriptor, network, and model to be assessed. We observe a wide variety of deviation magnitudes, both for the same descriptors across networks, and across descriptors. As expected, the DC-SBM results show systematically better agreement with the data when compared with the configuration model. Overall, the descriptors that show the worst agreement are the characteristic time of a random walk (τ) and the diameter ($\varnothing$), both of which are particularly high for networks that are embedded in two dimensions, and for which the DC-SBM is an inaccurate approximation. Nevertheless, there is no single descriptor that the DC-SBM does not capture for fewer than 50% of the networks. For descriptors like S, R_r, R_t and $\langle c \rangle$, the difference between the DC-SBM and the configuration model are relatively minor. This indicates that these descriptors can be captured to a substantial extent by the degree sequence alone.

When considering all descriptors simultaneously for each network, either by the median or mean of the absolute values of the z-score and relative deviation, we observe a substantial majority of networks showing a good agreement with the DC-SBM. On the contrary, a small minority of networks agree with the configuration model. The difference between the median and the mean indicates that there is a sizeable fraction of the networks where the agreement is spoiled by a few outlier descriptors — typically τ and $\varnothing$.

It is particularly interesting to see that in most cases the clustering coefficients are well reproduced by the DC-SBM, while it is commonly assumed that such model should not be able to capture the abundance of triangles often seen in empirical networks. The argument is that the DC-SBM becomes locally tree-like [175], with a vanishing probability of forming triangles, in the limit where the number of groups is much smaller than the total number of nodes. Therefore, one may imagine that the situations where there is an agreement with the DC-SBM are those where the clustering values are low. However, Fig. 3.4(a) to (d) suggest that this is not quite true, i.e., we observe good agreements even when the clustering values are high. This illustrates a point made in Ref. [49], that it is possible to obtain an abundance of triangles with the SBM simply by increasing the number of groups, in which case it can be explained as a byproduct of homophily. Indeed this is a situation we see in Fig. 3.4(a) to (d), where both the relative deviation and z-score values can be quite small even for extremal values of clustering. However, we do notice a substantial variability between agreements, and a fair amount of instances where the DC-SBM cannot capture the observed clustering values, even when they are moderate or even small. This seems to indicate that there are a variety of processes capable of resulting in high clustering values, with homophily being only one of them [49]. Overall, the mean local clustering values tend to be harder to reproduce than the global clustering values. In both cases, the z-scores are systematically high, indicating that the clustering values are in general a good criterion to reject the DC-SBM as a statistically plausible model, although the relative deviation values tend to be lower than what one would naively expect, meaning that the model can still serve as a reasonably accurate approximation for clustered networks in many

cases.[5]

On the contrary, we observe a different behavior for the diameter and characteristic time of a random walk, which are the least well reproduced descriptors, as shown Fig. 3.4(e) to (h). These descriptors are closely related, since a network with a large diameter will also tend to result in a slow mixing random walk. For both of them it is rare to find a network having very high empirical values, and simultaneously the DC-SBM being able to accurately describe it. Therefore it seems indeed that the DC-SBM offers an inadequate ansatz to describe the structure of these networks, even by optimally adjusting its complexity.

In Fig. 3.5, we show how the model assessment depends on the size of the network. As one could expect, the z-score values tend to increase for larger networks, as more evidence becomes available against the plausibility of the DC-SBM as the true generative model. Nevertheless, the values of the relative deviation do not change appreciably for larger networks, indicating that it remains a good approximation regardless of the size of the system.[6]

Furthermore, in Fig. 3.6 we show a summary of the fraction of all networks for which we obtain good agreement with either model, according to the network domains. Overall, we see that most domains show similar levels of agreements, except transportation and economic networks. Transportation networks are often embedded in two-dimensional spaces, resulting in large diameters and slow-mixing random walks. The economic networks considered also tend to show large values of these quantities, so the explanation for their discrepancy is the same.

Predicting quality of fit

Now we address the question of whether it is possible to predict the quality of fit of the DC-SBM and Configuration Model solely based on the empirical values of the networks descriptors. If we can isolate the descriptors which are most predictive, this would give us a general direction in which more accurate models could be constructed.

In order to evaluate such predictability, we frame it as a binary classification problem. For each network i, we ascribe a binary value $y_i = 0$ if we have simultaneously $|z_i| > 3$ and $|\Delta_i| > 0.05$, or otherwise $y_i = 1$ (the network is well-described by the model). The feature vector for each network is composed of the empirical values of the descriptors, $\boldsymbol{x}_i = (r, \langle c \rangle, C_l, C_g, \lambda_1^A, \lambda_1^H, \tau, \varnothing, R_r, R_t, S, E)$, with the addition of the number of edges E. For each network i, we train a random forest classifier on the entire corpus with that network removed, and evaluate the prediction score on the held-out network. We then repeat this procedure for

[5] See also Fig. A.3, where we show model deviations as a function of the number of nodes in the network. We also show that the inferred number of groups B generally increases, and for networks with high clustering, B is large rather than constant. Despite this growth, the scaling of B does not follow in a simple way.

[6] Sampling issues with MCMC could also contribute to the elevated z-scores for larger networks, as we discuss in Appendix A.1.

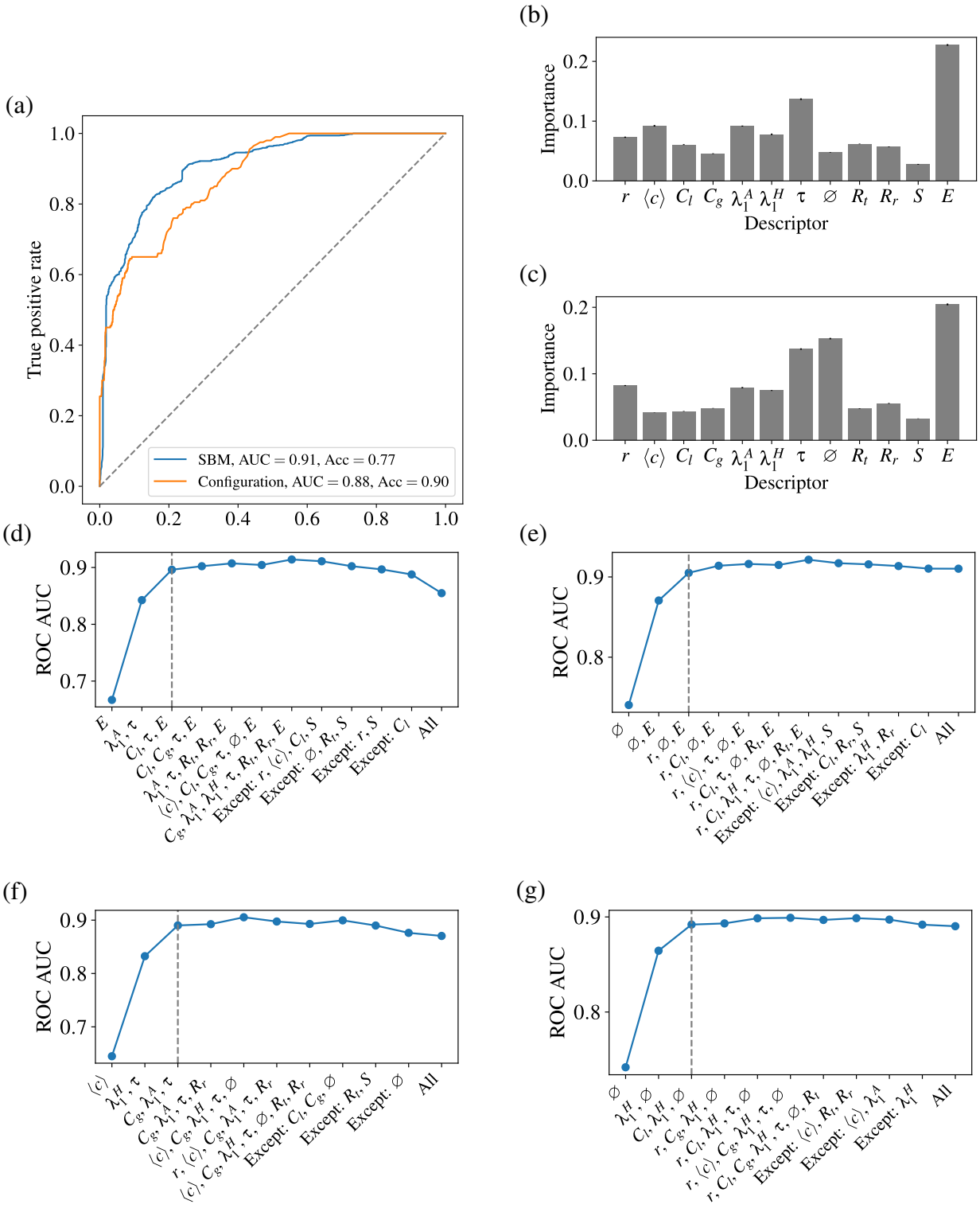


Figure 3.7: Predictiveness of the quality of fit of the generative models considered, according to the empirical descriptor values, framed as a binary classification problem, as described in the text. (a) ROC curve for a leave-one-out random-forest classifier, (b) Gini feature importance for the configuration model, (c) same as (b) but for the DC-SBM. Panels (d) and (e) show the best ROC AUC obtained for a set of descriptors of a given size, for the configuration model and DSCBM, respectively. Panels (f) and (g) show the same as (d) and (e), respectively, but with the number of edges excluded from the analysis.

all networks in the corpus, and evaluate how well the classifier is able to predict the binary label. We present the results of this experiment in Fig. 3.7 (top) which shows the receiver operating characteristic (ROC) curve, where the true positive rate and the false positive rate are plotted for all threshold values used to reach a classification. The area under the ROC curve (AUC), shown in the legend, can be equivalently interpreted as the probability that a randomly chosen true positive has a prediction score higher than a randomly chosen true negative. For the DC-SBM and configuration model, we obtain AUC values of 0.91 and 0.88, respectively. This indicates a fairly high predictability, from which we can conclude that it is indeed often possible to tell whether the models will provide a good or bad agreement, based only on the descriptor values.

Further insight can be obtained by inspecting the importance of each descriptor in the overall classification. We compute this via the so-called Gini importance [176], defined as the total decrease in node "impurity" (i.e. how often a node in decision tree contributes to a decision), weighted by the proportion of samples that reach that node, averaged over all trees in the classifier.[7] The results can be seen in Fig. 3.7 (b) and (c). In both cases, we note that the number of edges is the most predictive descriptor, which is compatible with what we had already seen in Fig. 3.5, namely that the larger the networks are, the easier it becomes to reject a model according to the z-score. Otherwise, as one would expect, the importance of the remaining descriptors is largely compatible with their reproducibility shown in Fig. 3.3, where the descriptors that agree the least with the inferred models tend to be the most useful at predicting quality of fit beforehand.

This analysis allows us to emphasize two points: the characteristic time of a random walk τ and the diameter $\varnothing$, both extremal quantities of the network structure that are closely related, are the most difficult descriptors to be captured by the DC-SBM. Therefore, an extension of the model that would cater for these properties would bring the most benefit across all networks. However, beyond these two descriptors, there is no substantial difference between the ones that remain, indicating that there is no obvious direction that would bring a systematic modelling improvement over all networks. On the other hand, as we show in Appendix A.2, the descriptor values and their predictive posterior deviations show nontrivial correlations, which means that if some of them are specifically targeted, it could potentially improve the quality of fit of other descriptors.

Finally, we would like to determine what is the minimal amount of information required to predict the suitability of both models, and in this way remove the redundancy provided by the different descriptors. In order to address this question, we computed the best ROC AUC obtained by a combination of descriptors of a given size, as shown in Fig. 3.7(d) and (e). In both

[7] We also computed different a measure, called permutation importance, which leads to very similar results (not shown).

cases, we see that the predictability is saturated by only few descriptors.[8] For the configuration model, most of the predictability is already achieved by a combination of (C_l, τ, E). For the DC-SBM we get instead $(r, \varnothing, E)$. If we remove the number of edges from the set of features (since it is not informative on the actual network structure), we obtain instead (C_g, λ_1^A, τ) and $(C_l, \lambda_1^H, \varnothing)$, for the configuration model and DC-SBM, respectively. It should be emphasized that if a descriptor does not appear in the minimal set, it does not mean that such descriptor is not predictive of the quality of fit, but only that it offers largely redundant information in that regard. Thus, for both models, replacing $\varnothing$ with τ or λ_1^H with λ_1^A, etc, yield similar results. This suggests that, besides spatial embeddedness (which influence $\varnothing$ and τ the most), the addition of explicit mechanisms for triangle formation (which affects $C_g, C_l, \lambda_1^H, \lambda_1^A$ directly) might improve the overall expressiveness of the DC-SBM — which in fact has been observed in a more limited dataset [49].

3.4 Concluding Remarks

We performed a systematic analysis of posterior predictive checks of the SBM on a diverse corpus of empirical networks, spanning a broad range of sizes and domains. Using a variety of network descriptors, we observed that the SBM is able to accurately capture the structure of the majority of networks in the corpus. The types of networks that show the worst agreement with DC-SBM tend to possess a large diameter and a slow mixing of random walks — features that are commonly associated with a low-dimensional spatial embedding, and a violation of the "small-world" property. For the other types of networks the agreement tends to be fairly good, even for many networks with an abundance of triangles. This contradicts what it is commonly assumed to be possible with this class of models.

We have also identified the minimal set of network descriptors capable of predicting the quality of fit of the SBM, which is composed of the network diameter and characteristic time of a random walk as the most important, followed by clustering as a secondary feature. This points to the most productive directions in which this class of models could be improved.

It is worth emphasizing that the consistency analysis that we have performed, which compares *a posteriori* the modelling assumptions with the actual properties seen in the data, is only possible if these assumptions are made explicitly via a generative model. Community detection methods that are only descriptive in nature (such as modularity maximization [177]) cannot be used for this purpose. Not only are these methods not guided by statistical evidence and prone to systematic overfitting, but they also provide no direct way to scrutinize the validity of their

[8] We optimized exhaustively for all descriptor combinations of a given size. Therefore, despite the leave-one-out cross-validation, care should be taken to avoid overfitting, because the optimization was performed on the same set of networks. Because of this, we always consider the smallest set of descriptors that reaches a ROC AUC close to the optimum, not the actual optimum which is likely to be overfitting.

implicit assumptions [11].

One of the limitations of our analysis is that it is conditioned on the set of descriptors used. Thus, shortcomings or successes of the model with respect to other properties not analyzed here are not uncovered. A natural extension of our work would be to consider an even broader set of descriptors that could reveal more relevant dimensions for the comparison. This kind of analysis is open ended, as there is no short supply of possible network descriptors. We hope our work will motivate further study in this direction, and with a larger variety of generative models within or beyond the SBM family.

Chapter 4

Reconstruction performance of the SBM in empirical networks

Empirical networks may contain errors or be incomplete. For instance, when measuring technological networks, one might encounter incomplete sampling and technical limitations [27–29]. Measurements of social networks might be affected by subjectivity, accuracy, and reliability of both participants and experimenters [30–32]. Natural variation in biological systems and inconsistent measurements in a lab might introduce large variability and discrepancies in the measurement of biological networks [33–35]. Thus, instead of directly analyzing the data and letting the error pollute the analysis pipeline, one should first infer — or reconstruct — the original network. In doing so, we may prevent misleading and erroneous conclusions [36–38].

Although network measurements are virtually always noisy, they are most often reported without any information on measurement uncertainty. In this situation, it becomes possible to reconstruct — or "denoise" — a network only if we possess suitable models for the measurement process together with generative models for the underlying network structure [43]. The stochastic block model (SBM) is a state-of-the-art approach for modelling network structure that has many useful characteristics for this purpose. Although it was initially motivated for community detection, and operates under the notion of preferences between groups of nodes, it can approximate a wide class of generative models when the number of such groups is suitably chosen [170]. In a recent comprehensive comparative analysis [50], the SBM was shown to consistently outperform alternative methods for link prediction. However, despite its high expressiveness, the SBM is not without limitations, as there might be underlying network patterns that the model cannot recover.

Most evaluations of network reconstruction or link-prediction methods are confined to relative comparisons between competing algorithms. In contrast, in our work we are interested in comparing the reconstruction performance of the SBM in absolute terms. We do so by not only computing its overall accuracy at recovering the missing edges and eliminating the spuri-

ous ones, but also in its ability to recover different kinds of descriptors that measure different aspects of the network structure. This kind of evaluation can determine the suitability of the reconstruction for different tasks, and provide different dimensions to judge its overall performance.

In this chapter, we evaluate the performance of the SBM in reconstructing network structure from a corpus of 248 networks spanning various domains and several orders of size magnitude. For this purpose, we assume each empirical network as error-free. Then we simulate a noisy measurement process that assumes the original network was measured once, and uniform error rates on both edges and non-edges. In this way, the noisy data would contain missing edges and spurious edges. Then we perform network reconstruction on the noisy network using the SBM. Finally, we assess the model by comparing the estimations of network descriptors with the descriptor values of the original network and those of the simulated noisy measurements. Large errors would mean that the SBM is incapable of reconstructing some aspect of the network. This outcome is complemented by comparing the error after reconstruction with the error before it. This comparison would tell us whether we gained something from the reconstruction procedure.

Our analysis reveals that the SBM yields small errors and more accurate estimates of the true network properties than those provided by the noisy observations. However, we also observe large reconstruction errors in networks having large diameter and slow-mixing random walks. These cases include many transportation networks — which are typically embedded in a low dimensional space — and some technological networks. Overall, our results show an encompassing delineation of the difficulty of the network reconstruction task and the suitability of the SBM for this purpose.

In the rest of this chapter, we describe the general framework of the network reconstruction procedure (Sec. 4.1.1), the setup of our evaluation (Sec. 4.1.2), our criteria to evaluate the accuracy of the reconstruction (Sec. 4.1.3), and the network corpus in which we assessed the model along with the results of our analysis (Sec. 4.2). Finally, we provide some concluding remarks in Sec. 4.3.

4.1 Network Reconstruction Framework

4.1.1 The goal of Network Reconstruction

Let $\boldsymbol{A}$ be a network and $\boldsymbol{D}$ the observed data, i.e., a noisy measurement that only contains indirect information about $\boldsymbol{A}$. The task of Bayesian network reconstruction consist on using statistical inference to obtain $\boldsymbol{A}$ from $\boldsymbol{D}$. In particular, the network can be reconstructed according to the posterior distribution [43]:

$$P(\boldsymbol{A},\boldsymbol{b}|\boldsymbol{D}) = \frac{P(\boldsymbol{D}|\boldsymbol{A})P(\boldsymbol{A},\boldsymbol{b})}{P(\boldsymbol{D})}, \tag{4.1}$$

where the likelihood $P(\boldsymbol{D}|\boldsymbol{A})$ models the measurement process, and the prior $P(\boldsymbol{A},\boldsymbol{b})$ is the SBM.

This means that when performing reconstruction, we sample both the community structure and the network itself from the posterior distribution. From it, we can obtain the marginal posterior probability of each edge (or non-edge), i.e.,

$$\pi_{ij} = \sum_{\boldsymbol{A},\boldsymbol{b}} A_{ij} P(\boldsymbol{A},\boldsymbol{b}|\boldsymbol{D}). \tag{4.2}$$

This allows us not only to infer the underlying network $\boldsymbol{A}$, but also to compute predictive indices (e.g., the area under the ROC curve) for model comparison purposes, or structural descriptors of the network to assess the model in absolute terms.[1]

4.1.2 Outline of the Analysis

The goal of this work is to test the reconstruction performance of the SBM in empirical networks using the framework described in section 4.1. In the following, we explain the pipeline of analysis for evaluating the SBM, and refer the reader to Fig. 4.1 for a schematic representation of the process.

We consider a simple undirected empirical network $\boldsymbol{A}$, with N nodes and E edges. We also take into account the model of the **measurement process** from Ref. [179], which assumes that the node pairs (i,j) were measured n_{ij} times, and an edge has been recorded x_{ij} times. A missing edge occurs with probability p, and a spurious edge occurs with probability q, uniformly for all node pairs, yielding a likelihood:

$$P(\boldsymbol{x}|\boldsymbol{n},\boldsymbol{A},p,q) = \prod_{i<j} \binom{n_{ij}}{x_{ij}} \left[(1-p)^{x_{ij}} p^{n_{ij}-x_{ij}}\right]^{A_{ij}} \times \left[q^{x_{ij}}(1-q)^{n_{ij}-x_{ij}}\right]^{1-A_{ij}}. \tag{4.3}$$

This equation can be rewritten in the following terms:

$$P(\boldsymbol{x}|\boldsymbol{n},\boldsymbol{A},p,q) = \left[\prod_{i<j} \binom{n_{ij}}{x_{ij}}\right] (1-p)^{\mathcal{T}} p^{\mathcal{E}-\mathcal{T}} q^{\mathcal{X}-\mathcal{T}} \times (1-q)^{\mathcal{M}-\mathcal{X}-\mathcal{E}+\mathcal{T}}, \tag{4.4}$$

where $\mathcal{M} = \sum_{i<j} n_{ij}$; $\mathcal{X} = \sum_{i<j} x_{ij}$; $\mathcal{E} = \sum_{i<j} n_{ij}A_{ij}$; $\mathcal{T} = \sum_{i<j} x_{ij}A_{ij}$.

[1] In this work, we made a correction of our estimates of marginal probabilities in a way that the sampled networks preserve the total number of edges of the original network. We provide further details in Appendix B.1.

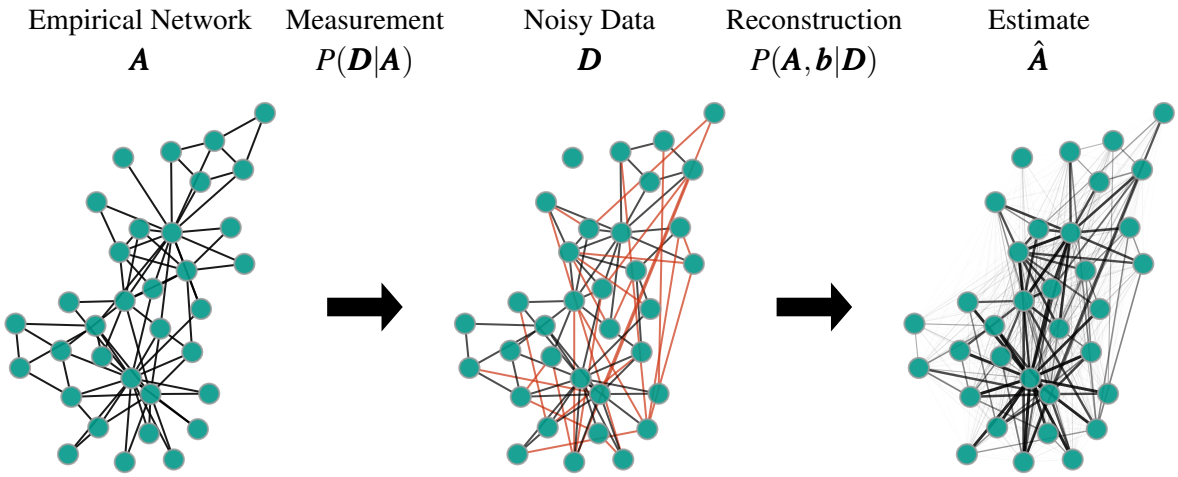


Figure 4.1: Schematic representation of our analysis on network reconstruction. We consider an empirical network $\boldsymbol{A}$ (here, the *karate club* [178]), add noise according to a measurement model, and try to reconstruct the original network from the noisy measurement $\boldsymbol{D}$, where some edges were deleted and spurious edges are colored in red. The result of the reconstruction procedure is a set of marginal probabilities of node pairs, which are represented as the thickness of edges in $\hat{\boldsymbol{A}}$.

We note that $\mathcal{M}$ is the total number of measurements (edge or non-edge), $\mathcal{X}$ is the total number of observed edges, and $\mathcal{T}$ is the total number of correctly observed edges. From these summary quantities, we can also identify the total number of false positives (spurious edges) as $\mathcal{X}-\mathcal{T}$ and the total number of false negatives (missing edges) as $\mathcal{E}-\mathcal{T}$.

In this work, we focus on the case of single measurements, i.e., $n_{ij}=1$ and $x_{ij}\in\{0,1\}$ for every pair (i,j), so that $\{x_{ij}\}$ corresponds to the reported adjacency matrix. Furthermore, we assume uniform error rates on edges and non-edges, setting both p and q to positive values, resulting in a mixture of edge- and non-edge denoising.[2] Here, $p\in\{0.1,0.3\}$, representing small and large values of noise level, respectively. The value of q is chosen so that the same number of affected edges and non-edges is on average the same, i.e., $q=pE/\left(\binom{N}{2}-E\right)$.

Given these positive error rates p and q, we simulate a measurement $\boldsymbol{x}$ (i.e., a noisy network) following Eq. (4.3). Specifically, we erase a fraction of edges (i,j) from $\boldsymbol{A}$ (according to p) and set $n_{ij}=0$, $x_{ij}=0$ for the affected entries, creating missing edges. Additionally, we add a fraction of non-edges (i,j) as spurious edges to $\boldsymbol{A}$ (according to q), and set $n_{ij}=1$, $x_{ij}=1$ for the corresponding entries. We note that the measured network has the same average density as the original network $\boldsymbol{A}$.

To conduct the reconstruction, we assume $P(\boldsymbol{A},\boldsymbol{b})$ to be the nested degree-corrected SBM (DC-

[2]Another possible scenario is *network completion*, where some edges or non-edges have not been observed, i.e., $n_{ij}=0$ and $x_{ij}=0$ for every pair (i,j). This is conceptually different to our setting since we assumed that measurements are performed, but they contain errors, while in network completion there are no measurements, and the error rates are zero. An evaluation of the SBM within this scenario is beyond the scope of this work. We refer the reader to Ref. [43] for further details.

SBM) [57, 110] and sample from the posterior of Eq. (4.1) using MCMC. During the MCMC sweeps, we collect the posterior probabilities of the affected entries π_{ij} and the group memberships of nodes.[3]

4.1.3 Assessing reconstruction performance

When assessing the reconstruction performance we are interested in understanding how close an inferred network $\hat{\boldsymbol{A}}$ is to the true network $\boldsymbol{A}$ underlying the data. There are several ways in which one can address this question. For instance, one could choose a measure of similarity or the area under the ROC curve (AUC) to assess the performance of the model in the reconstruction task.

In this work, we are interested in evaluating the model in its capacity to recover structural descriptors of the original network. This is done by getting estimates of such descriptors and quantifying the corresponding error of reconstruction, as described in the following.

Reconstruction error

Within the reconstruction framework of Eq. (4.1), we can compute estimates $\hat{y}$ of arbitrary scalar network properties $y(\boldsymbol{A})$ by averaging over the joint posterior $P(\boldsymbol{A},\boldsymbol{b}|\boldsymbol{D})$, i.e.,

$$\hat{y} = \sum_{\boldsymbol{A},\boldsymbol{b}} y(\boldsymbol{A}) P(\boldsymbol{A},\boldsymbol{b}|\boldsymbol{D}), \tag{4.5}$$

with uncertainties given by σ_y, such that

$$\sigma_y^2 = \sum_{\boldsymbol{A},\boldsymbol{b}} (y(\boldsymbol{A}) - \hat{y})^2 P(\boldsymbol{A},\boldsymbol{b}|\boldsymbol{D}). \tag{4.6}$$

Then we summarize the level of agreement via the relative error, which here we compute in two different ways,

$$\Delta_1 = \frac{y(\boldsymbol{A}^*) - \hat{y}}{y(\boldsymbol{A}^*)}, \quad \Delta_2 = \frac{y(\boldsymbol{A}^*) - \hat{y}}{y_{\max} - y_{\min}}, \tag{4.7}$$

depending on whether the descriptor values are bounded in a well defined interval $[y_{\min}, y_{\max}]$ (Δ_2) or not (Δ_1). In this work, we consider the same descriptors used in Chapter 3, which were listed in Table 3.1.

[3] The initial state of the MCMC algorithm is the best fit of the nested DC-SBM to the measurement $\boldsymbol{x}$. During the sampling, we monitor the equilibration of the description length and number of groups.

The procedure outlined above is repeated for several measurements, i.e., for a given empirical network $\boldsymbol{A}$, we consider k noisy measurements ($\boldsymbol{D_1}$, ..., $\boldsymbol{D_k}$), and perform network reconstruction from each D_i, $i = 1, ..., k$. Then we summarize the capacity of the SBM in reconstructing a network descriptor $y(\boldsymbol{A})$ as follows:

$$\bar{\Delta} = \frac{1}{k}\sum_i \Delta_i, \tag{4.8}$$

where Δ_i is the relative error (as in Eq. (4.7)) corresponding to the reconstruction from $\boldsymbol{D_i}$, the i-th noisy measurement. The uncertainty is given by $\sigma_{\bar{\Delta}} = \sigma_\Delta/\sqrt{k}$ with

$$\sigma_\Delta^2 = \frac{1}{k}\sum_i (\Delta_i - \bar{\Delta})^2. \tag{4.9}$$

We emphasize that in this work, we are interested in determining whether we can get accurate enough estimates using our framework, rather than showing that it is the best approach to network reconstruction when compared to other methods. Therefore, as a baseline, we use the value of the descriptor when no reconstruction is done along with its corresponding error. This would mimic a rather common practice in network data analysis, which is assuming the observed network is the true network. We present the results of these analyses in the next section.

4.2 Reconstruction performance of the SBM in empirical networks

We carry out our analysis in a corpus of 248 real-world networks spanning various domains and several orders of size magnitude (see Fig. 4.2). As before, when gathering the networks, we attempted to have networks in the corpus as structurally diverse as possible. Furthermore, every network in the corpus is a simple graph, i.e., we considered symmetrized versions of directed networks removing parallel edges and self-loops. These networks can be downloaded from the Netzschleuder repository [54].

For conciseness, in the remainder of this chapter, we focus on the results obtained with a level of noise $p = 0.1$, i.e., when on average 10% of edges have been removed. We include the results for a higher level of noise ($p = 0.3$) in Appendix B.2, where we observed qualitatively similar patterns. As expected, the performance of the algorithm deteriorates with a larger level of noise, resulting in larger errors and a lower percentage of reproduced networks. However, the improvement with respect to not doing any reconstruction at all becomes more noticeable.

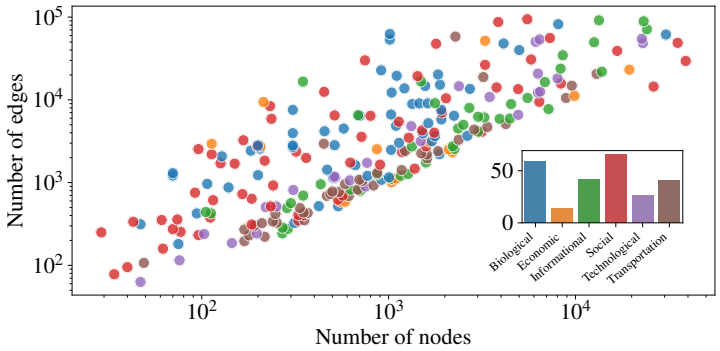


Figure 4.2: Number of nodes and edges of the networks in the corpus, and its domain composition.

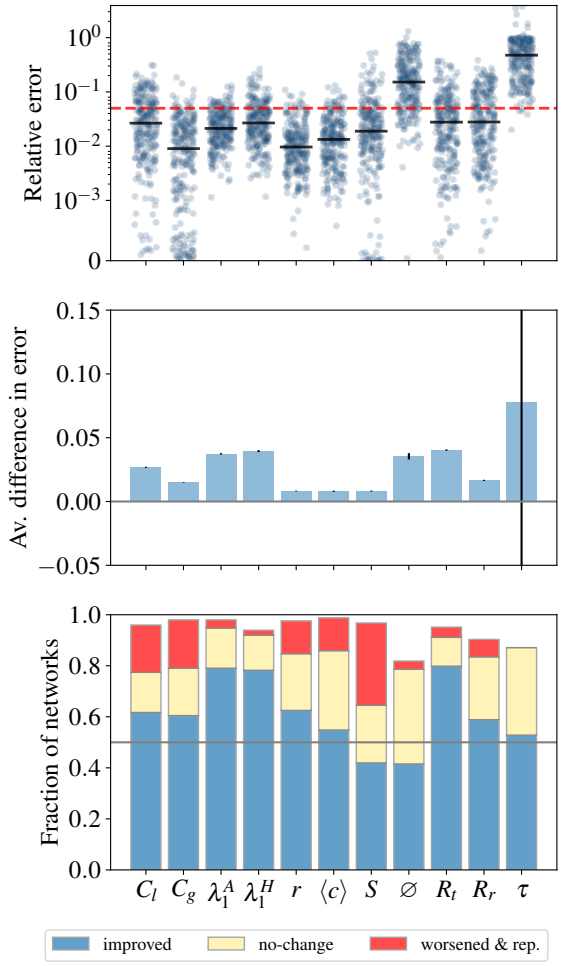


Figure 4.3: (Top) Distribution of relative error. The red dashed line corresponds to 0.05, and the black continuous lines to the medians. (Middle) Average difference between the error before reconstruction and after reconstruction. (Bottom) Percentage of networks whose error improved (i.e., the error after reconstruction is smaller than before reconstruction), did not change, or worsened but the descriptor was still reproduced (error is smaller than 0.05) for each descriptor. Noise level $p = 0.1$.

In Fig. 4.3, we show summaries of the reconstruction error for each descriptor and network. We observe a wide variety of error magnitudes, both for the same descriptors across networks, and across descriptors. Fig. 4.3 (top panel) shows that some descriptors are harder to estimate than others, i.e., the reconstruction approach has different accuracy. In particular, most errors are smaller than 0.05. However, the diameter ($\varnothing$) and the characteristic time of a random walk (τ) are the most challenging descriptors. We note that these descriptors were also the ones which showed the worst agreement in the previous chapter. We will return to this point later.

It should also be noted that we cannot assess the accuracy of reconstruction based only on the magnitude of the error. Instead, we should also take into account that some descriptors are more sensitive to noise than others, i.e., noise has different impact on different structural features. Thus the accuracy of estimations should be compared against a baseline. As mentioned, we consider the error in which we incur by not doing reconstruction and assuming the noisy network as the true one, as a baseline. As expected, on average, the SBM-reconstruction provides smaller errors compared with not reconstructing (see Fig. 4.3 (middle panel)). Once again, the exception is the characteristic time of a random walk (τ). Furthermore, we note that the best reproduced descriptors (e.g., r, $\langle c \rangle$, S), are not necessarily the same for which we obtained larger improvements in an absolute sense.

Overall, we would expect that our approach is accurate enough to the extent that, for every descriptor, it yields smaller errors, or at least not (drastically) worse than those of the baseline. In Fig. 4.3 (bottom panel), we observe that this is the case, since the algorithm is able to accurately estimate all descriptors for the majority of networks in the corpus.

In Chapter 3, we observed that the clustering coefficients were well reproduced by the DC-SBM, which contradicted the common assumption of the SBM being unable to capture abundance of triangles in empirical networks. The current scenario includes the presence of noise, so we expected it to be more challenging for the SBM to capture clustering coefficients. However, it is remarkable that in most cases the SBM-reconstruction approach not only is better than the baseline, but also yields relatively small errors.[4] In Fig. 4.4 (a) and (b), we observe that reconstruction errors are small when clustering is low. Moreover, when clustering is high, the reconstruction mostly yields better estimates than the baseline, and in some cases even small errors.

On the contrary, the performance of our approach decreases as the empirical values of the diameter ($\varnothing$) and the characteristic time of a random walk (τ) increase (see Fig. 4.4 (h) and (k)). This finding agrees with what we observed in the previous chapter, namely there exist large discrepancies between the DC-SBM and the network data when the empirical values of such descriptors are large. We expected that adding noise to the network make more challenging the task of getting accurate values of such descriptors, since noise removes information from the

[4] Peixoto (2018) [43] showed similar results for a network of political weblogs [3] and a network of flights among airports [180].

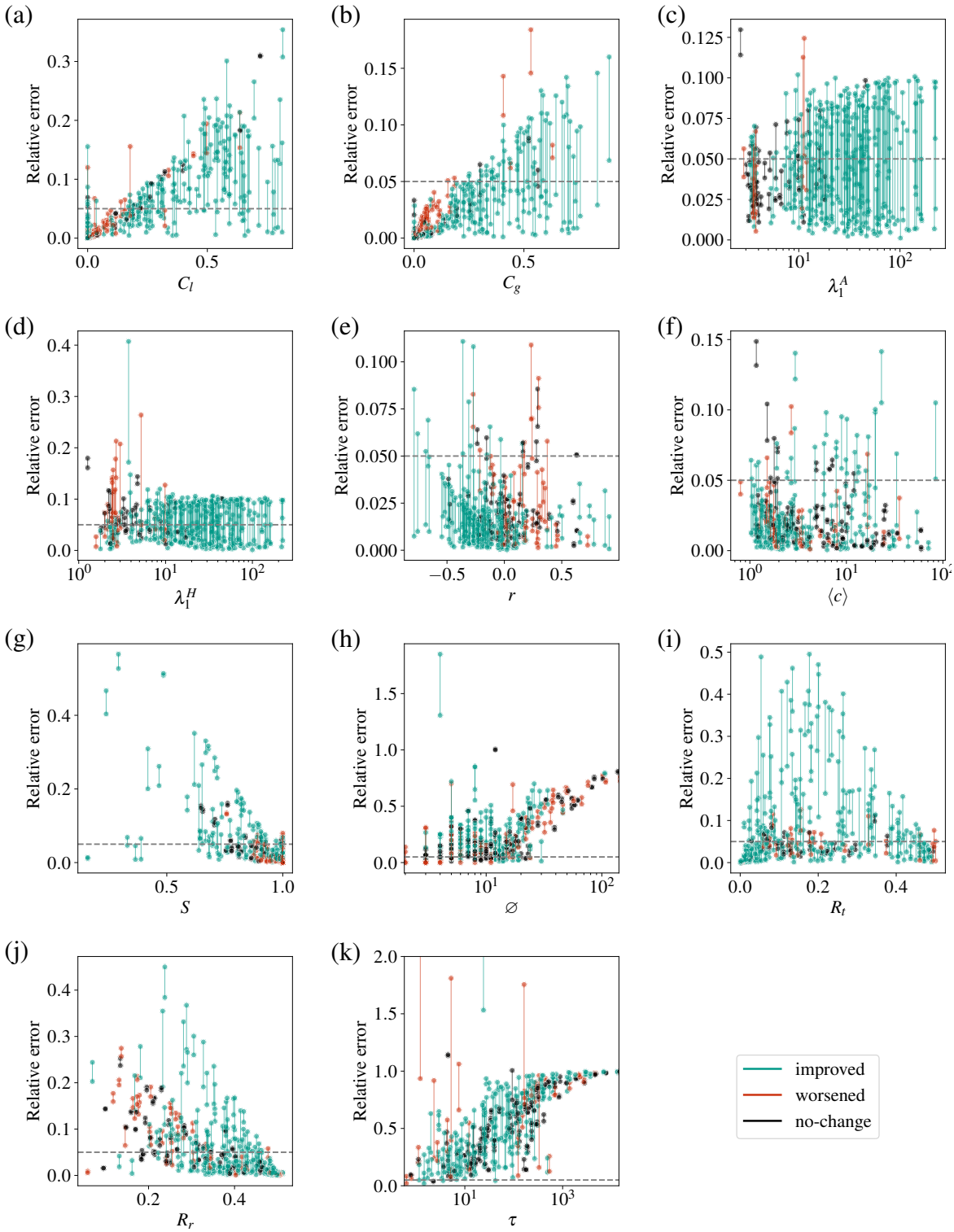


Figure 4.4: Relative error before and after reconstruction (joined by a line segment) as a function of the original value of the descriptor. The color indicates if the error after reconstruction is smaller than before doing it (i.e., there is improvement) or not. Dashed line at 0.05. Noise level $p = 0.1$.

data. Certainly, the reconstruction performance depends on how sensitive are such descriptors to noise. We observe that the diameter has high sensitivity, since a small amount of spurious

edges might significantly change the value of this descriptor. We illustrate this by an example of an urban street network (see Fig. 4.5). The reconstruction approach destroys spatial constraints, introducing links between distant nodes in the reconstructed networks, which would be unlikely in reality, and consequently yielding much smaller diameters than the original one. A similar explanation follows for τ, since it is closely related to $\varnothing$. This behaviour suggests that our assumptions may be incorrect for certain networks, particularly those with 2D embeddedness. Therefore, the SBM would not be a suitable prior for these networks when our focus is on $\varnothing$ and τ.

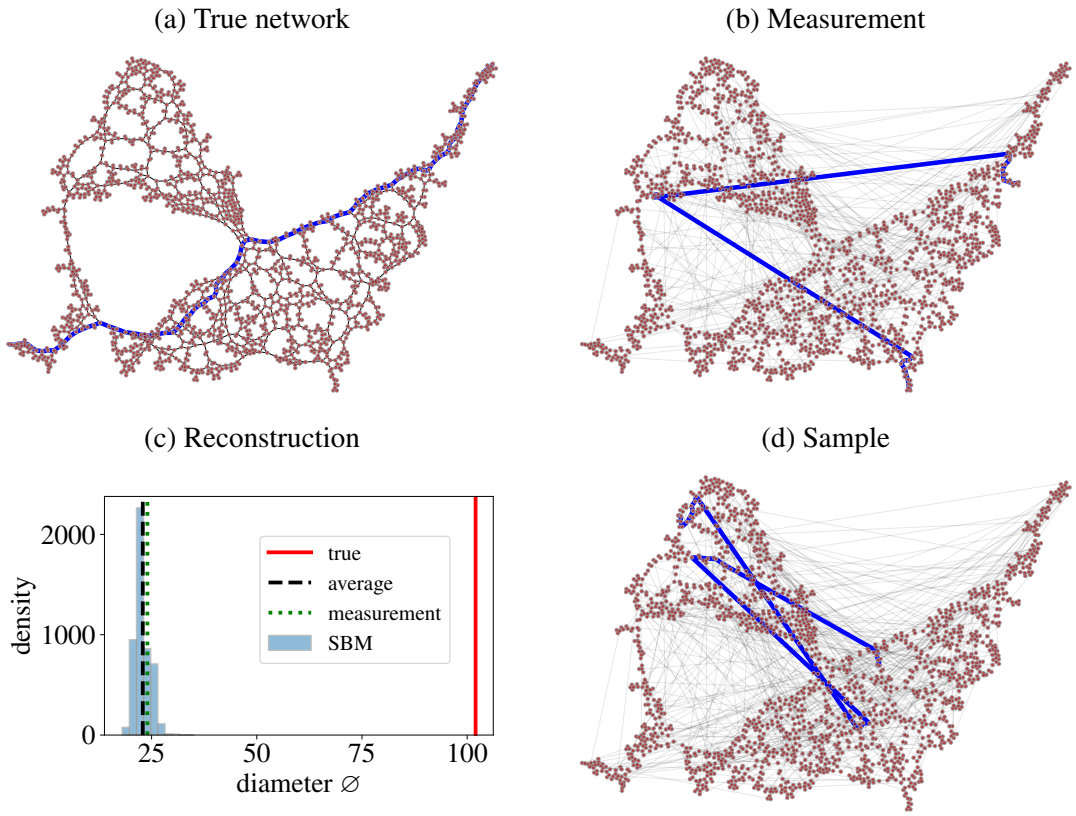


Figure 4.5: Estimation of diameter $\varnothing$ in Venice street network [167, 168] using SBM-reconstruction. (a) Original network, whose diameter is highlighted in blue. (b) Measured network obtained containing missing edges and spurious edges from (a). (c) Distribution of estimates of $\varnothing$. (d) One sample of the reconstructed networks with its diameter.

Furthermore, we summarize the reconstruction results by domain in Fig. 4.6. For most of the networks in every domain, we obtained accurate estimates of their descriptors, with the exception of $\varnothing$ and τ. Additionally, when comparing the SBM-reconstruction with a baseline, Transportation networks are systematically the harder to reconstruct, i.e., improvements in the error of reconstruction of all descriptors tend to be rare. Transportation networks are often embedded in two-dimensional spaces, resulting in large diameters and slow-mixing random walks. Some Technological networks have similar characteristics. Once again, this suggests that our assumptions are wrong for some networks, especially when having 2D embeddedness. In these

cases, our reconstruction would yield networks that violate spatial constraints, introducing links between distant nodes, creating triangles, and reducing the diameter of the network.

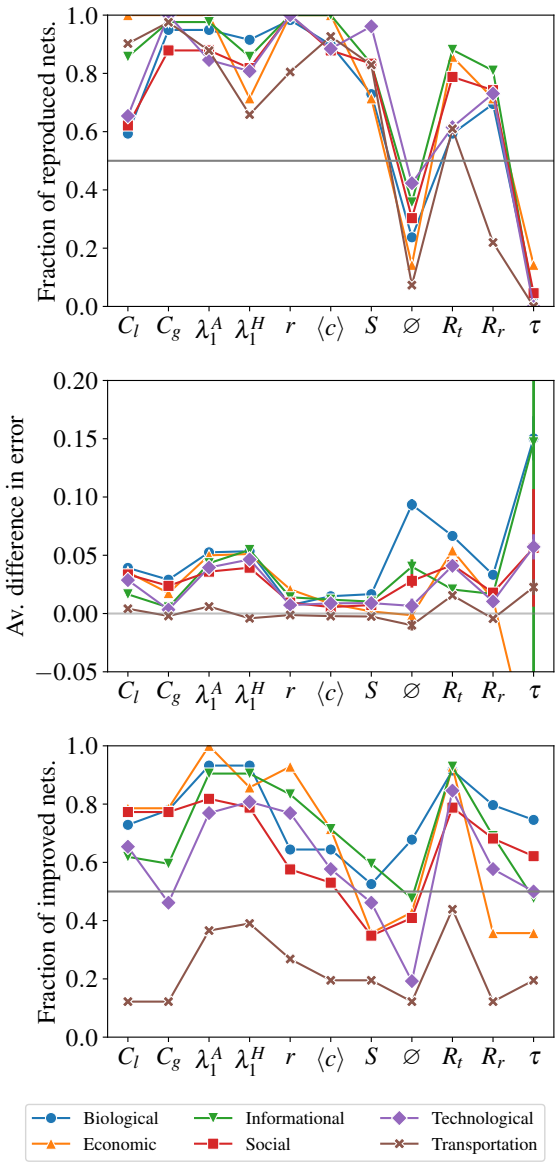


Figure 4.6: Reconstruction summaries by network domain. (Top) Fraction of reproduced networks (same criteria as before). (Middle) Average difference between error before reconstruction and after reconstruction. (Bottom) Fraction of networks whose error improved (same criteria as before). Noise level $p = 0.1$.

Increasing the number of measurements from 1 to 3

In the previous sections, we assumed that every node pair was measured once, i.e., $n = 1$. Although we showed that an accurate reconstruction was possible, we did not fully take advantage of our framework, since the procedure might have mostly relied on the model of network struc-

ture (SBM). If more repeated measurements n were available, there might have been sufficient information about the network structure, and therefore, the reconstruction procedure could have also relied on the error model. As a result, we would expect significant improvements in the performance of the reconstruction procedure.

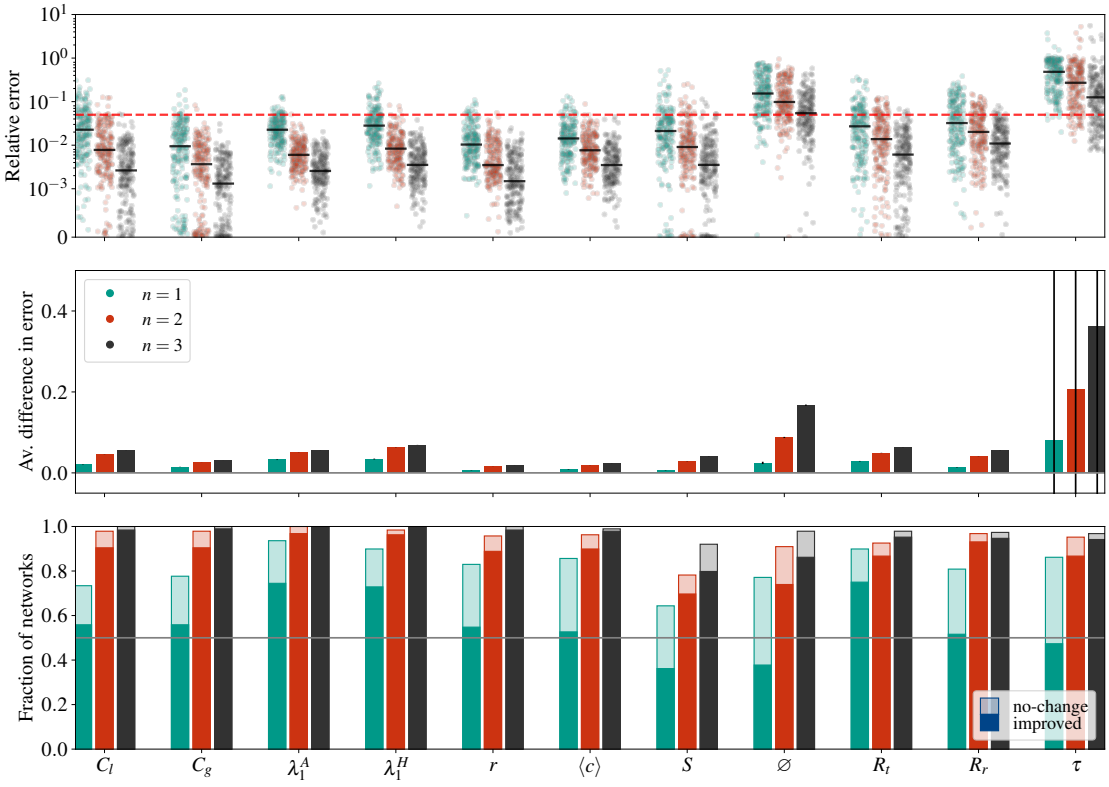


Figure 4.7: Summaries of the reconstruction error for 1, 2, and 3 measurements and noise level $p = 0.1$. (Top) Distribution of the reconstruction error. The red dashed line corresponds to 0.05, and the black continuous lines to the medians. (Middle) Average difference between the error before reconstruction and after reconstruction. (Bottom) Percentage of networks whose error improved (i.e., the error after reconstruction is smaller than before reconstruction), did not change, or worsened but the descriptor was still reproduced (error is smaller than 0.05) for each descriptor.

In this section, we attempt to obtain some insights on the effect of doing more measurements n on the reconstruction performance. We conduct a new analysis reverting to the denoising scenario described above. Here we set $n_{ij} = n$, $p_{ij} = p$, and $q_{ij} = pE/\left(\binom{N}{2} - E\right)$, such that the expected number of edges remain the same. Then we simulate x_{ij} from a binomial distribution with parameters $1-p$ and n if $A_{ij} = 1$, or parameters q and n otherwise.

We conducted this analysis on a subset of 188 networks of the original corpus, which correspond to those networks having at most 10^4 edges. We also considered the cases $n \in \{2, 3\}$.

In Fig. 4.7, we show that the quality of reconstruction of our framework systematically increases with a larger number of measurements n. The reconstruction error decreases in one

order of magnitude compared with the single measurement case (top panel). Furthermore, as n grows, the error after reconstruction is significantly smaller than before reconstruction.[5] When $n = 3$, almost all descriptors of the networks in our corpus are accurately estimated, with few exceptions occurring for τ.

It should also be noted that, the improvements obtained by increasing n from 1 to 2 are larger than from 2 to 3. This suggests that there might be a saturation after several increments of the number of measurements. Since researchers might have resource constraints for doing measurements (e.g., budget or time), it might be relevant to determine how many measurements one should make at most in order to obtain the maximum benefit from the reconstruction procedure. This issue is left as a further direction of research.

4.3 Concluding Remarks

In this chapter, we carried out a systematic analysis of the reconstruction performance of the SBM on a diverse corpus of empirical networks, spanning a broad range of sizes and domains. Using a variety of network descriptors, we observed that the SBM can provide accurate estimations of relevant features of empirical networks, even when we have only one measurement $n = 1$. Furthermore, for most networks and descriptors, we obtain smaller errors by doing reconstruction than by not doing so. The exceptions mainly include networks having large diameter and slow-mixing random walks. Most of them are transportation networks, which have a low-dimensional spatial embedding, where the "small-world" property is not fulfilled. We have also illustrated that by acknowledging the existence of errors in networks along with increasing the number of measurements, we can obtain significant improvements in our estimations. In this regard, it would be interesting to determine how fast this improvement changes, so that we obtain perfect or close to perfect reconstruction. This issue is left for future work.

It should also be noted that we can achieve such performance since our framework incorporates structured models (in this case, the SBM) that can recognize the structure in the data and extrapolate from it. Another benefit of our framework is that we can scrutinize the assumptions and assess the effectiveness of generative models incorporated as a prior. This kind of assessment is not possible under other non-statistical community detection methods which do not make their assumptions explicit.

Additionally, the conclusions obtained in this study are limited by the setup of our evaluation. First, we assumed uniform noise rates on node pairs. One possible direction for future work might consider other models of noise, e.g., having larger error rates around hubs. This might provide insights on how we can deal with systematic (or correlated) errors. Second, we focused

[5] For $n > 1$, the error before reconstruction is computed as the average error of an ensemble of graphs, such that for each entry (i, j) we sample an edge with probability x_{ij}/n_{ij}.

on two levels of noise. Thus, it would be interesting to consider larger levels of noise and study the sensitivity and reconstructability of structural descriptors. Another possibility consists on analyzing a larger set of descriptors that could reveal more relevant dimensions for the assessment, e.g., those related with dynamics happening on top of networks. Finally, we used the SBM as a prior, but our framework is flexible enough to incorporate other generative models. We hope to motivate future studies in this direction. In particular, it would be interesting to see if other models improve the estimation of some descriptors (e.g., diameter) and worsens others.

Chapter 5

Agreements and disagreements between compression and prediction of the SBM in empirical networks

In Chapter 3, we assessed the quality of fit of the SBM using posterior predictive checks. This allowed us to improve our understanding of the ways in which the model fits or deviates from data. Although this is an absolute assessment, one might be tempted to repurpose posterior predictive checks for model selection. Specifically, given network data and a set of competing models, we could compute discrepancy measures between summaries of fits and summaries of data. Then we could consider the most appropriate model to be the one that minimizes such discrepancies, i.e., the one that best reproduces the network summaries.

We argue that following such an approach might be misleading. If the goal is to get the model that is able to better reproduce network properties, we could consider an SBM that places each node in its own group. In this way, the adjacency matrix would be the same as the connectivity matrix between groups, and the features of sampled networks (there is only one possibility if the total number of edges is preserved) would be identical to the empirical one. Even though this model would achieve a perfect fit to the observed data, we would not learn anything meaningful about it, i.e., we would end up overfitting. Thus, to compare models and choose the most appropriate one, we need approaches that deal with overfitting and underfitting. In other words, to learn from the network data, we need models that are simple, but not too simple.

Two principled approaches to do model selection that address overfitting and underfitting are *compression* and *prediction*. The first approach relies on the Minimum Description Length (MDL) principle [24], described in Sec 2.3.5. It considers the best model as the one that most compresses the data, according to the description length. This approach penalizes overly complex models not supported by the data, thus preventing overfitting. In our initial example, when each node is placed in its own group by the model, there is no compression at all, as the model

merely encodes the network in a different form, using parameters (i.e., node memberships) instead. Consequently, we learn nothing about the network data from such model. To gain insights on the structure of the network, we need a simpler model that balances model fit and model complexity.

In contrast, the predictive approach to model selection favors the model that yields the best predictions. This approach can be framed in several ways, being link prediction one of the most common ones. Roughly speaking, link prediction involves using a model to predict links that have not been observed or have been removed. Returning to our initial example, when a model places each node in its own group, the link prediction performance within-sample is perfect, since the adjacency matrix, and consequently, summary descriptors are perfectly captured. However, the model will fail massively in predicting yet-to-be-observed links (e.g., true links that were not initially measured), thus lacking the ability to generalize. Once again, a simpler model (e.g., with fewer groups) is needed for achieving a better performance in the prediction task. The model should not be too simple (like an Erdős-Rényi model), because it would not be able to predict yet-to-be-observed links either.

Regardless of the approach we take to do model selection, we would like them to be *consistent*. In other words, if the true model, i.e., the model that generated the data, is among the set of competing models, we would expect that a given model selection approach favors the true model. Unfortunately, such consistency cannot always be achieved in every scenario. For instance, in the case of non-network data, Shao (1993) [181], and more recently Gronau and Wagenmakers (2019) [182], reported that data prediction using leave-one-out cross validation is not consistent, i.e., it does not favor the true model enough. However, Shao (1993) [181] also showed that consistency can be achieved when using instead k-fold cross validation. In the case of network data, Vallès-Català *et al.* (2018) [2] confirmed such inconsistency when removing one link, which would be the analogous of leave-one-out cross validation, and also when removing a fraction of links.

Although we may be willing to accept the limitations mentioned above, we may still expect that both compression and predictive approaches agree on the preferred model, since both attempt to deal with overfitting and underfitting. However, Vallès-Català *et al.* (2018) [2], using several variants of SBMs and network data, also showed that there might be discrepancies between model selection criteria, to the extent that, overly complex models give better predictions than more compressive models.

In this chapter, we revisit and expand upon such work in a more systematic way, by incorporating recent advances in SBMs. First, when doing link prediction, we utilize the network reconstruction framework from Peixoto (2018) [43], summarized in Chapter 4, as it allows us to address the predictive task in a principled and nonparametric manner. Second, we incorporate a measure of uncertainty for the predictive criterion, specifically the area under the ROC

curve (AUC), proposed by Hanley and McNeal (1982) [183]. Although the AUC is computed on the data, and thus has a corresponding uncertainty, this fact is often overlooked in studies evaluating the predictive performance of link prediction algorithms. Third, besides using the description length as compression criterion, we compare models according to the *model evidence* (see Eq. (2.15)). Although its computation is intractable, there have been attempts to approximate model evidence by first characterizing the posterior distribution of network partitions. We consider the approach proposed by Peixoto (2021) [52]. Finally, we harness the availability of more efficient MCMC algorithms to sample from the posterior distribution of network partitions [53], as well as the availability of more network datasets [54].

In this work, we aim to understand the frequency and magnitude of discrepancies between compression and predictive criteria for model selection, focusing on models of community detection in networks. To this end, we consider a corpus containing 392 empirical and synthetic networks, and fit two SBM variants (i.e., nested degree-corrected and nested non-degree corrected) to them. Then we obtain compression indices (description length and evidence) and predictive indices (AUCs obtained in a link prediction task, either by considering a point estimate of the node partition or the whole posterior distribution of node partitions). Based on these criteria we select the best model, and determine whether the most compressive model is the same as the most predictive one or not, when disagreements occur, and in which magnitude.

Our results show that, for synthetic networks, there is consistency between model selection criteria; the most compressive model is also the most predictive one. For empirical networks, consistency is prevalent, with few exceptions. Although agreements between model selection approaches are quite frequent, we observe that predictive criteria cannot always determine which model is better, as there are many cases in which the AUCs of competing models are statistically equivalent. On the contrary, both the description length and evidence consistently indicate which model compresses the data more and provide a degree of confidence for ruling out the alternative model. In this sense, the compression approach is more reliable for model selection in the context of community detection.

In the rest of this chapter, we describe the versions of the SBM being compared and the model selection criteria (Sec. 5.1), a motivating example where the predictive criteria yields misleading answers (Sec. 5.1.2), and the network corpus in which we fit the models along with the results of our analysis for both synthetic and empirical networks (Sec. 5.2). Finally, we provide some concluding remarks in Sec. 5.3.

5.1 Models and Model Selection Criteria

We focus on two versions of the SBMs, namely the hierarchical degree-correct SBM (HDC-SBM) from Eq. (2.33), and the non-degree corrected version (H-SBM), being the first one the

most complex model of the two.

As mentioned in Chapter 2, considering a model class (or variant) of the SBM with hierarchical partitions, denoted by $\mathcal{H}$, our inference framework consists on inferring the following posterior distribution:

$$P(\{\boldsymbol{b}_l\},\mathcal{H}|\boldsymbol{A}) = \frac{P(\boldsymbol{A}|\{\boldsymbol{b}_l\},\mathcal{H})P(\{\boldsymbol{b}_l\},\mathcal{H})}{P(\boldsymbol{A})}, \tag{5.1}$$

where $P(\boldsymbol{A}) = \sum_{\{\boldsymbol{b}_l\}} P(\boldsymbol{A}|\{\boldsymbol{b}_l\})P(\{\boldsymbol{b}_l\})$ is the model evidence.[1]

Under this approach it is possible to compare models and select the best among them. Here we consider two principled approaches to model selection, namely compression and prediction. Both of them attempt to deal with overfitting, although their goals are not the same. Under the first approach, the best model is the one that most compresses the data. Under the second approach, the best model is the one that yields better predictions. Besides this distinction, it is important to notice that SBMs are approximations to generative mechanisms of real-world networks. A single partition may not be a good fit for such networks, and therefore, it becomes important to consider less likely alternative partitions in order to fully capture the posterior uncertainty. In this regard, we consider two versions of each criterion: one that uses a single partition, which we call point estimate, and one that averages over partitions sampled from the posterior. We describe them in the rest of this section.

5.1.1 Compression Criterion

We present two ways in which we can do model selection using information theory, namely comparing single partitions or entire model classes. In our case, the comparison between two single partitions from the H-SBM and HDC-SBM is made via the ratio of posterior probabilities

$$\begin{aligned}\Lambda_1 &= \frac{P(\{\boldsymbol{b}_l\},\mathcal{H}_{\text{H-SBM}}|\boldsymbol{A})}{P(\{\boldsymbol{b}_l\}',\mathcal{H}_{\text{HDC-SBM}}|\boldsymbol{A})} \\ &= \frac{P(\boldsymbol{A},\{\boldsymbol{b}_l\}|\mathcal{H}_{\text{H-SBM}})}{P(\boldsymbol{A},\{\boldsymbol{b}_l\}'|\mathcal{H}_{\text{HDC-SBM}})} \times \frac{P(\mathcal{H}_{\text{H-SBM}})}{P(\mathcal{H}_{\text{HDC-SBM}})}.\end{aligned} \tag{5.2}$$

If we are *a priori* agnostic about how likely both model classes are, i.e., $P(\mathcal{H}_{\text{H-SBM}}) = P(\mathcal{H}_{\text{HDC-SBM}}) = 1/2$, then

[1] If we use a class of SBMs that does not consider a hierarchy of partitions, then $\{\boldsymbol{b}_l\}$ should be replaced by $\boldsymbol{b}$.

$$\Lambda_1 = \frac{P(\boldsymbol{A}, \{\boldsymbol{b}_l\}|\mathcal{H}_{\text{H-SBM}})}{P(\boldsymbol{A}, \{\boldsymbol{b}_l\}'|\mathcal{H}_{\text{HDC-SBM}})}. \tag{5.3}$$

We can use this criterion to choose the most plausible model given the data. More specifically, if $\Lambda_1 < 1$, then the evidence in the data favors the particular hierarchical partition $\{\boldsymbol{b}_l\}'$ together with the degree-corrected model variant. Contrarily, if $\Lambda_1 > 1$, then the alternative model along with its corresponding partition are favored. The value of 1 should not be taken as a hard threshold to immediately decide in favor of one model or another. Instead, we need to also consider the magnitude of Λ_1, which can be interpreted as the number of times one model is more likely than the alternative one, as an explanation for the data. In turn, this provides us the degree of confidence in taking such decision.

Importantly, the reader might have noticed that the numerator (and denominator) already appeared in Eq. (2.35), when referring to the description length Σ. In fact, the ratio of posterior probabilities can be written as

$$\Lambda_1 = \mathrm{e}^{-\Delta\Sigma}, \tag{5.4}$$

where $\Delta\Sigma = \Sigma_{\text{H-SBM}} - \Sigma_{\text{HDC-SBM}}$ is the difference in description length (in *nats*) considering one fitted partition per model class.

In this regard, choosing the most plausible model according to Λ_1 is equivalent to choosing the model that compresses the data the most, according to the MDL criterion. In this work, we consider the description length as the point estimate version of the compression criterion. It is important to note that using $\Delta\Sigma$ to do model selection within the same model class will favor the most compressive partition of such model class. Therefore, performing model selection among different model classes based on $\Delta\Sigma$ involves comparing the description lengths of the most compressive partitions $\{\boldsymbol{b}_l\}$ and $\{\boldsymbol{b}_l\}'$ for each model class.

Alternatively, we might want to compare entire model classes, i.e., we do not want to rely on a specific fit, but rather consider all possible fits. This approach becomes relevant when the posterior distribution is multimodal, i.e., when it contains several partitions having quite similar posterior probabilities. It is possible to compute summaries for a model class by averaging over all its possible partitions, weighting them by their posterior probability. As in the previous case, another ratio of posterior probabilities can be defined, i.e.,

$$\Lambda_2 = \frac{P(\mathcal{H}_{\text{H-SBM}}|\boldsymbol{A})}{P(\mathcal{H}_{\text{HDC-SBM}}|\boldsymbol{A})} = \frac{P(\boldsymbol{A}|\mathcal{H}_{\text{H-SBM}})}{P(\boldsymbol{A}|\mathcal{H}_{\text{HDC-SBM}})} \times \frac{P(\mathcal{H}_{\text{H-SBM}})}{P(\mathcal{H}_{\text{HDC-SBM}})}, \tag{5.5}$$

where

$$P(\boldsymbol{A}|\mathcal{H}) = \sum_{\{\boldsymbol{b}_l\}} P(\boldsymbol{A}|\{\boldsymbol{b}_l\},\mathcal{H})P(\{\boldsymbol{b}_l\}) = \sum_{\{\boldsymbol{b}_l\}} P(\boldsymbol{A},\{\boldsymbol{b}_l\}|\mathcal{H}) \tag{5.6}$$

is the *model evidence*, which appeared as the normalization constant of Eq. (5.1). If we have no prior preference for a model class, i.e., $P(\mathcal{H}_{\text{HDC-SBM}}) = P(\mathcal{H}_{\text{HDC-SBM}})$, then Λ_2 becomes the so-called *Bayes factor* [184]. It has a similar interpretation to Λ_1, but the statement about the model considers all its possible partitions. This ratio of posterior probabilities can be rewritten as

$$\Lambda_2 = \mathrm{e}^{-\Delta L}, \tag{5.7}$$

where $\Delta L = L_{\text{H-SBM}} - L_{\text{HDC-SBM}}$, with $L_{\mathcal{H}} = -\ln P(\boldsymbol{A}|\mathcal{H})$, is the difference in negative log-evidence.

In this work, we consider ΔL, instead of Λ_2, as the compression criterion that uses averages from the posterior distribution. We do so because ΔL has the advantage of having an information theoretical interpretation. In fact, it is possible to obtain a lower bound for the evidence from the most likely partition, i.e.,

$$P(\boldsymbol{A}) = \sum_{\{\boldsymbol{b}_l\}} P(\boldsymbol{A},\{\boldsymbol{b}_l\}) \geq \max_{\{\boldsymbol{b}_l\}} P(\boldsymbol{A},\{\boldsymbol{b}_l\}), \tag{5.8}$$

where we dropped the dependence on $\mathcal{H}$ to simplify the expressions. Since, maximizing the posterior is equivalent to minimizing the description length, the previous relation can be rewritten as

$$\begin{aligned} L = -\ln P(\boldsymbol{A}) &\leq \min_{\{\boldsymbol{b}_l\}} \Sigma(\boldsymbol{A},\{\boldsymbol{b}_l\}) \\ &= \min_{\{\boldsymbol{b}_l\}} -\ln P(\boldsymbol{A},\{\boldsymbol{b}_l\}) \\ &= \min_{\{\boldsymbol{b}_l\}} (-\ln P(\boldsymbol{A}|\{\boldsymbol{b}_l\}) - \ln P(\{\boldsymbol{b}_l\})). \end{aligned} \tag{5.9}$$

This means that, if we consider all possible partitions in the posterior distribution, instead of a single one, we can achieve a more efficient compression of the network data. It can be also noted that, the description length can be interpreted as the amount of information necessary to encode the network data in "two-parts", i.e., by first encoding a partition $\{\boldsymbol{b}_l\}$ and then the

network $\boldsymbol{A}$ conditioned on such partition. Analogously, L can be also interpreted as a description length, but considering a "one-part" encoding, since it does not need a specific partition to encode the network. Although it seems appealing to use the evidence for achieving better compressions and carry out model selection, its exact computation is intractable. The reason is that, for most cases of interest, such computation involves summing over a prohibitively large number of partitions. Unfortunately, the evidence cannot be estimated by sampling from the posterior with MCMC algorithms either. This issue can be shown by writing the logarithm of the evidence as the contribution of two terms, i.e.,

$$\ln P(\boldsymbol{A}) = \sum_{\{\boldsymbol{b}_l\}} \pi(\{\boldsymbol{b}_l\}) \ln P(\boldsymbol{A}, \{\boldsymbol{b}_l\}) - \sum_{\{\boldsymbol{b}_l\}} \pi(\{\boldsymbol{b}_l\}) \ln \pi(\{\boldsymbol{b}_l\}) \tag{5.10}$$

$$= \langle \ln P(\boldsymbol{A}, \{\boldsymbol{b}_l\}) \rangle + H(b), \tag{5.11}$$

where

$$\pi(\{\boldsymbol{b}_l\}) = \frac{P(\boldsymbol{A}, \{\boldsymbol{b}_l\})}{P(\boldsymbol{A})} \tag{5.12}$$

is the posterior distribution of Eq. (5.1). The first term is the mean joint log-probability computed over the posterior distribution,

$$\langle \ln P(\boldsymbol{A}, \{\boldsymbol{b}_l\}) \rangle = \sum_{\{\boldsymbol{b}_l\}} \pi(\{\boldsymbol{b}_l\}) \ln P(\boldsymbol{A}, \{\boldsymbol{b}_l\}), \tag{5.13}$$

and it can be estimated with MCMC methods by averaging $\ln P(\boldsymbol{A}, \{\boldsymbol{b}_l\})$ for sufficiently many samples. The second term is the entropy of the posterior distribution,

$$H(\{\boldsymbol{b}_l\}) = -\sum_{\{\boldsymbol{b}_l\}} \pi(\{\boldsymbol{b}_l\}) \ln \pi(\{\boldsymbol{b}_l\}), \tag{5.14}$$

which measures how concentrated is the posterior. Thus, if the posterior is concentrated in one partition, then the entropy would be close to zero, and therefore, $\Lambda_1 \approx \Lambda_2$, or equivalently, $\Sigma \approx L$. If this occurs for two model classes being compared, then the decision on which one is better will not change between the two versions of the compression criterion. However, if there are multiple partitions having similar probability in the posteriors, the decision of Λ_2 might lean towards the most entropic model class, even if their posterior probabilities are on average the same [57].

Importantly, the computation of $H(\{\boldsymbol{b}_l\})$ involves the computation of the log-posterior $\ln \pi(\{\boldsymbol{b}_l\})$

for every possible partition. In turn, this requires the value of $P(\boldsymbol{A})$ which is the quantity that we want to estimate. To overcome this problem, Peixoto (2021) [52] proposed an approach to estimate $\pi(\{\boldsymbol{b}_l\})$. It consists on fitting a mixed random label model to sampled partitions from the posterior distribution, which in turn, provides an approximation to the whole posterior distribution. With this tool in hand, we can compute the necessary terms to estimate the value of the evidence. In this work, we follow such approach, and describe it in App. C.1.

5.1.2 Predictive Criterion

Another approach to model selection relies on comparing the predictive power of the model candidates. One way in which this can be done is by carrying out link prediction in networks, i.e., identifying which edges have been deleted and which non-edges have been introduced as spurious edges. In this work, we focus in the case of edge denoising, where a fraction of edges have been deleted from the original network, so the corresponding entries of the adjacency matrix are assumed as observed but registered with zeros (i.e. as non-edges).

Thus, the predictive task consists on reconstructing the original network from a noisy network. This reconstruction is done by means of the inference framework of Ref. [43], which was also summarized in Sec. 4.1. More specifically, let $\boldsymbol{A}$ be the original network and assume that it was generated by a class of the SBM, denoted by $\mathcal{H}$. Let $\boldsymbol{D}$ be the observed data, which was obtained by removing edges from $\boldsymbol{A}$ with an error rate $p \in (0,1)$. The network $\boldsymbol{A}$ can be reconstructed according to the posterior of Eq. (4.1), and the posterior probability of an entry (i,j), conditioned on a partition $\{\boldsymbol{b}_l\}$, is given by

$$\pi_{ij}^{(1)} = \sum_{\boldsymbol{A}} A_{ij} P(\boldsymbol{A}|\{\boldsymbol{b}_l\},\mathcal{H},\boldsymbol{D}) P(\{\boldsymbol{b}_l\}|\mathcal{H},\boldsymbol{D}). \tag{5.15}$$

Here, we consider the partition of $\boldsymbol{D}$ that minimizes the description length for estimating $\pi_{ij}^{(1)}$, i.e.,

$$\{\boldsymbol{b}_l^*\} = \arg\max_{\{\boldsymbol{b}_l\}} P(\{\boldsymbol{b}_l\}|\mathcal{H},\boldsymbol{D}). \tag{5.16}$$

Under a binary classification task [185], these probabilities can be interpreted as ratings or scores, and used to compute indices that provide information about the predictive performance of the model. Here, we are interested in assessing the capacity of the model in distinguishing between missing edges and true non-edges. For this purpose, we compute $\pi_{ij}^{(1)}$ for each of them, and subsequently, compute a widely used index to measure the performance of a classifier, namely, the area under the ROC curve (AUC).

The ROC curve depicts the true positive rate (in this case, the percentage of missing edges

classified as edges) as a function of the false positive rate (in this case, the percentage of true non-edges classified as edges), which result from varying a threshold that discriminates between positive and negative instances. In this sense, better classifiers would have ROC curves being closer to the upper left corner, and consequently, the areas resulting from integrating such curves over all possible discrimination thresholds, i.e. the AUC, would be close to one. Importantly, the AUC can be also interpreted as the probability of correctly ranking a pair (missing-edge, true non-edge), i.e., the score of the missing-edge would be larger than the score of the true non-edge. In this regard, we expect that the AUC is 1/2 when the model is equally predictive as a random guess, while its value is 1 when the model provides a "perfect" ranking.

Since the estimates of the set of $\pi_{ij}^{(1)}$, and consequently of the AUC, were obtained with a single partition, we call them point estimates. Similarly, we can also obtain an estimate of the posterior probability of a node pair (i, j) by averaging from the posterior distribution, i.e.,

$$\pi_{ij}^{(2)} = \sum_{\boldsymbol{A},\{\boldsymbol{b}_l\}} A_{ij} P(\boldsymbol{A}|\{\boldsymbol{b}_l\},\mathcal{H},\boldsymbol{D}) P(\{\boldsymbol{b}_l\}|\mathcal{H},\boldsymbol{D}). \tag{5.17}$$

Thus, we get an estimate of the AUC that uses from posterior averages, by considering the set of $\pi_{ij}^{(2)}$ as scores. Regardless of which scoring rule is used, this process has to be repeated several times, in order to account for the possibility of having noisy networks with different structures. Then we summarize the resulting AUCs by computing their average and a measure of variability or uncertainty. Finally, when comparing two models, the best model would be the one that yields higher AUC in statistical terms.

Spurious AUCs from fluctuations

Both compression and predictive approaches to model selection incorporate regularization (via MDL Principle and cross-validation, respectively) and prevent overfitting. Thus, if we apply these criteria to randomly generated networks, we would expect that the true model is preferred over an overly complex model . However, we will demonstrate that it is possible to obtain better predictions (in terms of the AUC) by overfitting randomly generated networks when they are small.

In Sec. 2.2, we described the Erdős-Rényi model [97], which given fixed numbers of nodes N and edges E, generates networks by placing E distinct pairs of nodes uniformly at random from all possible pairs. Thus, this model can be also considered a special case of SBMs, in which all the nodes are placed in one group and there is no degree-correction. Since we use hierarchical priors, here we call this model class H-SBM.

We sampled hundreds of networks from the Erdős-Rényi model [97], varying the number of

nodes $N \in \{10, 20, 50, 100, 500, 100\}$ and average degree $\langle k \rangle \in \{4, 10, 20\}$. Furthermore, we considered 6 model candidates, which result from combining two classes of SBMs (H-SBM and its degree-correct variant HDC-SBM) and 3 possibilities for the number of groups B, namely B is not fixed *a priori* but inferred by minimizing the description length, or B is fixed to either 4 or 10 groups. We note that, the true model is among the set of competing models, which is the H-SBM with inference via MDL. For conciseness, we will present only the results for sampled networks having an average degree $\langle k \rangle = 4$, and include the other cases in App. C.2.

Since edges are placed uniformly at random in the network generation process, no model candidate should be able to predict missing edges better than random chance, resulting in an AUC of 0.5 for all candidates. Contrarily, Fig. 5.1 shows that for many cases, we can obtain better predictions by fitting more complex models to the data. Remarkably, there is not just one, but many ways in which we can achieve better predictions by overfitting, such as using the true model class (H-SBM) with more groups than necessary or using a more complex model class (HDC-SBM). Additionally, it is also possible to obtain better AUCs than those of the true model by using non-probabilistic methods, such as the the Jaccard similarity index [186] or the inverse log weighted similarity index [187], as shown in Fig. C.4.

We note that this inconsistency is more prevalent in smaller networks, where overfitting models yield better predictions in about 40% of sampled networks (see Figs. C.3 and C.5). This suggests that complex models might be able to exploit random fluctuations from edge removals in non-structured networks, particularly when these networks are small, making the AUC a potentially misleading model selection criterion. Only when averaging over the whole ensemble of networks we achieve consistency, i.e., no model outperforms the simplest one and neither they are better than random guessing. This rises concerns about the reliability of the AUC as a model selection criterion, especially because in real-world scenarios, we typically deal with one network rather than an ensemble.

Importantly, although this inconsistency is more prevalent in smaller networks, for larger networks we still observe small but statistically significant differences in AUC (see Fig. 5.1(c)). This prompts us to consider what constitutes a meaningful magnitude of such differences and how to properly quantify the uncertainty in the AUC to ensure predictive differences between models are not spurious. Here, we quantified the uncertainty in the AUC by the standard error of the mean, as is typically done in many studies comparing the predictive performance of link prediction algorithms.While this approach somewhat accounts for fluctuations in the AUC caused by random sampling of missing edges, it overlooks the fact that for *each* set of missing edges, the AUC is computed on the data, and therefore, also has an inherent uncertainty. Thus, by comparing models using the AUC without considering this aspect, we risk making erroneous conclusions due to spurious differences in AUC produced by random fluctuations, especially when networks are small.[2] This issue led us to consider a more suitable approach

[2] A related issue that might also interest the reader involves determining whether the values of the AUC obtained

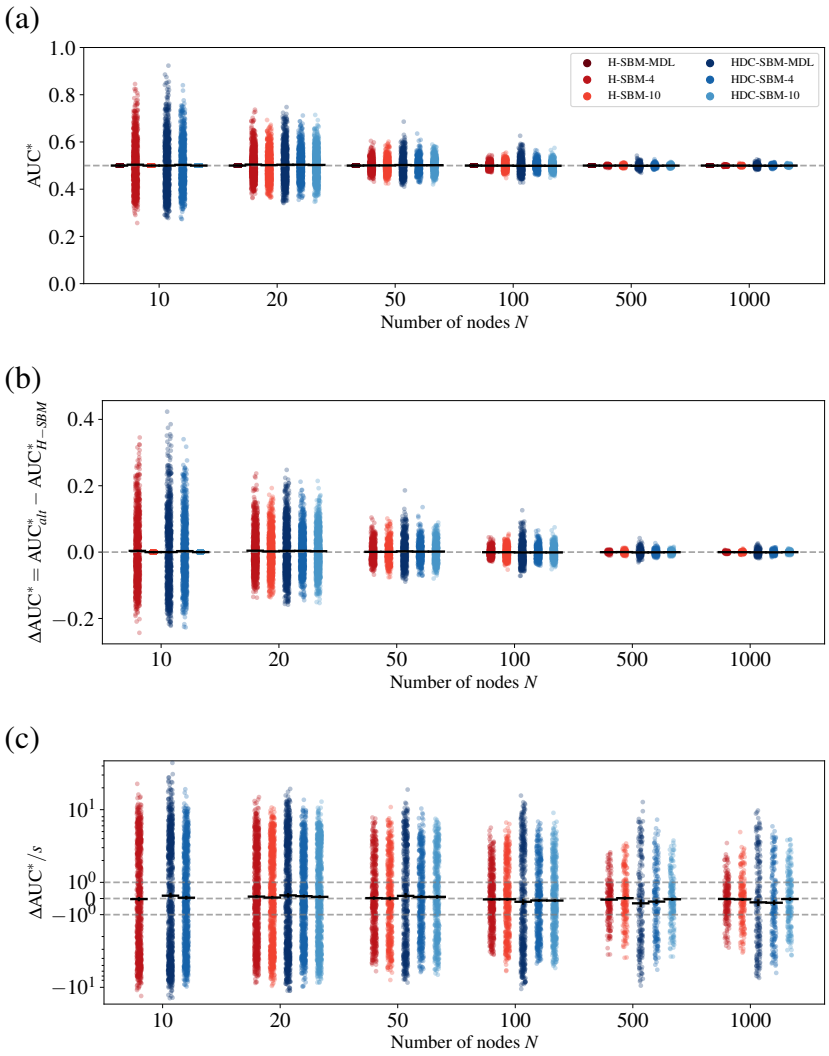


Figure 5.1: (a) AUC (point estimate) yielded by candidate models for several instances of the Erdős-Rényi model having average degree $\langle k \rangle = 4$, under an edge denoising task (at least 50 edge removal experiments removing 10% of edges on average were conducted). The point color indicates the model, which is a combination of the model variant (either H-SBM or HDC-SBM) and the number of groups. Each point corresponds to an instance of the Erdős-Rényi model, having N nodes and average degree $\langle k \rangle$. For $N \in \{10, 20\}$ and $\langle k \rangle = 4$, there are 1000 samples. For $N = \{50, 100\}$ and $\langle k \rangle = 4$, there are 500 samples. (b) Difference between the AUC (point estimate) yielded by simplest model AUC_{H-SBM} and the AUC yielded by more complex alternative models AUC_{alt}. The point color indicates the alternative model. (c) Ratio between the difference in AUC (in panel (b)) and the corresponding standard deviation of the mean AUC difference.

to quantify the uncertainty in AUC and avoid misleading conclusions from spurious predictive differences. We describe it in the following.

by generating random rankings (or scores) or reshuffling the original ones will be compatible with the AUC computed initially. Exploring these possibilities is beyond the scope of this work.

Uncertainty in the AUC

In this work, we quantify the uncertainty in the AUC for each realization following the approach of Hanley and McNeal (1982) [183], which is described in the following. Consider the scores of true positives X and true negatives Y as normally distributed according to

$$X \sim \mathcal{N}(\mu_+, \sigma_+^2) \qquad \text{and} \qquad Y \sim \mathcal{N}(\mu_-, \sigma_-^2). \tag{5.18}$$

For a particular cutoff value of a criterion variable, c, the true positive rate is given by

$$\mathrm{TPR}(c) = P(X > c) = 1 - \Phi\left(\frac{c-\mu_+}{\sigma_+}\right) = \Phi\left(\frac{\mu_+ - c}{\sigma_+}\right), \tag{5.19}$$

where $\Phi(z)$ is the cumulative distribution function of the standard normal distribution. Similarly, the false positive rate is given by

$$\mathrm{FPR}(c) = P(Y > c) = 1 - \Phi\left(\frac{c-\mu_-}{\sigma_-}\right) = \Phi\left(\frac{\mu_- - c}{\sigma_-}\right). \tag{5.20}$$

The ROC curve is defined by tracing out the functions

$$[\mathrm{TPR}(c), \mathrm{FPR}(c)] = \left[\Phi\left(\frac{\mu_+ - c}{\sigma_+}\right), \Phi\left(\frac{\mu_- - c}{\sigma_-}\right)\right]. \tag{5.21}$$

Then the area under the ROC curve (AUC) is defined as

$$\mathrm{AUC} = \int_{-\infty}^{\infty} \mathrm{TPR}(c)\mathrm{FPR}'(c)\mathrm{d}c \tag{5.22}$$

$$= \int_{-\infty}^{\infty} \Phi\left(\frac{\mu_+ - c}{\sigma_+}\right) \Phi\left(\frac{\mu_- - c}{\sigma_-}\right) \left(-\frac{1}{\sigma_-}\right) \mathrm{d}c. \tag{5.23}$$

Using this formulation, Hanley and McNeal (1982) [183] exploit the connection between the AUC and the Wilcoxon statistic [188] to derive a standard error of the AUC that depends on its estimated value and the imbalance between true positives and true negatives. Specifically, given an estimate $\widehat{\mathrm{AUC}}$ of the AUC, its standard error is given by

$$s_{\mathrm{AUC}} = \sqrt{\frac{\widehat{\mathrm{AUC}}(1-\widehat{\mathrm{AUC}}) + (n_e - 1)(Q_1 - \widehat{\mathrm{AUC}}^2) + (n_{ne} - 1)(Q_2 - \widehat{\mathrm{AUC}}^2)}{n_e \times n_{ne}}}, \tag{5.24}$$

where n_e is the number of true positives (in our case, removed true edges) and n_{ne} is the number of true negatives (true non-edges), and

$$Q_1 = \frac{\widehat{\text{AUC}}}{2 - \widehat{\text{AUC}}} \qquad \text{and} \qquad Q_2 = \frac{2\widehat{\text{AUC}}^2}{1 + \widehat{\text{AUC}}}. \tag{5.25}$$

Additionally, if we consider a simple graph having N nodes and average degree $\langle k \rangle$, removing a proportion f of edges results in

$$n_e = \frac{fN\langle k \rangle}{2} \tag{5.26}$$

missing edges, and

$$n_{ne} = \binom{N}{2} - \frac{N\langle k \rangle}{2} = \frac{N(N-1-\langle k \rangle)}{2} \tag{5.27}$$

true non-edges. Then Eq. (5.24) becomes

$$s_{\text{AUC}} = \sqrt{\frac{\widehat{\text{AUC}}(1-\widehat{\text{AUC}}) + \left(\frac{fN\langle k \rangle}{2} - 1\right)(Q_1 - \widehat{\text{AUC}}^2) + \left(\frac{N(N-1-\langle k \rangle)}{2} - 1\right)(Q_2 - \widehat{\text{AUC}}^2)}{\frac{fN^2\langle k \rangle(N-1-\langle k \rangle)}{4}}}, \tag{5.28}$$

with Q_1 and Q_2 defined as before. If $N \to \infty$, and the network is sparse, then s_{AUC} scales as $1/\sqrt{N}$. In Fig. 5.2(a,b), we illustrate how the standard error of the AUC varies with the number of nodes N. We note that, for the same number of nodes N, average degree $\langle k \rangle$, and fraction of removed edges f, there can be a significant difference in the magnitude of the standard errors as the AUC changes (here, from 0.7 to 0.9), particularly in smaller networks. Finally, in Fig. 5.2(c) we show how the standard deviation changes with different AUC estimates, for various number of nodes, while keeping $\langle k \rangle = 5$ and $f = 0.2$. Importantly, the observed magnitudes of s_{AUC} may have practical implications for interpreting the results from studies comparing the predictive performance of multiple algorithms. For example, for $N = 1000$ and AUC $= 0.9$, the value of s_{AUC} is around 0.01. In several studies, comparisons often rely on AUC differences even smaller than such s_{AUC} value (e.g., see Ref. [189–191]).

The final standard error, which involves several sets of edge removals, is obtained via error propagation. Specifically, if we have performed k sets of edge removals, and for each set $i \in \{1, ...k\}$ we compute the AUC and its uncertainty from Eq. (5.24) $(\widehat{\text{AUC}}_i, s_{\text{AUC}_i})$, then the final estimates for the average and standard error of the AUC are given by

$$\overline{\text{AUC}} = \frac{1}{k}\sum_{i=1}^{k} \widehat{\text{AUC}}_i, \tag{5.29}$$

and

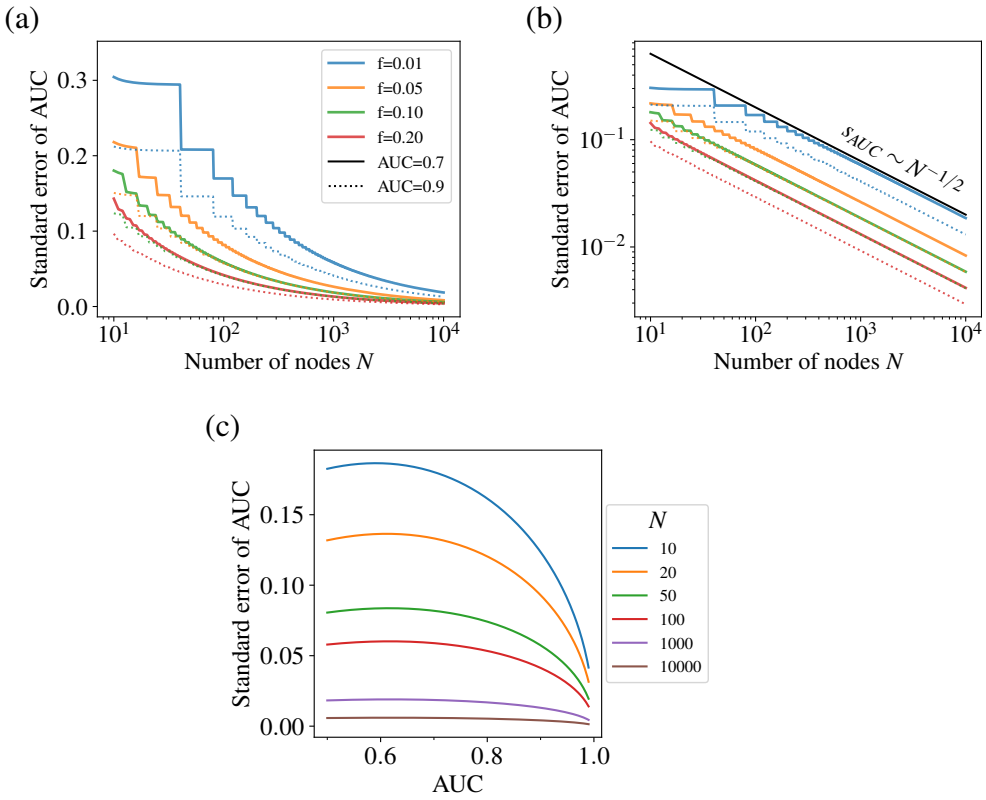


Figure 5.2: (a) Standard error of the AUC (see Eq. (5.28)) as a function of the number of nodes in the network N. The average degree $\langle k \rangle$ is 5. The line color indicates the percentage of removed edges f, and the line type indicates different values of AUC. (b) Similar to panel (a) with y-axis in logarithmic scale. The scaling $N^{-1/2}$ is shown in black. (c) Standard error of the AUC as a function of the AUC. The average degree $\langle k \rangle$ is 5, and the fraction of removed edges f is 0.1. The line color indicates different number of nodes.

$$s_{\overline{\text{AUC}}} = \sqrt{\frac{\sum_{i=1}^{k} (s_{\text{AUC}_i})^2}{k}}. \tag{5.30}$$

As shown in Fig 5.3, quantifying the uncertainty in the AUC in this way, mostly eliminates the inconsistencies. Thus, we will use this approach in the remaining of this chapter.

Before concluding, the reader may wonder whether the description length shows any inconsistency when used for model selection on random networks. Regarding our example of the Erdős-Rényi model [97], the theory states that fitting samples from this model using the Bayesian SBM framework described in Chapter 2.3.5 should correctly identify the true partition, where all nodes are placed in a single group. Within this framework, inferring a partition of the network is equivalent to compressing it, and overfitting can be avoided due to Shannon's source coding theorem [121]. As mentioned, this theorem states that the best compression can be achieved asymptotically only with the true model, which in this case is a SBM with one group.

Previous studies [11, 22] have empirically demonstrated this, and our results align with these

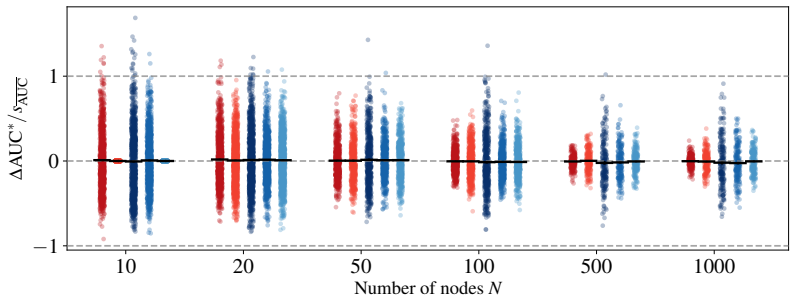


Figure 5.3: Ratio between the difference in AUC (Fig. 5.1(b)) and the corresponding standard deviation of Eq. 5.28, for several instances of the Erdős-Rényi model having average degree $\langle k \rangle = 4$. The point color indicates the alternative model.

findings. Specifically, Fig. 5.4 shows that according to description length, the true model — being the simplest— consistently achieves the smallest compression. This demonstrates that the compression approach to model selection is also consistent in practice, even for small networks. Notably, even when considering only a more complex model class (HDC-SBM), the compression approach would still favor a single-group partition. This agrees with the true model, but requires more information to describe the additional parameters. This underscores the usefulness of the Minimum Description Length (MDL) principle in providing meaningful partitions, even when the true model generating the data is unknown.

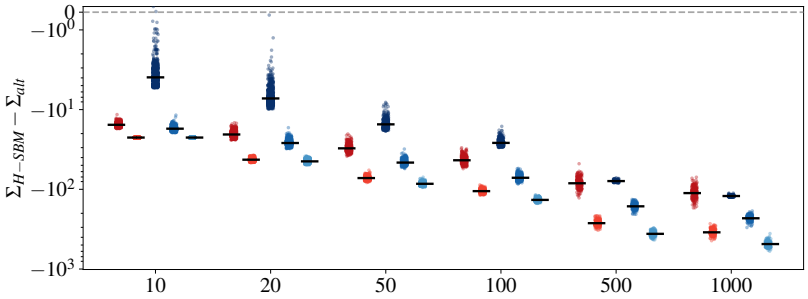


Figure 5.4: Difference between the description length of the simplest model Σ_{H-SBM} and the description length of more complex alternative models Σ_{alt}, for several instances of the Erdős-Rényi model. The point color indicates the alternative model, which is a combination of the model variant (either H-SBM or HDC-SBM) and the number of groups. Each point corresponds to an instance of the Erdős-Rényi model, having N nodes and average degree $\langle k \rangle$. For $N \in \{10, 20\}$ and $\langle k \rangle = 4$, there are 1000 samples. For $N = \{50, 100\}$ and $\langle k \rangle = 4$, there are 500 samples. For the remaining values of N and $\langle k \rangle$, there are 200 samples.

Finally, the reader might argue that samples from the Erdős-Rényi model are not representative examples of real-world networks, so these results should not be overemphasized. In the following sections, we will study more realistic scenarios, including empirical networks. As we will see, when dealing with more structured networks and incorporating Eq. (5.24) into the analysis, consistency can be expected, albeit with several nuances.

5.2 (Dis)agreements between Compression and Prediction

We carry out our analysis in a corpus of 196 real-world networks spanning various domains and several orders of size magnitude, as shown in Fig. 5.5. As before, when gathering the networks, we attempted to have networks as structurally diverse as possible. Additionally, every empirical network in the corpus is a simple graph, i.e., we considered symmetrized versions of directed networks removing parallel edges and self-loops. These networks can be downloaded from the Netzschleuder repository [54].

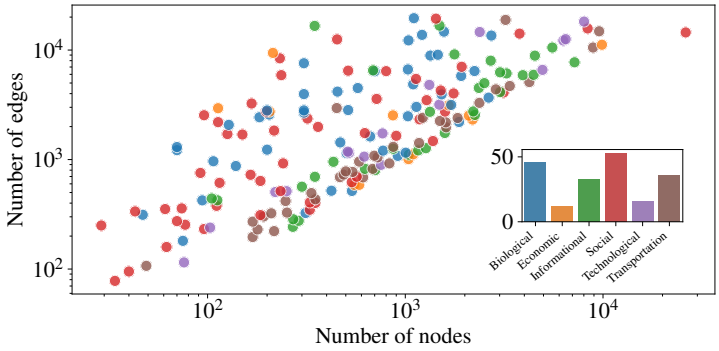


Figure 5.5: Number of nodes and edges of the networks in the corpus, and its domain composition.

Furthermore, we also considered a corpus of synthetic networks in our analysis. For each empirical network, a corresponding synthetic network was sampled from the SBM fit that minimized the description length (either HDC-SBM or H-SBM). For convenience, we first focus on model selection criteria based on point estimates, and then discuss the results using estimates obtained by averaging from the posterior distribution of network partitions.

5.2.1 Model Selection according to Point Estimates

Evaluation in Synthetic Networks

Although synthetic networks may not be exact representatives of empirical networks, studying them is valuable for two reasons. First, samples from SBM fits may contain relevant features of its empirical counterparts, as we have seen in Chapters 3 and 4. Second, in this scenario, we know the true model generating the network (i.e., an SBM variant), allowing us to exactly determine whether a model selection criteria favors the true model, and thus the correct one. This also provides a baseline for what to expect when the SBM is only an approximation of the true network generating mechanism, as is the case with empirical networks.

In Fig. 5.6(a), we show the percentage of networks for which description length and AUC (point estimate) agree or disagree. For almost all synthetic networks in the corpus, both criteria agree.

However, the standard error of the difference in AUC is typically large, resulting in roughly 40% of networks where the AUC cannot definitely rule out either the true or the alternative model. Thus, we cannot interpret these cases as disagreements, but neither can we consider them complete agreements. It is important to note that the magnitude of these standard errors cannot be reduced by increasing the number of edge removal sets. Each set has its own confidence interval, which depends on the corresponding AUC value and the imbalance between missing edges and actual non-edges.

In Fig. 5.6(c) we show the average difference in AUC (point estimate) as a function of the difference in description length. This plot shows that the description length consistently favors the true model, while the decisions given by the AUC are not always accurate. Besides the inconclusive cases mentioned above, there are two discrepancies for which the differences in the predictive criterion are at most 0.01. Importantly, the shape of the marginal distributions for difference in criteria differs significantly. The description length shows a bimodal distribution, where values do not concentrate around zero, whereas for the AUC, small differences are more frequent. Indeed, for around 40% of networks, this criterion cannot provide a decision, i.e., both the true and the alternative models are equally predictive, and neither can be discarded. This suggests that the AUC may not be sensitive enough to distinguish between competing variants of the SBM.

To compare both criteria across the whole corpus of synthetic networks, we use precision and recall indices. Specifically, we use a zero difference in criterion values as the threshold to prefer one model over another, then compute precision and recall indices relative to the true model. As shown in Fig. 5.6(b), the description length outperforms the AUC (point estimate) according to these indices. However, a zero value difference might be too stringent for model selection, as it does not account for the magnitude of the difference, i.e., the confidence in rejecting a model. To address this, we consider a simple classification task of synthetic networks. In this task, the target labels are the model variants that generated the networks, while the features are differences in description length or AUC. The ability of these criteria to distinguish between two competing models is measured by the AUC resulting from these classifiers (hence, an AUC of the AUC (point estimate)). This measure shows that both criteria are statistically equally capable of distinguishing between the true model and an alternative one.

These results suggest that both criteria are broadly consistent, although the description length is more sensitive than the AUC (point estimate) in selecting the true model. Before we proceed with empirical networks, we note that these results differ from those obtained for instances of the Erdős-Rényi model. The synthetic networks in our corpus are more structured than Erdős-Rényi instances, indicating as networks become more structured, there is less room for fluctuations that lead to overfitting. Additionally, we incorporated the standard error of the AUC from Eq. (5.24), which addresses the limitations of empirical studies that lack of a proper quantification of the uncertainty in AUC, reducing the risk of incorrectly discarding a model in

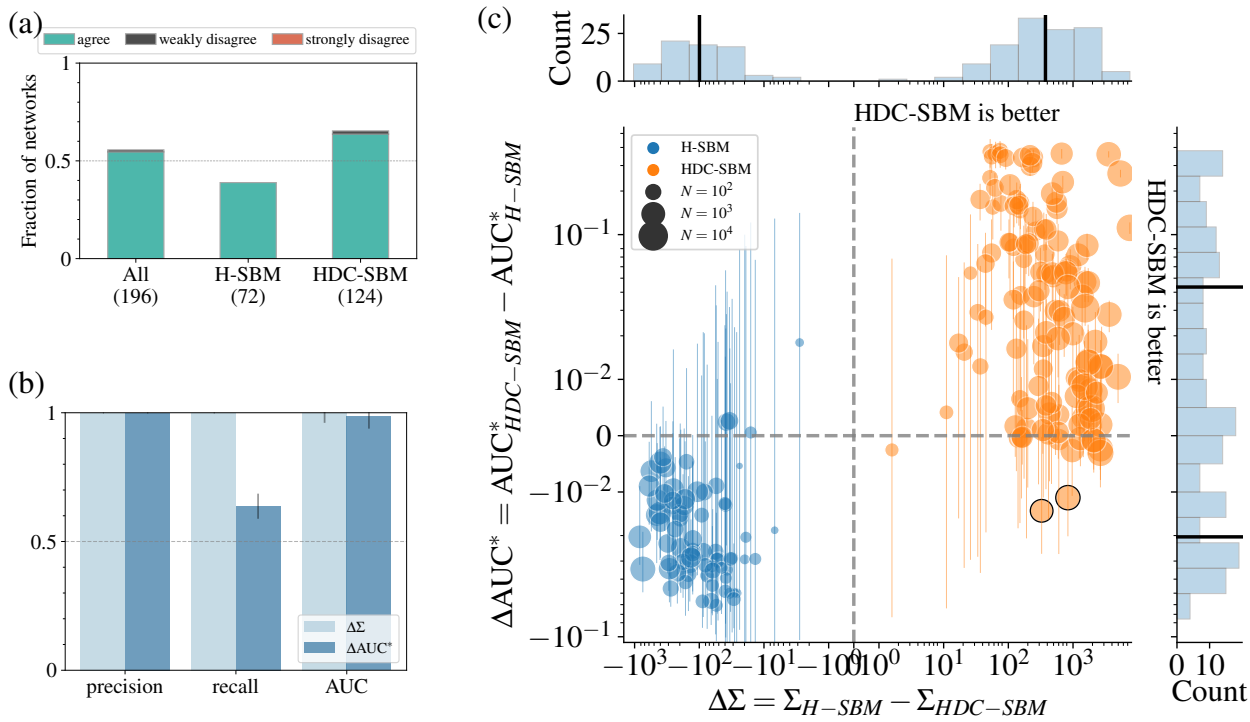


Figure 5.6: (a) Percentage of synthetic networks for which description length (Σ) and AUC (point estimate) (AUC*) agree, weakly disagree, or strongly disagree. We distinguish between the two latter categories to help interpretation and visualization. For a given network, there is a strong disagreement between model selection criteria when besides favoring different models, the difference in compression criteria is larger than 10 and the difference in predictive criteria is larger than 0.02. There is a weak disagreement when only one of the conditions is fulfilled. (b) Treating model selection as a classification task, we computed precision, recall, and AUC for both model selection criteria. Error bars are obtained by bootstrapping. (c) Average difference in AUC (best partition) vs difference in description length. Each point corresponds to a synthetic network. The error bar of points corresponds to the standard error of AUC of Eq. (5.24). The point color indicates the SBM variant from which the network was sampled. The point edge color highlights disagreements according to panel (a). Point sizes are proportional to the number of nodes N. We also include marginal histograms and the medians of positive and negative values in black lines.

the absence of statistically significant predictive differences.

Evaluation in Empirical Networks

Unlike the previous scenario with synthetic networks, we do not know the true generating mechanism of empirical networks. Since SBMs only provide an approximation, we can expect more disagreements between compression and predictive criteria. Furthermore, the reasons for discrepancies in empirical networks might also differ from those in synthetic cases.

We confirm such expectation in Fig. 5.7, which is an analog to Fig. 5.6 for empirical networks. Panel (a) shows that, overall, the description length and AUC (point estimate) agree. However, in at least 50% of these cases, the AUC cannot distinguish between competing models. It

becomes more challenging for the AUC to select a model when networks are empirical rather than synthetic. In panel (b), we note that the shape of marginal histograms differ between criteria. In particular, the bimodality persists for differences in description length, allowing for more confident decisions using this criterion. Differently, the AUC tends to favor the more complex model (HDC-SBM), regardless of the network domain (see Fig. C.8(b)). Despite the difference in shapes, it is noteworthy that, values around zero are more likely for both criteria than in the case of synthetic networks. This suggests that even for the description length, distinguishing among SBM variants might be harder with empirical network data. Additionally, we find some disagreements, although the magnitude of these discrepancies is usually not large. The exception corresponds to a small friendship network among 29 seventh grade students in Victoria, Australia [192].

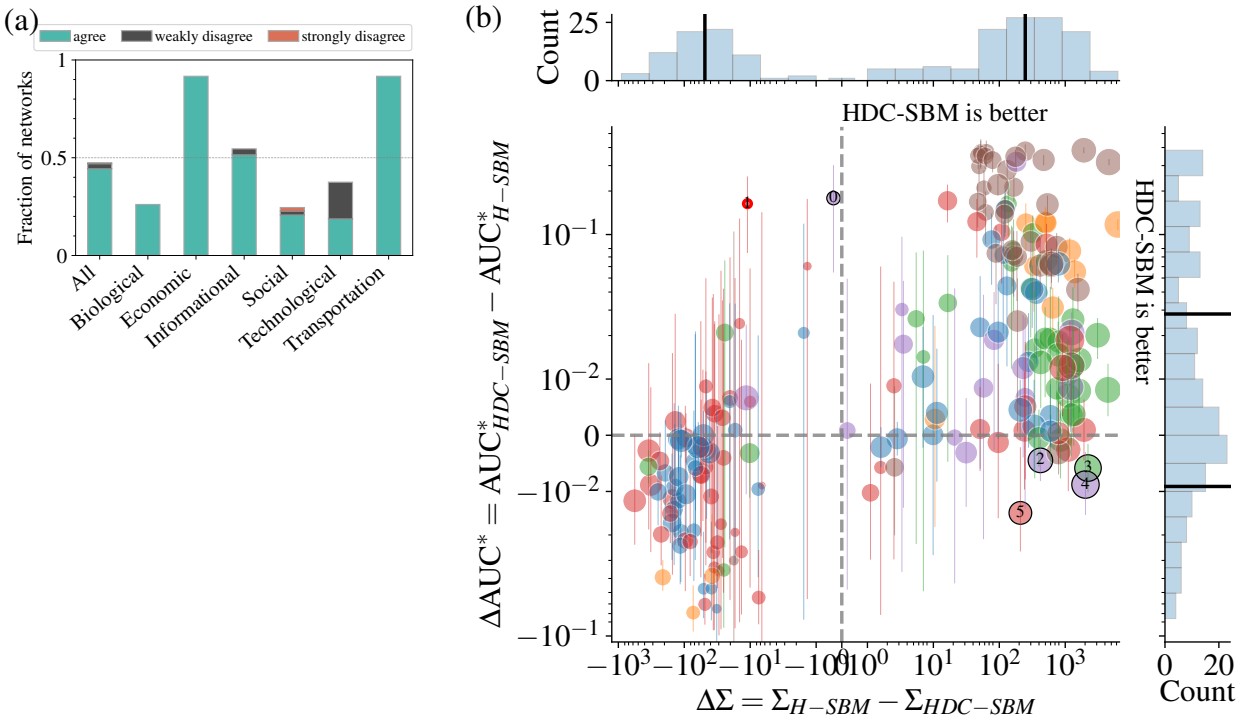


Figure 5.7: (a) Percentage of empirical networks for which description length (Σ) and AUC (point estimate) (AUC^*) agree, weakly disagree, or strongly disagree. We distinguish between the two latter categories to help interpretation and visualization. For a given network, there is a strong disagreement between model selection criteria when besides favoring different models, the difference in compression criteria is larger than 10 and the difference in predictive criteria is larger than 0.02. There is a weak disagreement when only one of the conditions is fulfilled. (b) Average difference in AUC (best partition) vs difference in description length. Each point corresponds to an empirical network. The error bar of points corresponds to the standard error of AUC of Eq. (5.24). The point color indicates the domain to which a network belongs to (as in Fig. 4.2). The point edge color highlights disagreements according to panel (a). Point sizes are proportional to the number of nodes N. We also include marginal histograms and the medians of positive and negative values in black lines.

Furthermore, the results for network domains, shown in Fig. 5.7(a), prompt a reevaluation of the

conclusions from previous chapters. Specifically, we observed that the HDC-SBM could not accurately capture structural features of Transportation networks (e.g., see Figs. 3.6 and 4.6). Interestingly, this is one of the domains with more agreement between description length and AUC, with the HDC-SBM being the most favored variant by both criteria across the corpus. This might occur because the degree distributions of transportation networks (especially urban street networks) are highly homogeneous, even more so than a Poisson distribution, making it difficult for the non degree-corrected SBM to account for them. This underscores the idea that a model can be the best model among a set of alternatives but still perform poorly in capturing certain aspects of network structure.

On the contrary, we noted in previous chapters that the HDC-SBM tends to perform well in capturing properties of social networks. However, this is one of the domains with fewer agreements between criteria. The non degree-corrected SBM is mostly preferred by the MDL approach, while the more complex version is preferred by the AUC (see Figs. C.8(a-b)). This highlights the potential unsuitability of using posterior predictive checks as a model selection criterion, as more parsimonious models might capture network properties as well as their more complex counterparts, while also offering a simpler explanation for the data.

5.2.2 Model Selection according to posterior averages

From point estimates to the posterior distribution

In the previous section, we relied on a single network partition for model selection. However, a single partition only provides a partial view of the data and models. More specifically, if the posterior distribution of network partitions is concentrated around a single partition, then considering alternative partitions, computing the corresponding versions of compression and prediction indices, and carrying out model selection, would yield similar results to what we have already seen. Nonetheless, it is possible for the posterior to contain few closely plausible partitions, representing the modes of the distribution, which offer different explanations for the data. Another possibility is a significantly broad posterior distribution with no single dominating partition, which might be a sign of model misspecification. In any of the latter cases, we have no guarantee that the same agreements or disagreements observed with point estimates versions of model selection criteria will hold.

To gain insight into the broadness of the posterior distributions of the models discussed in this chapter, we computed their number of modes using the approach described in App. C.1. Fig. 5.8(b) shows the distribution of the effective number of modes[3] for both SBM variants across synthetic and empirical networks. We observe that the posterior distribution of net-

[3]The effective number of groups B_e of a network partition is defined as: e^H, with $H = -\sum_r \frac{n_r}{N} \ln \frac{n_r}{N}$, where n_r is the number of nodes in group r and N is the number of nodes in the network. The effective number of modes is

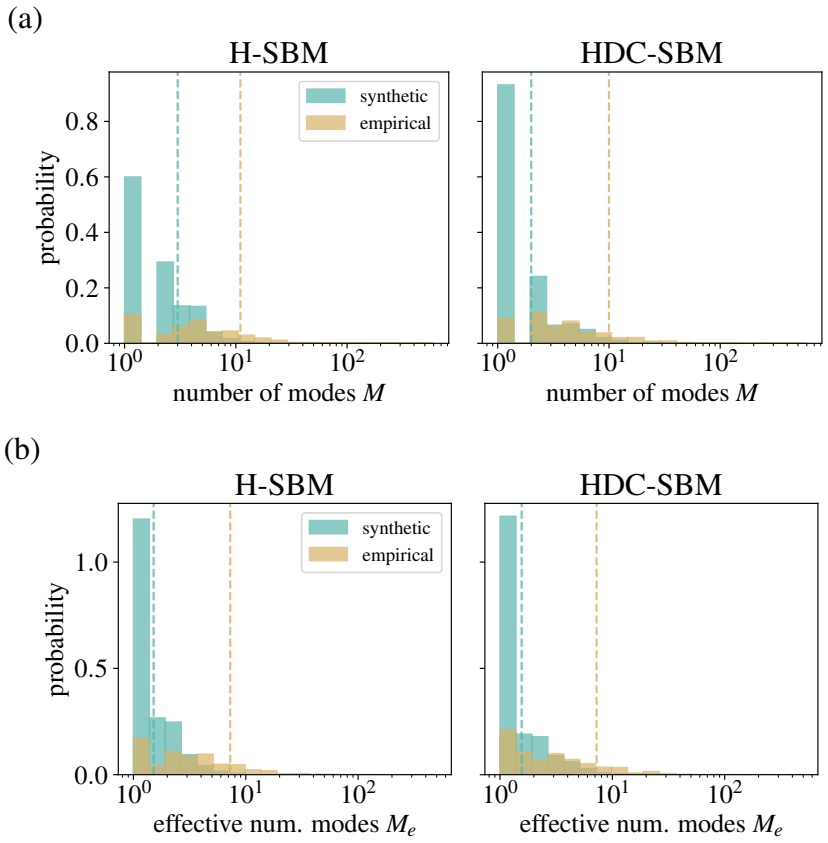


Figure 5.8: Distribution of the number of modes M (a) and effective number of modes M_e (b) of the posterior distributions obtained by fitting SBM variants to synthetic and empirical networks. Vertical dashed lines indicate the corresponding medians.

work partitions is generally broader for empirical networks compared with synthetic ones, with median effective number of modes 7.2 and 1.5, respectively. This finding agrees with the expectation that the SBM better fits its samples than other types of networks. Another indication of the multimodality or broadness of the posterior distributions comes from analyzing the AUC. Fig. C.12 shows that averaging tends to improve the AUC, especially for the simplest model variant (H-SBM) and empirical networks. This suggests that a single partition might be an inaccurate description of the data, and other alternative explanations need to be incorporated to achieve better predictions.

Overall, we may expect that the patterns observed on synthetic networks, i.e., agreement between compression and predictive criteria, hold. Therefore, a single-point estimate approach could be enough to carry out model comparison on this type of networks. However, it is not clear what we should expect for empirical networks, since they can be seen as a combination of structure and noise, and for which the SBM is an approximation. In this section, we address this issue by considering the model evidence as compression criteria and the AUC that results from sampling partitions from the posterior distribution as predictive criteria.

defined analogously, with $H = -\sum_k \omega_k \ln \omega_k$, where ω_k is the proportion of partitions that belong to mode k.

Evidence and AUC (from posterior averages)

Compression and prediction fully agree in synthetic networks, as shown in Fig. 5.9(a). In Fig. 5.9(c), there are three cases where the evidence appears to favor the wrong model. However, because the differences are small, and we have few confidence in discarding the true model, these cannot be considered as mistakes of the criterion.[4] Furthermore, for each model selection approach, its point estimate version is consistent with its posterior averaging version (see Figs. C.10 and C.11). This result suggests that SBMs are reliable models, to the extent that there is consistency in the model selection approaches when networks are sampled from the SBM.

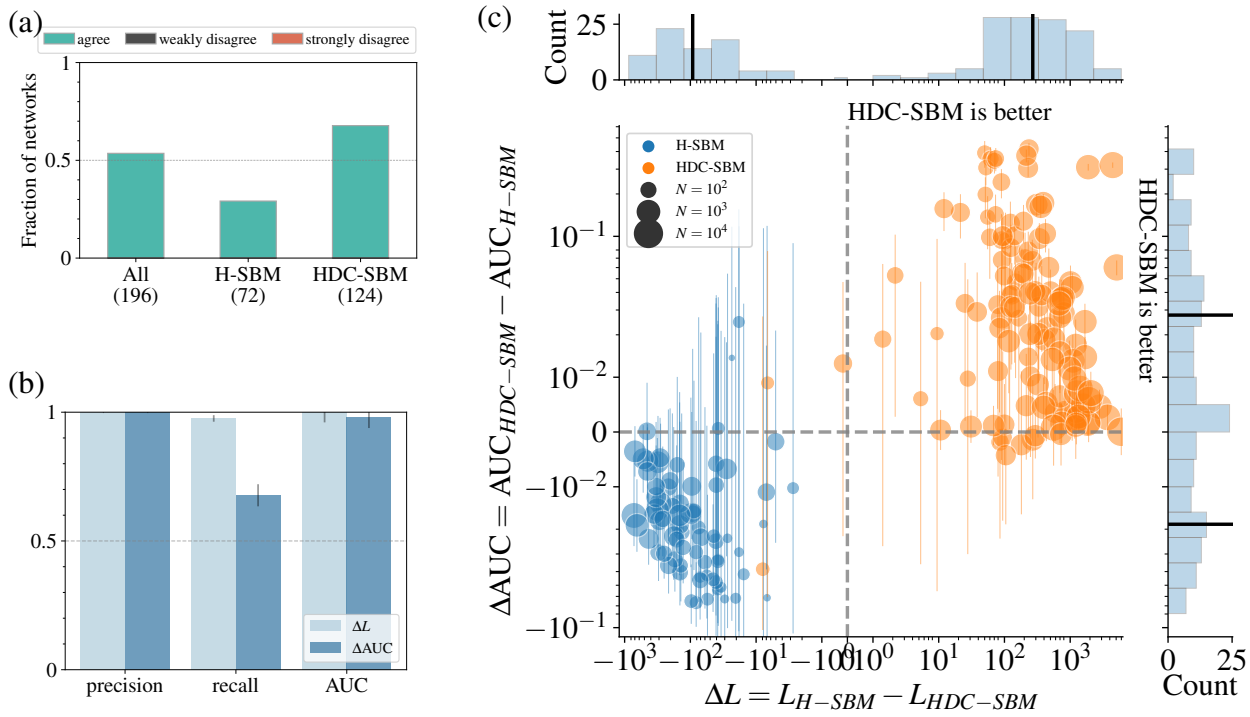


Figure 5.9: Panels (a), (b), and (c) have the same explanation as in Fig. 5.6, but in this case, we consider the difference in -log-evidence (ΔL), as compression criterion, and the difference in AUC (from posterior averages) (ΔAUC), as predictive criterion.

For empirical networks, Fig. 5.10(a) shows that there are mostly agreements between criteria, with a few exceptions. In particular, Fig. 5.10(b) shows that there are three networks with significant disagreements: two networks of social interactions among university students within the Copenhagen Networks Study (*copenhagen/calls*, *copenhagen/sms*) [193]), and a network of international "E-roads" (*euroroad*) [194]. Since both versions of the AUC are consistent (see Fig. C.11(b)), these discrepancies can be attributed solely to disagreements between their description length and model evidence, as shown in Fig. C.10(b).

These discrepancies between different versions of the compression criterion may not be a

[4]These networks have 96, 112, and 433 nodes, respectively. The effective number of groups (see footnote 3) in the true model of these networks is 1.96, 1, and 2.11, respectively.

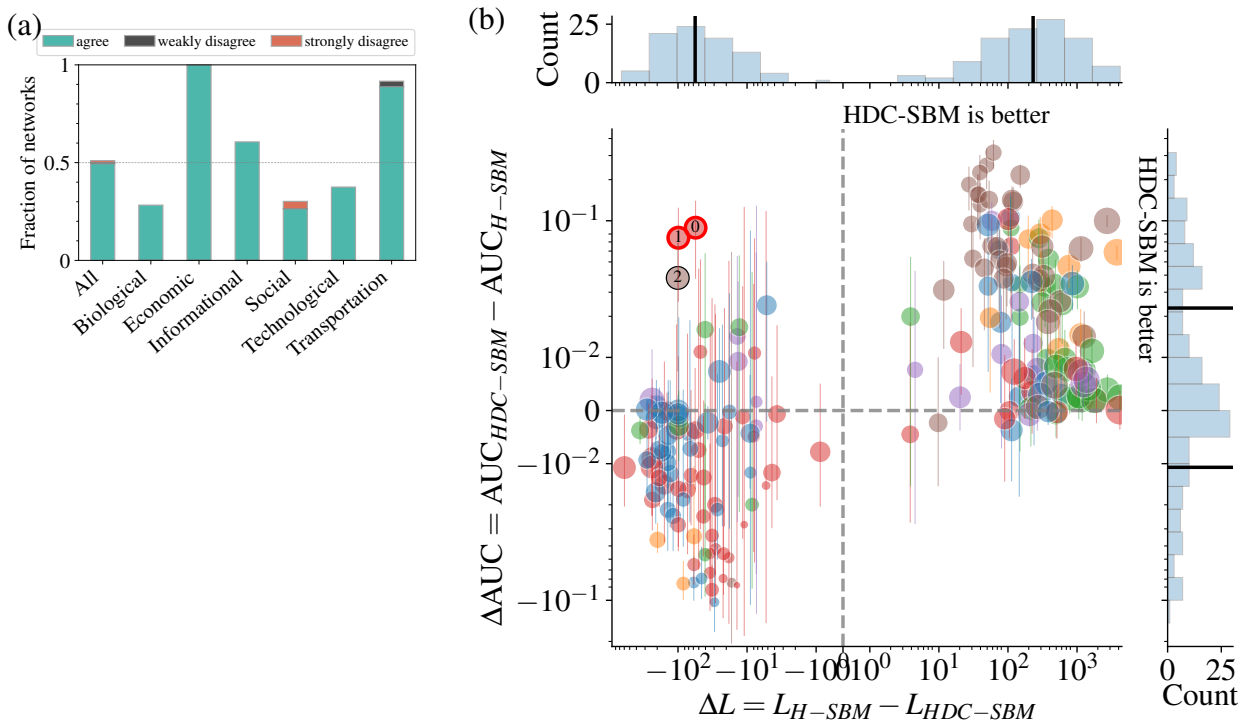


Figure 5.10: Panels (a) and (b) have the same explanation as in Fig. 5.7, but in this case, we consider the difference in -log-evidence (ΔL), as compression criterion, and the difference in AUC (from posterior averages) (ΔAUC), as predictive criterion.

methodological issue, but rather a sign of model misspecification. For example, consider the social network *copenhagen/calls* [193], which has 536 nodes and an average degree of 2.32. Fig. 5.11 shows summaries of the corresponding posterior distributions of network partitions. Initially, we observe that the posterior distributions of both SBM variants are broad, containing at least 100 modes (panel (a)). In other words, for each model variant, there is no single mode that dominates the corresponding posterior. Furthermore, the partition with minimum description length is not among the more representative modes (panel (d)). Even though the difference in the number of groups between the mode partitions and the partition having minimum description length is small (panel (b)), the difference in description length is significant.

Additionally, Fig. 5.12 allows for a closer inspection to the network partitions of *copenhagen/calls* [193]. In some parts of the network, the mode partitions coincide[5], particularly in the largest connected component. However, discrepancies between partitions are more evident in the smaller components. This behavior might be attributed to the sparsity of the network, i.e., this network is so sparse that neither variant of the SBM can accommodate one or few suitable explanations for it, which suggests that the SBM might be misspecified for this type of networks. A similar explanation follows for the other two networks with disagreements (e.g., see the case of *euroroad* in Figs. C.13 and C.14), which are larger than the previous example, but

[5]For each mode, this partition corresponds to the maximum marginal group membership for each node at lowest level.

have a similar average degree (i.e., no larger than 3).

Before concluding this section, it should be noted that, agreements between evidence and AUC (from posterior averages) occur more frequently than between their point estimate versions, even when stratified by domains. The main reason is that, while for some cases in which the AUC (point estimate) cannot provide a decision, the AUC (from posterior averages) can provide one, and this decision agrees with the decision based on model evidence. This highlights the importance of considering alternative partitions when modelling network data.

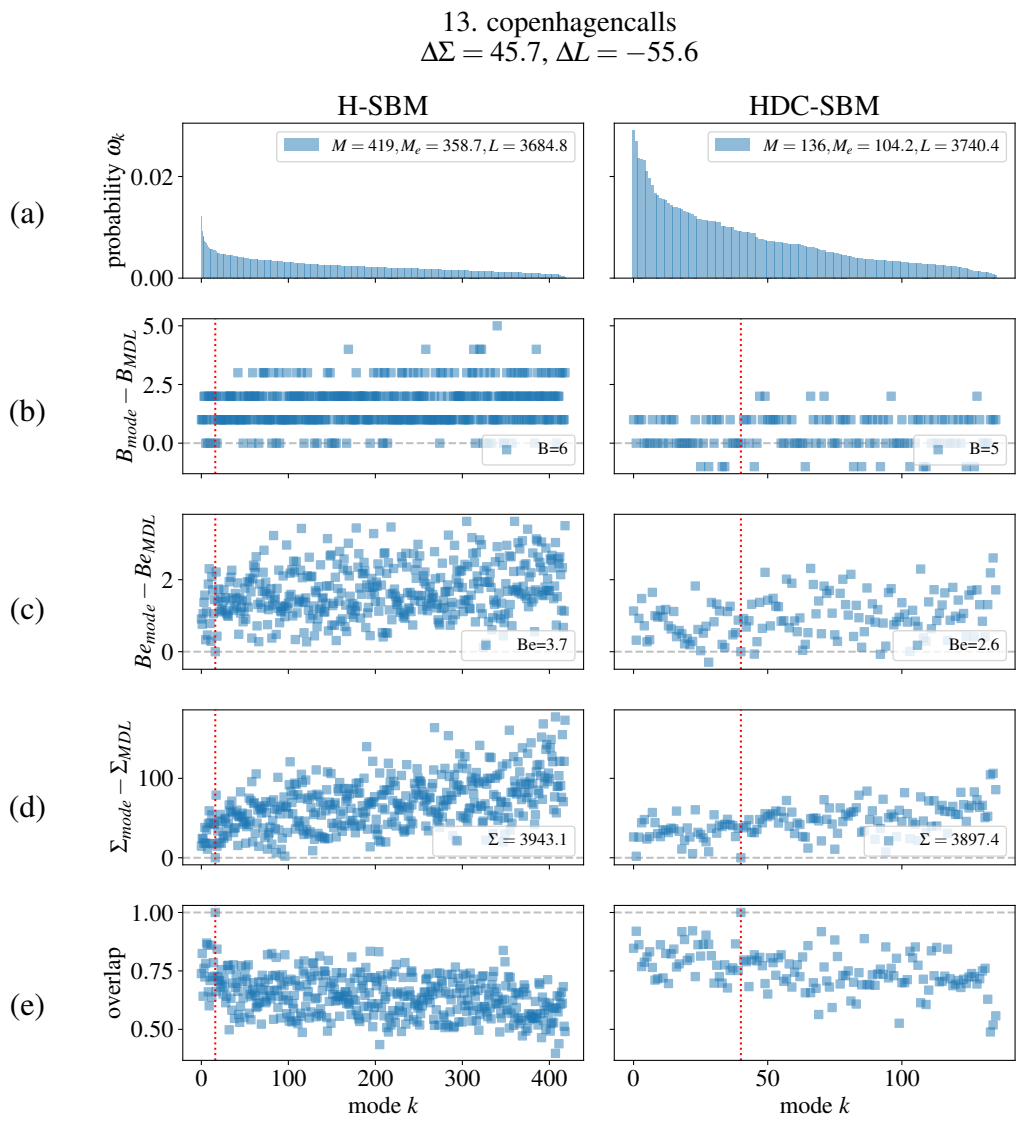


Figure 5.11: Summaries of the modes of the posterior distribution of node partitions of *copenhagen/calls* network for H-SBM (left) and HDC-SBM (right). (a) Mode fractions ω_k. In the legend, we include the number of modes M, the effective number of modes M_e (see footnote 3) and the negative log-evidence L. (b) Difference in the number of groups B corresponding to partition modes and the MDL partition. (c) Difference in the effective number of groups B_e corresponding to partition modes and the MDL partition. (d) Difference in description length Σ corresponding to partition modes and the MDL partition. (e) Overlap between partition modes and the MDL partition. For panels (b) to (e), the vertical red line indicates the mode to which the MDL fit belongs. The legend indicates the summary corresponding to the MDL fit.

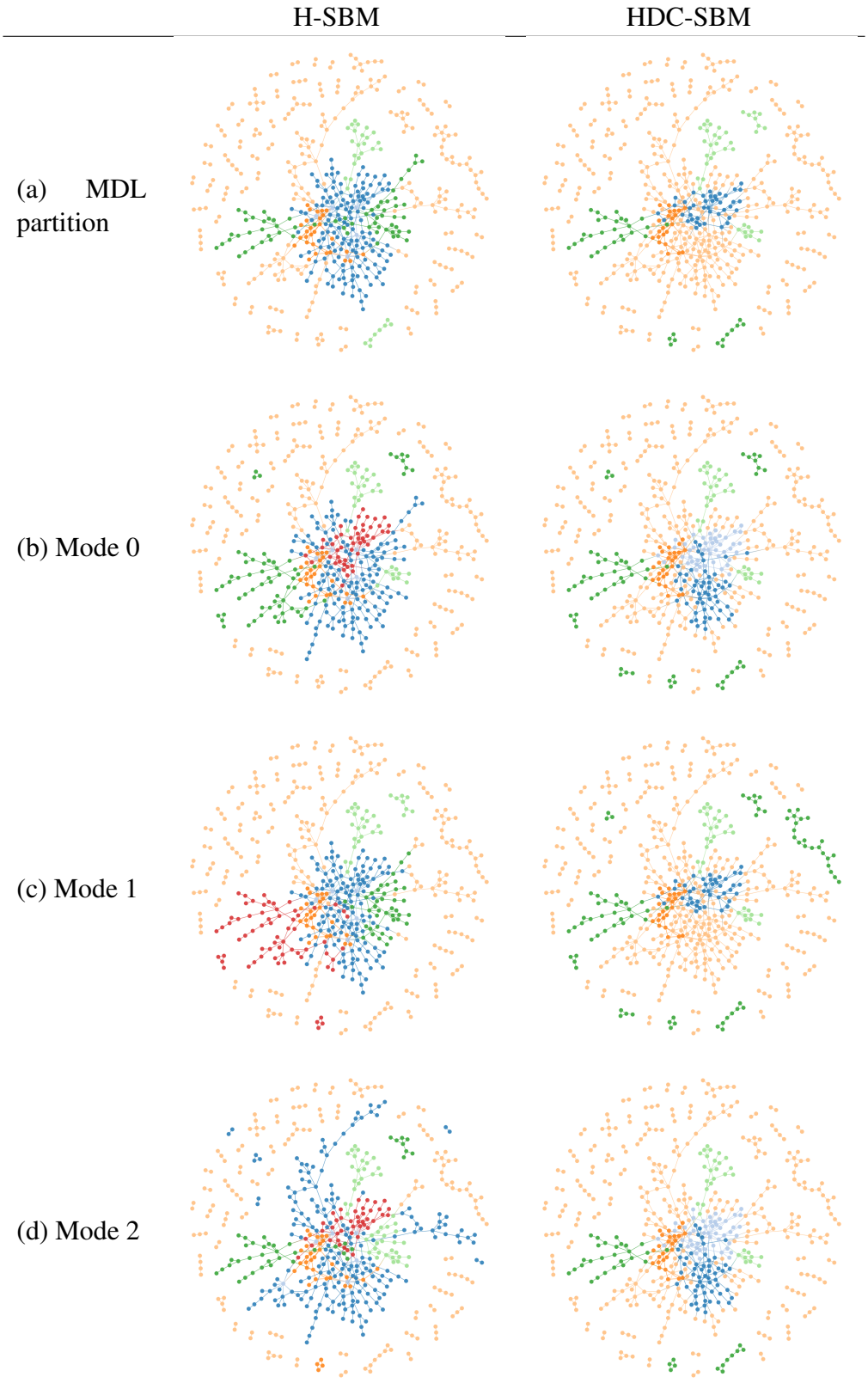


Figure 5.12: Several partitions for *copenhagen/calls* (a network of social interactions among university students within the Copenhagen Networks Study [193]) obtained with the H-SBM and HDC-SBM from (a) minimizing the description length, (b-d) fitting a mixture model to characterize the posterior distribution of node partitions and obtaining its modes. The three most likely modes are shown here.

5.3 Concluding Remarks

In this chapter, we revisited and expanded upon the work of Vallès-Català *et al.* (2018) [2] to study the consistency of compression and predictive criteria when selecting among SBM variants. While we incorporated some advancements from the literature on SBMs, our results overall align with the conclusions of that work, namely that both criteria agree in practice. However, our analysis has revealed further nuances.

For synthetic networks, we find consistency between model selection criteria, i.e., the most compressive model is also the most predictive. For empirical networks, consistency is also common, though there are a few exceptions. Although agreements between model selection approaches are quite frequent, we cannot claim that both approaches are equally reliable. The AUC might be incapable of preferring a single model, as there are many cases where the predictiveness of competing models is statistically equivalent. In contrast, both the description length or evidence tells us which model compresses more the data, and provides a degree of confidence for ruling out the alternative model. In this sense, the compression approach to model selection is more reliable.

This does not mean that predictive criteria should be abandoned in network modeling. Rather, it highlights the need to understand their capabilities for model selection. If our goal is to select the model that best explains the data, compression criteria should be preferred. If our goal is to predict, we need not limit ourselves to the set of competing models and choose one among them. Instead, combining models, for example via *stacking* [195], may yield a better predictive approach.

Finally, we note that this analysis is limited to two variants of the SBM. We expect that future research explores other variants or families of generative models. Furthermore, the standard error of the AUC used in this work relies on distributional assumptions of normality. A more suitable approximation might be needed to accurately quantify the uncertainty in AUC and determine which magnitudes of difference should considered "significant". Such approximation can be particularly valuable in large scale studies comparing the predictive power of various algorithms, helping to prevent misleading claims of superiority of a method based on spurious differences in predictive performance.

Chapter 6

Conclusion

In this dissertation, we aimed to deepen our understanding of the behavior, capabilities, and limitations of Stochastic Block Models (SBMs) as approximations of true underlying models of real-world networks. To this end, we conducted large scale studies of SBM fits to hundreds of empirical networks to uncover systematic patterns in SBM performance. We employed two complementary approaches to assess the quality of the model, namely model checking and model comparison. In a Bayesian framework, these approaches are used in parallel to obtain reliable inferences and valuable insights into model behavior. On one hand, model checking reveals aspects of the data that the model may not accurately describe, thereby evaluating the quality of fit of the model to the data. On the other, model comparison helps to identify potential problems of overfitting and underfitting among competing models. It is important to note that such scrutiny is feasible in inferential approaches to community detection, because their assumptions are made explicit. In descriptive approaches, assumptions are implicit, making it difficult to test their validity directly. In the rest of this chapter, we summarize our results and suggest avenues for future research.

In Chapter 3, we observed that the SBM accurately captures structural descriptors for most networks in our corpus. The largest discrepancies between the SBM and the data typically occur in networks with large diameter and slow mixing of random walks, which are often embedded in a low-dimensional space. Interestingly, we also found that for the other types of networks, including many networks with an abundance of triangles, the SBM shows a fairly good agreement, contrary to common assumptions about the capabilities of the model. We also identified a minimal set of network descriptors that can predict the quality of fit of the SBM, with the most important predictors being the network diameter and characteristic time of a random walk, followed by clustering as a secondary feature. This result points to potentially beneficial directions for model improvement.

The conclusions of Chapter 3 are constrained by the set-up of the assessment, and consequently, the analysis can be extended in several ways. First, it might be beneficial to consider a larger

set of descriptors than the ones considered here. This would provide further insights into the model strengths and weaknesses. Additionally, we only evaluated one variant of the SBM, namely the degree corrected SBM (DC-SBM) with hierarchical priors. Future analysis could include other generative models within or beyond the SBM family. Of particular interest are the non-degree corrected versions and non-hierarchical versions of SBMs, since this would not only offer insights on the quality of fit of such models, but also may inform about the impact of degree-correction and priors on the model fit to data. Finally, we have considered simple networks as inputs of the model. Extending the analysis to directed networks, multilayer networks, or other types of networks could also be beneficial.

In Chapter 4, we observed that the SBM can accurately estimate relevant features of empirical networks whose measurements are noisy, even when we have only one network measurement at our disposal. Furthermore, we observed that in most cases, the reconstruction procedure is beneficial in terms of the error magnitude, compared with taking the data and not doing reconstruction. The exceptions primarily include networks with large diameter and slow-mixing random walks, being most of them transportation networks, which have a low-dimensional spatial embedding, where the "small-world" property is not fulfilled. We also illustrated how including more measurements of the network benefits the reconstruction accuracy. We observed that reconstruction errors using one measurement are one order of magnitude larger than when using more measurements. However, there seems to be diminishing improvements as we gradually increase the number of measurements. It might be useful to study how fast this improvement changes, in order to achieve perfect or close to perfect reconstruction.

Furthermore, this analysis can be extended in several ways. First, we considered uniform rates of noise. Taking into account other models of noise, e.g., having larger error rates around hubs, might provide insights on how we can handle systematic (or correlated) errors. We also focused on two levels of noise, which only gives a partial picture of the capabilities and limitations of the reconstruction framework. Thus, it would be interesting to consider larger levels of noise not only to complete such picture, but also to study the sensitivity of descriptors to noise and the difficulty of reconstructing structural descriptors. Another possibility consists on analyzing a larger set of descriptors that could reveal more relevant dimensions for the assessment, e.g., those related with dynamics happening on top of networks. Finally, although we used the SBM as a prior for network structure, our framework is flexible enough to allow for other generative models. We hope to motivate further research in this direction and understand how other models may improve or worsen the estimation of descriptors of interest (e.g., the diameter).

In Chapter 5, we observed consistency between model selection criteria under a community detection task. Specifically, we found that the most compressive model is often the most predictive when comparing two SBM variants fitted to both synthetic and empirical networks. However, there are few exceptions in empirical networks for which the SBM may be misspecified. Furthermore, we also observed differences in the degree of reliability of these criteria. In

many cases, both SBM variants were equally predictive in terms of their AUC, and therefore, it was not possible to select a model with this approach. Contrarily, both the description length and evidence provided a decision on the most compressive model along with a degree of confidence for ruling out the alternative model. In that sense, the compression criterion is more reliable for model selection within a community detection task.

Our analysis focused on two variants of the SBM, but future research can extend this by incorporating other generative models of network structure. It would be interesting to study how preferred models change and whether the agreements between model selection criteria persist or not, and why. Additionally, the predictive criterion relied on the AUC, however other indices such as accuracy, precision, and recall might yield different results. Furthermore, we also advised caution in using discrepancies between model and data from posterior predictive checks as a model selection criterion. Although we expect this criterion to be misleading, it remains to be determined the magnitude of such issue, and the characteristics of the networks in which it occurs.

Besides what has been mentioned, there are other avenues of research that could also improve our understanding of SBMs. One of them is to evaluate their suitability in other inference frameworks. Previously, we suggested to extend our evaluation of reconstruction performance to study the accuracy of estimates of dynamical aspects occurring on top of the reconstructed network. In such case the input was a noisy network. Differently, there are situations in which only indirect measurements of the network are available. Typically, the observations correspond to functional behavior, i.e., to a dynamical process taking place on such network. Therefore, the network of interactions must be inferred. Examples include inferring cortical neuronal network structure from neuronal activity [196], gene regulatory networks from expression assays [197], infection propagation networks from epidemic data [198], and financial networks from the activity of financial institutions [199]. Peixoto (2019) [200], and more recently Peixoto (2024) [201, 202], proposed Bayesian approaches to infer such connections along with the community structure of the network. In this sense, it would be interesting to determine how accurate the estimates of summaries of network structure are, in which instances the reconstruction procedure is easier or more challenging, and why.

Finally, an important modelling aspect which was not covered here, but offers several prospects, is the study of the priors. Understanding how changes in priors affect model performance and the practical implications for different types of data could offer valuable insights. For certain networks, changing the prior might significantly impact the inferred partition, while for others, it may have little to no impact. Additionally, it has been shown that using a hierarchy of priors can help addressing the underfitting issue. However, it remains unclear what is the depth of the hierarchy of (hyper)priors needed to obtain reliable inferences in empirical data. These analyses should also study the computational costs involved in using different priors, and weigh them against potential inferential advantages.

Bibliography

[1] F. Vaca-Ramírez and T. P. Peixoto, "Systematic assessment of the quality of fit of the stochastic block model for empirical networks," *Physical Review E*, vol. 105, no. 5, p. 054311, 2022.

[2] T. Vallès-Català, T. P. Peixoto, M. Sales-Pardo, and R. Guimerà, "Consistencies and inconsistencies between model selection and link prediction in networks," *Physical Review E*, vol. 97, p. 062316, June 2018.

[3] L. A. Adamic and N. Glance, "The political blogosphere and the 2004 us election: divided they blog," in *Proceedings of the 3rd international workshop on Link discovery*, pp. 36–43, ACM, 2005.

[4] S. Fortunato, "Community detection in graphs," *Physics reports*, vol. 486, no. 3-5, pp. 75–174, 2010.

[5] P. Holme, M. Huss, and H. Jeong, "Subnetwork hierarchies of biochemical pathways," *Bioinformatics*, vol. 19, no. 4, pp. 532–538, 2003.

[6] R. Guimera and L. A. N. Amaral, "Functional cartography of complex metabolic networks," *nature*, vol. 433, no. 7028, p. 895, 2005.

[7] C. Cortes, D. Pregibon, and C. Volinsky, "Communities of interest," in *International Symposium on Intelligent Data Analysis*, pp. 105–114, Springer, 2001.

[8] L. S. Haggerty, P.-A. Jachiet, W. P. Hanage, D. A. Fitzpatrick, P. Lopez, M. J. O'Connell, D. Pisani, M. Wilkinson, E. Bapteste, and J. O. McInerney, "A pluralistic account of homology: adapting the models to the data," *Molecular biology and evolution*, vol. 31, no. 3, pp. 501–516, 2013.

[9] S. Fortunato and M. E. Newman, "20 years of network community detection," *Nature Physics*, vol. 18, no. 8, pp. 848–850, 2022.

[10] M. Schaub, J. Delvenne, M. Rosvall, and R. Lambiotte, "The many facets of community detection in complex networks. vol. 2, issue 1," *Appl Netw Sci*, p. 4, 2017.

[11] T. P. Peixoto, *Descriptive vs. inferential community detection in networks: Pitfalls, myths and half-truths*. Cambridge University Press, 2023.

[12] M. E. Newman and M. Girvan, "Finding and evaluating community structure in networks," *Physical review E*, vol. 69, no. 2, p. 026113, 2004.

[13] R. Guimerà, M. Sales-Pardo, and L. A. N. Amaral, "Modularity from fluctuations in random graphs and complex networks," *Physical Review E*, vol. 70, p. 025101, Aug. 2004.

[14] C. McDiarmid and F. Skerman, "Modularity in random regular graphs and lattices," *Electronic Notes in Discrete Mathematics*, vol. 43, pp. 431–437, Sept. 2013.

[15] J. P. Bagrow, "Communities and bottlenecks: Trees and treelike networks have high modularity," *Physical Review E*, vol. 85, p. 066118, June 2012.

[16] S. Fortunato and M. Barthélemy, "Resolution limit in community detection," *Proceedings of the National Academy of Sciences*, vol. 104, pp. 36–41, Feb. 2007.

[17] B. H. Good, Y.-A. de Montjoye, and A. Clauset, "Performance of modularity maximization in practical contexts," *Physical Review E*, vol. 81, p. 046106, Apr. 2010.

[18] P. W. Holland, K. B. Laskey, and S. Leinhardt, "Stochastic blockmodels: First steps," *Social Networks*, vol. 5, pp. 109–137, June 1983.

[19] S. E. Fienberg, M. M. Meyer, and S. S. Wasserman, "Statistical analysis of multiple sociometric relations," *Journal of the american Statistical association*, vol. 80, no. 389, pp. 51–67, 1985.

[20] K. Faust and S. Wasserman, "Blockmodels: Interpretation and evaluation," *Social networks*, vol. 14, no. 1-2, pp. 5–61, 1992.

[21] C. J. Anderson, S. Wasserman, and K. Faust, "Building stochastic blockmodels," *Social networks*, vol. 14, no. 1-2, pp. 137–161, 1992.

[22] T. P. Peixoto, "Bayesian Stochastic Blockmodeling," in *Advances in Network Clustering and Blockmodeling*, pp. 289–332, John Wiley & Sons, Ltd, 2019.

[23] T. Funke and T. Becker, "Stochastic block models: A comparison of variants and inference methods," *PloS one*, vol. 14, no. 4, p. e0215296, 2019.

[24] P. D. Grünwald, *The Minimum Description Length Principle*. The MIT Press, Mar. 2007.

[25] T. P. Peixoto, "Hierarchical Block Structures and High-Resolution Model Selection in Large Networks," *Physical Review X*, vol. 4, p. 011047, Mar. 2014.

[26] A. Gelman, A. Vehtari, D. Simpson, C. C. Margossian, B. Carpenter, Y. Yao, L. Kennedy, J. Gabry, P.-C. Bürkner, and M. Modrák, "Bayesian workflow," *arXiv preprint arXiv:2011.01808*, 2020.

[27] A. Clauset and C. Moore, "Accuracy and scaling phenomena in internet mapping," *Physical Review Letters*, vol. 94, no. 1, p. 018701, 2005.

[28] A. Lakhina, J. W. Byers, M. Crovella, and P. Xie, "Sampling biases in ip topology measurements," in *IEEE INFOCOM 2003. Twenty-second Annual Joint Conference of the IEEE Computer and Communications Societies (IEEE Cat. No. 03CH37428)*, vol. 1, pp. 332–341, IEEE, 2003.

[29] K. G. Leyba, J. J. Daymude, J.-G. Young, M. Newman, J. Rexford, and S. Forrest, "Cutting through the noise to infer autonomous system topology," in *IEEE INFOCOM 2022-IEEE Conference on Computer Communications*, pp. 1609–1618, IEEE, 2022.

[30] P. Killworth and H. Bernard, "Informant accuracy in social network data," *Human organization*, vol. 35, no. 3, pp. 269–286, 1976.

[31] P. V. Marsden, "Network data and measurement," *Annual review of sociology*, vol. 16, no. 1, pp. 435–463, 1990.

[32] C. T. Butts, "Network inference, error, and informant (in) accuracy: a bayesian approach," *social networks*, vol. 25, no. 2, pp. 103–140, 2003.

[33] E. Sprinzak, S. Sattath, and H. Margalit, "How reliable are experimental protein–protein interaction data?," *Journal of molecular biology*, vol. 327, no. 5, pp. 919–923, 2003.

[34] N. J. Krogan, G. Cagney, H. Yu, G. Zhong, X. Guo, A. Ignatchenko, J. Li, S. Pu, N. Datta, A. P. Tikuisis, *et al.*, "Global landscape of protein complexes in the yeast saccharomyces cerevisiae," *Nature*, vol. 440, no. 7084, pp. 637–643, 2006.

[35] S. J. Wodak, S. Pu, J. Vlasblom, and B. Seéraphin, "Challenges and rewards of interaction proteomics," *Molecular & Cellular Proteomics*, vol. 8, no. 1, pp. 3–18, 2009.

[36] L. Peel, T. P. Peixoto, and M. De Domenico, "Statistical inference links data and theory in network science," *Nature Communications*, vol. 13, p. 6794, Nov. 2022.

[37] G. Kossinets, "Effects of missing data in social networks," *Social networks*, vol. 28, no. 3, pp. 247–268, 2006.

[38] N. Erman and L. Todorovski, "The effects of measurement error in case of scientific network analysis," *Scientometrics*, vol. 104, pp. 453–473, 2015.

[39] A. Clauset, C. Moore, and M. E. Newman, "Hierarchical structure and the prediction of missing links in networks," *Nature*, vol. 453, no. 7191, pp. 98–101, 2008.

[40] M. Kim and J. Leskovec, "The network completion problem: Inferring missing nodes and edges in networks," in *Proceedings of the 2011 SIAM international conference on data mining*, pp. 47–58, SIAM, 2011.

[41] R. Guimerà and M. Sales-Pardo, "Missing and spurious interactions and the reconstruction of complex networks," *Proceedings of the National Academy of Sciences*, vol. 106, pp. 22073 –22078, Dec. 2009.

[42] A. Ghasemian, H. Hosseinmardi, and A. Clauset, "Evaluating Overfit and Underfit in Models of Network Community Structure," *IEEE Transactions on Knowledge and Data Engineering*, pp. 1–1, 2019.

[43] T. P. Peixoto, "Reconstructing Networks with Unknown and Heterogeneous Errors," *Physical Review X*, vol. 8, p. 041011, Oct. 2018.

[44] J.-G. Young, G. T. Cantwell, and M. Newman, "Bayesian inference of network structure from unreliable data," *Journal of Complex Networks*, vol. 8, no. 6, p. cnaa046, 2020.

[45] L. Peel, D. B. Larremore, and A. Clauset, "The ground truth about metadata and community detection in networks," *Science advances*, vol. 3, no. 5, p. e1602548, 2017.

[46] D. Hric, R. K. Darst, and S. Fortunato, "Community detection in networks: Structural communities versus ground truth," *Physical Review E*, vol. 90, p. 062805, Dec. 2014.

[47] A. Lancichinetti, S. Fortunato, and F. Radicchi, "Benchmark graphs for testing community detection algorithms," *Physical Review E*, vol. 78, p. 046110, Oct. 2008.

[48] J. Moody, "Peer influence groups: identifying dense clusters in large networks," *Social Networks*, vol. 23, pp. 261–283, 2001.

[49] T. P. Peixoto, "Disentangling homophily, community structure, and triadic closure in networks," *Physical Review X*, vol. 12, no. 1, p. 011004, 2022.

[50] A. Ghasemian, H. Hosseinmardi, A. Galstyan, E. M. Airoldi, and A. Clauset, "Stacking models for nearly optimal link prediction in complex networks," *Proceedings of the National Academy of Sciences*, vol. 117, pp. 23393–23400, Sept. 2020.

[51] M. Contisciani, E. A. Power, and C. De Bacco, "Community detection with node attributes in multilayer networks," *Scientific reports*, vol. 10, no. 1, p. 15736, 2020.

[52] T. P. Peixoto, "Revealing Consensus and Dissensus between Network Partitions," *Physical Review X*, vol. 11, p. 021003, Apr. 2021. Publisher: American Physical Society.

[53] T. P. Peixoto, "Merge-split Markov chain Monte Carlo for community detection," *Physical Review E*, vol. 102, p. 012305, July 2020.

[54] T. P. Peixoto, "The Netzschleuder network catalogue and repository.," 2020. Accessible at https://networks.skewed.de.

[55] M. Newman, *Networks*. Oxford university press, 2018.

[56] E. Kolaczyk, "Statistical analysis of network data: Methods and models," *Springer Series In Statistics*, p. 386, 2009.

[57] T. P. Peixoto, "Nonparametric Bayesian inference of the microcanonical stochastic block model," *Physical Review E*, vol. 95, p. 012317, Jan. 2017.

[58] J. L. Moreno, *Application of the group method to classification*. National committee on prisons and prison labor, 1932.

[59] J. L. Moreno and H. H. Jennings, "Statistics of social configurations," *Sociometry*, pp. 342–374, 1938.

[60] H. Jeong, S. P. Mason, A.-L. Barabási, and Z. N. Oltvai, "Lethality and centrality in protein networks," *Nature*, vol. 411, pp. 41–42, May 2001. Number: 6833 Publisher: Nature Publishing Group.

[61] E. Bullmore and O. Sporns, "The economy of brain network organization," *Nature Reviews Neuroscience*, vol. 13, pp. 336–349, Apr. 2012.

[62] N. D. Martinez, "Artifacts or Attributes? Effects of Resolution on the Little Rock Lake Food Web," *Ecological Monographs*, vol. 61, no. 4, pp. 367–392, 1991.

[63] S. González-Bailón, J. Borge-Holthoefer, A. Rivero, and Y. Moreno, "The dynamics of protest recruitment through an online network," *Scientific reports*, vol. 1, no. 1, pp. 1–7, 2011.

[64] B. Zhang, R. Liu, D. Massey, and L. Zhang, "Collecting the internet as-level topology," *ACM SIGCOMM Computer Communication Review*, vol. 35, no. 1, pp. 53–61, 2005.

[65] G. Boeing, "U.S. Street Network Shapefiles, Node/Edge Lists, and GraphML Files," 2017.

[66] G. Boeing, "A multi-scale analysis of 27,000 urban street networks: Every us city, town, urbanized area, and zillow neighborhood," *Environment and Planning B: Urban Analytics and City Science*, p. 2399808318784595, 2018.

[67] M. Barthélemy, "Spatial networks," *Physics Reports*, vol. 499, pp. 1–101, Feb. 2011.

[68] M. Kivelä, A. Arenas, M. Barthelemy, J. P. Gleeson, Y. Moreno, and M. A. Porter, "Multilayer networks," *Journal of Complex Networks*, vol. 2, pp. 203–271, Jan. 2014.

[69] S. Boccaletti, G. Bianconi, R. Criado, C. I. del Genio, J. Gómez-Gardeñes, M. Romance, I. Sendiña-Nadal, Z. Wang, and M. Zanin, "The structure and dynamics of multilayer networks," *Physics Reports*, vol. 544, pp. 1–122, Nov. 2014.

[70] C. C. Hyland, Y. Tao, L. Azizi, M. Gerlach, T. P. Peixoto, and E. G. Altmann, "Multilayer networks for text analysis with multiple data types," *EPJ Data Science*, vol. 10, no. 1, p. 33, 2021.

[71] M. S. Granovetter, "The Strength of Weak Ties," *American Journal of Sociology*, vol. 78, pp. 1360–1380, May 1973.

[72] G. Bianconi, R. K. Darst, J. Iacovacci, and S. Fortunato, "Triadic closure as a basic generating mechanism of communities in complex networks," *Physical Review E*, vol. 90, p. 042806, Oct. 2014. Publisher: American Physical Society.

[73] M. McPherson, L. Smith-Lovin, and J. M. Cook, "Birds of a Feather: Homophily in Social Networks," *Annual Review of Sociology*, vol. 27, no. 1, pp. 415–444, 2001.

[74] J. Moody, "Race, School Integration, and Friendship Segregation in America," *American Journal of Sociology*, vol. 107, pp. 679–716, Nov. 2001. Publisher: The University of Chicago Press.

[75] K. M. Harris, C. T. Halpern, E. Whitsel, J. Hussey, J. Tabor, P. Entzel, and J. R. Udry, "The national longitudinal study of adolescent to adult health: Research design," *See http://www. cpc. unc. edu/projects/addhealth/design (accessed 9 April 2015)*, 2009.

[76] P. S. Bearman, J. Moody, and K. Stovel, "The structure of adolescent romantic and sexual networks," *Handbook of Applied System Science*, p. 164, 2016.

[77] P. Holme and J. Saramäki, "Temporal networks," *Physics Reports*, vol. 519, pp. 97–125, Oct. 2012.

[78] P. J. Mucha, T. Richardson, K. Macon, M. A. Porter, and J.-P. Onnela, "Community Structure in Time-Dependent, Multiscale, and Multiplex Networks," *Science*, vol. 328, pp. 876–878, May 2010.

[79] T. P. Peixoto and M. Rosvall, "Modelling sequences and temporal networks with dynamic community structures," *Nature Communications*, vol. 8, p. 582, Sept. 2017.

[80] M. Cha, H. Haddadi, F. Benevenuto, and K. Gummadi, "Measuring user influence in twitter: The million follower fallacy," in *Proceedings of the international AAAI conference on web and social media*, vol. 4, pp. 10–17, 2010.

[81] J. Yang and J. Leskovec, "Defining and evaluating network communities based on ground-truth," in *Proceedings of the ACM SIGKDD Workshop on Mining Data Semantics*, MDS '12, (New York, NY, USA), pp. 3:1–3:8, ACM, 2012.

[82] H. A. Simon, "On a class of skew distribution functions," *Biometrika*, vol. 42, no. 3/4, pp. 425–440, 1955.

[83] D. J. d. S. Price, "Networks of Scientific Papers," *Science*, vol. 149, pp. 510–515, July 1965.

[84] A.-L. Barabási and R. Albert, "Emergence of Scaling in Random Networks," *Science*, vol. 286, pp. 509–512, Oct. 1999.

[85] D. J. Watts and S. H. Strogatz, "Collective dynamics of 'small-world' networks," *Nature*, vol. 393, pp. 409–10, 1998.

[86] P. D. Hoff, A. E. Raftery, and M. S. Handcock, "Latent space approaches to social network analysis," *Journal of the american Statistical association*, vol. 97, no. 460, pp. 1090–1098, 2002.

[87] C. Matias and S. Robin, "Modeling heterogeneity in random graphs through latent space models: a selective review," *ESAIM: Proceedings and Surveys*, vol. 47, pp. 55–74, 2014.

[88] E. McFowland III and C. R. Shalizi, "Estimating Causal Peer Influence in Homophilous Social Networks by Inferring Latent Locations," *Journal of the American Statistical Association*, vol. 0, pp. 1–27, July 2021. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/01621459.2021.1953506.

[89] T. A. Snijders, P. E. Pattison, G. L. Robins, and M. S. Handcock, "New specifications for exponential random graph models," *Sociological methodology*, vol. 36, no. 1, pp. 99–153, 2006.

[90] G. Robins, P. Pattison, Y. Kalish, and D. Lusher, "An introduction to exponential random graph (p*) models for social networks," *Social networks*, vol. 29, no. 2, pp. 173–191, 2007.

[91] S. Chatterjee and P. Diaconis, "Estimating and Understanding Exponential Random Graph Models," *1102.2650*, Feb. 2011.

[92] N. Kashtan, S. Itzkovitz, R. Milo, and U. Alon, "Efficient sampling algorithm for estimating subgraph concentrations and detecting network motifs," *Bioinformatics*, vol. 20, no. 11, pp. 1746–1758, 2004.

[93] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon, "Network Motifs: Simple Building Blocks of Complex Networks," *Science*, vol. 298, pp. 824–827, Oct. 2002.

[94] O. Frank and T. Snijders, "Estimating the size of hidden populations using snowball sampling," *Journal of Official Statistics-Stockholm-*, vol. 10, pp. 53–53, 1994.

[95] R. Solomonoff and A. Rapoport, "Connectivity of random nets," *The bulletin of mathematical biophysics*, vol. 13, pp. 107–117, 1951.

[96] E. N. Gilbert, "Random Graphs," *The Annals of Mathematical Statistics*, vol. 30, pp. 1141–1144, Dec. 1959.

[97] P. Erdős and A. Rényi, "On random graphs, I," *Publicationes Mathematicae (Debrecen)*, vol. 6, pp. 290–297, 1959.

[98] P. Erdős, A. Rényi, *et al.*, "On the evolution of random graphs," *Publ. math. inst. hung. acad. sci*, vol. 5, no. 1, pp. 17–60, 1960.

[99] P. Erdős and A. Rényi, "On the strength of connectedness of a random graph," *Acta Mathematica Hungarica*, vol. 12, no. 1, pp. 261–267, 1961.

[100] B. Bollobás, "A Probabilistic Proof of an Asymptotic Formula for the Number of Labelled Regular Graphs," *European Journal of Combinatorics*, vol. 1, pp. 311–316, Dec. 1980.

[101] B. K. Fosdick, D. B. Larremore, J. Nishimura, and J. Ugander, "Configuring Random Graph Models with Fixed Degree Sequences," *arXiv:1608.00607 [physics, q-bio, stat]*, Aug. 2016. arXiv: 1608.00607.

[102] T. A. B. Snijders and K. Nowicki, "Estimation and Prediction for Stochastic Blockmodels for Graphs with Latent Block Structure," *Journal of Classification*, vol. 14, pp. 75–100, Jan. 1997.

[103] K. Nowicki and T. A. B. Snijders, "Estimation and Prediction for Stochastic Blockstructures," *Journal of the American Statistical Association*, vol. 96, pp. 1077–1087, Sept. 2001.

[104] B. Söderberg, "General formalism for inhomogeneous random graphs," *Physical Review E*, vol. 66, p. 066121, Dec. 2002.

[105] B. Bollobás, S. Janson, and O. Riordan, "The phase transition in inhomogeneous random graphs," *Random Structures & Algorithms*, vol. 31, pp. 3–122, Aug. 2007.

[106] A. Condon and R. M. Karp, "Algorithms for graph partitioning on the planted partition model," *Random Structures & Algorithms*, vol. 18, no. 2, pp. 116–140, 2001. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/1098-2418%28200103%2918%3A2%3C116%3A%3AAID-RSA1001%3E3.0.CO%3B2-2.

[107] M. Boguñá and R. Pastor-Satorras, "Class of correlated random networks with hidden variables," *Physical Review E*, vol. 68, no. 3, p. 036112, 2003.

[108] J.-J. Daudin, F. Picard, and S. Robin, "A mixture model for random graphs," *Statistics and Computing*, vol. 18, pp. 173–183, June 2008.

[109] G. Bianconi, P. Pin, and M. Marsili, "Assessing the relevance of node features for network structure," *Proceedings of the National Academy of Sciences*, vol. 106, pp. 11433 –11438, July 2009.

[110] B. Karrer and M. E. J. Newman, "Stochastic blockmodels and community structure in networks," *Physical Review E*, vol. 83, p. 016107, Jan. 2011.

[111] T. P. Peixoto, "Efficient Monte Carlo and greedy heuristic for the inference of stochastic block models," *Physical Review E*, vol. 89, p. 012804, Jan. 2014.

[112] M. P. Young, "The organization of neural systems in the primate cerebral cortex," *Proceedings of the Royal Society of London. Series B: Biological Sciences*, vol. 252, no. 1333, pp. 13–18, 1993.

[113] X. Yan, Y. Zhu, J.-B. Rouquier, and C. Moore, "Active Learning for Hidden Attributes in Networks," *arXiv:1005.0794*, May 2010.

[114] E. Côme and P. Latouche, "Model selection and clustering in stochastic block models based on the exact integrated complete data likelihood," *Statistical Modelling*, vol. 15, pp. 564–589, Dec. 2015.

[115] M. E. J. Newman and G. Reinert, "Estimating the Number of Communities in a Network," *Physical Review Letters*, vol. 117, p. 078301, Aug. 2016.

[116] T. P. Peixoto, "Parsimonious Module Inference in Large Networks," *Physical Review Letters*, vol. 110, p. 148701, Apr. 2013.

[117] A. Clauset, C. R. Shalizi, and M. E. J. Newman, "Power-Law Distributions in Empirical Data," *SIAM Review*, vol. 51, pp. 661–703, Jan. 2009.

[118] G. E. Andrews, *The Theory of Partitions*. Cambridge: Cambridge University Press, 1984.

[119] T. P. Peixoto, "Entropy of stochastic blockmodel ensembles," *Physical Review E*, vol. 85, p. 056122, May 2012.

[120] J. Rissanen, "Modeling by shortest data description," *Automatica*, vol. 14, pp. 465–471, Sept. 1978.

[121] C. E. Shannon, "A mathematical theory of communication," *Bell Syst Tech. J*, vol. 27, no. 379, p. 623, 1948.

[122] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, "Equation of State Calculations by Fast Computing Machines," *The Journal of Chemical Physics*, vol. 21, no. 6, p. 1087, 1953.

[123] W. K. Hastings, "Monte Carlo sampling methods using Markov chains and their applications," *Biometrika*, vol. 57, pp. 97 –109, Apr. 1970.

[124] S. Kirkpatrick, C. D. Gelatt Jr, and M. P. Vecchi, "Optimization by simulated annealing," *Science*, vol. 220, no. 4598, p. 671, 1983.

[125] T. P. Peixoto, "The `graph-tool` python library," *figshare*, 2014. Available at https://graph-tool.skewed.de.

[126] K. Xu and A. Hero, "Dynamic Stochastic Blockmodels for Time-Evolving Social Networks," *IEEE Journal of Selected Topics in Signal Processing*, vol. 8, pp. 552–562, Aug. 2014.

[127] T. P. Peixoto, "Inferring the mesoscale structure of layered, edge-valued, and time-varying networks," *Physical Review E*, vol. 92, p. 042807, Oct. 2015.

[128] A. Ghasemian, P. Zhang, A. Clauset, C. Moore, and L. Peel, "Detectability Thresholds and Optimal Algorithms for Community Structure in Dynamic Networks," *Physical Review X*, vol. 6, p. 031005, July 2016.

[129] N. Stanley, S. Shai, D. Taylor, and P. J. Mucha, "Clustering Network Layers with the Strata Multilayer Stochastic Block Model," *IEEE Transactions on Network Science and Engineering*, vol. 3, pp. 95–105, Apr. 2016.

[130] E. A. Power, "Collective ritual and social support networks in rural south india," *Proceedings of the Royal Society B: Biological Sciences*, vol. 285, no. 1879, p. 20180023, 2018.

[131] A.-C. Gavin, M. Bösche, R. Krause, P. Grandi, M. Marzioch, A. Bauer, J. Schultz, J. M. Rick, A.-M. Michon, C.-M. Cruciat, *et al.*, "Functional organization of the yeast proteome by systematic analysis of protein complexes," *Nature*, vol. 415, no. 6868, pp. 141–147, 2002.

[132] E. M. Airoldi, D. M. Blei, S. E. Fienberg, and E. P. Xing, "Mixed Membership Stochastic Blockmodels," *J. Mach. Learn. Res.*, vol. 9, pp. 1981–2014, June 2008.

[133] E. A. Erosheva, *Grade of membership and latent structure models with application to disability survey data*. PhD thesis, Carnegie Mellon University, 2002.

[134] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, Mar. 2003.

[135] L. Fei-Fei and P. Perona, "A bayesian hierarchical model for learning natural scene categories," in *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, vol. 2, pp. 524–531, IEEE, 2005.

[136] E. M. Airoldi, D. M. Blei, E. A. Erosheva, and S. E. Fienberg, *Handbook of mixed membership models and their applications*. CRC press Boca Raton, FL, 2015.

[137] M. E. J. Newman and E. A. Leicht, "Mixture models and exploratory analysis in networks," *Proceedings of the National Academy of Sciences*, vol. 104, pp. 9564 –9569, June 2007.

[138] J. J. Ramasco and M. Mungan, "Inversion method for content-based networks," *Physical Review E—Statistical, Nonlinear, and Soft Matter Physics*, vol. 77, no. 3, p. 036122, 2008.

[139] B. Ball, B. Karrer, and M. E. J. Newman, "Efficient and principled method for detecting communities in networks," *Physical Review E*, vol. 84, no. 3, p. 036103, 2011.

[140] P. K. Gopalan and D. M. Blei, "Efficient discovery of overlapping communities in massive networks," *Proceedings of the National Academy of Sciences*, vol. 110, pp. 14534–14539, Mar. 2013.

[141] J. Yang, J. McAuley, and J. Leskovec, "Detecting cohesive and 2-mode communities indirected and undirected networks," in *Proceedings of the 7th ACM international conference on Web search and data mining*, pp. 323–332, 2014.

[142] C. De Bacco, E. A. Power, D. B. Larremore, and C. Moore, "Community detection, link prediction, and layer interdependence in multilayer networks," *Physical Review E*, vol. 95, p. 042317, Apr. 2017.

[143] C. De Bacco, M. Contisciani, J. Cardoso-Silva, H. Safdari, G. Lima Borges, D. Baptista, T. Sweet, J.-G. Young, J. Koster, C. T. Ross, *et al.*, "Latent network models to account for noisy, multiply reported social network data," *Journal of the Royal Statistical Society Series A: Statistics in Society*, vol. 186, no. 3, pp. 355–375, 2023.

[144] M. Contisciani, F. Battiston, and C. De Bacco, "Inference of hyperedges and overlapping communities in hypergraphs," *Nature communications*, vol. 13, no. 1, p. 7229, 2022.

[145] T. P. Peixoto, "Model Selection and Hypothesis Testing for Large-Scale Network Models with Overlapping Groups," *Physical Review X*, vol. 5, p. 011033, Mar. 2015.

[146] T. P. Peixoto, "Latent Poisson models for networks with heterogeneous density," *Physical Review E*, vol. 102, p. 012309, July 2020.

[147] T. P. Peixoto, "Ordered community detection in directed networks," *Physical Review E*, vol. 106, no. 2, p. 024305, 2022.

[148] G. Palla, I. Derényi, I. Farkas, and T. Vicsek, "Uncovering the overlapping community structure of complex networks in nature and society," *Nature*, vol. 435, pp. 814–818, June 2005.

[149] A. R. Benson, D. F. Gleich, and J. Leskovec, "Higher-order organization of complex networks," *Science*, vol. 353, pp. 163–166, July 2016. Publisher: American Association for the Advancement of Science Section: Reports.

[150] H. Yin, A. R. Benson, J. Leskovec, and D. F. Gleich, "Local Higher-Order Graph Clustering," in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '17, (New York, NY, USA), pp. 555–564, Association for Computing Machinery, Aug. 2017.

[151] A. E. Wegner and S. Olhede, "Atomic subgraphs and the statistical mechanics of networks," *Physical Review E*, vol. 103, p. 042311, Apr. 2021. Publisher: American Physical Society.

[152] J.-G. Young, G. Petri, and T. P. Peixoto, "Hypergraph reconstruction from network data," *Communications Physics*, vol. 4, pp. 1–11, June 2021. Bandiera_abtest: a Cc_license_type: cc_by Cg_type: Nature Research Journals Number: 1 Primary_atype: Research Publisher: Nature Publishing Group Subject_term: Complex networks;Computational science Subject_term_id: complex-networks;computational-science.

[153] M. S. Handcock, G. Robins, T. Snijders, J. Moody, and J. Besag, "Assessing degeneracy in statistical models of social networks," tech. rep., Working paper, 2003.

[154] M. Schweinberger, "Instability, sensitivity, and degeneracy of discrete exponential families," *Journal of the American Statistical Association*, vol. 106, no. 496, pp. 1361–1370, 2011.

[155] R. Fischer, J. C. Leitão, T. P. Peixoto, and E. G. Altmann, "Sampling Motif-Constrained Ensembles of Networks," *Physical Review Letters*, vol. 115, p. 188701, Oct. 2015.

[156] J. Dall and M. Christensen, "Random geometric graphs," *Physical review E*, vol. 66, no. 1, p. 016121, 2002.

[157] M. L. Huson and A. Sen, "Broadcast scheduling algorithms for radio networks," in *Proceedings of MILCOM'95*, vol. 2, pp. 647–651, IEEE, 1995.

[158] A. Fabrikant, E. Koutsoupias, and C. H. Papadimitriou, "Heuristically optimized tradeoffs: A new paradigm for power laws in the internet," in *Automata, Languages and Programming: 29th International Colloquium, ICALP 2002 Málaga, Spain, July 8–13, 2002 Proceedings 29*, pp. 110–122, Springer, 2002.

[159] M. T. Gastner and M. E. J. Newman, "Optimal design of spatial distribution networks," *Physical Review E*, vol. 74, p. 016117, July 2006. Publisher: American Physical Society.

[160] M. T. Gastner and M. E. Newman, "Shape and efficiency in spatial distribution networks," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2006, no. 01, p. P01015, 2006.

[161] A. Bejan and G. Ledezma, "Streets tree networks and urban growth: optimal geometry for quickest access between a finite-size volume and one point," *Physica A: Statistical Mechanics and its Applications*, vol. 255, no. 1-2, pp. 211–217, 1998.

[162] M. Barthélemy and A. Flammini, "Modeling urban street patterns," *Physical review letters*, vol. 100, no. 13, p. 138702, 2008.

[163] M. Barthélemy and A. Flammini, "Co-evolution of density and topology in a simple model of city formation," *Networks and spatial economics*, vol. 9, pp. 401–425, 2009.

[164] T. Courtat, C. Gloaguen, and S. Douady, "Mathematics and Morphogenesis of the City: A Geometrical Approach," *1010.1762*, Oct. 2010.

[165] J. R. Banavar, A. Maritan, and A. Rinaldo, "Size and form in efficient transportation networks," *Nature*, vol. 399, no. 6732, pp. 130–132, 1999.

[166] A. Maritan, F. Colaiori, A. Flammini, M. Cieplak, and J. R. Banavar, "Universality classes of optimal channel networks," *Science*, vol. 272, no. 5264, pp. 984–986, 1996.

[167] P. Crucitti, V. Latora, and S. Porta, "Centrality measures in spatial networks of urban streets," *Physical Review E*, vol. 73, no. 3, p. 036125, 2006.

[168] V. Latora, V. Nicosia, and G. Russo, *Complex networks: principles, methods and applications*. Cambridge University Press, 2017.

[169] B. Fosdick, D. Larremore, J. Nishimura, and J. Ugander, "Configuring Random Graph Models with Fixed Degree Sequences," *SIAM Review*, vol. 60, pp. 315–355, Jan. 2018.

[170] S. C. Olhede and P. J. Wolfe, "Network histograms and universality of blockmodel approximation," *Proceedings of the National Academy of Sciences*, vol. 111, pp. 14722–14727, Oct. 2014.

[171] L. Lovász, "Random walks on graphs: A survey," *Combinatorics, Paul Erdos is Eighty*, vol. 2, no. 1, pp. 1–46, 1993.

[172] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin, *Bayesian data analysis*. CRC press, 2013.

[173] R. McElreath, *Statistical rethinking: A Bayesian course with examples in R and Stan*. CRC press, 2020.

[174] E. Kreyszig, H. Kreyszig, and E. J. Norminton, *Advanced Engineering Mathematics*. Hoboken, NJ: John Wiley & Sons Ltd, 10 ed., Dec. 2010.

[175] A. Decelle, F. Krzakala, C. Moore, and L. Zdeborová, "Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications," *Physical Review E*, vol. 84, p. 066106, Dec. 2011.

[176] C. Gini, "Measurement of Inequality of Incomes," *The Economic Journal*, vol. 31, no. 121, pp. 124–126, 1921. Publisher: [Royal Economic Society, Wiley].

[177] M. E. J. Newman, "Modularity and community structure in networks," *Proceedings of the National Academy of Sciences*, vol. 103, pp. 8577–8582, June 2006.

[178] W. W. Zachary, "An Information Flow Model for Conflict and Fission in Small Groups," *Journal of Anthropological Research*, vol. 33, pp. 452–473, Dec. 1977.

[179] M. E. Newman, "Estimating network structure from unreliable measurements," *Physical Review E*, vol. 98, no. 6, p. 062321, 2018.

[180] "The openflights.org website," 2019.

[181] J. Shao, "Linear Model Selection by Cross-Validation," *Journal of the American Statistical Association*, vol. 88, pp. 486–494, June 1993.

[182] Q. F. Gronau and E.-J. Wagenmakers, "Limitations of bayesian leave-one-out cross-validation for model selection," *Computational brain & behavior*, vol. 2, no. 1, pp. 1–11, 2019.

[183] J. Hanley and B. McNeil, "The meaning and use of the area under a receiver operating characteristic (ROC) curve," *Radiology*, vol. 143, no. 1, pp. 29–36, 1982.

[184] H. Jeffreys, *Theory of Probability*. Oxford Oxfordshire : New York: Oxford University Press, auflage: third. ed., Mar. 2000.

[185] L. Lü and T. Zhou, "Link prediction in complex networks: A survey," *Physica A: Statistical Mechanics and its Applications*, vol. 390, pp. 1150–1170, Mar. 2011.

[186] Wikipedia contributors, "Jaccard index — Wikipedia, the free encyclopedia." https://en.wikipedia.org/w/index.php?title=Jaccard_index&oldid=1228322016, 2024. [Online; accessed 18-June-2024].

[187] L. A. Adamic and E. Adar, "Friends and neighbors on the Web," *Social Networks*, vol. 25, pp. 211–230, July 2003.

[188] D. Bambar, "The area above the ordinal dominance graph and the area below the receiver operating graph," *Journal of Mathematical Psychology*, vol. 12, no. 4, pp. 387–415, 1975.

[189] A. Ghasemian, H. Hosseinmardi, and A. Clauset, "Evaluating overfit and underfit in models of network community structure," *IEEE Transactions on Knowledge and Data Engineering*, 2019.

[190] A. Ghasemian, H. Hosseinmardi, A. Galstyan, E. M. Airoldi, and A. Clauset, "Stacking models for nearly optimal link prediction in complex networks," *arXiv preprint arXiv:1909.07578*, 2019.

[191] X. He, A. Ghasemian, E. Lee, A. C. Schwarze, A. Clauset, and P. J. Mucha, "Link prediction accuracy on real-world networks under non-uniform missing-edge patterns," *Plos one*, vol. 19, no. 7, p. e0306883, 2024.

[192] M. Vickers and S. Chan, "Representing classroom social structure," *Victoria Institute of Secondary Education, Melbourne*, 1981.

[193] P. Sapiezynski, A. Stopczynski, D. D. Lassen, and S. Lehmann, "Interaction data from the copenhagen networks study," *Scientific Data*, vol. 6, no. 1, p. 315, 2019.

[194] L. Šubelj and M. Bajec, "Robust network community detection using balanced propagation," *The European Physical Journal B*, vol. 81, pp. 353–362, 2011.

[195] Y. Yao, A. Vehtari, D. Simpson, and A. Gelman, "Using stacking to average Bayesian predictive distributions," *Bayesian Analysis*, vol. 13, Sept. 2018. arXiv: 1704.02030.

[196] H. F. Po, A. M. Houben, A.-C. Haeb, D. R. Jenkins, E. J. Hill, H. R. Parr, J. Soriano, and D. Saad, "Inferring structure of cortical neuronal networks from firing data: A statistical physics approach," *arXiv preprint arXiv:2402.18788*, 2024.

[197] P. D'haeseleer, S. Liang, and R. Somogyi, "Genetic network inference: from co-expression clustering to reverse engineering," *Bioinformatics*, vol. 16, no. 8, pp. 707–726, 2000.

[198] A. Braunstein, A. Ingrosso, and A. P. Muntoni, "Network reconstruction from infection cascades," *arXiv:1609.00432 [cond-mat, physics:physics, q-bio]*, Sept. 2016. arXiv: 1609.00432.

[199] G. Cimini, T. Squartini, D. Garlaschelli, and A. Gabrielli, "Systemic risk analysis on reconstructed economic and financial networks," *Scientific reports*, vol. 5, no. 1, p. 15758, 2015.

[200] T. P. Peixoto, "Network Reconstruction and Community Detection from Dynamics," *Physical Review Letters*, vol. 123, p. 128301, Sept. 2019.

[201] T. P. Peixoto, "Scalable network reconstruction in subquadratic time," *arXiv preprint arXiv:2401.01404*, 2024.

[202] T. P. Peixoto, "Network reconstruction via the minimum description length principle," *arXiv preprint arXiv:2405.01015*, 2024.

[203] J. Parkkinen, J. Sinkkonen, A. Gyenge, and S. Kaski, "A block model suitable for sparse graphs," in *Proceedings of the 7th International Workshop on Mining and Learning with Graphs (MLG 2009), Leuven*, 2009.

[204] K. Rohe, J. Tao, X. Han, and N. Binkiewicz, "A note on quickly sampling a sparse matrix with low rank expectation," *The Journal of Machine Learning Research*, vol. 19, no. 1, pp. 3040–3052, 2018. Publisher: JMLR. org.

[205] M. E. J. Newman, "Mixing patterns in networks," *Phys. Rev. E*, vol. 67, p. 026126, Feb. 2003.

[206] V. Batagelj and M. Zaveršnik, "Fast algorithms for determining (generalized) core groups in social networks," *Advances in Data Analysis and Classification*, vol. 5, pp. 129–145, July 2011.

[207] K.-i. Hashimoto, "Zeta Functions of Finite Graphs and Representations of p-Adic Groups," in *Automorphic Forms and Geometry of Arithmetic Varieties* (K. Hashimoto and Y. Namikawa, eds.), vol. 15 of *Advanced Studies in Pure Mathematics*, pp. 211–280, Academic Press, Jan. 1989.

Appendix A

Supplementary Material for Chapter 3

A.1 Posterior predictive sampling

As described in the main text, we obtain samples from the posterior predictive distribution of Eq. (3.1) by first sampling from the posterior distribution of Eq. (2.44) using MCMC and then generating new networks from the inferred models. More specifically, we sample $(\boldsymbol{A},\boldsymbol{k},\boldsymbol{e},\boldsymbol{b})$ from

$$P(\boldsymbol{A},\boldsymbol{k},\boldsymbol{e},\boldsymbol{b}|\boldsymbol{G}) = \frac{P(\boldsymbol{G}|\boldsymbol{A})P(\boldsymbol{A}|\boldsymbol{k},\boldsymbol{e},\boldsymbol{b})P(\boldsymbol{k},\boldsymbol{e},\boldsymbol{b})}{P(\boldsymbol{G})}, \tag{A.1}$$

using the merge-split MCMC of Ref. [53], together with the agglomerative initialization heuristic of Refs. [25, 111], and the multigraph edge moves of Ref. [146]. For networks of size up to $E = 10^5$ edges we observe good equilibration of the MCMC runs, but for large networks it becomes too slow. For these large networks we settle for a point estimate of the partition $\boldsymbol{b}$ obtained by several runs of the initialization algorithm and keeping the best result, and then we equilibrate the chain according to $\boldsymbol{A}$ alone (which affects $\boldsymbol{k}$ and $\boldsymbol{e}$), which tends to happen quickly. We have verified that performing this calculation several times yields very similar results. The only noticeable outcome of this shortcut for larger networks is that it tends to reduce the variance of the posterior predictive distributions, which can potentially contribute to the elevated z-scores we obtained in our analysis. However, since the relative deviation values we obtained did not seem to depend on the size of the network, this gives us confidence that this approach does not introduce significant biases.

Given a sample $(\boldsymbol{A},\boldsymbol{k},\boldsymbol{e},\boldsymbol{b})$, we are interested only in $(\boldsymbol{k},\boldsymbol{e},\boldsymbol{b})$ (and hence samples from their marginal distribution), so we discard $\boldsymbol{A}$ and sample a new multigraph $\boldsymbol{A}'$ from the model of Eq. (2.16). This can be done exactly with an efficient algorithm that works similarly to what was proposed in Refs. [203, 204], but is valid for the microcanonical model: Given the parameters

$(\boldsymbol{k},\boldsymbol{e},\boldsymbol{b})$ we proceed by creating for each group r a multiset of candidate nodes $\boldsymbol{v}_r$, containing k_i copies of each node i with $b_i = r$. Then, for each group pair (r,s) with $r \leq s$ and $e_{rs} > 0$, we repeat the following three steps for an e_{rs} number of times (or $e_{rs}/2$ if $r = s$):

1. We sample a node i from the multiset $\boldsymbol{v}_r$ uniformly at random, and we remove it from the multiset.

2. We sample a node j from the multiset $\boldsymbol{v}_s$ uniformly at random, and we remove it from the multiset.

3. We add an edge (i, j) to $\boldsymbol{A}$ (i.e. increment A_{ij} by one, or two if $i = j$).

The resulting multigraph $\boldsymbol{A}$ is sampled exactly with a probability given by Eq.(2.16). Since the number of nonzero entries of $\boldsymbol{e}$ cannot be larger than the total number of edges E, the whole algorithm finishes in time $O(N+E)$, where N is the number of nodes.

Given a sample $\boldsymbol{A}$, we obtain a simple graph $\boldsymbol{G}$ simply by removing all self-loops and truncating the edge multiplicities, i.e.

$$G_{ij} = \begin{cases} 1, \text{ if } A_{ij} > 0 \text{ and } i \neq j, \\ 0, \text{ otherwise.} \end{cases} \tag{A.2}$$

Finally, given $\boldsymbol{G}$ we compute the network descriptor $f(\boldsymbol{G})$ of interest.

A C++ implementation of every algorithm used in this analysis is freely available as part of the `graph-tool` library [125].

A.2 Network descriptors

Below are the definitions of the descriptors used in our analyses.

Degree assortativity, r Defined as [205]

$$r = \frac{\sum_{kk'} kk'(m_{kk'} - m_k m_{k'})}{\sigma_k \sigma_{k'}},$$

where $m_{kk'}$ is the fraction of edges with endpoints of degree k and k', $m_k = \sum_{k'} m_{kk'}$, and σ_k is the standard deviation of m_k.

Mean k-core, $\langle c \rangle$ The k-core is a maximal set of vertices such that its induced subgraph only contains vertices with degree larger than or equal to k. The k-core value c_i of node i is

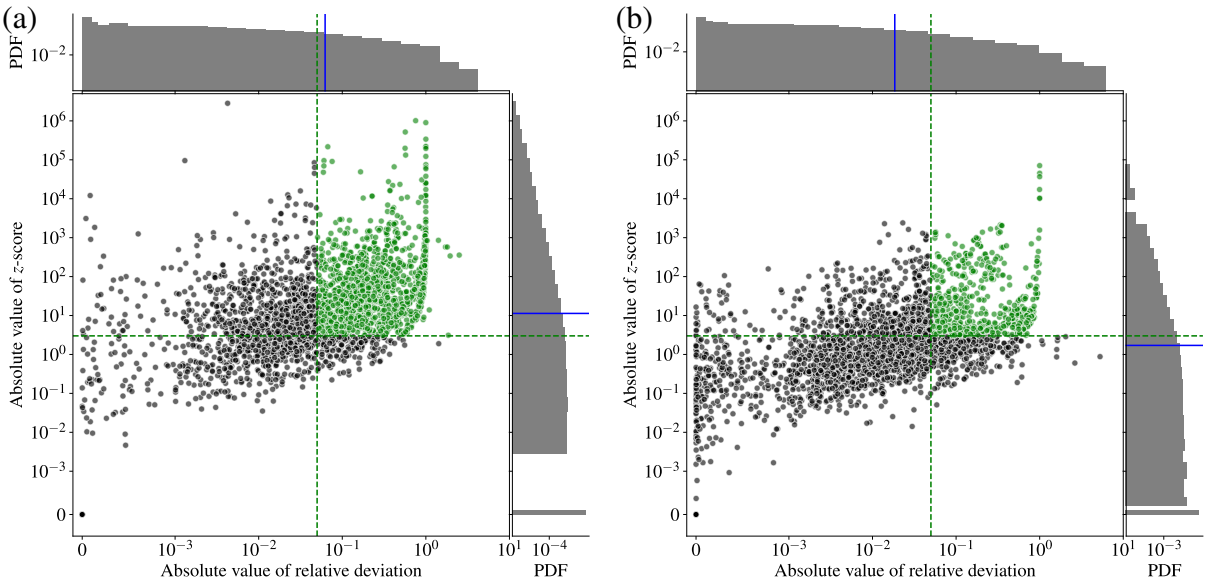


Figure A.1: Absolute value of the z-score versus absolute value of relative deviation, for every descriptor value and network in the corpus, according to (a) the configuration model and (b) the DCSBM. The dashed lines mark the values $|z| = 3$ and $|\Delta| = 0.05$, and the histograms the marginal distributions. The solid blue lines mark the median values.

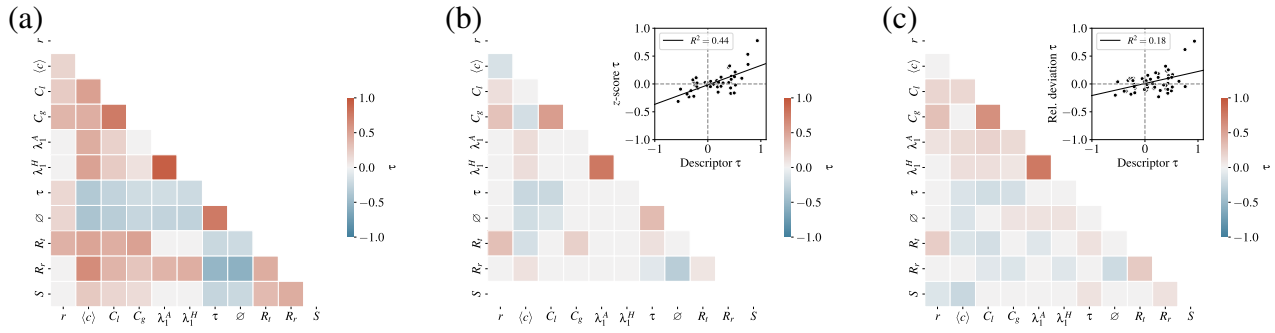


Figure A.2: (a) Kendall's correlation coefficient τ between pairs of descriptor values across all networks in the corpus. Panels (b) and (c) show the same but for z-score and relative deviation values, respectively, according to the DCSBM. The insets show the correlation between coefficients from each respective panel and panel (a).

the largest value of k for which i belongs to the k-core. The mean value is then

$$\langle c \rangle = \frac{1}{N} \sum_i c_i.$$

This can be computed in time $O(N+E)$ according to the algorithm of Ref. [206].

Mean local clustering coefficient, C_l The local clustering coefficient [85] of node i is given by

$$C_i = \frac{\sum_{jk} G_{ij} G_{ki} G_{jk}}{k_i(k_i-1)}.$$

It measures the fraction of pairs of neighbors that are also connected. The mean value is then just

$$C_l = \frac{1}{N} \sum_i C_i.$$

Global clustering coefficient, C_g The global clustering coefficient of is given by

$$C_g = \frac{\sum_{ijk} G_{ij} G_{ki} G_{jk}}{\sum_i k_i(k_i-1)}.$$

It measures the fraction of connected triads that close to form a triangle.

Leading eigenvalue of adjacency matrix, λ_1^A The leading eigenvalue of the adjacency matrix is the largest value of λ which solves

$$\boldsymbol{G}\boldsymbol{x} = \lambda \boldsymbol{x},$$

where $\boldsymbol{x}$ is the associated eigenvector.

Leading eigenvalue of Hashimoto matrix, λ_1^H The leading eigenvalue of the Hashimoto (a.k.a. non-backtracking) matrix [207] is the largest value of λ which solves

$$\boldsymbol{H}\boldsymbol{x} = \lambda \boldsymbol{x},$$

where $\boldsymbol{x}$ is the associated eigenvector, and $\boldsymbol{H}$ is an asymmetric $E \times E$ matrix with entries defined as

$$H_{k\to l, i\to j} = \begin{cases} 1 & \text{if } G_{kl} = G_{ij} = 1, l = i, k \neq j, \\ 0 & \text{otherwise.} \end{cases}$$

Characteristic time of a random walk, τ The characteristic time of a random walk is obtained via the second largest eigenvalue $\lambda_2^T \in [0,1]$ of the transition matrix $\boldsymbol{T}$, with entries

$$T_{ij} = \frac{G_{ij}}{k_j},$$

where $k_i = \sum_j G_{ji}$. It is defined as

$$\tau = -\ln \lambda_2^T.$$

If the network is disconnected, we compute τ only on the largest component.

Pseudo-diameter, $\varnothing$ The pseudo-diameter is an approximate graph diameter. It is obtained by starting from an arbitrary source node, and finding a target node that is farthest away from the source. This process is repeated by treating the target as the new starting node, and ends when the graph distance no longer increases. This graph distance is taken to be the pseudo-diameter. The algorithm runs in time $O(N+E)$.

If the network is disconnected, $\varnothing$ is taken as the maximum of pseudo-diameters of the connected components.

Node percolation profile (random removal), R_r We chose a random node order, and remove nodes sequentially from the graph according to it. If S_i is the fraction of nodes in the largest component after the i-th removal, then the profile value is

$$R_r = \frac{1}{N}\sum_i S_i.$$

The value is averaged over several node orderings.

Node percolation profile (targeted removal), R_t The computation is the same as R_r, but the nodes are always removed in decreasing order of the degree.

Fraction of nodes in the largest component, S A component is a maximal set of nodes that are connected by a path. The largest component is the component with the largest number of nodes, and S is the fraction of all nodes that belong to it.

In Fig. A.1 we show how the z-scores and relative deviation values are related for every network descriptor, according to both models used. In Fig. A.2 we show Kendall's τ correlation coefficient among the descriptor values themselves, as well as their z-scores and relative deviations, according to the DCSBM. The insets show how the correlations among the deviations are themselves also correlated with the descriptor correlations.

A.3 Model deviations in clustering coefficient

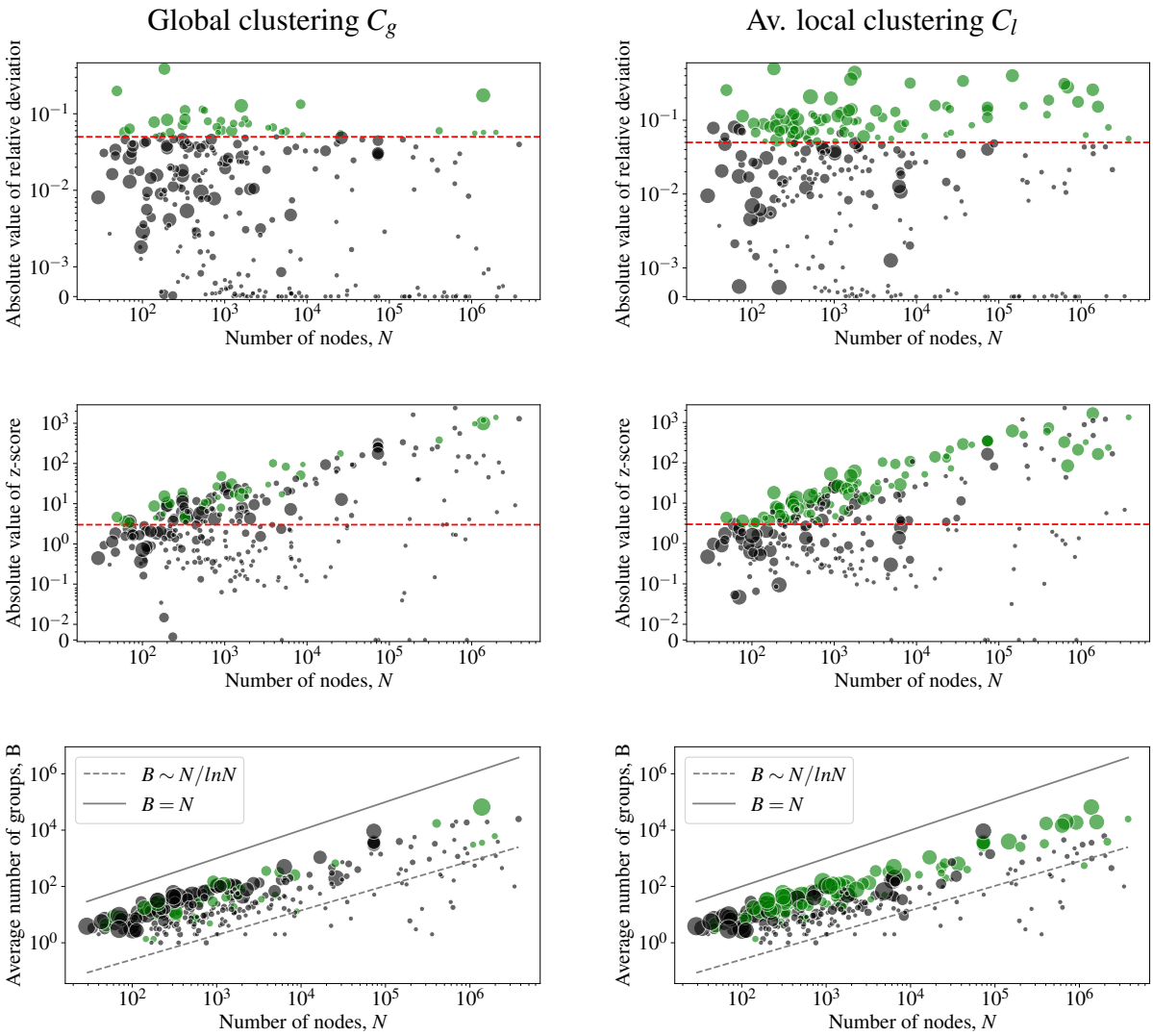


Figure A.3: Absolute value of the relative deviation (top), z-score (middle), and average number of inferred groups (bottom), as a function of the number of nodes, for every network in the corpus. The dashed red line marks the values of $|\Delta| = 0.05$ and $|z| = 3$. The points in green color correspond to the descriptors that are not reproduced. The point size is proportional to the empirical value of the descriptor.

Appendix B

Supplementary Material for Chapter 4

B.1 Correction of marginal probabilities

To correct our estimates of marginal probabilities $\pi(\boldsymbol{A})$ and sample networks that preserve the total number of edges, we look for $P(\boldsymbol{A})$ that can be obtained via the Lagrangian

$$L = \sum_{\boldsymbol{A}} P(\boldsymbol{A}) \ln \frac{P(\boldsymbol{A})}{\pi(\boldsymbol{A})} - \beta \left(\sum_{\boldsymbol{A}} \sum_{i<j} A_{ij} P(\boldsymbol{A}) - E \right) - \lambda \left(\sum_{\boldsymbol{A}} P(\boldsymbol{A}) - 1 \right), \tag{B.1}$$

where β and λ are Lagrange multipliers that enforce the constraints of total number of edges and normalization, respectively.

Obtaining the saddle point $\{\partial L/\partial P(\boldsymbol{A}) = 0, \partial L/\partial \beta = 0, \partial L/\partial \lambda = 0\}$ yields

$$P(\boldsymbol{A}) = \frac{\pi(\boldsymbol{A}) e^{\beta \Sigma_{i<j} A_{ij}}}{Z}, \tag{B.2}$$

where Z is a normalization constant given by

$$Z = \sum_{\boldsymbol{A}} \pi(\boldsymbol{A}) e^{\beta \Sigma_{i<j} A_{ij}}. \tag{B.3}$$

By enforcing the constraint on the total number of edges we get

$$\sum_{\boldsymbol{A}} \frac{\pi(\boldsymbol{A}) e^{\beta \Sigma_{i<j} A_{ij}} \sum_{i<j} A_{ij}}{Z} = E. \tag{B.4}$$

Assume

$$\pi(A) = \prod_{i<j} \pi_{ij}^{A_{ij}} (1 - \pi_{ij})^{1 - A_{ij}}. \tag{B.5}$$

This assumption is justified by observing that sampling networks directly from the posterior of Eq. (4.1). and by considering a Bernoulli process on each entry (i, j) with probability of success equal to the corresponding marginal probability (Eq. (4.2))on each yields similar results.

Then Eq. (B.2) can be written as:

$$P(\boldsymbol{A}) = \frac{\prod_{i<j} \pi_{ij}^{A_{ij}} (1-\pi_{ij})^{1-A_{ij}} e^{\beta \sum_{i<j} A_{ij}}}{Z}. \tag{B.6}$$

Furthermore, Eq. (B.3) can be written as:

$$Z = \sum_{\boldsymbol{A}} \prod_{i<j} \pi_{ij}^{A_{ij}} (1-\pi_{ij})^{1-A_{ij}} e^{\beta \sum_{i<j} A_{ij}}. \tag{B.7}$$

Since $e^{\beta \sum_{i<j} A_{ij}} = \prod_{i<j} e^{\beta A_{ij}}$, the last two equations become:

$$P(\boldsymbol{A}) = \frac{\prod_{i<j} (e^{\beta} \pi_{ij})^{A_{ij}} (1-\pi_{ij})^{1-A_{ij}}}{Z}, \tag{B.8}$$

$$Z = \sum_{\boldsymbol{A}} \prod_{i<j} (e^{\beta} \pi_{ij})^{A_{ij}} (1-\pi_{ij})^{1-A_{ij}}. \tag{B.9}$$

Z can be written as

$$Z = \prod_{i<j} e^{\beta} \pi_{ij} + 1 - \pi_{ij}. \tag{B.10}$$

Then

$$P(\boldsymbol{A}) = \prod_{i<j} \left(\frac{e^{\beta} \pi_{ij}}{e^{\beta} \pi_{ij} + 1 - \pi_{ij}} \right)^{A_{ij}} \left(\frac{1-\pi_{ij}}{e^{\beta} \pi_{ij} + 1 - \pi_{ij}} \right)^{1-A_{ij}}, \tag{B.11}$$

and the probability of generating and edge for the pair (i, j) is given by

$$p_{ij} = \frac{e^{\beta} \pi_{ij}}{e^{\beta} \pi_{ij} + 1 - \pi_{ij}}. \tag{B.12}$$

Note that the correction is made only for entries having positive probability π_{ij}.

Also note that, using Eq. (B.9), $\partial \ln Z / \partial \beta = E$, i.e., we get the constraint of total number of edges.

Using Eq. (B.10), $\partial \ln Z / \partial \beta = \sum_{i<j} e^{\beta} \pi_{ij} / (e^{\beta} \pi_{ij} + 1 - \pi_{ij})$.

To obtain β, we solve numerically the equation

$$\sum_{i<j} \frac{e^{\beta}\pi_{ij}}{e^{\beta}\pi_{ij} + 1 - \pi_{ij}} = E. \tag{B.13}$$

Increasing the number of measurements from 1 to 3

Although we need to make a correction in the marginal probabilities in a way that our reconstruction approach preserves the total number of edges, it should be mentioned that such correction is less necessary as number of measurements n increase. Fig. B.1 shows that, the uncorrected number of edges of reconstructed networks is closer to the *true* value when $n > 1$.

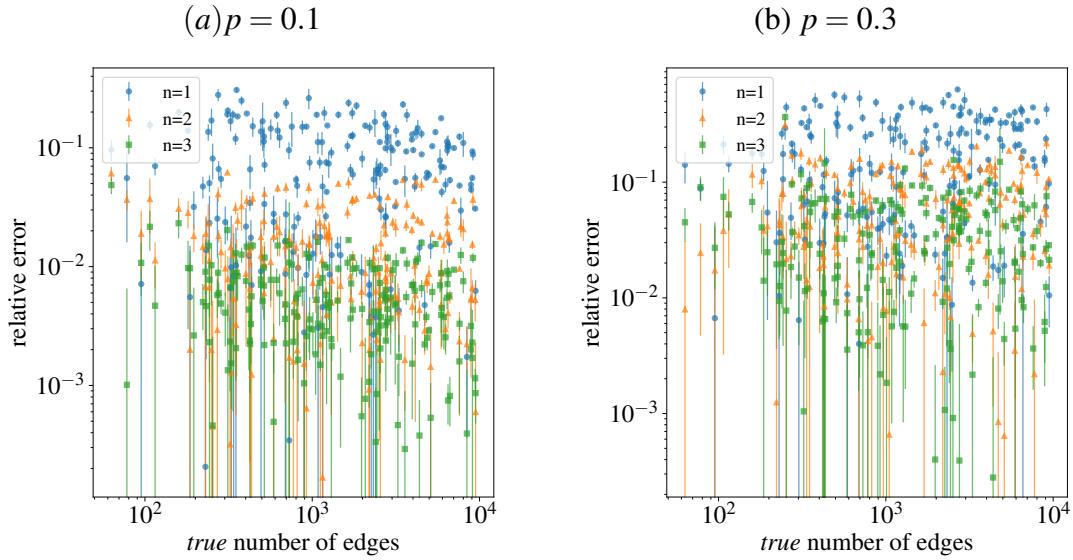


Figure B.1: Average error in the number of edges (*true* vs inferred but uncorrected) as a function of a network index where the *true* number of edges is sorted in increasing order. (a) and (b) indicate different noise levels.

B.2 Results for noise level $p = 0.3$

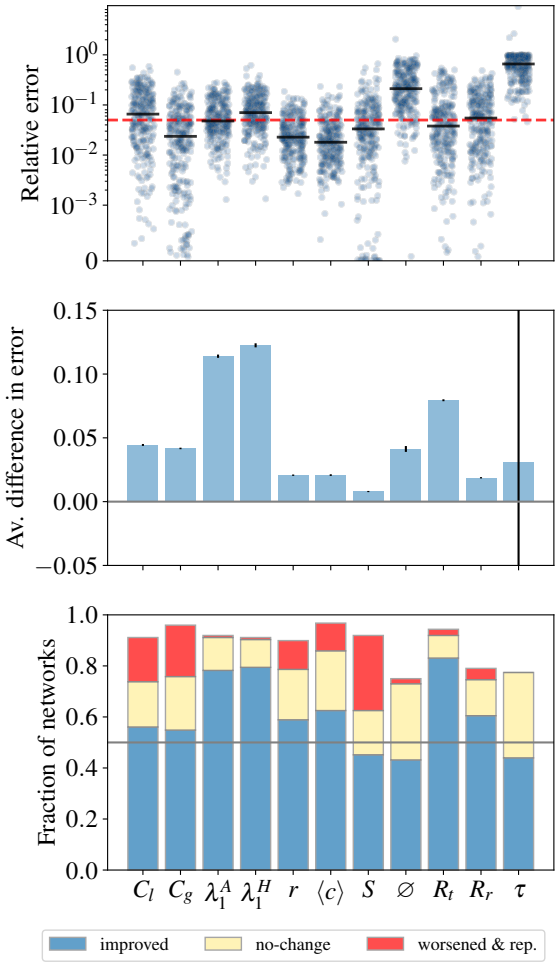


Figure B.2: Same as Fig. 4.3 for noise level $p = 0.3$.

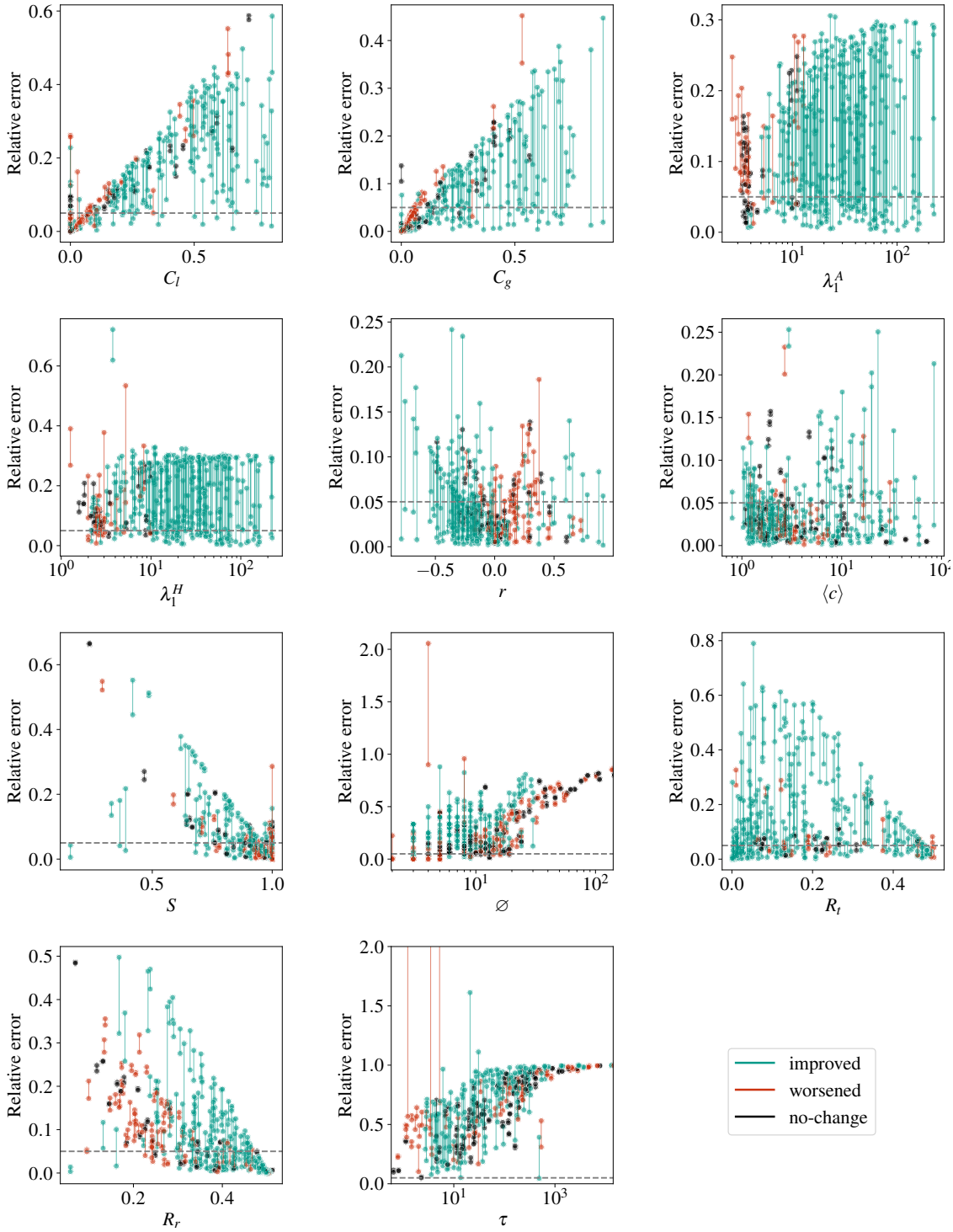


Figure B.3: Same as Fig. 4.4 for noise level $p = 0.3$.

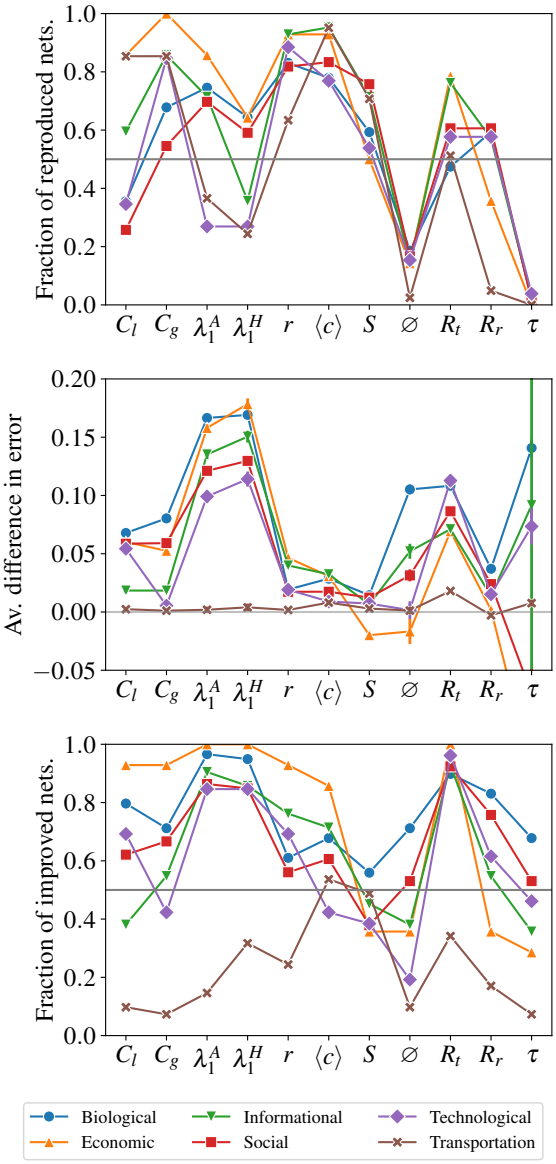


Figure B.4: Same as Fig. 4.6 for noise level $p = 0.3$.

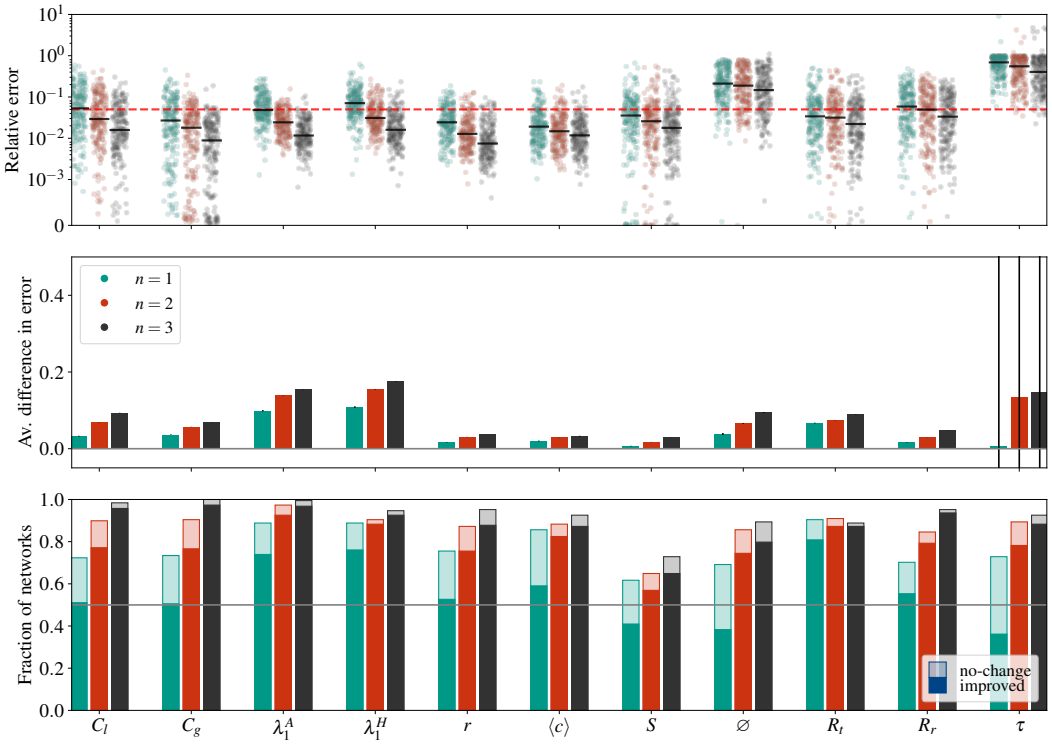


Figure B.5: Same as Fig. 4.7 for noise level $p = 0.3$

Appendix C

Supplementary Material for Chapter 5

C.1 Mixed Random Label Model

As a starting point, we assume a non-hierarchical partition $\boldsymbol{b} = \{b_i\}$. Let $\mu(r)$ be a bijective function, such that

$$\mu(b_i) = c_i, \quad \forall i, \tag{C.1}$$

i.e., there is a partition $\boldsymbol{c}$ that is identical to $\boldsymbol{b}$ up to a random label permutation. We denote this equivalence by the indicator function

$$[\boldsymbol{b} \sim \boldsymbol{c}] = \begin{cases} 1 & \text{if } \boldsymbol{b} \text{ is a label permutation of } \boldsymbol{c}, \\ 0 & \text{otherwise.} \end{cases} \tag{C.2}$$

Since the posterior distribution is invariant to label permutations, then

$$\pi(\boldsymbol{b}) = \pi(\boldsymbol{c}). \tag{C.3}$$

Although this relation also holds for hierarchical partitions, we need to introduce additional details, since an invariant label permutation in a given level also affects the node labels in the immediately superior level. In particular, consider a hierarchical partition $\{\boldsymbol{b}_l\}$. Consider also a bijection $\mu(r)$ for labels at level l, such that $\mu(b_i^l) = c_i^l$, and change the membership in level $l+1$ to $b_i^{l+1} = c_{\mu(i)}^{l+1}$. Then the hierarchical partitions $\{\boldsymbol{b}_l\}$ and $\{\boldsymbol{c}_l\}$ are identical up to a relabeling of the groups, which is denoted by

$$[\{\boldsymbol{b}_l\} \sim \{\boldsymbol{c}_l\}] = \begin{cases} 1 & \text{if } \{\boldsymbol{b}_l\} \text{ is identical to } \{\boldsymbol{c}_l\} \text{ up to a label permutation,} \\ 0 & \text{otherwise.} \end{cases} \tag{C.4}$$

Additionally, $\pi(\{\boldsymbol{b}_l\}) = \pi(\{\boldsymbol{c}_l\})$.

Considering these preliminary ideas, the posterior distribution of partitions can be modeled with a mixture model, in which each partition can belong to one of K clusters, or "modes". More specifically, the posterior $\pi(\{\boldsymbol{b}_l\})$ can be approximated by

$$\pi(\{\boldsymbol{b}_l\}) \approx \sum_{k,\{\boldsymbol{c}_l\}} P(\{\boldsymbol{b}_l\}|\{\boldsymbol{c}_l\})P(\{\boldsymbol{c}_l\}|k)P(k), \tag{C.5}$$

where $P(k) = \omega_k$ determines the mode mixture, i.e., the relative size of mode k, with $\sum_k \omega_k = 1$;

$$P(\{\boldsymbol{c}_l\}|k) = \prod_l \prod_i p_i^{(l,k)}(c_i^l) \tag{C.6}$$

are the independent[1] marginal distributions of mode k, where $p_i^{(l,k)}(r)$ is the probability that a node i has group label r in level l and mode k; and

$$P(\{\boldsymbol{b}_l\}|\{\boldsymbol{c}_l\}) = \frac{[\{\boldsymbol{b}_l\} \sim \{\boldsymbol{c}_l\}]}{\prod_l q(\boldsymbol{b}_l)!} \tag{C.7}$$

is the random relabeling of groups, with $q(\boldsymbol{b})!$ being the total number of label permutations of $\boldsymbol{b}$. If the modes are significantly separated, we can assume that

$$\pi(\{\boldsymbol{b}_l\}) \approx \max_{k,\{\boldsymbol{c}_l\}} P(\{\boldsymbol{b}_l\}|\{\boldsymbol{c}_l\})P(\{\boldsymbol{c}_l\}|k)P(k), \tag{C.8}$$

which in turn, allows us to write the entropy as

$$H(\{\boldsymbol{b}_l\}) \approx H(\{\boldsymbol{b}_l\},\{\boldsymbol{c}_l\},k) = H(\{\boldsymbol{b}_l\}|\{\boldsymbol{c}_l\}) + H(\{\boldsymbol{c}_l\}|k) + H(k), \tag{C.9}$$

where

[1] The assumption is that, at every level, the labels are sampled independently.

$$H(k) = -\sum_k w_k \ln w_k \tag{C.10}$$

is the entropy of the mode mixture distribution;

$$H(\{\boldsymbol{b}_l\}|k) = -\sum_k w_k \sum_{\{\boldsymbol{b}_l\}} P(\{\boldsymbol{b}_l\}|k) \ln P(\{\boldsymbol{b}_l\}|k) \tag{C.11}$$

$$= -\sum_k w_k \sum_l \sum_i \sum_r p_i^{(l,k)}(r) \ln p_i^{(l,k)}(r) \tag{C.12}$$

is the entropy of mode k; and

$$H(\{\boldsymbol{b}_l\}|\{\boldsymbol{c}_l\}) = -\sum_{\{\boldsymbol{c}_l\}} P(\{\boldsymbol{c}_l\}) \sum_{\{\boldsymbol{b}_l\}} P(\{\boldsymbol{b}_l\}|\{\boldsymbol{c}_l\}) \ln P(\{\boldsymbol{b}_l\}|\{\boldsymbol{c}_l\}) \tag{C.13}$$

$$= \sum_{\{\boldsymbol{c}_l\}} P(\{\boldsymbol{c}_l\}) \sum_l \ln q(\boldsymbol{c}_l)! = \sum_{\{\boldsymbol{b}_l\}} P(\{\boldsymbol{b}_l\}) \sum_l \ln q(\boldsymbol{b}_l)!, \tag{C.14}$$

is the relabeling entropy.

Considering all these terms, the mixed random label model provides the following approximation for the log-evidence

$$\ln P(\boldsymbol{A}) \approx \langle \ln P(\boldsymbol{A}, \{\boldsymbol{b}_l\}) \rangle + \sum_l \left\langle \ln q(\boldsymbol{b}^l)! \right\rangle - \sum_k w_k \ln w_k - \sum_k w_k \sum_l \sum_i \sum_r p_i^{(l,k)}(r) \ln p_i^{(l,k)}(r). \tag{C.15}$$

In order to compute each of the terms of Eq. (C.15), we first need to sample M partitions from the posterior distribution.[2] The first two terms can be computed directly by averaging the corresponding quantities. For the remaining ones, we need to fit the mixed random label model to the sampled partitions, obtain the estimates of ω and $\{\boldsymbol{p}_l\}$, and use them in the computation of such terms. It should be noted that, the estimate for ω_k corresponds to the ratio between the number of sampled partitions in mode k and the total number of sampled partitions, i.e. $\omega_k = M_k/M$. In turn, this estimate can be interpreted as the relative posterior plausibility of mode k serving as an alternative explanation for the data.

[2] In this work, depending on the network, we use between 10^4 and 10^5 partitions.

C.2 Supplementary Figures

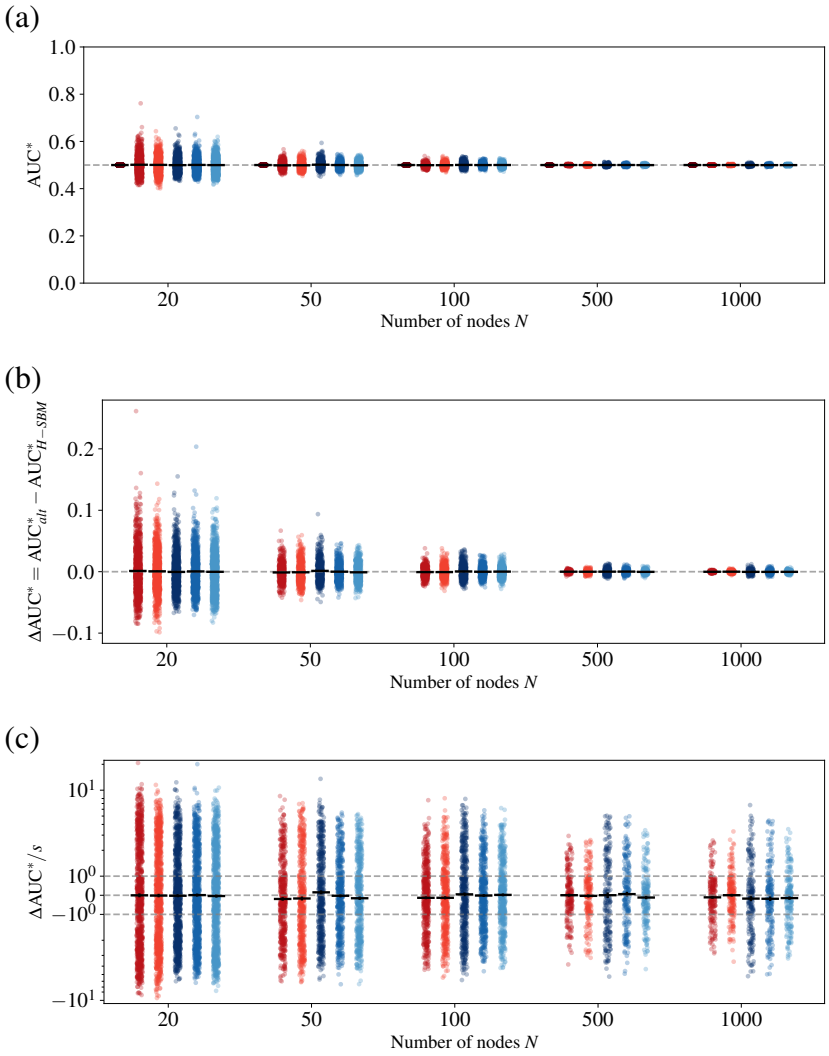


Figure C.1: (a) AUC (point estimate) yielded by candidate models for several instances of the Erdős-Rényi model having average degree $\langle k \rangle = 10$, under an edge denoising task (at least 50 edge removal experiments removing 10% of edges on average were conducted). The point color indicates the model, which is a combination of the model variant (either H-SBM or HDC-SBM) and the number of groups. Each point corresponds to an instance of the Erdős-Rényi model, having N nodes and average degree $\langle k \rangle$. For each combination of N and $\langle k \rangle$, there are 200 samples. (b) Difference between the AUC (point estimate) yielded by simplest model AUC_{H-SBM} and the AUC yielded by more complex alternative models AUC_{alt}. The point color indicates the alternative model. (c) Ratio between the difference in AUC (in panel (b)) and the corresponding standard deviation of the mean AUC difference.

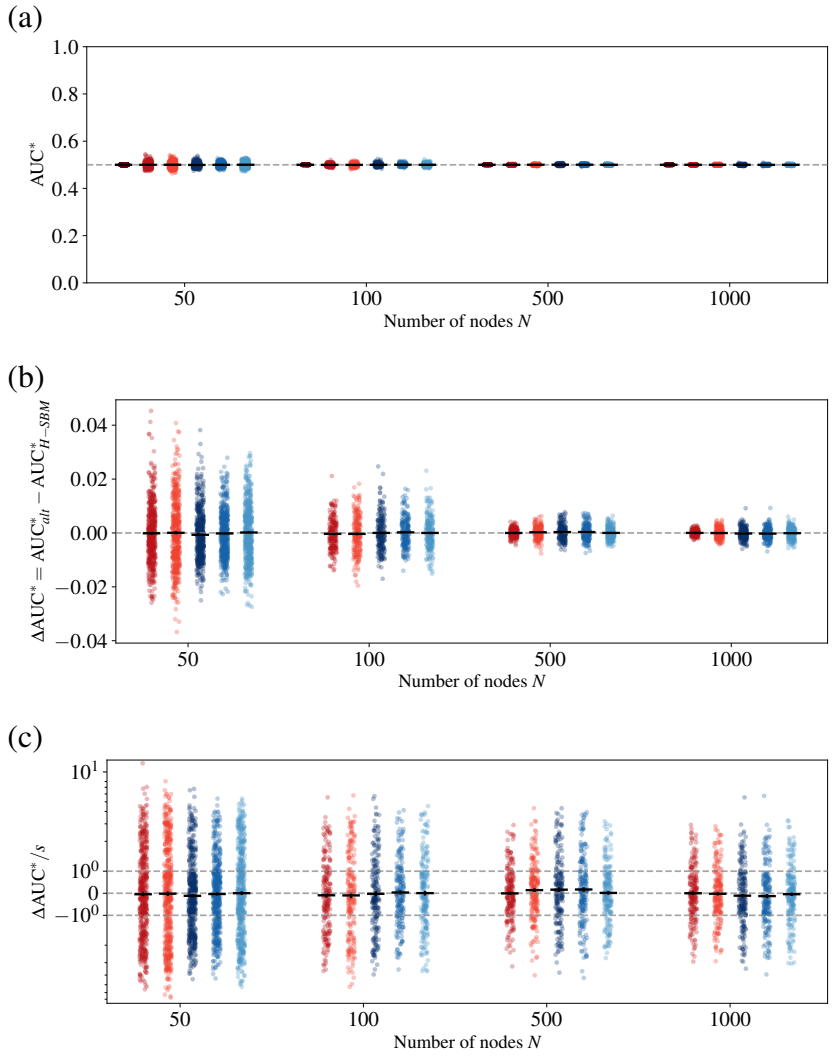
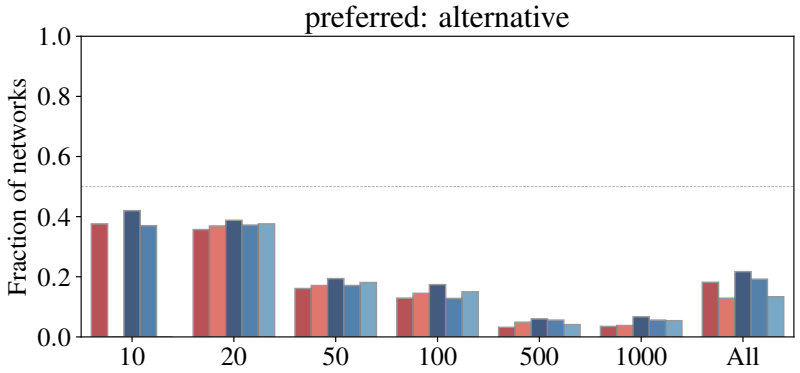
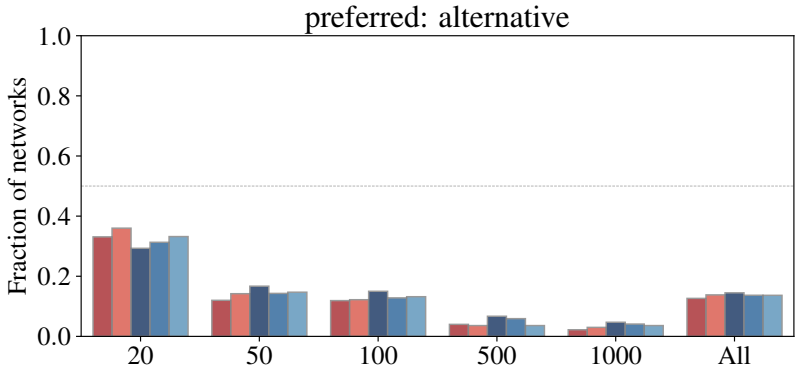


Figure C.2: (a) AUC (point estimate) yielded by candidate models for several instances of the Erdős-Rényi model having average degree $\langle k \rangle = 20$, under an edge denoising task (at least 50 edge removal experiments removing 10% of edges on average were conducted). The point color indicates the model, which is a combination of the model variant (either H-SBM or HDC-SBM) and the number of groups. Each point corresponds to an instance of the Erdős-Rényi model, having N nodes and average degree $\langle k \rangle$. For each combination of N and $\langle k \rangle$, there are 200 samples. (b) Difference between the AUC (point estimate) yielded by simplest model AUC_{H-SBM} and the AUC yielded by more complex alternative models AUC_{alt}. The point color indicates the alternative model. (c) Ratio between the difference in AUC (in panel (b)) and the corresponding standard deviation of the mean AUC difference.

(a) $\langle k \rangle = 4$



(b) $\langle k \rangle = 10$



(c) $\langle k \rangle = 20$

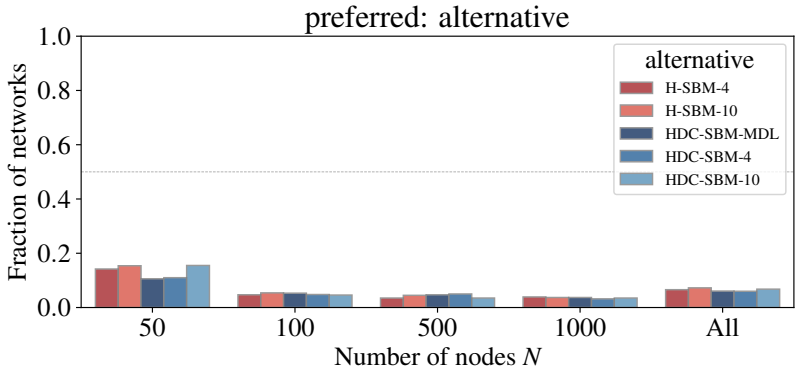
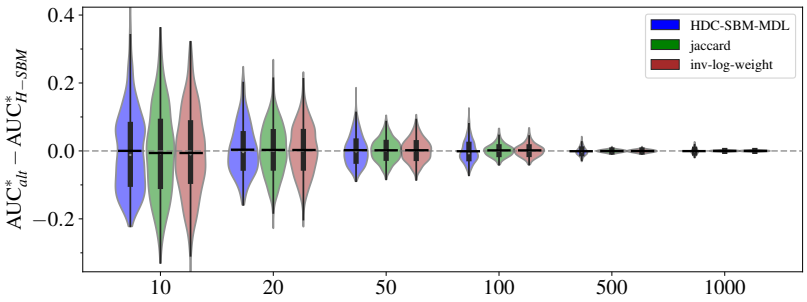
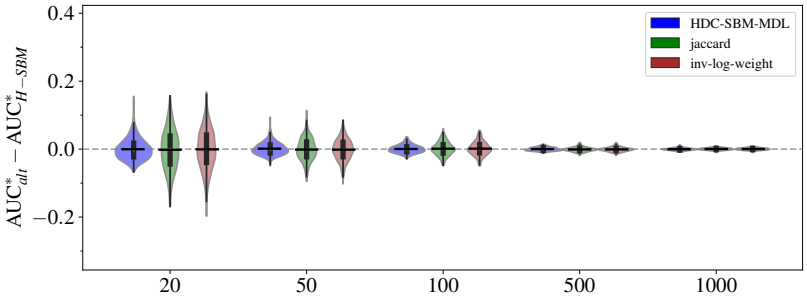


Figure C.3: Percentage of instances of the Erdős-Rényi model for which alternative more complex model have better AUC than the true model. The bar color indicates the alternative model, which is a combination of the model variant (either H-SBM or HDC-SBM) and the number of groups. For $N \in \{10, 20\}$ and $\langle k \rangle = 4$, there are 1000 samples. For $N = \{50, 100\}$ and $\langle k \rangle = 4$, there are 500 samples. For the remaining values of N and $\langle k \rangle$, there are 200 samples.

(a) $\langle k \rangle = 4$



(b) $\langle k \rangle = 10$



(c) $\langle k \rangle = 20$

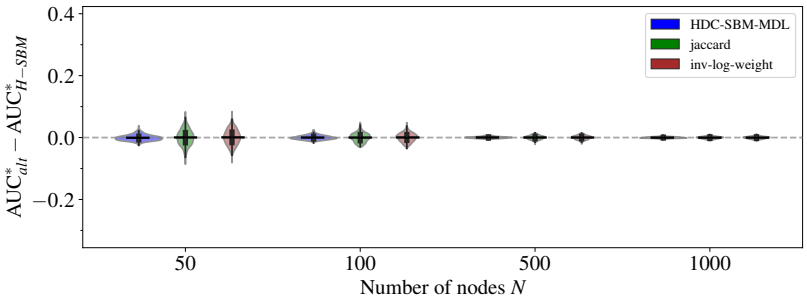
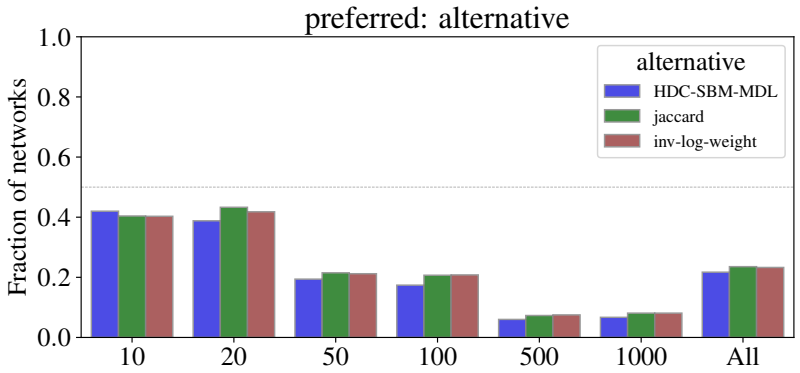
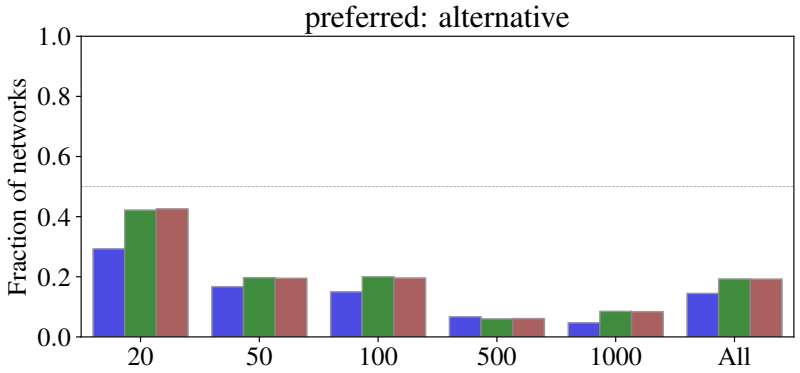


Figure C.4: Difference between the AUC yield by simplest model AUC_{H-SBM} and other non probabilistic strategies AUC_{alt}. We also include the results for HDC-SBM. The point color indicates the which approach was used for link prediction. Each point corresponds to an instance of the Erdős-Rényi model, having N nodes and average degree $\langle k \rangle$. For $N \in \{10, 20\}$ and $\langle k \rangle = 4$, there are 1000 samples. For $N = \{50, 100\}$ and $\langle k \rangle = 4$, there are 500 samples. For the remaining values of N and $\langle k \rangle$, there are 200 samples. In link prediction, 10% of edges were removed.

(a) $\langle k \rangle = 4$



(b) $\langle k \rangle = 10$



(c) $\langle k \rangle = 20$

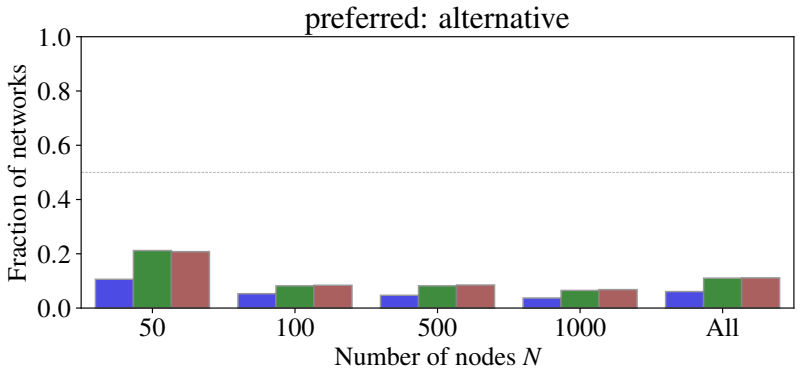


Figure C.5: Percentage of instances of the Erdős-Rényi model for which other non probabilistic strategies have better AUC than the true model. The bar color indicates the which approach was used for link prediction. For $N \in \{10, 20\}$ and $\langle k \rangle = 4$, there are 1000 samples. For $N = \{50, 100\}$ and $\langle k \rangle = 4$, there are 500 samples. For the remaining values of N and $\langle k \rangle$, there are 200 samples.

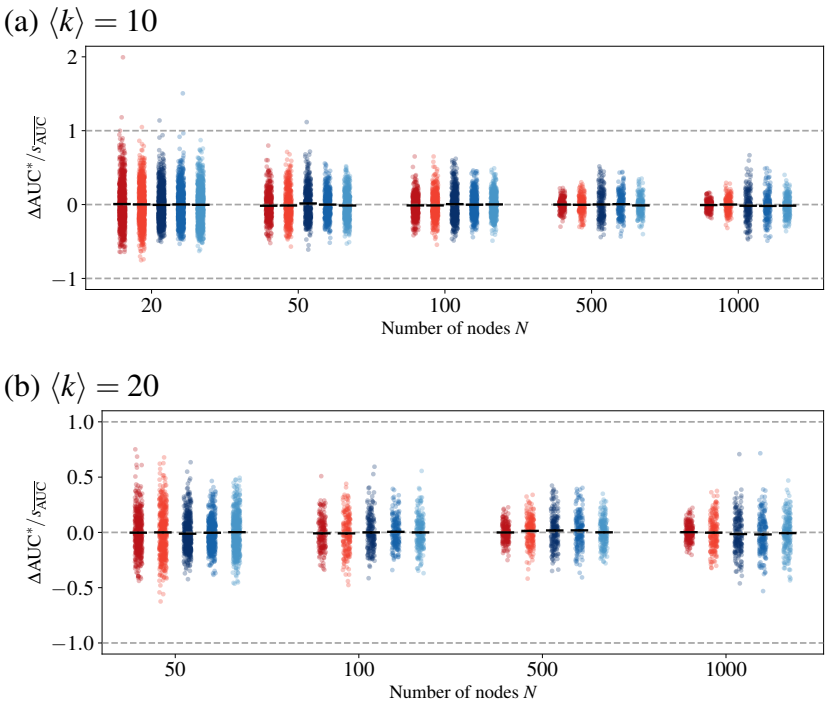


Figure C.6: Ratio between the difference in AUC (Fig. 5.1(b)) and the corresponding standard deviation of Eq. 5.28, for several instances of the Erdős-Rényi model. The point color indicates the alternative model.

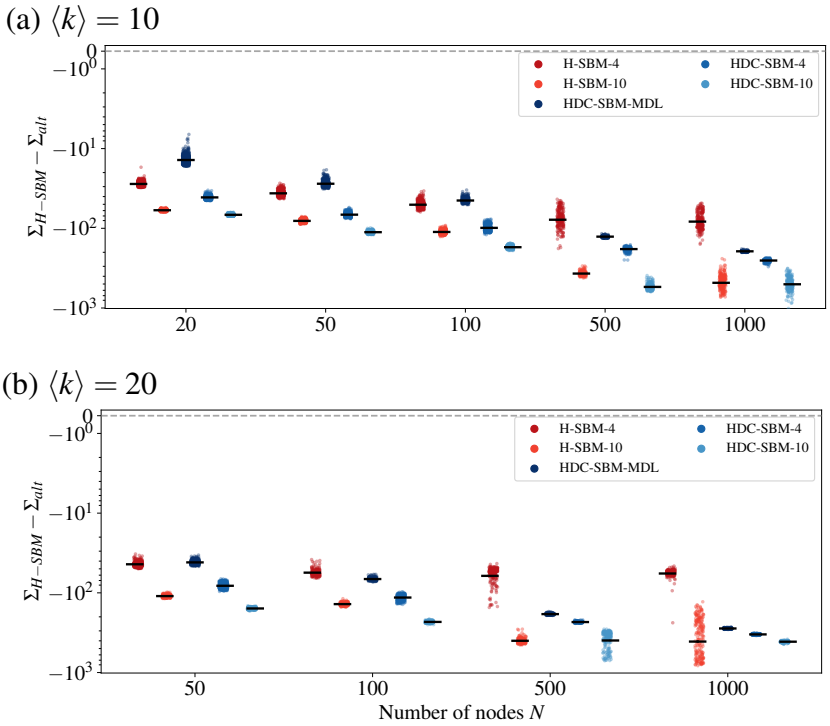


Figure C.7: Difference between the description length of the simplest model Σ_{H-SBM} and the description length of more complex alternative models Σ_{alt}, for several instances of the Erdős-Rényi model. The point color indicates the alternative model, which is a combination of the model variant (either H-SBM or HDC-SBM) and the number of groups. Each point corresponds to an instance of the Erdős-Rényi model, having N nodes and average degree $\langle k \rangle$. For each combination of N and $\langle k \rangle$, there are 200 samples.

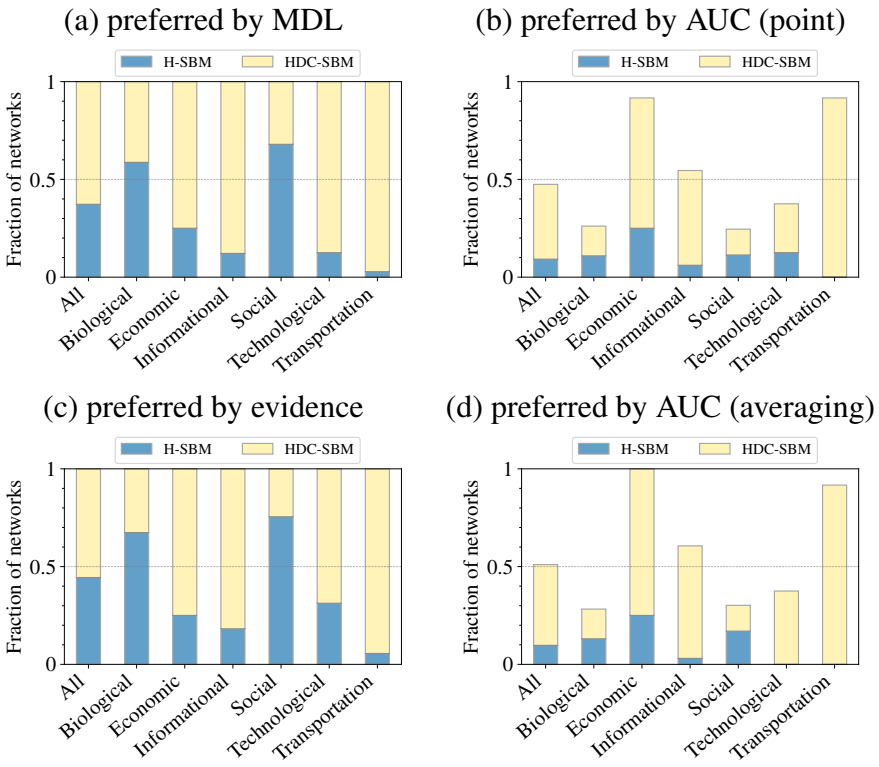


Figure C.8: Fraction of empirical networks for which a model is preferred by considered criteria. Results are reported for all networks in the corpus and for each network domain.

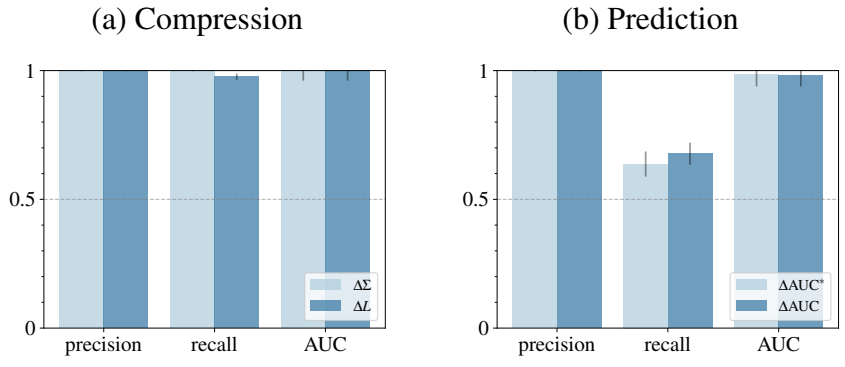


Figure C.9: Precision, recall, and AUC obtained by treating model selection as a classification task. The targets corresponds to the true model, and the predictors to differences in (a) compression criteria or (b) predictive criteria.

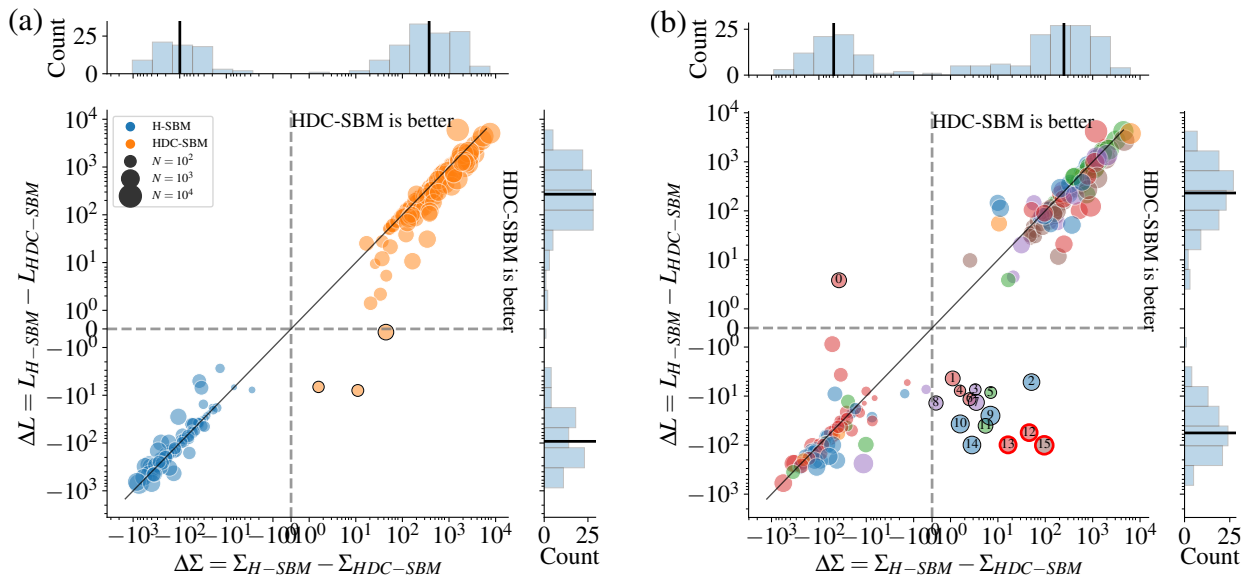


Figure C.10: Difference in -log-evidence (ΔL) as a function of the difference in description length ($\Delta\Sigma$) for (a) synthetic networks and (b) empirical networks. The format of points follows the same rules as in previous plots.

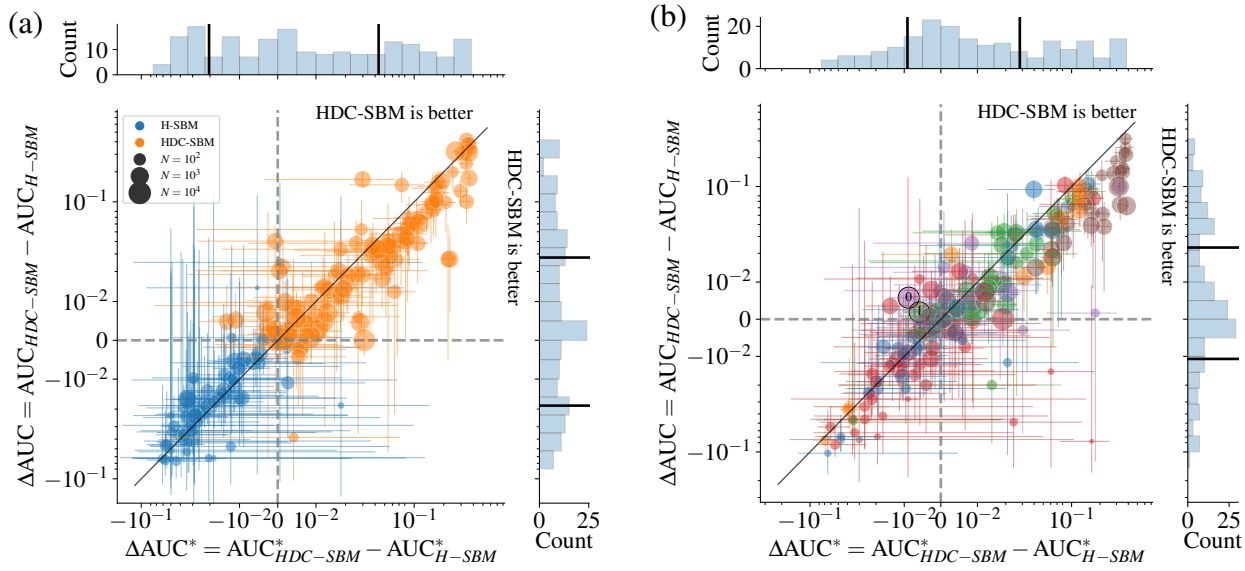


Figure C.11: Difference in AUC from posterior averages as a function of the difference in AUC point estimate (AUC^*) for (a) synthetic networks and (b) empirical networks. The format of points follows the same rules as in previous plots.

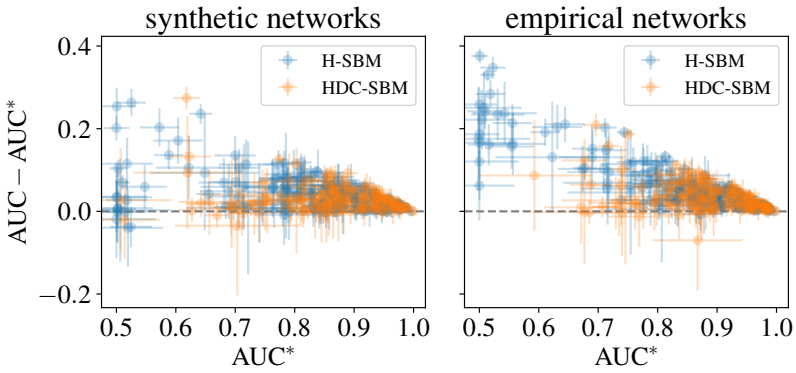


Figure C.12: Difference between AUC obtained from posterior averages (AUC) and point estimate (AUC*) as a function of the latter. The color indicates which model was used in the prediction task.

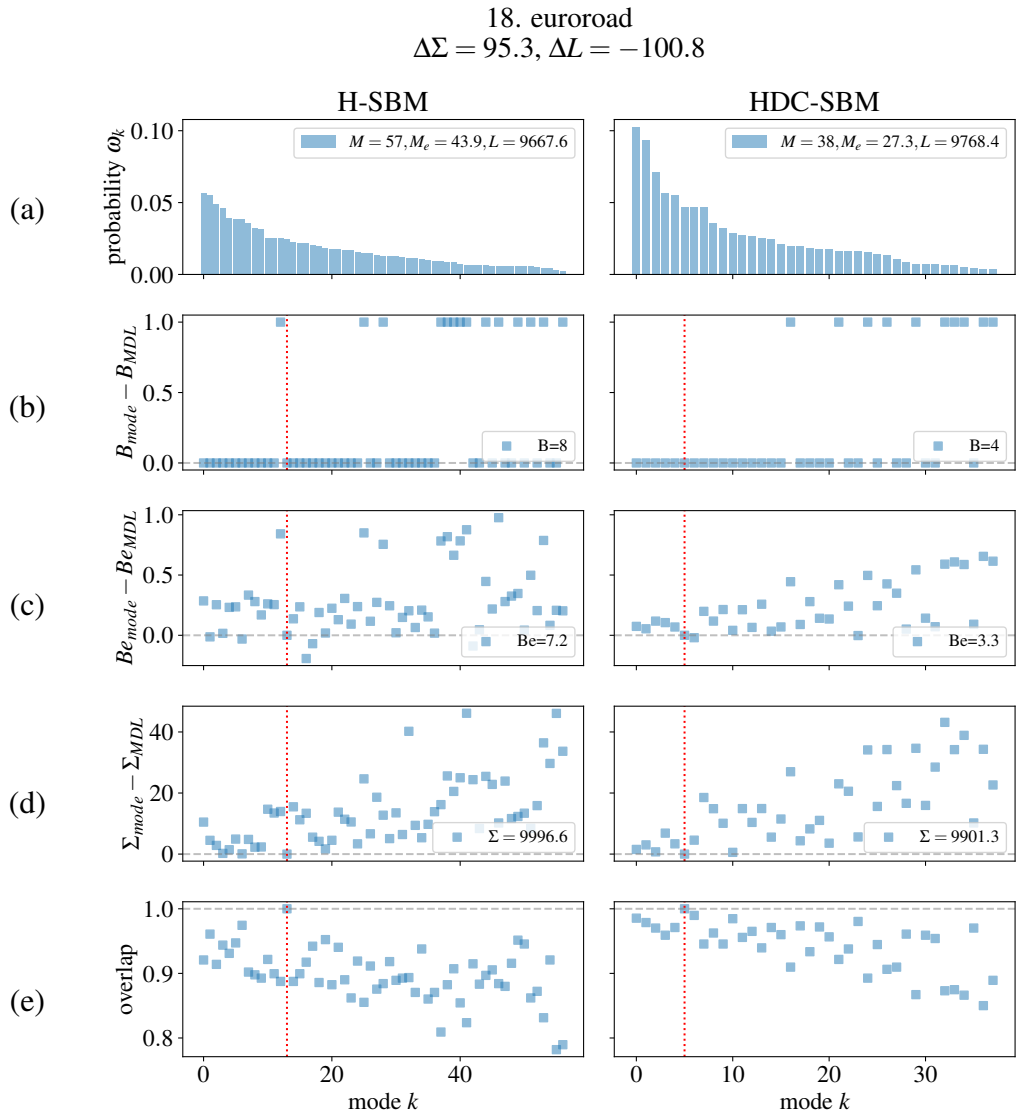


Figure C.13: Summaries of the modes of the posterior distribution of node partitions of *euroroad* network for H-SBM (left) and HDC-SBM (right). (a) Mode fractions ω_k. In the legend, we include the number of modes M, the effective number of modes M_e and the negative log-evidence L. (b) Difference in the number of groups B corresponding to partition modes and the MDL partition. (c) Difference in the effective number of groups B_e corresponding to partition modes and the MDL partition. (d) Difference in description length Σ corresponding to partition modes and the MDL partition. (e) Overlap between partition modes and the MDL partition. For panels (b) to (e), the vertical red line indicates the mode to which the MDL fit belongs. The legend indicates the summary corresponding to the MDL fit.

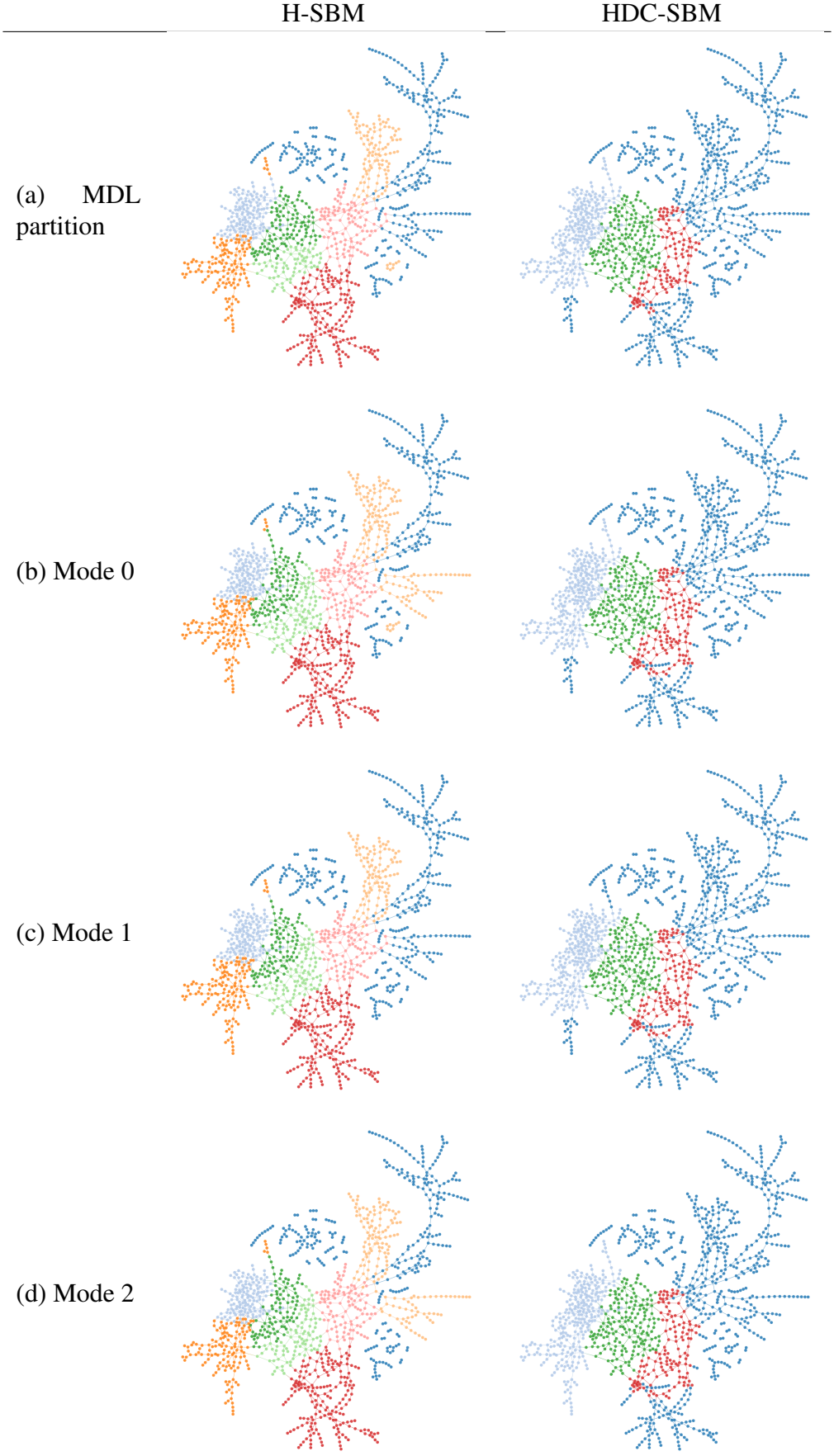


Figure C.14: Several partitions for *euroroad* (a network of international "E-roads", mostly in Europe [194]) obtained with the H-SBM and HDC-SBM from (a) minimizing the description length, (b-d) fitting a mixture model to characterize the posterior distribution of node partitions and obtaining its modes. The three most likely modes are shown here.