DOCTOR OF SCIENCE (DSC)

Realistic constraints, model selection, and detectability of modular network structures

Zhang, Lizhi

Award date:
2023

Awarding institution:
University of Bath

Realistic constraints, model selection, and detectability of modular network structures

submitted by

Lizhi Zhang

for the degree of Doctor of Philosophy

of the

University of Bath

Department of Mathematical Science

January 25, 2023

Contents

List of Figures

Selected notations

$\mathbb{Z}$	set of integers
$\mathbb{R}$	set of real numbers
G	a network
$\mathcal{V}$	a set of vertices (nodes)
$\mathcal{E}$	a set of edges
N	number of vertices (nodes)
E	number of edges
B	number of groups (communities)
$\boldsymbol{A}$	adjacency matrix
$\langle A_{uv} \rangle$	expected number of edges between node u and v
$\boldsymbol{b}$	network partition
$\boldsymbol{e}$	edge count matrix
e_{rs}	number of edges between groups r and s
$\langle e_{rs} \rangle$	expected number of edges between groups r and s
e_r	number of edges attaching to group r, which is equal to $\sum_s e_{rs}$
e_{in}	total number of within-group edges
e_{out}	total number of between-group edges
Σ	description length
$\boldsymbol{n}$	group (community) size vector
n_r	number of nodes in group r
u, v, i, j	indicies for vertices (nodes)
(u, v)	an edge connecting u and v
r, s	indices for groups (communities)
k_u	degree of node u
$\boldsymbol{k}$	degree sequence vector
$\langle k \rangle$	average degree
θ_u	degree propensity of node u
$\hat{\theta}_r$	sum of degree propensities of nodes in group r
$\boldsymbol{\theta}$	degree propensity vector
$\mathcal{B}(\cdot, \cdot)$	Beta function
$\Gamma(\cdot)$	Gamma function
δ	Kronecker delta function
$S(., .)$	Stirling's number of second kind

Abstract

Many real-world systems are complex, consisting of many entities with interactions among them. Our understanding of real-world complex systems has been significantly advanced by modelling these systems as networks. A *network* is a mathematical abstraction of complex systems, representing entities and interactions by nodes and edges. Recent years have witnessed a rapid growth in the demand for analysing networks data, driven by the increased availability of large-scale, quality datasets. A common task in network analysis is to identify the "building blocks" of a network by finding divisions of nodes, such that nodes in the same division connect with the rest of the network in a similar way. This task is often referred to as *community detection* in networks. Community detection methods allow researchers to characterise network data from the perspective of connection pattern, which could convey important information about the functional and evolutionary mechanism of the underlying systems.

Recently, Bayesian inference based on generative network model has attracted great attention as a community detection method, which is mainly due to its principle inference nature and formal implementation of the Occam's razor. However, this method often relies on *general models* that simultaneously account for different kinds of community structure. If the dominant structure in data is in fact restricted and simple, using general models could lead to sub-optimal fit to data.

This thesis concerns with developing Bayesian inference community detection methods that are tailored for a particular kind of structure - the *assortative structure*. A network is said to be assortative if it can be divided into subgroups of nodes, such that connections inside each of division are dense while between distinct divisions are sparse. To this end, we develop the Bayesian formulation of the degree-corrected *planted partition model*. Such model assumes the probability of an edge between a pair of nodes is dependant on whether they are from the same subgroups as well as their node-wise propensity of receiving an edge. This formulation leads to a novel method for

extracting assortative structures and this method is one of the main contributions of this thesis. Compared with other existing methods, our proposed method has the advantage of being robust against overfitting, which means our method will not report spurious community structures in random networks while other non-statistical, heuristic methods usually do. In deriving our proposed method, we clarify on an established equivalence between the popular *modularity maximisation* approach and maximum likelihood inference. Our analysis shows that the equivalence result is tenuous, since it relies on subjective choices of model parameters which lack of principle justifications.

We demonstrate the performance of our proposed method in both synthetic and empirical networks. In particular, we construct a large network corpus consisting of datasets which are diverse in terms of size and density. Using this network corpus, we find evidence that the degree-corrected planted partition model has the ability of achieving better quality of fit in some empirical networks compared to existing models in some cases,. Moreover, the degree-corrected planted partition model has the potential of providing additional insight into data regarding high-resolution community structure. Moreover, by conducting model selection in our network corpus, we find that assortativity is often too simplistic to be the dominant pattern in empirical networks.

Finally, we study the detectability of assortative community structures. In networks where all nodes receive identical number of edges on average, there exists a detectability threshold of the strength of community structure, below which no polynomial algorithms can detect the planted community structure better than random guessing. We conduct a numerical study to examine the effect of heterogeneity in the number of edges attaching to nodes on the detectability of assortative structures. Such effect has been analytically studied in a special case where networks have two equal-size communities. Our results provide further numerical evidence for the existing theoretical analysis and open the door to investigation about the detectability of community structures in more general settings, e.g. in networks consisting of more than two communities, which could have different extents of heterogeneity in degree distribution.

Acknowledgements

I am fortunate to have a group of wonderful people who support me throughout my PhD journey. First and foremost, I want to thank my lead supervisor Tiago Peixoto who introduced me to the field of network science. He is always supportive, patient, and has helped me build confidence in challenging the well-known and exploring the unknown. Thank you Tiago, for your supervision as well as the guidance on making perfect lattes.

I want to thank everyone in the SAMBa CDT at Bath, where I spent most of my PhD time. Special thanks to my co-supervisor Matthew Nunes and the co-director of the SAMBa programme Susie Douglas who have provided valuable advice in different stages of my research. I also want to thank my cohort in the CDT, especially Stefano Bruno for his humour, Jordan Taylor for spontaneous chats over machine learning and his authentic Singaporean curry, and Daniel Ng for our restaurant adventure at Bath.

My gratitude also goes to people I met during my research placement in London. I would like to thank Dr Maria Bazzi and Dr Hao Ni for their encouragement. Many thanks to Jin Wang, Tiejun Wei, Shaoxiong Hu for their company and discussions about basically anything.

Finally, I would not be able to make it this far without the love and support from my family, especially my parents Yongcheng Zhang and Mingmei Li, thank you mum and dad. And thank you Chenxia for always being by my side.

Chapter 1

Introduction

Many real-world systems are complex, consisting of many entities with interactions among them. Examples of complex systems include the human society in which people socialise with each other, the system of human brain where information exchanges among billions of neurones, and power grids where electricity is transmitted between power stations. One common practice of studying complex systems is to map them into *networks*, in which entitles and interactions are represented by *nodes* and *edges*. Networked representation of complex systems has the advantage of removing distractions of fine-level details and allowing us to focus on the connection pattern, which often plays a prominent role in the behaviour of the underlying systems [1]. Our understanding of real-world complex systems has been significantly advanced by modelling them as networks, with examples including explaining social dynamical processes like disease spreading and scientific innovation emergence [2], revealing the underlying mechanism of brains [3], providing guidance for mitigation of failures escalating in infrastructure systems [4], etc. Recent years have witnessed a rapid growth in the demand for analysing network data, driven by the increased availability of large-scale, quality datasets [5–7].

Among other important properties of networked systems, their *community structure* has received considerable interest. Roughly speaking, communities in networks are groups of nodes sharing similar connection patterns. The most widely studied pattern is the *assortativity*. A network is said to have assortative structure if it can be divided into subgroups of nodes such that connections inside each group are denser than that between distinct pair of groups. There are other types of community structure that are of equally importance as well, for example, the core-periphery structure manifests itself

by a densely connected core surrounded by layers of loosely connected periphery nodes. Note that core-periphery structure could consist of more than two communities, with multiple pair of core and periphery groups or multiple layers of periphery nodes [8, 9]. Another commonly studied structure is the bipartite structure, where nodes belong to two intrinsic classes and edges are allowed only between nodes from distinct classes [10]. Community structures in networks are often related to the functional or evolutional mechanisms of the underlying systems. For example, scientific collaboration networks often acquire communities which are well correlated with disciplinary lines [11, 12]. In biological networks like metabolic networks or protein-protein interaction networks, communities structures could represent basic functional modules [13–15]. Motivated by the prevalence of community structures in empirical networks [16–18], much research effort has been devoted to developing methods that divide networks into groups based on the observed network topology. These methods are often known as the *community detection* methods. Community detection has become one of the most fundamental tool for analysing network data. We refer to [19–22] for complete surveys about community detection methods.

Community detection methods have been successfully applied to complex systems from a variety of disciplines, including but not limited to social science [23–26], biology [27–30], economics [31–35] and computer science [36, 37]. In addition to being a powerful tool for analysing natural networked systems, community detection also receives interests because it is getting common to construct similarity graphs of any kind of data, such as text [38], images [39] and time series [40], with a hope to expose the relation pattern in them. Then, finding the community structures in similarity graphs can be useful for downstream applications. For example, outcomes of community detection methods can be fed to machine learning algorithms as a feature of the original data [41]. A specific application of community detection is the design of recommendation systems, where a network of users and items to be recommended is constructed and inferred community structures are used to assist making recommendation [42, 43]. More examples of applications of community detection methods can be found in [44, 45].

One of the most widely used method for community detection is the *modularity maximisation* [46]. This method aims to identify assortative structures by assigning nodes into groups such that the number of within-group connections exceeds that in a statistical null model as much as possible. Despite modularity maximisation being intuitive to comprehend and easy to implement, it has several limitations. Firstly, modularity maximisation is notorious for *overfitting* data, i.e. it claims presence of community

structures in networks which are known be non-modular, e.g. in tree-like networks, in lattices [47] and even in fully random networks [48]. To address this problem, Zhang and Moore [49] proposed a message-passing algorithm to search for a consensus of many partitions with high-modularity values. The idea is to construct a Gibbs distribution of network partitions, with the energy function of partitions being dependant on their modularity values. Then, nodes are labeled according to the marginal distributions of the Gibbs distribution, which in essence gathers information from many partitions with high modularity values. In random networks, there could exist a large amount of high-modularity partitions which are uncorrelated with each other. As a result, random networks admit no meaningful consensus. Applying the message-passing algorithm to random networks will return uniform marginal distributions, indicating the absence of statistically significant structures. However, in networks with known statistically significant structure, this method might mistakenly conclude that no consensus exists. Whether the correct structure can be correctly detected depends on the value of an inverse temperature being used in the construction of the Gibbs distribution, which is generally unknown in practice [50].

Besides overfitting, modularity maximisation paradoxically has the risk of *underfitting* data. That means, the method might conclude overly simplistic structures compared to the actual pattern in data. As a result of the undesired tendency of underfitting, modularity maximisation has an intrinsic limit on the number of detectable communities, which grows with the size of networks. If the number of communities in a network is above this limit, then modularity maximisation will fail to identify all of the communities, often merging small communities into large ones regardless how significant the small-size communities are. This observation is often referred to as the *resolution limit* problem [51], since the inferred structure is deficient in resolution in the sense that only coarse-level structures can be recovered and detailed structural information will be missing.

There have been many attempts to resolve the resolution limit of modularity maximisation. Arenas et al. [52] proposed a multi-resolution detection method, which relies on adding self-loops with a weight to each node in the original network. The weight of self-loops is a tunable parameter, representing our subjective bias toward the resolution of structure to be detected. The partition is then given by optimising the modularity function over the modified network. Another similar but different suggestion given by Reichardt and Bornholdt [53] is to introduce a resolution parameter, which adjusts the contribution from the null model in the original definition of modularity. One common property of these two modified modularity measures is that they both require subjec-

tive choice of parameters as input. Unfortunately, we often do not have the knowledge about what are proper choices of related parameters, and more importantly, there is generally not a single optimum value to use [50, 54]. Although it is common to try out many different resolution parameters, that often leads to a large amount of competing results and it is not clear how to select from them. The modularity density [55, 56] is another quality metric for finding partition that shows the ability of addressing resolution limit and does not require subjective choice of parameter. However, none of these variants of modularity are ideal solutions, because all of them are heuristic approaches, lacking of theoretical ground. In addition, all of these modularity-based methods fail to consider statistical significance of their results, thereby suffering from the overfitting problem just as the original modularity does [50, 54].

Finally, modularity maximisation implicitly assumes that every community is "similar" to each other. In particular, Newman [57] showed that maximising modularity is equivalent to applying maximum likelihood principle (MLE) with a particular generative network model, which assumes the numbers of edges inside communities are identical for different communities. Moreover, unlike other MLE approach in [58, 59] which involve extra parameters controlling the size of communities, modularity maximisation methods do not consider this perspective and therefore implicitly assumes the sizes of communities is uniform. These limitations are rooted in the very definition of the modularity measure and shared by all of its variants [50, 57]. For the reasons stated above, modularity-based methods will bias toward regular assortative structures, where the number of edges and nodes in each community are similar. Therefore, when communities have non-uniform sizes, modularity-based methods are expected to have degenerate performance. Due to their restricted modelling capacity and tendency of overfittng and underfitting data, despite their popularity, modularity-based methods are not reliable solution for community detection.

More recently, the Bayesian inference approach for community detection based on generative models has gained great attention, mainly due to its ability to provide principled inference and its built-in Occam's razor effect of preventing overfitting [60]. The idea is to construct some generative networked models, which allow us to generate networks with desired community structures. For an observed network, we assume it is a sample generated from our model. Then, the community detection problem becomes a model inference problem and we can solve the inference problem by fitting models to data. Specifically, stochastic blockmodels (SBMs) [61] are arguably the most commonly applied models in this context. In particular, among the many variants of SBMs, the degree-corrected stochastic blockmodel (DC-SBM) is one of the most widely variant

due to its capability of modelling heterogeneous degree distribution in real world networks as well as its ease for analysis. According to SBMs, communities in networks are groups of nodes that are statistically equivalent. That means, nodes from the same group have the same probability of being connected to the rest of the network. This definition of community is a general one, encapsulating a series of commonly studied structures and even the mixture of them. Therefore, Bayesian inference based on SBMs is a versatile approach compared to modularity maximisation, since it is able to recover not just assortativity, but many different kinds of community structures simultaneously, as long as they exist in data.

Compared to modularity-based methods, the main advantage of Bayesian inference with SBMs is that it will not overfit data. This is because each possible model will be given a probability weight, in a way that complicated models will be penalised for model complexity. As a result, a complicated model will not be favoured over a simple one, unless using the complicated model can bring significant reward in the model fit. As a result, unlike other non-statistical approaches, Bayesian inference with SBMs will not claim spurious communities in fully a random network. In addition, although the Bayesian inference approach with SBMs can also run into the underfitting problem [62, 63], the root cause of this problem has been well understood. The underfitting problem occurs to the Bayesian inference approach when a naive choice of uninformative prior is used. This problem can be circumvented by employing a hierarchical prior [63], leading to the *hierarhical* or *nested* variant of SBMs. Overall, Bayesian inference approach based on SBMs should be preferred over modularity-based methods, since it is robust against overfitting and underfitting, and has the ability of revealing different types of community structure.

The versatility of the Bayesian inference approach based on general SBMs is a strength, because it allows practitioners to be agnostic about what kind of structure to be inferred. However, general models are suboptimal when it comes to networks where a particular kind of structure dominates. This is due to exactly the Bayesian Occam's razor effect, which states that complicated models should not be preferred over simpler ones, given their quality of fit being equal. Moreover, with only general models at hand, we will not be able to decide which particular structure dominates in data, which is of great importance for characterising the underlying systems. As a result, there is a pressing need for *restricted* models which are tailored for particular kinds of community structure. There have been works on Bayesian inference with restricted variants of SBMs focusing on bipartite [10] and core-periphery structure [64]. However, an assortative-constrained variant is still lacking. The main motivation of this thesis is to

fill this gap by developing a Bayesian formulation of an assortative constrained variant of general SBMs, which leads to a novel method for detecting statistically significant assortative structures in networks. We will provide details of the method in Chapter 3. In Chapter 4, we will further study the overfitting and underfitting properties of our proposed method in a large set of empirical networks.

In addition to the development of useful model variants, there is another line of research concerning the detectability of community structure in networks. Research in this direction exploded after the work by Decelle *et al.* [65] in which a *phase transition* phenomenon was shown in community detection: Networks generated from SBMs show different phases which are related to the detectability of structures used in the data generating process. Decelle *et al.* provided an estimate of the position where the transition happens through a stability analysis of the belief propagation (BP) algorithm for SBMs. Based on their analysis, Decelle *et al.* conjectured that there exists a non-trivial detectability threshold of the strength of community structure, below which networks generated from SBM are in an *undetectable phase* and no polynomial algorithms can perform better than a random guess. Moreover, the position of the detectability threshold was found to be related to the average number of edges attaching to each node, which is known as the *average degree* of the network. The estimate of the position of the detectability threshold implies that, as the average degree increases, the detectability threshold is expected to decrease, i.e. community detection becomes easier as we observe more data (i.e. more edges sampled according to the model) generated from the model. Such conjecture regarding the detectability of community structures has inspired a series of theoretical works that have provided rigorous proofs of the detectability threshold [59,66,67]. These results are unexpected and remarkable, because they tell us that structural information which is fundamentally different from randomness might remain undetectable, regardless of which inference algorithms we use.

When the phase transition phenomenon was firstly reported in [65], the authors focused on symmetric communities, where nodes from each community receive identical number of edges on average. In other word, the average degree of each community is the same across the entire network. Nevertheless, this assumption is rather strict and communities are more likely to be asymmetric in empirical networks. Intuitively, symmetry of communities makes the detection problem more challenging than when the symmetry breaks. This is because if nodes from different communities have different degrees, then the discrepancy in degree provides extra local information that can assist us in inference. The detectability of asymmetric communities was studied by Zhang

and Moore in [68]. The phase transition phenomenon was found to *disappear* as the asymmetry of communities increases, i.e. we can find partitions of nodes that are positively related to the planted community structures as long as they exist. For the same reason, phase-transition does not appear in the task of detecting core-periphery structure [69], in which nodes in the core naturally acquire higher degree than those in the periphery. The detectability phase-transition also disappears in the semi-supervised community detection task, in which the correct labelling of a fraction of nodes is given as *a priori* knowledge [70].

All of previous studies regarding the detectability of SBMs assumed that the degree distribution of nodes within communities are homogeneous, which means the expected number of edges connecting to each node inside each community is the same. However, real-world networks often acquire heterogeneous distribution, which are commonly in the form of power-law [71]. In [72], Massoulé et al. derived rigorously the detectability threshold of community structure in networks generated from SBMs, when the degree distribution is heterogeneous. It turns out that the detectability threshold will vanish as the second moment of the degree distribution diverges, i.e. the detectability phase-transition will disappear as in the unequal average degree setting and the semi-supervised setting. However, such theoretical result only applies to networks with two communities. Extending such result to more general setting - e.g. in networks with more than two communities - with the same theoretical technique is hard. In order to explore the complete picture of the detectability phase-transition, as will be explained in Chapter 5, we instead adopt the numerical approximation methods proposed in [65] by Decelle et al. and [73] by Yan et al. Our numerical results are consistent with existing theoretical result in [72] and will serve as a stepping stone to better understanding the detectability phase-transition in networks with heterogeneous degree distributions.

1.1 Contributions of the thesis

My motivation during my PhD project is to study community structures in networks with a focus on assortative structures, using tools from statistical physics and Bayesian inference. The objective of this thesis is to develop a novel method for extracting assortative structures from networks without overfitting or underfitting data. We summarise the main original results in the order of their presence in the rest of this thesis.

1. *We develop a Bayesian formulation of the degree-corrected planted partition model (PP model) - an assortative-constrained variant of SBMs - which leads to a novel algorithm for detecting assortative structures in networks.* Compared to Bayesian

inference with general SBMs, our proposed method focuses on assortative structures and will achieve better performance than general models if the dominant structure in data is indeed assortativity. At the same time, PP model has the ability to recover more general assortative structures than modularity-based methods do. Specifically, unlike modularity-based methods which are restrictive to regular assortativity, PP model has the potential to resolve heterogeneous assortativity, which means the sizes of communities can vary across the entire network. In deriving our proposed method, we also look into an established result regarding the equivalence between MLE with PP model and the modularity maximisation approach [57]. It turns out this equivalence is tenuous and we discuss to what extent the equivalence holds, which is crucial but has not been examined in the literature. These results have been summarised in this paper

> Statistical inference of assortative community structures, Lizhi Zhang and Tiago P. Peixoto, *Physical Review Research*, 2020.

We shall explain these results in details in Chapter 3.

2. *We study the underfitting properties and the quality of fit to data of our proposed PP model by conducting a meta study of empirical networks.* With the Bayesian formulation of PP model at hand, we conduct model selection to find out whether assortativity is the dominant pattern in data. We do the comparison between our PP model and general SBMs, as well as the modularity maximisation method. Our results show that although general models achieve the best fit most of the time, in several illuminating examples, our assortative-constrained variant can achieve better fit than its general counterparts. We also show that our proposed approach is free from the underfitting problem of DC-SBM and modularity maximisation, and investigate the extent to which underfitting problem happens in practice. It turns out that DC-SBM systematically underfits compared to the nested version of DC-SBM (Nested DC-SBm) in our network corpus. In networks where assortativity is the dominant pattern, our PP model shows the ability to recover much detailed structures than both DC-SBM and Nested DC-SBM. We shall present these results in Chapter 4.

3. *We adapt the BP algorithm for DC-SBM and use the algorithm to investigate the effect of heterogeneous degree distribution on the detectability of community structures.* Specifically, in networks generated from the PP model, we observe that as the heterogeneity in degree distribution increases, the detectability threshold decreases and the area of undetectable phase of the model shrinks. These results

are detailed in Chapter 5.

1.2 Organisation of the thesis

The remaining chapters of this thesis are organised as follows.

In Chapter 2, we shall define the SBM and several its variants and explain how to conduct Bayesian inference with them for community detection.

In Chapter 3, we shall focus on the detection of assortative structures in networks. We provide the Bayesian formulation of the PP model and demonstrate its use in synthetic and empirical networks. We shall also clarify the established equivalence between the modularity maximisation approach and MLE for the PP model.

In Chapter 4, we characterise the underfitting and overfitting behaviour of different variants of SBMs by fitting them to a large empirical network corpus. By conducting model selection, we also show that our proposed assortative-constrained model achieve better quality of fit compared to general SBMs in networks where the assortative structure dominates.

In Chapter 5, we shall turn to the detectability phase-transition in community detection. We firstly introduce the belief propagation algorithm for community detection with SBMs and review existing results of detectability in community detection. We then demonstrate the effect of heterogeneity in degree distribution on the detectability phase-transition by applying the BP algorithm to networks with heterogeneous degree distribution.

In Chapter 6, we summarise the contributions of this thesis and outline potential directions for future work.

Chapter 2

Background

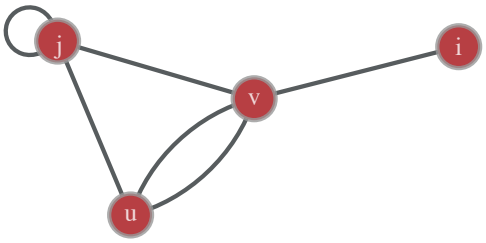
This chapter aims to provide a basic introduction to the Bayesian inference approach for community detection in networks. We will define the stochastic blockmodel and a few of its variants that will be used in the following chapters. Having specified the models, we then explain how to use them to infer community structure in networks, following closely the steps in [60,74]. Specifically, for a given model, we need to derive the posterior probability distribution of all possible network partitions given the observed network. Although posterior distributions rarely permit direct sampling or maximisation, we can still make useful inference from them using numerical approximation algorithms.

We start with notations in Section 2.1, then define stochastic blockmodels in Section 2.2. In Section 2.3, we use the degree-corrected stochastic blockmodel as an example to explain how to derive the posterior probability distribution, followed by Section 2.4 where we introduce another two variants of the stochastic blockmodel, i.e. the microcanonical and the nested variant. In Section 2.5, we explain the numerical algorithm for drawing samples from posterior distributions. Section 2.6 considers comparing different model variants under the Bayesian framework and we wrap up this chapter in Section 2.7 with concluding remarks and motivations for the next chapter.

2.1 Preliminaries

2.1.1 Networks

We introduce some terminologies to use throughout the thesis. A *network*, or a *graph*



$$\boldsymbol{A} = \begin{array}{c} \\ \text{u} \\ \text{v} \\ \text{i} \\ \text{j} \end{array} \begin{array}{c} \begin{array}{cccc} \text{u} & \text{v} & \text{i} & \text{j} \end{array} \\ \begin{pmatrix} 0 & 2 & 0 & 1 \\ 2 & 0 & 1 & 1 \\ 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 2 \end{pmatrix} \end{array}$$

Figure 2-1: A toy example of network G with $N = 4$ nodes and $E = 6$ edges. Elements in the adjacency matrix $\boldsymbol{A} = \{A_{uv}\}_{4\times 4}$ are the number of edges between nodes. The set of nodes is $\mathcal{V} = \{u, v, i, j\}$ and the set of edges is $\mathcal{E} = \{(u, v), (u, j), (j, j), (j, v), (v, i)\}$. The degree of nodes are $k_u = 3, k_v = 4, k_i = 1, k_j = 4$.

is a tuple $G = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ is the node set of size $N = |\mathcal{V}|$, and $\mathcal{E}$ is the list of edges of size $E = |\mathcal{E}|$. We denote elements in the node set $\mathcal{V}$ by u, v or i, j, and elements in the edge set $\mathcal{E}$ by tuples in the form of (u, v). A network is *directed* if its edges have directions. In this thesis, we will restrict ourselves to *undirected* networks where edges have no directions.

An undirected network G can be uniquely encoded by an *adjacency matrix* $\boldsymbol{A} = \{A_{uv}\} \in \mathbb{Z}^{N\times N}$ whose entries A_{uv} represent the number of edges between node u and v if $u \neq v$, or twice that number if $u = v$ for convention. This convention allows us to write the number of half-edges[1] adjacent to node u as follows

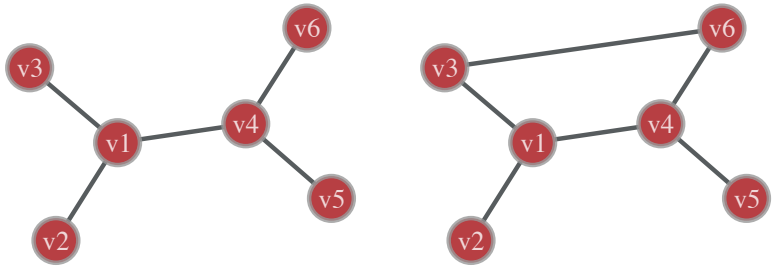
$$k_u = \sum_{v}^{N} A_{uv}. \tag{2.1}$$

The value of k_u is called the *degree* of node u. An example of network and its adjacency matrix is given in Fig 2-1. We denote by $\langle k \rangle$ the average degree of a network, which can be computed as $\langle k \rangle = 2E/N$.

A *walk* in a network consists of a sequence of edges $\{(u_n, v_n)\}$, such that the end node of the previous edge is the same as the starting node of the next edge, i.e. $v_n = u_{n+1}$. The *length* of a walk is the number of edges required to form the walk. A *cycle* is a special walk that needs to satisfy the following conditions: it starts and ends at the same node; its length is larger or equal to three; the walk does not visit any of its nodes twice except for the starting (ending) node.

An undirected network contains no cycles[2] is called a *tree*. For example, in Fig. 2-2(a),

[1] A self-loop of a node contributing two half-edges attaching to the node
[2] No cycles nor self-loops.



(a) (b)

Figure 2-2: (a) An example of tree with 5 nodes (b) Adding an edge between node $v3$ and $v6$ in the tree (a) destroys the tree structure and forms a cycle of length 4, consisting of edges $\{(v1, v3), (v3, v6), (v6, v4), (v4, v1)\}$.

we show an example of tree with 5 vertices. In Fig. 2-2(b), we add an edge between node $v3$ and $v6$, then the resulted graph is not a tree anymore, since the new edge leads to a cycle of length 4.

There are other kinds of network representation which are used for specific problems. For example, it is not uncommon to see *weighted networks*, where edges are equipped with weights, representing attributes like counts, distance, or strength [75, 76]. *Multiplex networks* are networks with multiple layers, in which all layers share the same set of nodes but acquire different set of edges [77]. *Temporal networks* with time-dependent nodes and edges also consist of multiple layers, but the layers are strictly ordered according to the time-stamp of layers [78, 79]. In addition to mapping systems' connection pattern, sometimes it is useful to integrate the *attributes*, or *properties* associated with nodes and edges [80, 81] into analysis as well. Although these modelling tools are of great importance in practice, they are beyond the scope of this thesis. We will focus on *undirected, static, single-layered* networks *without* any associated properties of nodes or edges.

2.1.2 Network partition

The main theme of this thesis is identifying community structures in networks. The community structure of a network is commonly described by a *partition* of the network, which assigns nodes to non-overlapping groups. A partition of a network of size N can be represented by a vector $\boldsymbol{b} = \{b_u\} \in \mathbb{Z}^N$, where $b_u \in \{1, 2, .., B\}$ represents to which group node u belongs and B is the corresponding number of groups. As an example, we visualise a partition of the Grevy's zebra network [82] in Fig. 2-3, where the colouring of nodes indicates a specific partition of the network. This partition is inferred by fitting the planted partition model, which will be introduced in the Chapter 3.

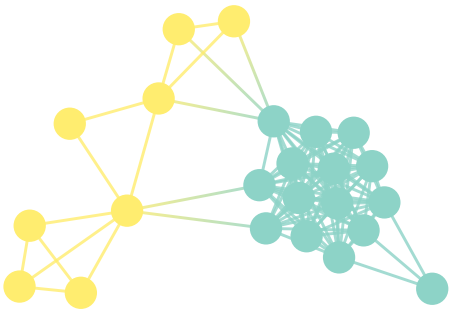


Figure 2-3: A social network of interactions among zebras with 23 nodes and 105 edges [82]. Each node represents a Grevy's zebra and an edge between two nodes means the two zebras were observed to appear together in the field study. The colouring of nodes indicates the network partition $\boldsymbol{b} = (1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 2, 2)$, which is inferred by fitting the planted partition model [83].

The layout of the visualisation in Fig. 2-3 is obtained with a *force-directed* graph drawing algorithm [84]. At a high level, the algorithm assumes that there exist attractive forces between connected nodes and repulsive forces among all nodes in the network. Then, the positions of nodes and edges are determined by minimising the energy of the system. For the rest of network visualisations in this thesis, we will use the same drawing algorithm unless otherwise stated.

For a given network partition $\boldsymbol{b}$ of a network, we denote by $\boldsymbol{n} = \{n_r\} \in \mathbb{Z}^B$ the number of nodes in each group (or community). The matrix $\boldsymbol{e} = \{e_{rs}\} \in \mathbb{Z}^{B\times B}$ is the *edge count* matrix, in which entries e_{rs} counts the number of edges between group r and s. We define $e_r = \sum_s e_{rs}$ which is the number of edges attaching to group r. For example, for the zebra network in the Fig. 2-3, we have $B = 2$ and

$$\boldsymbol{n} = (15, 8), \quad \boldsymbol{e} = \begin{pmatrix} 176 & 5 \\ 5 & 24 \end{pmatrix}, \tag{2.2}$$

and

$$e_1 = 176 + 5 = 181, \quad e_2 = 5 + 24 = 29. \tag{2.3}$$

Sometimes it is useful to consider *overlapping partitions* of networks where nodes can belong to multiple communities at the same time [85, 86]. Although the discussion in this thesis will be exclusively devoted to non-overlapping setting, our main results and conclusions should hold in overlapping setting as well and we expect to work toward this direction in the future.

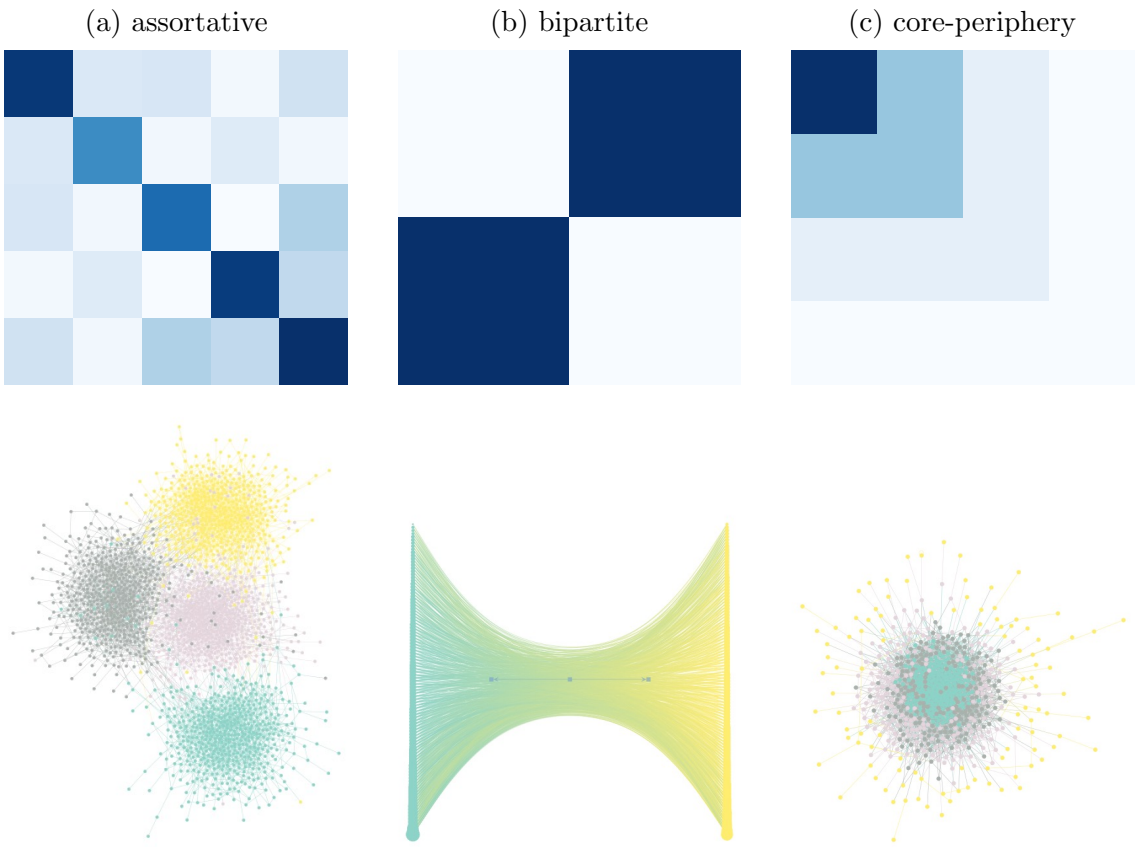
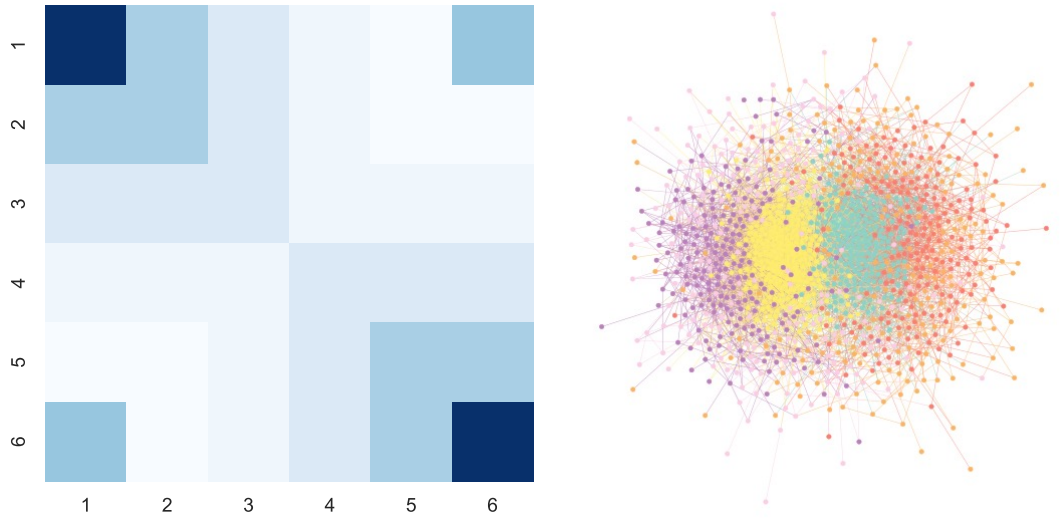


Figure 2-4: Examples of networks with community structures generated from the family of stochastic blockmodels with **(a)** assortative **(b)** bipartite structure and **(c)** core-periphery. Colouring of nodes indicate the community membership of nodes that are used to generate the observed networks.

2.2 Stochastic blockmodels

The stochastic blockmodel (SBM) is a generative model that allows generation of network samples that contain community structure. Since the origin of the model in [61], many different variants of the SBM have been developed but sharing the same spirit. Specifically, the family of SBMs assumes nodes in a network belong to groups (or communities) and edges are independently placed among nodes following some probability distributions. The probability distributions of edge occurrence depend on the group membership of nodes, leading to modular structures in networks sampled from the model. This generating process also implies what constitutes a community in the SBMs sense: a community is a subgroup of nodes which are stochastically equivalent, i.e. they have the same probabilities of being connected to the rest of the network.



(a) (b)

Figure 2-5: A toy example of network with a mixture of assortative and core-periphery structures. As indicated by the affinity matrix in subfigure (a), the probability of edges is relatively large inside the each of the group 1 and 6 but small between them. In addition, groups 2 and 3 are periphery surrounding group 1 while group 6 is surrounded by group 4 and 5.

We begin with the stochastic blockmodel for simple networks. A network is *simple* if it contains no multiple edges and no self-loops. Therefore, in a simple network, the number of edges between any two nodes is either 0 or 1. Following the idea of SBMs, for any two nodes u and v in the network, we assume the probability of an edge is $p_{b_u b_v}$, and non-edge is $1 - p_{b_u b_v}$. The matrix $\boldsymbol{p} = \{p_{rs}\} \in [0,1]^{B\times B}$ is known as the *affinity matrix*, playing the role of determining the modular structures of network samples. For example, we can generate networks with assortative structures by choosing diagonally dominant affinity matrices satisfying $p_{rr} > p_{rs}$ for $r, s \in \{1, 2, .., B\}$. Similarly, one can generate networks with other kinds of community structure like bipartite, core-periphery or even the mixture of them by tuning the affinity matrix (see Fig. 2-4 and Fig. 2-5). The probability of generating an observed network G with adjacency matrix $\boldsymbol{A} = \{A_{uv}\} \in \mathbb{Z}^{N\times N}$ from the model we just describe is

$$P(\boldsymbol{A}|\boldsymbol{p}, \boldsymbol{b}) = \prod_{u<v} p_{b_u b_v}^{A_{uv}} (1 - p_{b_u b_v})^{1-A_{uv}}. \tag{2.4}$$

We will refer to this model as the *Bernoulli SBM* in the rest of the thesis, since the generating process above implies that the counts of edges between nodes are independent Bernoulli random variables.

The Bernoulli SBM has one major shortcoming which restricts its use in practice. Real-world networks often possess heterogeneous degree distribution [71], i.e. the number of edges attaching to each node can vary significantly. However, degree distributions have little variation in networks generated by the Bernoulli SBM. Indeed, the Bernoulli SBM implicitly assumes that the expected degree of nodes from the same communities are identical. To see this, under the assumption that the size of each community is sufficiently large, the expected degree of any node u is approximately a Poisson variable with mean $\langle k_u \rangle$, which has the following expression

$$\langle k_u \rangle = \sum_{v}^{N} \langle A_{uv} \rangle = \sum_{r}^{B} n_r p_{b_u r}, \tag{2.5}$$

where $\langle A_{uv} \rangle$ is the expected number of edges between node u and v. Since $\langle k_u \rangle$ depends on the node-wise index u only via its community membership b_u, any two nodes u and v from the same community $r = b_u = b_v$ then have identical expected degree $\langle k_u \rangle = \langle k_v \rangle$. As a result, the degree distribution is expected to be homogeneous inside each community. For this reason, the Bernoulli SBM is often inadequate for modelling empirical networks.

A better alternative to Bernoulli SBM for modelling real-world networks is the *Degree-Corrected SBM* (DC-SBM) [87], which relies on the *Poisson SBM*. The Poisson SBM assumes the number of edges between any two nodes are Poisson rather than Bernoulli random variables

$$A_{uv} \sim \mathrm{Poi}(\lambda_{b_u b_v}), \tag{2.6}$$

where $\lambda_{b_u b_v}$ is the probability of an edge between two nodes from group b_u and b_v respectively (or twice that number if $u = v$)[3]. The likelihood of the Poisson SBM for an observed network $\boldsymbol{A}$ reads as

$$P(\boldsymbol{A}|\boldsymbol{\lambda}, \boldsymbol{b}) = \prod_{u<v} e^{-\lambda_{b_u b_v}} \frac{\lambda_{b_u b_v}^{A_{uv}}}{A_{uv}!} \prod_{u} e^{-\lambda_{b_u b_v}/2} \frac{(\lambda_{b_u b_v}/2)^{A_{uu}/2}}{A_{uu}!!}. \tag{2.7}$$

To accommodate the degree heterogeneity, the DC-SBM extends the Poisson SBM by introducing a *degree propensity* parameter $\boldsymbol{\theta} = \{\theta_u\}$ for each node in the network. The number of edges between two nodes u and v is still a Poisson random variable, but the mean of the Poisson variable changes to $\theta_u \theta_v \lambda_{b_u b_v}$, which means

$$A_{uv} \sim \mathrm{Poi}(\theta_u \theta_v \lambda_{b_u b_v}). \tag{2.8}$$

[3]This is due to the aforementioned convention A_{uu} is twice the number of self-loops connecting to node u

The extra degree propensity parameter allows us to control the number of edges connecting to each node, since the expected degree of a node u under the DC-SBM is

$$\langle k_u \rangle = \sum_v \theta_u \theta_v \lambda_{b_u b_v} = \theta_u \sum_v \theta_v \lambda_{b_u b_v}. \tag{2.9}$$

The likelihood of DC-SBM is then

$$P(\boldsymbol{A}|\boldsymbol{\theta}, \boldsymbol{\lambda}, \boldsymbol{b}) = \prod_{u<v} e^{-\theta_u \theta_v \lambda_{b_u b_v}} \frac{(\theta_u \theta_v \lambda_{b_u b_v})^{A_{uv}}}{A_{uv}!} \prod_u e^{-\theta_u^2 \lambda_{b_u b_v}/2} \frac{(\theta_u^2 \lambda_{b_u b_v}/2)^{A_{uu}/2}}{A_{uu}!!}. \tag{2.10}$$

Note that the model has not been fully defined yet, since in the last expression, θ_u and $\lambda_{b_u b_v}$ always appear in the form of the product $\theta_u \theta_v \lambda_{b_u b_v}$, leaving a freedom of the multiplying constant to be fixed. If every θ_u is multiplied by a factor C, the probability of generating a network $\boldsymbol{A}$ is unchanged if we decrease every λ_{rs} by a factor C^2. To fully determine the model, it is convenient to define the quality $\hat{\theta}_r = \sum_u \delta_{r b_u}, \forall r \in \{1, 2, .., B\}$ where δ is the Kronecker delta function and impose the following constraints

$$\hat{\theta}_r = 1, \quad \forall r \in \{1, 2, .., B\}. \tag{2.11}$$

This choice leads to straightforward interpretations of model parameters $\boldsymbol{\theta} = \{\theta_u\}$ and $\boldsymbol{\lambda} = \{\lambda_{rs}\}$: θ_u is the probability of choosing node u from its group b_u, and λ_{rs} is the expected number of edges between group r and s (or twice that number of $r = s$), since

$$\langle e_{rs} \rangle = \frac{1}{2} \sum_{u,v} \langle A_{uv} \rangle = \frac{1}{2} \sum_{u,v} \theta_u \theta_v \lambda_{b_u b_v} \delta_{r b_u} \delta_{s b_v} = \hat{\theta}_r \hat{\theta}_s \lambda_{rs} = \lambda_{rs}. \tag{2.12}$$

Notice that the non-degree-corrected Poisson SBM can be viewed as a special case nested within the DC-SBM with parameters $\boldsymbol{\theta} = \{\theta_u\}$ being uniform within each community, i.e. under the constraint in equation (2.11), the non-degree-corrected model is equivalent to setting $\theta_u = 1/n_{b_u}$, with n_r being the number of nodes in community r.

2.3 Bayesian inference: the posterior probability of the DC-SBM

With an observed network $\boldsymbol{A}$, our goal is to make inference about the community structure $\boldsymbol{b}$, assuming that the network is generated from a SBM. In this section we explain how to compute the posterior probability distribution of the DC-SBM, following the steps in [60]. Although these results have been presented in the literature, they

are worth a review since we will make use of similar ideas and techniques in Chapter 3, where we shall introduce a the main contribution of the thesis.

To begin with, the *Bayes' rule* allows us to express our knowledge about the unknown community structure in a succinct formula,

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

The $P(\boldsymbol{b}|\boldsymbol{A})$ is called the *posterior* probability distribution, which represents the uncertainties of the unknown community structure conditioned on the network (data) we observe. In words, the Bayes' rule says that the status of our understanding is a combination of our prior knowledge, represented by the *prior* distribution $P(\boldsymbol{b})$, and the information presented in the observed data, represented by the *likelihood* function $P(\boldsymbol{A}|\boldsymbol{b})$. Sometimes the likelihood $P(\boldsymbol{A}|\boldsymbol{b})$ is also referred to as the *marginal likelihood* to emphasise the fact that it requires marginalisation of all model parameters $\boldsymbol{\Theta}$ except for the community structure $\boldsymbol{b}$,

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

The term $P(\boldsymbol{A})$ is called the *model evidence*

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

which serves as a normalising constant. Although $P(\boldsymbol{A})$ is generally intractable, because $P(\boldsymbol{A})$ is the same for every possible partition $\boldsymbol{b}$, most of the time it is sufficient to conduct inference of the community structure $\boldsymbol{b}$ as long as we can evaluate $P(\boldsymbol{b}|\boldsymbol{A})$ up to its normalising constant $P(\boldsymbol{A})$. Hence, the necessary pieces for inference are the expressions of the marginal likelihood $P(\boldsymbol{A}|\boldsymbol{b})$ and the prior distribution $P(\boldsymbol{b})$. Here, we take the DC-SBM as an example to explain how to obtain the posterior probability (up to its normalising constant) of the community structure.

Marginal likelihood for the DC-SBM

In DC-SBM, except for the network partition $\boldsymbol{b}$, the model takes the average number of connections between groups $\boldsymbol{\lambda} = \{\lambda_{rs}\}$ and the degree propensity of nodes $\boldsymbol{\theta} = \{\theta_u\}$ as parameters. Assuming that $\boldsymbol{\lambda}$ and $\boldsymbol{\theta}$ are conditionally independent given $\boldsymbol{b}$, our goal

is then to compute the following integral to derive the marginal likelihood,

$$P(\boldsymbol{A}|\boldsymbol{b}) = \int P(\boldsymbol{A}, \boldsymbol{\lambda}, \boldsymbol{\theta}|\boldsymbol{b}) d\boldsymbol{\lambda} d\boldsymbol{\theta} = \int P(\boldsymbol{A}|\boldsymbol{\lambda}, \boldsymbol{\theta}, \boldsymbol{b}) P(\boldsymbol{\lambda}|\boldsymbol{b}) P(\boldsymbol{\theta}|\boldsymbol{b}) d\boldsymbol{\lambda} d\boldsymbol{\theta}. \tag{2.16}$$

The $P(\boldsymbol{\theta}|\boldsymbol{b})$ and $P(\boldsymbol{\lambda}|\boldsymbol{b})$ are prior distributions of parameters $\boldsymbol{\lambda}$ and $\boldsymbol{\theta}$ respectively, conveying our knowledge about the model before any data is observed. Prior distributions can be set to incorporate *a priori* knowledge into inference, for example results from previous studies or experts' domain knowledge. However, network data often comes as a single, unique object (e.g. the World Wide Web, power station transmission grid), in which no past knowledge can be directly applied to present analysis. In this case, it is reasonable to choose "uninformative" priors, which are as agnostic about the pattern in data as possible, such that any significant pattern in data can be fully revealed without being suppressed by subjective bias. To this end, we will choose prior probability distributions according to the *maximum entropy principle* [88, 89]. Intuitively, because entropy describes the amount of uncertainty in a probability distribution, maximum entropy distributions then are the least informative choices for priors and therefore meet our requirement the best. Maximum entropy priors are derived by solving a constrained optimisation problem with the following form,

$$\begin{aligned} &\underset{p(X)}{\text{maximise}} \quad -\int_X p(X) \ln p(X) dX \\ &\text{subject to} \int_X p(X) M_i(X) dX = c_i \ \text{ for all constraints } M_i. \end{aligned} \tag{2.17}$$

The zeroth moment constraint

$$\int_X p(X) = c_0 = 1, \tag{2.18}$$

is a must to make sure the sum of probability is equal to one.

To derive the maximum entropy prior for the degree propensity parameter $\boldsymbol{\theta}$, we need to optimise the following Lagrangian function w.r.t. $p(\boldsymbol{\theta})$

$$\begin{aligned} L(p(\boldsymbol{\theta}), \xi_0) = &-\int_{\boldsymbol{\theta} \in \boldsymbol{C}} p(\boldsymbol{\theta}) \ln p(\boldsymbol{\theta}) d\boldsymbol{\theta} + \xi_0 \left(\int_{\boldsymbol{\theta} \in \boldsymbol{C}} p(\boldsymbol{\theta}) d\boldsymbol{\theta} - 1 \right), \\ &\text{with } \boldsymbol{C} = \{\boldsymbol{\theta} : \hat{\theta}_r = \sum_u \theta_u \delta_{b_u r} = 1, \ \ \forall r = 1, 2, .., B\}. \end{aligned} \tag{2.19}$$

The integral is restricted to the space $\boldsymbol{C}$ due to the group-wise normalisation constraints

defined in equation (2.11). It is easy to derive[4] that the maximum entropy distribution for $\boldsymbol{\theta}$ is simply a product of uniform distribution over $n_r - 1$ regular simplices for $r \in \{1, 2, .., B\}$

$$P(\boldsymbol{\theta}|\boldsymbol{b}) = \prod_{r=1}^{B}(n_r - 1)! \times \delta_{\sum_u \theta_{b_u r},1}. \tag{2.20}$$

For the expected number of connections between groups λ_{rs} which take values in $[0, \infty]$, defining the maximum entropy prior requires an extra constraint on the first moment. We adopt the empirical Bayes approach [90] in which the mean of the λ_{rs} is set to be the value of observed average number of connections between any pair of groups $\bar{\lambda} = 2E/B(B+1)$.[5] Because all λ_{rs} are equivalent and independent, the maximum entropy distribution of each λ_{rs} is the stationary point of the Lagrangian function

$$\begin{aligned} &L(p(\lambda_{rs}), \xi_0, \xi_1) \\ &\quad = -\int_0^\infty p(\lambda_{rs}) \ln p(\lambda_{rs}) d\lambda_{rs} + \xi_0 \left(\int_0^\infty p(\lambda_{rs}) d\lambda_{rs} - 1 \right) + \xi_1 \left(\int_0^\infty \lambda_{rs} p(\lambda_{rs}) d\lambda - \bar{\lambda}_{rs} \right). \end{aligned} \tag{2.21}$$

It is easy to show [92] the solution of this constrained optimisation is the exponential distribution with mean $\bar{\lambda}$,

$$p(\lambda_{rs}|\bar{\lambda}) = \begin{cases} \frac{1}{(1+\delta_{rs})\bar{\lambda}} e^{\left[-\lambda_{rs}/(1+\delta_{rs})\bar{\lambda}\right]}, & \lambda_{rs} \in [0, \infty] \\ 0, & \text{otherwised} \end{cases}, \tag{2.22}$$

where the δ_{rs} term is used to accommodate the fact that λ_{rr} are twice the number of edges inside group r.

Having made our choice of the priors for $\boldsymbol{\lambda}$ and $\boldsymbol{\theta}$, we are ready to compute the marginal likelihood $P(\boldsymbol{A}|\boldsymbol{b})$ by plugging the equations (2.20) and (2.22) into (2.16), which gives the following expression

$$P(\boldsymbol{A}|\boldsymbol{b}) = \frac{\bar{\lambda}^E}{(\bar{\lambda}+1)^{E+B(B+1)/2}} \times \frac{\prod_{r<s} e_{rs}! \prod_r e_{rr}!!}{\prod_{u<v} A_{uv}! \prod_u A_{uu}!!} \times \prod_r \frac{(n_r - 1)!}{(e_r + n_r - 1)!} \prod_u k_u!. \tag{2.23}$$

We leave the detailed derivation of this expression to Appendix A.2.

[4] Please see derivation in Appendix A.1.

[5] Empirical Bayes approach is relatively straightforward to implement, with hyperparameters being set to be estimates obtained from data. By contrast, fully Bayesian inference requires to integrate hyperparameters out, which might not admit analytical solution. However, Empirical Bayes approach is generally a reasonable approximate to the fully Bayesian approach [91]. Since we will focus on the microcanonical variant later, which does not require actual computation of marginalisation, we just choose the empirical Bayes approach for the canonical DC-SBM here for the ease of presentation.

Prior for the network partition

Choosing the prior distribution for the network partition $\boldsymbol{b}$ requires more work compared to what we have done for the parameters $\boldsymbol{\lambda}$ and $\boldsymbol{\theta}$. Since we generally do not have any prior knowledge about the community structure, we should again consider the most uninformative distribution. However, applying the maximum entropy principle might still introduce undesired bias. To begin with, we consider the following Lagrangian function

$$L(p(\boldsymbol{b}), \xi_0) = -\sum_{\boldsymbol{b}} p(\boldsymbol{b}) \ln p(\boldsymbol{b}) + \xi_0 \left(\sum_{\boldsymbol{b}} p(\boldsymbol{b}) - 1 \right), \tag{2.24}$$

where the summation in the last equation goes through all possible network partitions $\boldsymbol{b}$. The stationary point of the Largrangian function above is the uniform distribution

$$P(\boldsymbol{b}) = \frac{1}{\sum_{b'} 1} = \frac{1}{a_N}, \tag{2.25}$$

where in the denominator $a_N$is the total number of possible partitions with N nodes,

$$a_N = \sum_{B=1}^{N} S(N, B) B!, \tag{2.26}$$

with $S(N, B)$ being the Stirling's number of second kind, counting the number of ways to assign N nodes into B indistinguishable blocks [93]. Despite being the maximum entropy solution, this uniform distribution actually carries strong bias regarding the number of communities. As shown in Fig. 2-6, the values of $S(N, B)$ are significantly smaller when B is either very small or very large. In practice, the number of communities B is rarely compatible with the total number of nodes N. In the more practically relevant region where B is much smaller than N, the left tail of the "bell shape" given in Fig. 2-6 indicates that the prior in equation (2.25) has preference toward the partitions with large values of B. Because the number of communities B is usually an important aspect we like to infer from data, it is necessary to choose a prior which does not carry bias regarding the number of communities.

One general solution for removing undesired bias is to construct *hierarchical priors* [74]. The idea is to model the higher-oder perspective with which we like to be agnostic, e.g. the number of communities, by a *hyperprior*. Then, the prior to be used is a parametric probability distribution which is conditioned on the values being sampled from the hyperprior. In particular, we view the number of communities B as a *hyperparameter*

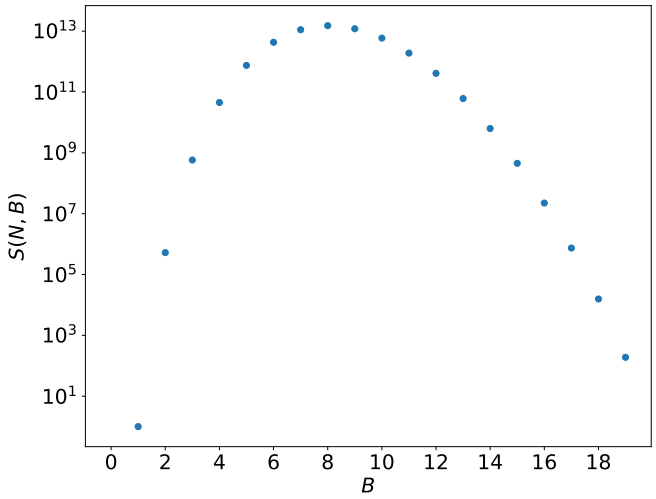


Figure 2-6: Stirling number of the second kind $S(N, B)$ as explained in the text for $N = 20$ and $B \in [1, 19]$. As suggested by the bell shape of the points in the figure, the uniform distribution in equation (2.25) carries strong bias regarding certain values of the number of blocks.

that is sampled from a uniform hyperprior, e.g. $P(B) = 1/N$. The network partition $\boldsymbol{b}$ then can be drawn from the following uniform distribution,

$$P(\boldsymbol{b}|B) = \frac{1}{S(N, B)B!}, \tag{2.27}$$

which leads us to a hierarchical prior

$$P(\boldsymbol{b}, B) = P(\boldsymbol{b}|B)P(B) = \frac{1}{S(N, B)B!}\frac{1}{N}. \tag{2.28}$$

However, the hierarchical prior above it is still not a good choice. This is because the hierarchical prior still carries bias regarding the sizes of communities. If we draw samples from the prior (2.28), most of partitions will acquire approximately identical size of communities. Since we generally do not want to assume a uniform distribution of community sizes, to represent our agnosticism regarding the size of communities, we can take the same remedy as before. Specifically, we adapt the prior by treating the group sizes $\boldsymbol{n} = \{n_r\}$ s.t. $\sum_r n_r = N$ as another hyperparameter, then draw $\boldsymbol{n} = \{n_r\}$ from a uniform distribution with equal probability,

$$P(\boldsymbol{n}|B) = \binom{N-1}{B-1}^{-1}, \tag{2.29}$$

where $\binom{N-1}{B-1}$ counts the number of ways to assign N nodes into B nonempty groups.

The network partition is then being drawn from a uniform distribution which assigns equal probability to every possible configuration of $\boldsymbol{b}$, conditioned on the communities' sizes $\boldsymbol{n} = \{n_r\}$,

$$P(\boldsymbol{b}|\boldsymbol{n}) = \frac{\prod_r n_r!}{N!}. \tag{2.30}$$

The arguments above leads to the following Bayesian hierarchical prior[6]

$$P(\boldsymbol{b}) = P(\boldsymbol{b}|\boldsymbol{n})P(\boldsymbol{n}|B)P(B) = \frac{\prod_r n_r!}{N!}\binom{N-1}{B-1}^{-1}\frac{1}{N}. \tag{2.31}$$

We will stick with this choice of prior for the network partition in the rest of the thesis. Although it is always possible to find bias of the prior at higher-level perspectives, there is a good reason to stop developing the hierarchy further. As argued in [60, 74], constructing higher hierarchy of the prior will bring only vanishingly small reward. Specifically, if we take the logarithm of the prior (2.31) and assume that community sizes are large enough to allow Stirling's factorial approximation $\ln x! = x\ln x - x$, as well as $B \ll N$, then

$$\ln P(\boldsymbol{b}) \approx -NH(\boldsymbol{n}) - \mathcal{O}(\ln N), \tag{2.32}$$

where $H(\boldsymbol{n}) = -\sum_r (n_r/N)\ln(n_r/N)$ is the entropy of the community size distribution. The first term $-NH(\boldsymbol{n})$ in the last equation is in fact an information-theoretical limit of $\ln P(\boldsymbol{b})$ [94]: for sufficient data, the log-probability of the data generating process approaches the entropy of the data. Therefore, no matter how we further refine the prior in (2.31), the improvement will be no larger than the scale of $\mathcal{O}(\ln N)$, which will make little practical difference. Therefore, we choose (2.31) as our final choice of prior for the partition $\boldsymbol{b}$.

Now we have obtained all necessary ingredients for inference in the Bayes' formula (2.13). In summary, we can evaluate the posterior probability of the DC-SBM up to a normalising constant via the following expression

$$\begin{aligned} P(\boldsymbol{b}|\boldsymbol{A}) \propto \frac{\bar{\lambda}^E}{(\bar{\lambda}+1)^{E+B(B+1)/2}} \times \frac{\prod_{r<s} e_{rs}! \prod_r e_{rr}!!}{\prod_{u<v} A_{uv}! \prod_u A_{uu}!!} \\ \times \prod_r \frac{(n_r-1)!}{(e_r+n_r-1)!} \prod_u k_u! \times \frac{\prod_r n_r!}{N!}\binom{N-1}{B-1}^{-1}\frac{1}{N}. \end{aligned} \tag{2.33}$$

[6] Here we write the prior probability of network partition $\boldsymbol{b}$ as $P(\boldsymbol{b})$ rather than $P(\boldsymbol{b}, \boldsymbol{n}, B)$ to be consistent with the literature [60, 74]. This convention comes from the fact that $\boldsymbol{n}$ and B are hyperparameters which are fixed for a particular network partition $\boldsymbol{b}$. In other words, $P(\boldsymbol{b}) = \sum_{\boldsymbol{n},B} P(\boldsymbol{b}, \boldsymbol{n}, B) = P(\boldsymbol{b}, \boldsymbol{n}^*, B^*)$, where $\{\boldsymbol{n}*, B^*\}$ is the pair of hyperparameters compatible with $\boldsymbol{b}$.

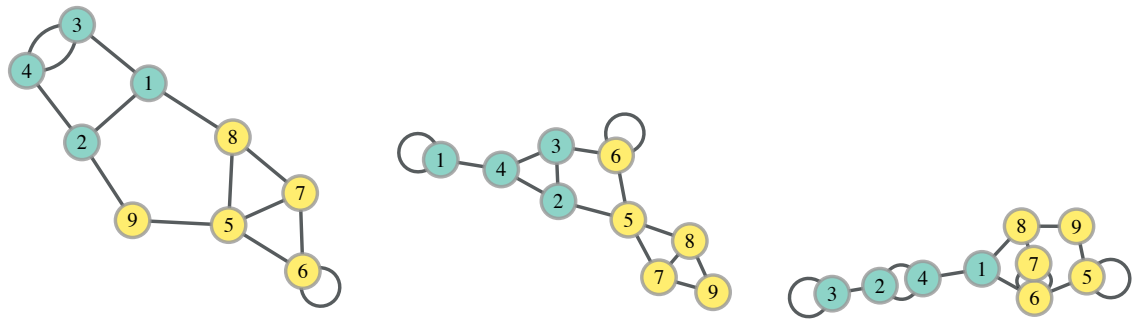


Figure 2-7: Three samples drawn from the same microcanonical SBM with parameters $\boldsymbol{b} = (1,1,1,1,2,2,2,2,2), \boldsymbol{e} = \begin{pmatrix} 10 & 2 \\ 2 & 14 \end{pmatrix}, \boldsymbol{k} = (3,3,3,3,4,4,3,3,2)$. Note that the edge count matrix $\boldsymbol{e}$ and degree sequence $\boldsymbol{k}$ are satisfied exactly across all the examples.

As we shall explain later, when we can only evaluate a probability distribution up to the normalising constant, we can make useful inference via numerical approximate algorithms. Before we look at the inference algorithm in Section 2.5, in the next section, we introduce another two important variants of SBMs to be used later.

2.4 Microcanonical SBM, description length and Nested SBM

2.4.1 Microcanonical SBM

The term "microcanonical" has its origin in the field of physics, suggesting that model parameters are set to satisfy hard constraints *without* variation, as opposed to canonical models where parameters are only required to obey on average. In particular, parameters of the *microcanonical* DC-SBM [74] include a network partition $\boldsymbol{b}$, the edge count matrix $\boldsymbol{e} = \{e_{rs}\} \in \mathbb{Z}^{B\times B}$ whose entries e_{rs} are numbers of edges between group r and s, and the degree sequence of each node $\boldsymbol{k} = \{k_u\} \in \mathbb{Z}^N$, with $\sum_{rs} e_{rs} = \sum_u k_u = 2E$. As a concrete example, in Fig 2-7, we consider a microcanonical DC-SBM with 2 communities and visualise three samples drawn from the same model. The point is that the edge count matrix $\boldsymbol{e}$ and the node degree sequence $\boldsymbol{k}$ remain the same across different samples. The sample space of a microcanonical DC-SBM model consists of all compatible (with the predefined parameters $\boldsymbol{b}, \boldsymbol{e}, \boldsymbol{k}$) networks, and each of them acquires equal probability of being seen under the model.

To write down the probability of generating an observed network from the microcanonical DC-SBM, we simply need to count the number of all possible compatible networks.

Firstly, note that the total number of edge count matrices satisfying the predefined model parameter $\boldsymbol{e}$ is

$$\Omega(\boldsymbol{e}) = \frac{\prod_r e_r!}{\prod_{r<s} e_{rs}! \prod_r e_{rr}!!}. \tag{2.34}$$

Since the number of adjacency matrices $\boldsymbol{A}$ satisfying a fixed edge count matrix $\boldsymbol{e} = \{e_{rs}\}$ is

$$\Xi(\boldsymbol{A}) = \frac{\prod_u k_u!}{\prod_{u<v} A_{uv}! \prod_u A_{uu}!!}, \tag{2.35}$$

the probability of generating a network $\boldsymbol{A}$ from the microcanonical SBM is then

$$P(\boldsymbol{A}|\boldsymbol{e}, \boldsymbol{k}, \boldsymbol{b}) = \frac{\Xi(\boldsymbol{A})}{\Omega(\boldsymbol{e})} = \frac{\prod_{r<s} e_{rs}! \prod_r e_{rr}!! \prod_u k_u!}{\prod_{u<v} A_{uv}! \prod_u A_{uu}!! \prod_r e_r!}. \tag{2.36}$$

To obtain the posterior probability $P(\boldsymbol{b}|\boldsymbol{A})$ of the microcanonical SBM, we need to go through the same steps we have done for the DC-SBM: choose appropriate priors for the parameters $\boldsymbol{\Theta} = \{\boldsymbol{e}, \boldsymbol{k}\}$, then marginalise the likelihood $P(\boldsymbol{A}|\boldsymbol{e}, \boldsymbol{k}, \boldsymbol{b})$ to derive $P(\boldsymbol{A}|\boldsymbol{b})$. One important property of the microcanonical variant is that the derivation of its marginal likelihood does not require any actual computation of marginalisation. To see this, we write down the expression of the marginal likelihood of the microcanonical SBM according to equation (2.16),

$$P(\boldsymbol{A}|\boldsymbol{b}) = \sum_{\boldsymbol{e},\boldsymbol{k}} P(\boldsymbol{A}, \boldsymbol{e}, \boldsymbol{k}|\boldsymbol{b}) = P(\boldsymbol{A}|\boldsymbol{e}, \boldsymbol{k}, \boldsymbol{b}) P(\boldsymbol{k}|\boldsymbol{e}, \boldsymbol{b}) P(\boldsymbol{e}|\boldsymbol{b}), \tag{2.37}$$

where the integral in (2.16) becomes summation here because the parameters $\boldsymbol{e}$ and $\boldsymbol{k}$ are discrete under the microcanonical formulation. The second equals sign in the last equation holds due to the microcanonical nature of the model parameters: Conditioned on network partition $\boldsymbol{b}$, there is only one pair of $(\boldsymbol{e}, \boldsymbol{k})$ that matches the observed data and all other inconsistent parameters have zero probability (recall the three example networks shown in Figure 2-7). Such property is an advantage of the microcanonical variant compared to its canonical counterpart, since the marginal likelihood is easy to derive without any computation involved. To complete the marginal likelihood of the microcanonical DC-SBM, we just need to pick appropriate prior distributions for $P(\boldsymbol{e}|\boldsymbol{b})$ and $P(\boldsymbol{k}|\boldsymbol{e}, \boldsymbol{b})$.

Interestingly, one can make specific choices of priors $P(\boldsymbol{e}|\boldsymbol{k}, \boldsymbol{b})$ and $P(\boldsymbol{k}|\boldsymbol{b})$ such that the microcanonical DC-SBM equivalent to the canonical DC-SBM [74]. In particular, consider the geometric distribution with parameter p, which states that the probability of getting x tails when the first head occurs in a series of identical Bernoulli trails with

probability p is

$$P(x|p) = (1-p)^x p, \quad \text{for } x \in \{0, 1, 2, ...\}. \tag{2.38}$$

If we assume e_{rs} - the number of edges between group r and s - are independent variables drawn from geometric distributions with the same parameter $p = 1/(\bar{\lambda}+1)$, then we have

$$P(\boldsymbol{e}|\boldsymbol{b}) = \prod_{r<s} \frac{\bar{\lambda}^{e_{rs}}}{(\bar{\lambda}+1)^{e_{rs}+1}} \prod_r \frac{\bar{\lambda}^{e_{rr}/2}}{(\bar{\lambda}+1)^{e_{rr}/2+1}} = \frac{\bar{\lambda}^E}{(\bar{\lambda}+1)^{E+B(B+1)/2}}. \tag{2.39}$$

Suppose we choose the mean parameter[7] $\bar{\lambda} = 2E/(B(B+1))$, and set the prior for the degree sequence $P(\boldsymbol{k}|\boldsymbol{e}, \boldsymbol{b})$ to be the following uniform distribution

$$P(\boldsymbol{k}|\boldsymbol{e}, \boldsymbol{b}) = \prod_r \frac{e_r!(n_r-1)!}{(e_r+n_r-1)!} = \prod_r \left(\!\binom{n_r}{e_r}\!\right)^{-1}, \tag{2.40}$$

where $\left(\!\binom{n}{m}\!\right) = \binom{n+m-1}{m}$ counts the number of m-combinations with repetitions from a set of size n, we can write down the marginal likelihood of the microcanonical DC-SBM in equation (2.37),

$$\begin{aligned}
P(\boldsymbol{A}|\boldsymbol{b}) &= P(\boldsymbol{A}|\boldsymbol{e}, \boldsymbol{k}, \boldsymbol{b}) P(\boldsymbol{k}|\boldsymbol{e}, \boldsymbol{b}) P(\boldsymbol{e}|\boldsymbol{b}) \\
&= \text{Eq. (2.37)} \times \text{Eq.(2.41)} \times \text{Eq.(2.40)} \\
&= \frac{\prod_{r<s} e_{rs}! \prod_r e_{rr}!! \prod_u k_u!}{\prod_{u<v} A_{uv}! \prod_u A_{uu}!! \prod_r e_r!} \times \prod_r \frac{e_r!(n_r-1)!}{(e_r+n_r-1)!} \times \frac{\bar{\lambda}^E}{(\bar{\lambda}+1)^{E+B(B+1)/2}} \\
&= \frac{\bar{\lambda}^E}{(\bar{\lambda}+1)^{E+B(B+1)/2}} \times \frac{\prod_{r<s} e_{rs}! \prod_r e_{rr}!!}{\prod_{u<v} A_{uv}! \prod_u A_{uu}!!} \times \prod_r \frac{(n_r-1)!}{(e_r+n_r-1)!} \prod_u k_u!.
\end{aligned} \tag{2.41}$$

The last equation is identical to the marginal likelihood of the canonical DC-SBM we derived in equation (2.23). Such an observation implies that, despite the microcanonical and canonical SBM prescribing different data generating processes, we will not be able to determine the observed data is generated from which of the two models.

The choice of prior in equation (2.39) is made simply to show the connections between the canonical and microcanonical models. However, with the microcanonical generating process in mind, it is natural to use microcanonical priors instead. For the edge count

[7]The geometric distribution is not the only possible choice for the prior distribution of e_{rs}. We consider geometric distributions with the particular choice of $\bar{\lambda}$ as explained in the text is simply because they make microcanonical and canonical SBMs equivalent. We shall replace this prior later with a microcanonical prior because it the microcanonical prior does require any hyper-parameters, making the final expression of the posterior distribution non-parametric.

matrix $\boldsymbol{e} = \{e_{rs}\}$, the microcnanonical prior is the uniform distribution over all possible ways to assign the total number of edges E among B communities:

$$P(\boldsymbol{e}|\boldsymbol{b}) = \left(\binom{B(B+1)/2}{E}\right)^{-1}. \tag{2.42}$$

Notice that this microcanonical prior does not require any extra hyperparameter, while the mixture geometric prior in equation (2.39) needs a subjective estimate of the mean value $\bar{\lambda}$. For this reason, the microcanonical prior in equation (2.42) is a better choice[8] in the sense that the resulted posterior probability $P(\boldsymbol{b}|\boldsymbol{A})$ is independent of any extra information not included in $\boldsymbol{b}$, except for the total number of edges E. Since the total number of edges E is independent of the network partition $\boldsymbol{b}$, the prior $P(E)$ can be chosen arbitrarily, which only amounts to a multiplying constant without affecting the posterior distribution. This is the other advantage of using the microcanonical variant: the posterior probability of the microcanonical SBM is fully non-parametric, requiring no subjective estimate of any hyperparameters. We summarise the posterior probability of the microcanonical DC-SBM (up to the normalising constant) as follows,

$$\begin{aligned} P(\boldsymbol{b}|\boldsymbol{A}) \propto & \frac{\prod_{r<s} e_{rs}! \prod_r e_{rr}!! \prod_u k_u!}{\prod_{u<v} A_{uv}! \prod_u A_{uu}!! \prod_r e_r!} \times \prod_r \frac{e_r! n_r!}{(e_r + n_r - 1)!} \\ & \times \prod_r \left(\binom{n_r}{e_r}\right)^{-1} \times \left(\binom{(B(B+1)/2)}{E}\right)^{-1} \times \frac{\prod_r n_r!}{N!} \binom{N-1}{B-1}^{-1} \frac{1}{N}. \end{aligned} \tag{2.43}$$

2.4.2 Description length

The Bayesian inference approach has a built-in Occam's razor effect, making it robust against overfitting data. The microcanonical formulation of the DC-SBM allows us to examine the Bayesian Occam's razor effect in a direct way, which requires some concepts from the field of information theory. Specifically, consider a discrete random variable X generated from some source with probability distribution $P(X)$, then the amount of information required to compress an observed outcome $X = x$ is approximately $-\ln P(x)$ (nats) units. Then, we can rewrite the joint probability $P(\boldsymbol{A}, \boldsymbol{b})$ as follows

$$\ln P(\boldsymbol{A}, \boldsymbol{b}) = e^{-\Sigma}, \tag{2.44}$$

[8] We will see later in Chapter 4 that this prior is still not ideal, causing the "resolution limit" underfitting problem. This problem can be resolved by replacing the uniform prior by a hierarchical prior, which leads to the nested variant of variant.

where

$$\Sigma = -\ln P(\boldsymbol{A}|\boldsymbol{e},\boldsymbol{k},\boldsymbol{b}) - \ln P(\boldsymbol{e},\boldsymbol{k},\boldsymbol{b}) \tag{2.45}$$

is called the *description length* of the data [95,96]. Description length is approximately the total amount of information required to describe the data, as well as the model used to help data compression. As can be read from the last equation, the expression of description length consists of two parts: the first term $-\ln P(\boldsymbol{A}|\boldsymbol{e},\boldsymbol{k},\boldsymbol{b})$ is the asymptotic amount of information required to describe the data, conditioned on the model described by the parameters $\{\boldsymbol{e},\boldsymbol{k},\boldsymbol{b}\}$; the second term $-\ln P(\boldsymbol{e},\boldsymbol{k},\boldsymbol{b})$ is the amount of information required to describe the model itself. As the model becomes more complicated with more modelling parameters, the modelling capacity should increase, making the amount of information required to describe the data decrease. Nevertheless, the increase in model complexity should cause increase in the amount of information required to describe the model itself, playing a role of penalising the use of complicated models. Therefore, the quality of model fit and model complexity fight with each other, and as a result, the partitions with high posterior probability will be those reaching a good balance between the two. For this reason, complicated models will not be selected unless the reward in the model fit exceeds the penalty of model complexity. This trade-off between the model fit and model complexity functions as an instantiation of the Occam's razor that prevents us from overfitting data.

Note that the equivalence between the Bayesian inference approach and the *minimum description length* (MDL) principle holds in general [96] and does not rely on the use of the microcanonical model. Conducting Bayesian inference with other variants of SBM also has the advantage of being robust against overfitting, but the microcanonical model allows the connection, therefore the robustness of the Bayesian approach, to reveal itself in a more evident way.

2.4.3 Nested DC-SBM

The mirocanonical variant of DC-SBM makes it convenient to define the *hierarchical*, or the *Nested DC-SBM* [63]. The main motivation behind the Nested DC-SBM is that inference with the DC-SBM has the risk of underfitting data. We say underfitting occurs if what we get from the algorithm is overly simplistic compared to the actual pattern in data. The underfitting behaviour of DC-SBM leads to a so-called "resolution limit" of the number of detectable communities. If the number of communities is above the resolution limit, no matter how significant the community structure is, DC-SBM will only be able to partially identify the communities structure, often merging communities of small sizes into large ones. As explained by Peixoto in [60] and [63], the root cause

of the underfitting behaviour lies in the use of the maximum entropy prior of the edge count matrix $\boldsymbol{e} = \{e_{rs}\}$ as defined in equation (2.42), or the canonical version in equation (2.39). Samples being drawn from these two choices of prior for $\boldsymbol{e}$ tend to have identical expected number of connections between any pair of communities. When the pattern in data deviates from this property, the penalty effect caused by the prior could overtake the signal in data, causing the loss of detailed structure information. Later in Chapter 4, we shall demonstrate the underfititng problem of DC-SBM in synthetic networks with clear community structures.

The Nested DC-SBM proposed by Peixoto in [63] resolves the underfitting problem of DC-SBM by considering the connection pattern at community level as another aspect to be modelled. The idea behind the Nested DC-SBM is similar to the one for developing the hierarchical prior for the network partition as defined in equation (2.31). Because the undesired bias is about the connections among communities, we can remove the bias by firstly sampling the connection matrix $\boldsymbol{e}$ from some hyperpriors. One trick here is to exploit the fact that the edge count matrix $\boldsymbol{e} = \{e_{rs}\}$ of a network partition $\boldsymbol{b}$ can be viewed as the adjacency matrix of a meta-graph, where meta-nodes are communities and edges are placed according to the connections between meta-nodes. Then, this meta-graph can be modelled by another SBM at one level above the original model, serving as a prior for the edge count matrix at the bottom level. This procedure can carry on recursively, modelling the edge count matrix of the graph at the current level by another SBM at a higher level, until we reach the highest level with a single meta-node. Fig 2-8 visualises the hierarchical construction of a Nested DC-SBM with three levels. To write down the posterior distribution of the Nested DC-SBM, assume the total number of levels is L. Let $\boldsymbol{b}_l$ be the network partition and $\boldsymbol{e}_l$ be the edge count matrices at the l^{th} level. Then the probability of generating the hierarchical construction $\{\boldsymbol{e}_l\}, l \in \{1, 2, .., L\}$ can be written as

$$P(\{e_l\}|\{b_l\}) = \prod_{l=1}^{L} P(e_l|e_{l+1}, \boldsymbol{b}_l), \tag{2.46}$$

and at each level l, we have

$$P(\boldsymbol{e}_l|\boldsymbol{e}_{l+1}, \boldsymbol{b}_l) = \prod_{r<s} \left(\!\!\binom{n_r^l n_s^l}{e_{rs}^{l+1}}\!\!\right)^{-1} \prod_r \left(\!\!\binom{n_r^l (n_r^l + 1)/2}{e_{rr}^{l+1}/2}\!\!\right)^{-1} \tag{2.47}$$

being the probability of sampling a multigraph from the microcanonical SBM.

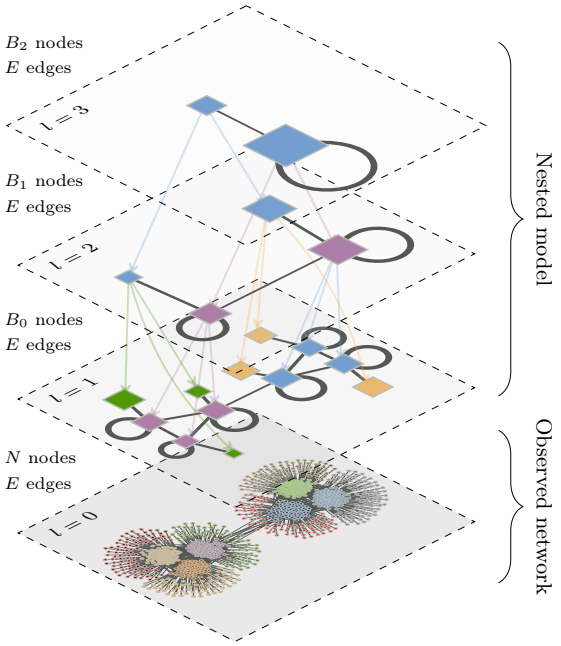


Figure 2-8: Visualisation of the hierarchal construction of a nested variant of SBM with three levels, reproduced from [60].

This leads to the joint distribution

$$\begin{aligned} P(\boldsymbol{b}|\boldsymbol{A},\boldsymbol{k},\{\boldsymbol{e}_l\},\{\boldsymbol{b}_l\}) &\propto P(\boldsymbol{A}|\boldsymbol{k},\boldsymbol{e}_0,\boldsymbol{b}_1)\times P(\boldsymbol{k},\boldsymbol{e}_1,\boldsymbol{b}_1)\times P(\{e_l\})\times P(\{b_k\}) \\ &= \frac{\prod_u k_u!\prod_{r<s}e_{rs}!\prod_r e_{rr}!!}{\prod_r e_r!\prod_{u<v}A_{uv}!\prod_u A_{uu}!!}\times\prod_r\left(\!\!\binom{n_r}{e_r}\!\!\right)^{-1}\times\prod_{l=1}^{L}\prod_{r<s}\left(\!\!\binom{n_r^l n_s^l}{e_{rs}^{l+1}}\!\!\right)^{-1} \\ &\times\prod_r\left(\!\!\binom{n_r^l(n_r^l+1)/2}{e_{rr}^{l+1}/2}\!\!\right)^{-1}\times\frac{\prod_r n_r^l!}{B_{l-1}!}\binom{B_{l-1}-1}{B_l-1}^{-1}\frac{1}{B_{l-1}}. \end{aligned} \tag{2.48}$$

Despite the Nested DC-SBM requiring more modelling parameters compared to the vanilla DC-SBM, under the Bayesian framework, we do not need to worry about over-fitting data. As before, we can see the Occam's razor is in place by writing down the description length of data in terms of the joint probability $P(\boldsymbol{A},\boldsymbol{k},\{e_l\},\{\boldsymbol{b}_l\})$,

$$\Sigma = -\ln P(\boldsymbol{A},\boldsymbol{k},\{\boldsymbol{e}_l\},\{\boldsymbol{b}_l\}) = -\ln P(\boldsymbol{A}|\boldsymbol{k},\{\boldsymbol{e}_l\},\{\boldsymbol{b}_l\}) - \ln P((\boldsymbol{k},\{\boldsymbol{e}_l\},\{\boldsymbol{b}_l\}). \tag{2.49}$$

Since the vanilla DC-SBM is in fact a special case of the Nested DC-SBM with a single layer $L = 1$, the Nested DC-SBM will always perform at least as well as the vanilla DC-SBM, but having the potential to achieve better fit to data. We will explain how to conduct model selection to decide which model is better in Section 2.6 after we

explain the numerical approximate algorithm for making inference from the posterior distribution of SBMs in the next section.

2.5 Inference algorithm

The goal of the inference task is to extract information about the network partition $\boldsymbol{b}$ via the posterior distribution $P(\boldsymbol{b}|\boldsymbol{A})$, which requires us to draw samples, or to maximise $P(\boldsymbol{b}|\boldsymbol{A})$. Although we can write down the expression of the posterior distribution of SBM, direct sampling or maximisation could be challenging. In practice, inference from the posterior distribution often relies on numerical approximation algorithms. In particular, we can use techniques based on the *Markov Chain Monte Carlo* (MCMC) method. The idea is that we can construct a Markov chain that acquires our target distribution as its equilibrium distribution. After we realise the chain for a sufficiently long time and the chain is in its equilibrium distribution, samples drawn from this chain will serve as approximation of samples from our target distribution. Specifically, we firstly start with some random initialisation $\boldsymbol{b}_0$, then successively make moves from current state $\boldsymbol{b}$ to a new state $\boldsymbol{b}'$ using some proposal distribution $P(\boldsymbol{b}'|\boldsymbol{b})$. If the proposal distribution is ergodic, i.e. every state has a non-zero probability to be visited, and we accept samples from $P(\boldsymbol{b}'|\boldsymbol{b})$ according to the Metroplis-Hasting criterion [97], which states that we accept a proposed sample with probability

$$a = \min\left\{1, \frac{P(\boldsymbol{b}'|\boldsymbol{A})P(\boldsymbol{b}|\boldsymbol{b}')}{P(\boldsymbol{b}|\boldsymbol{A})P(\boldsymbol{b}'|\boldsymbol{b})}\right\}, \tag{2.50}$$

then the chain will have an equilibrium distribution being our target distribution $P(\boldsymbol{b}|\boldsymbol{A})$. To find the *maximum a priori* (MAP) solution

$$\boldsymbol{b}^* = \underset{\boldsymbol{b}}{\operatorname{argmax}}\ P(\boldsymbol{b}|\boldsymbol{A}), \tag{2.51}$$

we can use the simulated annealing scheme [98, 99]. This can be done by replacing the $P(\boldsymbol{b}|\boldsymbol{A})$ in (2.50) with $P(\boldsymbol{b}|\boldsymbol{A})^{\pi}$, where π is called the inverse temperature, which should increase at each iteration and gradually grow to infinity (see Algorithm 2.1). When π is small, the algorithm is in an exploring status with a strong interest to search in a broad region. As π becomes large, the algorithm becomes greedy and exploits the region with as large increment in the objective function as possible.

Algorithm 2.1: Simulated annealing

1. Start from a random initialisation $\boldsymbol{b}_0$; choose the number of iterations of the simulated annealing n_{iter} and the minimum and maximum values of the inverse temperature, $\pi_{\min}$ and $\pi_{\max}$

2. Consider an exponentially decrease scheme for the temperature, where the speed of increase of the inverse temperature is $\Delta\pi = \exp((\ln \pi_{\max} - \ln \pi_{\min})/n_{\text{iter}})$

3. The simulated annealing proceed as follows

   for $i = 1, 2, .., n_{\text{iter}}$ **do**
   $\pi \leftarrow \pi_{\min}(\Delta\pi)^i$
   for $u = 1, 2, ..., N$ **do**
   move node u from group b_u to b'_u with a probability a, where

$$a = \min\left\{1, \left(\frac{P(\boldsymbol{b}'|\boldsymbol{A})}{P(\boldsymbol{b}|\boldsymbol{A})}\right)^{\pi} \frac{P(\boldsymbol{b}|\boldsymbol{b}')}{P(\boldsymbol{b}'|\boldsymbol{b})}\right\}$$

   end for
   end for

Despite its theoretical guarantee of convergence, it is not uncommon to see MCMC methods require prohibitively long convergence time in practice when the algorithm is naively implemented. For example, there are two typical factors that can significantly affect the efficiency of MCMC. The first one is the choice of the initial state $\boldsymbol{b}_0$. Intuitively, if we start from a "bad" position where all of its surrounding regions acquire low posterior probability, then the chain takes a longer time to travel to high-probability regions where the mass of probability density concentrates. In a worse situation, if our goal is to find the MAP solution, starting from a bad position might result in getting stuck in local optimums. The other important factor for the convergence speed of MCMC is the quality of the proposal distribution $P(\boldsymbol{b}'|\boldsymbol{b})$. A proposal with poor quality can cause an overly high rejection rate, wasting time on generating samples that end up with being rejected. For instance, suppose we make proposal from the uniform distribution

$$P(b'_u = r|\boldsymbol{b}) = \frac{1}{B+1}, \tag{2.52}$$

where $B+1$ is total number of possible communities (including the possibility of creating a new community). Then, most of the proposal will be rejected when the networks have

relatively well-defined structure and large number of communities.

Fortunately, for the task of community detection with SBMs, recent advances in the field have equipped us with various strategies for facilitating MCMC [100, 101], making it possible to conduct Bayesian inference in networks with millions of nodes and edges. Firstly, regarding the initial state of simulating the Markov chain, rather than starting from a random network partition, one intuitive remedy is to start from some educated guesses, e.g. partitions given by heuristic algorithms or spectral clustering methods [102]. Moreover, instead of using the random proposal as given in equation (2.52), a more effective strategy is to use proposals which leverage the information of the graph structure and the current state of Markov chain. For example, making proposals of moving a node based on its neighbours' community membership has been proved to an effective technique [100]. In particular, to move a node u in the network to group r, we can get hints from one of its neighbours v and make the move with the following probability,

$$P(r|u, \boldsymbol{b}) = \frac{e_{rb_v} + \epsilon}{e_{b_v} + (B+1)\epsilon}, \tag{2.53}$$

where the ϵ parameter is used for maintaining the ergodicity (i.e. even when $e_{b_v r}$ is zero, the probability of proposing a move toward group r is still non-zero). Since the probability of a random neighbour of u being in group s is

$$y_s^u = \sum_v A_{uv}\delta_{sb_v}/k_u, \tag{2.54}$$

the proposal distribution is

$$P_e(b_u = r|\boldsymbol{b}) \coloneqq \sum_s y_s^u \frac{e_{sr} + \epsilon}{e_s + \epsilon(B+1)}. \tag{2.55}$$

In words, the proposal in equation (2.55) tends to recommend the groups with which most of node u's neighbours are connecting. Computing this proposal takes $\mathcal{O}(k_u)$ times for node u, as long as the edge count matrix $\boldsymbol{e} = \{e_{rs}\}$ is tracked, then the overall complexity of one sweep the MCMC takes $\mathcal{O}(E)$ which is scalable to large-scale systems. In [100], the author showed that start simulating MCMC from a partition given by an agglomerative heuristic with the smart proposal defined above, MCMC shows decent performance in terms of convergence time and not getting stuck at local optimums.

In the rest of this thesis, when it comes to fitting SBMs to data with MCMC, we make use the *graph-tool* library [103] unless we state otherwise. The graph-tool li-

brary provides efficient implementation of MCMC, combining the smart proposal defined in equation (2.55), an agglomerative heuristic for finding good initial state for MCMC [100], and a merge-split proposal for fast traversing low-probability barriers in the solution space [101]. In our numerical experiments, with the implementations available in the graph-tool library, we manage to conduct inference in networks at the sizes of 10^6 nodes and 10^7 edges.

2.6 Model selection

Applying different variants of SBMs to the same dataset is likely to yield different results. To select from different partitions given by different models, we can conduct pairwise comparison of partition-model pairs by computing the posterior probability ratio [89]. Specifically, for two partitions $\boldsymbol{b}_1$ and $\boldsymbol{b}_2$ obtained with models $\mathcal{M}_1$ and $\mathcal{M}_2$ respectively, the posterior probability ratio is defined as

$$\Lambda = \frac{P(\boldsymbol{b}_1, \mathcal{M}_1|\boldsymbol{A})}{P(\boldsymbol{b}_2, \mathcal{M}_2|\boldsymbol{A})} = \frac{P(\boldsymbol{A}, \boldsymbol{b}_1|\mathcal{M}_1)P(\mathcal{M}_1)}{P(\boldsymbol{A}, \boldsymbol{b}_2|\mathcal{M}_2)P(\mathcal{M}_2)}, \tag{2.56}$$

where $P(\mathcal{M}_1)$ and $P(\mathcal{M}_2)$ are prior probabilities representing our bias toward the two models. The pair $(\boldsymbol{b}_1, \mathcal{M}_1)$ is preferred if $\Lambda > 1$, otherwise $(\boldsymbol{b}_2, \mathcal{M}_2)$ is preferred and the magnitude of Λ indicates the level of confidence of model selection. When we have no preference toward any of the two models, we might set $P(\mathcal{M}_1) = P(\mathcal{M}_2)$, then the ratio can be rewritten in terms of the description length of the two models

$$\Lambda = \exp(\Sigma_2 - \Sigma_1), \tag{2.57}$$

with $\Sigma_1 = -\ln P(\boldsymbol{A}, \boldsymbol{b}_1|\mathcal{M}_1)$ and $\Sigma_2 = -\ln P(\boldsymbol{A}, \boldsymbol{b}_2|\mathcal{M}_2)$ being the description length of data under the two models respectively. This shows how the idea of posterior probability ratio coincides with the *Minimum Description Length* (MDL) principle for model selection: we should choose the partition-model pair which achieves the shortest description length. In Chapter 3 and 4, we will extensively make use of the MDL principle to find the best fitting model among different variants of SBMs in empirical networks.

2.7 Concluding remarks

In this chapter we define various variants of SBMs and explain how to use them to infer community structures in networks. Taking the DC-SBM as an example, we explained how to choose prior distributions properly to avoid intrinsic bias in the process of deriving the posterior distribution of network partitions. Once we can evaluate the

posterior probability up to a normalising constant, MCMC algorithms allow us to draw samples or to obtain the maximum a posteriori estimate solution. When we are faced with different network partitions given by different models, we can compare them by computing the posterior odds ratio, which is equivalent to making use of the MDL principle.

All models covered so far are designed to account for general community structures. On one hand, the versatility of general models is a strength because it allows data to speak for themselves: Bayesian inference with general SBMs can detect not just the typical assortativity, but also many other fundamental community structures, as long as they exist in data. On the other hand, for a class of networks in which a particular kind of structure dominates, using general models could be suboptimal. This is due to the Bayesian Occam's razor effect, which penalises complicated models and prefers simpler models if they have identical ability of explaining the data. Under the Bayesian inference framework, restricted variants of SBMs for particular structures like bipartite and core-periphery structure have been developed. These restricted model variants have demonstrated the ability of achieving better quality of fit to data than general model. Although assortative constrained variant of SBMs has been widely studied in the literacture, no work has considered Bayesian inference with the assortative-constrained variant of SBMs. In the next chapter, we will fill this gap in the literature by providing the Bayesian formulation of an assortative-constrained variant of DC-SBM.

Chapter 3

Statistical inference of assortative structures

The assortative structure is probably the most intensively studied community structure [104]. Roughly speaking, a network is said to be assortative, or to have assortative structure, if it can be divided into subgroups of nodes such that nodes in the same group are more connected to themselves than to nodes in other groups. Most of the popular community detection methods search for assortative structure exclusively, ignoring other fundamental structures that are equally important. The widespread use of these methods has led to an impression that assortative structure is a ubiquitous property shared by networks across different domains [105]. However, it is not clear yet whether the assortative structures reported in the literature are simply an artefact caused by using methods that can only find assortative structures. One reason for the doubt on the prevalence of assortativity is that most of assortative community detection methods do not take statistical significance of their results into account. As a result, these methods often return spurious communities in networks known to be non-modular [47, 48]. Such tendency of presenting exaggerated results is known as the *overfitting* behaviour: overfitting algorithms tend to report overly complicated results compared to the actual pattern in data.

The main advantage of the Bayesian inference approach over non-statistical community detection methods is its robustness against overfitting [60]. We have explained the robustness of Bayesian inference from the information-theoretic perspective in Section 2.4.2 and this point is supported by a series of empirical studies as well [106, 107]. The models we have covered so far are general models, which account for general com-

munity structure including assortativity as a special case. The versatility of general models is a strength, because it allows practitioners to be agnostic about the kind of structure to be inferred. However, having only general models at our disposal also has disadvantages. Firstly, due to the Bayesian Occam's razor effect [108], Bayesian inference approach will prefer simpler models over complicated models if they have identical ability of describing data. As a result, using general models will be suboptimal in the case where a particular kind of community structure dominates. In addition, there are questions that we can not answer with only general models at hand. For example, we might want to know whether an observed network acquires a dominant structure, which happens to be a particular kind of structure, say assortativity. If assortativity is in fact not the dominant structure, then applying methods that can only search for assortative structures could return misleading results. Besides, when assortativity is indeed the dominant structure, we might want to know to what extent using a general models is unnecessarily complicated. Answers to these questions are valuable for deepening our understanding of data, but they are difficult to be obtained with general models alone, unless we can have *restricted models* that are designed for particular structures of interest.

In this chapter, we propose a novel method for extracting statistically significant assortative structures in networks. Our method is based on the *planted partition model* (PP model), an assortative-constrained variant of the DC-SBM. We will stick with the Bayesian approach for the inference of the planted partition model, which has the benefit of automatically preventing from overfitting data. Compared to conducting Bayesian inference with DC-SBM and Nested DC-SBM, using PP model is equivalent to adjusting the prior distribution according to the planted partition constraint, which should be detailed later. We demonstrate via analysis and numerical experiments that our method can indeed extract assortative structures and perform robustly against overfitting.

In deriving our method, we also clarify on a claimed equivalence between maximum likelihood inference (MLE) of the planted partition model and the celebrated *modularity maximisation* approach for detecting assortative structure [57]. Modularity maximisation has been one of the most widely applied community detection methods in network analysis. Despite its popularity, modularity maximisation receives criticism because of its heuristic nature and tendency of overfitting data. The connection between modularity maximisation and MLE with PP model was considered as a principle deviation of the former, which has inspired a series of extended community detection methods [109, 110]. However, our analysis shows that this equivalence does not hold in

general, since it requires subjective choices of model parameters, which lack principled justification. Even in the narrow regime of model parameters where the equivalence holds, the equivalence result implies several limitations of the modularity maximisation approach. Specifically, the model to which modularity maximisation is equivalent assumes the number of edges inside each community is identical. Besides, MLE is notorious for overfitting data. Being equivalent to MLE means modularity maximisation will have exactly the same problem with no improvement. In comparison, our proposed method not only will not overfit (i.e. will not report spurious communities in fully random networks), but also have the ability of resolving more general assortative structure, with the number of edges inside different communities being non-uniform. For the reasons listed above, despite the established connection between the two methods, applying Bayesian inference with the PP model is different from, and superior to both maximum likelihood inference and modularity maximisation for detecting *statistically significant* assortative structures.

There are several related works which also introduced assortative constraints to general SBMs. Lu and Szymanski [111] proposed a *regularised* SBM, which associates each node with two different degree-propensity parameters, one for within-group and the other for between group connections. Then, assortativity is enforced by setting higher propensity of having within-group edges than that of between group-edges. One key difference between our proposed method and the regularised SBM is that the latter carry subjective *a posteriori* bias toward finding assortative structures, while our proposed approach adapts *a priori* constraint. In particular, regularised SBM comes with tunable parameters, which control the strength of assortative structures to be inferred in data. When applying this method to different datasets, practitioners generally need to adjust the value of corresponding tunable parameters to search for some "desired" assortative partitions. Therefore, when we say the regularised SBM carries *a posteriori* bias toward assortativity, we mean that their assortative-constraints are dependent on observed data. In comparison, our proposed method adapts *a priori* constraints which are independent of data. In the other related work given by Gabriel et al. [112], constraints on model parameters are explicitly enforced such that the probability of within-group edges always exceed that of between-group edges. Such constraints are stricter than the one we apply in the PP model, and they are generally not appropriate unless the networks of interest are known to have assortative structure. By contrast, our method is arguably more suitable if the goal is to objectively assess whether statistically significant assortative structures exist, and if so how the structures look like.

Our method is amenable to model selection, which allows us to verify the prevalence of

assortative structure. To this end, we compare the performance of the two assortative-constrained models to that of general models, including the DC-SBM and Nested DC-SBM as defined in Chapter 2, using a set of empirical networks from various scientific domains. Such comparison is possible because all these models are developed under the Bayesian inference framework, and they share the same underlying model (i.e. the DC-SBM) and differ from each other only in the perspective of the choice of prior distributions. We find that assortative-constrained variants manage to achieve better fit to data in a few illuminating examples. Nevertheless, general models outperform the constrained variants most of the time, implying that assortative structure is often too simplistic to sufficiently describe the pattern in empirical networks. Our results suggest that the ubiquitousness of assortative structure has been exaggerated in the literature. Therefore, it is worth considering the possibility of general community structure in the design and application of community detection methods.

The rest of this chapter is organised as follows. In Section 3.1, we introduce the maximum likelihood inference with the PP model, followed by the clarification of the established equivalence between maximum likelihood inference and the modularity maximisation approach. We then move to derive the Bayesian inference approach with PP model in Section 3.2. In Section 3.3, we present numerical results in synthetic and empirical networks.

3.1 The planted partition model and modularity maximisation

3.1.1 Maximum likelihood inference with the planted partition model

Before we look into the assortative-constrained variant of SBM, it is useful to remind the DC-SBM and discuss the maximum likelihood inference with it. Following the ideas in [74, 87], we can rewrite the probability of generating a network $\boldsymbol{A}$ from the (canonical) DC-SBM - i.e. the likelihood function of DC-SBM - as follows

$$P(\boldsymbol{A}|\boldsymbol{\theta},\boldsymbol{\lambda},\boldsymbol{b}) = \prod_{u<v} e^{-\theta_u\theta_v\lambda_{b_ub_v}} \frac{(\theta_u\theta_v\lambda_{b_ub_v})^{A_{uv}}}{A_{uv}!} \prod_u e^{-\theta_u^2\lambda_{b_ub_u}/2} \frac{(\theta_u^2\lambda_{b_ub_u}/2)^{A_{uu}/2}}{(A_{uu}/2)!!} \tag{3.1}$$

$$= \prod_{r<s} e^{-\hat{\theta}_r\hat{\theta}_s\lambda_{rs}} \lambda_{rs}^{e_{rs}} \prod_r e^{-\hat{\theta}_r^2\lambda_{rr}/2} \lambda_{rr}^{e_{rr}/2} \times \frac{\prod_u \theta_u^{k_u}}{\prod_{u<v} A_{uv}! \prod_u A_{uu}!!}, \tag{3.2}$$

with $\hat{\theta}_r$ being the sum of degree propensity parameters θ_u inside community r. That is,

$$\hat{\theta}_r = \sum_u^N \theta_u \delta_{r b_u}. \tag{3.3}$$

From equation (3.1) to (3.2), we simply change the order of multiplications from being node-wise to being community-wise. We provide detailed derivation of such result in Appendix A.3.

To obtain the maximum likelihood estimators of model parameters $\boldsymbol{\lambda} = \{\lambda_{rs}\}$ and $\boldsymbol{\theta} = \{\theta_u\}$, it is more convenient to work with the log-likelihood,

$$\ln P(\boldsymbol{A}|\boldsymbol{\theta}, \boldsymbol{\lambda}, \boldsymbol{b}) = -\frac{1}{2}\sum_{rs} \hat{\theta}_r \hat{\theta}_s \lambda_{rs} - \frac{1}{2}\sum_{rs} e_{rs}(\ln \lambda_{rs} - \delta_{rs} \ln 2) + \sum_u k_u \ln \theta_u, \tag{3.4}$$

with the constant $-\ln(\prod_{u<v} A_{uv}! \prod_u A_{uu}!!)$ being dropped. Differentiating the last equation and setting the derivatives to zeros gives[1]

$$\lambda^*_{rs} = \frac{e_{rs}}{\hat{\theta}_r \hat{\theta}_s},\ \theta^*_u = \frac{k_u}{e_{b_u}} \hat{\theta}_{b_u}. \tag{3.5}$$

Looking at the expression of the maximum likelihood estimators λ^*_{rs} and θ^*_u, we notice that the inferred probability of an edge in the network is independent of the values of $\hat{\theta}_r$, since

$$p^*_{uv} = \lambda^*_{b_u b_v} \theta^*_{b_u} \theta^*_{b_v} = \frac{e_{b_u b_v}}{e_{b_u} e_{b_v}} k_u k_v. \tag{3.6}$$

This property reflects the fact that the DC-SBM requires extra constraints on $\hat{\theta}_r$ to fully define the model. We do have the freedom to choose arbitrary values of $\hat{\theta}_r$ at our convenience and the model does not change.

The assortative-constrained model to be considered is called the *planted partition* (PP) model. In the literature, the term planted partition is refered to a special constraint on the parameters of SBMs. In a general Poisson SBM with B communities, the connection pattern is described by its connection matrix $\boldsymbol{\lambda} = \{\lambda_{rs}\}$, which consists of $B(B+1)/2$ unique elements. In comparison, under the planted partition constraint, only two distinct values are allowed in the connection matrix. All of the diagonal elements take the same value $\lambda_{rr} = \lambda_{\text{in}}$, representing twice the expected number of edges inside each community. Similarly, all of the off-diagonal elements in the connection matrix share the same value $\lambda_{rs} = \lambda_{\text{out}}$[2], representing the expected number of connections between

[1] We leave derivations to Appendix A.4

[2] Although it is possible to assume the number of edges between communities are dependant on the

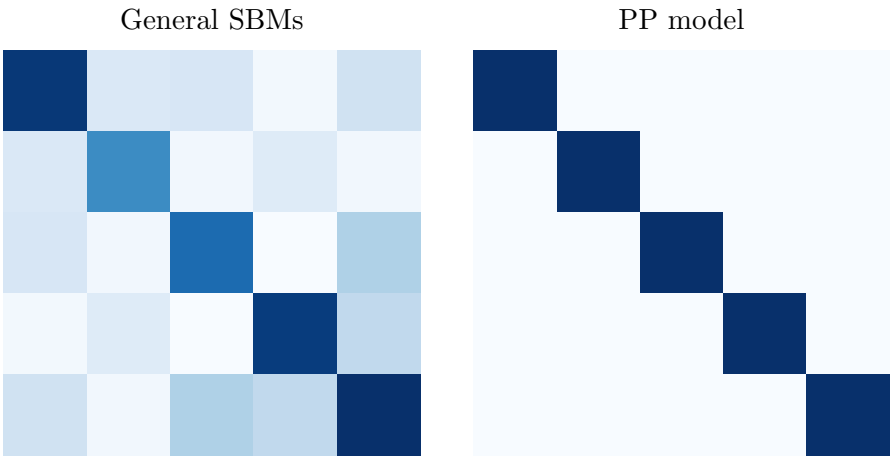


Figure 3-1: Visualisation of two possible connection matrices $\boldsymbol{\lambda} = \{\lambda_{rs}\}$. Both matrices represent assortative connection pattern since both of them are diagonally dominant. However, in principle, general SBMs allow arbitrary mixing structure, which describes by $B(B+1)/2$ uniques values in the connection matrix $\boldsymbol{\lambda}$. Therefore, general SBMs have the ability of describing much more general pattern by varying the elements in $\boldsymbol{\lambda}$. In comparison, the PP model is a restricted model. Under the planted partition constraint, the elements of $\boldsymbol{\lambda}$ can take only two distinct values: λ_{in} for diagonal elements and λ_{out} for off-diagonal elements. The PP model is only able to describe assortative structure when $\lambda_{\text{in}} > \lambda_{\text{out}}$, or disassortative otherwise.

any pair of distinct communities. In other words, the planted partition constraint states that the elements in the connection matrix $\boldsymbol{\lambda} = \{\lambda_{rs}\}$ can be written as

$$\lambda_{rs} = \lambda_{\text{in}}\delta_{rs} + \lambda_{\text{out}}(1 - \delta_{rs}). \tag{3.7}$$

In Fig. 3-1, we show a pictorial comparison of the connection matrix $\boldsymbol{\lambda}$ between general SBMs and the PP model. Although this planted partition constraint significantly reduces the flexibility of the general DC-SBM, it maintains the ability of generating networks with assortative structures. To generate a network with dense within-group connections and sparse between-group connections, we just need to tune the value of $\lambda_{\text{in}}, \lambda_{\text{out}}$ and $\boldsymbol{\theta} = \{\theta_u\}$ such that the inequality $\lambda_{\text{in}} \sum_r \hat{\theta}_r^2/2 > \lambda_{\text{out}} \sum_{r<s} \hat{\theta}_r\hat{\theta}_s$ holds.

The likelihood function of the PP model can be obtained by substituting the planted partition constraint into the likelihood of DC-SBM in (3.2),

$$P(\boldsymbol{A}|\lambda_{\text{in}}, \lambda_{\text{out}}, \boldsymbol{\theta}, \boldsymbol{b}) = e^{-\lambda_{\text{out}} \sum_{r<s} \hat{\theta}_r\hat{\theta}_s} \lambda_{\text{out}}^{e_{\text{out}}} e^{-\lambda_{\text{in}} \sum_r \hat{\theta}_r^2/2} \lambda_{\text{in}}^{e_{\text{in}}} \frac{\prod_u \theta_u^{k_u}}{\prod_{u<v} A_{uv}! \prod_u A_{uu}!!}. \tag{3.8}$$

number of nodes in related communities, e.g. $\lambda_{rs} = n_r n_s \lambda_{\text{out}}$, that might make further analysis tedious. Moreover, such modelling choice is partially due to a historical reason: the planted partition model was originally studies with the assumption that every community has the same number of nodes and the probabilities of an edge between or within communities are the same across the entire network [113].

In the last equation, e_{in} and e_{out} are the number of edges within and between communities respectively,

$$e_{\text{in}} = \frac{1}{2}\sum_{uv} A_{uv}\delta_{b_u b_v}, \tag{3.9}$$

$$e_{\text{out}} = E - e_{\text{in}} = \frac{1}{2}\sum_{uv} A_{uv}(1-\delta_{b_u b_v}). \tag{3.10}$$

We would like to make a comment on the maximum likelihood estimator of the degree propensity parameter $\{\theta_u\}$ for the PP model. The expression of the maximum likelihood estimator can be obtained again by finding the stationary point of the log-likelihood

$$\begin{aligned}\ln P(\boldsymbol{A}|\lambda_{\text{in}},\lambda_{\text{out}},\boldsymbol{\theta},\boldsymbol{b}) = -\lambda_{\text{in}}\sum_r \frac{\hat{\theta}_r^2}{2} - \lambda_{\text{out}}\sum_{r<s}\hat{\theta}_r\hat{\theta}_s + e_{\text{in}}\ln\lambda_{\text{in}} + e_{\text{out}}\ln\lambda_{\text{out}} \\ + \sum_u k_u \ln\theta_u.\end{aligned} \tag{3.11}$$

Differentiating $\ln P(\boldsymbol{A}|\lambda_{\text{in}},\lambda_{\text{out}},\boldsymbol{\theta},\boldsymbol{b})$ above with respect to parameters except for the network partition $\boldsymbol{b}$ leads to following results,

$$\frac{\partial \ln P(\boldsymbol{A}|\lambda_{\text{in}},\lambda_{\text{out},\boldsymbol{\theta},\boldsymbol{b})}}{\partial\lambda_{\text{in}}} = -\sum_r \frac{\hat{\theta}_r^2}{2} + \frac{e_{\text{in}}}{\lambda_{\text{in}}}, \tag{3.12}$$

$$\frac{\partial \ln P(\boldsymbol{A}|\lambda_{\text{in}},\lambda_{\text{out},\boldsymbol{\theta},\boldsymbol{b})}}{\partial\lambda_{\text{out}}} = -\sum_{r<s}\hat{\theta}_r\hat{\theta}_s + \frac{e_{\text{out}}}{\lambda_{\text{out}}}, \tag{3.13}$$

$$\frac{\partial \ln P(\boldsymbol{A}|\lambda_{\text{in}},\lambda_{\text{out},\boldsymbol{\theta},\boldsymbol{b})}}{\partial\theta_u} = -\lambda_{\text{in}}\hat{\theta}_{b_u} - \lambda_{\text{out}}\sum_{s\neq b_u}\hat{\theta}_s + \frac{k_u}{\theta_u}. \tag{3.14}$$

Equating the three equations above to zeros gives,

$$\lambda_{\text{in}}^* = \frac{2e_{\text{in}}}{\sum_r \hat{\theta}_r^2}, \tag{3.15}$$

$$\lambda_{\text{out}}^* = \frac{e_{\text{out}}}{\sum_{r<s}\hat{\theta}_r\hat{\theta}_s}, \tag{3.16}$$

$$\theta_u^* = k_u\left[\frac{2e_{\text{in}}\hat{\theta}_{b_u}}{\sum_s\hat{\theta}_s^2} + \frac{e_{\text{out}}\sum_{s\neq b_u}\hat{\theta}_s}{\sum_{r<s}\hat{\theta}_r\hat{\theta}_s}\right]^{-1}. \tag{3.17}$$

Unfortunately, these equations do not permit analytical solutions and require numerical methods for approximation. Moreover, compared to the maximum likelihood estimator for $\{\theta_u\}$ in the DC-SBM in equation (3.5), here we do not have the freedom to choose

the value of $\hat{\theta}_r$ anymore. This is because, by definition, the maximum likelihood solution $\hat{\theta}_r^*$ should satisfy the following equations

$$\hat{\theta}_r^* = \sum_u \theta_u^* \delta_{r b_u} = e_r \left[\frac{2e_{\text{in}} \hat{\theta}_r}{\sum_s \hat{\theta}_s^2} + \frac{e_{\text{out}} \sum_{s \neq b_u} \hat{\theta}_s}{\sum_{t<s} \hat{\theta}_t \hat{\theta}_s} \right]^{-1}. \tag{3.18}$$

These equations generally do not permit analytical solutions, except for the special case where all communities have the same total degree. That is, if we do have $e_r = 2E/B$, then our freedom in choosing the value of $\hat{\theta}_r$ is partially restored, since the equation (3.18) will hold as long as all $\hat{\theta}_r$ are identical, i.e. $\hat{\theta}_r = \hat{\theta}$. Otherwise, the maximum likelihood estimates of $\hat{\theta}_r$ are determined by equation (3.18) and we need numerical approximation algorithm to obtain the values of $\hat{\theta}_r^*$. Since the estimator λ_{in}^* and λ_{out}^* also depends on the value of $\hat{\theta}_r$, overall, imposing the planted partition constraint to the DC-SBM complicates the maximum likelihood inference of the model.

However, note that we actually have the freedom to make *a priori* constraints on the values of $\hat{\theta}_r$ as a part of the model. Then, as we should see below, maximum likelihood solutions for parameters of the constrained PP model turn out to have simple expression. The maximum likelihood estimators of the new model are the stationary points of the Lagrangian function,

$$\begin{aligned} \ln P(\boldsymbol{A}|\lambda_{\text{in}}, \lambda_{\text{out}}, \boldsymbol{\theta}, \boldsymbol{b}) = -\lambda_{\text{in}} \sum_r \frac{\hat{\theta}_r^2}{2} - \lambda_{\text{out}} \sum_{r<s} \hat{\theta}_r \hat{\theta}_s + e_{\text{in}} \ln \lambda_{\text{in}} + e_{\text{out}} \ln \lambda_{\text{out}} \\ + \sum_u k_u \ln \theta_u + \sum_r \xi_r (\hat{\theta}_r - \hat{\theta}'_r), \end{aligned} \tag{3.19}$$

where $\hat{\theta}'_r$ are any pre-defined value. Then, the maximum likelihood estimators have the following expressions,

$$\lambda_{\text{in}}^* = \frac{2e_{\text{in}}}{\sum_r \hat{\theta}_r^2}, \tag{3.20}$$

$$\lambda_{\text{out}}^* = \frac{e_{\text{out}}}{\sum_{r<s} \hat{\theta}_r \hat{\theta}_s}, \tag{3.21}$$

$$\theta_u^* = \frac{k_u}{e_{b_u}} \hat{\theta}_{b_u}. \tag{3.22}$$

The estimators λ_{in}^* and λ_{out}^* for the PP model are the same as before when there is no extra constraints of $\hat{\theta}_r$, however, the estimator θ_u^* now changes to be the same as that for the DC-SBM in equation (3.5). We emphasise that the constraints on $\hat{\theta}_r$ need to be predefined before any data is seen. It is really because the values $\{\hat{\theta}_r\}$ are associated

with the model can we restore the convenience of the inference.

Despite that constrained PP model and DC-SBM have the same expression of the maximum likelihood estimator for $\{\theta_u^*\}$, the former is not able to produce consistent degree-correction as the latter does. Specifically, the expected degree of a node in networks generated from an inferred DC-SBM matches with the observed degree in data, i.e.

$$\langle k_u \rangle = \sum_u \theta_u \theta_v \lambda_{b_u b_v} = \frac{k_u}{e_{b_u}} \hat{\theta}_{b_u} \sum_s \hat{\theta}_s \frac{e_{b_u s}}{\hat{\theta}_{b_u} \hat{\theta}_s} = k_u. \tag{3.23}$$

By contrast, the constrained PP model fails to provide accurate degree correction, with the expected degree depending on the predefined constraints $\hat{\theta}_r$. The expression for the average degree in the constrained PP model is

$$\begin{aligned} \langle k_u \rangle &= \sum_v \theta_u \theta_v [\lambda_{\text{in}} \delta_{b_u b_v} + \lambda_{\text{out}} (1 - \delta_{b_u b_v})] \\ &= \theta_u \Big[\lambda_{\text{in}} \hat{\theta}_{b_u} + \lambda_{\text{out}} \sum_{r \neq b_u} \hat{\theta}_r \Big]. \end{aligned} \tag{3.24}$$

If we plug the maximum likelihood estimators $\lambda_{\text{in}}^*, \lambda_{\text{out}}^*$ and θ_u^* into the last equation, we get

$$\langle k_u \rangle = \frac{k_u \hat{\theta}_{b_u}}{e_{b_u}} \left[\frac{2 e_{\text{in}} \hat{\theta}_{b_u}}{\sum_r \hat{\theta}_r^2} + \frac{e_{\text{out}} \sum_{r \neq b_i} \hat{\theta}_r}{\sum_{r<s} \hat{\theta}_r \hat{\theta}_s} \right]. \tag{3.25}$$

The expression of $\langle k \rangle$ in equation (3.25) tells us that networks generated from the inferred PP model will not have the property $\langle k_u \rangle = k_u$, unless the communities are uniform in the sense that the total degree $e_r = \sum_u k_u \delta_{r b_u} = 2E/B$ and $\hat{\theta}_r = \hat{\theta}$ for all communities.

Overall, we have seen that adding the planted partition constraint complicates the maximum likelihood solution of model parameters of DC-SBM. It is important to keep the trade-off between the ease of inference and accurate degree-correction in mind as we move on to discuss the celebrated modularity maximisation approach, which is closely related to the maximum likelihood inference of the PP model.

3.1.2 Modularity maximisation

Modularity maximisation is arguably the most widely used method for community detection. This method is built on the *modularity measure*, which is a quality function measuring how modular a network partition is [46]. Modularity takes the adjacency $\boldsymbol{A}$ and a network partition $\boldsymbol{b}$ as input and returns a numerical score Q, which is defined

as

$$Q(\boldsymbol{A}, \boldsymbol{b}) = \sum_{uv}^{N} (A_{uv} - \langle A_{uv} \rangle) \delta_{b_u b_v}. \tag{3.26}$$

In the definition of the modularity above, the term $\langle A_{uv} \rangle$ is the expected number of edges between node u and v in a chosen *null model*. A null model in the context of community detection is a model that generates network samples with no community structures. Modularity measure compares the observed number of within-group connections to the expectation of the same quantity in a random network ensemble. Therefore, a network partition with large modularity value is modular in the sense that the density of within-group connections is higher than what is expected in a random network.

The choice of the null model can vary depending on the problem at hand. One of the most popular choice is the *configuration model*. The configuration model attaches each node with k_u half-edges according to a given degree sequence $\{k_u\}$, then randomly pairs half-edges together. For a half-edge attaching to node u, its probability of being connected to another node v with degree k_v is $k_v/(2E-1)$. Therefore, the expected number of edges between node u and another node v in the configuration model is

$$\langle A_{uv} \rangle = \sum_{i=1}^{k_u} \frac{k_v}{2E-1} \approx \frac{k_u k_v}{2E}, \tag{3.27}$$

assuming the number of edges E is sufficiently large. Substituting the $\langle A_{uv} \rangle$ in equation (3.27) into the equation (3.26) and adding a normalising factor $1/2E$ in the front, we obtain the celebrated *Newman-Girvan modularity*

$$Q(\boldsymbol{A}, \boldsymbol{b}) = \frac{1}{2E} \sum_{uv}^{N} \left(A_{uv} - \frac{k_u k_v}{2E} \right) \delta_{b_u b_v}, \tag{3.28}$$

where $\delta_{b_u b_v}$ is the Kronecker delta function. The normalising constant scales to the modularity value such that $Q \in [-1, 1]$. The expression of the Newman-Girvan modularity is often written in a slightly different form,

$$Q(\boldsymbol{A}, \boldsymbol{b}) = \frac{1}{2E} \sum_{r}^{B} \left(e_{rr} - \frac{e_r^2}{2E} \right), \tag{3.29}$$

where the summation now goes through each group of network partition $r \in \{1, 2, .., B\}$. This rewritten expression of the Newman-Girvan modularity implies the requirement of forming a community. Because the partition that puts all of the nodes into the same

group has modularity value zero,

$$Q(\boldsymbol{A}, \boldsymbol{b}) = \frac{1}{2E}\left(2E - \frac{(2E)^2}{2E}\right) = 0, \tag{3.30}$$

a community in networks consists of group of nodes satisfying

$$e_{rr} - \frac{e_r^2}{2E} > 0. \tag{3.31}$$

The last inequality means that the observed number of edges inside the group of nodes exceeds the same quantity in a network generated from the corresponding configuration model. Although modularity was originally proposed as a quality function for evaluating network partitions, it was soon realised that we can use modularity as an objective function to search for assortative communities [114]. This is the central idea behind the class of modularity-based methods for community detection. Despite the fact that exact optimisation of modularity is NP-hard [115], many approximate algorithms have been proposed, which can provide good estimates in practice [116,117]. As an example, we explain the Louvain algorithm in Appendix A.8 for finding the maximum modularity solution and show its results in several empirical networks.

However, maximising the Newman-Girvan modularity measure in equation (3.28) is known to suffer from a *resolution limit* problem [51]. The name of the problem is referred to an undesired property of the method: in networks with large size, modularity maximisation might fail to identify communities with small sizes, regardless how strong the community structure is. We will look into the resolution limit problem in Chapter 4. To get around the resolution limit problem, it is common to use the *generalised modularity* Q_γ instead, which is defined as

$$Q_\gamma(\boldsymbol{A}, \boldsymbol{b}) = \frac{1}{2E}\sum_{uv}^{N}\left(A_{uv} - \gamma\frac{k_u k_v}{2E}\right)\delta_{b_u b_v}, \tag{3.32}$$

or equivalently

$$Q_\gamma(\boldsymbol{A}, \boldsymbol{b}) = \frac{1}{2E}\sum_{r}^{B}\left(e_{rr} - \gamma\frac{e_r^2}{2E}\right). \tag{3.33}$$

The generalised modularity is identical to the Newman-Girvan modularity except for an extra *resolution parameter* γ, which was multiplied to the expected number of connections in configuration model. Setting γ to different values has the effect of tuning the resolution of the inferred community structure. If γ is small, then the weight of the negative contribution in Q_γ is small, lowering the criterion for forming

new communities. As a result, maximising Q_γ with a small value of γ will lead to high-resolution network partitions, with relatively larger number of communities and small community sizes. On the contrary, large γ will lead to community structure with low-resolution. The Newman-Girvan modularity is a special case of the generalised modularity with the resolution limit parameter being set to one.

The modularity measure as well as its generalisation had triggered a burst of interest in identifying community structure in real-world networked systems. Despite their widespread use, the modularity maximisation approach also receives criticism because of its heuristic nature. Recently, the generalised modularity function (3.32) was found to be equivalent to the likelihood function of the PP model [118] under certain choices of model parameters. This equivalence result was considered as a theoretical justification of the modularity maximisation approach and it has inspired further extension of the equivalence results, as well as new algorithms which are built on the equivalence [109, 110]. However, as we are going to explain, the extent to which this equivalence holds is rather limited. Moreover, the equivalence result also implies that modularity maximisation shares several limitations of the maximum likelihood approach, making it an unreliable tool for community detection.

3.1.3 On the equivalence between the planted partition model and generalised modularity

We firstly revisit the results developed in [57], in which the generalised modularity Q_γ in equation (3.32) is found to be equivalent to the log-likelihood function of the PP model in equation (3.36). The derivation of the equivalence result begins with writing down the log-likelihood of the DC-SBM,

$$\ln P(\boldsymbol{A}|\boldsymbol{\lambda},\boldsymbol{\theta},\boldsymbol{b}) = \frac{1}{2}\sum_{uv}(A_{uv}(\log\lambda_{b_ub_v} - \theta_u\theta_v\lambda_{b_ub_v}) + \sum_u k_u \ln\theta_u, \tag{3.34}$$

with constants independent of the network partition being discarded. Then, we can obtain the log-likelihood of the PP model by plugging the planted partition constraint into the last equation, which gives

$$\begin{aligned}\ln P(\boldsymbol{A}|\lambda_{\text{in}},\lambda_{\text{out}},\boldsymbol{\theta},\boldsymbol{b}) = \frac{1}{2}\sum_{uv}\Big[A_{uv}(\delta_{b_ub_v}\ln\lambda_{\text{in}} + (1-\delta_{b_ub_v})\ln\lambda_{\text{out}} \\ - \theta_u\theta_v(\lambda_{\text{in}}\delta_{b_ub_v} + \lambda_{\text{out}}(1-\delta_{b_ub_v})\Big] + \sum_u k_u\ln\theta_u.\end{aligned} \tag{3.35}$$

Based on the observation that the modularity in equation (3.32) involves the summation of terms multiplying to the Kronecker delta function, we can expose the relation between the likelihood function in equation (3.36) and the modularity by grouping terms multiplying to the Kronecker delta, which leads to the following expression

$$\begin{aligned}\ln P(\boldsymbol{A}|\lambda_{\text{in}},\lambda_{\text{out}},\boldsymbol{\theta},\boldsymbol{b}) = \frac{\psi}{2}\sum_{uv}\left(A_{uv}-\frac{\lambda_{\text{in}}-\lambda_{\text{out}}}{\log\lambda_{\text{in}}-\ln\lambda_{\text{out}}}\theta_u\theta_v\right)\delta_{b_u b_v} \\ +E\ln\lambda_{\text{out}}-\frac{\lambda_{\text{out}}}{2}\Big(\sum_u\theta_u\Big)^2+\sum_u k_u\ln\theta_u,\end{aligned} \tag{3.36}$$

with $\psi \coloneqq \ln\lambda_{\text{in}} - \log\lambda_{\text{out}}$. Note that only the first summation depends on the network partition $\boldsymbol{b}$ via the Kronecker delta function. Therefore, conditioned on the values of the degree propensity $\boldsymbol{\theta} = \{\theta_u\}$, discarding the last three terms in the last equation does not affect the result of maximising the likelihood with respect to the network partition $\boldsymbol{b}$. If now we set the constraints of $\hat{\theta}_r$ as follows

$$\hat{\theta}'_r = \frac{e_r}{\sqrt{2E}}, \tag{3.37}$$

then according to our analysis in Section 3.1.1 regarding the maximum likelihood inference with PP model, the maximum likelihood estimator of θ^*_u in equation (3.22) becomes

$$\theta^*_u = k_u/\sqrt{2E}. \tag{3.38}$$

Substituting the last equation back to the log-likelihood function in (3.36), we get

$$\ln P(\boldsymbol{A}|\gamma^*,\boldsymbol{\theta'},\boldsymbol{b}) = \frac{\psi}{2}\sum_{uv}\left(A_{uv}-\gamma^*\frac{k_u k_v}{2E}\right) \propto Q_\gamma(\boldsymbol{A},\boldsymbol{b}), \tag{3.39}$$

with independent constants $E\log\lambda_{\text{out}}$, $-\lambda_{\text{out}}(\sum_u\theta_u)^2$ and $\sum_u k_u\log\theta_u$ being dropped. The resolution limit parameter γ^* has the following expression:

$$\gamma^* = \frac{\lambda_{\text{in}}-\lambda_{\text{out}}}{\ln\lambda_{\text{in}}-\ln\lambda_{\text{out}}}. \tag{3.40}$$

Equation (3.39) implies that, conditioned on the values of λ_{in} and λ_{out}, the network partition $\boldsymbol{b}^*$ that maximises the generalised modularity Q_{γ^*} also maximises the likelihood function of the PP model, when the choice of constraint in (3.37) is made. Based on this observation, the author of [57] claimed that modularity maximisation approach is equivalent to implementing the maximum likelihood principle with the PP model.

The establishment of the equivalence result has several implications. Firstly, although

modularity maximisation was heuristically motivated, being equivalent to the maximum likelihood approach means the former is a consistent method in a sense that the inferred partition will converge to the underlying truth as we observed sufficient data. Moreover, the expression of γ^* in equation (3.40) can be used as a principled way to choose the resolution limit parameter. When the values of λ_{in} and λ_{out} are not available, as proposed in [57], we might estimate them from data using the following expressions

$$\lambda^*_{\text{in}} = \frac{e_{\text{in}}}{\sum_r e_r^2/2E},\ \lambda^*_{\text{out}} = \frac{e_{\text{out}}}{\sum_{r<s} e_r e_s/2E}, \tag{3.41}$$

which are simply the maximum likelihood estimators in equation (3.20) with $\hat{\theta}_r$ being set to $e_r/\sqrt{2E}$. The corresponding estimate of the resolution parameter is then

$$\gamma^*(\lambda^*_{\text{in}}, \lambda^*_{\text{out}}) = \frac{\lambda^*_{\text{in}} - \lambda^*_{\text{out}}}{\ln \lambda^*_{\text{in}} - \ln \lambda^*_{\text{out}}}. \tag{3.42}$$

Although the equivalence result above provides useful insight into the modularity maximisation approach, we argue that it should not be interpreted as the advocacy for modularity-based methods. In fact, the extent to which the equivalence holds has been overlooked. In particular, one key step for developing the equivalence is to make the choices of constraints on $\hat{\theta}_r = e_r/\sqrt{2E}$. These choices are made simply for constructing the equivalence result, but they do not have any principled justifications. More importantly, recall that when we derivate maximum likelihood estimators for the PP model in Section 3.1.1, we also make subjective choices for the values of $\hat{\theta}_r$ such that the maximum likelihood solutions admit cleaner expressions. However, our choices have to be determined before any data is observed. By contrast, the choices of $\hat{\theta}_r$ which induces the equivalence as given in equation (3.37) obviously depend on the observed network $\boldsymbol{A}$ via the degree sequence $\boldsymbol{k}$,

$$\hat{\theta}'_r = e_r/\sqrt{2E} = \sum_u k_u \delta_{r b_u}/\sqrt{2E}. \tag{3.43}$$

In general, any choices of $\hat{\theta}_r$ that differs from $\hat{\theta}_r = e_r/\sqrt{2E}$ will invalidate the equivalence result. When no constraints of $\hat{\theta}_r$ is made, the maximum likelihood principle requires us to use the estimates $\hat{\theta}^*_r$ satisfying equation (3.18), which is unlikely to coincide with the choice made by modularity maximisation in equation (3.37), with the only exception that all $e_r = 2E/B$. Last but not least, when we apply the maximum likelihood principle, the parameters of λ^*_{in}, λ^*_{out} and θ^*_u need to be inferred simultaneously, and their maximum likelihood estimators are all dependent of the network partition $\boldsymbol{b}$. Therefore, optimising the objective function in equation (3.36) is princi-

pally different from optimising the one in equation (3.39). For the reasons above, we argue that the claimed equivalence between modularity maximisation and maximum likelihood approach in [57] is an overstatement. Such nuance was firstly brought up in [83] and a more recent work [50] provides numerical evidence for the discrepancy in the performance of modularity maximisation and the maximum likelihood inference with the PP model.

Furthermore, suppose we take a different proposition that modularity maximisation is just related to maximum likelihood inference, this relation actually implies that modularity maximisation inherits the disadvantages of the maximum likelihood approach. One problem of the maximum likelihood approach is the tendency of overfitting data. As a result, maximum likelihood usually requires proper regularisation to work well in practice. Being equivalent, or just related to the maximum likelihood just means modularity maximisation also suffers from the overfitting problem, which has been widely reported in the literature [16, 119]. Besides, the planted partition constraint in equation (3.7) implies a rather restricted pattern: the ratio of within- and between-group connections is the same for every community. Since modularity maximisation is equivalent to conducting inference when such restricted constraint is in place, the performance of modularity maximisation could degenerate especially in networks with properties at odds with the prescribed restrictive pattern.

In short, although the equivalence result provides an alternative derivation of the modularity maximisation method, the equivalence is rather tenuous and it implies modularity maximisation inherits weaknesses from the maximum likelihood approach with no improvements. In the next section, we shall provide a better solution for extracting assortative structures. Our method is based on Bayesian inference with the PP model. We will show our approach is advantageous in term of preventing overfitting and also in the ability of modelling more general assortative pattern.

3.2 Bayesian inference: posterior probability of planted partition models

Instead of doing maximum likelihood, we propose to consider Bayesian inference with the PP model, where the goal is to draw samples or to optimise the posterior distribution

$$P(\boldsymbol{b}|\boldsymbol{A}) = \frac{P(\boldsymbol{A}|\boldsymbol{b})P(\boldsymbol{b})}{P(\boldsymbol{A})} \tag{3.44}$$

For the PP model, the marginal likelihood is obtained by computing the integral

$$\begin{aligned} P(\boldsymbol{A}|\boldsymbol{b}) &= \int P(\boldsymbol{A}, \lambda_{\text{in}}, \lambda_{\text{out}}, \boldsymbol{\theta}|\boldsymbol{b}) d\lambda_{\text{in}} d\lambda_{\text{out}} d\boldsymbol{\theta} \\ &= \int P(\boldsymbol{A}|\lambda_{\text{in}}, \lambda_{\text{out}}, \boldsymbol{\theta}, \boldsymbol{b}) P(\lambda_{\text{in}}, \lambda_{\text{out}}|\boldsymbol{b}) P(\boldsymbol{\theta}|\boldsymbol{b}) d\lambda_{\text{in}} \lambda_{\text{out}} d\boldsymbol{\theta}, \end{aligned} \tag{3.45}$$

where $P(\boldsymbol{A}|\lambda_{\text{in}}, \lambda_{\text{out}}, \boldsymbol{\theta}, \boldsymbol{b})$ is the likelihood function in equation (3.8), and $P(\lambda_{\text{in}}, \lambda_{\text{out}}|\boldsymbol{b})$, $P(\boldsymbol{\theta}|\boldsymbol{b})$ are priors for model parameters $\boldsymbol{\theta}$, and λ_{in} and λ_{out}. Just like what we have done for general models in the Chapter 2, we need to specify the priors carefully such that we do not introduce any intrinsic bias in the posterior distribution. We begin with the prior for the degree propensity parameter $\boldsymbol{\theta}$. For the ease of inference, we consider the following constraints

$$\hat{\theta}_r = 1, \ \ \forall r \in \{1, 2.., B\}. \tag{3.46}$$

This choice is associated with the model and does not depend on data. With these constraints on $\hat{\theta}_r$, the parameter λ_{in} becomes twice the expected degree (equivalently twice the number of edges) within each community, since

$$\langle e_{rr} \rangle = \sum_{uv} \theta_u \theta_v \lambda_{\text{in}} \delta_{r b_u} \delta_{s b_v} = \hat{\theta}_r^2 \lambda_{\text{in}} = \lambda_{\text{in}}. \tag{3.47}$$

Similarly, λ_{out} is the expected number of edges between any pair of distinct communities. With the constraint on $\{\theta_u\}$ in equation (3.46), maximum-entropy prior of $\boldsymbol{\theta}$ is the uniform distribution on $n_r - 1$ regular simplex,

$$P(\boldsymbol{\theta}|\boldsymbol{b}) = \prod_r (n_r - 1)! \delta_{\sum_u \theta_u \delta_{r b_u}, 1}. \tag{3.48}$$

For λ_{in}, λ_{out} the maximum-entropy prior is the exponential distribution

$$\begin{aligned} P(\lambda_{\text{in}}|\bar{\lambda}) &= e^{-\lambda_{\text{in}}/2\bar{\lambda}}/(2\bar{\lambda}), \\ P(\lambda_{\text{out}}|\bar{\lambda}) &= e^{\lambda_{\text{out}}/\bar{\lambda}}/\bar{\lambda}, \end{aligned} \tag{3.49}$$

where $\bar{\lambda} = 2E/(B(B+1))$. Performing the integral in equation (3.45) with the likelihood function in equation (3.8) and priors in equations (3.48) - (3.49) gives[3]

$$P(\boldsymbol{A}|\bar{\lambda}, \boldsymbol{b}) = \frac{e_{\text{in}}! e_{\text{out}}!}{2\bar{\lambda}^2 \left[\frac{B}{2} + \frac{1}{2\bar{\lambda}}\right]^{e_{\text{in}}+1} \left[\binom{B}{2} + \frac{1}{\bar{\lambda}}\right]^{e_{\text{out}}+1}} \times \prod_r \frac{(n_r - 1)!}{(e_r + n_r - 1)!} \times \frac{\prod_u k_u!}{\prod_{u<v} A_{uv}! \prod_u A_{uu}!!}. \tag{3.50}$$

[3]Details of the derivation are provided in Appendix A.6.

Just like the canonical DC-SBM, the marginal likelihood of the PP model has an alternative microcanonical formulation. The marginal likelihood above can be rewritten as follows

$$P(\boldsymbol{A}|\bar{\lambda},\boldsymbol{b}) = P(\boldsymbol{A}|\boldsymbol{e},\boldsymbol{k},\boldsymbol{b})P(\boldsymbol{k}|\boldsymbol{e},\boldsymbol{b})P(\boldsymbol{e}|e_{\text{in}},e_{\text{out}},\boldsymbol{b})P(e_{\text{in}}|\bar{\lambda})P(e_{\text{out}}|\bar{\lambda}), \tag{3.51}$$

with

$$P(\boldsymbol{A}|\boldsymbol{e},\boldsymbol{k},\boldsymbol{b}) = \frac{\prod_{r<s} e_{rs}! \prod_r e_{rr}!!}{\prod_r e_r!!} \times \frac{\prod_u k_u!}{\prod_{u<v} A_{uv}! \prod_u A_{uu}!!} \tag{3.52}$$

being the likelihood of the microcanonical DC-SBM as we introduced in Section 2.3, and the other four terms correspond to the priors of the degree sequence $\boldsymbol{k}$ and $\boldsymbol{e}$:

$$P(\boldsymbol{k}|\boldsymbol{e},\boldsymbol{b}) = \prod_r \frac{e_r!(n_r-1)!}{(e_r+n_r-1)!}, \tag{3.53}$$

$$P(\boldsymbol{e}|e_{\text{in}},e_{\text{out}},\boldsymbol{b}) = \frac{e_{\text{in}}!}{B^{e_{\text{in}}}\prod_r (e_{rr}/2)!} \times \frac{e_{\text{out}}!}{\binom{B}{2}^{e_{\text{out}}}\prod_{r<s} e_{rs}!}, \tag{3.54}$$

$$P(e_{\text{in}}|\bar{\lambda},\boldsymbol{b}) = \frac{(B\bar{\lambda})^{e_{\text{in}}}}{(B\bar{\lambda}+1)^{e_{\text{in}}+1}}, \tag{3.55}$$

$$P(e_{\text{out}}|\bar{\lambda},\boldsymbol{b}) = \frac{\left(\binom{B}{2}\bar{\lambda}\right)^{e_{\text{in}}}}{\left(\binom{B}{2}\bar{\lambda}+1\right)^{e_{\text{out}}+1}}. \tag{3.56}$$

Note that because the microcnanoical model generates networks with the exact degree sequence $\boldsymbol{k}$, the inferred model will have accurate degree-correction even though we have made pre-defined constraints on $\hat{\theta}_r$. That is, Bayesian inference with the PP model does not have the problem of inaccurate degree-correction which occurs to the maximum likelihood inference. Moreover, the microcanonical interpretation allows us to replace the parametric priors with non-parametric ones. Specifically, the prior for e_{in} and e_{out} in equation (3.55) and (3.56) are geometric distributions with mean $\bar{\lambda}$. We can proceed to replace the geometric priors $P(e_{\text{in}},e_{\text{out}}|\bar{\lambda},\boldsymbol{b}) = P(e_{\text{in}}|\bar{\lambda},\boldsymbol{b})P(e_{\text{out}}|\bar{\lambda},\boldsymbol{b})$ with the following non-parametric prior

$$P(e_{\text{in}},e_{\text{out}}|\boldsymbol{b}) = P(e_{\text{in}},e_{\text{out}}|E,\boldsymbol{b})P(E), \tag{3.57}$$

where

$$P(e_{\text{in}},e_{\text{out}}|E,\boldsymbol{b}) = \left(\frac{1}{E+1}\right)^{1-\delta_{B,1}} \tag{3.58}$$

is a uniform distribution of splitting total E edges into e_{in} within-group edges and e_{out} between-group edges. The prior for the number of edges $P(E)$ can be arbitrarily

chosen since it will just introduce a multiplying constant independent of the network partition $\boldsymbol{b}$. In summary, the marginal probability distribution of the PP model takes the following decomposition

$$P(\boldsymbol{A}|\boldsymbol{b}) = P(\boldsymbol{A}|\boldsymbol{e},\boldsymbol{k},\boldsymbol{b})P(\boldsymbol{k}|\boldsymbol{e},\boldsymbol{b})P(\boldsymbol{e}|e_{\text{in}},e_{\text{out}},\boldsymbol{b})P(e_{\text{in}},e_{\text{out}}|E,\boldsymbol{b})P(E) \tag{3.59}$$

which reads as

$$P(\boldsymbol{A}|\boldsymbol{b}) = \frac{e_{\text{in}}!e_{\text{out}}!}{\left(\frac{B}{2}\right)^{e_{\text{in}}}\binom{B}{2}^{e_{\text{out}}}(E+1)^{1-\delta_{B,1}}} \times \prod_r \frac{(n_r-1)!}{(e_r+n_r-1)!} \times \frac{\prod_u k_u!}{\prod_{u<v}A_{uv}!\prod_u A_{uu}!!}. \tag{3.60}$$

Together with the prior for network partition $P(\boldsymbol{b})$ we defined in equation (2.31), we can now fit the PP model to data with the MCMC algorithm we described in Section 2.5. Fitting the PP model to data by sampling or maximising the posterior probability will not overfit data, and we can assure this point by exploiting the connection between Bayesian inference and information theory, just as we did in Section 2.4.2. Alternatively, we can see why the Bayesian approach is more robust than the maximum likelihood approach by directly looking at the joint probability

$$P(\boldsymbol{A},\boldsymbol{b}) = \underbrace{P(\boldsymbol{A}|\boldsymbol{e},\boldsymbol{k},\boldsymbol{b})}_{\text{model likelihood}}\underbrace{P(\boldsymbol{e},\boldsymbol{k},\boldsymbol{b})}_{\text{prior}}, \tag{3.61}$$

where the prior is

$$P(\boldsymbol{e},\boldsymbol{k},\boldsymbol{b}) = P(\boldsymbol{k}|\boldsymbol{e},\boldsymbol{b})P(\boldsymbol{e}|e_{\text{in}},e_{\text{out}}|\boldsymbol{b})P(e_{\text{in}},e_{\text{out}}|E,\boldsymbol{b})P(\boldsymbol{b}). \tag{3.62}$$

The maximum likelihood approach only cares about the quality of fit, which corresponds to the likelihood term in the joint probability in equation (3.61). From a Bayesian point of view, that is equivalent to adopting an inappropriate constant prior $P(\boldsymbol{e},\boldsymbol{k},\boldsymbol{b}) = 1$ for all possible parameter combinations $\{\boldsymbol{k},\boldsymbol{e},\boldsymbol{b}\}$. Because the modelling capacity of the model naturally grows with the order of the model, which represented by the number of communities B, maximum likelihood is likely to prefer models which are overly complicated. The Bayesian approach, by contrast, chooses the prior $P(\boldsymbol{e},\boldsymbol{k},\boldsymbol{b})$ carefully to represent our prior knowledge. The prior knowledge plays a role of regularisation on the maximum likelihood approach: although complicated models can achieve better fit to data, i.e. large likelihood values $P(\boldsymbol{A}|\boldsymbol{e},\boldsymbol{k},\boldsymbol{b})$, their corresponding prior value $P(\boldsymbol{e},\boldsymbol{k},\boldsymbol{b})$ are small. This is because as models become complicated (e.g. the number of communities B increases), the number of possible parameter combinations $\{\boldsymbol{e},\boldsymbol{k},\boldsymbol{b}\}$ increases. Since the sum of the prior probability must be one, complex

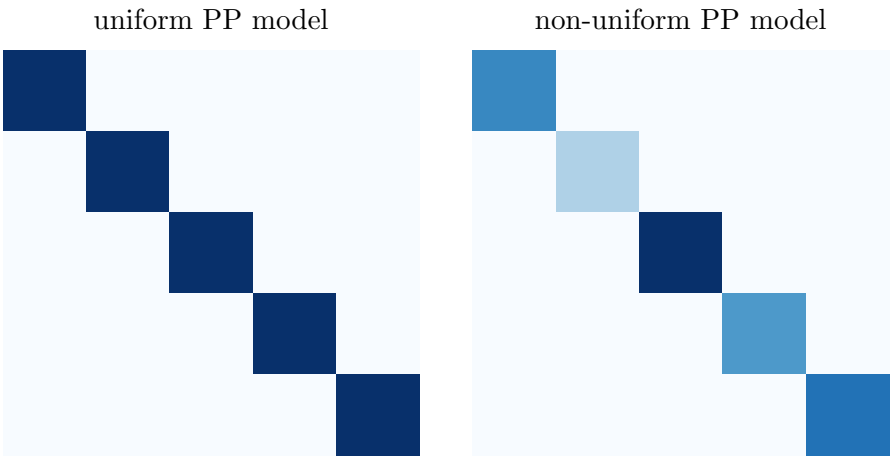


Figure 3-2: Visualisation of possible connection matrices $\boldsymbol{\lambda} = \{\lambda_{rs}\}$ under the uniform and non-uniform planted partition constraints.

models (large number of communities B) permit large number of possible parameter combinations, resulting in smaller prior probabilities for each possible parameter combination in the uniform prior distribution. Such comparison between maximum likelihood and Bayesian inference, together with the information-theoretical explanation in Section 2.4.2 about the Bayesian Occam's razor effect, assure us to use Bayesian inference without worrying about overfitting data. In the next section, we will provide further numerical evidence for the robustness of Bayesian inference approach for community detection.

Non-uniform PP model

Reminds that our goal is to develop a restricted version of SBMs which focuses on assortative structures. The planted partition constraint in equation (3.7) indeed reduces the flexibility of the general model, leaving the assortativity and disassortativity as the only two structures that the model can describe. However, the assortative structures prescribed by the planted partition constraint seems overly restricted, with every community having the same within- and between-communities connection rate. As a compromise between the general model and the regular assortativity, we propose to adopt a non-uniform planted partition constraint,

$$\lambda_{rs} = \lambda_r \delta_{rs} + \lambda_{\text{out}}(1 - \delta_{rs}). \tag{3.63}$$

As illustrated in Fig. 3-2, compared to the uniform case, the constraint in equation (3.63) allows each community to acquire its own expected number of within-

community connections. The marginal likelihood of the DC-SBM under the non-uniform planted partition constraint reads as

$$P(\boldsymbol{A}|\boldsymbol{b}) = \frac{e_{\text{out}}! \prod_r e_{rr}!!}{\binom{B}{2}^{e_{\text{out}}} (E+1)^{1-\delta_{B,1}}} \times \binom{B + e_{\text{in}} - 1}{e_{\text{in}}}^{-1} \times \prod_r \frac{(n_r - 1)!}{(e_r + n_r - 1)!} \times \frac{\prod_u k_u!}{\prod_{u<v} A_{uv}! \prod_u A_{uu}!!}. \tag{3.64}$$

The derivation of this expression is similar to that of the uniform case and we defer the details to Appendix A.7. From now on, we will refer to the DC-SBM with the non-uniform planted partition constraint as the non-uniform PP model, and to its uniform counterpart as the uniform PP model. Having defined our models, in the next section, we will demonstrate their performance in synthetic and empirical networks.

3.3 Numerical experiments

3.3.1 Results for synthetic networks

We firstly show how PP models perform in synthetic networks with known community structures. We compare the result of the PP models with the DC-SBM, as well as the modularity maximisation approach with different values of the resolution parameter. We will focus on the risk of overfitting data, i.e. the potential of identifying non-existing communities. To this end, we generate networks with known assortative structures from the uniform PP model. We have examined networks with a various number of communities while the community sizes are set to be identical, i.e. $n_r = N/B,\ \forall r \in \{1, 2, .., B\}$. Nodes are assumed to have uniform degree propensities

$$\theta_u = \frac{1}{n_r} = \frac{B}{N},\ \ \forall u \in \{1, 2, .., N\}\ , \forall r \in \{1, 2, .., B\}. \tag{3.65}$$

The expected number of edges between communities are parameterised as follows

$$\lambda_{\text{in}} = \Big(1 + (B-1)\epsilon\Big)\langle k \rangle \frac{N}{B}, \quad \lambda_{\text{out}} = (1-\epsilon)\langle k \rangle \frac{N}{B}. \tag{3.66}$$

The ϵ parameter takes values in the interval $[0, 1]$ and it controls the strength of assortative structures. When $\epsilon = 0$, the expected probability of an edge within-community is the same as that of an between-community edge,

$$\langle p_{\text{in}} \rangle = \theta_u \theta_v \lambda_{\text{in}} = B\frac{\langle k \rangle}{N} = \lambda_{\text{out}} \theta_u \theta_w = \langle p_{\text{out}} \rangle, \tag{3.67}$$

for any $u, v, w \in \mathcal{N}$ with $b_u = b_v$ and $b_u \neq b_w$. Therefore, networks generated from the model with $\epsilon = 0$ are random networks with no community structures. When $\epsilon = 1$,

$\lambda_{\text{in}} > 0$ but $\lambda_{\text{out}} = 0$, we have an extreme case of assortativty, in which edges are only allowed within communities. Values of ϵ between 0 and 1 correspond to all other intermediate cases between randomness and perfect assortativity. To fit the data to PP model and DC-SBM, we find the MAP solution by running the simulated annealing scheme as described in Section 2.5 and we refer to the documentation of the *graph-tool* library [103] for details of the implementation.

For maximising modularity, we sample from the target distribution

$$P(\boldsymbol{b}|\boldsymbol{A}) = \frac{e^{\beta Q_\gamma(\boldsymbol{b},\boldsymbol{A})}}{Z(\boldsymbol{A})}, \tag{3.68}$$

where $Z(\boldsymbol{A}) = \sum_{\boldsymbol{b}} e^{\beta Q_\gamma(\boldsymbol{b},\boldsymbol{A})}$ is a normalising constant and $Q_\gamma(\boldsymbol{b},\boldsymbol{A})$ is the generalised modularity as defined in equation (3.32). With the connection between the generalised modularity and the PP model in mind, the inverse temperature parameter β is set to be $E(\ln\lambda_{\text{in}} - \ln\lambda_{\text{out}})$ such that the posterior will be proportional to the likelihood of the true underlying model, i.e. $P(\boldsymbol{b}|\boldsymbol{A}) \propto P(\boldsymbol{A}|\lambda_{\text{in}}, \lambda_{\text{out}}, \boldsymbol{\theta}, \boldsymbol{b})$, as long as we choose the

$$\gamma = \gamma_{\text{true}} = \frac{\lambda_{\text{in}} - \lambda_{\text{out}}}{\ln\lambda_{\text{in}} - \ln\lambda_{\text{out}}}. \tag{3.69}$$

We also examine the results obtained with the original Newman-Giravn modularity $Q_{\gamma=1}$, as well as the generalised modularity with the maximum likelihood estimate of γ,

$$\gamma = \gamma_{\text{fit}} = \frac{\lambda^*_{\text{in}} - \lambda^*_{\text{out}}}{\ln\lambda^*_{\text{in}} - \ln\lambda^*_{\text{out}}}, \tag{3.70}$$

where the value of λ^*_{in} and λ^*_{out} are

$$\lambda^*_{\text{in}} = \frac{Be_{\text{in}}}{E}, \quad \lambda_{\text{out}} = \frac{Be_{\text{out}}}{(B-1)E}, \tag{3.71}$$

using the maximum likelihood estimators we derived in equation (3.20) and making the following assumption

$$\hat{\theta}_r = \frac{e_r}{\sqrt{2E}} = \frac{\sqrt{2E}}{B}. \tag{3.72}$$

The inferred number of communities is plotted against the true number of groups in Fig. 3-3. In our experiment, we set the assortative parameter sufficiently strong ($\epsilon = 0.8$) such that the structure is easy to detect [4]. As can be seen from Fig. 3-

[4]Community detection in networks generated from the uniform PP model undergoes a *phase transition* phenomenon. The phases of the model are related to the detectability of the planted community structures. Under our parameterisation, the uniform PP model is in an *undetectable phase* if the as-

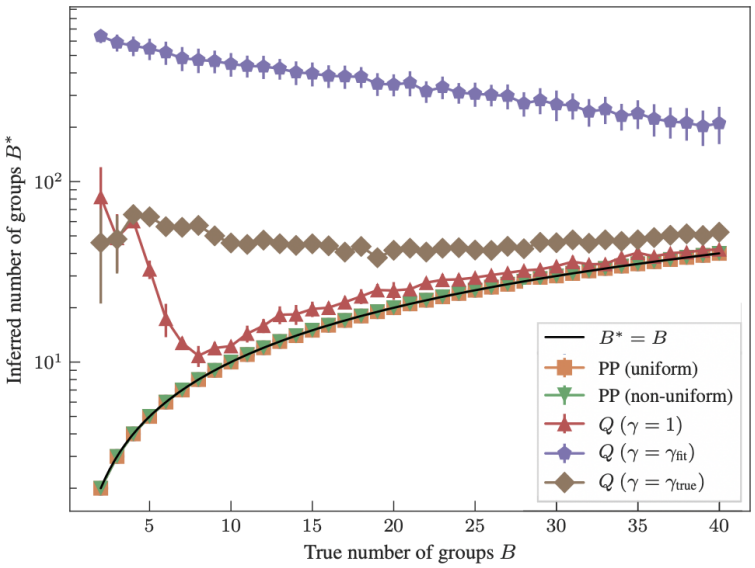


Figure 3-3: Inferred number of groups as a function of the true number of groups. Networks are generated from the uniform PP model with the parameterisation given in text. The error bars show the standard deviation of the distribution. Networks consist of 10^5 nodes and all communities have the same size $n_r = N/B$, average degree $\langle k \rangle = 5$.

3, Bayesian inference of the PP models (both uniform and non-uniform) consistently manage to identify the correct number of groups. By contrast, all versions of modularity maximisation systematically overfit, manifested by significantly larger number of groups compared to the planted number of communities.

The difference between the performance of different versions of modularity is a bit bizarre. Having seen the connection between the generalised modularity and the likelihood function of the uniform PP model, one might expect the generalised modularity Q_γ to perform better than the original modularity $Q_{\gamma=1}$, because it resembles the implementation of the maximum likelihood principle. However, from the perspective of overfitting data, the original modularity $Q_{\gamma=1}$ is the least problematic one among all three versions of modularity we have tested. The modularity with the fitted value $\gamma = \gamma_{\text{fit}}$ gives the worst performance, which are orders of magnitude wrong compared to the correct number of communities as indicated by the purple pentagons. Although maximising the generalised modularity $Q_{\gamma_{\text{true}}}$ is equivalent to maximising the likelihood of the underlying model, because maximum likelihood has the tendency of overfitting, it is not surprising to see $Q_{\gamma_{\text{true}}}$ prefers exaggerated results (given by brown diamond

sortative strength ϵ is below the threshold $\epsilon^* = 1/\sqrt{k}$ [65]. We will discuss more on the detectability phase-transition in Chapter 5. For our experiment here, we set the value of ϵ such that the model is in the detectable phase.

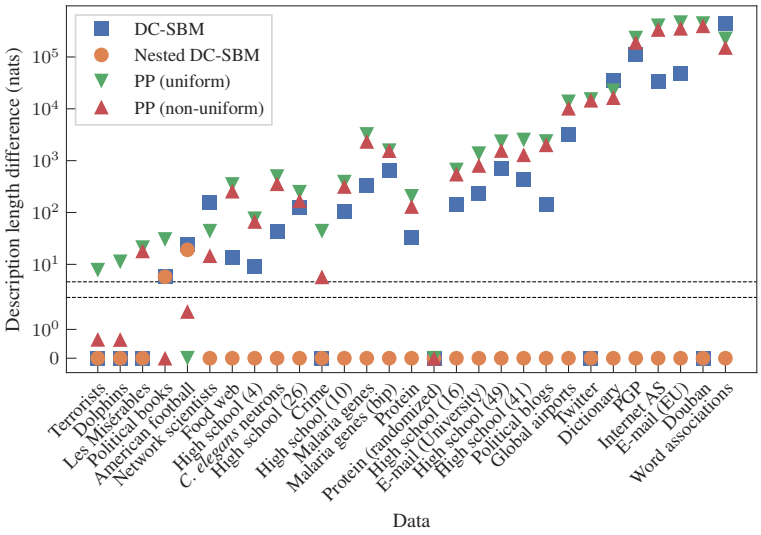


Figure 3-4: Difference in description length between the best fitting model and other model variants of interest. Models are fitted to a selection of 29 networks obtained from the KONECT repository [120]. The best fitting model appears in the bottom. For reference, the values in $\ln 10$ and $\ln 100$ are shown as dashed lines.

markers). Although the overfitting problem seems to be less severe for the original modularity $Q_{\gamma=1}$ (given by red traingles) compared to its generalised generalised versions, the discrepancy between the true and inferred number of communities is notable, especially when the true number of communities is small, say when $B < 8$.

3.3.2 Results for real-world networks

The majority of traditional community detection algorithms search for assortative structures exclusively and ignore other possible structures. However, it is possible that there exist other non-assortative structures being the better description of data. Here, we conduct model selection with our assortative-constrained variants and general SBMs, including the DC-SBM as well as the Nested DC-SBM, using a set of empirical networks. A comparison study like this allows us to investigate whether assortativity is indeed the dominant pattern in data.

We choose 29 networks from a variety of scientific domains, obtained from the KONECT repository [120]. For each network, we find the MAP solution of network partition given by different models, then compare the description length of each pair of model and their inferred network partition. Fig. 3-4 summaries the difference in description length between each of model variants and the model with the best fit to data. Points in the bottom indicate the best fitting model in each network. It is clear that most of the time

general DC-SBM are selected as the best model with a high level of confidence. This implies that simple assortative structures (both uniform and non-uniform) prescribed by the PP models are too simplistic to describe the pattern in our dataset.

However, there are exceptions where PP models are preferred over the general models. One example is an online co-purchase networks of American political books around the time of 2004 presidential election [121]. In this network, nodes are political books available in an online bookselling website and edges represent frequent co-purchases of books by the same customers. As shown in the top panel in Fig. 3-5, the Nested DC-SBM finds a partition with three groups, aligning closely to the known labelling of books: books are either liberal, neutral or conservative. However, the non-uniform PP model manages to compress the data even further by splitting the group in the middle into two sub-groups.

Another example is the American college football network [11] which represents the schedule of matches among college football teams. This is the only case we find the uniform PP model is selected as the best fitting model. In the bottom panel in Fig 3-5, we show the community structure inferred by the uniform PP model and the Nested DC-SBM respectively. Although the partition given by the two models are quite similar to each other, the uniform PP model achieves a slightly better fit in terms of description length. The advantage of the uniform PP model is well supported by the generating process of this network. When this dataset was constructed, college football teams were divided into conferences. Matches are arranged more frequently between teams in the same conference, leading to the highly assortative structure in the observed network. Moreover, the assortativity is relatively regular because edges in the network represent match relationship during the regular season of games. Therefore, it is not surprised to have small variance in the arrangement of number of matches across different teams.

Whether assortative structures are absolutely the dominant pattern in these two examples is debatable, as the advantage of PP models in the description length are not significant, and the level of confidence for rejecting alternative hypothesis is always subjective. However, it is clear that the Bayesian approach does not always favour complicated models, and simple models can be selected as long as they are sufficient to describe the pattern in data. Notice that in Fig. 3-4. for the terrorists network [122] and the social network of dolphins [123], general SBMs achieve better fit to the data than PP models but with only little advantage. This means there is no sufficient information that allows us to conclude which kinds of structure provides the best summary of the pattern in data. In these cases, we should consider general SBMs and PP models as equally possible generating processes of the observed networks.

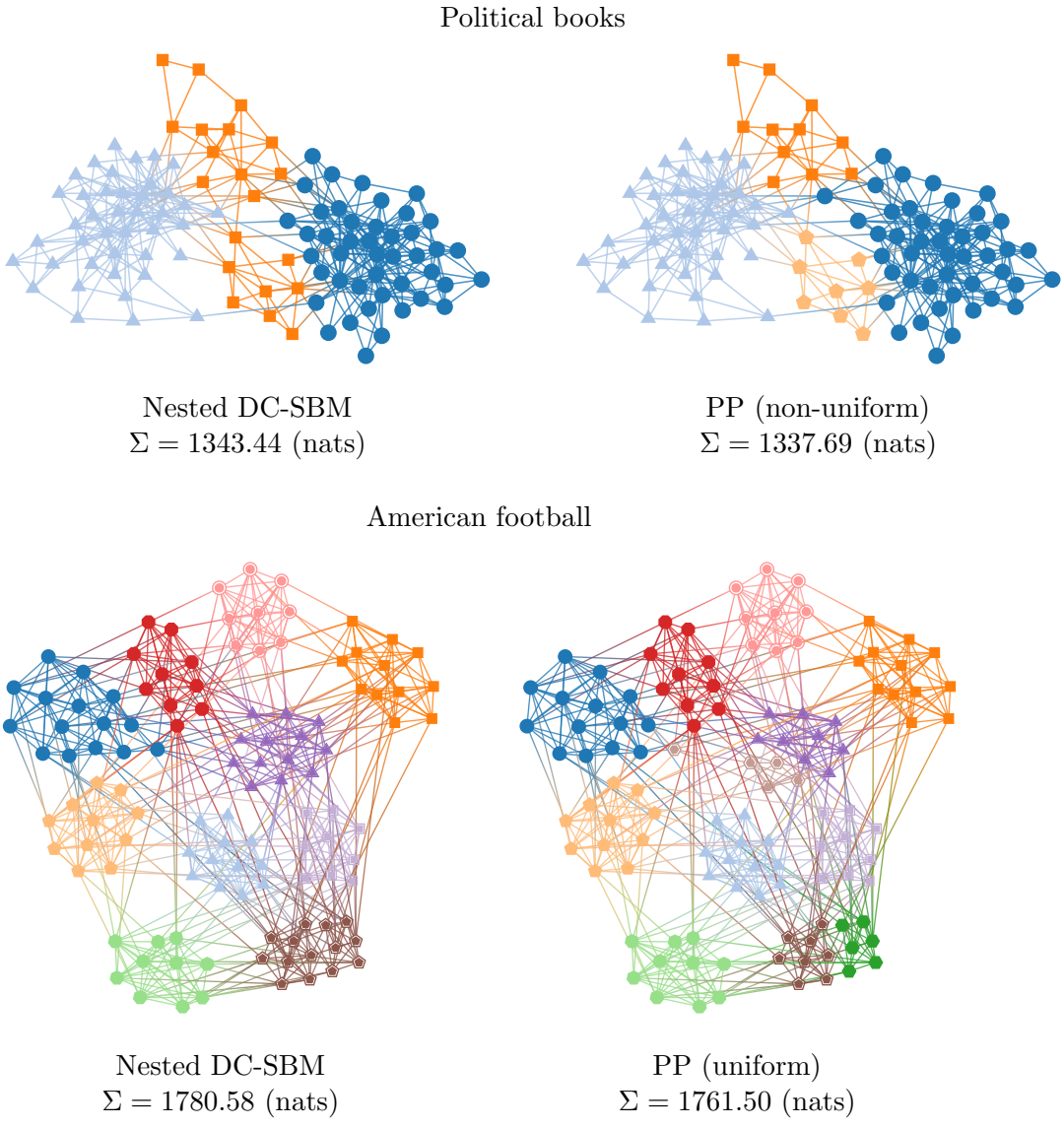


Figure 3-5: Community structures inferred by the Nested DC-SBM and PP models in a network of co-purchases political books [121] and American college football team [11] respectively. Legends show the description length of partitions and their corresponding models.

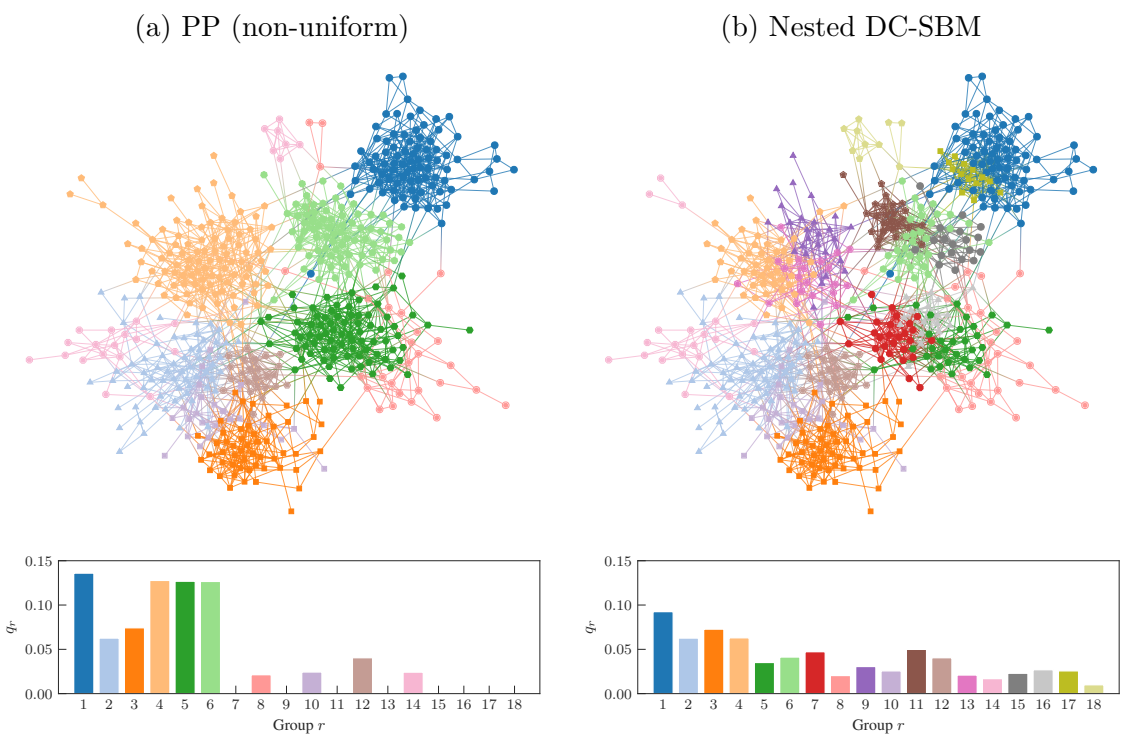


Figure 3-6: Inferred community structure in a social network of high school students [124], obtained with the non-uniform PP model and the Nested DC-SBM. The bottom panels show the community-wise modularity value as defined in the text. The group colours are chosen to maximise the matching between both partitions, as described in Ref. [125], and the same colours are used in the bottom panels.

In the rest of networks we have examined (25 out of 29), general models significantly outperform PP models in terms of the description length. Interestingly, general model can achieve better fit to data even when the uncovered structure is indeed very assortative. This is exemplified by a social network of high school students [124] and we visualise the inferred community structure in this network in Fig. 3-6. Although the overall modularity value of the partition given by the Nested DC-SBM is not as high as that given by the non-uniform PP model, each group of the partition actually acquires positive community-wise modularity value, which is computed as

$$q_r = \frac{1}{2E}\left(e_{rr} - \frac{e_r^2}{2E}\right), \tag{3.73}$$

such that $\sum_r q_r = Q$ holds. However, the Nested DC-SBM allows better data compression, therefore is prefered according to the MDL principle. If we look into the results of the two model closely, to a large extent, the partition given by the Nested DC-SBM can be obtained by subdividing that of the non-uniform PP model. For instance, in Fig 3-6, the No.7 community (nodes filled with red colouring) in the partition given by the Nested DC-SBM is merged into the No.5 community (nodes filled with darkgreen colouring) in that of the non-uniform PP model. This can be explained by the fact that the DC-SBM can leverage the preference of connections between different communities as additional evidence for their existence alongside the assortative pattern.

In Fig. 3-7, we provide further information about the results obtained on the set of empirical networks, including the number of inferred communities B, the modularity value of partitions Q, and the normalised maximum overlap distant [125] $d(\boldsymbol{b}, \boldsymbol{b}')$ between the best fitting partition $\boldsymbol{b}$ and other partitions $\boldsymbol{b}'$. The overlap distant is computed as

$$d(\boldsymbol{b}, \boldsymbol{b}') = 1 - \frac{1}{N}\max_{\phi}\sum_{u}\delta_{b_u,\phi(b'_u)}, \tag{3.74}$$

where $\phi(r)$ here is a bijection between the group labels of $\boldsymbol{b}$ and $\boldsymbol{b}'$ such that the distance is the maximum value over all possible permutations of labels. In terms of the inferred number of communities, PP models generally give conservative results, concluding smaller number of groups compared to general SBMs. We also find that the partitions given by PP models and general SBMs are rarely similar according to the partition distance, even when the modularity values of the corresponding partitions are close. If we compare the modularity values of the best fitting model (most of the time the Nested DC-SBM) to that of PP models, they are similar in some cases, but in examples like the Douban social network [126], political blogs [127], and internet

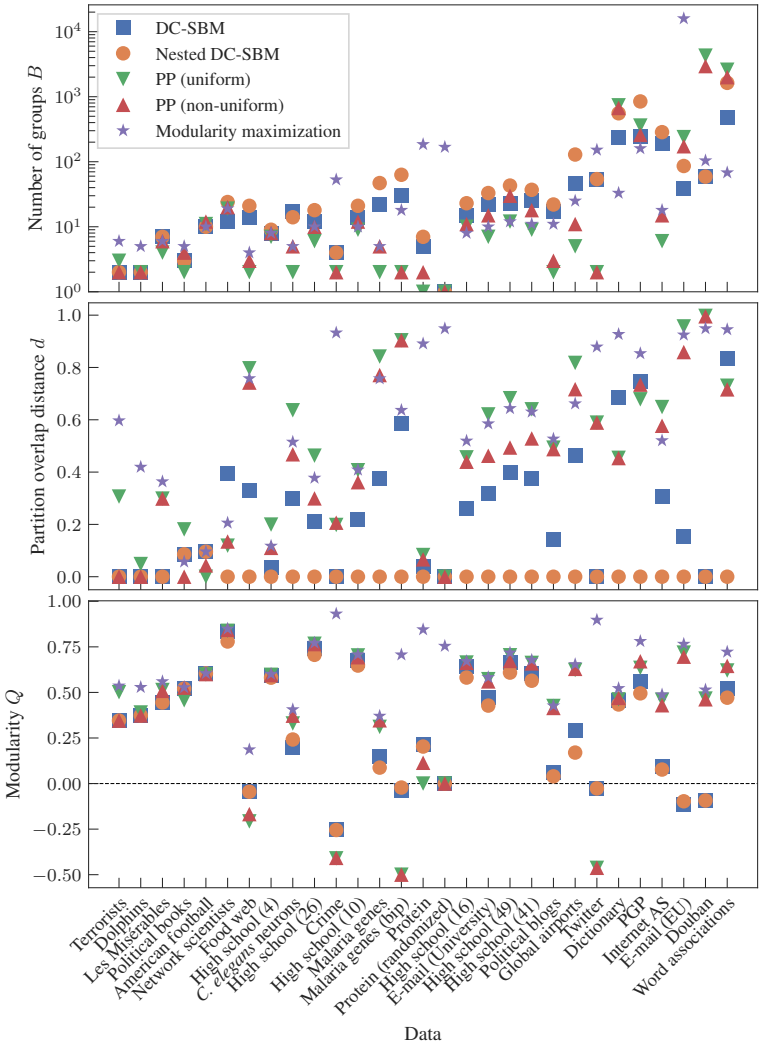


Figure 3-7: More details about the inference results on the set of 29 empirical networks, including the number of communities found with each method (top panel), the normalised maximum overlap distance [125] between the best fitting model and other model variants (middle panel), and the modularity value of the partitions (bottom panel).

at the autonomous system level (Internet AS) [128], the modularity value obtained with the best fitting model is significantly smaller than that with the PP models, indicating that the dominant patterns are not assortativity in these networks. Thus, it will be imprudent to apply a community detection algorithm that blindly searches for assortativity to these networks, e.g. the modularity maximisation approach, which will lead to significantly biased results.

We also include results obtained with the modularity maximisation approach in Fig 3-7. Having seen that the overfitting behaviour of modularity maximisation in synthetic networks in Section 3.3.1, one might expect that modularity maximisation will always find more communities than Bayesian inference approach does. Indeed, in some networks like the E-mail network of undisclosed European institution [129], protein-protein interaction [130], and bipartite person-crime association [131], modularity maximisation finds significantly larger number of communities than what is given by the Bayesian inference approach, signifying the risk of overfitting in these datasets. Although we are not to able to directly evaluate the extent of overfitting, because there is no underlying truth associated with real-world networks for evaluation, we can still get a rough idea about how much information in the inferred community structures is due to random fluctuations rather than to statistically significant structures. To do so, we take the protein-protein interaction network [130] as an example. We apply modularity maximisation and the non-uniform PP model to this network and show the inferred community structures in Fig 3-8. Maximising modularity approach returns a partition with over 200 communities and a high modularity value $Q = 0.84$. On the contrary, the non-uniform PP model is much more conservative and return only two communities with modularity value $Q = 0.11$. Then, we randomise this network according to the configuration model by randomly pairing edges attaching to each node. The resulted network contains no community structures, because the probability of an edge between any two nodes only depends on their degree and nothing else. However, in the randomised network, modularity maximisation still finds a partition with over a hundred communities and a large modularity value ($Q = 0.75$). In comparison, the non-uniform PP model correctly returns a single community in the random network, indicating there is no structures in the data.

It is possible that in a different realisation of the same configuration model, modularity maximisation returns some partitions with modularity values which are much smaller than the value obtained in the original network. If that is the case, one might argue that the problem of modularity might not be as problematic as it seems in the example show above. Indeed, there are actually works try to conduct statistical test for

modularity maximisation, concluding the presence of community structure only when the modularity value is significantly large compared to a population of modularity values obtained in random networks [132]. However, the point is that Bayesian inference should correctly puts all of the nodes into a single group in every random networks, given that networks of consideration are sufficiently large. The robust performance of the Bayesian inference approach is backed by the powerful Shannon's source coding theorem [133], which states that it is impossible to compress the outcomes $\boldsymbol{x}$ of a probability source $P(\boldsymbol{x})$ more than using the code associated with the source probability distribution. Recalling the correspondence between data compression and model inference we have introduced in Section 2.4.2, the theorem can be translated as that there is no other models can achieve higher posterior probability than the trivial SBM with one single block in random networks, since the trivial SBM is exactly the source probability for generating the data. The exaggerated result of modularity maximisation in this experiment brings up the caveat that, to a non-negligible extent, community structures given by modularity maximisation are not statistically significant and can be simply explained by the degree sequence of nodes.

On the other hand, despite the tendency of overfitting, modularity maximisation sometimes finds rather conservative results in terms of the inferred number of communities compared to the Bayesian inference approach. General SBMs finds more communities because they can identify non-assortative structures, which are not the targets of modularity maximisation. However, if we focus on the comparison between modularity maximisation and PP models which are designed for extracting assortative structures, for networks like the Douban social network [126] and word associations [134], PP models find roughly 100 times larger number of communities than modularity maximisation does. In other words, modularity maximisation seems to massively "underfit" in these two datasets even though it generally tends to overfit. In fact, despite being vulnerable to overfitting, modularity maximisation ironically has the tendency of underfitting data at the same time [51]. In the next chapter, we will look into the underfitting problem in community detection. We will show that our PP models are better alternatives for detecting assortative structures because they are robust against not only the overfitting but also the underfitting problems.

3.4 Concluding remarks

In this chapter, we revisit the equivalence result between the modularity maximisation approach and the maximum likelihood inference with the uniform PP model. We clarify that this equivalence does not hold in general because it relies on subjective

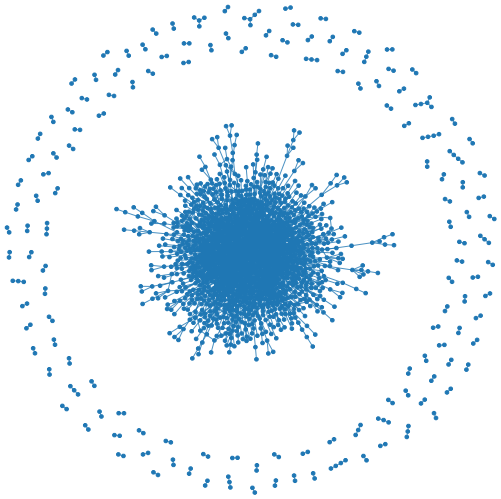
Original network

Modularity maximization
$B = 185$, $Q = 0.84$

PP (non-uniform)
$B = 2$, $Q = 0.11$

Randomized network

Modularity maximization
$B = 168$, $Q = 0.75$



PP (non-uniform)
$B = 1$, $Q = 0$

Figure 3-8: Inferred community structures in a network of protein-protein interactions [130], using modularity maximisation and Bayesian inference with the non-uniform PP model. In the top panel, we show the results obtained with the original network, while in the bottom panel the results with a randomised network generated from the configuration model.

choices of model parameters lacking of principle justifications. In fact, even when the equivalence holds, it implies that limitations of applying the maximum likelihood pass over to modularity maximisation, e.g. the tendency of overfitting and the restricted modelling capacity.

We develop the Bayesian formulation of the PP models which can be used to extract assortative structures in networks. Our Bayesian approach based on the PP models has the advantages of being robust against overfitting and describing more general assortative patterns. In making inference with the PP models, the change in the posterior likelihood is not more complicated to compute than that of the modularity measure. Therefore, in principle, the existing collection of modularity maximisation heuristics can be exploited by adopting the posterior probability of PP models as their objective functions, which should provide better regularisations for their results[5].

Our PP models are amenable to model selection technique which allows us to compare different model variants in a principle way. We compare PP models with general SBMs and discuss how this comparison allows us to determine whether assortativity is the dominant structural pattern in networks. Our results suggest that assortativity is often too simplistic to appropriately describe the community structure in empirical networks, at least for the dataset we have considered so far. We have found a few instructive examples which the assortative-constrained models manage to fit better than general models. In these examples, the difference in description length between PP modelss and general SBMs provides a quantitative measure which reflects how much modelling capacity is unnecessarily wasted by adopting general models.

Note that in the examples where PP models achieve the best fit to data, the difference in the inferred network partitions is not notable between PP models and general SBMs. In addition, in these examples, PP models' advantages over general models in terms of description length is not significant. These observations promotes the question that whether PP models are actually useful in practice. If general SBMs can always return results which are similar to that of PP models, should we just stick with the general models and forget about the PP models? The doubt on the necessity of PP models also comes from the observation that general models might achieve better fit than PP models when the inferred network partition is indeed assortative. This happens because our PP models cannot describe all possible assortative structures. The planted partition constraint, either uniform or non-uniform, rules out not just the possibility of other

[5]For example, we manage to adapt the celebrated Louvain algorithm [116] for modularity maximisation to find the MAP solution of the uniform PP model. For further details, please see the Appendix A.8.

kinds of community structures, but also the general form of assortativity with varying connections rate between different groups. For these reasons, PP models might seem practically redundant. However, as we are going to see in the next chapter, PP models in fact are the best fitting model in some networks while the difference in the inferred network partition are significant between PP models and general SBMs. Therefore, it is definitely worth including PPM models into our analysis toolbox, since they can provide extra insight into data, which would haven been overlooked by general models.

Chapter 4

Assessment of underfitting, overfitting, and model selection for modular network structure

Although the Bayesian inference approach keeps us safe from overfitting data, we still need to be careful with the risk of *underfitting* data. As the term suggests, we say a method underfits if its result is overly simplistic compared to the actual pattern in data, just the opposite to overfitting. In particular, we are interested in the underfitting behaviour of community detection algorithms in terms of the order of inferred models, which is reflected by the number of inferred communities. Underfitting community detection methods often find themselves being able to extract structural information at a coarse-grained level, but struggling with resolving fine-resolution details. Specifically, it is common to see an underfitting method merges communities of small sizes into large ones, returning partitions with overly conservative number of communities.

Bayesian inference with DC-SBM is vulnerable to underfitting [63]. The root cause of this problem is in the use of an inappropriate uninformative prior for edge placement. The uninformative prior has a penalising effect on the model complexity, which is reflected by the number of communities. When the underlying assumption of the uninformative prior is not compatible with the pattern in data, the penalty caused by the uninformative prior could be excessively strong, which might suppress significant structural pattern. The tendency of underfitting data leads to a *resolution limit* of DC-SBM. The limit is related to the number of detectable communities and it grows with the size of networks. When the actual number of communities is above the resolution

limit, DC-SBM will fail to recover complete structural pattern in data.

Addressing the underfitting problem of DC-SBM is the main motivation behind the development of the Nested DC-SBM [63]. The Nested DC-SBM does not have the underfitting problem because it replaces the uninformative prior for edges with a hierarchical prior. Unlike the uninformative prior which assumes edges placement is completely random among communities, the Nested DC-SBM explicitly models the connections between communities as a multigraph above the original network. The multigraph is then assumed to be generated from another SBM. This procedure can carry on recursively until there is a single node at the highest level. By doing so, the models inferred at upper levels form a hierarchy of priors for the inference at the bottom level. The hierarchical prior is a more realistic prior for the edge placements than the uninformative one, because the hierarchical prior can adapt its structure according to data. As a result, the hierarchal construction allows Nested DC-SBM to detect much more refined structures compared to DC-SBM. Interestingly, the PP models we developed in the last chapter also have the advantage of not underfitting data [83]. This is because PP models are restricted models with less modelling complexity. Hence, even when simple uninformative priors are in place, PP models have less severe induced penalty than DC-SBM, allowing detection of statistically significant structures of arbitrary sizes.

The underfitting behaviour of community detection methods is often studied in synthetic networks with well-defined structures, which are mainly assortative for the ease of analysis. However, these synthetic examples are not satisfactory representatives of real-world networks. Community structures in real-world networks usually constitute a mixture of randomness and pattern, with different kinds of community structures being present at the same time. Amir et al. [106] evaluated the underfitting behaviour of DC-SBM together with other 15 different community detection methods in an empirical network corpus. However, the evaluation was done in an indirect way, where algorithms' underfitting behaviour was measured by their performance in a downstream application. Peixoto [63] directly compared the difference in the inferred number of communities between DC-SBM and Nested DC-SBM, but the comparison was only done in a small set of empirical networks and PP models were not included in the comparison. Therefore, it is still not clear yet how widespread the underfitting problem of the DC-SBM occurs in practice, and to what extent DC-SBM behaves differently from non-underfitting methods like Nested DC-SBM and PP models.

In this chapter, we aim to reveal the underfitting behaviour of DC-SBM in practice by comparing its performance to that of Nested DC-SBM and PP models on a large

empirical network corpus. We construct a network corpus with 263 empirical networks which are diverse in both their sizes and scientific domains. Although underlying truth of community structures generally do not exist in empirical networks [135], since both Nested DC-SBM and PP models are known to be free from the underfitting problem of DC-SBM, we propose to measure the underfitting behaviour of DC-SBM by conducting *model selection*. We know underfitting occurs to DC-SBM when non-underfitting methods manage to achieve better quality of fit, or equivalently shorter description length of data, with more detailed partitions. Our results suggest that DC-SBM systematically underfits and using Nested DC-SBM is able to find significantly more detailed structures, especially in networks with large sizes. Although PP models do not have the underfitting problem, their results are usually more conservative compared to that of DC-SBM. This is due to the fact that PP models are restricted models focusing on assortative structures, but assortativity is usually not the dominant pattern in empirical networks. When assortativity is indeed the dominant pattern, we find that PP models are able to extract detailed structures that are not visible to general SBMs. Finally, using the same network corpus, we show that the modularity maximisation approach also suffers from the problem of underfitting data, even though its underfitting behaviour is often covered by its tendency of overfitting.

We begin this chapter with an introduction to the resolution limit problem of DC-SBM in Section 4.1. Then in Section 4.2, we explain why the Nested DC-SBM and the PP models are free from the resolution limit. In Section 4.3, we compare the results of DC-SBM to that of Nested DC-SBM and PP models. Section 4.4 focuses on comparing PP models to general models. Finally, Section 4.5 concerns the underfitting behaviour of the modularity maximisation approach.

4.1 The resolution limit underfitting problem

We start with introducing the underfitting problem of the DC-SBM, which is often demonstrated by considering synthetic networks with clear community structures. For example, in Fig 4-1, we show a network consisting of 64 isolated cliques. A *clique* is a fully connected subgraph. Intuitively, we should assign each clique into it its own community, concluding an extreme assortative partition. However, the partition with the maximum posterior probability of DC-SBM, or equivalently the MDL solution, is the one with every two cliques being merged together. Although this result partially recover the clique structure, it is an inadequate fit to the data, because the inferred model with only 32 communities has a vanishingly small probability of generating the observed example. This counter-intuitive behaviour is referred to as underfitting,

Figure 4-1: A network consists of 64 cliques of size 10. Shading ovals as well as the colouring of nodes imply the partition given by applying the MDL principle with the DC-SBM. This is undesired behaviour of DC-SBM is known as the resolution limit problem. We should derive the resolution limit of DC-SBM later in equation (4.10). Figure reproduced from [60].

because the inferred model is overly simplistic, even though DC-SBM has the ability of describing the assortative structure in the data. As explained in [63], DC-SBM suffers from the underfitting problem because the choice of the uninformative prior for edge placements as defined in equation (2.42) is inappropriate. According to this uninformative prior, edges are randomly assigned among communities with an equal probability. As a result, it is implicitly assumed that the expected number of edges among different pairs of communities are identical. When this assumption violates the pattern in data, using this uninformative prior might cause an over-penalising effect. To see how the uninformative prior leads us to the undesired solution given in Fig. 4-1, we firstly write down the expression of the posterior probability $P(\boldsymbol{b}|\boldsymbol{A})$ for the correct partition $\boldsymbol{b}^*$, which identifies all 64 cliques as 64 disconnected communities. For $\boldsymbol{b}^*$, we have $e_r = 2E/B^*$ and $n_r = N/B^*$, with $B^* = 64$. Then, the posterior probability $P(\boldsymbol{b}^*|\boldsymbol{A})$ is proportional to the joint probability $P(\boldsymbol{A}, \boldsymbol{b}^*)$, which involves the likelihood of DC-SBM

$$P(\boldsymbol{A}|\boldsymbol{k}, \boldsymbol{e}, \boldsymbol{b}^*) = \frac{(2E/B^*)!!^{B^*}}{(2E/B^*)!^{B^*}} \times \frac{\prod_u k_u!}{\prod_{u<v} A_{uv}! \prod_u A_{uu}!!}, \tag{4.1}$$

and the prior probabilities for parameters $\boldsymbol{b}, \boldsymbol{e}, \boldsymbol{k}$:

$$P(\boldsymbol{b}^*) = \frac{(N/B^*)!^{B^*}}{N!} \times \binom{N-1}{B^*-1}^{-1} \frac{1}{N}, \tag{4.2}$$

$$P(\boldsymbol{e}|\boldsymbol{b}^*) = \left(\!\!\binom{B^*(B^*+1)/2}{E}\!\!\right)^{-1} \tag{4.3}$$

$$P(\boldsymbol{k}|\boldsymbol{e},\boldsymbol{b}^*) = \left(\!\!\binom{N/B^*}{2E/B^*}\!\!\right)^{-B^*}. \tag{4.4}$$

The logarithm of the joint probability then has the following expression[1]

$$\ln P(\boldsymbol{A},\boldsymbol{k},\boldsymbol{e},\boldsymbol{b}) \approx (E-N)\ln B^* - (E+{B^*}^2/2)g\left(\frac{E}{E+{B^*}^2/2}\right), \tag{4.5}$$

where $g(x) = -x\ln x - (1-x)\ln(1-x)$ and the Stirling's approximation is used. The first term in the log-probability above corresponds to the model likelihood while the second term corresponds to the priors of the model parameters. The ${B^*}^2$ term comes from the uninformative prior of edges placement $P(\boldsymbol{e})$ and it is exactly the number of unknown parameters in the connection matrix $\boldsymbol{e} = \{e_{rs}\}$. The expression in (4.5) suggests that the uninformative prior has a penalising effect that increases quadratically with the number of communities B.

Having seen the contribution of the uninformative prior in the joint probability in equation (4.5), we can try to understand a bit more about why the correct partition $\boldsymbol{b}^*$ is missed by the MDL approach. To this end, we try to find some partitions $\boldsymbol{b}'$ which are different from the correct partition $\boldsymbol{b}^*$, such that the posterior probability with $\boldsymbol{b}'$ is higher than that of $\boldsymbol{b}^*$. Rather than exploring all possible network partitions, we will restrict ourselves to a special set of partitions, in which correct communities (cliques) are merged into larger communities of equal sizes. Although this special set only takes up a small fraction of the entire solution space of network partitions, restricting to this special set makes our analysis easier. Specifically, for any partition $\boldsymbol{b}'$ in this special set, its likelihood function and prior distributions have exactly the same expressions as those of $\boldsymbol{b}^*$ in equations (4.1) to (4.4), except for the number of communities B needs to be rescaled correspondingly to B'. Consider $B' \in [1, B]$, we then have the logarithm of the joint probability with $\boldsymbol{b}'$ as a function with the only variable being the number of communities B',

$$\ln P(\boldsymbol{A},\boldsymbol{k},\boldsymbol{e},\boldsymbol{b}(B')) \approx (E-N)\ln B' - (E+{B'}^2/2)g\left(\frac{E}{E+{B'}^2/2}\right). \tag{4.6}$$

[1] We have provided the erivation of the expression is provided in Appendix B.1, further details please refer to [63, 74].

Then, finding the best partition is equivalent to finding the best number of communities $B_{\max}$ which leads to the highest posterior probability, or equivalently

$$B_{\max} = \underset{B'}{\operatorname{argmax}} \frac{P(\boldsymbol{A}, \boldsymbol{k}, \boldsymbol{e}, \boldsymbol{b}(B'))}{P(\boldsymbol{A})}. \tag{4.7}$$

To proceed, we might treat B' as a continuous variable and differentiate the equation (4.6) with respective to B'. The derivative of $\ln P(\boldsymbol{A}, \boldsymbol{k}, \boldsymbol{e}, \boldsymbol{b}(B'))$ of B' reads as

$$\frac{dP(\boldsymbol{A}, \boldsymbol{k}, \boldsymbol{e}, \boldsymbol{b}(B'))}{dB'} = \frac{E-N}{B'} + B' \ln \frac{B'^2/2}{E + B'^2/2}. \tag{4.8}$$

In the case where $E - N \leq 0$ or equivalently $\langle k \rangle = 2E/N \leq 2$, the derivative is negative regardless the value of B'. This means, for an observed network with an average degree smaller than 2, the amount of information in data will not be sufficient to support any modular structures[2], and the most plausible partition is the one that puts all of the nodes in a single group,

$$B_{\max} = \min_{\mathbb{Z}} B' = 1. \tag{4.9}$$

When average degree $\langle k \rangle$ is larger than 2, setting the derivate in equation (4.8) equal to zero gives

$$B_{\max} = x(\langle k \rangle)\sqrt{N}, \tag{4.10}$$

where $x(\langle k \rangle)$ is the solution of the following equation

$$\langle k \rangle - 2 = 2x^2 \ln \frac{\langle k \rangle + x^2}{x^2}. \tag{4.11}$$

Although there is no analytic expression of $x(\langle k \rangle)$, it is clear that the solution $x(\langle k \rangle)$ is a function of the average degree $\langle k \rangle$, which does not grow with the size of network N. Therefore, the expression of $B_{\max}$ in equation (4.10) implies the optimal number of communities has an intrinsic scale at $\mathcal{O}(\sqrt{N})$. This scaling is often referred as the *resolution limit* and many community detection methods are found to suffer from similar limits of the number of detectable communities. Empirically, the solution of equation (4.11) is at the scale of $\mathcal{O}(\sqrt{k})$, leading the entire scale of the resolution limit to be $\mathcal{O}(\sqrt{N\langle k \rangle}) = \mathcal{O}(\sqrt{E})$[3]. When the correct number of communities B is larger than $B_{\max}$ in equation (4.10), we will not able to recover the correct partition by finding the

[2]We emphasis that our discussion here is restricted to the special "isolated cliques" networks only.

[3]We used the Newton-Raphson method [136] to obtain numerical estimates of the resolution limit $B_{\max}$ for datasets in the empirical network corpus to be analysed later in this chapter. Our results support that the resolution limit of DC-SBM is roughly as the scale $\mathcal{O}(\sqrt{E})$. We present the results of approximated resolution limit in Appendix B.2.

maximum a posterior solution. This is exactly what happens in the network of cliques in Fig 4-1, where the numerical estimate of the resolution limit $B_{\max} = 36.06$, which is smaller than the correct number of communities 64. With the resolution limit in mind, we should be careful with the interpretation of the community structures inferred by DC-SBM, since significant structures might remain undetected below the resolution limit.

4.2 Models that do not suffer from the resolution limit

4.2.1 Nested DC-SBM

The Nested DC-SBM we introduced in Section 2.4 addresses the resolution limit problem of DC-SBM by replacing the uninformative prior of edge placement with a hierarchical prior. The hierarchical prior is more realistic compared to the non-uninformative prior since it does not make any particular assumption about the connections between communities. Instead, the hierarchical prior explicitly includes the connection pattern at community-level as a part of the model. As a result, the structures inferred at higher levels will serve as more appropriate priors for the inference of community structure at bottom levels, leading to a better description of data.

We can try to deduce the limit of the inferred number of communities for Nested DC-SBM, if any exists, just as we did for DC-SBM. In the network of cliques, following the steps in [74], let the number of communities given by Nested DC-SBM be $B' \in [1, B]$ and assume a uniform hierarchical division where at each level the number of groups decreases by a factor σ, i.e. $B_l = B'/\sigma^l$. Then, the number of nodes in each level l is $N_l = B_{l-1} = B'/\sigma^{l-1}$ and the number of nodes in each group is $n_r^l = N_l/B_l = \sigma$. By construction, the top level of the hierarchy should have one community with a single node, which means $B'^L = 1$. Therefore, the height of the hierarchy should satisfy $L = \ln_\sigma B'$. Then the hierarchical prior in equation (2.46) for the 64 cliques example in Fig 4-1 is

$$\begin{aligned} P(\boldsymbol{e}) &= \prod_l^L P(e_l|\boldsymbol{e}_{l-1}, \boldsymbol{b}_{l-1}) P(\boldsymbol{b}_{l-1}) \\ &= \prod_l^{\ln_\sigma B'} \prod_r^{B'/\sigma^l} \left(\binom{\sigma(\sigma+1)/2}{2E\sigma^l/B'} \right)^{-1} \times \frac{\sigma!^{B'/\sigma^l}}{(B'/\sigma^{l-1})!} \binom{B'/\sigma^{l-1}-1}{B'/\sigma^l - 1}^{-1} \frac{1}{B'/\sigma^{l-1}-1}. \end{aligned} \tag{4.12}$$

Since the underfitting problem is more likely to occur in network with large number of

communities, we are mainly interested in the regime where the number of communities in the network is at the scale of the network's size, i.e. $B = \mathcal{O}(N)$. Then, assume $B \gg \sigma$ and σ is sufficiently large such that we can make use of the Stirling's approximation, making use of results in [63, 74], $\ln P(\boldsymbol{e})$ has the following expression

$$\ln P(\boldsymbol{e}) \approx -\frac{\sigma(\sigma+1)}{2(\sigma-1)} B' \ln E, \tag{4.13}$$

and hence the description length

$$\ln P(\boldsymbol{A}, \boldsymbol{k}, \boldsymbol{e}, \boldsymbol{q}) \approx (E-N) \ln B' - \frac{\sigma(\sigma+1)}{2(\sigma-1)} B' \ln E. \tag{4.14}$$

Differentiating the last equation with respective to B' and setting the derivative equal to zero gives the optimum number of communities

$$B_{\max} = \frac{(\sigma-1)(\langle k \rangle - 2)}{\sigma(\sigma+1)} \times \frac{N}{\ln N}. \tag{4.15}$$

This expression of $B_{\max}$ implies that the scale of maximum number of detectable communities for Nested DC-SBM is $\mathcal{O}(N/\ln N)$. Notice that the scale of this limit is not only significantly larger than the limit of DC-SBM, but also compatible with the maximum number of communities N, since there is at most $B = N$ groups in a network with N nodes. Therefore, the Nested DC-SBM virtually resolves the underfitting problem of DC-SBM and is able to find arbitrarily large number of communities as long as they exist in data.

4.2.2 The planted partition models

Conducting inference with PP models as we explained in Chapter 3 also does not have the resolution limit problem of DC-SBM [83]. The Nested DC-SBM solves the underfitting problem by adopting a hierarchical prior, whereas PP models are free from the problem even when a uninformative prior is used. This is because PP models have less modelling parameters than general SBMs. Recall that DC-SBM allows $B(B+1)/2$ distinct values in its connection matrix for the number of connections between all possible $B(B+1)/2$ pair of groups. In comparison, the uniform PP model only takes two parameters regarding the edge placements among communities, i.e. e_{in} and e_{out}. As a result, the uninformative prior $P(\boldsymbol{e})$ for the uniform PP model has less severe penalty than that for DC-SBM. The prior for the uniform assortativity is the one in

equation (3.54) and it has the following expression in the network of cliques

$$P(\boldsymbol{e}) = P(\boldsymbol{e}|e_{\text{in}} = E, e_{\text{out}} = 0, \boldsymbol{b})P(e_{\text{in}} = E, e_{\text{out}} = 0|\boldsymbol{b}) = \frac{E!}{{B'}^E \prod_r^B (E/B')!} \frac{1}{(E+1)^{\delta_{B',1}}}. \tag{4.16}$$

Notice that this uninformative prior has a constant contribution to the description length:

$$\ln P(\boldsymbol{e}) \approx -E \ln E. \tag{4.17}$$

This reflects the fact that the uniform PP model only requires a constant number of parameters to generate the edge count matrix $\boldsymbol{e} = \{e_{rs}\}$. Hence, the description length of the model with the use of the uniform planted partition prior has the following expression

$$\ln P(\boldsymbol{A}, \boldsymbol{e}, \boldsymbol{k}, \boldsymbol{b}(B')) = (E - N) \ln B', \tag{4.18}$$

which is simply an increasing function of B' under the assumption B' in $[1, B]$. Therefore, the uniform PP model will not incorrectly merge cliques as DC-SBM does.

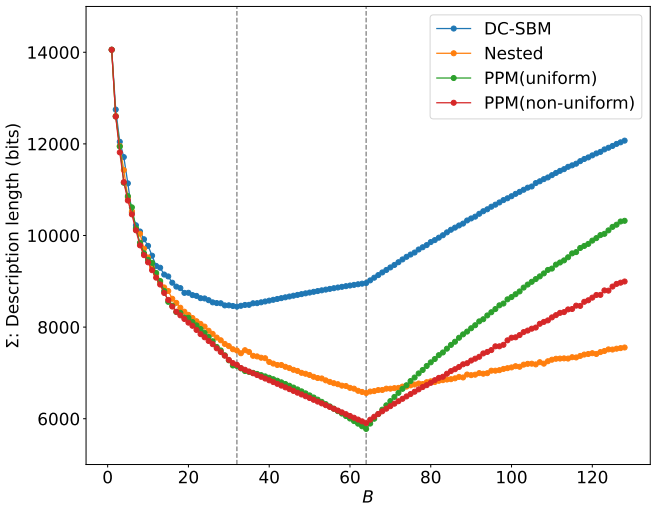


Figure 4-2: Description length as a function of the number of communities obtained in the 64 cliques example in Fig 4-1. The DC-SBM suffers from the resolution limit underfitting problem, merging every two cliques into a large group. In comparison, both the Nested DC-SBM and the PP models are able to correctly identify 64 cliques.

Similarly, if we consider the non-uniform PP model, the prior for the edge count matrix is then

$$\begin{aligned} P(\boldsymbol{e}) &= P(\boldsymbol{e}|\{e_{rr}\}, e_{\text{out}} = 0, \boldsymbol{b})P(\{e_{rr}\}, e_{\text{out}} = 0|\boldsymbol{b}) \\ &= \left(\!\!\binom{B}{E}\!\!\right)^{-1} \frac{1}{(E+1)^{\delta_{B',1}}}. \end{aligned} \tag{4.19}$$

Compared to DC-SBM, the penalty caused by the uninformative prior reduces from a quadratic function dependent on B' to a term which is only linearly dependant on B' with $\ln P(\boldsymbol{e}) \approx -(E + B')\ln E$. The description length of the non-uniform PP model of the network of cliques is

$$\ln P(\boldsymbol{A}, \boldsymbol{e}, \boldsymbol{k}, \boldsymbol{b}(B')) = (E - N)\ln B' - (E + B')\ln E, \tag{4.20}$$

where the linear penalty term corresponds to the linearly growing modelling complexity that comes from the $\{e_{rr}\}$ parameters. The maximum of the equation above is achieved at

$$B_{\max} = \frac{\langle k\rangle - 2}{2\log\langle k\rangle}\frac{N}{\ln N}, \tag{4.21}$$

which has the scale at $\mathcal{O}(N/\ln N)$. This is again significantly larger than the resolution limit of DC-SBM and is similar to that of Nested DC-SBM as we derived in equation (4.15). In Fig 4-2, we show the result of fitting the DC-SBM, Nested DC-SBM and uniform PP model to the 64 cliques network. Both Nested DC-SBM and PP models correctly distinguish 64 cliques, while the DC-SBM places the optimum at a partition with only 32 communities.

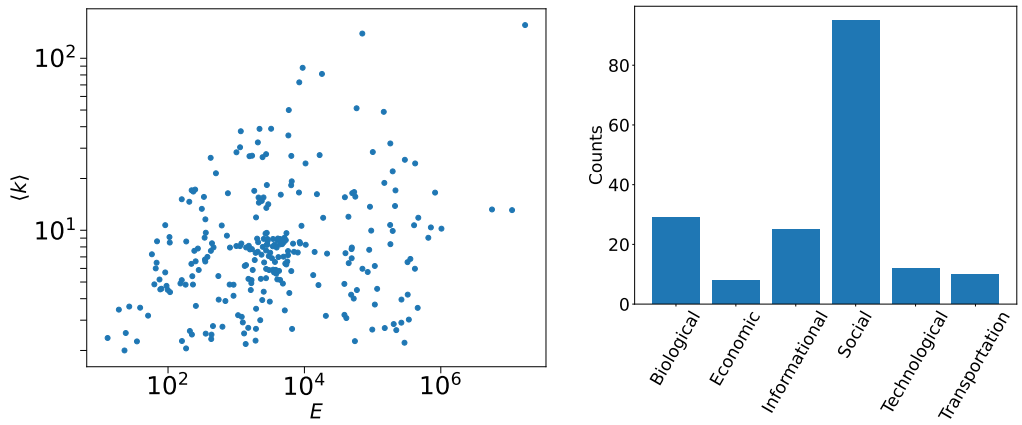
Having seen that Nested DC-SBM and PP models manage to get around the resolution limit of DC-SBM in the network of cliques, it is natural to ask how DC-SBM differs from non-underfitting models in practice. After all, real-world networks are rarely to have structures that are as clear as the cliques example. Besides, it is not clear whether DC-SBM still underfits when networks consist of non-assortative structures, or when the true number of communities is below the resolution limit, and if so, to what extent DC-SBM performs differently from non-underfitting models. It is not impossible that DC-SBM actually provides similar results compared to non-underfitting models in empirical networks, and that the underfitting problem we observed in synthetic networks is just an extreme case with little practical relevance. To answer the questions above, we will compare the results of DC-SBM to that of Nested DC-SBM and PP models using an empirical network corpus.

4.3 Underfitting in empirical networks

We compare the performance of the DC-SBM to that of Nested DC-SBM as well as PP models (uniform and non-uniform) by fitting them to a network corpus with 263 empirical networks. We constructed the corpus by gathering networks from the Netzschleuder network dataset repository [7]. We collected every available network

in the repository with a cut-off of the largest number nodes at 10^6. For datasets that consist of multiple networks, we picked only one of them to avoid closely related network samples. One exception is the Adolescent health dataset [137]. We included all of the 84 networks in this dataset into our network corpus, because there are no correlations among these networks, which correspond to social networks of students in different schools. As shown in Fig. 4-3, our network corpus spans a wide range of network sizes and density (average degree), but with a majority of networks coming from the social science domain. This skewed distribution of network domains prevents us from learning the correlation between algorithmic behaviours and the source of network data, but it is not a problem for our purpose of comparing the results of DC-SBM to non-underfitting models. For each network, we consider its simple version (no self-loops and no multiple edges) and focus on their largest connected component. When we fit models to our network corpus, we find the MAP solution using the MCMC algorithm as explained in Section 2.5. To fully explore the solution space, we ran the inference algorithm for multiple times with different initial states and recorded the inferred community structure that achieves the shortest description length. Our experiment was done with the Balena high performance computing system at the University of Bath.

Unlike in synthetic networks where we know the correct community structures, underlying truth of structures in empirical networks are not only unavailable to us, but also generally do not exist [135]. For this reason, we are not able to assess the underfitting behaviour of models at an absolute level. Nevertheless, we can obtain evidence of model's tendency of underfitting at a relative level by conducting model selection. In particular, we are interested in how the difference in description length between DC-SBM and non-underfitting models relates to their difference in inferred number of communities. The idea is that, following the connection between statistical inference and data compression as we explained in Section 2.4.2, when DC-SBM finds a partition with longer description length than that of a non-underfitting model, the difference in description length is approximately the amount of structural information not being captured under the parameterisation of DC-SBM. We say DC-SBM underfits relative to non-underfitting models if its inferred community structure has longer description length and smaller inferred number of communities than that of non-underfitting models. We emphasis that this definition consists of two parts - one regarding the difference in description length and the other regarding the number of inferred communities. What we care is the disadvantage in description caused by the limitation of the model which restricted its ability to detect statistically significant structure, although we are only exploring the correlation here. According to this criterion, looking at the result of fitting the network of cliques in Fig 4-2, we say DC-SBM underfits compared to Nested



(a) (b)

Figure 4-3: (a) The average degree $\langle k \rangle$ against the number of edges E in our network corpus. (b) The distribution of network domains of datasets in our network corpus. Note that the 84 networks from the the Adolescent health dataset are counted as one contribution to the social network domain in the histogram.

DC-SBM and PP models, because DC-SBM misses the chance of compressing data (equivalently better quality of fit) by leveraging the existence of significant structure. When comparing the results of PP models to that of Nested DC-SBM, we do not say the latter underfits compared to the former, since they find exactly the same network partition, despite the difference in description length between them.

To reveal how much DC-SBM underfits in our network corpus, we compare its inferred number of communities in our network corpus to that of the Nested DC-SBM. Because DC-SBM can be viewed as a special Nested DC-SBM with a single layer, Nested DC-SBM should always perform at least as good as, if not better than, the single-layered DC-SBM. What is not clear yet is the correlation between Nested DC-SBM's advantage over DC-SBM in the quality of fit, or equivalently the description length of data, and their difference in the inferred number of communities. In Fig. 4-4, we plot the difference in inferred number of communities between the DC-SBM the Nested DC-SBM, $B_{\text{Nested}} - B_{\text{DCSBM}}$, for each dataset in our network corpus. The x-axis is the indices of networks which lists networks in the increasing order of their sizes (number of edges). For networks with small sizes, the difference in inferred number of communities between the two models is minor, so is the difference in the description length indicated by the colouring of points. As the network size increases, there is a trend that Nested DC-SBM resolves significantly larger number of communities, with differences being in orders of magnitude. At the same time, the advantage of the nested model in description length also grows as the sizes of networks become large. From Fig. 4-5 we can see clearly

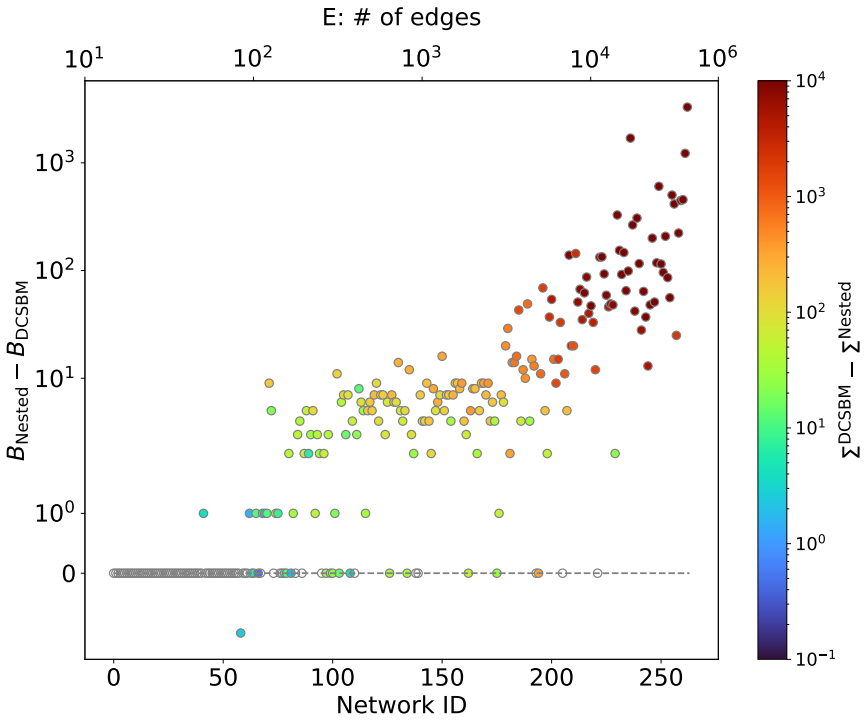


Figure 4-4: Inferred number of communities given by the Nested DC-SBM subtracted from that given by the single layer DC-SBM. Networks indices are ordered in the increasing order of network sizes (number of edges) and the colouring of points indicates the description length difference: $\Sigma_{\text{DCSBM}} - \Sigma_{\text{Nested}}$. The middle point of the colour map is $\ln 100$ such that points with reddish colouring should be interpreted as the Nested DC-SBM is significantly preferred over the single-layered DC-SBM according to MDL.

the positive correlation between the difference in inferred number of communities and the difference in description length. The seemly power-law shape in Fig. 4-5 might be explained by considering the analytical expression of $\Sigma_{\text{DCSBM}} - \Sigma_{\text{Nested}}$ in the clique example in Section 4.1. By subtracting equation (4.10) from (4.15), we get

$$\Sigma_{\text{DCSBM}} - \Sigma_{\text{Nested}} = -(E + B^{*^2})g\left(\frac{E}{E + B^{*^2}/2}\right) + \frac{\sigma(\sigma+1)}{2(\sigma-1)}B' \ln E = \mathcal{O}(E). \quad (4.22)$$

Similarly, $B_{\text{Nested}} - B_{\text{DCSBM}}$ roughly have the following expression

$$B_{\text{Nested}} - B_{\text{DCSBM}} = \frac{(\sigma-1)(\langle k\rangle - 2)}{\sigma(\sigma+1)}\frac{N}{\ln N} - \sqrt{N}x(\langle k\rangle) = \mathcal{O}(N). \quad (4.23)$$

The last two equation justifies the power-law shape we observed in Fig. 4-5. These observations imply that DC-SBM systematically underfits in our network corpus, and using Nested DC-SBM is able to resolve much more detailed structures.

One comment should be made to the Fig. 4-4 is that there is no correspondence between the inferred number of communities B and the description length of the corresponding network partition. Therefore, although we know the Nested DC-SBM should always outperform the single-layer DC-SBM in terms of description length, this knowledge does not imply that the former will always find more communities than the latter. In our network corpus, there is only one such example in which Nested DC-SBM finds less communities than the single-layer DC-SBM, corresponding to the only point landing below the horizontal line $y = 0$ in Fig. 4-4. This example is the E. coli transcription network [138] and we visualise the inferred communities in this network in Fig. 4-6. The inferred number of communities for the Nested DC-SBM is 4 while that for the single-layered DC-SBM is 5. Both of these two partitions are local optimum solutions for both the nested and single-layer variants, but the hierarchical construction brings more improvement in description length for the partition with less communities. This indeed can happen in practice and it is just by chance that we only have one such example in our network corpus.

Although PP models do not have the resolution limit of DC-SBM in the clique example in Fig. 4-1, recall that PP models are restricted models and can only extract assortative structures. Therefore, we should expect PP models to show advantage in terms of resolving detailed structures only when the dominant patterns in data are assortative. Otherwise, DC-SBM might find more communities but simply because it can identify non-assortative structures which will be hidden from PP models. For instance, in the top panel of Fig 4-7, we show the inferred community structures in the network

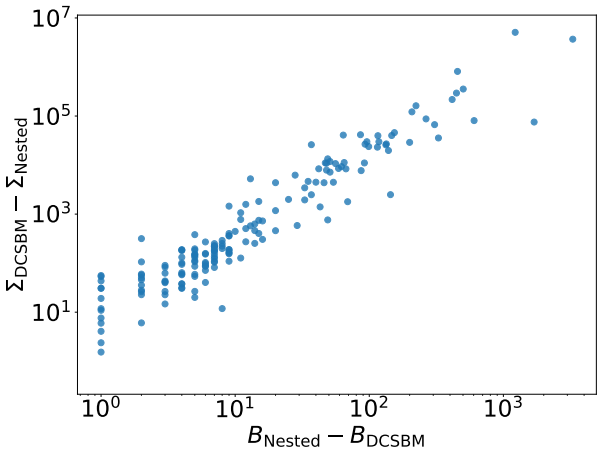


Figure 4-5: Difference in inferred number of communities as a function of the difference in the description length between DC-SBM and Nested DC-SBM in our network corpus with 263 empirical networks.

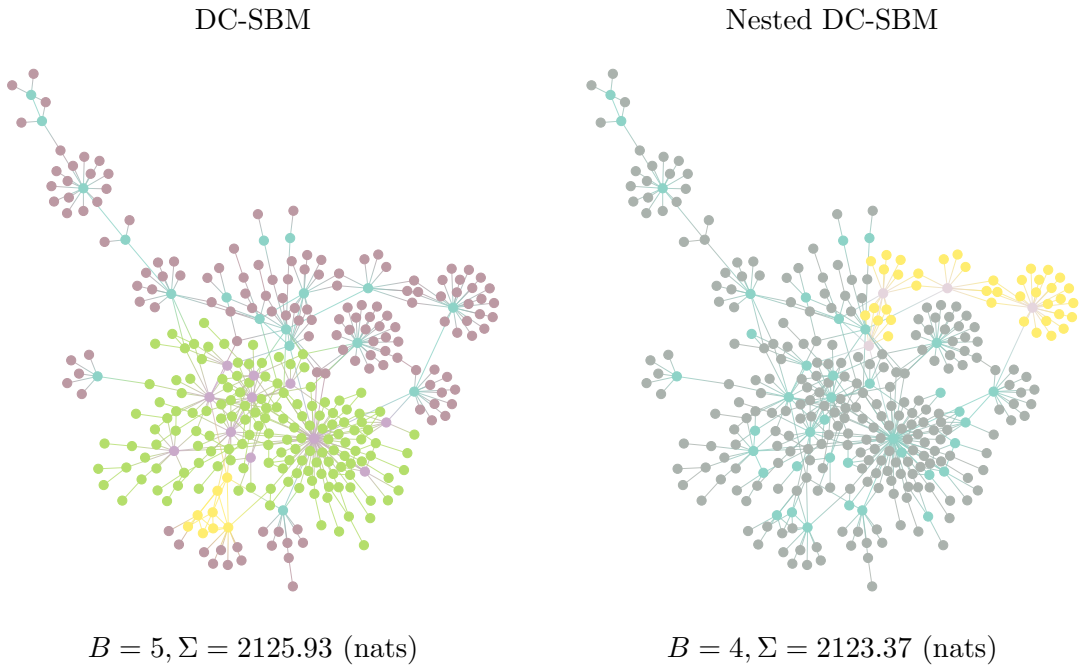


Figure 4-6: Inferred community structures for the network of E. coli transcription. Legends give the description length and inferred number of communities of the corresponding network partitions. Constructing a hierarchical partition with the bottom level partition being the one inferred by the single-layered DC-SBM has description length $\Sigma = 2125.86$, which is larger than that of the result given by Nested DC-SBM.

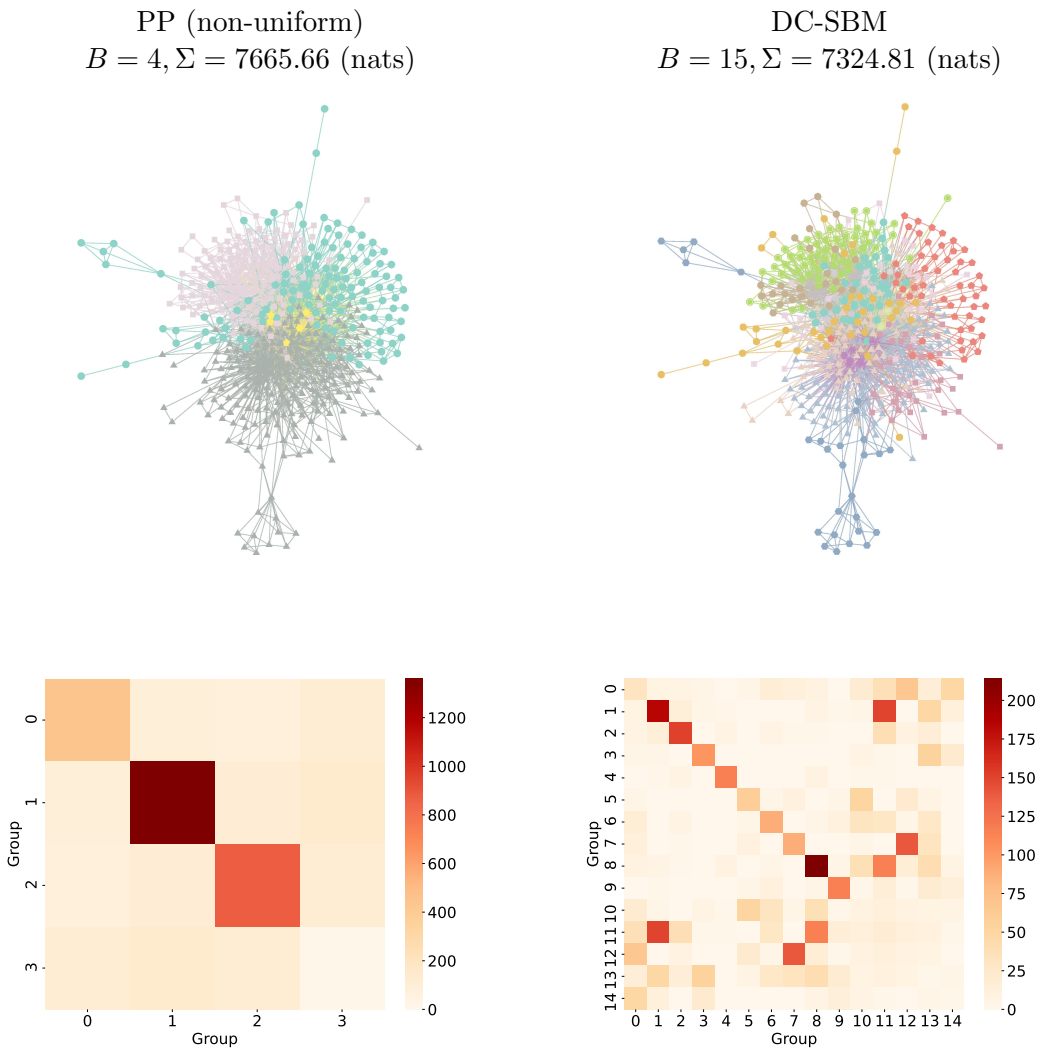


Figure 4-7: Top panel: inferred community structures in the network of the nematode C. elegans [139]. Bottom panel: the edge count matrices of partitions inferred given by the non-uniform PP model (left) and the DC-SBM (right).

of the nematode (roundworm) C. elegans [139]. In this network, non-uniform PP model returns a partition with 4 communities while the DC-SBM finds 15. In the bottom panel of Fig 4-7, we visualise the edge count matrices of inferred partitions given by the two models. It is clear that the non-uniform PP model returns a typical assortative structure with little variations in connections between distinct communities, while the DC-SBM concludes a hybrid structure, consisting both of assortativity and heterogeneous disassortativity.

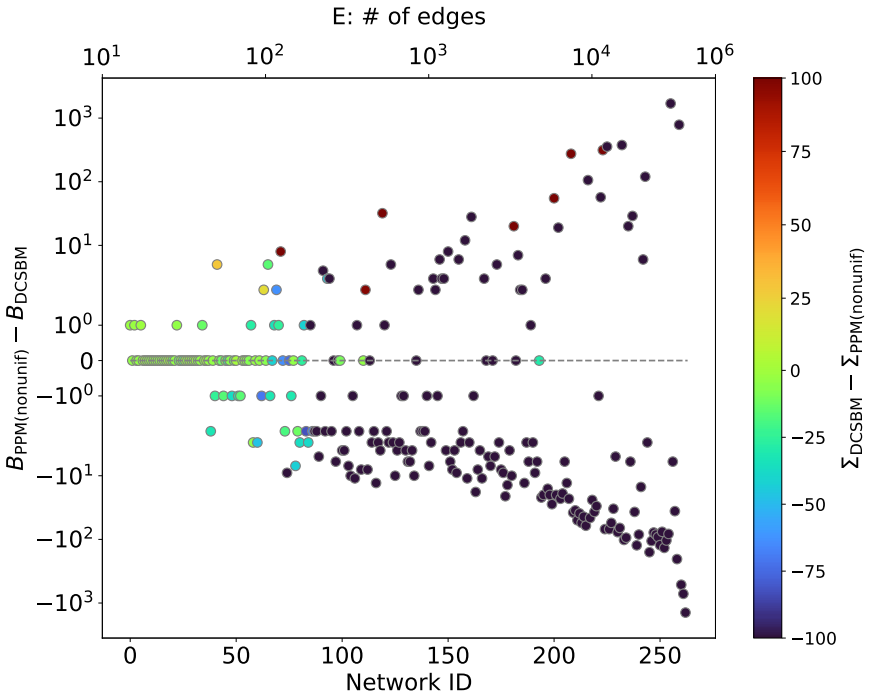


Figure 4-8: Inferred number of communities given by the non-uniform PP model subtracted from that given by the single layer DC-SBM. Networks indices are ordered in the increasing order of network sizes (number of edges) and the colouring of points indicates the description length difference: $\Sigma_{\mathrm{DCSBM}} - \Sigma_{\mathrm{PPM}}$. The results for the comparison between the uniform PP model and DC-SBM is similar and it is given in the Appendix B-2.

Therefore, when we compare the inferred number communities of DC-SBM to that of PP models, it is not surprised to see that PP models generally return more conservative results than DC-SBM, as shown in Fig 4-8. The difference seems to be a function of the number of edges E, reflecting the fact that the number of detectable communities is a function of $\sqrt{E}$ for DCSBM or E for PP models. The colouring of points reflects the difference in description length between the two models, indicating that assortativity is often less plausible compared to the general pattern described by DC-SBM according to the MDL principle. That means PP models often "underfits" data relative to DC-SBM,

finding less communities than general SBMs, but in an expected way: PP models can only identify assortative structures and fail to detect other kinds of structures, which can be detected by general models. Nevertheless, as we are going to see in the next section, in networks where the dominant pattern is indeed assortative, we find that PP models are able to achieve better fit to data with higher-resolution network partitions compared to general models.

4.4 Are PP models redundant?

Both PP models and Nested DC-SBM can address the resolution limit problem of DC-SBM, but the former are restricted while the latter is more general with better modelling capacity. In the network of cliques in Fig 4-1, although PP models achieve shorter description length than Nested DC-SBM, they conclude the same partition which corresponds to the correct structure. In addition, recall that in Section 3.3 we found PP models only achieve better fit to data in a minority of examples, i.e. the political books and American college football in Fig. 3-5. In these two examples, PP models only win by marginal advantage in terms of the description length and the inferred partitions given by PP models and general SBM are very similar. These observations promote the question that whether PP models are practically redundant if our goal is to infer the structures in data rather than justifying assortativity. As we we going to show, general models could actually "underfit" compared to PP models in some empirical networks. That means, there exist empirical networks where PP models achieve the best fit and their inferred network partitions have larger number of groups than that of general models.

We firstly find the best fitting model among DC-SBM, Nested DC-SBM, and PP models according to the MDL principle for every dataset in our network corpus. In the left panel of Fig. 4-9, we plot an indicator variable which represents the best fitting model for each network. In the right panel of Fig 4-9 is a histogram showing the frequency of each model being the best fitting model. Similar to what we have seen in Section 3.3, Nested DC-SBM achieves the best fit to data most of the time, suggesting that real-world networks often possess structures that are more general than simple assortativity. In addition, Fig 4-9 shows that general models are more likely to be the best fitting model in networks of relatively large size, reflecting by the distribution of points along the x-axis. Interestingly, this time we find more examples where PP models are preferred over general SBMs (32 examples this time compared to only 2 in Chapter 3). We will take a closer look at these examples and investigate how PP models perform differently compared to general models.

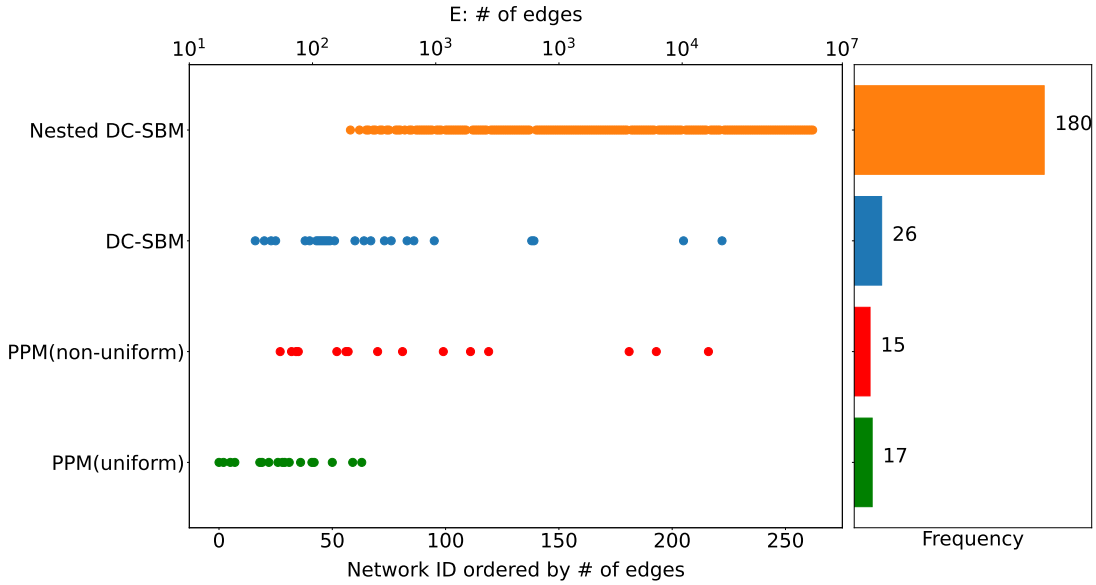


Figure 4-9: Frequency of models being selected as the best model according to the MDL principle. Networks are indexed in the increasing order of number of edges as indicated in the top axis. There are 25 networks in which all models agree on the lack of any modular structures (i.e. $B = 1$) and we exclude them from the this histogram.

We find PP models achieve the best fit in 32 out of 263 networks. Among the 32 networks where PP models win, 12 of them are known to have bipartite structure. In Table 4.1, we show the description length of each model. The uniform and non-uniform PP model show almost identical results in these bipartite networks, achieving shorter description length compared to general models. However, in fact, the difference between PP models and general models is also minor, as can be read from Table 4.2 and Table 4.3, showing the inferred number of communities and the partition overlap between each model variant and the best fitting model respectively. The only exceptions are the first three networks in Table 4.2 : the south African companies [140], wikipedia book edits [141] and plant-pollinator interaction at Safariland [142]. If we look at the inferred number of communities in the top panel of Table 4.3, it seems that PP models manage to resolve the bipartite structure in these three networks while general SBMs fail.

However, much of information in the posterior distribution has been overlooked when we only look at the MAP solution. If we draw samples from the the posterior distribution of DC-SBM and record the inferred number of communities and the description length, we observe that the difference between PP models and general DC-SBM is actually not significant. General SBMs actually also acknowledges the bipartite structure as a

	PP (uniform)	PP(non-uniform)	DCSBM	Nested DCSBM
South African companies	**35.00**	**35.00**	+1.37	+1.37
Wikipedia book edits	**75.97**	**75.97**	+0.38	+0.38
Plant-pollinator webs	**106.52**	**106.52**	+2.27	+2.27
Wiktionary edits	**99.62**	**99.62**	+3.82*	+3.82*
CEO Club Memberships	**258.41**	**258.41**	+3.88*	+3.88*
Elite affiliations	**294.39**	**294.39**	+3.92*	+3.92*
American revolution	**337.31**	**337.31**	+4.39**	+4.39**
Boards of directors	**749.43**	**749.43**	+4.54**	+4.54**
Swingers and parties	**746.50**	**746.50**	+4.84**	+4.84**
Kidnappings	**1673.93**	**1673.93**	+5.19**	+5.19**
EU procurement contract	+8.48*	**10831.68**	+2.96*	+2.96*
Foursquare NYC restaurants	+34.54**	**93569.10**	+8.81**	+8.81**

Table 4.1: Description length achieved by models in a set of 12 bipartite networks where PP models achieve better fit compared to general models. For best fitting models in each network, the value of description length is provided in bold. For other models, the difference between its description length and the best fitting model is shown. For reference purpose, description length differences are marked with marked with one asterisk $*$ if they are larger than $\ln 10 = 2.30$ (bits) or two asterisks $**$ if larger than $\ln 100 = 4.61$ (bits).

plausible explanation of structures in these three networks[4]. Therefore, despite being able to achieve better fit to data, the results of PP models are similar to general SBMs in networks with bipartite structures.

When we restrict our analysis to unipartite networks, we find the uniform PP model is the best fitting model for 7 networks. We show the description length, inferred number of communities and partition overlap between each model variant and the best model in Table 4.4 - 4.5. As can been seen from Table 4.4, uniform PP model only achieves marginal advantage in terms of description length, except for the student cooperation [143] and the American college football network [11]. We have seen in Section 3.3 the results of PP models and general SBMs are similar in the American college football example. However, this is not the case in the student cooperation network. In Fig 4-10, we visualise the communities found by different variants of SBMs in the student cooperation network. In this example, PP models identify more communities with small sizes by subdividing the communities identified by general SBMs. If we check from the angle of posterior distribution[5], in addition to the student cooperation network, the uniform PP model also concludes very different results from general models in the network of Illinois high school student and physician trust. Whereas in the other three networks in the set of networks where the uniform PP model outperforms (the US

[4]Results are given in the Appendix Fig. B-3 and Fig. B-5.
[5]Results are given in the Appendix Fig. B-4 and Fig. B-6.

	PP (uniform)	PP(non-uniform)	DCSBM	Nested DCSBM
South African companies	1.00	1.0	0.55	0.55
Wikipedia book edits	1.00	1.0	0.74	0.74
Plant-pollinator webs	1.00	1.0	0.74	0.74
Wiktionary edits	1.00	1.0	1.00	1.00
CEO Club Memberships	1.00	1.0	1.00	1.00
Elite affiliations	1.00	1.0	1.00	1.00
American revolution	1.00	1.0	1.00	1.00
Boards of directors	1.00	1.0	1.00	1.00
Swingers and parties	1.00	1.0	1.00	1.00
Kidnappings	1.00	1.0	1.00	1.00
EU procurement contract	0.97	1.0	1.00	1.00
Foursquare NYC restaurants	1.00	1.0	1.00	1.00

Table 4.2: The partition overlap between partitions given by each model and the best fitting model. Notice that uniform PP model and non-uniform PP model produce almost identical results. Regarding the comparison between general models and PP models, their inferred partitions are the same in 9 out of the 12 networks in this set, with exceptions being the first three small networks.

	PP (uniform)	PP (non-uniform)	DCSBM	Nested DCSBM
South African companies	2	2	1	1
Wikipedia book edits	2	2	1	1
Plant-pollinator webs	2	2	1	1
Wiktionary edits	2	2	2	2
CEO Club Memberships	2	2	2	2
Elite affiliations	2	2	2	2
American revolution	2	2	2	2
Boards of directors	2	2	2	2
Swingers and parties	2	2	2	2
Kidnappings	2	2	2	2
EU procurement contract	2	2	2	2
Foursquare NYC restaurants	2	2	2	2

Table 4.3: The inferred number of communities in the set of 12 bipartite networks where PP models achieve the best fitting models.

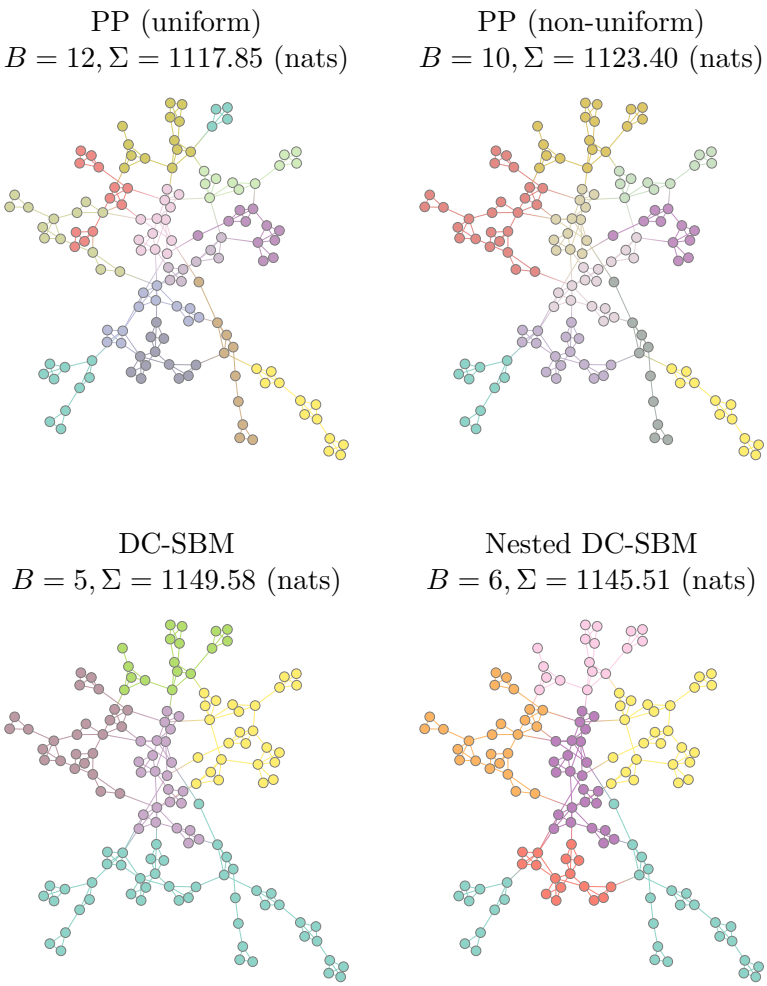


Figure 4-10: Inferred community structures in a network of computer science student cooperation [143]. The legends show the inferred number of communities as well as the description length of the corresponding partitions.

contiguous state [144], Adolescent health No.76 [145], PDZ-domain interactome [146]), the difference is not notable.

In comparison, in networks where the non-uniform PP model is the best fitting model, the difference between PP models and general SBMs is more evident, as shown in Table 4.7 - 4.9. Among these networks, there is a set of transportation and infrastructure networks, including the road network of streets in Abbeville [147], Alaska [148], Europe [149] and the American western states power grid [150]. Note that the generating process of these networks involve spatial constraints which are not incorporated in any variant of SBMs we have considered here. However, being suboptimal for describing these datasets does not prevent us from making comparison among them. In fact, it

	PP (uniform)	PP (non-uniform)	DCSBM	Nested DCSBM
US contiguous states	**377.17**	+0.5	+0.43	+0.43
Adolescent health No.76	**504.86**	+0.7	+0.3	+0.3
PDZ-domain interactome	**920.19**	+0.63	+3.5*	+3.5*
Student cooperation	**1117.86**	+7.14**	+31.72**	+27.65**
Illinois high school students	**832.94**	+4.41*	+2.94*	+2.94*
Physician trust	**1728.93**	+1.05	+0.56	+0.56
Football	**1761.50**	+1.63	+24.42**	+19.80**

Table 4.4: Description length of models in the set of 7 networks where the uniform PP model achieves the best fitting model. For best fitting models in each network, the value of description length is provided in bold. For other models, the difference between its description length and the best fitting model is shown. For reference purpose, description length differences are marked with marked with one asterisk $*$ if they are larger than $\ln 10 = 2.30$ (bits) or two asterisks $**$ if larger than $\ln 100 = 4.61$ (bits).

	PP (uniform)	PP (non-uniform)	DCSBM	Nested DCSBM
US contiguous states	1.0	0.86	0.88	0.88
Adolescent health No.76	1.0	0.53	0.53	0.53
PDZ-domain interactome	1.0	1.00	1.00	1.00
Student cooperation	1.0	0.80	0.44	0.52
Illinois high school students	1.0	1.00	1.00	1.00
Physician trust	1.0	0.86	0.86	0.86
Football	1.0	0.96	0.90	0.90

Table 4.5: Partition overlap between each model and the uniform PP model in a set of 7 networks where the latter is the best fitting model.

	PP (uniform)	PP (non-uniform)	DCSBM	Nested DCSBM
US contiguous states	2	3	2	2
Adolescent health No.76	2	1	1	1
PDZ-domain interactome	2	2	2	2
Student cooperation	12	10	5	6
Illinois high school students	4	4	4	4
Physician trust	2	2	2	2
Football	11	12	10	10

Table 4.6: The inferred number of communities in a set of 7 networks where the uniform PP model is the best fitting model.

	PP (uniform)	PP (non-uniform)	DCSBM	Nested DCSBM
Blumenau drug	+0.62	**634.30**	+0.8	+0.8
Adolescent health No.1	+3.52*	**704.07**	+0.19	+0.19
Marvel partnerships	+14.6**	**1148.40**	+5.62 **	+5.89 **
Freshwater stream webs	+2.0	**682.90**	+2.34*	+2.34*
Adolescent health No.6	+13.61 **	**1308.05**	+1.16	+1.16
Abbeville city streets	+8.03 **	**2461.24**	+5.07 **	+6.17**
Political books network	+30.19**	**1337.69**	+5.75**	+5.75 **
Adolescent health No.67	+55.65**	**4723.52**	+34.04 **	+23.11 **
Euroroad	+31.68 **	**8919.45**	+8.66**	+6.86**
Adolescent health No.27	+35.37 **	**15241.26**	+248.4 **	+30.83**
Western US Power Grid	+463.64 **	**51927.76**	+870.71 **	+556.42 **
Roads in Alaska	+8812.95**	**568568.02**	+8234.43 **	+531.96**

Table 4.7: Description length in a set of 13 networks where the non-uniform PP model is the best fitting model. For best fitting models in each network, the value of description length is provided in bold. For other models, the difference between its description length and the best fitting model is shown. For reference purpose, description length differences are marked with marked with one asterisk $*$ if they are larger than $\ln 10 = 2.30$ (bits) or two asterisks $**$ if larger than $\ln 100 = 4.61$ (bits).

	PP (uniform)	PP (non-uniform)	DCSBM	Nested DCSBM
Blumenau drug	0.84	1.0	1.00	1.00
Adolescent health No.1	1.00	1.0	1.00	1.00
Marvel partnerships	0.93	1.0	0.73	0.73
Freshwater stream webs	0.95	1.0	1.00	1.00
Adolescent health No.6	0.71	1.0	0.78	0.78
Abbeville city streets	1.00	1.0	1.00	1.00
Political books network	0.82	1.0	0.91	0.91
Adolescent health No.67	0.71	1.0	0.87	1.00
Euroroad	0.67	1.0	1.00	1.00
Adolescent health No.27	0.64	1.0	0.42	0.46
Western US Power Grid	0.75	1.0	0.57	0.69
Chicago road	0.68	1.0	0.44	1.00
Roads in Alaska	0.29	1.0	0.43	0.93

Table 4.8: Partition overlap between each model and the non-uniform PP model in a set of 13 networks where the latter is the best fitting model.

	PP (uniform)	PP (non-uniform)	DCSBM	Nested DCSBM
Blumenau drug	2	2	2	2
Adolescent health No.1	4	4	4	4
Marvel partnerships	3	4	3	3
Freshwater stream webs	2	2	2	2
Adolescent health No.6	4	4	5	5
Abbeville city streets	3	3	3	3
Political books network	2	4	3	3
Adolescent health No.67	10	8	7	8
Euroroad	3	4	4	4
Adolescent health No.27	46	49	17	24
Western US Power Grid	23	34	14	16
Chicago road	82	91	36	90
Roads in Alaska	29	131	25	112

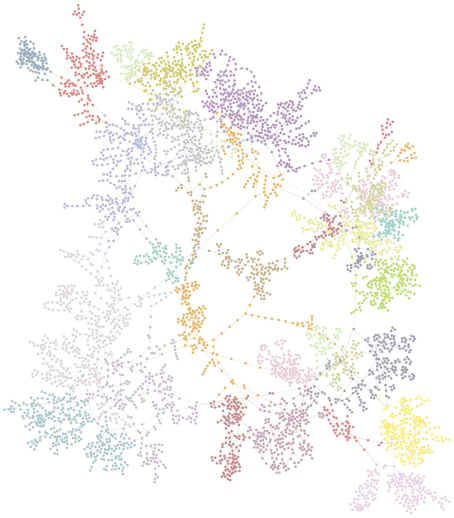
Table 4.9: Number of inferred communities in in a set of 13 networks where the non-uniform PP model achieves the best fit.

is an understandable result that the non-uniform PP model stands out in these cases. These spatial networks can be easily divided into locally densely connected groups which correspond to different spatial regions. As an example, we visuialise the inferred communities in the American western state power grid [150] in Fig 4-11. Based on visual inspection, the partition given by the non-uniform PP model seems to largely agree with that of the DC-SBM at a coarse-grained level. Both uniform and non-uniform PP models manage to find more detailed structures (23 and 33 communities given by uniform and non-uniform PP model respectively) with shorter description length of data than general SBMs. Although Nested DC-SBM is able to make improvements over DC-SBM, the inferred number of communities given by Nested DC-SBM is just about a half of that of the non-uniform PP model (14 and 16 for DC-SBM and Nested DC-SBM respectively), with a not negligible difference in the description length.

There are non-spatial networks where non-uniform PP model returns significantly different results from that of general SBMs as well. One example is the a social network from the Adolescent health dataset [137][6]. We demonstrate how the non-uniform PP model differs from general SBMs by plotting the edge count matrices of their inferred partitions. As shown in Fig. 4-12, all model variants agree on the dominance of assortative structure in this network. The DC-SBM clearly underfits compared to others, concluding a partition with only 17 communities. Using Nested DC-SBM leads to a partition with 24 communities and a large reduction in description length, from $\Sigma_{\text{DC-SBM}} = 15489.66$ (nats) to $\Sigma_{\text{Nested}} = 15272.09$ (nats). However, the non-uniform PP model manages to find an even more detailed partition with 49 commu-

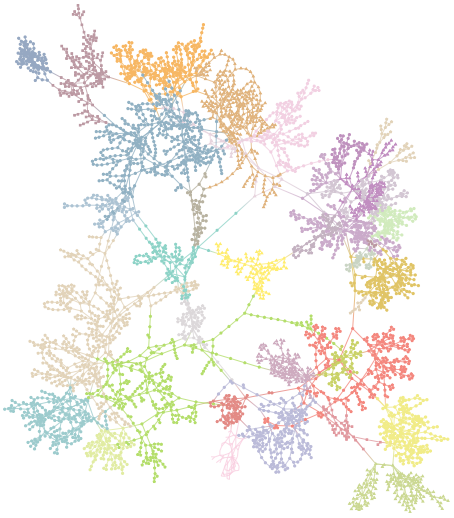
[6]This is the 27th network in the dataset.

PP (uniform)



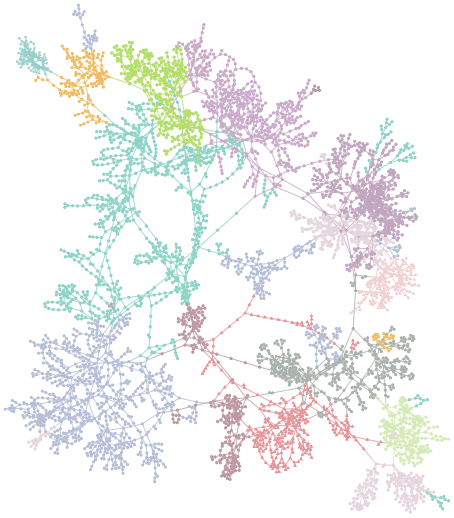
$B = 23, \Sigma = 52391.40$ (nats)

PP (non-uniform)



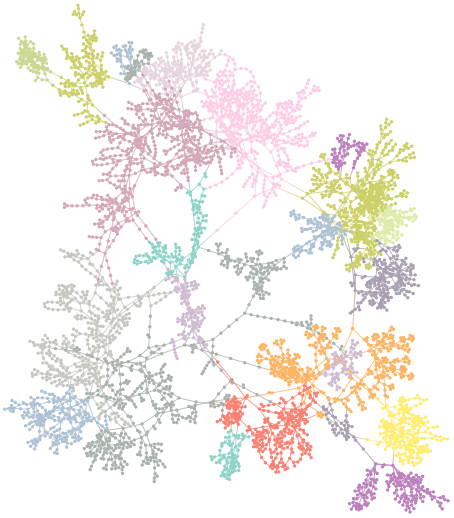
$B = 33, \Sigma = 51927.76$ (nats)

DC-SBM



$B = 14, \Sigma = 52798.47$ (nats)

Nested DCSBM



$B = 16, \Sigma = 52482.17$ (nats)

Figure 4-11: Inferred community structures in the American western states power grid network [150] given by PP models, DC-SBM and Nested DCSBM. Nodes are transforms or power relay points and edges represent power transmission relationship.

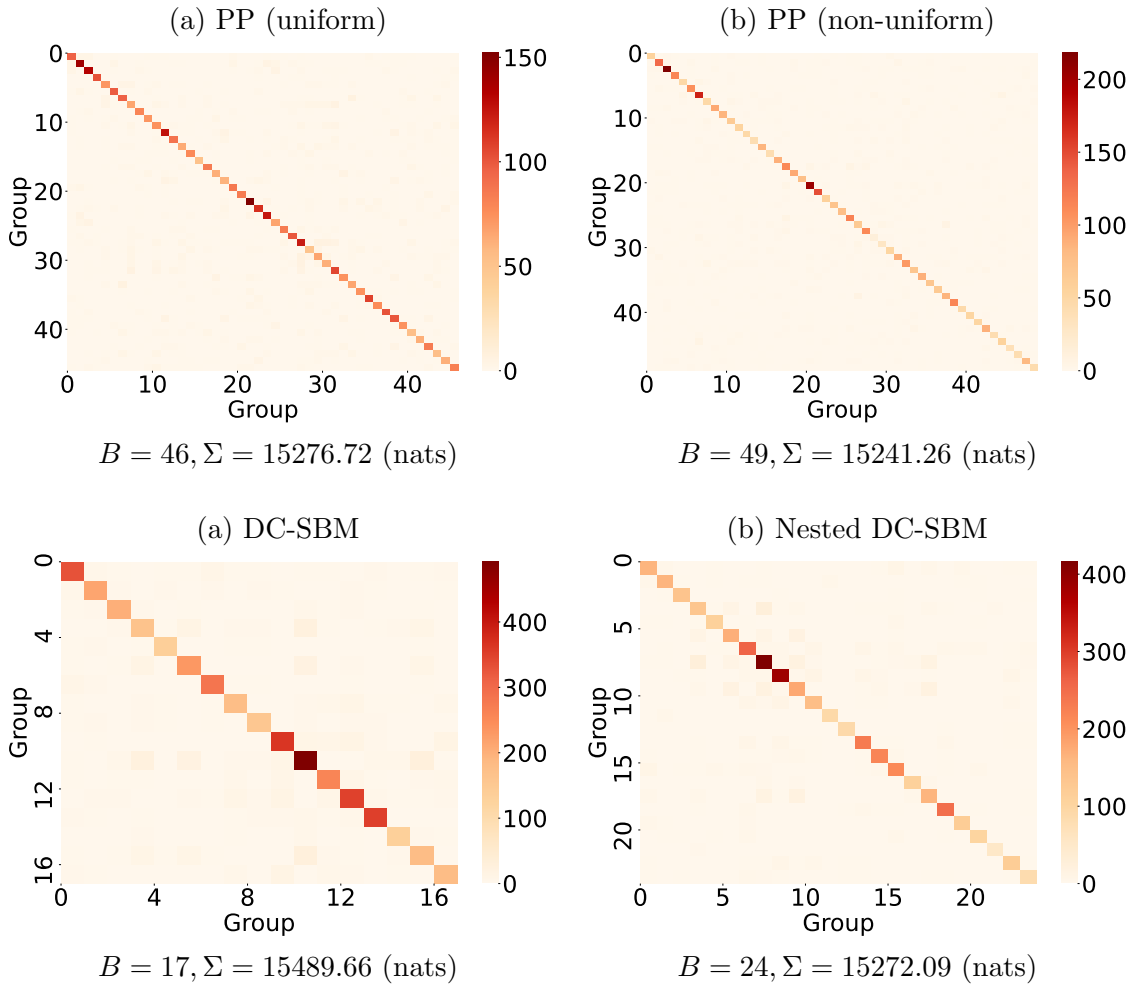


Figure 4-12: Edge count matrices of inferred partitions of the No.27 network from the Adolescent health dataset [137] given by PP models, DC-SBM and Nested DC-SBM. The legends show the inferred number of communities B and description length of the corresponding partitions.

nities and achieve the shortest description length of data among all model variants, $\Sigma_{\text{non-uniform PP}} = 15241.25$ (nats). In other words, the Nested DC-SBM "underfits" compared to the non-uniform PP model in this example, despite the former being known to not have the underfitting problem. Although it is debatable whether PP models' advantage in description length is significant enough to completely reject general models, it is clear that the non-uniform PP model manages to find a more refined partition of the network, which is not less plausible than that given by general models. Although Nested DC-SBM seems to be the most powerful variant with the most sophisticated construction, PP models demonstrate its ability of achieving the best fit if the dominant pattern in networks are assortative, or if the true network generating process significantly deviates from the process described by SBMs (e.g. spatial networks). Therefore, in practice, when we have no prior knowledge about the network, we might just try each of model variants available to us and select the one with the shortest description length. As a result, far from being redundant extensions, PP models are important complement to the general models since they have the potential of providing extra insight into data.

4.5 The underfitting and overfitting behaviour of modularity maximisation

4.5.1 Underfitting in synthetic networks

The modularity maximisation approach also has the problem of underfitting data. As a result, modularity maximisation has its own resolution limit [51]. The resolution limit of modularity maximisation is commonly demonstrated by considering a network consisting of a ring of cliques, as shown in Fig. 4-13. Each clique has only two edges connecting them with two neighbouring cliques. Because the connections within each cliques are much denser than that between distinct cliques, it is intuitive to expect community detection algorithms to assign each clique into its own group. However, the maximum modularity approach favours the partition that merges every two neighbouring communities together. We can obtain an estimate of the resolution limit of modularity maximisation just as we did for SBMs. Consider a special set of partitions in which cliques are merged into equal-size groups. For these network partitions, we have $e_r = (2E - B)/B$ and $e_{rr} = 2E/B$. Then, the modularity measure of these special network partitions defined in equation (3.29) becomes a function of the number of

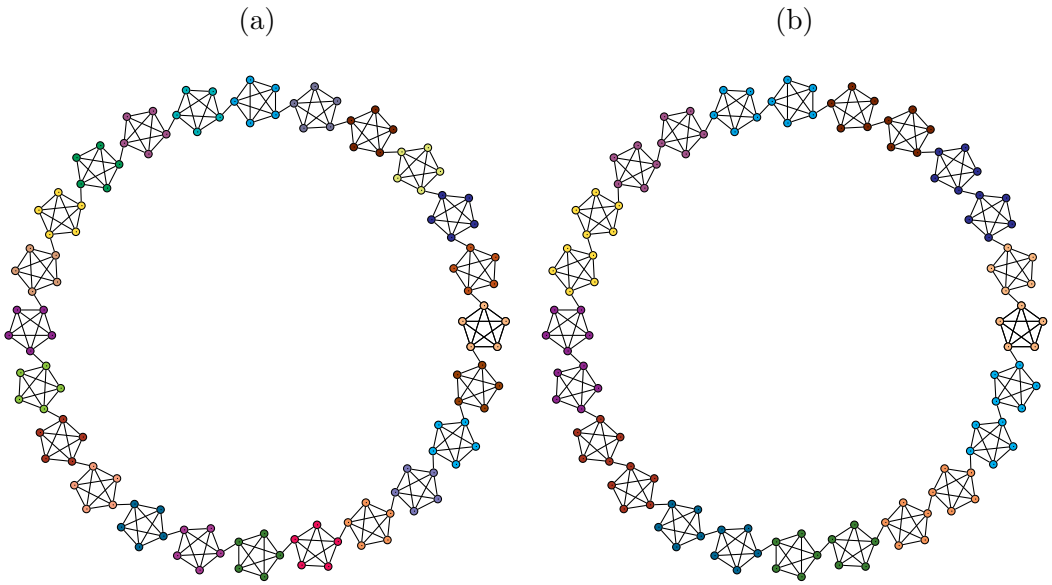


Figure 4-13: Two different partitions of a ring of cliques network with 24 cliques of size 5. Colouring of nodes indicates the corresponding network partition. Legends show the corresponding number of communities B, modularity value Q and the description length of the uniform PP model.

communities:

$$Q(\boldsymbol{A}, \boldsymbol{b}) = \frac{1}{2E} \sum_{r}^{B} \left(\frac{2E - 2B}{B} - \frac{(2E/B)^2}{2E} \right) = \frac{1}{2E} \left(2E - 2B - \frac{2E}{B} \right)$$
$$\Rightarrow Q(\boldsymbol{A}, \boldsymbol{b}) = 1 - \frac{B}{E} - \frac{1}{B}. \tag{4.24}$$

Treating B as a continuous variable and setting the derivative with respect to B to zero gives

$$B_{\max} = \sqrt{E}. \tag{4.25}$$

This means the modularity maximisation has a resolution limit which is at the same scale $\mathcal{O}(\sqrt{E})$ as the DC-SBM. In comparison, we have seen in Section 4.2 that PP models do not have the resolution limit problem. In this ring of cliques example, the correct partition indeed has a shorter description length than the merged partition according to the uniform PP model, shown in the legends in Fig 4-13.

Recall that the generalised modularity we introduced in Chapter 3 was proposed to address the resolution limit problem of modularity maximisation. The generalised

modularity Q_γ for partitions that merge cliques together has the following expression

$$Q_\gamma(B) = 1 - \frac{B}{E} - \frac{\gamma}{B}. \tag{4.26}$$

Then the optimal number of communities becomes

$$B_{\max} = \sqrt{\gamma E}. \tag{4.27}$$

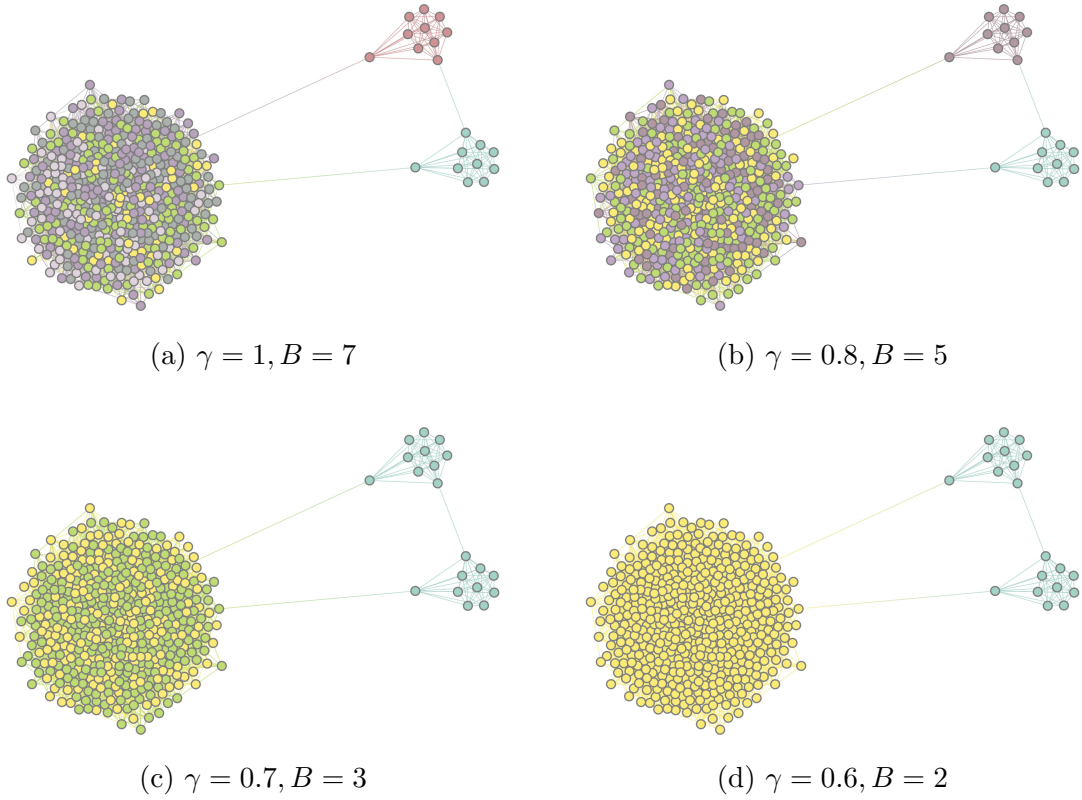
With this estimate of the resolution limit in mind, it was argued that we can resolve detailed structure which would have been blinded from the original modularity by adjusting the resolution parameter γ to some values larger than one [53].

Unfortunately, the generalised modularity Q_γ is not an ideal solution for two reasons. Firstly, the resolution limit parameter is an extra input required by the algorithm and this information is usually unavailable in practice. The author in [57] argued that we can estimate the value of the resolution parameter γ by exploiting the connection between modularity and the uniform PP model. That means, we might use the γ_{fit} as defined equation (3.70), which reads as

$$\gamma_{\text{fit}} = \frac{\lambda_{\text{in}^*} - \lambda^*_{\text{out}}}{\ln \lambda^*_{\text{in}} - \ln \lambda^*_{\text{out}}}, \tag{4.28}$$

where $\lambda^*_{\text{in}}, \lambda^*_{\text{out}}$ are maximum likelihood estimators of the λ_{in} and λ_{out} parameters of the uniform PP model. However, as we have explained in Section 3.1, the connection between modularity and uniform PP model does not hold in general. Therefore, there is no reason to believe that estimating γ with γ_{fit} is better than other arbitrary choices.

Secondly, the generalised modularity measure Q_γ can only search for structures at one single resolution at a time. Nevertheless, real-world networks often posses structures which vary in size, thereby exhibiting structures at multiple resolutions at the same time [14, 52]. Lancichinetti and Fortunato [54] demonstrated the deficient performance of generalised modularity in networks with heterogenous distribution of community sizes. If the resolution limit parameter strongly biases toward high-resolution structures, then significant structures of large size will be mistakenly split into small communities in order to accommodate the subjective bias. As a concrete example, we generate a synthetic network consisting of two small and one large communities. The two small communities are cliques of size 10, while the large one has 512 nodes with internal edges being randomly placed such that its average is 10. In Fig 4-14, we show the inferred community structures given by maximising the generalised modularity Q_γ with varying values of the resolution limit parameter γ. When $\gamma = 1$, the two small



(a) $\gamma = 1, B = 7$ (b) $\gamma = 0.8, B = 5$

(c) $\gamma = 0.7, B = 3$ (d) $\gamma = 0.6, B = 2$

Figure 4-14: Inferred community structures in a synthetic network with three communities of different sizes, using the generalised modularity Q_γ with varying values of the resolution parameter γ. Legends give the values of γ as well as the corresponding inferred number of communities.

cliques are correctly identified, while the third large community is mistakenly divided into several sub-communities. Decreasing the value of γ introduces bias toward partitions with large-size communities. However, before all of the sub-communities in the large community are correctly merged together, the two small cliques are grouped into a larger group. In fact, Lancichinetti and Fortunato showed in [54] that there is no a single value of γ that allows detection all of three communities in this synthetic example.

The undesired performance of the generalised modularity in networks with unequal communities is understandable: The underlying model of modularity maximisation is the uniform PP model, which is restrictive to regular assortative structures. However, we point out that the tendency of splitting the large community in the last example is not only due to the restrictive nature of the underlying model, but also due to the fact that statistical significance of inferred partitions is not properly taken into account.

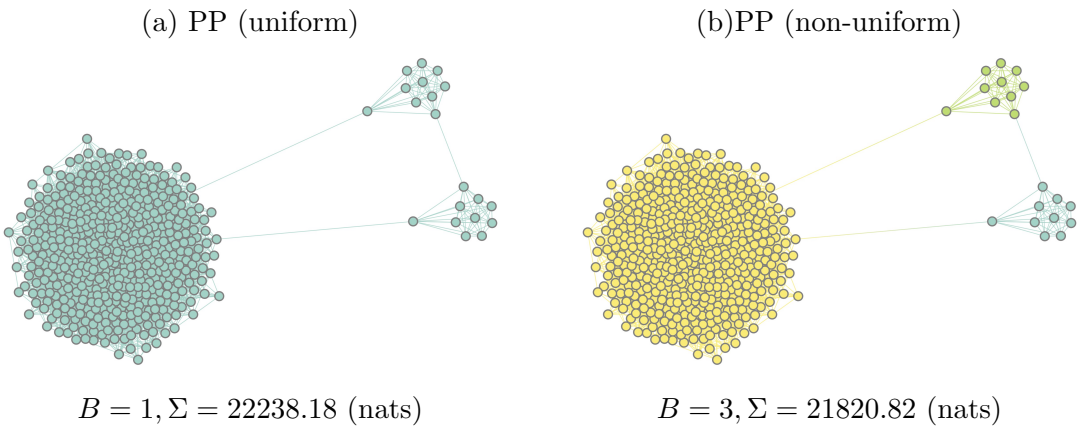


Figure 4-15: Inferred community structures in a synthetic network with three communities of different sizes, using the uniform and non-uniform PP models. Legends give the corresponding description length Σ of data and inferred number of communities.

For this reason, the generalised modularity Q_γ suffers from the overfitting problem[7] just as the original modularity does. For comparison, we fit the uniform PP model to this example and the most plausible partition is the one assigning all of the nodes into a single group, as shown in 4-15(a). Although this is still not recovering the correct structure, we know the uniform PP model is insufficient to describe the structure in this network, and replacing the uniform PP model with the non-uniform PP model can successfully identify all of three communities with a shorter description length than the trivial partition given by the uniform PP model (see Fig 4-15(b)). Therefore, the non-uniform PP model is generally better than its uniform counterpart as well as other modularity-based methods, because it not only fixes overfitting and underfitting at once, but also enables the detection of multiple-resolution assortativity.

4.5.2 Underfitting in empirical networks

We also want to know whether modularity maximisation underfits in empirical networks and whether underfitting could happen below the $\mathcal{O}(E)$ scaling of the resolution limit. We apply modularity maximisation to the network corpus we have considered in Section 4.3. We only compare the modularity maximisation to Bayesian inference with the uniform PP model since these two methods share the same underlying model. In Fig 4-16, we plot the difference in the inferred number of communities between the uniform PP model and modularity maximisation, $B_{\text{ppm}} - B_{\text{modualrity}}$, for each dataset in our network corpus. We find that modularity maximisation finds larger number

[7]Here we abuse the term "overfitting" to refer to the fact that generalised modularity finds exaggerated results, while more strictly we say a method overfits if it returns structures in fully random networks.

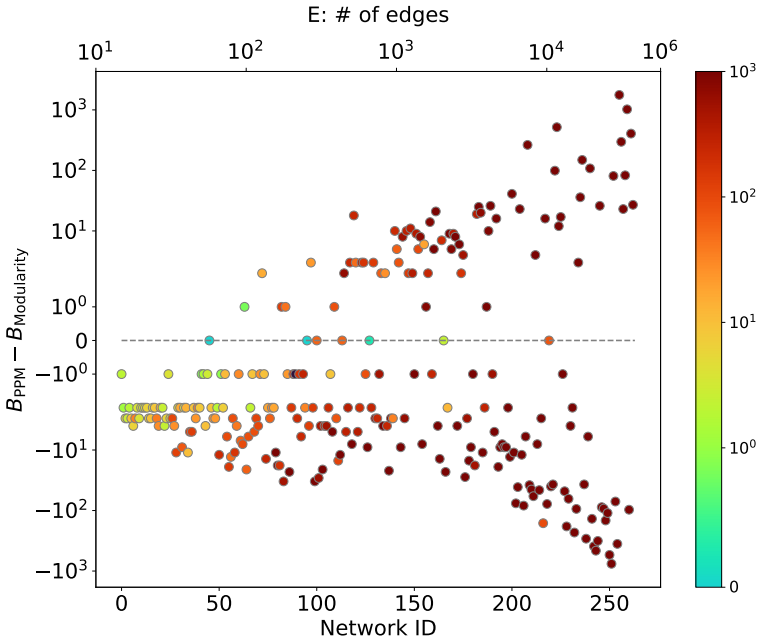


Figure 4-16: Difference in the inferred number of communities between the uniform PP model and the modularity maximisation approach, i.e. $B_{\mathrm{PPM(unif)}} - B_{\mathrm{Modularity}}$. The colouring of the points indicates the difference in the description length $\Sigma_{\mathrm{Modularity}} - \Sigma_{\mathrm{PPM(unif)}}$, where the description length of the partition given by modularity maximisation is computed using the same expression as the uniform PP model.

of communities than that of the uniform PP model in over a half of networks in our dataset (194 out of 263). This result does not contradict our previous analysis that modularity maximisation underfits while uniform PP model does not, because the partitions given by modularity maximisation are less plausible than that of the uniform PP model as indicated by the description length difference in the colorer on the right. The synthetic example in Fig 4-14 already hints us about why the inferred number of communities given by modularity maximisation outnumbers that of the uniform PP model: modularity maximisation lacks of proper regularisation and therefore has the tendency of returning exaggerated results.

To further illustrate the extent to which modularity maximisation overfits, we extend our experiment with the protein-protein interaction example in Section 3.3 to all networks in our network corpus. We apply the modularity maximisation approach to networks in our corpus as well as their randomised counterparts, which are generated from the configuration model. As shown in Fig 4-17, modularity maximisation returns spurious communities in random networks. For comparison, the uniform PP model correctly concludes that there is no structures. Moreover, in Fig. 4-18 (a), we plot

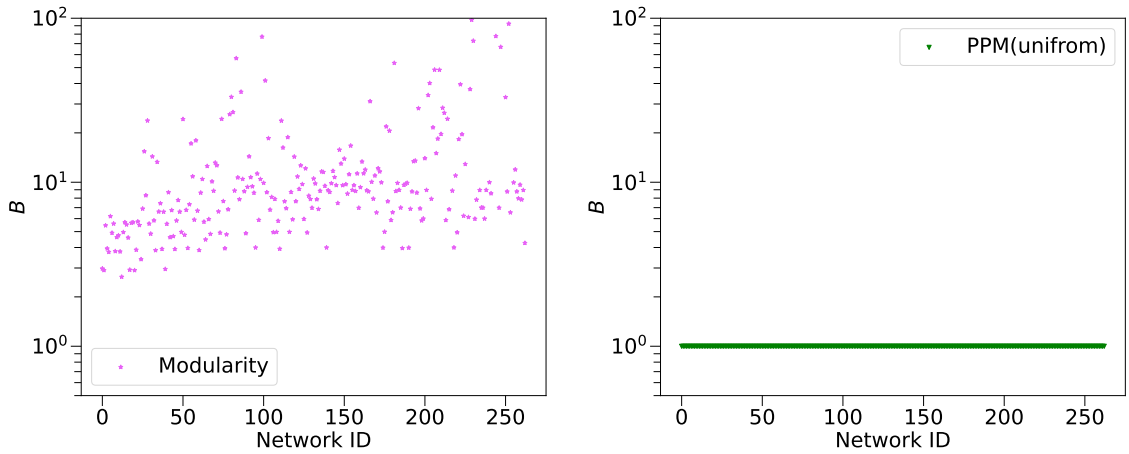
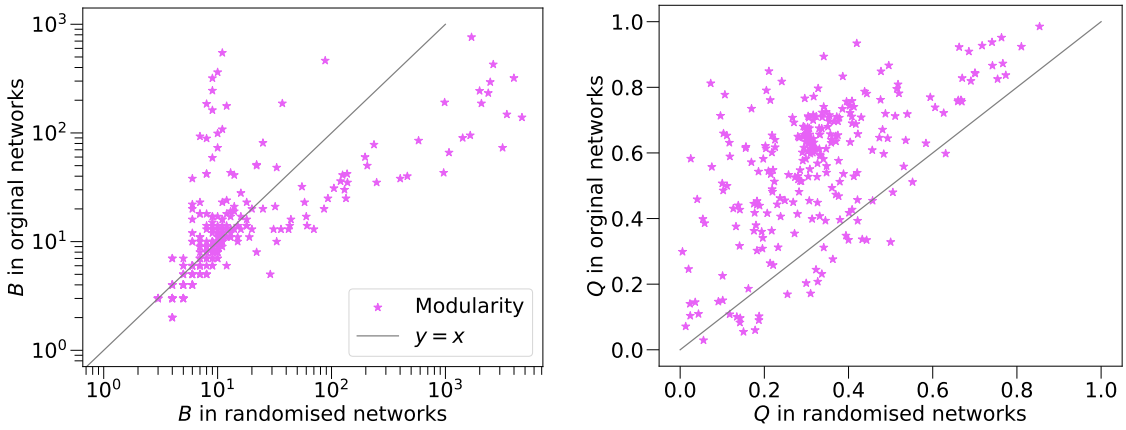


Figure 4-17: Inferred number of communities in randomised networks, using (a) modularity maximisation (b) uniform PP model. The results of DC-SBM, Nested DC-SBM and non-uniform PP model are given in Appendix B-7.

the inferred number of communities given by modularity maximisation in the original networks against that in randomised networks. There is a strong correlation, indicating that a non-negligible portion of the modular structures detected in the original networks can be simply explained by the degree sequence of nodes. Similarly, Fig. 4-18 (b) shows the overfitting behaviour of modularity maximisation from a different perspective: even in randomised networks, modularity maximisation always finds partitions with positive modularity values, and the modularity values found in original networks seems to be positively related to that in randomised networks.

Having seen that modularity maximisation systematically overfits in our network corpus, we expect that modularity maximisation runs into the underfitting problem with a large chance in those cases where its inferred number of communities is smaller than that of the uniform PP model. This is because modularity maximisation's tendency of overfitting makes it more likely to return more communities than what actually exists in data. In comparison, Bayesian inference with the uniform PP model is generally more conservative and should be closer to the correct structure, if any existed, according to the minimum description principle. As a result, seeing that generally exaggerated modularity maximisation outnumbers the conservative uniform PP model is a strong sign of underfitting problem occurring to modularity maximisation. From Fig 4-16, we can see the underfitting problem of modularity maximisation is more severe in networks with larger sizes. Moreover, we identify networks where modularity maximisation underfits relative to the uniform PP model, and plot the inferred number of communities given



(a) (b)

Figure 4-18: (a) Inferred number of communities given by modularity maximisation in the original network as a function of that obtained in the randomised network for every dataset in our network corpus. The grey line is the identity function $y = x$. (b) Modularity values found in original networks against that of randomised networks.

by the latter in Fig 4-19. It turns out that the inferred number of communities given by the uniform PP model are often not close to the $\mathcal{O}(\sqrt{E})$ scale of the resolution limit of modularity maximisation. Although we should not treat the results of uniform PP model as the underlying truth, this observation provides strong evidence that modularity maximisation could suffer from the underfitting problem even when the number of communities are below its resolution limit.

4.6 Concluding remarks

By analysing a large empirical network corpus, we confirm that the underfitting problem of DC-SBM is widespread. Specifically, we find that there exists a positive correlation between the difference in inferred number of communities and the difference in description length between the single-layer and Nested DC-SBM. Using Nested DC-SBM generally allows us to extract more detailed structures compared to DC-SBM. Although the PP models do not have the underfititng problem, since they are assortative-constrained variants and most real-world networks have more general community structures, PP models often find more conservative results compared to general models. Therefore, when it comes to characterising different community detection algorithms, it is important to keep in mind that algorithms designed for different structures might intrinsically have different behaviours. When assortativity is the dominant pattern in data, we find several examples where PP models reveal notably different results compared to general SBMs. Our results suggest that PP models are important extension of the existing col-

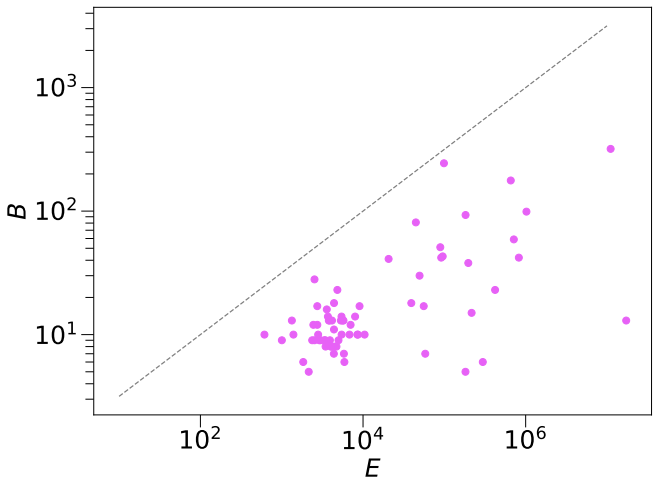


Figure 4-19: Inferred number of communities found by the uniform PP model in networks where $B_{\text{PPM(unif)}} - B_{\text{Modularity}} > 0$. The grey dashed line is the resolution limit of the modularity maximisation approach $\sqrt{E}$.

lection of network models, since they have the potential of providing extra insight into data. In general, we expect Nested DC-SBM to be a good starting point for identifying large-scale community structures in networks. If the network of interest is relatively small and sparse, then it is worth considering simpler models like DC-SBM and PP models. If we are only interested in assortative structures, then PP models should be the preferred over general models.

Modularity maximisation suffers from the overfitting and underfititng problem at the same time. Even though the underfitting problem of modularity maximisation is often covered by its overfitting behaviour, we find evidence that modularity maximisation could underfit in practice, even when the number of communities is below the theoretical estimate of resolution limit. Adapting the generalised modularity does not solve the problem, since the resolution limit parameter is generally unknown and it does not address the tendency to overfit data. Our PP models are better alternatives for extracting assortative structures, because they address the overfitting and underfitting problem at once, and the non-uniform PP model is able to detect assortative structures at multiple resolutions without the need for any ad-hoc parameter.

Chapter 5

Detectability of community structures in SBMs

Although Nested DC-SBM and PPMs assure us that significant structures will not be missed, it is worth pointing out that there exists a fundamental limit of community detection in networks generated from SBMs. This limit was firstly demonstrated by Decelle et al. in [65], where a *phase-transition* phenomenon was discovered in community detection: there exists a non-trivial threshold of the strength of community structure. Below the threshold, SBMs are in an *undetectable* phase, in which no polynomial community detection algorithms can do better than random guessing. In other words, even in networks that contain community structures, we might still not be able to recover the structures, unless their strength are sufficiently strong. This result was revealed by analysing the stability of the *belief propagation* algorithm - a semi-parametric algorithms for computing the marginal distribution of variables on graphs. The belief propagation algorithm allows Decelle et al. to demonstrate the detectability phase-transition phenomenon via numerical simulations, which has been rigorously proved for networks with two equal-sized communities [59, 67].

Knowing the existence of the detectability phase-transition affects the way we interpret the results of community detection. Intuitively, if fitting a SBM to an observed network leads to a trivial partition where every node is in a single group, we might conclude that there is no community structure in the network and edges are just randomly placed among nodes. However, having been told that there is an undetectable phase of SBMs, we should adjust our conclusion. When applying the inference approach with SBMs finds no communities, it is possible that certain structure exists, but its strength is

below the detectability threshold.

The detectability phase-transition in community detection was firstly reported in networks where every community is assumed to share the same average degree [65]. There were a series of following works that studied the detectability of community detection under different, arguably more realistic conditions. In [70], the authors considered a semi-supervised setting where the correct labels of a fraction of nodes are provided as side information for the inference. In the semi-supervised learning setting, it turns out the phase-transition phenomenon disappears. That means, as long as community structures exist, we can find network partitions that are well-correlated to the correct structures, no matter how weak the structures are. The detectability phase-transition also disappears in networks with asymmetric communities [68,69]. Asymmetry of communities affects the detectability of community structure via the average degree of nodes: If nodes in different communities have different average degrees, then the average degree can be leveraged to facilitate the inference when the strength of community structure is weak. These studies around the detectability phase-transition not only have deepened our understanding of the limitation of community detection in practical applications, but also have provided guidance for the development of novel detection algorithms [151].

All of the aforementioned works are restricted to networks with homogenous degree distribution. However, degree distribution in real-world networks are usually heterogeneous. Therefore, previous results have provided limited insight into the fundamental limit of community detection in realistic scenarios. In [72], Guilders et al. provided rigorous proof of the detectability of community structures in networks with heterogeneous degree distribution in SBMs. It turns out that the detectability threshold will decrease as the heterogeneity in degree distribution increases. The positive part of this detectability results is later confirmed in [152] where a spectral algorithm was shown to recover planted structures up to the theoretical detectability threshold. However, the analysis in [72] has only considered networks with two communities and is hard to be generalised to general cases where networks contain more than two communities. It has been shown that the detectability phase-transition phenomenon can be significantly different when it comes to networks with more than two communities in homogeneous case [65, 68]. In this Chapter, in order to explore the detectability phase-transition phenomenon in more general cases, we consider an alternative approach for studying the detectability phase-transition. In particular, we apply the belief propagation (BP) algorithm for DC-SBM to networks with customised heterogeneous degree distributions. Our numerical results confirm the effect of the heterogeneous degree distribution

on the detectability of community structures: as the heterogeneity in degree distribution increases, we find that planted structures become more detectable in the sense that the area of undetectable phase shrinks. Our results and modified algorithm shall serve as the stepping stone to further investigation of detectability phase-transition in more general settings, e.g. in networks with more than two communities or asymmetric degree distribution.

Following this introduction, in Section 5.1, we introduce the BP algorithm for the Bernoulli SBM. In Section 5.2, we apply the algorithm to demonstrate the detectability phase-transition when the degree distribution is homogeneous. In Section 5.3, we explain our experiment for studying the effect of heterogeneity of degree distribution on detectability and present our numerical results.

5.1 Belief propagation

Studies of phase-transition phenomenon in community detection often relies on the belief propagation (BP) algorithm [94]. BP is a message-passing method that provides estimate of the marginal distribution of variables in graphs. By applying BP to the inference of SBMs, one can obtain the the marginal posterior distribution of the group assignment of each node u,

$$P(b_u = r|\boldsymbol{A}) = \sum_{\boldsymbol{b}\backslash u} P(\boldsymbol{b}|\boldsymbol{A}) := q_r^u. \tag{5.1}$$

The *marginal estimator* given by this marginal distribution

$$b_u^* = \underset{r}{\operatorname{argmax}}\ q_r^u \tag{5.2}$$

is the optimal estimator in terms of the number of nodes being correctly labeled [153]. Although the MCMC algorithm we introduced in Chapter 1 can provide estimate of the marginal posterior distribution as well, BP is more efficient because it computes the marginal distribution $\{q_r^u\}$ directly, while MCMC relies on drawing large amount of samples. In the following, we define the BP algorithm for the inference of the Bernoulli SBM, which will be used to illustrate the detectability phase-transition.

Recall that the Bernoulli SBM takes the probability of connections among communities $\{p_{rs}\} \in \mathbb{R}_{[0,1]}^{B\times B}$ as input. Furthermore, we assume nodes are independently assigned into one of $\{1, 2, .., B\}$ groups according to a prior distribution $\{\eta_r\}$ with $\sum_r^B \eta_r = 1$. Then we go through all possible pairs of nodes in the network, place an edge between a pair

of nodes u and v with probability $p_{b_u b_v}$. The probability of generating an observed network $\boldsymbol{A}$, together with the network partition $\boldsymbol{b}$, is then

$$P(\boldsymbol{A}, \boldsymbol{b}|\{\eta_r\}, \{p_{rs}\}) = \prod_u \eta_r \prod_{u<v} p_{b_u b_v}^{A_{uv}} (1 - p_{b_u b_v})^{1-A_{uv}}. \tag{5.3}$$

Because real-world networks are usually sparse, we assume the probabilities p_{rs} are in the order of $\mathcal{O}(1/N)$, such that networks generated from the model do not become denser as the size of networks increases. It is therefore convenient to work with the rescaled connection matrix $\{c_{rs}\}$, where each entry c_{rs} satisfies $c_{rs} = p_{rs}N$, hence having a scale at $\mathcal{O}(1)$. We will further assume that we know the true modelling parameters $\{c_{rs}\}$, $\{\eta_r\}$, and the number of communities B. Then, the task of community detection is simply to decide how to assign nodes into $\{1, 2, .., B\}$ groups according to the observed network $\boldsymbol{A}$. This assumption that we have access to the underlying data generating process might not seem realistic. Indeed, we rarely have knowledge about how our data is generated in practice[1]. However, assuming we know the generating process is useful for the purpose of understanding the fundamental limit of the detection problem. If we are not able to do well in this kind of best-case scenario, there is no reason to believe we can do better when the true modelling parameters are not available.

The BP algorithm consists of recursively computing a set of BP equations that are satisfied by the *conditional marginal* distributions, or *messages*. Messages are defined for all possible interactions among nodes. For any two nodes u and v in the network, the message sending from node u to v conveys the probability of node u belonging to one of the B communities, when node v is removed from the network (or equivalently when we do not know whether u and v are connected). We motivate the definition of messages for the Bernoulli SBM below and refer to [94] for further details. Our goal is to compute the marginal probability

$$P(b_u|\boldsymbol{A}) = \sum_{\boldsymbol{b}\backslash u} P(\boldsymbol{b}|\boldsymbol{A}) \propto \sum_{\boldsymbol{b}\backslash u} P(\boldsymbol{A}, \boldsymbol{b}|b_u = r) P(b_u = r). \tag{5.4}$$

Under the conditionally independent assumption which states that neighbours of a node u are independent with each other conditioned on the label of node u, the likelihood term $P(\boldsymbol{A}, \boldsymbol{b}|b_u = r)$ can be arranged into factors. Each factor consists of marginalisation of nodes in the branch where the starting point of the branch is a neighbour of

[1]We can still conduct inference of SBMs even when we have no knowledge about the parameters used to generate the data by adopting the expectation-maximisation procedure [65], where BP can provide estimates in the expectation step.

node u. In particular, we might rewrite the last equation as follows

$$P(b_u|\boldsymbol{A}) \propto P(b_u = r) \prod_{v\in\partial u}\left[\sum_{b_w:w\in\mathcal{B}(v)}\left(\prod_w P(b_w)\prod_{i<j:i,j\in\mathcal{B}(v)} P(A_{ij},b_i,b_j|b_u=r)\right)\right]$$
$$= \eta_r \prod_{v\in\partial u}\left[\sum_{b_w:w\in\mathcal{B}(v)}\left(\prod_{w\in\mathcal{B}(v)}\eta_{b_w}\prod_{i<j:i,j\in\mathcal{B}(v)} c_{b_ib_j}^{A_{ij}}(1-c_{b_ib_j}/N)^{1-A_{ij}}\right)\right], \quad (5.5)$$

where $\mathcal{B}(v)$ represents the set of nodes in the branch starting with node v. Then, in the square bracket in the last equation, we further rearrange the summation by leveraging the conditionally independent assumption. We can explicitly write out the marginalisation with respect to v, i.e. the set of neighbours of node u, and let $\mu_r^{v\to u}$ denote the messages, which are obtained by marginalising over neighbours of node v except for node u. Then, the expression in the last equation becomes

$$P(b_u|\boldsymbol{A}) \propto \eta_r \prod_{v\in\partial u}\left[\sum_{s=1}^{B}\mu_s^{w\to u}c_{rs}^{A_{uv}}\left(1-\frac{c_{rs}}{N}\right)^{1-A_{uv}}\right]. \quad (5.6)$$

To obtain $P(b_u|\boldsymbol{A})$, we just to need to update the values of messages $\{\mu_r^{u\to v}\}$ recursively, with each message having the following expression

$$\mu_r^{u\to v} = \frac{\eta_r}{Z^{u\to v}}\prod_{w\in\partial u\setminus v}^{N}\left[\sum_{s=1}^{B}\mu_s^{w\to u}c_{rs}^{A_{uv}}\left(1-\frac{c_{rs}}{N}\right)^{1-A_{uv}}\right], \quad (5.7)$$

where $Z^{u\to v}$ is the normalising constant ensuring $\sum_r \mu_r^{u\to v} = 1$.

The conditional independence assumption holds if the network contains no cycles, which are also referred to as *trees*. Although this requirement seems difficult to be satisfied in practice, BP is approximately exact in locally tree-like networks [65]. Locally tree-like means the typical length of cycles in the network is large, such that the neighbourhood of any nodes in the network looks like a tree. Networks generated from SBM are locally tree-like, because in the sparse regime where the probabilities between any pair of nodes are in the order of $\mathcal{O}(1/N)$, the typical length of a cycle $\mathcal{L}$ in the network is in the order of $\log N$. To see this, firstly consider the probability of forming a cycle of length 3 containing node u,

$$p_3^u = \sum_{v\neq u}^{N} p_{uv}\sum_{w\neq u,v}^{N} p_{vw}p_{wu} = \sum_{v\neq u} p_{uv}\mathcal{O}\left(\frac{\langle k\rangle}{N}\right) = \mathcal{O}\left(\frac{\langle k\rangle^2}{N}\right). \quad (5.8)$$

The second and third equal signs in the last equation make use of the assumption that

every node has the same expected degree $\langle k\rangle = \sum_v p_{uv} = \sum_v p_{wv}$[2]. According to the expression of p_3^u above, the density of cycles of length will be vanishingly small as the size of network becomes large. More generally, the probability of forming a cycle of length $\mathcal{L}$ has the following expression

$$p_{\mathcal{L}}^u = \mathcal{O}\left(\frac{\langle k\rangle^{\mathcal{L}-1}}{N}\right) \tag{5.9}$$

and such probability will not vanish as long as $\langle k\rangle^{L-1} = \mathcal{O}(N)$, which implies that $L = \mathcal{O}(\log_{\langle k\rangle} N)$. Therefore, the length typical cycles in networks generated from SBMs will grow as the size of networks increase, making neighbourhoods of nodes locally-treelike and hence we can apply BP to make inference of SBMs.

The BP messages defined in (5.7) leads to a series of equations which can be solved iteratively. The estimate of marginal distribution $\{q_r^u\}$ is then given by

$$\mu_r^u = \frac{\eta_r}{Z^u}\prod_{w\in\partial w}^{N}\left[\sum_{s=1}^{B}\mu_s^{w\to u}c_{rs}^{A_{uw}}\left(1-\frac{c_{rs}}{N}\right)^{1-A_{uw}}\right]. \tag{5.10}$$

However, there is a practical issue in computing BP equations in (5.7). Recall the generating process of SBMs, there are interactions between every pair of nodes in the network. As a result, there are in total $N(N-1) = \mathcal{O}(N^2)$ messages to update in each round of iteration, which is prohibitively expensive to track. We can get around this issue by noticing that all the messages on non-edges are in fact identical, except for an error up to the order $\mathcal{O}(1/N)$. To see this, consider u and v are not connected and we split the product in the message $\mu_r^{u\to v}$ into two parts, one for interactions along observed edges and the other for non-edges. The message $\mu_r^{u\to v}$ in equation (5.7) then becomes

$$\begin{aligned}
\mu_r^{u\to v} &= \frac{\eta_r}{Z^{u\to v}}\left(\prod_{w\notin\partial u}\sum_{s}^{B}\mu_s^{w\to u}\left(1-\frac{c_{rs}}{N}\right)\right)\left(\prod_{w\in\partial u}\sum_{s=1}^{B}\mu_s^{w\to u}c_{rs}\right)\\
&= \frac{\eta_r}{Z^{u\to v}}\left(\prod_{w\notin\partial u}1-\sum_{s}^{B}\mu_s^{w\to u}\frac{c_{rs}}{N}\right)\left(\prod_{w\in\partial u}\sum_{s=1}^{B}\mu_s^{w\to u}c_{rs}\right)\\
&\approx \frac{\eta_r}{Z^{u\to v}}\exp\Big(-\sum_{w}\sum_{s}\mu_s^{w\to u}\frac{c_{rs}}{N}\Big)\prod_{w\in\partial u}\sum_{s=1}^{B}\mu_s^{w\to u}c_{rs}+\mathcal{O}\left(\frac{1}{N}\right),
\end{aligned} \tag{5.11}$$

where ∂u is the set of neighbouring nodes of u. In the last equation, we make use

[2]Indeed, when it comes to networks with heterogeneous degree distribution, extra care is needed for verifying the validity of BP. This will be discussed later in Section 5.3

of the approximation $e^{-x} \approx 1 - x$ when x is sufficiently small. The approximation in equation (5.11) suggests that the messages on non-edges are independent of the destination node. On the other hand, if u and v are connected, then the message is indeed dependent on the destination node v,

$$\mu_r^{u\to v} = \frac{\eta_r}{Z^{u\to v}} \left(\prod_{w\notin\partial u} 1 - \sum_s^B \mu_s^{w\to u} \frac{c_{rs}}{N} \right) \left(\prod_{w\in\partial u/v} \sum_{s=1}^B \mu_s^{w\to u} c_{rs} \right). \tag{5.12}$$

With the observation above, we can rewrite the BP equations defined in (5.7) as follows

$$\mu_r^{u\to v} = \frac{\eta_r}{Z^{u\to v}} e^{-h_r} \prod_{w\in\partial u/v} \sum_s c_{rs} \mu_r^{w\to u}. \tag{5.13}$$

with terms up to order $\mathcal{O}(1/N)$ being ignored. The exponent h_r has the following expression

$$h_r = \frac{1}{N} \sum_w^N \sum_s^B \mu_s^w c_{rs}, \tag{5.14}$$

which is easy to track, since it only requires $\mathcal{O}(1)$ computation to update. After each time μ_s^w is updated, the h_r should be adjusted as follows

$$h_r^{\text{new}} = h_r^{\text{old}} - \sum_s (\mu_s^w)^{\text{old}} c_{rs} + \sum_s (\mu_s^w)^{\text{new}} c_{rs}. \tag{5.15}$$

And the final estimate of $\{q_u^r\}$ is given by

$$\mu_r^u = \frac{\eta_r}{Z^u} e^{-h_r} \prod_{w\in\partial u} \sum_s c_{rs} \mu_r^{w\to u}. \tag{5.16}$$

As a result, we only need to record in total $2E$ messages and each message requires $\mathcal{O}(\langle k\rangle B)$ computations. For sparse networks in which the average degree is significantly smaller than the size of the network, i.e. $\langle k\rangle \ll N$ and the degree distribution of nodes is homogeneous, the total time complexity is $\mathcal{O}(E)$, which is highly scalable for networks of large size. We summarise the BP algorithm for the Bernoulli SBM below. In the next section, we apply this algorithm to demonstrate the detectability phase-transition in community detection.

Algorithm 5.1: BP inference for Bernoulli SBM

1. Initialise $\{\mu_r^{u\to v}\}$ for every edge (u,v) in $\mathcal{E}$

2. Compute $\{\mu_r^u\}$ according to equation (5.16) and the $\{h_r\}$ according to equation (5.14)

3. For every edge $(u,v)\in\mathcal{E}$, update $\mu_r^{u\to v}$ according to equation (5.13), then the message μ_r^v according to equation (5.16), and the $\{h_r\}$ according to (5.15)

4. Repeat step 3 until convergence

5.2 Detectability phase-transition in factorised SBMs

The detectability phase-transition in community detection was firstly demonstrated in networks where the *factorised condition* holds. The factorised condition states that the average degree of each community are identical. That means,

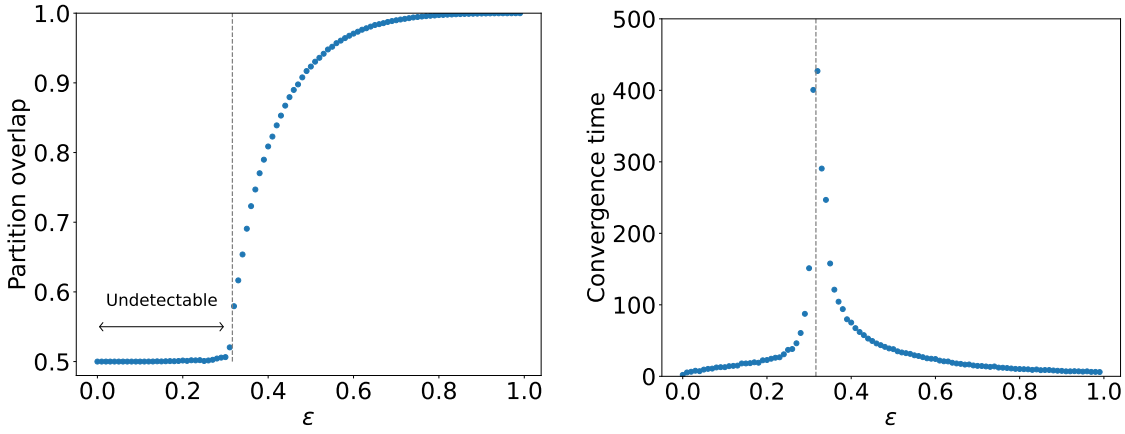
$$\langle k\rangle = \sum_a^B c_{ra}\eta_a = \sum_b^B c_{sb}\eta_b, \quad \forall r,s\in\{1,2,..,B\}. \tag{5.17}$$

The name of this condition comes from the fact that the prior probability $\boldsymbol{\eta}=\{\eta_r\}$ of network partition $\boldsymbol{b}$ is always a fixed point of belief propagation equations. Indeed, if we substitute $\mu_r^{u\to v}=\eta_r$ into belief propagation equations in (5.13), with the factorised condition in equation (5.17),

$$\mu_r^{u\to v} = \frac{\eta_r}{Z_u^{u\to v}} e^{-h_r} \prod_{w\in\partial u/v}\Big(\sum_s c_{rs}\mu_r^{w\to u}\Big) = \frac{\eta_r}{Z^{u\to v}} e^{-\langle k\rangle}\langle k\rangle^{k_u-1} \propto \eta_r. \tag{5.18}$$

In the literature, a fixed point where messages are independent of their source and destination nodes is called a *factorised* fixed point. This is where the name of the condition comes from. A SBM satisfies the fixed point condition is called a *factorised SBM*.

When the factorised fixed point is the correct marginal distribution, community detection is in principle impossible in the sense that we are not able to make better decision than guessing the labels of nodes according to the prior distribution $\{\eta_r\}$. Community structures are only detectable when the factorised fixed point becomes un-



(a) (b)

Figure 5-1: (a) Partition overlap and the (b) convergence time of BP as a function of the strength of community structure. Here, networks are generated with $N = 10^5$ nodes, $B = 2$ communities, and the average degree $\langle k \rangle = 10$. The vertical dashed line corresponds to the estimated position of critical point $\epsilon^* = 1/\sqrt{\langle k \rangle}$.

stable, and other fixed points which are correlated with the correct structure emerge. To demonstrate the detectability phase-transition phenomenon, we generate networks from the Bernoulli SBM and apply the BP algorithm we introduced in the last section. We restrict to networks with assortative structures by imposing the planted partition constraint

$$p_{rs} = p_{\text{in}}\delta_{rs} + p_{\text{out}}(1 - \delta_{rs}). \tag{5.19}$$

To control the strength of assortativity, we further parameterise the parameters p_{in} and p_{out} as follows

$$p_{\text{in}} = \frac{(1 + (B-1)\epsilon)\langle k \rangle}{N}, \quad p_{\text{out}} = \frac{(1-\epsilon)\langle k \rangle}{N}. \tag{5.20}$$

The assortativity parameter ϵ takes values in the interval [0,1]. If $\epsilon = 0$, the probability of an edge is the same across the entire network, $p = p_{\text{in}} = p_{\text{out}}$, then the resulted networks are random networks with no community structures. Whenever the assortativity parameter $\epsilon > 0$, networks generated from the Bernoulli SBM contain assortative structures, because the probability of an edge inside communities is larger than that between distinct communities.

In Fig. 5-1, we show the results of applying the BP algorithm to synthetic networks with $N = 10^5$ nodes, average degree $\langle k \rangle = 10$, $B = 2$ communities and different strengths of assortativity. We measure the quality of inference by computing the partition overlap

between the correct structure $\boldsymbol{b}^*$ and the inferred partition $\boldsymbol{b}$, which is defined as

$$\text{overlap}(\boldsymbol{b}^*, \boldsymbol{b}) = 1 - d(\boldsymbol{b}^*, \boldsymbol{b}) = \frac{1}{N} \max_{\phi} \sum_{u} \delta_{b_u^*, \phi(b_u)}, \tag{5.21}$$

where $d(\boldsymbol{b}^*, \boldsymbol{b})$ is the partition overlap distanced in equation (3.74) and $\phi(r)$ is a bijection between the group labels of $\boldsymbol{b}$ and $\boldsymbol{b}^*$. The partition overlap for the trivial partition $\boldsymbol{b}^{\text{trivial}}$ which randomly assigns nodes according to the prior distribution $\boldsymbol{\eta}$ is

$$\text{overlap}(\boldsymbol{b}^*, \boldsymbol{b}^{\text{trivial}}) = \frac{1}{N} \sum_{u} \Big(\sum_{b_u^*=0,1} \sum_{b_u^{\text{trivial}}=0,1} \delta_{b_u^*, b_u^{\text{trivial}}} P(b_u^*, b_u^{\text{trivial}}) \Big) = 0.5. \tag{5.22}$$

Although our synthetic networks are fundamentally different from random networks whenever $\epsilon > 0$, there is a non-trivial regime of the assortativity parameter ϵ, where the BP algorithm converges to the factorised fixed point. As can be seen from Fig 5-1(a), the overlap value remains at 0.5 when ϵ are small. When ϵ approaches a critical position ϵ^*, the overlap abruptly jumps to values above 0.5, then gradually grows to 1 as the assortativity parameter ϵ increases. This kind of change is referred to as *phase-transition* and the two regimes of the assortativity parameter are called the *phases* of the model. The value ϵ^* at which the transition happens is called the *critical point*. When $\epsilon < \epsilon^*$, the SBM is said to be in the *undetectable* phase, because even the correct marginal posterior distribution (the prior distribution $\{\eta_r\}$) fails to detect the planted structures.

Generally, the critical point of phase-transition does not permit analytical expression and usually are estimated by using numerical approaches. One rule of thumb for locating the critical point is that the phase-transition is often associated to the *critical slowing down* [154], e.g. the divergence of BP algorithm. However, for factorised SBMs, it is actually possible to obtain analytical estimate of the critical point. This is done by analysing the stability of the factorised fixed point of the BP equations [65], which leads to the detectability condition

$$|p_{\text{in}} - p_{\text{out}}| > B\sqrt{\langle k \rangle} \Rightarrow \epsilon > \frac{1}{\sqrt{k}}. \tag{5.23}$$

The detectability condition means the detectability of community structures in factorised SBMs is dependant on the average degree $\langle k \rangle$, which effectively represents the amount of data related the latent community structure.

Algorithm 5.2: Semi-supervised BP inference for SBM

1. Initialise $\{\mu_r^{u\to v}\}$ for every edge (u, v) in $\mathcal{E}$

2. Compute $\{\mu_r^u\}$ and $\{h_r\}$ as in the normal BP, but for nodes with known correct label, set $\mu_r^u = \delta_{b_u r}$

3. Update $\{\mu_r^{u\to v}\}$, $\{\mu_r^u\}$ and the $\{h_r\}$ as in the normal BP except for nodes with known labelling; Keep $\mu_r^u = \delta_{b_u r}$ unchanged

4. Repeat step 3 until convergence

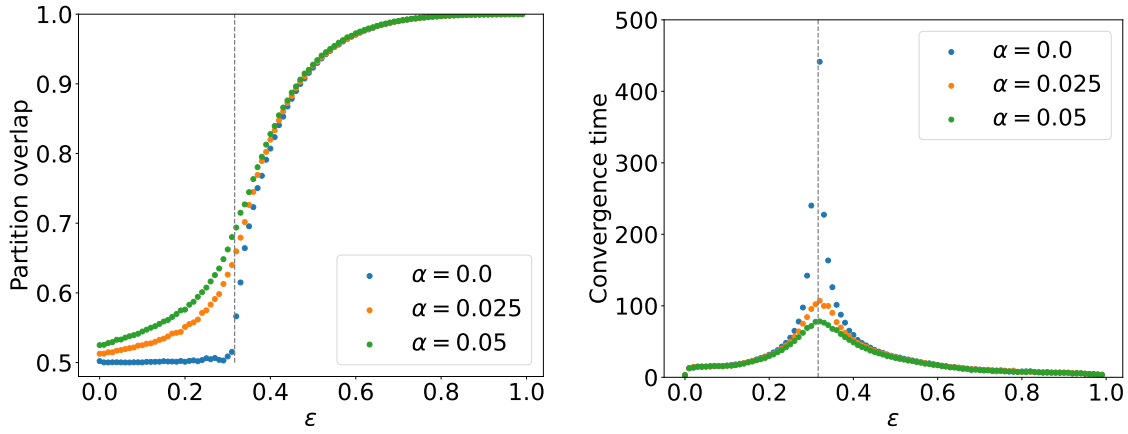
Following the discovery of the detectability phase-transition in factorised SBMs, there are a series of following works looking at the detectability phase-transition under different conditions. For example, in a semi-supervised setting where we have access to the correct labelling of a fraction of nodes, the detectability phase-transition vanishes [70]. It is straightforward to adapt the BP algorithm we defined in Algorithm 5.1 to the semi-supervised learning setting. In each round of iteration, the messages of nodes with known correct labels will be fixed as

$$\mu_r^u = \delta_{b_u^*, r}. \tag{5.24}$$

Only the rest of unknown messages are updated until convergence. The semi-supervised version of BP is summarised in Algorithm 5.2.

In Fig 5-2, we show the results of the BP algorithm under the semi-supervised setting. Suppose there is a fraction $\alpha \in [0, 1]$ of nodes with known correct labelling. When $\alpha = 0$, we reduce back to the unsupervised learning setting. The detectability phase-transition is clearly signified by the abrupt change of the overlap function as well as the divergence of the BP algorithm. In comparison, when $\alpha > 0$, the partition overlap becomes a smooth increasing function of the assortativity parameter ϵ. Since we know the correct labelling of a fraction of nodes, the partition overlap is slightly above 0.5 even when there is not community structures (i.e. when $\epsilon = 0$). Whenever $\epsilon > 0$, the partition overlap are always larger than 0.5 and gradually increases as the assortatitvity parameter ϵ increases. It seems that the BP algorithm can leverage the fixed correct labelling of nodes to propagate the success of detections to nodes with unknown labelling. Moreover, the convergence time of BP does not diverge anymore. Since we cannot identify distinct phases of the model according to the behaviour of the

overlap function and no critical slowing occurs in the BP convergence time, we say the detectability phase-transition disappears.



(a) (b)

Figure 5-2: (a) Partition overlap (b) the convergence time of semi-supervised BP as a function of the strength of community structure. The parameter α controls the fraction of nodes being set to acquire correct marginal message $\{\mu_r^u\}$ as described in the text. Networks are generated with $N = 10^5$ nodes, $B = 2$ communities and average degree $\langle k \rangle = 10$.

5.3 Detectability phase-transition in networks with heterogeneous degree distribution

All of previous works on the detectability of community structures assume the degree distribution is homogenous. Since most of real-world networks have heterogeneous degree distributions, we are interested in how heterogeneity in degree distribution affects the detectability phase-transition. We expect the heterogeneity in degree distribution to enhance the detectability of community structures. This is because heterogeneous degree distribution implies that some nodes will have larger degrees than others. As a result, the detection task should be relatively easy in some (denser) regions than what is implied by the detectability condition in equation (5.23), which was derived in the homogeneous case. We hypothesise that the undetectable phase should shrink, if not disappear completely, when the degree distribution of nodes become heterogeneous.

5.3.1 Belief propagation for DC-SBM

To verify our hypothesis, we consider instead the DC-SBM which allows us to generate networks with customised degree distributions. Nodes are assumed to be sampled from a prior distribution $\{\eta_r\}$, and the probability of generating an observed network $\boldsymbol{A}$

together with its network partition $\boldsymbol{b}$ is

$$P(\boldsymbol{A}, \boldsymbol{b}|\{\lambda_{rs}\}, \boldsymbol{\theta}) = \prod_u \eta_{b_u} \prod_{u<v} e^{-\theta_u\theta_v\lambda_{b_ub_v}} \frac{(\theta_u\theta_v\lambda_{b_ub_v})^{A_{uv}}}{A_{uv}!} \prod_u e^{-\theta_u^2\lambda_{b_ub_u}/2} \frac{(\theta_u^2\lambda_{b_ub_u})^{A_{uu}/2}}{(A_{uu}/2)!}. \tag{5.25}$$

We can fit the model with the BP algorithm for DC-SBM as introduced in [73]. For the ease of presentation, denote the probability of the number of edges between two nodes u and v as a function g,

$$g(\theta_u, \theta_v, \lambda_{rs}, A_{uv}) = e^{-\theta_u\theta_v\lambda_{rs}} \frac{(\theta_u\theta_v\lambda_{rs})^{A_{uv}}}{A_{uv}!}. \tag{5.26}$$

The edge-wise messages for the DC-SBM have the following expression,

$$\mu_r^{u\to v} = \frac{\eta_r}{Z^{u\to v}} e^{-H_r^u} \prod_{w\in\partial u\setminus v} \frac{\sum_{s=1}^B \mu_s^{w\to u} g(\theta_w, \theta_u, \lambda_{rs}, A_{uw})}{\sum_{s=1}^B \mu_s^w g(\theta_w, \theta_u, \lambda_{rs}, 0)}, \tag{5.27}$$

where H_r^u is a defined as

$$H_r^u = -\sum_w^N \ln\Big(\sum_{s=1}^B \mu_s^w g(\theta_w, \theta_u, \lambda_{rs}, 0)\Big). \tag{5.28}$$

Finally, the node-wise messages are

$$\mu_r^u = \frac{\eta_r}{Z^u} e^{-H_r^u} \prod_{w\in\partial u} \frac{\sum_{s=1}^B \mu_s^{w\to u} g(\theta_w, \theta_u, \lambda_{rs}, A_{uw})}{\sum_{s=1}^B \mu_s^w g(\theta_w, \theta_u, \lambda_{rs}, 0)}. \tag{5.29}$$

It is worth noting that there are important nuances in the time complexity of BP between fitting networks with homogeneous and heterogeneous degree distribution. Firstly, consider generating networks with the degree propensity parameter being uniform, $\theta_u = \theta$. Then, the resulting networks are expected to acquire homogenous degree distributions. When applying BP to these networks, the function H_r^u in equation (5.28) is the same for every node, $H_r^u = H_r$. By contrast, when different nodes acquire different degree propensity parameters θ^u, computations for updating $\{H_r^u\}$ increases linearly as the number of unique degree propensity increases. However, notice that for two nodes u and v with the same degree propensity $\theta_u = \theta_v$, we have $H_r^u = H_r^v$. Therefore, we only need to maintain $\{H_r^u\}$ for unique degree propensity parameters in the network[3]. This is generally not an issue in practice, since the number of unique

[3]When we do not know the true degree propensity parameter, we can make inference with BP by adopting the expectation-maximisation procedure. In that case, θ_u will be estimated by the maximum

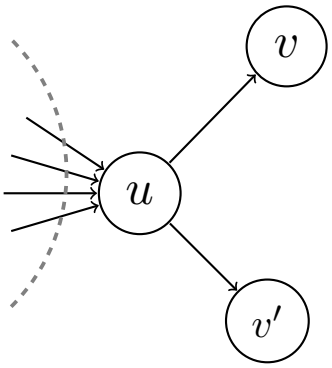


Figure 5-3: When we update the messages sending out from the node u to two of its neighbours v and v', the ratio in equation (5.32) remains unchanged but will be recomputed for all neighbours of u except for v, v', as indicated by the grey dashed curved.

degree propensity is much smaller than the size of network.

The other issue is in computing the messages in Section (5.3.1). Note that the product involves all the neighbours of node u, requiring Bk_u computations, where k_u is the degree of node u. The total time complexity of computing the messages $\{\mu_r^{u\to v}\}$is then

$$\sum_u^N \sum_{v\in\partial u} k_u B = \sum_u^N k_u^2 B. \tag{5.30}$$

For sparse networks with homogeneous degree distribution, the expected degree of all of the nodes are identical, and much smaller than the total number of nodes,

$$\langle k_u \rangle = \langle k \rangle \ll N. \tag{5.31}$$

The order of the number of computations in equation (5.30) is therefore roughly linearly dependent on the size of the network, $\mathcal{O}(N\langle k\rangle^2) \approx \mathcal{O}(N)$. Nevertheless, in the heterogenous case, some nodes might have significantly larger degrees than others. The expression in equation (5.30) means that the time complexity of updating BP messages will increase quadratically as the heterogeneity in degree distribution (so does the number of nodes with large degrees) increases.

It turns out that we can avoid the issue mentioned above by adopting a slightly different way of updating BP messages. In particular, we point out that many computations are unnecessarily repeated when we update messages according to equations (5.27)

likelihood estimator, which will be dependent on the degree of node u only. Therefore, the BP algorithm remains scalable as long as the number of distinct degrees is much smaller than the size of the network, which generally holds in practice.

and (5.29). According to the definition of messages in equation (5.27), for messages sending out from node u to two of its neighbours v and v', the following ratio remains unchanged but will be computed in both $\mu_r^{u\to v}$ and $\mu_r^{u\to v'}$,

$$\frac{\sum_{s=1}^{B}\mu_s^{w\to u}g(\theta_w,\theta_u,\lambda_{rs},A_{uw})}{\sum_{s=1}^{B}\mu_s^{w}g(\theta_w,\theta_u,\lambda_{rs},0)},\quad w\in\partial u\setminus v,v'. \tag{5.32}$$

As shown in Figure 5-3, it is easy to see that the amount of unnecessary computations increases with the degree of node u. This increase in the unnecessary computations is the main cause of the increase in the time complexity in networks with heterogeneous degree distribution. With this observation in mind, we can precompute the product in equation (5.29)

$$\mathcal{I}_r^u := \prod_{w\in\partial u}\frac{\sum_{s=1}^{B}\mu_s^{w\to u}g(\theta_w,\theta_u,\lambda_{rs},A_{uw})}{\sum_{s=1}^{B}\mu_s^{w}g(\theta_w,\theta_u,\lambda_{rs},0)} \tag{5.33}$$

for every node u and store $\{\mathcal{I}_r^u\}$ in a table. Then, we can perform exactly the same update of the message $\mu_r^{u\to v}$ according to (5.27) but with the following expression,

$$\mu_r^{u\to v}=\frac{\eta_r}{Z^{u\to v}}e^{-H_r}\times\mathcal{I}_r^u\times\left(\frac{\sum_{s=1}^{B}\mu_s^{w\to u}g(\theta_w,\theta_u,\lambda_{rs},A_{uw})}{\sum_{s=1}^{B}\mu_s^{w}g(\theta_w,\theta_u,\lambda_{rs},0)}\right)^{-1}, \tag{5.34}$$

and the node-wise messages are

$$\mu_r^u=\frac{\eta_r}{Z^u}e^{-H_r}\mathcal{I}_r^u. \tag{5.35}$$

Updating each message $\mu_r^{u\to v}$ in this improved way only requires $\mathcal{O}(1)$ computations (assume the number of communities $B\ll N$), making the total time complexity at the scale of $\mathcal{O}(E)$. Although this modified updating scheme induces extra computations for maintaining the table of $\{\mathcal{I}_r^u\}$, we find that the implementation of BP is more efficient when the modification is in place, even when the degree distribution is homogeneous[4]. We summarise the scheme of updating the BP for DC-SBM in Algorithm 5.3. In Section 5.3.3, we will make use of this BP algorithm to investigate the effect of heterogeneity in degree distribution on the detectability phase-transition in DC-SBM.

[4]We have compared the running time of BP in networks with heterogeneous degree distributions, using the original and modified updating schemes. Details of the comparison can be found in Appendix C.1.

Algorithm 5.3: BP inference for DC-SBM

1. Initialise $\{\mu_r^{u\to v}\}$ for every edge (u,v) in $\mathcal{E}$

2. Compute $\{\mu_r^u\}$ according to equation (5.16), and $\{H_r\}$ according to equation (5.28) as well as the $\{\mathcal{I}_r^u\}$ defined in equation (5.33)

3. For every edge $(u,v)\in\mathcal{E}$, update $\mu_r^{u\to v}$ according to equation (5.34);
   Update the value of $\mathcal{I}_r^v$

4. Update the message μ_r^v according to equation (5.35);
   Update the value of $\mathcal{I}_r^w$ for all w in the neighbouring set of node v and $\{H_r\}$

5. Repeat step 3-4 until convergence

5.3.2 Generating heterogeneous degree propensity

We generate networks from the DC-SBM with the following parametrisation. Remind that the DC-SBM assumes the number of edges between any pair of nodes is a Poisson variable

$$A_{uv}\sim \mathrm{Poi}(\theta_u\theta_v\lambda_{b_ub_v}), \tag{5.36}$$

where $\{\theta_u\}$ is the degree propensity parameter for each node and $\{\lambda_{rs}\}$ are the expected number of connections between group r and s when the following group-wise normalisation is imposed,

$$\hat{\theta}_r=\sum_u \theta_u\delta_{rb_u}=1,\ \ \forall r\in\{1,2,..,B\}. \tag{5.37}$$

We will set a hard constraint on the size of communities

$$n_r=\frac{N}{B}\in\mathbb{Z},\quad \forall r\in\{1,2,..,B\}, \tag{5.38}$$

which allows us to impose the same degree propensity across different communities

$$\theta_u=\theta_{u+nN/B},\ \ \forall u\in\{1,2,..,N/B\},\ \ \forall n\in\{1,2,..,B-1\}. \tag{5.39}$$

As a result, the prior distribution of assigning a node u to group r is the uniform distribuion

$$\eta_{b_u}=\frac{n_r}{N}=\frac{1}{B}. \tag{5.40}$$

Furthermore, we enforce the planted partition constraint

$$\lambda_{rs} = \lambda_{\text{in}}\delta_{rs} + \lambda_{\text{out}}(1 - \delta_{rs}), \tag{5.41}$$

and parameterise λ_{in} and λ_{out} as follows

$$\begin{aligned} \lambda_{\text{in}} &= \left(\frac{N}{B}\right)^2 p_{\text{in}} = \frac{N}{B}(1 + \epsilon(B-1))\langle k\rangle, \\ \lambda_{\text{out}} &= \left(\frac{N}{B}\right)^2 p_{\text{out}} = \frac{N}{B}(1 - \epsilon)\langle k\rangle, \end{aligned} \tag{5.42}$$

where the ϵ parameter controls the strength of the assortativity just as it did in the equation (5.20).

To obtain networks with heterogeneous degree distribution, we need to manipulate the degree propensity parameter $\{\theta_u\}$. We can examine the effect of the value θ_u on the expected degree of node u by noticing that

$$\langle k_u\rangle = \sum_{v}^{N} \theta_u\theta_v\lambda_{b_u b_v} = \theta_u \sum_{r}^{B} \hat{\theta}_r\lambda_{rb_u}. \tag{5.43}$$

Substituting λ_{in}, λ_{out} into the expression of the expected degree in equation (5.43) and making use of the normalisation we imposed in equation (5.37), we get

$$\langle k_u\rangle = \theta_u \frac{N}{B}\langle k\rangle. \tag{5.44}$$

This expression of $\langle k_u\rangle$ implies that our model also satisfies the factorised condition in equation (5.17), because the expected degree of nodes in any communities r is

$$\sum_{u}\langle k_u\rangle\delta_{rb_u}/n_r = \frac{N}{B}\langle k\rangle(n_r)^{-1} = \langle k\rangle, \tag{5.45}$$

no matter what are the values of $\{\theta_u\}$. When $\theta_u = \theta = (N/B)^{-1}$, we reduce to the uniform case where each node has the same average degree as the global average degree $\langle k\rangle$. As θ_u becomes heterogenous, since the expected degree $\langle k_u\rangle$ is proportional to θ_u, the degree distribution in the network will also become heterogeneous. The extent to which $\langle k_u\rangle$ is scaled up or down compared to the global average degree $\langle k\rangle$ depends on how much the degree propensity parameter θ_u differs from the uniform choice $\theta_u = \theta = (N/B)^{-1}$.

Since most of real-world networks possess power-law degree distributions [71], a natural

choice for the degree propensity parameter is to use samples drawn from the *Zipf's* distribution, which is defined as follows

$$f_X(x) = \begin{cases} x^{-\zeta} / \sum_{x_{\min}}^{\infty} x^{-\zeta}, & \text{if } x_{\min} \leq x;\ x \in \mathbb{Z} \\ 0, & \text{otherwise} \end{cases} . \tag{5.46}$$

Without the loss of generality, we can set $x_{\min} = 1$. The value of the ζ parameter controls the strength of the heterogeneity of the distribution. We will restrict ourselves to the choice of ζ in $[2, \infty]$, since the mean of the Zipf's distribution is not defined when $\zeta \leq 2$. We are particularly interested in the case when ζ takes a value in the interval $[2, 3]$, because when ζ in this interval the Zipf's distribution has a finite mean but diverging variance, which aligns best with most of empirical networks with locally dense but globally sparse connections. To satisfy the normalisation constraint in equation (5.37), we draw samples $\{x_u\} \in \mathbb{Z}^{N/B}$ from the Zipf's distribution and then set $\{\theta_u\}$ to the following normalised values,

$$\theta_{u+nN/B} = \theta_u = \frac{x_u}{\sum_{v=1}^{N/B} x_v}, \quad \forall u \in \{1, 2, .., N/B\}, \quad \forall n \in \{1, 2, .., B-1\}. \tag{5.47}$$

There are two caveats on using samples from the Zip's distribution for the degree propensity parameter $\{\theta_u\}$. Firstly, because we like to learn the effect of heterogeneous degree distribution on the detectability of community structure, it is important to control other perspectives which might affect the detectability. When generate networks with $\{\theta_u\}$ obtained via equation (5.47), there is a risk that the average degree of the network is out of control, especially in the case where the heterogeneity is strong (ζ close to 2). This is because the values of $\{\theta_u\}$ could be extremely small when there are some dominantly large samples in $\{x_u\}$. As a result, there is a great chance to observe a non-negligible amount of isolated nodes, making the average degree in the connected component much larger than the value we expect. To make sure our experiments can properly reflect the effect of heterogeneous degree distribution rather the network density, it is therefore important to monitor the size of the connected component and make sure the network density is under control.

Secondly, as pointed out in [155], the number of short cycles could diverge when the degree distribution follows a power-law distribution with the exponent of the power-law $\zeta \in$ being in $[2, 3]$. The existence of short cycles disqualifies the use of the BP algorithm, because short-length loops violates the conditional dependance assumption that is used in deriving BP equations. Since we want to take advantage of the efficiency

of BP, instead of the Zipf's distribution, we turn to the truncated Zipf's distribution

$$f_X(x) = \begin{cases} x^{-\zeta} / \sum_{x=1}^{x_{\max}} x^{-\zeta}, & \text{if } 1 \leq x \leq x_{\max}; x \in \mathbb{Z} \\ 0, & \text{otherwise} \end{cases} . \tag{5.48}$$

Compared to the Zip's distribution, the truncated Zip's distribution introduce a cut-off $x_{\max}$ for the values of samples $\{x_u\}$. Note that both the parameter ζ and $x_{\max}$ have effect on the heterogeneity of the distribution. If we set $x_{\max} = 1$, then we reduce back to the homogeneous case, regardless of the value of ζ. For a fixed $x_{\max}$ which is larger than 1 and ζ in $[2, \infty]$, the distribution becomes more heterogeneous as ζ decreases. When the value ζ is fixed, increasing the value of $x_{\max}$ also increases the heterogeneity in samples of $\{x_u\}$. We want to select a cut-off value for $x_{\max}$ such that the resulted networks are also locally-treelike and therefore BP is still a valid inference tool. To this end, similar to the arguments we used in Section 5.1, we need the probability of an edge between any pair of nodes to be at the scale of $\mathcal{O}(1/N)$. When such condition holds, in our DC-SBM, the probability of having edges between nodes i and j is approximately

$$1 - P(A_{ij} = 0 | \boldsymbol{\lambda}, \boldsymbol{\theta}, \boldsymbol{b}) = 1 - e^{-\theta_i \theta_j \lambda_{b_i b_j}} \approx \theta_i \theta_j \lambda_{b_i b_j}. \tag{5.49}$$

and hence the following condition needs to be satisfied

$$\theta_i \theta_j \lambda_{b_i b_j} \leq \theta_{\max}^2 \lambda_{b_i b_j} = \theta_{\max}^2 \mathcal{O}(N), \tag{5.50}$$

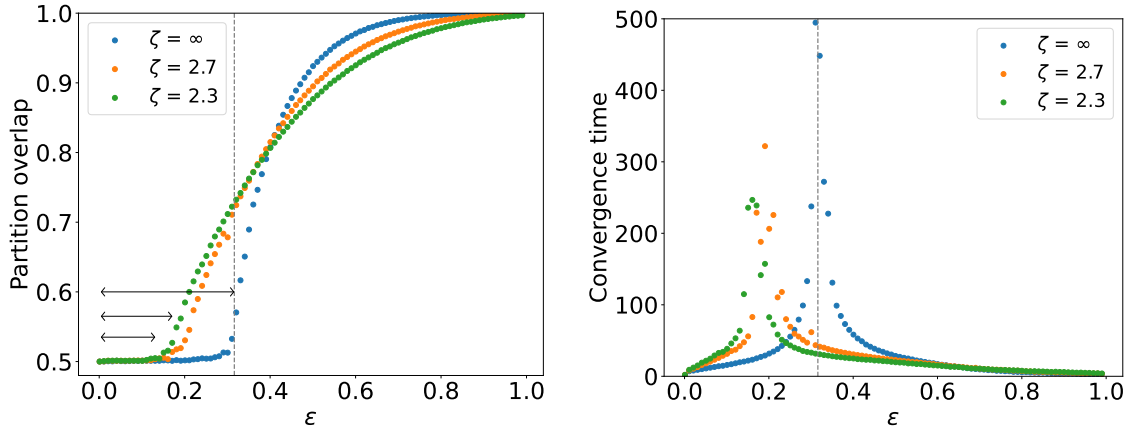
where the equation sign makes use of our parameterisation in equation (5.42). Therefore, we need

$$\theta_{\max}^2 \mathcal{O}(N) = \mathcal{O}(1/N) \Rightarrow \theta_{\max} = \mathcal{O}(1/N) \tag{5.51}$$

Notice that for samples $\{x_u\}$ which are drawn from the truncated Zipf's distribution, we can obtain an upper bound for the largest degree propensity parameter

$$\theta_{\max} = \frac{\max_u \{x_u\}}{\sum_u x_u} \leq \frac{x_{\max}}{n_r - 1 + x_{\max}} \leq \frac{x_{\max}}{n_r}. \tag{5.52}$$

With such upper bound of $\langle \theta_{\max} \rangle$ in mind, if we set the cut-off value $x_{\max}$ at the scale $\mathcal{O}(1)$, then the expected largest degree $\theta_{\max} = \mathcal{O}(1/N)$ holds and can use BP to make inference.



(a) (b)

Figure 5-4: Results obtained with the BP algorithm in networks generated from the DC-SBM with the degree propensity parameter $\{\theta_u\}$ being sampled from the truncated Zipf's distribution. The cutoff value $x_{\max}$ of the truncated Zipf's distribution is set to be 50. We choose $B = 2$, $N = 10^5$ and $\langle k \rangle = 10$. (a) Partition overlap and (b) the convergence time of BP as a function of the strength of community structure. The vertical dashed line indicates the estimate of the critical point given by the detectability condition in equation (5.23).

5.3.3 Numerical results

We applied the BP algorithm defined in Section 5.3.1 to networks generated from DC-SBM with degree propensity parameters being proportional to samples of the truncated Zipf's distribution. Our results show that, when the network density (global average degree) is fixed, increase in heterogeneity of degree distribution enhances the detectability of communities structures. In Fig. 5-4, we show the partition overlap achieved by BP for different values of ζ. When $\zeta = \infty$, the truncated Zipf's distribution is equivalent to a uniform distribution, degree propensity parameter $\{\theta_u\}$ is uniform and the detectability phase-transition is clearly signified by the sharp change of the overlap function as well as the divergence of BP algorithm. When ζ takes values in $[2, 3]$, it seems that the detectability phase-transition still occurs, but the critical position shifts toward the region where the assortative structure is weaker. The extent of the shift of the critical point is larger when there are more variation in the degree propensity parameter.

Remind that in the homogeneous case, according to the detectability condition in the equation (5.23), the critical point also shifts if we increase the average degree in the networks. In Fig 5-5, we show how the critical point of the detectability phase-transition changes as we increase the average degree from 10 to 15 and 20. Indeed, as the average degree $\langle k \rangle$ increases, the critical point moves to the left and the area

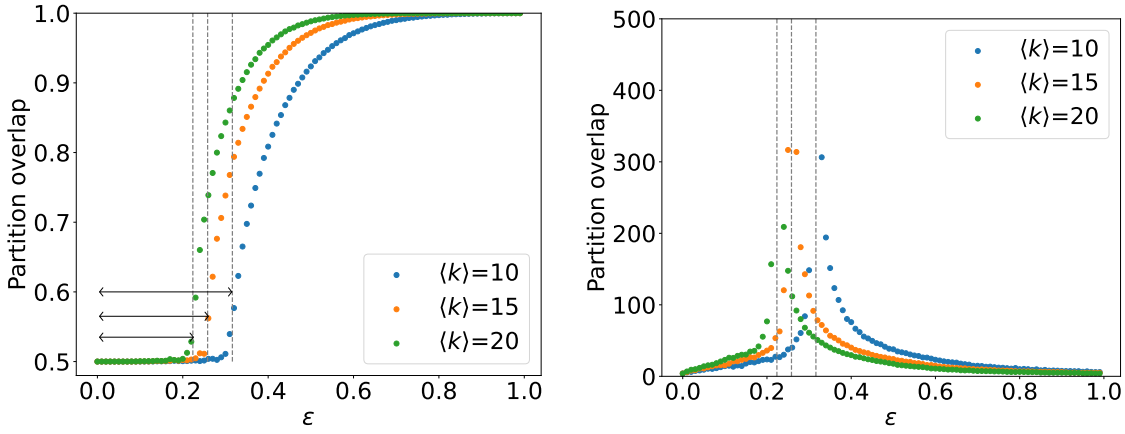


Figure 5-5: Results obtained with the BP algorithm in networks generated from the DC-SBM with uniform propensity parameter $\theta_u = B/N$. Networks are generated with $B = 2$, $N = 10^5$ and average degree $\langle k \rangle$ being 10, 15, and 20. (a) Partition overlap the (b) convergence time of BP as a function of the strength of community structure. The vertical dashed line indicates the estimate of the critical point given by the detectability condition in equation (5.23).

of undetectable phases shrinks. We emphasis that, for results in the Fig. 5-4 where we examine the effect of heterogeneous degree distribution, the global average degree $\langle k \rangle = 10$ is fixed. However, we also observe the reduce in the undetectable phase of model as the heterogeneity in degree distribution increases. Therefore, we conclude that the global average degree $\langle k \rangle = 2E/N$ alone does not determine the detectability phase-transition when nodes acquire heterogeneous degree distribution: Heterogeneity in degree distribution can enhance the detectability of the community structures.

Nevertheless, as the old saying goes, "there is no such thing as free lunch". Notice that in Fig. 5-4, the curves for partition overlap cross over at some point. That means, for ϵ larger than certain value (say 0.5), the partition overlap achieved by BP is lower when the extent of heterogeneity in degree distribution is stronger. To further understand this trade-off between detectability of community structure and accuracy, in Fig 5-6, we plot the histogram of 5×10^4 samples from the truncated Zipf's distribution with different values of ζ. Because the total number of samples is fixed, the area under the histogram remains the same when we change the values of ζ. When the value of ζ decreases, there are more samples x become larger than the rest, making more nodes acquire large degree propensity parameter θ_u. In other words, as the degree distribution becomes more heterogeneous, the fixed amount of input is reorganised among nodes, forcing some nodes to receive less edges than others. The degree propensity parameter θ_u acts like a filter: It magnifies the signals for some nodes at the price of shrinking the

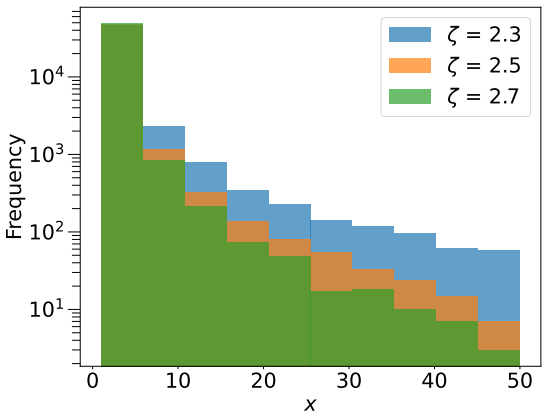


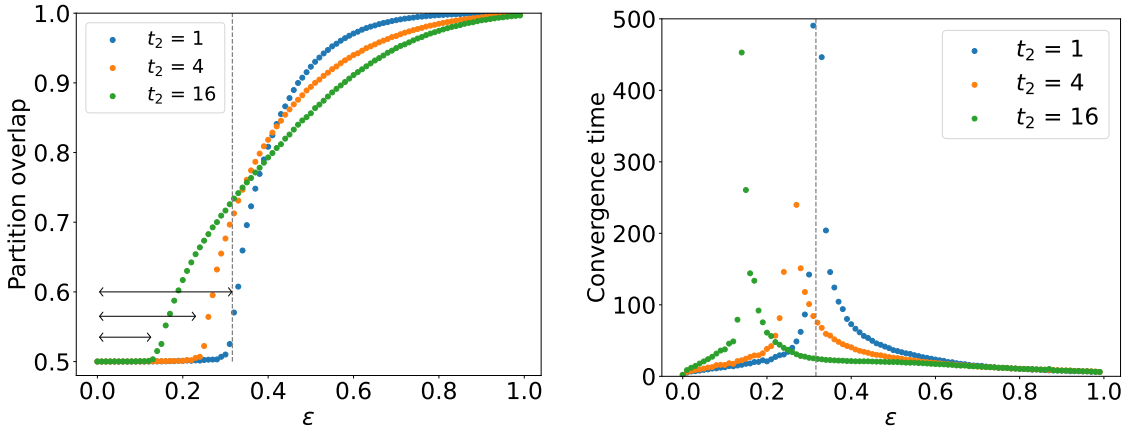
Figure 5-6: Histogram of 5×10^4 samples x drawn from the truncated Zipf's distribution with varying values of the shape parameter ζ and a cut-off at $x_{\max} = 50$.

signals for others. As a result, when the global average degree $\langle k \rangle$ is fixed, detecting the correct labelling of nodes with large degrees is easier than what is implied by the detectability condition, which is derived for networks with homogeneous degree distribution. The "lunch" to pay is that the correct labelling of nodes with small degree propensity parameter are more difficult to detect, even when the strength of assortative structure described by ϵ is high. The filtering effect of the degree propensity parameter wastefully assigns too much observations to large-degree-propensity nodes, whose correct labelling could have been detected even when a portion of their received edges were passed over to other nodes with low-degree-propensity.

According to our explanation above, we expect that the increase in detectability of community structures does not require the degree propensity parameter to follow the power-law distribution. As a way to verify this postulation, we consider an alternative heterogeneous degree distribution, where $\{\theta_u\}$ are computed by normalising samples of a bimodal distribution

$$P(t_u = t_1) = \tau, \quad P(t_u = t_2) = 1 - \tau, \quad \tau \in [0, 1] \tag{5.53}$$

where t_1 and t_2 are two positive values. Without loss of generality, suppose $t_1 \leq t_2$. Choosing degree propensity $\{\theta_u\}$ according to this bimodal distribution means nodes can have either low-degree propensity with probability τ, or high-degree propensity with probability $1 - \tau$. The value of difference the $t_2 - t_1$ controls the extent to which high degree nodes exceed the average degree $\langle k \rangle$. In light of the results we have obtained with the truncated Zipf's distribution, we expect the detectability of community structure to increase as the value of difference $t_2 - t_1$ increases. With the



(a) (b)

Figure 5-7: Results obtained with the BP algorithm in networks generated from the DC-SBM with $B = 2$, $N = 10^5$, $\langle k \rangle = 10$. The degree propensity parameter $\{\theta_u\}$ are sampled from the bimodal distribution with the parameter t_2 being 1, 4, and 16. When $t_2 = 1$, the degree propensity parameter is uniform. (a) Partition overlap and (b) convergence time of BP as a function of the strength of community structure

group-wise normalisation constraint, there are only two possible values of the degree propensity parameter

$$\theta_{\min} = \frac{B}{N}\frac{t_1}{(\tau t_1 + (1-\tau)t_2)}, \quad \theta_{\max} = \frac{B}{N}\frac{t_2}{(\tau t_1 + (1-\tau)t_2)}. \tag{5.54}$$

We set $t_1 = 1$, then the absolute value of t_2 alone controls the extent of heterogeneity.

Because we want to use the BP algorithm for inference, the value of t_2 cannot be arbitrarily large, which could make the BP an invalid tool to use. We should examine the largest expected degree to get what are the valid choices of t_2. Under the parameterisation in equation (5.53) and (5.54), the expected degree of nodes can take one of the following two values

$$\begin{aligned} \langle k_{\min} \rangle &= \theta_{\min}\frac{N}{B}\langle k \rangle = \frac{1}{(\tau + (1-\tau)t_2)}\langle k \rangle, \\ \langle k_{\max} \rangle &= \theta_{\max}\frac{N}{B}\langle k \rangle = \frac{t_2}{(\tau + (1-\tau)t_2)}\langle k \rangle. \end{aligned} \tag{5.55}$$

Notice that, when τ is fixed, $\langle k_{\max} \rangle$ will saturate as t_2 increases. Therefore, we further parameterise τ as follows

$$\tau = 1 - t_2^{-1}, \tag{5.56}$$

such that $\langle k_{\max} \rangle$ is always a increasing function of t_2. The expected largest degree in

networks generated with the bimodal degree distribution becomes

$$\langle k_{\max} \rangle = \theta_u \frac{N}{B} \langle k \rangle = \frac{t_2}{2 - t_2^{-1}} \langle k \rangle. \tag{5.57}$$

The last equation indicates that as long as we choose t_2 at the order of $\mathcal{O}(1)$, networks generated with the bimodal degree propensity distribution will have vanishingly small number of short-length loops, which justifies the use of the BP algorithm. We present in Fig 5-7 the results of applying BP to networks generated with the degree propensity being sampled from the bimodal distribution. Similar to what we have seen before when the truncated Zipf's distribution is in place, larger values of t_2 causes more significant reduce in the area of the undetectable phase. These results corroborate our hypothesis that variation in degree distribution affects the detectability phase-transition, causing the undetectable phase of the model to shrink compared to that in networks with homogenous degree distribution.

5.4 Concluding remarks

In this chapter, we investigated the effect of heterogeneous degree distribution on the detectability of community structures. To this end, we apply the BP algorithm to networks generated from the uniform PP model with customised degree distributions. When the global average degree is fixed, we show that enforcing heterogeneous degree distribution makes the detectability of community structures increases, manifested by the reduced area of the undetectable phase of the model. This is because nodes with large degree propensity receive more edges than the rest of the network, forming some relatively dense regions where the detection task is less challenging. The price to pay for the increased detectability is that there are more nodes with low degree whose correct labelling are difficult to infer. Overall, our numerical results provide further confirmation for the existing theoretical analysis in [72] which states that the global average degree alone does not determine the detectability of community structure in real-world networks, where degree distributions are often heterogeneous.

Our results and the adapted BP algorithm open the door to investigation about detectability phase-transition in more general settings. For example, one possible direction to pursue is to extend our experiments by considering networks with more than two communities. This is interesting because the detectability phase-transition phenomenon in networks with homogeneous degree distribution is found to change significantly as we move from two communities to more than two communities [65, 70]. Moreover, another interesting question to ask is that whether the detectability phase-

transition will vanish just as it does in the semi-supervised learning setting, when the heterogeneity in degree distribution becomes asymmetric across distinct communities. The asymmetric degree distribution is a more realistic assumption. For example, consider the situation where the truncated Zip'f distribution is used for sampling the degree propensity parameters but with different heterogeneity parameter ζ_r, or different cut-off values $x^r_{\max}$ for different communities $r \in \{1, 2, .., B\}$. With this parameterisation, the symmetry of communities breaks despite the fact that their average degree could be the same. When different communities acquire different average degree but with homogeneous degree distributions, the detectability phase-transition is known to completely disappear [68, 69]. It is not clear yet how the detectability phase-transition changes when the asymmetry of communities arises in the degree distribution. We leave this question to future work.

Chapter 6

Conclusions and future work

We close this thesis by summarising contributions and outlining some potential revenues for future work. In terms of theoretical contributions, we hope this thesis can provide new insights about the modularity-based community detection methods as well as their relation to the statistical inference method that relies on generative network models. From the practical perspective, this thesis contributes to the field of network inference by providing a novel method to extract statistically significant assortative structures in network data. Our method has the advantages of not overfitting and not underfitting data, as well as being able to simultaneously resolve assortative structures at multiple resolutions. In networks where the dominant pattern is assortative and the sizes of networks are large, our proposed assortative-constrained models can achieve better quality of fit than their general counterparts.

In Chapter 3, we clarify on the equivalence between the maximum likelihood inference and the popular modularity maximisation approach for detecting assortative structures. Modularity-based methods dominate in the early stage of network analysis, as evidenced by a large amount of citations and a series of extensions that are built on the original modularity measure. For instance, there are modularity measures for temporal networks and multilayer networks [79], spatial networks [156], hypergraph [157, 158], etc. Given the clarification we have provided here on the relation between modularity maximisation and the maximum likelihood inference, it is worth re-examining the results and conclusions which rely on the use of the modularity maximisation approach, especially from the perspective of overfitting data and the bias toward uniform assortative structures.

As an example, it is useful to generalised our PP models to multilayer networks. In

contrast to the ordinary graphs (monolayer networks) that we have been discussed throughout the thesis, multilayer networks are collections of interrelated ordinary networks. Multilayer networks are commonly used to describe temporal interactions or different kinds of interactions occur in the same systems. To detect assortative structures in multilayer networks, researchers have come up with the multilayer version of modularity measure [77] and SBMs [159,160]. In light of the equivalence result between the monolayer version of modularity maximisation and the inference of the planted partition model given by Newman in [57], a recent work showed the equivalence between the multilayer version of the two methods [109]. However, as we have discussed in Chapter 3, Newman's equivalence result is tenuous and our arguments can be easily applied to challenge the equivalence result in multilayer networks as well. Instead of modifying the modularity measure, we believe that our proposed PP models will serve as a more reliable base model for developing community detection method in multilayer networks.

In Chapter 4, we found that Bayesian inference with the degree-corrected stochastic block models (DC-SBM) systematically underfits when we apply it to empirical networks. In comparison, using Nested DC-SBM can systematically lift the resolution limit of DC-SBM. Our PP models also do not have the underfitting problem and are able to identify arbitrarily large number of assortative communities, as long as they exist in data. By conducting model selection, we find that assortative-constrained variants are the best fitting model in only a minority of datasets in our network corpus. Most of the time, general SBMs achieve the best quality of fit according to the MDL principle, indicating that assortativity is often too simplistic to account for the structures in empirical networks. We are well aware that our experiment only gives us the best model in a relative sense, and it will not be a surprise to see other models that we have not considered in this thesis to achieve better fit, especially when assumptions of SBMs are not compatible with the true generating processes. It is easy to envision that other variants of SBMs to be proposed to better match data with special characteristic. For example, we have seen that PP models outperform general models in a set of infrastructure networks where edges between nodes are likely to be subjective to spatial constraints. One potential direction to pursue is to develop spatial variant of SBMs with spatial constraints being appleid to the placement of edges between nodes.

Understanding the relation between community structures and network domains will facilitate the design and use of the appropriate community detection methods in practice. Unfortunately, the network corpus we have constructed in this thesis does not allow us to make any conclusions related to the source of networks, because our corpus

has a skew distribution of network domains. Hence, in order to understand what are the most typical structures for each network domain, we need to expand our analysis with a careful selection of data such that the source of networks is sufficiently diverse.

In addition, it could be helpful to conduct a comparative study with a larger collection of community detection algorithms. We have only compared our proposed PP models to the modularity maximisation as well as the generalised modularity variant, mainly due to the fact that modularity has a close relation to our uniform PP model and modularity-based methods are widely used. However, there are a plethora of other methods for extracting assortative community structures in networks, such as the modularity density [56], spectral algorithms [161], and methods that consider dynamical processes taking place on networks, e.g. the Markov stability [162] and Infomap [163]. A comprehensive comparison will deepen our understanding of the strengths and limits of our proposed PP models in practice.

In Chapter 5, we study the detectability of community structures when the degree distribution is heterogeneous. We observe that the area of undetectable phase of the uniform PP model shrinks when the heterogeneity in degree distribution increases. To obtain a complete picture of the detectability phase-transition when degree distribution is heterogeneous, it is worth extending our experiment in Chapter 5 to networks with more than two communities. This is because the number of communities has an effect on the types of detectability phase-transition phenomenon [65,68]. Moreover, our analysis in Chapter 5 is restricted to the situation where different communities acquire identical degree propensity parameters. A more realistic setup is to assume each community has its own way of adjusting the degree propensity of nodes. This asymmetric degree propensity setting is expected to affect the detectability phase-transition as well, since average degrees in the dense regions of different communities will be different, even though the average degree of each community could remain identical. Because the vanishing of detectability phase-transition has been observed in networks where different communities acquire different average degrees [68, 69], it is interesting to examine how the detectability phase-transition changes as the symmetry of heterogeneity in degree distribution breaks.

Bibliography

[1] Mark Newman. *Networks*. Oxford university press, 2018.

[2] Alain Barrat, Marc Barthelemy, and Alessandro Vespignani. *Dynamical processes on complex networks*. Cambridge university press, 2008.

[3] Q. K. Telesford, S. L. Simpson, J. H. Burdette, S. Hayasaka, and P. J. Laurienti. The brain as a complex system: Using network science as a tool for understanding the brain. *Brain Connectivity*, 1(4):295–308, 2011.

[4] Mert Korkali, Jason G Veneman, Brian F Tivnan, James P Bagrow, and Paul DH Hines. Reducing cascading failure risk by increasing infrastructure network interdependence. *Scientific reports*, 7(1):1–13, 2017.

[5] J Kunegis. American revolution network dataset, 2016.

[6] Aaron Clauset, Ellen Tucker, and Matthias Sainz. The colorado index of complex networks. *Retrieved July*, 20(2018):22, 2016.

[7] TP Peixoto. The netzschleuder network catalogue and repository, 2020.

[8] Bing-Bing Xiang, Zhong-Kui Bao, Chuang Ma, Xingyi Zhang, Han-Shuang Chen, and Hai-Feng Zhang. A unified method of detecting core-periphery structure and community structure in networks. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 28(1):013122, 2018.

[9] Sadamori Kojaku and Naoki Masuda. Core-periphery structure requires something else in the network. *New Journal of physics*, 20(4):043012, 2018.

[10] Tzu-Chi Yen and Daniel B. Larremore. Community detection in bipartite networks with stochastic block models. *Physical Review E*, 102(3):032309–, 09 2020.

[11] Michelle Girvan and Mark EJ Newman. Community structure in social and biological networks. *Proceedings of the national academy of sciences*, 99(12):7821–7826, 2002.

[12] P Chen and Sidney Redner. Community structure of the physical review citation network. *Journal of Informetrics*, 4(3):278–290, 2010.

[13] Erzsébet Ravasz, Anna Lisa Somera, Dale A Mongru, Zoltán N Oltvai, and A-L Barabási. Hierarchical organization of modularity in metabolic networks. *science*, 297(5586):1551–1555, 2002.

[14] Anna CF Lewis, Nick S Jones, Mason A Porter, and Charlotte M Deane. The function of communities in protein interaction networks at multiple scales. *BMC systems biology*, 4(1):1–14, 2010.

[15] Sarvenaz Choobdar, Mehmet E Ahsen, Jake Crawford, Mattia Tomasoni, Tao Fang, David Lamparter, Junyuan Lin, Benjamin Hescott, Xiaozhe Hu, Johnathan Mercer, et al. Assessment of network module identification across complex diseases. *Nature methods*, 16(9):843–852, 2019.

[16] Santo Fortunato. Community detection in graphs. *Physics reports*, 486(3-5):75–174, 2010.

[17] Mason A Porter, Jukka-Pekka Onnela, Peter J Mucha, et al. Communities in networks. *Notices of the AMS*, 56(9):1082–1097, 2009.

[18] Marko Jusup, Petter Holme, Kiyoshi Kanazawa, Misako Takayasu, Ivan Romić, Zhen Wang, Sunčana Geček, Tomislav Lipić, Boris Podobnik, Lin Wang, et al. Social physics. *Physics Reports*, 948:1–148, 2022.

[19] Santo Fortunato. Community detection in graphs, 2 2010.

[20] Santo Fortunato and Darko Hric. Community detection in networks: A user guide. *Physics reports*, 659:1–44, 2016.

[21] Michael T Schaub, Jean-Charles Delvenne, Martin Rosvall, and Renaud Lambiotte. The many facets of community detection in complex networks. *Applied network science*, 2(1):4, 2017.

[22] Clement Lee and Darren J Wilkinson. A review of stochastic block models and extensions for graph clustering. *Applied Network Science*, 4(1):1–50, 2019.

[23] Núria Rovira-Asenjo, Tània Gumí, Marta Sales-Pardo, and Roger Guimera. Predicting future conflict between team-members with parameter-free models of social networks. *Scientific reports*, 3(1):1–6, 2013.

[24] Clement Lee and Darren J Wilkinson. A social network analysis of articles on social network analysis. *arXiv preprint arXiv:1810.09781*, 2018.

[25] Elizabeth E Bruch and MEJ Newman. Structure of online dating markets in us cities. *Sociological Science*, 6:219–234, 2019.

[26] Joel Dyer and Blas Kolic. Public risk perception and emotion on twitter during the covid-19 pandemic. *Applied Network Science*, 5(1):1–32, 2020.

[27] Joshua Faskowitz, Xiaoran Yan, Xi-Nian Zuo, and Olaf Sporns. Weighted stochastic block models of the human connectome across the life span. *Scientific reports*, 8(1):1–16, 2018.

[28] Javier O Garcia, Arian Ashourvan, Sarah Muldoon, Jean M Vettel, and Danielle S Bassett. Applications of community detection techniques to brain graphs: Algorithmic considerations and implications for neural function. *Proceedings of the IEEE*, 106(5):846–867, 2018.

[29] Richard F Betzel, Maxwell A Bertolero, and Danielle S Bassett. Non-assortative community structure in resting and task-evoked functional brain networks. *Biorxiv*, page 355016, 2018.

[30] Leonardo Morelli, Valentina Giansanti, and Davide Cittaro. Nested stochastic block models applied to the analysis of single cell data. *BMC bioinformatics*, 22(1):1–19, 2021.

[31] Mel MacMahon and Diego Garlaschelli. Community detection for correlation matrices. *arXiv preprint arXiv:1311.1924*, 2013.

[32] Paolo Barucca and Fabrizio Lillo. Disentangling bipartite and core-periphery structure in financial networks. *Chaos, Solitons & Fractals*, 88:244–253, 2016.

[33] Cazabet Remy, Baccour Rym, and Latapy Matthieu. Tracking bitcoin users activity using community detection on a network of weak signals. In *International conference on complex networks and their applications*, pages 166–177. Springer, 2017.

[34] Andrew Elliott, Mihai Cucuringu, Milton Martinez Luaces, Paul Reidy, and Gesine Reinert. Anomaly detection in networks with application to financial transaction networks. *arXiv preprint arXiv:1901.00402*, 2019.

[35] Longfeng Zhao, Chao Wang, Gang-Jin Wang, H Eugene Stanley, and Lin Chen. Community detection and portfolio optimization. *arXiv preprint arXiv:2112.13383*, 2021.

[36] Brian Wilson Kernighan. *Some graph partitioning problems related to program segmentation*. Princeton University, 1969.

[37] Chihiro Watanabe, Kaoru Hiramatsu, and Kunio Kashino. Understanding community structure in layered neural networks. *Neurocomputing*, 367:84–102, 2019.

[38] Yun Gu, Xueming Qian, Qing Li, Meng Wang, Richang Hong, and Qi Tian. Image annotation by latent community detection and multikernel learning. *IEEE Transactions on Image Processing*, 24(11):3450–3463, 2015.

[39] Symeon Papadopoulos, Christos Zigkolis, Giorgos Tolias, Yannis Kalantidis, Phivos Mylonas, Yiannis Kompatsiaris, and Athena Vakali. Image clustering through community detection on hybrid image similarity graphs. In *2010 IEEE International Conference on Image Processing*, pages 2353–2356. IEEE, 2010.

[40] Leonardo N Ferreira and Liang Zhao. Time series clustering via community detection in networks. *Information Sciences*, 326:227–242, 2016.

[41] Zhiqiang Zhong, Cheng-Te Li, and Jun Pang. Hierarchical message-passing graph neural networks. *arXiv preprint arXiv:2009.03717*, 2020.

[42] Antonia Godoy-Lorite, Roger Guimerà, Cristopher Moore, and Marta Sales-Pardo. Accurate and scalable social recommendation using mixed-membership stochastic block models. *Proceedings of the National Academy of Sciences*, 113(50):14207–14212, 2016.

[43] Data Study Group team. Data study group final report: Entale, January 2022.

[44] Arzum Karataş and Serap Şahin. Application areas of community detection: A review. In *2018 International congress on big data, deep learning and fighting cyber terrorism (IBIGDELFT)*, pages 65–70. IEEE, 2018.

[45] Xing Su, Shan Xue, Fanzhen Liu, Jia Wu, Jian Yang, Chuan Zhou, Wenbin Hu, Cecile Paris, Surya Nepal, Di Jin, et al. A comprehensive survey on commu-

nity detection with deep learning. *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

[46] Mark EJ Newman and Michelle Girvan. Finding and evaluating community structure in networks. *Physical review E*, 69(2):026113, 2004.

[47] Colin McDiarmid and Fiona Skerman. Modularity of regular and treelike graphs. *Journal of Complex Networks*, 6:596–619, 8 2018.

[48] Roger Guimerà, Marta Sales-Pardo, and Luís A Nunes Amaral. Modularity from fluctuations in random graphs and complex networks. *Physical Review E*, 70:25101, 8 2004.

[49] Pan Zhang and Cristopher Moore. Scalable detection of statistically significant communities and hierarchies, using message passing for modularity. *Proceedings of the National Academy of Sciences*, 111(51):18144–18149, 2014.

[50] Tiago P Peixoto. Descriptive vs. inferential community detection: pitfalls, myths and half-truths. *arXiv preprint arXiv:2112.00183*, 2021.

[51] Santo Fortunato and Marc Barthélemy. Resolution limit in community detection. *Proceedings of the National Academy of Sciences*, 104(1):36–41, 2007.

[52] Alex Arenas, Alberto Fernandez, and Sergio Gomez. Analysis of the structure of complex networks at different resolution levels. *New journal of physics*, 10(5):053039, 2008.

[53] Jörg Reichardt and Stefan Bornholdt. Statistical mechanics of community detection. *Physical review E*, 74(1):016110, 2006.

[54] Andrea Lancichinetti and Santo Fortunato. Limits of modularity maximization in community detection. *Physical review E*, 84(6):066122, 2011.

[55] Mingming Chen, Konstantin Kuzmin, and Boleslaw K Szymanski. Community detection via maximization of modularity and its variants. *IEEE Transactions on Computational Social Systems*, 1(1):46–65, 2014.

[56] Mingming Chen, Tommy Nguyen, and Boleslaw K Szymanski. A new metric for quality of network community structure. *arXiv preprint arXiv:1507.04308*, 2015.

[57] M E J Newman. Equivalence between modularity optimization and maximum likelihood methods for community detection. *Physical Review E*, 94:52315, 11 2016.

[58] Aurelien Decelle, Florent Krzakala, Cristopher Moore, and Lenka Zdeborová. Inference and phase transitions in the detection of modules in sparse networks. *Physical Review Letters*, 107(6):065701, 2011.

[59] Elchanan Mossel, Joe Neeman, and Allan Sly. Reconstruction and estimation in the planted partition model. *Probability Theory and Related Fields*, 162(3):431–461, 2015.

[60] Tiago P Peixoto. Bayesian stochastic blockmodeling. *Advances in network clustering and blockmodeling*, pages 289–332, 2019.

[61] Paul W. Holland, Kathryn Blackmond Laskey, and Samuel Leinhardt. Stochastic blockmodels: First steps. *Social Networks*, 5(2):109–137, 1983.

[62] Tiago P. Peixoto. Parsimonious module inference in large networks. *Phys. Rev. Lett.*, 110:148701, Apr 2013.

[63] Tiago P Peixoto. Hierarchical block structures and high-resolution model selection in large networks. *Physical Review X*, 4(1):011047, 2014.

[64] Ryan J Gallagher, Jean-Gabriel Young, and Brooke Foucault Welles. A clarified typology of core-periphery structure in networks. *Science Advances*, 7(12):eabc9800, 2021.

[65] Aurelien Decelle, Florent Krzakala, Cristopher Moore, and Lenka Zdeborová. Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, 84, 12 2011.

[66] Laurent Massoulié. Community detection thresholds and the weak ramanujan property. In *Proceedings of the forty-sixth annual ACM symposium on Theory of computing*, pages 694–703, 2014.

[67] Elchanan Mossel, Joe Neeman, and Allan Sly. A proof of the block model threshold conjecture. *Combinatorica*, 38(3):665–708, 2018.

[68] Pan Zhang, Cristopher Moore, and M. E. J. Newman. Community detection in networks with unequal groups. 8 2015.

[69] Z. Xiao, T. Martin, and Mej Newman. Identification of core-periphery structure in networks. *Physical Review E*, 91(3-1):784321–784321–10, 2014.

[70] Pan Zhang, Cristopher Moore, and Lenka Zdeborová. Phase transitions in semisupervised clustering of sparse networks. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, 90, 11 2014.

[71] Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *science*, 286(5439):509–512, 1999.

[72] Lennart Gulikers, Marc Lelarge, and Laurent Massoulié. An impossibility result for reconstruction in the degree-corrected stochastic block model. *The Annals of Applied Probability*, 28(5):3002–3027, 2018.

[73] Xiaoran Yan, Cosma Shalizi, Jacob E Jensen, Florent Krzakala, Cristopher Moore, Lenka Zdeborová, Pan Zhang, and Yaojia Zhu. Model selection for degree-corrected block models. *Journal of Statistical Mechanics: Theory and Experiment*, 2014(5):P05007, 2014.

[74] Tiago P Peixoto. Nonparametric bayesian inference of the microcanonical stochastic block model. *Physical Review E*, 95(1):012317, 2017.

[75] Christopher Aicher, Abigail Z Jacobs, and Aaron Clauset. Learning latent block structure in weighted networks. *Journal of Complex Networks*, 3(2):221–248, 2015.

[76] Tiago P Peixoto. Nonparametric weighted stochastic block models. *Physical Review E*, 97(1):012306, 2018.

[77] Peter J Mucha, Thomas Richardson, Kevin Macon, Mason A Porter, and Jukka-Pekka Onnela. Community structure in time-dependent, multiscale, and multiplex networks. *science*, 328(5980):876–878, 2010.

[78] Petter Holme and Jari Saramäki. Temporal networks. *Physics reports*, 519(3):97–125, 2012.

[79] Marya Bazzi, Mason A Porter, Stacy Williams, Mark McDonald, Daniel J Fenn, and Sam D Howison. Community detection in temporal multilayer networks, with an application to correlation networks. *Multiscale Modeling & Simulation*, 14(1):1–41, 2016.

[80] Jaewon Yang, Julian McAuley, and Jure Leskovec. Community detection in networks with node attributes. In *2013 IEEE 13th international conference on data mining*, pages 1151–1156. IEEE, 2013.

[81] Martina Contisciani, Eleanor A Power, and Caterina De Bacco. Community detection with node attributes in multilayer networks. *Scientific reports*, 10(1):1–16, 2020.

[82] Siva R Sundaresan, Ilya R Fischhoff, Jonathan Dushoff, and Daniel I Rubenstein. Network metrics reveal differences in social organization between two fission–fusion species, grevy's zebra and onager. *Oecologia*, 151(1):140–149, 2007.

[83] Lizhi Zhang and Tiago P Peixoto. Statistical inference of assortative community structures. *Physical Review Research*, 2(4):043271, 2020.

[84] Yifan Hu. Efficient, high-quality force-directed graph drawing. *Mathematica journal*, 10(1):37–71, 2005.

[85] Jierui Xie, Stephen Kelley, and Boleslaw K Szymanski. Overlapping community detection in networks: The state-of-the-art and comparative study. *Acm computing surveys (csur)*, 45(4):1–35, 2013.

[86] Vincenzo Nicosia, Giuseppe Mangioni, Vincenza Carchiolo, and Michele Malgeri. Extending the definition of modularity to directed graphs with overlapping communities. *Journal of Statistical Mechanics: Theory and Experiment*, 2009(03):P03024, 2009.

[87] Brian Karrer and Mark EJ Newman. Stochastic blockmodels and community structure in networks. *Physical review E*, 83(1):016107, 2011.

[88] E T Jaynes. Information theory and statistical mechanics. *Physical Review*, 106:620–630, 5 1957.

[89] E T Jaynes. Prior probabilities. *IEEE Transactions on Systems Science and Cybernetics*, 4:227–241, 1968.

[90] Bradley P Carlin and Thomas A Louis. *Bayesian methods for data analysis*. CRC Press, 2008.

[91] S. Petrone, J. Rousseau, and C. Scricciolo. Bayes and empirical bayes: Do they merge? *Biometrika*, 101:285–302, 2014.

[92] MTCAJ Thomas and A Thomas Joy. *Elements of information theory*. Wiley-Interscience, 2006.

[93] Neil JA Sloane. The on-line encyclopedia of integer sequences. In *Towards mechanized mathematical assistants*, pages 130–130. Springer, 2007.

[94] Marc Mezard and Andrea Montanari. *Information, physics, and computation*. Oxford University Press, 2009.

[95] Jorma Rissanen. Modeling by shortest data description. *Automatica*, 14(5):465–471, 1978.

[96] Peter D Grünwald. *The minimum description length principle*. MIT press, 2007.

[97] W Keith Hastings. Monte carlo sampling methods using markov chains and their applications. 1970.

[98] Scott Kirkpatrick, C Daniel Gelatt, and Mario P Vecchi. Optimization by simulated annealing. *science*, 220(4598):671–680, 1983.

[99] Roger Guimera and Luís A Nunes Amaral. Functional cartography of complex metabolic networks. *nature*, 433(7028):895–900, 2005.

[100] Tiago P Peixoto. Efficient monte carlo and greedy heuristic for the inference of stochastic block models. *Physical Review E*, 89(1):012804, 2014.

[101] Tiago P Peixoto. Merge-split markov chain monte carlo for community detection. *Physical Review E*, 102(1):012305, 2020.

[102] Pan Zhang, Florent Krzakala, Jörg Reichardt, and Lenka Zdeborová. Comparative study for inference of hidden classes in stochastic block models. *Journal of Statistical Mechanics: Theory and Experiment*, 2012(12):P12021, 2012.

[103] Tiago P. Peixoto. The graph-tool python library. *figshare*, 2014.

[104] M. E. J. Newman. Mixing patterns in networks. *Phys. Rev. E*, 67:026126, Feb 2003.

[105] Mark EJ Newman. Communities, modules and large-scale structure in networks. *Nature physics*, 8(1):25–31, 2012.

[106] Amir Ghasemian, Homa Hosseinmardi, and Aaron Clauset. Evaluating overfit and underfit in models of network community structure. *IEEE Transactions on Knowledge and Data Engineering*, 32(9):1722–1735, 2020.

[107] Filipi N Silva, Aiiad Albeshri, Vijey Thayananthan, Wadee Alhalabi, and Santo Fortunato. Robustness modularity in complex networks. *Physical Review E*, 105(5):054308, 2022.

[108] William H Jefferys and James O Berger. Ockham's razor and bayesian analysis. *American scientist*, 80(1):64–72, 1992.

[109] A Roxana Pamfil, Sam D Howison, Renaud Lambiotte, and Mason A Porter. Relating modularity maximization and stochastic block models in multilayer networks. *SIAM Journal on Mathematics of Data Science*, 1(4):667–698, 2019.

[110] Ryan Gibson and Peter Mucha. Finite-state parameter space maps for pruning partitions in modularity-based community detection. 2022.

[111] Xiaoyan Lu and Boleslaw K Szymanski. A regularized stochastic block model for the robust community detection in complex networks. *Scientific reports*, 9(1):1–9, 2019.

[112] Daniel Gribel, Thibaut Vidal, and Michel Gendreau. Assortative-constrained stochastic block models. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 6212–6218. IEEE, 2021.

[113] Anne Condon and Richard M Karp. Algorithms for graph partitioning on the planted partition model. *Random Structures & Algorithms*, 18(2):116–140, 2001.

[114] Mark EJ Newman. Fast algorithm for detecting community structure in networks. *Physical review E*, 69(6):066133, 2004.

[115] U Brandes, D Delling, M Gaertler, R Görke, M Hoefer, Z Nikoloski, and D Wagner. On modularity-np-completeness and beyond. 2006, 2006.

[116] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008:P10008, 2008.

[117] V A Traag, L Waltman, and N J van Eck. From louvain to leiden: guaranteeing well-connected communities. *Scientific Reports*, 9:5233, 2019.

[118] Mark EJ Newman. Equivalence between modularity optimization and maximum likelihood methods for community detection. *Physical Review E*, 94(5):052315, 2016.

[119] Benjamin H Good, Yves-Alexandre de Montjoye, and Aaron Clauset. Performance of modularity maximization in practical contexts. *Physical Review E*, 81:46106, 4 2010.

[120] Jérôme Kunegis. Konect: The koblenz network collection. In *Proceedings of the 22nd International Conference on World Wide Web*, WWW '13 Companion, pages 1343–1350, New York, NY, USA, 2013. Association for Computing Machinery.

[121] V. Krebs. Political books network. *unpublished*.

[122] Valdis Krebs. Uncloaking terrorist networks. *First Monday*, 2002.

[123] David Lusseau, Karsten Schneider, Oliver J. Boisseau, Patti Haase, Elisabeth Slooten, and Steve M. Dawson. The bottlenose dolphin community of doubtful sound features a large proportion of long-lasting associations. *Behavioral Ecology and Sociobiology*, 54(4):396–405, 2003.

[124] Kathleen Mullan Harris, Carolyn T Halpern, Eric Whitsel, Jon Hussey, Joyce Tabor, Pamela Entzel, and J Richard Udry. The national longitudinal study of adolescent to adult health: Research design. 2009.

[125] Tiago P. Peixoto. Revealing consensus and dissensus between network partitions. *Phys. Rev. X*, 11:021003, Apr 2021.

[126] Zafarani Reza and Liu Huan. Social computing data repository. 2009.

[127] Lada A Adamic and Natalie Glance. The political blogosphere and the 2004 us election: divided they blog. In *Proceedings of the 3rd international workshop on Link discovery*, pages 36–43, 2005.

[128] Brian Karrer, Mark EJ Newman, and Lenka Zdeborová. Percolation on sparse networks. *Physical review letters*, 113(20):208702, 2014.

[129] Jure Leskovec, Jon Kleinberg, and Christos Faloutsos. Graph evolution: Densification and shrinking diameters. *ACM transactions on Knowledge Discovery from Data (TKDD)*, 1(1):2–es, 2007.

[130] Hawoong Jeong, Sean P Mason, A-L Barabási, and Zoltan N Oltvai. Lethality and centrality in protein networks. *Nature*, 411(6833):41–42, 2001.

[131] University of Missouri-St. Louis, Saint Louis (Mo.), Saint Louis (Mo.). Metropolitan Police Department, and Missouri. Department of Health. *The St. Louis Homicide Project: Local Responses to a National Problem*. University, 1991.

[132] Jörg Reichardt and Stefan Bornholdt. When are networks truly modular? *Physica D: Nonlinear Phenomena*, 224(1-2):20–26, 2006.

[133] Claude Elwood Shannon. A mathematical theory of communication. *ACM SIGMOBILE mobile computing and communications review*, 5(1):3–55, 2001.

[134] George A Miller. *WordNet: An electronic lexical database*. MIT press, 1998.

[135] Leto Peel, Daniel B Larremore, and Aaron Clauset. The ground truth about metadata and community detection in networks. *Science advances*, 3(5):e1602548, 2017.

[136] Endre Süli and David F Mayers. *An introduction to numerical analysis*. Cambridge university press, 2003.

[137] James Wilson Moody. *The structure of adolescent social relations: modeling friendship in dynamic social settings*. The University of North Carolina at Chapel Hill, 1999.

[138] Shai S Shen-Orr, Ron Milo, Shmoolik Mangan, and Uri Alon. Network motifs in the transcriptional regulation network of escherichia coli. *Nature genetics*, 31(1):64–68, 2002.

[139] Jordi Duch and Alex Arenas. Community detection in complex networks using extremal optimization. *Physical review E*, 72(2):027104, 2005.

[140] John Atkinson Hobson. *The Evolution of Modern Capitalism (Routledge Revivals): A Study of Machine Production*. Routledge, 2013.

[141] Wikimedia foundation. wikimedia downloads.

[142] Diego P Vázquez and Daniel Simberloff. Changes in interaction biodiversity induced by an introduced ungulate. *Ecology Letters*, 6(12):1077–1083, 2003.

[143] Michael Fire, Gilad Katz, Yuval Elovici, Bracha Shapira, and Lior Rokach. Predicting student exam's scores by analyzing social network data. In *International Conference on Active Media Technology*, pages 584–595. Springer, 2012.

[144] Donald Ervin Knuth. *The Stanford GraphBase: a platform for combinatorial computing*, volume 1. AcM Press New York, 1993.

[145] James Moody. Peer influence groups: identifying dense clusters in large networks. *Social networks*, 23(4):261–283, 2001.

[146] Thijs Beuming, Lucy Skrabanek, Masha Y Niv, Piali Mukherjee, and Harel Weinstein. Pdzbase: a protein–protein interaction database for pdz-domains. *Bioinformatics*, 21(6):827–828, 2005.

[147] Geoff Boeing. Street network models and measures for every us city, county, urbanized area, census tract, and zillow-defined neighborhood. *Urban Science*, 3(1):28, 2019.

[148] 9th dimacs implementation challenge - shortest paths.

[149] Lovro Šubelj and Marko Bajec. Robust network community detection using balanced propagation. *The European Physical Journal B*, 81(3):353–362, 2011.

[150] Duncan J Watts and Steven H Strogatz. Collective dynamics of 'small-world'networks. *nature*, 393(6684):440–442, 1998.

[151] Emmanuel Abbe. Community detection and stochastic block models: Recent developments. *Journal of Machine Learning Research*, 18(177):1–86, 2018.

[152] Lennart Gulikers, Marc Lelarge, and Laurent Massoulié. Non-backtracking spectrum of degree-corrected stochastic block models. *arXiv preprint arXiv:1609.02487*, 2016.

[153] Yukito Iba. The nishimori line and bayesian statistics. *Journal of Physics A: Mathematical and General*, 32(21):3875, 1999.

[154] M Newman and G Barkema. Monte carlo methods in statistical physics chapter 1-4. *New York, USA*, 1999.

[155] Ginestra Bianconi and Matteo Marsili. Loops of any size and hamilton cycles in random scale-free networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2005(06):P06005, 2005.

[156] Paul Expert, Tim S Evans, Vincent D Blondel, and Renaud Lambiotte. Uncovering space-independent communities in spatial networks. *Proceedings of the National Academy of Sciences*, 108(19):7663–7668, 2011.

[157] Renaud Lambiotte, Martin Rosvall, and Ingo Scholtes. From networks to optimal higher-order models of complex systems. *Nature physics*, 15(4):313–320, 2019.

[158] Tarun Kumar, Sankaran Vaidyanathan, Harini Ananthapadmanabhan, Srinivasan Parthasarathy, and Balaraman Ravindran. Hypergraph clustering by iteratively reweighted modularity maximization. *Applied Network Science*, 5(1):1–22, 2020.

[159] Marya Bazzi, Lucas GS Jeub, Alex Arenas, Sam D Howison, and Mason A Porter. A framework for the construction of generative models for mesoscale structure in multilayer networks. *Physical Review Research*, 2(2):023100, 2020.

[160] Tiago P Peixoto. Inferring the mesoscale structure of layered, edge-valued, and time-varying networks. *Physical Review E*, 92(4):042807, 2015.

[161] Florent Krzakala, Cristopher Moore, Elchanan Mossel, Joe Neeman, Allan Sly, Lenka Zdeborová, and Pan Zhang. Spectral redemption in clustering sparse networks. *Proceedings of the National Academy of Sciences*, 110(52):20935–20940, 2013.

[162] Renaud Lambiotte, Jean-Charles Delvenne, and Mauricio Barahona. Random walks, markov processes and the multiscale modular organization of complex networks. *IEEE Transactions on Network Science and Engineering*, 1(2):76–90, 2014.

[163] Martin Rosvall and Carl T Bergstrom. Maps of information flow reveal community structure in complex networks. *arXiv preprint physics.soc-ph/0707.0609*, 2007.

[164] Thomas Aynaud. python-louvain x.y: Louvain algorithm for community detection. https://github.com/taynaud/python-louvain, 2020.

[165] David Lusseau, Karsten Schneider, Oliver J Boisseau, Patti Haase, Elisabeth Slooten, and Steve M Dawson. The bottlenose dolphin community of doubtful sound features a large proportion of long-lasting associations. *Behavioral Ecology and Sociobiology*, 54(4):396–405, 2003.

[166] Phillip Bonacich and Philip Lu. *Introduction to mathematical sociology*. Princeton University Press, 2012.

[167] James Coleman, Elihu Katz, and Herbert Menzel. The diffusion of an innovation among physicians. *Sociometry*, 20(4):253–270, 1957.

[168] S Atran. John jay & artis transnational terrorism database. *College of Criminal Justice*, 2009.

[169] Simon Knight, Hung X Nguyen, Nickolas Falkner, Rhys Bowden, and Matthew Roughan. The internet topology zoo. *IEEE Journal on Selected Areas in Communications*, 29(9):1765–1775, 2011.

[170] Robust Action. the rise of the medici. *American Journal of Sociology*, 98:1259–1319, 1993.

[171] Eric Schwimmer. *Exchange in the social structure of the Orokaiva: traditional and emergent ideologies in the Northern District of Papua*. Angus & Robertson, 1973.

[172] Kenneth E Read. Cultures of the central highlands, new guinea. *Southwestern Journal of Anthropology*, 10(1):1–43, 1954.

[173] Tom AB Snijders, Gerhard G Van de Bunt, and Christian EG Steglich. Introduction to stochastic actor-based models for network dynamics. *Social networks*, 32(1):44–60, 2010.

[174] CJ Rhodes and P Jones. Inferring missing links in partially observed social networks. In *OR, Defence and Security*, pages 256–271. Springer, 2015.

[175] Jermain Kaminski, Michael Schober, Raymond Albaladejo, Oleksandr Zastupailo, and César Hidalgo. Moviegalaxies-social networks in movies. 2018.

[176] Donald S Sade. Sociometrics of macaca mulatta i. linkages and cliques in grooming matrices. *Folia primatologica*, 18(3-4):196–223, 1972.

[177] Karine Descormiers and Carlo Morselli. Alliances, conflicts, and contradictions in montreal's street gang landscape. *International Criminal Justice Review*, 21(3):297–314, 2011.

[178] Wayne W Zachary. An information flow model for conflict and fission in small groups. *Journal of anthropological research*, 33(4):452–473, 1977.

[179] Wouter De Nooy. A literary playground: Literary criticism and balance theory. *Poetics*, 26(5-6):385–404, 1999.

[180] TR Grant. Dominance and association among members of a captive and a free-ranging group of grey kangaroos (macropus giganteus). *Animal Behaviour*, 21(3):449–456, 1973.

[181] Katherine Faust. Centrality in affiliation networks. *Social networks*, 19(2):157–191, 1997.

[182] Roy Barnes and Tracy Burkett. Structural redundancy and multiplicity in corporate networks. *Connections*, 30(2):4–20, 2010.

[183] Valdis E Krebs. Mapping networks of terrorist cells. *Connections*, 24(3):43–52, 2002.

[184] David Krackhardt. Cognitive social structures. *Social networks*, 9(2):109–134, 1987.

[185] Rion Brattig Correia, Luciana P de Araújo Kohler, Mauro M Mattos, and Luis M Rocha. City-wide electronic health records reveal gender and age biases in administration of known drug–drug interactions. *NPJ digital medicine*, 2(1):1–13, 2019.

[186] Cathrine Seierstad and Tore Opsahl. For the few not the many? the effects of affirmative action on presence, prominence, and social capital of women directors in norway. *Scandinavian journal of management*, 27(1):44–54, 2011.

[187] Martin W Schein and Milton H Fohrman. Social dominance relationships in a herd of dairy cattle. *The British Journal of Animal Behaviour*, 3(2):45–55, 1955.

[188] Dale F Lott. Dominance relations and breeding rate in mature male american bison. *Zeitschrift für Tierpsychologie*, 49(4):418–432, 1979.

[189] Twitter (2018). Ulf aslak (@ulfaslak) tweet.

[190] Ross M Thompson and CR Townsend. Impacts on stream food webs of native and exotic forest: an intercontinental comparison. *Ecology*, 84(1):145–161, 2003.

[191] Anne-Marie Niekamp, Liesbeth AG Mercken, Christian JPA Hoebe, and Nicole HTM Dukers-Muijrers. A sexual affiliation network of swingers, heterosexuals practicing risk behaviours that potentiate the spread of sexually transmitted infections: a two-mode approach. *Social Networks*, 35(2):223–236, 2013.

[192] Christine C Hass. Social status in female bighorn sheep (ovis canadensis): expression, development and reproductive correlates. *Journal of Zoology*, 225(3):509–523, 1991.

[193] Brian Hayes. Connecting the dots. *American Scientist*, 94(5):400–404, 2006.

[194] M Vickers and S Chan. Representing classroom social structure. melbourne: Victoria institute of secondary education. 1981.

[195] James Samuel Coleman et al. Introduction to mathematical sociology. *Introduction to mathematical sociology.*, 1964.

[196] Malcolm P Young. The organization of neural systems in the primate cerebral cortex. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 252(1333):13–18, 1993.

[197] Linton C Freeman, Sue C Freeman, and Alaina G Michaelson. On human social intelligence. *Journal of Social and Biological Structures*, 11(4):415–425, 1988.

[198] Andrew Beveridge and Jie Shan. Network of thrones. *Math Horizons*, 23(4):18–22, 2016.

[199] Matteo Magnani, Barbora Micenkova, and Luca Rossi. Combinatorial analysis of multiple networks. *arXiv preprint arXiv:1303.4986*, 2013.

[200] Luke M Gerdes, Kristine Ringler, and Barbara Autin. Assessing the abu sayyaf group's strategic and learning capacities. *Studies in Conflict & Terrorism*, 37(3):267–293, 2014.

[201] Gerhard G Van de Bunt, Marijtje AJ Van Duijn, and Tom AB Snijders. Friendship networks through time: An actor-oriented dynamic statistical network model. *Computational & Mathematical Organization Theory*, 5(2):167–192, 1999.

[202] Mark EJ Newman. Finding community structure in networks using the eigenvectors of matrices. *Physical review E*, 74(3):036104, 2006.

[203] Manlio De Domenico, Albert Solé-Ribalta, Sergio Gómez, and Alex Arenas. Navigability of interconnected networks under random failures. *Proceedings of the National Academy of Sciences*, 111(23):8351–8356, 2014.

[204] Moses C Kiti, Michele Tizzoni, Timothy M Kinyanjui, Dorothy C Koech, Patrick K Munywoki, Milosch Meriac, Luca Cappa, André Panisson, Alain Barrat, Ciro Cattuto, et al. Quantifying social contacts in a household setting of rural kenya using wearable proximity sensors. *EPJ data science*, 5(1):1–21, 2016.

[205] Goylette F Chami, Sebastian E Ahnert, Narcis B Kabatereine, and Edridah M Tukahebwa. Social network fragmentation and community health. *Proceedings of the National Academy of Sciences*, 114(36):E7425–E7431, 2017.

[206] Sean Slattery and Mark Craven. Combining statistical and relational methods for learning in hypertext domains. In *International Conference on Inductive Logic Programming*, pages 38–52. Springer, 1998.

[207] Piotr Sapiezynski, Arkadiusz Stopczynski, David Dreyer Lassen, and Sune Lehmann. Interaction data from the copenhagen networks study. *Scientific Data*, 6(1):1–10, 2019.

[208] Mathieu Génois, Christian L Vestergaard, Julie Fournet, André Panisson, Isabelle Bonmarin, and Alain Barrat. Data on face-to-face contacts in an office building suggest a low-cost vaccination strategy based on community linkers. *Network Science*, 3(3):326–347, 2015.

[209] Emmanuel Lazega et al. *The collegial phenomenon: The social mechanisms of cooperation among peers in a corporate law partnership*. Oxford University Press on Demand, 2001.

[210] Ron Milo, Shai Shen-Orr, Shalev Itzkovitz, Nadav Kashtan, Dmitri Chklovskii, and Uri Alon. Network motifs: simple building blocks of complex networks. *Science*, 298(5594):824–827, 2002.

[211] Philippe Vanhems, Alain Barrat, Ciro Cattuto, Jean-François Pinton, Nagham Khanafer, Corinne Régis, Byeul-a Kim, Brigitte Comte, and Nicolas Voirin. Estimating potential infection transmission routes in hospital wards using wearable proximity sensors. *PloS one*, 8(9):e73970, 2013.

[212] Yukio Takahata. Diachronic changes in the dominance relations of adult female japanese monkeys of the arashiyama b group. *The monkeys of Arashiyama. State University of New York Press, Albany*, pages 123–139, 1991.

[213] Makoto Kato, Takehiko Kakutani, Tamiji Inoue, and Takao Itino. Insect-flower relationship in the primary beech forest of ashu, kyoto: an overview of the flowering phenology and the seasonal pattern of insect visits. *Contributions from the biological laboratory, Kyoto University*, 27(4):309–376, 1990.

[214] Scott Decker, Carol W Kohfeld, Richard Rosenfeld, and John Sprague. St. louis homicide project: Local responses to a national problem. *A report made to the community*, pages 22–23, 1991.

[215] Nicolas Simonis, Jean-François Rual, Anne-Ruxandra Carvunis, Murat Tasan, Irma Lemmens, Tomoko Hirozane-Kishikawa, Tong Hao, Julie M Sahalie, Kavitha Venkatesan, Fana Gebreab, et al. Empirically controlled mapping of the caenorhabditis elegans protein-protein interactome network. *Nature methods*, 6(1):47–54, 2009.

[216] William Gray Roncal, Zachary H Koterba, Disa Mhembere, Dean M Kleissas, Joshua T Vogelstein, Randal Burns, Anita R Bowles, Dimitrios K Donavos, Sephira Ryman, Rex E Jung, et al. Migraine: Mri graph reliability analysis and inference for connectomics. In *2013 IEEE Global Conference on Signal and Information Processing*, pages 313–316. IEEE, 2013.

[217] Julie Fournet and Alain Barrat. Contact patterns among high school students. *PloS one*, 9(9):e107878, 2014.

[218] Linton C Freeman, Cynthia M Webster, and Deirdre M Kirke. Exploring social structure using dynamic three-dimensional color images. *Social networks*, 20(2):109–118, 1998.

[219] Stéphane Coulomb, Michel Bauer, Denis Bernard, and Marie-Claude Marsolier-Kergoat. Gene essentiality and the topology of protein interaction networks. *Proceedings of the Royal Society B: Biological Sciences*, 272(1573):1721–1725, 2005.

[220] Benjamin F Maier and Dirk Brockmann. Cover time for random walks on arbitrary complex networks. *Physical Review E*, 96(4):042307, 2017.

[221] Robert E Ulanowicz and Donald L DeAngelis. Network analysis of trophic dynamics in south florida ecosystems. *US Geological Survey Program on the South Florida Ecosystem*, 114:45, 2005.

[222] Augustin Chaintreau, Pan Hui, Jon Crowcroft, Christophe Diot, Richard Gass, and James Scott. Impact of human mobility on opportunistic forwarding algorithms. *IEEE Transactions on Mobile Computing*, 6(6):606–620, 2007.

[223] Lorenzo Isella, Juliette Stehlé, Alain Barrat, Ciro Cattuto, Jean-François Pinton, and Wouter Van den Broeck. What's in a crowd? analysis of face-to-face behavioral networks. *Journal of theoretical biology*, 271(1):166–180, 2011.

[224] Samin Aref and Mark C Wilson. Balance and frustration in signed networks. *Journal of Complex Networks*, 7(2):163–189, 2019.

[225] Mikael Huss and Petter Holme. Currency and commodity metabolites: their identification and relation to the modularity of metabolic networks. *IET systems biology*, 1(5):280–285, 2007.

[226] Michael Fire and Rami Puzis. Organization mining using online social networks. *Networks and Spatial Economics*, 16(2):545–578, 2016.

[227] United States Federal Aviation Administration. Air traffic control system command center.

[228] Neo D Martinez. Artifacts or attributes? effects of resolution on the little rock lake food web. *Ecological monographs*, 61(4):367–392, 1991.

[229] Jure Leskovec and Julian Mcauley. Learning to discover social circles in ego networks. *Advances in neural information processing systems*, 25, 2012.

[230] Pablo M Gleiser and Leon Danon. Community structure in jazz. *Advances in complex systems*, 6(04):565–573, 2003.

[231] Daniel B Larremore, Aaron Clauset, and Caroline O Buckee. A network approach to analyzing highly recombinant malaria parasite genes. *PLoS computational biology*, 9(10):e1003268, 2013.

[232] Alessio Cardillo, Jesús Gómez-Gardenes, Massimiliano Zanin, Miguel Romance, David Papo, Francisco del Pozo, and Stefano Boccaletti. Emergence of network features from multiplexity. *Scientific reports*, 3(1):1–6, 2013.

[233] Ulrich Stelzl, Uwe Worm, Maciej Lalowski, Christian Haenig, Felix H Brembeck, Heike Goehler, Martin Stroedicke, Martina Zenkner, Anke Schoenherr, Susanne Koeppen, et al. A human protein-protein interaction network: a resource for annotating the proteome. *Cell*, 122(6):957–968, 2005.

[234] Radosław Michalski, Sebastian Palus, and Przemysław Kazienko. Matching organizational structure and social network extracted from email communication. In *International conference on business information systems*, pages 197–206. Springer, 2011.

[235] Steven J Cook, Travis A Jarrell, Christopher A Brittin, Yi Wang, Adam E Bloniarz, Maksim A Yakovlev, Ken CQ Nguyen, Leo T-H Tang, Emily A Bayer, Janet S Duerr, et al. Whole-animal connectomes of both caenorhabditis elegans sexes. *Nature*, 571(7763):63–71, 2019.

[236] Stephen Kosack, Michele Coscia, Evann Smith, Kim Albrecht, Albert-László Barabási, and Ricardo Hausmann. Functional structures of us state governments. *Proceedings of the National Academy of Sciences*, 115(46):11748–11753, 2018.

[237] University of oregon route views project, 2001.

[238] Roger Guimera, Leon Danon, Albert Diaz-Guilera, Francesc Giralt, and Alex Arenas. Self-similar community structure in a network of human interactions. *Physical review E*, 68(6):065103, 2003.

[239] Rossana Mastrandrea, Julie Fournet, and Alain Barrat. Contact patterns in a high school: a comparison between data collected using wearable sensors, contact diaries and friendship surveys. *PloS one*, 10(9):e0136497, 2015.

[240] Juliette Stehlé, Nicolas Voirin, Alain Barrat, Ciro Cattuto, Lorenzo Isella, Jean-François Pinton, Marco Quaggiotto, Wouter Van den Broeck, Corinne Régis, Bruno Lina, et al. High-resolution measurements of face-to-face contact patterns in a primary school. *PloS one*, 6(8):e23176, 2011.

[241] Jean-François Rual, Kavitha Venkatesan, Tong Hao, Tomoko Hirozane-Kishikawa, Amélie Dricot, Ning Li, Gabriel F Berriz, Francis D Gibbons, Matija Dreze, Nono Ayivi-Guedehoussou, et al. Towards a proteome-scale map of the human protein–protein interaction network. *Nature*, 437(7062):1173–1178, 2005.

[242] Jennifer A Dunne, Conrad C Labandeira, and Richard J Williams. Highly resolved early eocene food webs show development of modern trophic structure after the end-cretaceous extinction. *Proceedings of the Royal Society B: Biological Sciences*, 281(1782):20133280, 2014.

[243] Rob M Ewing, Peter Chu, Fred Elisma, Hongyan Li, Paul Taylor, Shane Climie, Linda McBroom-Cerajewski, Mark D Robinson, Liam O'Connor, Michael Li, et al. Large-scale mapping of human protein–protein interactions by mass spectrometry. *Molecular systems biology*, 3(1):89, 2007.

[244] Manlio De Domenico, Andrea Lancichinetti, Alex Arenas, and Martin Rosvall. Identifying modular flows on multilayer networks reveals highly overlapping organization in interconnected systems. *Physical Review X*, 5(1):011027, 2015.

[245] A. Calderone. A wikipedia based map of science, 2020.

[246] Sean R Collins, Patrick Kemmeren, Xue-Chu Zhao, Jack F Greenblatt, Forrest Spencer, Frank CP Holstege, Jonathan S Weissman, and Nevan J Krogan. Toward a comprehensive atlas of the physical interactome of saccharomyces cerevisiae. *Molecular & Cellular Proteomics*, 6(3):439–450, 2007.

[247] Department of Economic United Nations and Population Division Social Affairs. Trends in international migrant stock: The 2015 revision.

[248] C. Harrison C. Römhild. Bible cross-references.

[249] Manlio De Domenico, Vincenzo Nicosia, Alexandre Arenas, and Vito Latora. Structural reducibility of multilayer networks. *Nature communications*, 6(1):1–9, 2015.

[250] J. Kunegis. Dnc emails co-recipients, 2016.

[251] Dingqi Yang, Daqing Zhang, Zhiyong Yu, and Zhiwen Yu. Fine-grained preference-aware location search leveraging crowdsourced digital footprints from lbsns. In *Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing*, pages 479–488, 2013.

[252] Michael Fire and Carlos Guestrin. The rise and fall of network stars: Analyzing 2.5 million graphs to reveal how high-degree vertices emerge over time. *Information Processing & Management*, 57(2):102041, 2020.

[253] C. Robertson. Flowers and insects: lists of visitors to four hundred and fifty-three flowers, 1929.

[254] Manlio De Domenico, Mason A Porter, and Alex Arenas. Muxviz: a tool for multilayer analysis and visualization of networks. *Journal of Complex Networks*, 3(2):159–176, 2015.

[255] Zachary P Neal. A sign of the times? weak and strong polarization in the us congress, 1973–2016. *Social Networks*, 60:103–112, 2020.

[256] The openflights.org website.

[257] B Stabler. Transportation network test problems, 2021.

[258] Srijan Kumar, Francesca Spezzano, VS Subrahmanian, and Christos Faloutsos. Edge weight prediction in weighted signed networks. In *2016 IEEE 16th International Conference on Data Mining (ICDM)*, pages 221–230. IEEE, 2016.

[259] Luis EC Rocha, Fredrik Liljeros, and Petter Holme. Simulated epidemics in an empirical spatiotemporal network of 50,185 sexual contacts. *PLoS computational biology*, 7(3):e1001109, 2011.

[260] Paolo Massa, Martino Salvetti, and Danilo Tomasoni. Bowling alone and trust decline in social network sites. In *2009 Eighth IEEE International Conference on Dependable, Autonomic and Secure Computing*, pages 658–663. IEEE, 2009.

[261] Matei Ripeanu, Ian Foster, and Adriana Iamnitchi. Mapping the gnutella network: Properties of large-scale peer-to-peer systems and implications for system design. *arXiv preprint cs/0209028*, 2002.

[262] Martina Morris. *HIV Transmission Network Metastudy Project: An Archive of Data From Eight Network Studies, 1988–2001*. 2011.

[263] Ron Milo, Shalev Itzkovitz, Nadav Kashtan, Reuven Levitt, Shai Shen-Orr, Inbal Ayzenshtat, Michal Sheffer, and Uri Alon. Superfamilies of evolved and designed networks. *Science*, 303(5663):1538–1542, 2004.

[264] Mark EJ Newman. The structure of scientific collaboration networks. *Proceedings of the national academy of sciences*, 98(2):404–409, 2001.

[265] Michael Fire, Rami Puzis, and Yuval Elovici. Link prediction in highly fractional data sets. In *Handbook of computational approaches to counterterrorism*, pages 283–300. Springer, 2013.

[266] Furkan Gursoy and Dilek Gunnec. Influence maximization in social networks under deterministic linear threshold model. *Knowledge-Based Systems*, 161:111–123, 2018.

[267] Michael Ley. The dblp computer science bibliography: Evolution, research issues, perspectives. In *International symposium on string processing and information retrieval*, pages 1–10. Springer, 2002.

[268] Lovro Šubelj and Marko Bajec. Software systems through complex networks science: Review, analysis and applications. In *Proceedings of the First International Workshop on Software Mining*, pages 9–16, 2012.

[269] The caida as relationships dataset. http://www.caida.org/data/as-relationships/.

[270] J. Kunegis. Jdk dependency. https://doi.org/10.1145/2487788.2488173, 2016.

[271] Kaggle. Chess ratings - elo versus the rest of the world. https://www.kaggle.com/c/chess/data.

[272] Bureau of Transportation Statistics. T-100 domestic market. https://www.transtats.bts.gov/DL_SelectFields.asp?Table_ID=310, 2017.

[273] Tad A Dallas, A Alonso Aguirre, Sarah Budischak, Colin Carlson, Vanessa Ezenwa, Barbara Han, Shan Huang, and Patrick R Stephens. Gauging support for macroecological patterns in helminth parasites. *Global Ecology and Biogeography*, 27(12):1437–1447, 2018.

[274] Balázs Szalkai, Csaba Kerepesi, Bálint Varga, and Vince Grolmusz. The budapest reference connectome server v2. 0. *Neuroscience Letters*, 595:60–62, 2015.

[275] GroupLens Research. Movielens data sets. http://grouplens.org/datasets/movielens/.

[276] Munmun De Choudhury, Hari Sundaram, Ajita John, and Dorée Duncan Seligmann. Social synchrony: Predicting mimicry of user actions in online social media. In *2009 International conference on computational science and engineering*, volume 4, pages 151–158. IEEE, 2009.

[277] Andrew Kachites McCallum, Kamal Nigam, Jason Rennie, and Kristie Seymore. Automating the construction of internet portals with machine learning. *Information Retrieval*, 3(2):127–163, 2000.

[278] Vladimir Batagelj, Andrej Mrvar, and Matjaz Zaversnik. *Network analysis of texts*. University of Ljubljana, Inst. of Mathematics, Physics and Mechanics ..., 2002.

[279] Ricardo Alberich, Joe Miro-Julia, and Francesc Rosselló. Marvel universe looks almost like a real social network. *arXiv preprint cond-mat/0202174*, 2002.

[280] Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. Dbpedia: A nucleus for a web of open data. In *The semantic web*, pages 722–735. Springer, 2007.

[281] Jure Leskovec, Daniel Huttenlocher, and Jon Kleinberg. Governance in social media: A case study of the wikipedia promotion process. In *Fourth international AAAI conference on weblogs and social media*, 2010.

[282] Beichuan Zhang, Raymond Liu, Daniel Massey, and Lixia Zhang. Collecting the internet as-level topology. *ACM SIGCOMM Computer Communication Review*, 35(1):53–61, 2005.

[283] Kevin Gullikson. Python dependency analysis. http://kgullikson88.github.io/blog/pypi-analysis.html.

[284] Vicenç Gómez, Andreas Kaltenbrunner, and Vicente López. Statistical analysis of the social network and discussion threads in slashdot. In *Proceedings of the 17th international conference on World Wide Web*, pages 645–654, 2008.

[285] G Joshi-Tope, Marc Gillespie, Imre Vastrik, Peter D'Eustachio, Esther Schmidt, Bernard de Bono, Bijay Jassal, GR Gopinath, GR Wu, Lisa Matthews, et al. Reactome: a knowledgebase of biological pathways. *Nucleic acids research*, 33(suppl_1):D428–D432, 2005.

[286] Gergely Palla, Illés J Farkas, Péter Pollner, Imre Derényi, and Tamás Vicsek. Directed network modules. *New journal of physics*, 9(6):186, 2007.

[287] Bryan Klimt and Yiming Yang. The enron corpus: A new dataset for email classification research. In *European conference on machine learning*, pages 217–226. Springer, 2004.

[288] Robert West, Hristo S Paskov, Jure Leskovec, and Christopher Potts. Exploiting social network structure for person-to-person sentiment analysis. *Transactions of the Association for Computational Linguistics*, 2:297–310, 2014.

[289] Bimal Viswanath, Alan Mislove, Meeyoung Cha, and Krishna P Gummadi. On the evolution of user interaction in facebook. In *Proceedings of the 2nd ACM workshop on Online social networks*, pages 37–42, 2009.

[290] Oliver Richters and Tiago P Peixoto. Trust transitivity in social networks. *PloS one*, 6(4):e18384, 2011.

[291] J. Kunegis. Linux. https://doi.org/10.1145/2487788.2488173, 2016.

[292] James H Fowler and Sangick Jeon. The authority of supreme court precedent. *Social networks*, 30(1):16–30, 2008.

[293] Anna Evtushenko and Michael T Gastner. Beyond fortune 500: women in a global network of directors. In *International Conference on Complex Networks and Their Applications*, pages 586–598. Springer, 2019.

[294] George R Kiss, Christine Armstrong, Robert Milroy, and James Piper. An associative thesaurus of english and its computer analysis. *The computer and literary studies*, pages 153–165, 1973.

[295] Elisa Omodei, Manlio De De Domenico, and Alex Arenas. Characterizing interactions in online social networks during exceptional events. *Frontiers in Physics*, 3:59, 2015.

[296] J. Kunegis. Douban network dataset. https://doi.org/10.1145/2487788.2488173, 2016.

[297] Dingqi Yang, Daqing Zhang, Longbiao Chen, and Bingqing Qu. Nationtelescope: Monitoring and visualizing large-scale collective behavior in lbsns. *Journal of Network and Computer Applications*, 55:170–180, 2015.

[298] Matthew Richardson, Rakesh Agrawal, and Pedro Domingos. Trust management for the semantic web. In *International semantic Web conference*, pages 351–368. Springer, 2003.

[299] S. Chacon. The 2009 github contest. https://doi.org/10.1007/978-1-4842-0076-6_6.

[300] Johannes Gehrke, Paul Ginsparg, and Jon Kleinberg. Overview of the 2003 kdd cup. *Acm Sigkdd Explorations Newsletter*, 5(2):149–151, 2003.

[301] Jérôme Kunegis, Andreas Lommatzsch, and Christian Bauckhage. The slashdot zoo: mining a social network with negative edges. In *Proceedings of the 18th international conference on World wide web*, pages 741–750, 2009.

[302] Christiane Fellbaum. Wordnet. In *Theory and applications of ontology: computer applications*, pages 231–243. Springer, 2010.

[303] Silviu Maniu, Talel Abdessalem, and Bogdan Cautis. Casting a web of trust over wikipedia: an interaction-based approach. In *Proceedings of the 20th international conference companion on World wide web*, pages 87–88, 2011.

[304] Julia Preusse, Jérôme Kunegis, Matthias Thimm, Steffen Staab, and Thomas Gottron. Structural dynamics of knowledge networks. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 7, pages 506–515, 2013.

[305] Yutao Zhang and Jie Tang. Social network integration: towards constructing the social graph. *arXiv preprint arXiv:1311.2670*, 2013.

[306] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. What is twitter, a social network or a news media? In *Proceedings of the 19th international conference on World wide web*, pages 591–600, 2010.

[307] Daniele Perito, Claude Castelluccia, Mohamed Ali Kaafar, and Pere Manils. How unique and traceable are usernames? In *International Symposium on Privacy Enhancing Technologies Symposium*, pages 1–17. Springer, 2011.

[308] Terence Chen, Mohamed Ali Kaafar, Arik Friedman, and Roksana Boreli. Is more always merrier? a deep dive into online social footprints. In *Proceedings of the 2012 ACM workshop on Workshop on online social networks*, pages 67–72, 2012.

[309] The caida ucsd macroscopic skitter topology dataset. https://www.caida.org/tools/measurement/skitter/.

[310] Jérôme Kunegis, Gerd Gröner, and Thomas Gottron. Online dating recommender systems: The split-complex number approach. In *Proceedings of the 4th ACM RecSys workshop on Recommender systems and the social web*, pages 37–44, 2012.

Appendix A

Supplementary materials for Chapter 3

A.1 Maximum entropy distribution for the degree-propensity parameter

The Lagrangian function to be optimised with respect to $\boldsymbol{\theta}$ is the equation (2.19), which reads as the following

$$\begin{aligned} L(p(\boldsymbol{\theta}|\boldsymbol{b}), \xi_0) = &- \int_{\boldsymbol{\theta}\in\boldsymbol{C}} p(\boldsymbol{\theta}|\boldsymbol{b}) \ln p(\boldsymbol{\theta}|\boldsymbol{b}))d\boldsymbol{\theta} + \xi_0 \left(\int_{\boldsymbol{\theta}\in\boldsymbol{C}} p(\boldsymbol{\theta}|\boldsymbol{b})d\boldsymbol{\theta} - 1 \right), \\ &\text{with } \boldsymbol{C} = \{\boldsymbol{\theta} : \hat{\theta}_r = \sum_u \theta_u \delta_{b_u r} = 1, \ \ \forall r = 1, 2, .., B\}. \end{aligned} \tag{A.1}$$

Differentiating $L(p(\boldsymbol{\theta}), \xi_0$ with respect to $p(\boldsymbol{\theta}|\boldsymbol{b})$ leads to the following expression

$$\frac{\partial L(p(\boldsymbol{\theta}|\boldsymbol{b}), \xi_0)}{\partial p(\boldsymbol{\theta}|\boldsymbol{b})} = -\ln p(\boldsymbol{\theta}|\boldsymbol{b})) - 1 + \xi_0. \tag{A.2}$$

Setting the derivate equal to zero, we obtain the form of the maximum entropy distribution

$$\ln p(\boldsymbol{\theta}|\boldsymbol{b})) = \xi_0 - 1 \Rightarrow p(\boldsymbol{\theta}|\boldsymbol{b}) = e^{\xi_0 - 1}. \tag{A.3}$$

Since the expression of $p(\boldsymbol{\theta}|\boldsymbol{b})$ in the last equation only depends on a constant ξ_0 (rather than $\boldsymbol{\theta}$ itself), the maximum entropy distribution $p(\boldsymbol{\theta}|\boldsymbol{b}))$ is simply the uniform distribution which assigns the same probability to every valid $\boldsymbol{\theta}$ in $\boldsymbol{C} = \{\boldsymbol{\theta} : \hat{\theta}_r = \sum_u \theta_u \delta_{b_u r} = 1, \ \ \forall r = 1, 2, .., B\}$. The probability is then the reciprocal of the volume of $\boldsymbol{C}$, which can be obtained by multiplying the volumes of B regular simplexes,

$$p(\boldsymbol{\theta}|\boldsymbol{b}) = \frac{1}{\text{Vol}(C)} = \prod_r^B (n_r - 1)! \times \delta_{\sum_u \theta_{b_u r}, 1}. \tag{A.4}$$

The second equality in the last equation is implied by a result to be proved later in equation (A.11).

A.2 Marginal likelihood of DC-SBM

The marginal likelihood of DC-SBM can be obtained by computing the integral

$$P(\boldsymbol{A}|\boldsymbol{b}) = \int P(\boldsymbol{A}|\boldsymbol{\lambda},\boldsymbol{\theta},\boldsymbol{b})P(\boldsymbol{\lambda}|\boldsymbol{b})P(\boldsymbol{\theta}|\boldsymbol{b})d\boldsymbol{\lambda}d\boldsymbol{\theta}. \tag{A.5}$$

As shown at the beginning of Section 3.1, the likelihood function of DC-SBM can be rewritten as follows

$$\begin{aligned} P(\boldsymbol{A}|\boldsymbol{\lambda},\boldsymbol{\theta},\boldsymbol{b}) &= \prod_{u<v} e^{-\theta_u\theta_v\lambda_{b_ub_v}} \frac{(\theta_u\theta_v\lambda_{b_ub_v})^{A_{uv}}}{A_{uv}!} \prod_u e^{-\theta_u^2\lambda_{b_ub_u}/2} \frac{(\theta_u^2\lambda_{b_ub_u}/2)^{A_{uu}/2}}{(A_{uu}/2)!!} \\ &= \prod_{r<s} e^{-\hat{\theta}_r\hat{\theta}_s\lambda_{rs}}\lambda_{rs}^{e_{rs}} \prod_r e^{-\hat{\theta}_r^2\lambda_{rr}/2}\lambda_{rr}^{e_{rr}/2} \times \frac{\prod_u \theta_u^{k_u}}{\prod_{u<v}A_{uv}!\prod_u A_{uu}!!}. \end{aligned} \tag{A.6}$$

To proceed, consider the following uninformative priors,

$$\begin{aligned} P(\lambda_{rs}|\bar{\lambda}) &= e^{-\lambda_{rs}/\bar{\lambda}(1+\delta_{rs})}/\bar{\lambda}(1+\delta_{rs}), \quad \text{for } \lambda_{rs}\in[0,\infty], \\ P(\boldsymbol{\theta}|\boldsymbol{b}) &= \prod_r (n_r-1)! \times \delta_{\hat{\theta}_r,1}, \end{aligned} \tag{A.7}$$

the integral in equation (A.5) becomes

$$\begin{aligned} P(\boldsymbol{A}|\bar{\lambda},\boldsymbol{b}) &= \prod_{r<s}\int e^{-\lambda_{rs}(1+1/\bar{\lambda})}\frac{\lambda_{rs}^{e_{rs}}}{\bar{\lambda}}d\lambda_{rs} \times \prod_r \int e^{-\lambda_{rr}(1/2+1/2\bar{\lambda})}\frac{(\lambda_{rr}/2)^{e_{rr}/2}}{2\bar{\lambda}}d\lambda_{rr} \\ &\qquad \times \prod_r (n_r-1)! \int_{\hat{\theta}_r=1} \prod_{u:b_u=r}\theta_u^{k_u}d\boldsymbol{\theta} \times \frac{1}{\prod_{u<v}A_{uv}!\prod_u A_{uu}!!} \\ &:= \frac{I\times II\times III}{\prod_{u<v}A_{uv}!\prod_u A_{uu}!!}. \end{aligned} \tag{A.8}$$

The first two products in the numerator in the last equation can be obtained using integration by parts,

$$\begin{aligned} I &= \prod_{r<s}\int e^{-\lambda_{rs}(1+1/\bar{\lambda})}\frac{\lambda_{rs}^{e_{rs}}}{\bar{\lambda}}d\lambda_{rs} = \prod_{r<s}\frac{\bar{\lambda}^{e_{rs}}}{(\bar{\lambda}+1)^{e_{rs}+1}}e_{rs}!, \\ II &= \prod_r \int e^{-\lambda_{rr}(1/2+1/2\bar{\lambda})}\frac{(\lambda_{rr}/2)^{e_{rr}/2}}{2\bar{\lambda}}d\lambda_{rr} = \prod_r \frac{\bar{\lambda}_{rr}^{e_{rr}/2}}{(\bar{\lambda}+1)^{e^{rr}/2+1}}e_{rr}!!. \end{aligned} \tag{A.9}$$

For the third product III in which we integrate the degree propensity parameter $\boldsymbol{\theta}$,

$$III = \prod_r (n_r - 1)! \int_{\hat{\theta}_r=1} \prod_{u:b_u=r} \theta_u^{k_u} d\boldsymbol{\theta} \tag{A.10}$$

we need to make use of the following result

$$\int_{y\in\Omega} \prod_{u=1}^{n} y_u^{a_u-1} \mathrm{d}y = \frac{h^{(\sum_u a_u - 1)} \times \prod_{u=1}^n \Gamma(a_u)}{\Gamma(\sum_{u=1}^n a_u)}, \qquad \Omega = \{y | y_u \geq 0, \sum_{u=1}^{n} y_u = h\}. \tag{A.11}$$

Let's assume this result hold for the moment and apply it to each integral in the equation (A.10) with $a_u = k_u + 1$ and $h = 1$, we have

$$III = \prod_u k_u! \prod_r \frac{(n_r - 1)!}{(e_r + n_r - 1)!}. \tag{A.12}$$

Substituting I, II and III back to the equation (A.8) gives us the expression of the marginal likelihood of DC-SBM in equation (2.23).

In the following we derivate the result in equation (A.11). Firstly we define

$$\int_{y\in\Omega} \prod_{u=1}^{n} y_u^{a_u-1} \mathrm{d}y \triangleq I_n(h). \tag{A.13}$$

Then we can change the variables of integration as the following:

$$\begin{aligned}
I_n(h) &= \int_{y\in\mathcal{R}_+^n} y_1^{a_1-1} y_2^{a_2-1} \cdots y_n^{a_n-1} \times \boldsymbol{\delta}(y_1 + y_2 + \cdots + y_n - h) \quad \mathrm{d}y_1 \mathrm{d}y_2 \cdots \mathrm{d}y_n \\
&= (h^{\sum_u a_u - 1}) \int_{y\in\mathcal{R}_+^n} \left(\frac{y_1}{h}\right)^{a_1-1} \cdots \left(\frac{y_n}{h}\right)^{a_n-1} \times \boldsymbol{\delta}\left(\frac{y_1}{h} + \frac{y_2}{h} + \cdots + \frac{y_n}{h} - 1\right) \quad \mathrm{d}\frac{y_1}{h} \mathrm{d}\frac{y_2}{h} \cdots \mathrm{d}\frac{y_n}{h} \\
&= (h^{\sum_u a_u - 1}) \int_{u\in\mathcal{R}_+^n} g_1^{a_1-1} g_2^{a_2-1} \cdots g_n^{a_n-1} \times \boldsymbol{\delta}(g_1 + \cdots + g_n - 1) \quad \mathrm{d}g_1 \cdots \mathrm{d}g_n \qquad \left(g_u = \frac{y_u}{h}\right) \\
&= (h^{\sum_u a_u - 1}) I_n(1). \qquad (*)
\end{aligned}$$

Now our task is just to compute the integral $I_n(1)$. To proceed, notice that we can rewrite $I_n(1)$ as follows

$$
\begin{aligned}
I_n(1) &= \int_{u\in\mathcal{R}_+^n} g_1^{a_1-1} g_2^{a_2-1}\cdots g_n^{a_n-1} \times \boldsymbol{\delta}(g_1+g_2+\cdots+g_n-1) \quad \mathrm{d}g_1\mathrm{d}g_2\cdots\mathrm{d}g_n \\
&= \int_0^1 g_1^{a_1-1}\Big(\int_0^{(1-g_1)} u_2^{a_2-1}\cdots \int_0^{(1-g_1)(1-\sum_{u=2}^{n-1} g_u/(1-g_1))} g_n^{a_n-1} \quad \mathrm{d}g_n\mathrm{d}g_{n-1}\cdots\mathrm{d}g_2\Big)\mathrm{d}g_1 \\
&= \int_0^1 g_1^{a_1-1} I_{n-1}(1-g_1)\mathrm{d}g_1.
\end{aligned}
$$

From the result in $(*)$, we have $I_{n-1}(1-g_1) = I_{n-1}(1)(1-g_1)^{\sum_{g=2} a_g - 1}$, then the last equation becomes

$$
\begin{aligned}
I_n(1) &= I_{n-1}(1)\int_0^1 u_1^{a_1-1}(1-g_1)^{\sum_{u=2}^n a_u - 1}\mathrm{d}u_1 \\
&= I_{n-1}(1)\mathcal{B}(a_1, \sum_{u=2}^n a_u),
\end{aligned}
$$

where $\mathcal{B}$ is the Beta function. This gives us the recursion of $I_n(1)$. Recall the relationship between the Beta function and gamma function is

$$
\mathcal{B}(x,y) = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}.
$$

Then, the recursion of $I_n(1)$ can be written as

$$
\begin{aligned}
I_n(1) &= I_{n-1}(1)\frac{\Gamma(a_1)\Gamma(\sum_{u=2}^n a_u)}{\Gamma(\sum_{u=1}^n a_u)} \\
&= I_{n-2}(1)\frac{\Gamma(a_2)\Gamma(\sum_{u=3}^n a_u)}{\Gamma(\sum_{u=2}^n a_u)}\frac{\Gamma(a_1)\Gamma(\sum_{u=2}^n a_u)}{\Gamma(\sum_{u=1}^n a_u)} \\
&\cdots \\
&= \frac{\prod_{u=1}^n \Gamma(a_u)}{\Gamma(\sum_{u=1}^n a_u)}. \qquad (**)
\end{aligned}
$$

Substituting $(**)$ back into $(*)$ leads to the result in equation (A.11).

A.3 Simplifying the likelihood function of DC-SBM

Here, we explain how to derive the expression of the likelihood function of DC-SBM in equation (3.2)

$$P(\boldsymbol{A}|\boldsymbol{\theta},\boldsymbol{\lambda},\boldsymbol{b}) = \prod_{r<s} e^{-\hat{\theta}_r\hat{\theta}_s\lambda_{rs}}\lambda_{rs}^{e_{rs}} \prod_r e^{-\hat{\theta}_r^2\lambda_{rr}/2}\lambda_{rr}^{e_{rr}/2} \times \frac{\prod_u \theta_u^{k_u}}{\prod_{u<v} A_{uv}! \prod_u A_{uu}!!} \tag{A.14}$$

from equation (3.1)

$$P(\boldsymbol{A}|\boldsymbol{\theta},\boldsymbol{\lambda},\boldsymbol{b}) = \prod_{u<v} e^{-\theta_u\theta_v\lambda_{b_ub_v}} \frac{(\theta_u\theta_v\lambda_{b_ub_v})^{A_{uv}}}{A_{uv}!} \prod_u e^{-\theta_u^2\lambda_{b_ub_u}/2} \frac{(\theta_u^2\lambda_{b_ub_u}/2)^{A_{uu}/2}}{(A_{uu}/2)!!}. \tag{A.15}$$

The main idea is to rearrange terms properly, changing the order of multiplications from being node-wise to being community-wise (i.e. according to the community membership of nodes). For the illustration purpose, we firstly consider a toy example - a graph consists of **only two nodes** u and v. The reason why we look at this trivial example is that we can easily write down every term in equation (A.15) and equation (A.14) for the toy example such that we can get the inspiration about deriving the latter from the former in a general setting.

Likelihood of DC-SBM for the toy example with nodes belonging to the same group

We firstly consider a parameterisation $\boldsymbol{b} = (b_u, b_v)$ where node u and v belong to the same group denoted by r, i.e.

$$r = b_u = b_v. \tag{A.16}$$

Then, we write down the likelihood function in equation (A.15) for the toy example as follows

$$P(\boldsymbol{A}|\boldsymbol{\lambda},\boldsymbol{\theta},\boldsymbol{b}) = \left[e^{-\theta_u\theta_v\lambda_{rr}} \frac{(\theta_u\theta_v\lambda_{rr})^{A_{uv}}}{A_{uv}!}\right] \times \left[e^{-\theta_u^2\lambda_{rr}/2} \frac{(\theta_u^2\lambda_{rr}/2)^{A_{uu/2}}}{(A_{uu}/2)!!} \times e^{-\theta_v^2\lambda_{rr}/2} \frac{(\theta_v^2\lambda_{rr}/2)^{A_{vv}/2}}{(A_{vv}/2)!}\right] \tag{A.17}$$

$$= \left(e^{-(\theta_u+\theta_v)^2\lambda_{rr}/2}\right) \times \left(\lambda_{rr}^{(A_{uv}+A_{uu}/2+A_{vv}/2)}\right) \times \left(\frac{\theta_u^{(A_{uv}+A_{uu})}\theta_v^{(A_{uv}+A_{vv})}}{A_{uv}!\left[2^{A_{uu}/2}(A_{uu}/2)!\right]\left[(2^{A_{vv}/2}A_{vv}/2)!\right]}\right). \tag{A.18}$$

One key observation in the last equation is that once exponential functions are combined together, the exponent of the exponential function is in the form of square of the sum of degree parameters θ_u. Furthermore, since there are only two nodes in the entire graph, we have the following identities hold

$$k_u = \sum_{w=1}^{N=2} A_{uw} = A_{uv} + A_{uu}, \tag{A.19}$$

$$e_{rr} = \sum_{u,v=1}^{N=2} A_{uv}\delta_{b_u b_v} = A_{uv} + A_{vu} + A_{uu} + A_{vv}, \tag{A.20}$$

where k_u is the degree of node u and e_{rr} is twice the number of edges inside group r. Remind that the definition of the adjacency matrix $\boldsymbol{A}$ of a network takes the convention that A_{uv} is the number of edges between nodes u and v, while A_{uu} is twice the number of self-loops of node u. Since we have been restricted to symmetric networks, we have

$$A_{uv} = A_{vu}, \forall u, v. \tag{A.21}$$

and therefore equation (A.20) indicates the following equations hold

$$e_{rr}/2 = A_{uv} + A_{uu}/2 + A_{vv}/2 \tag{A.22}$$

$$= A_{vu} + A_{uu}/2 + A_{vv}/2. \tag{A.23}$$

With these results, equation (A.18) can be further written as

$$P(\boldsymbol{A}|\boldsymbol{\lambda}, \boldsymbol{\theta}, \boldsymbol{b}) = e^{-\hat{\theta}_r^2 \lambda_{rr}/2} \lambda_{rr}^{e_{rr}/2} \times \frac{\theta_u^{k_u} \theta_v^{k_v}}{A_{uv}!(A_{uu}!!)(A_{vv}!!)}, \tag{A.24}$$

with

$$\hat{\theta}_r = \sum_{u=1}^{N=2} \theta_u \delta_{r_{b_u}} = \theta_u + \theta_v \tag{A.25}$$

being the sum of degree parameters θ_u inside group r, and $A_{uu}!! = 2^{A_{uu}/2}(A_{uu}/2)!$. It can be easily verified that the expression derived in equation (A.24) is consistent with equation (A.14).

Likelihood of DC-SBM for the toy example with nodes belonging to distinct groups

Consider the same toy example as we discuss above, but with a different parameterisation $\boldsymbol{b}'$ where nodes are assumed to belong to distinct groups. That is,

$$r = b_u \neq b_v = s. \tag{A.26}$$

Again, we can write down every term of the likelihood function in equation (A.15) for the toy example,

$$P(\boldsymbol{A}|\boldsymbol{\lambda},\boldsymbol{\theta},\boldsymbol{b}) = \left[e^{-\theta_u\theta_v\lambda_{rs}}\frac{(\theta_u\theta_v\lambda_{rs})^{A_{uv}}}{A_{uv}!}\right] \times \left[e^{-\theta_u^2\lambda_{rr}/2}\frac{(\theta_u^2\lambda_{rr}/2)^{A_{uu}/2}}{(A_{uu}/2)!!} \times e^{-\theta_v^2\lambda_{rr}/2}\frac{(\theta_v^2\lambda_{rr}/2)^{A_{vv}/2}}{(A_{vv}/2)!}\right] \tag{A.27}$$

$$= \left(e^{-(\theta_u^2+\theta_v^2)\lambda_{rr}/2}e^{-\theta_u\theta_v\lambda_{rs}}\right) \times \left(\lambda_{rs}^{A_{uv}}\lambda_{rr}^{A_{uu}/2}\lambda_{ss}^{A_{vv}/2}\right) \times \left(\frac{\theta_u^{(A_{uv}+A_{uu})}\theta_v^{(A_{uv}+A_{vv})}}{A_{uv}!\left[2^{A_{uu}/2}(A_{uu}/2)!\right]\left[(2^{A_{vv}/2}A_{vv}/2)!\right]}\right). \tag{A.28}$$

In the last line above, we do nothing but rearranging terms in equation (A.15) with a hope to simplify the expression by combining similar terms together. Given that there are only two nodes in the graph and each of them form its own group (only two groups, i.e. r and s respectively), we have the following properties

$$\hat{\theta}_r = \sum_w^{N=2} \theta_w\delta_{rb_w} = \theta_u, \tag{A.29}$$

$$\hat{\theta}_s = \sum_w^{N=2} \theta_w\delta_{sb_w} = \theta_v, \tag{A.30}$$

$$\hat{\theta}_r\hat{\theta}_s = \theta_u\theta_v, \tag{A.31}$$

and

$$e_{rs} = \sum_{u,v}^{N=2} A_{uv}\delta_{rb_u}\delta_{sb_v} = A_{uv}, \tag{A.32}$$

$$e_{rr} = \sum_{u,v}^{N=2} A_{uv}\delta_{rb_u}\delta_{rb_v} = A_{uu}, \tag{A.33}$$

$$e_{ss} = \sum_{u,v}^{N=2} A_{uv}\delta_{sb_u}\delta_{sb_v} = A_{uu}. \tag{A.34}$$

Substituting equation (A.29)-(A.34) to (A.28) gives

$$P(\boldsymbol{A}|\boldsymbol{\lambda},\boldsymbol{\theta},\boldsymbol{b}) = \left(e^{-(\hat{\theta}_r^2+\hat{\theta}_s^2)\lambda_{rr}/2}e^{-\hat{\theta}_r\hat{\theta}_s\lambda_{rs}}\right) \times \left(\lambda_{rs}^{e_{rs}}\lambda_{rr}^{e_{rr}/2}\lambda_{ss}^{e_{ss}/2}\right) \times \left(\frac{\theta_u^{k_u}\theta_v^{k_v}}{A_{uv}!A_{uu}!!A_{vv}!!}\right), \tag{A.35}$$

which can be easily verified to be consistent with equation (A.14).

Likelihood of DC-SBM for any graph

Given our discussion above regarding the toy example, now we explain how to derive the expression in equation (A.14) from equation (A.15) in a general setting. Similar to what we have done above, we begin with rearranging terms in equation (A.14):

$$\begin{aligned} P(\boldsymbol{A}|\boldsymbol{\lambda},\boldsymbol{\theta},\boldsymbol{b}) &= \prod_{u<v}^{N} e^{-\theta_u\theta_v\lambda_{b_ub_v}}\frac{(\theta_u\theta_v\lambda_{b_ub_v})^{A_{uv}}}{A_{uv}!}\prod_{u}^{N} e^{-\theta_u^2\lambda_{b_ub_u}/2}\frac{(\theta_u^2\lambda_{uu}/2)^{A_{uu}/2}}{(A_{uu}/2)!} \\ &= \left[\prod_{u<v}^{N} e^{-\theta_u\theta_v\lambda_{b_ub_v}}\prod_{u}^{N} e^{-\theta_u^2\lambda_{b_ub_u}/2}\right] \times \left[\prod_{u<v}(\theta_u\theta_v)^{A_{uv}}\prod_u \theta_u^{A_{uu}})\right] \times \left[\prod_{u<v}^{N}\lambda_{b_ub_v}^{A_{uv}}\prod_{u}^{N}\lambda_{uu}^{A_{uu}/2}\right] \\ &\qquad \times \frac{1}{\prod_{u<v}A_{uv}!\prod_u 2^{A_{uu}/2}(A_{uu}/2)!} \\ &:= I \times II \times III \times \frac{1}{\prod_{u<v}A_{uv}!\prod_u A_{uu}!!}. \end{aligned} \tag{A.36}$$

To proceed, we work on each of I, II, III in the last equation separately.

- For I,

$$I = \prod_{u<v}^{N} e^{-\theta_u\theta_v\lambda_{b_ub_v}}\prod_{u}^{N} e^{-\theta_u^2\lambda_{b_ub_u}/2} \tag{A.37}$$

$$= \left[\prod_{u<v}^{N}\left(e^{-\theta_u\theta_v\lambda_{b_ub_v}}\right)^{1-\delta_{b_ub_v}}\left(e^{-\theta_u\theta_v\lambda_{b_ub_v}}\right)^{\delta_{b_ub_v}}\right] \times \prod_{u}^{N} e^{-\theta_u^2\lambda_{b_ub_u}/2} \tag{A.38}$$

$$= \prod_{u<v}^{N}\left(e^{-\theta_u\theta_v\lambda_{b_ub_v}}\right)^{1-\delta_{b_ub_v}}\prod_{u<v}^{N}\left(e^{-\theta_u\theta_v\lambda_{b_ub_v}}\right)^{\delta_{b_ub_v}} \times \prod_{u}^{N} e^{-\theta_u^2\lambda_{b_ub_v}/2}. \tag{A.39}$$

From equation (A.37) to (A.38), we simply split the first product in (A.37) into two parts: the first part corresponds to contribution from edges with end-nodes coming from distinct groups, and the second part corresponds to contribution from edges with end-nodes belonging to the same groups.

Notice that we can rewrite the first product in equation (A.39) by changing the order of the product,

$$\prod_{u<v}^{N}\left(e^{-\theta_u\theta_v\lambda_{b_ub_v}}\right)^{1-\delta_{b_ub_v}} = \prod_{r\neq s}^{B} e^{-(\lambda_{rs}/2)\sum_{u,v:u\neq v}^{N}\theta_u\theta_v\delta_{rb_u}\delta_{sb_v}} \tag{A.40}$$

$$= \prod_{r\neq s}^{B} e^{-(\lambda_{rs}/2)\left[\sum_{u=1}^{N}\theta_u\delta_{rb_u}\left(\sum_{v\neq u}^{N}\theta_v\delta_{sb_v}\right)\right]} \tag{A.41}$$

$$= \prod_{r\neq s}^{B} e^{(-\lambda_{rs}/2)\left[\sum_{u=1}^{N}\theta_u\delta_{rb_u}\left(\sum_{v=1}^{N}\theta_v\delta_{sb_v}\right)\right]} \tag{A.42}$$

$$= \prod_{r\neq s}^{B} e^{-\hat{\theta}_r\hat{\theta}_s\lambda_{rs}/2} \tag{A.43}$$

$$= \prod_{r<s}^{B} e^{-\hat{\theta}_r\hat{\theta}_s\lambda_{rs}}. \quad \text{(due to the parameterisation fact that } \lambda_{rs}=\lambda_{sr}\text{)} \tag{A.44}$$

From equation (A.41) to equation (A.42), we use the fact that

$$\left[\sum_{u=1}^{N}\theta_u\delta_{rb_u}\left(\sum_{v\neq u}^{N}\theta_v\delta_{sb_v}\right)\right] = \left[\sum_{u=1}^{N}\theta_u\delta_{rb_u}\left(\sum_{v=1}^{N}\theta_v\delta_{sb_v}\right)\right] \tag{A.45}$$

when $r \neq s$.

Then, in equation (A.39), the last two products can be manipulated as follows,

$$\prod_{u<v}^{N}\left(e^{-\theta_u\theta_v\lambda_{b_ub_v}}\right)^{\delta_{b_ub_v}} = \prod_{r}^{B} e^{-\lambda_{rr}\sum_{u,v}^{N}\theta_u\theta_v\delta_{rb_u}\delta_{rb_v}}, \tag{A.46}$$

$$\prod_{u}^{N} e^{-\theta_u^2\lambda_{b_ub_u}/2} = \prod_{r}^{B} e^{-(\lambda_{rr}/2)\sum_{u}^{N}\theta_u^2\delta_{rb_u}}, \tag{A.47}$$

and these two products can be combined such that exponents of exponential terms are

in the form of square of the community-wise sums of degree propensity parameter θ_u:

$$\prod_r^B e^{-\lambda_{rr}\sum_{u,v}^N \theta_u\theta_v\delta_{rb_u}\delta_{rb_v}} \prod_r^B e^{-(\lambda_{rr}/2)\sum_u^N \theta_u^2\delta_{rb_u}} = \prod_r^B e^{-(\sum_u^N \theta_u\delta_{ru})^2\lambda_{rr}/2} = \prod_r^B e^{-\hat{\theta}_r^2\lambda_{rr}/2}. \tag{A.48}$$

As a result, substituting equation (A.44), (A.48) into (A.39) leads us to the following expression of I defined in equation (A.36):

$$I = \prod_{r<s}^B e^{-\hat{\theta}_r\hat{\theta}_s\lambda_{rs}} \prod_r^B e^{-\hat{\theta}_r^2\lambda_{rr}/2}. \tag{A.49}$$

- For II,

$$\begin{aligned}
II &= \prod_{u<v}^N (\theta_u\theta_v)^{A_{uv}} \prod_u^N \theta_u^{A_{uu}} \\
&= \left(\prod_{u<v}^N \theta_u^{A_{uv}}\right)\left(\prod_{u<v}^N \theta_v^{A_{uv}}\right)\prod_u^N \theta_u^{A_{uu}} \\
&= \left(\prod_u^N \theta_u^{\sum_{v:u<v}^N A_{uv}}\right)\left(\prod_v^N \theta_v^{\sum_{u:u<v}^N A_{uv}}\right)\left(\prod_u^N \theta_u^{A_{uu}}\right) \\
&= \left(\prod_u^N \theta_u^{\sum_{v:v\neq u}^N A_{uv}}\right)\left(\prod_u^N \theta_u^{A_{uu}}\right) \\
&= \prod_u^N \theta_u^{\sum_v^N A_{uv}} \\
&= \prod_u^N \theta_u^{k_u}
\end{aligned} \tag{A.50}$$

- For III,

$$III = \prod_{u<v}^N \lambda_{b_ub_v}^{A_{uv}} \prod_u^N \lambda_{b_ub_u}^{A_{uu}/2} = \left(\prod_{r<s}^B \lambda_{rs}^{\sum_{u<v}^N A_{uv}\delta_{rb_u}\delta_{sb_v}}\right) \times \left(\prod_r^B \lambda_{rr}^{\sum_u^N A_{uu}/2}\right) = \prod_{r<s}^B \lambda_{rs}^{e_{rs}} \prod_r^B \lambda_{rr}^{e_{rr}/2} \tag{A.51}$$

Substituting what we have for I, II, III in equation (A.49) - (A.51) back into equation (A.36), we reach the expression in equation (A.14).

A.4 Maximum likelihood inference with DC-SBM

The log-likelihood of DC-SBM has the following expression

$$\ln P(\boldsymbol{A}|\boldsymbol{\theta},\boldsymbol{\lambda},\boldsymbol{b}) = -\frac{1}{2}\sum_{rs}\hat{\theta}_r\hat{\theta}_s\lambda_{rs} + \frac{1}{2}\sum_{rs}e_{rs}(\ln\lambda_{rs} - \delta_{rs}\ln 2) + \sum_u k_u \ln\theta_u. \quad \text{(A.52)}$$

Differentiating the log-likelihood above with respect to λ_{rs} and θ_u respectively gives

$$\frac{\partial}{\partial\lambda_{rs}}\ln P(\boldsymbol{A}|\boldsymbol{\theta},\boldsymbol{\lambda},\boldsymbol{b}) = -\hat{\theta}_r\hat{\theta}_s + \frac{e_{rs}}{\lambda_{rs}}, \quad \text{(A.53)}$$

$$\frac{\partial}{\partial\theta_u}\ln P(\boldsymbol{A}|\boldsymbol{\theta},\boldsymbol{\lambda},\boldsymbol{b}) = -\sum_r \hat{\theta}_r\lambda_{rb_u} + \frac{k_u}{\theta_u}. \quad \text{(A.54)}$$

Firstly setting the derivates respective to λ_{rs} to zeros gives the MLE of λ_{rs}

$$\lambda^*_{rs} = \frac{e_{rs}}{\hat{\theta}^*_r\hat{\theta}^*_s}. \quad \text{(A.55)}$$

For the MLE of θ_u, notice that the following equality needs to hold

$$\frac{k_u}{\theta^*_u} = \sum_r \hat{\theta}_r\lambda^*_{rb_u} = \sum_r \frac{e_{rb_u}}{\hat{\theta}^*_{b_u}} = \frac{e_{b_u}}{\hat{\theta}^*_{b_u}}, \quad \text{(A.56)}$$

for every node u in the same group b_u, which implies that

$$\theta^*_u = \frac{k_u}{e_{b_u}}\hat{\theta}^*_{b_u}. \quad \text{(A.57)}$$

The solution for θ^* remains undetermined at this point, since there are many different valid solutions which differ from each other up to some multiplying constants. This is because for some particular choices of $\{\theta^*_u\}$ satisfying the last equation, $\{c\theta^*_u\}$ is also valid for an arbitrary constant c. To fully determine the MLE for θ, we therefore need to specify a normalising constant $\hat{\theta}_r$ for each group r. The convention $\hat{\theta}_r = 1, \forall r \in \{1,2,..,B\}$ makes the λ_{rs} parameter becomes the expected number of connections between group r and s (or twice the number if $r = s$),

$$\langle e_{rs}\rangle = \frac{1}{2}\sum_{uv}\theta_u\theta_v\lambda_{b_ub_v}\delta_{rb_u}\delta_{sb_v} = \hat{\theta}_r\hat{\theta}_s\lambda_{rs} = \lambda_{rs}. \quad \text{(A.58)}$$

A.5 Maximum likelihood inference with the uniform PP model

The log-likelihood of the uniform PP model reads as

$$\ln P(\boldsymbol{A}|\lambda_{\text{in}}, \lambda_{\text{out}}, \boldsymbol{\theta}, \boldsymbol{b}) = -\lambda_{\text{in}} \sum_r \frac{1}{2}\hat{\theta}_r^2 - \lambda_{\text{out}} \sum_{r<s} \hat{\theta}_r \hat{\theta}_s + e_{\text{in}} \ln \lambda_{\text{in}} + e_{\text{out}} \ln \lambda_{\text{out}} + \sum_u k_u \ln \theta_u. \tag{A.59}$$

Differentiating the log-likelihood function above with respective to $\lambda_{\text{in}}, \lambda_{\text{out}}$ and θ_u, we get

$$\frac{\partial}{\partial \lambda_{\text{in}}} \ln P(\boldsymbol{A}|\lambda_{\text{in}}, \lambda_{\text{out}}, \boldsymbol{\theta}, \boldsymbol{b}) = -\sum_r \frac{1}{2}\hat{\theta}_r^2 + \frac{e_{\text{in}}}{\lambda_{\text{in}}}, \tag{A.60}$$

$$\frac{\partial}{\partial \lambda_{\text{out}}} \ln P(\boldsymbol{A}|\lambda_{\text{in}}, \lambda_{\text{out}}, \boldsymbol{\theta}, \boldsymbol{b}) = -\sum_{r<s} \hat{\theta}_r \hat{\theta}_s + \frac{e_{\text{out}}}{\lambda_{\text{out}}}, \tag{A.61}$$

$$\frac{\partial}{\partial \theta_u} \ln P(\boldsymbol{A}|\lambda_{\text{in}}, \lambda_{\text{out}}, \boldsymbol{\theta}, \boldsymbol{b}) = -\lambda_{\text{in}} \hat{\theta}_{b_u} - \lambda_{\text{out}} \sum_{s \neq b_u} \hat{\theta}_s + k_u/\theta_u. \tag{A.62}$$

The MLE for model parameters then are obtained by equating these derivates to zeros, leading to the following estimators

$$\lambda_{\text{in}}^* = \frac{2e_{\text{in}}}{\sum_r (\hat{\theta}_r^*)^2}, \tag{A.63}$$

$$\lambda_{\text{out}}^* = \frac{e_{\text{out}}}{\sum_{r<s} \hat{\theta}_r^* \hat{\theta}_s^*}, \tag{A.64}$$

$$\theta_u^* = \frac{k_u}{\lambda_{\text{in}}^* \hat{\theta}_{b_u}^* + \lambda_{\text{out}}^* \sum_{s \neq b_u} \hat{\theta}_s^*}. \tag{A.65}$$

We can substitute the MLE λ_{in}^* and λ_{out}^* into θ_u^*, which gives us exactly the expression of θ_u we present in equation (3.18).

A.6 Marginal likelihood of the uniform PP model

Here we provide detailed derivation of the marginal likelihood of the uniform PP model presented in equation (3.50). We need to conduct the integral in equation (3.45), which reads as

$$\begin{aligned} P(\boldsymbol{A}|\bar{\lambda},\boldsymbol{b}) &= \int P(\boldsymbol{A}|\lambda_{\text{in}},\lambda_{\text{out}},\boldsymbol{\theta},\boldsymbol{b})P(\lambda_{\text{in}},\lambda_{\text{out}}|\bar{\lambda})P(\boldsymbol{\theta}|\boldsymbol{b})\,d\lambda_{\text{in}}d\lambda_{\text{out}}d\boldsymbol{\theta} \\ &= \int e^{-\lambda_{\text{out}}\sum_{r<s}\hat{\theta}_r\hat{\theta}_s}\lambda_{\text{out}}^{e_{\text{out}}}e^{-\lambda_{\text{in}}\sum_r\hat{\theta}_r^2/2}\lambda_{\text{in}}^{e_{\text{in}}}\frac{\prod_u\theta_u^{k_u}}{\prod_{u<v}A_{uv}!\prod_u A_{uu}!!}\times e^{\lambda_{\text{out}}/\bar{\lambda}}/\bar{\lambda}\times e^{\lambda_{\text{in}}/(2\bar{\lambda})}/(2\bar{\lambda}) \\ &\qquad \times\prod_r(n_r-1)!\delta_{\sum_u\theta_u\delta_{rb_u},1}d\lambda_{\text{in}}d\lambda_{\text{out}}d\boldsymbol{\theta} \end{aligned} \tag{A.66}$$

Since the parameters $\lambda_{\text{in}},\lambda_{\text{out}}$ and $\boldsymbol{\theta}$ are mutually independent, the integral in the last equation can be decomposed as follows,

$$\begin{aligned} P(\boldsymbol{A}|\bar{\lambda},\boldsymbol{b}) = \int e^{-\lambda_{\text{out}}\binom{B}{2}}\lambda_{\text{out}}^{e_{\text{out}}}\times e^{\lambda_{\text{out}}/\bar{\lambda}}/\bar{\lambda}\,d\lambda_{\text{out}}\times\int e^{-\lambda_{\text{in}}B}\lambda_{\text{in}}^{e_{\text{in}}}\times e^{\lambda_{\text{in}}/(2\bar{\lambda})}/(2\bar{\lambda})\,d\lambda_{\text{in}}\times \\ \int\prod_u\theta_u^{k_u}\prod_r(n_r-1)!\delta_{\sum_u\theta_u\delta_{rb_u},1}\,d\boldsymbol{\theta}\times\frac{1}{\prod_{u<v}A_{uv}!\prod_u A_{uu}!!}. \end{aligned} \tag{A.67}$$

The first two integrals are in the form of

$$\int_0^\infty e^{-ax}x^b dx = \frac{b!}{a^{b+1}}, \tag{A.68}$$

so we can easily obtain

$$\int e^{-\lambda_{\text{out}}\binom{B}{2}}\lambda_{\text{out}}^{e_{\text{out}}}\times e^{\lambda_{\text{out}}/\bar{\lambda}}/\bar{\lambda}\,d\lambda_{\text{out}} = \frac{e_{\text{out}}!}{\bar{\lambda}\left[\binom{B}{2}+\frac{1}{\bar{\lambda}}\right]^{e_{\text{out}}+1}}, \tag{A.69}$$

$$\int e^{-\lambda_{\text{in}}B/2}\lambda_{\text{in}}^{e_{\text{in}}}\times e^{\lambda_{\text{in}}/(2\bar{\lambda})}/\bar{\lambda}\,d\lambda_{\text{in}} = \frac{e_{\text{in}}!}{2\bar{\lambda}\left[\frac{B}{2}+\frac{1}{2\bar{\lambda}}\right]^{e_{\text{in}}+1}}. \tag{A.70}$$

Finally, making use of the result in equation (A.11), we

$$\int\prod_u\theta_u^{k_u}\prod_r(n_r-1)!\delta_{\sum_u\theta_u\delta_{rb_u},1}\,d\boldsymbol{\theta} = \prod_u k_u!\prod_r\frac{(n_r-1)!}{(e_r+n_r-1)!}. \tag{A.71}$$

Combining equations (A.69) - (A.71), we get the desired expression

$$P(\boldsymbol{A}|\bar{\lambda},\boldsymbol{b}) = \frac{e_{\text{in}!}e_{\text{out}}!}{2\bar{\lambda}^2\left[\frac{B}{2}+\frac{1}{2\bar{\lambda}}\right]^{e_{\text{in}}+1}\left[\binom{B}{2}+\frac{1}{\bar{\lambda}}\right]^{e_{\text{out}}+1}} \times \prod_r \frac{(n_r-1)!}{(e_r+n_r-1)!} \times \frac{k_u!}{\prod_{u<v} A_{uv}! \prod_u A_{uu}!!}. \tag{A.72}$$

A.7 Marginal likelihood of the non-uniform PP model

The model likelihood of the non-uniform PP model described in the text can be written as follows

$$P(\boldsymbol{A}|\{\boldsymbol{\lambda_r}\},\lambda_{\rm out},\boldsymbol{\theta},\boldsymbol{b}) = e^{-\lambda_{\rm out}\sum_{r<s}\hat{\theta}_r\hat{\theta}_s}\lambda_{\rm out}^{e_{\rm out}} \times \prod_r e^{-\lambda_r\hat{\theta}_r^2/2}\lambda_r^{e_{rr}/2} \times \frac{\prod_u \theta_u^{k_u}}{\prod_{u<v}A_{uv}!\prod_u A_{uu}!!}. \tag{A.73}$$

Enforcing the constraint $\hat{\theta}_r = 1$ and using the maximum-entropy priors

$$P(\lambda_r|\bar{\lambda}) = e^{-\lambda_r/2\bar{\lambda}}/(2\bar{\lambda}), \tag{A.74}$$

$$P(\lambda_{\rm out}|\bar{\lambda}) = e^{\lambda_{\rm out}/\bar{\lambda}}/\bar{\lambda}, \tag{A.75}$$

$$P(\boldsymbol{\theta}|\boldsymbol{b}) = \prod_r (n_r-1)!\delta(\sum_r \theta_u\delta_{b_u r}-1), \tag{A.76}$$

where $\bar{\lambda}$ is a hyperparameter for the expected number of edges between any pair of distinct groups. As we have explained in the text, it is possible to take the empirical Bayes' approach and set $\bar{\lambda}$ equal to the average number of connections among all B communities, $\bar{\lambda} = 2E/(B+1)B$. By computing the following integral,

$$P(\boldsymbol{A}|\bar{\lambda},\boldsymbol{b}) = \int P(\boldsymbol{A}|\{\lambda_r\},\lambda_{\rm out},\boldsymbol{\theta},\boldsymbol{b})P(\{\lambda_r\},\lambda_{\rm out}|\bar{\lambda})P(\boldsymbol{\theta}|\boldsymbol{b})\,d\lambda_r d\lambda_{\rm out} d\boldsymbol{\theta} \tag{A.77}$$

we obtain the marginal likelihood of the non-uniform PP model,

$$P(\boldsymbol{A}|\bar{\lambda},\boldsymbol{b}) = \frac{e_{\rm out}!\prod_r (e_{rr}/2)!}{\left(\frac{1}{2}+\frac{1}{2\bar{\lambda}}\right)^{e_{\rm in}+B}\left[\binom{B}{2}+\frac{1}{\bar{\lambda}}\right]^{e_{\rm out}+1}} \times \frac{\prod_u k_u!}{\prod_{u<v}A_{uv}!\prod_u A_{uu}!!}. \tag{A.78}$$

Just as we have seen for the DC-SBM and uniform PP model, the marginal likelihood of the non-uniform has an alternative interpretation, which is based on the microcanonical formulation of SBM. The marginal likelihood in the last equation can be rewritten as

$$P(\boldsymbol{A}|\{\lambda_r\},\boldsymbol{b}) = P(\boldsymbol{A}|\boldsymbol{e},\boldsymbol{k},\boldsymbol{b})P(\boldsymbol{k}|\boldsymbol{e},\boldsymbol{b})P(\boldsymbol{e}|\{e_{rr}\},e_{\rm out},\boldsymbol{b})P(\{e_{rr}\}|\bar{\lambda},\boldsymbol{b})P(e_{\rm out}|\bar{\lambda},\boldsymbol{b})P(E), \tag{A.79}$$

where $P(\boldsymbol{A}|\boldsymbol{e},\boldsymbol{k},\boldsymbol{b})$ is the likelihood function of the microcanonical SBM in equation (2.36) and the priors are

$$P(\boldsymbol{e}|\{e_{rr}\},e_{\rm out},\boldsymbol{b}) = \frac{e_{\rm out}!}{\binom{B}{2}^{e_{\rm out}}\prod_{r<s}e_{rs}!}, \tag{A.80}$$

$$P(\{e_{rr}\}|\bar{\lambda},\boldsymbol{b}) = \prod_r \frac{\bar{\lambda}^{e_{rr}}}{(\bar{\lambda}+1)^{e_{rr}+1}}, \tag{A.81}$$

$$P(e_{\text{out}}|\bar{\lambda},\boldsymbol{b}) = \frac{\left[\bar{\lambda}\binom{B}{2}\right]^{e_{\text{out}}}}{\left[\bar{\lambda}\binom{B}{2}+1\right]^{e_{\text{out}}+1}}, \tag{A.82}$$

and $P(E)$ is just a constant independent of the network partition $\boldsymbol{b}$ hence can be arbitrarily chosen. With the microcanonical interpretation in mind, we can replace the parametric priors $P(\{e_{rr}\}|\bar{\lambda},\boldsymbol{b})$ and $P(e_{\text{out}}|\bar{\lambda},\boldsymbol{b})$ with the following microcanonical prior,

$$\begin{aligned} P(\{e_{rr}\}, e_{\text{out}}|\boldsymbol{b}, E) &= P(\{e_{rr}\}|e_{\text{in}},\boldsymbol{b})P(e_{\text{in}}|E,\boldsymbol{b}) \\ &= \binom{B+e_{\text{in}}-1}{e_{\text{in}}}^{-1}\left(\frac{1}{E+1}\right)^{1-\delta_{B,1}}. \end{aligned} \tag{A.83}$$

This prior firstly samples the number of intra- and inter-group connections e_{in} and e_{out} from a uniform distribution $P(e_{\text{in}}, e_{\text{out}}|E) = (E+1)^{-1}$, then assigns equal probability to every possible configuration of $\{e_{rr}\}$ such that $\sum_r e_{rr} = 2e_{\text{in}}$. The marginal probability of sampling a network $\boldsymbol{A}$ from the non-uniform PP model is then

$$P(\boldsymbol{A}|\boldsymbol{b}) = P(\boldsymbol{A}|\boldsymbol{e},\boldsymbol{k},\boldsymbol{b})P(\boldsymbol{k}|\boldsymbol{e},\boldsymbol{b})P(\boldsymbol{e}|\{e_{rr}\},e_{\text{out}},\boldsymbol{b})P(\{e_{rr}\},e_{\text{out}}|\boldsymbol{b},E)P(E), \tag{A.84}$$

which has the expression as we introduce in equation (3.64).

A.8 Louvain algorithm with uniform PP model refinement

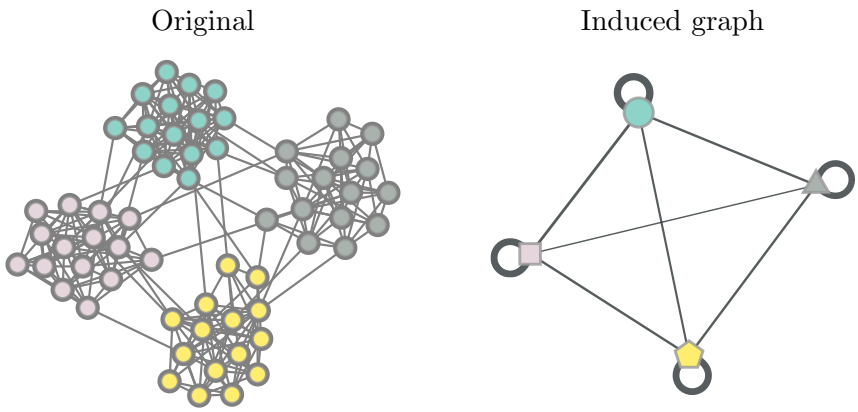


Figure A-1: The induced graph required in the second phase of the Louvain algorithm is constructed by merging nodes in the same groups together.

Recall that the Newman-Girvan modularity is defined as

$$Q = \frac{1}{2E} \sum_{uv} \left(A_{uv} - \frac{k_u k_v}{2E} \right) \delta_{uv}. \tag{A.85}$$

The *Louvain algorithm* is a greedy heuristic algorithm for finding the network partition b with the maximum modualrity. The algorithm starts from a partition where every node is put in its own community, i.e. $B = N$. To proceed, the algorithm iterates between two different phases. Firstly, in the search phase, we go through every node in the network and try to move the node to one of its neighbouring group. The criterion for whether we make the move is based on the change in the modularity value once the move is made. We should move each node to its neighbouring group which leads to the maximum increase in the modularity value. Such local search continues until there is no a move of a single node can make the modularity value increase. Then, in the induced graph phase, an induced graph is constructed by considering communities obtained in the first phase as meta-nodes, and meta-edges are placed accordingly (see Fig. A-1). The induced graph is then used as the input to conduct local search in the first phase. The algorithm iterates the two phases until the modularity value converges. The main advantage of the Louvain algorithm is in its implementation speed. The key of the efficient implementation is in the fact that it takes only constant time to evaluate

the change in the modularity value after moving a node u from group r to group s,

$$\Delta Q = \frac{1}{2E}\left(e_r + 2k_u^r - \frac{(e_r + k_u^r)^2}{2E}\right) - \frac{1}{2E}\left(e_r - \frac{e_r^2}{2E} - \frac{k_u^2}{2E}\right). \tag{A.86}$$

Moreover, it was argued that the induced graphs generated at different iterations lead to a hierarchical partition of nodes which could convey useful information of the underlying systems.

Recall that the expression of the posterior probability of the PP model is

$$P(\boldsymbol{b}|\boldsymbol{A}) = \frac{e_{\text{in}}! e_{\text{out}}!}{\left(\frac{B}{2}\right)^{e_{\text{in}}}\binom{B}{2}^{e_{\text{out}}}(E+1)^{1-\delta_{B,1}}} \times \prod_r \frac{(n_r-1)!}{(e_r+n_r-1)!} \times \frac{\prod_u k_u!}{\prod_{u<v} A_{uv}! \prod_i A_{uu}!!} \times \frac{\prod_r n_r!}{N}\binom{B-1}{N-1}^{-1}\frac{1}{N}. \tag{A.87}$$

To evaluate the change in the posterior probability, we just need to track the following properties of the network partition: $e_{\text{in}}, e_{\text{out}}, \{e_r\}, \{n_r\}, B$, which are no more difficult to obtain compared to the change in the modularity value. Therefore, we can in fact use the change in the posterior probability of the uniform PP model as the criterion that is used in the search phase of the Louvain algorithm. By doing so, the output of the modified Louvain algorithm is an approximate of the MAP solution of the uniform PP model.

We provide a Python implementation of the Louvain algorithm for finding the MAP estimate of the uniform PP model[1]. The main purpose of this implementation is to show how practical it is to adapt existed optimisation heuristics to perform inference based on statistically principled models. Moreover, refining the results obtained with the Louvain algorithm according to the uniform PP model can be used as a sanity check in terms of overfitting data. As we are going to see, if we start with the partitions given by modularity maximisation and replace the objective function with the posterior probability of the uniform PP model, the Louvain heuristic will continue to merge communities, signalising the noises that have been included when the modularity measure is used as the objective function.

We firstly consider synthetic networks with known community structures. In light of the fact that the underlying models of the modularity maximisation approach is the uniform PP model, we compare the performance of modularity maximisation and

[1]Code available in author's Github repository. The implementation of the Louvain heuristic is built on the python-louvain package [164].

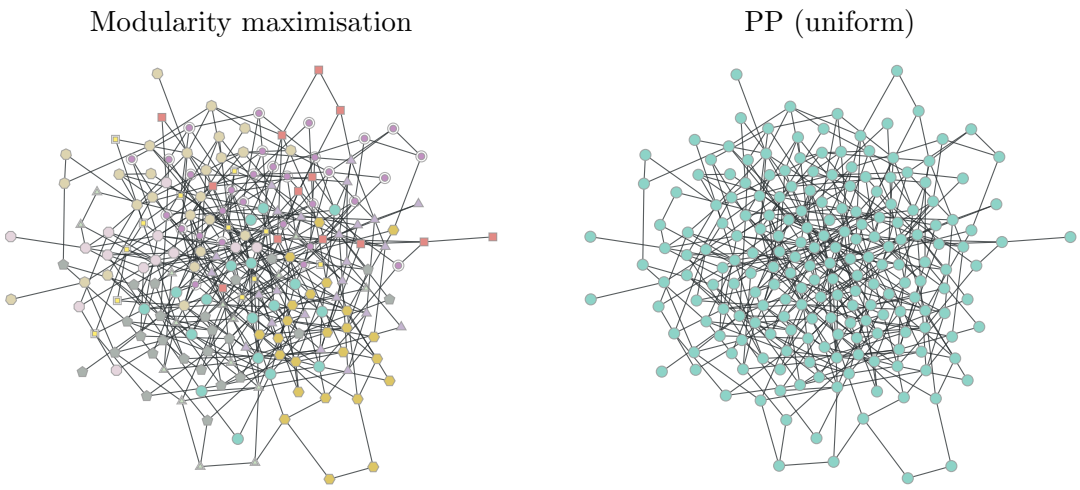


Figure A-2: Inferred community structures in a ER network with $N = 200$ and $\langle k \rangle = 5$. We run the Louvain algorithm with different objective function: (left) Newman-Girvan modularity maximisation and (right) the posterior probability of the uniform PP model.

the uniform PP model in a sample of the uniform PP model. Recall that we can parameterise the parameters of PP model as follows,

$$p_{\text{in}} = \frac{(1 + (B-1)\epsilon)\langle k \rangle}{N}, \quad p_{\text{out}} = \frac{(1-\epsilon)\langle k \rangle}{N}, \tag{A.88}$$

such that the ϵ parameters controls the strength of assortative structures. When $\epsilon = 0$, $p_{\text{in}} = p_{\text{out}}$, the PP model is equivalent to the ER model and there is no community structures in the networks sampled from the model. As shown in Fig A-2, running the Louvain algorithm for modularity maximisation gives us a partition with $B = 10$ communities, $Q = 0.461$. Then, start with this partition, we continue the Louvain heuristic but use the posterior likelihood of the PP model as the objective function. Refining the partition in this way leads to a partition in which all nodes are correctly put in a single community.

Then, we generate a network with two equal-size assortative communities. We set the average degree and the assorativity parameter as $\langle k \rangle = 5$ and $\epsilon = 0.85$ such that the assortative structure is above the detectability threshold. The network with true labelling as well as the inferred community structures given by the Louvain with modularity and the uniform PP model are given in the Fig. A-3. Having seen the overfitting behaviour in the random network, it is not surprised to see modularity maximisation finds an overly complicated partition here, in which the correct communities are sub-

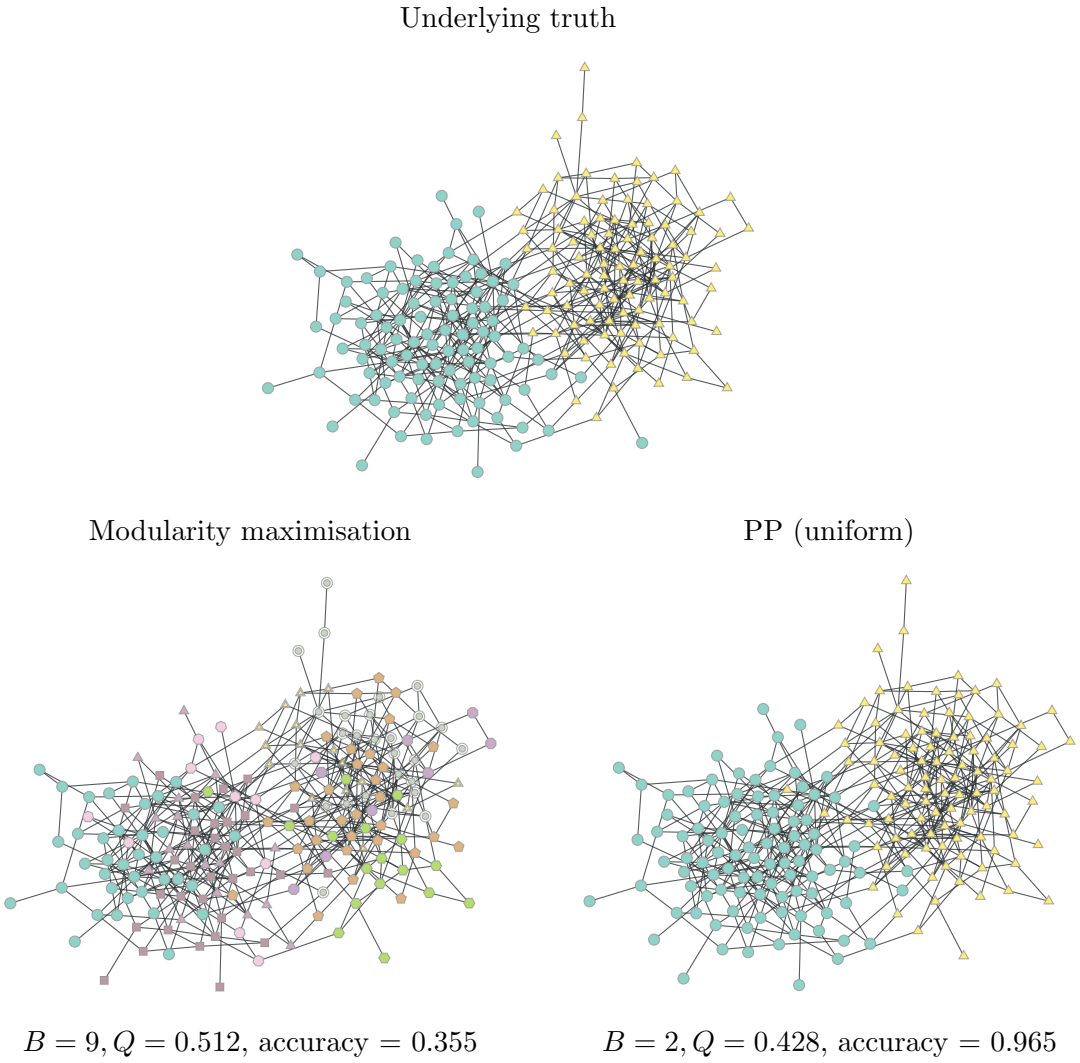


Figure A-3: Inferred community structures in a network generated from the uniform PP model with $B = 2$, $n_1 = n_2 = 100$, $\langle k \rangle = 5$. The strength of the assortative structure ϵ is defined as introduced in the text. We choose $\epsilon = 5$ such that the assortative structure is far above the detectability threshold. We run the Louvain algorithm with different objective functions: (left) Newman-Girvan modularity and (right) the posterior probability of the uniform PP model.

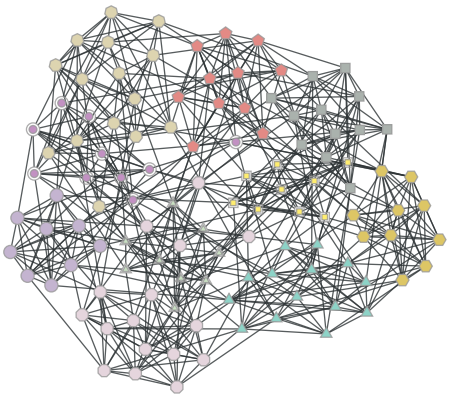
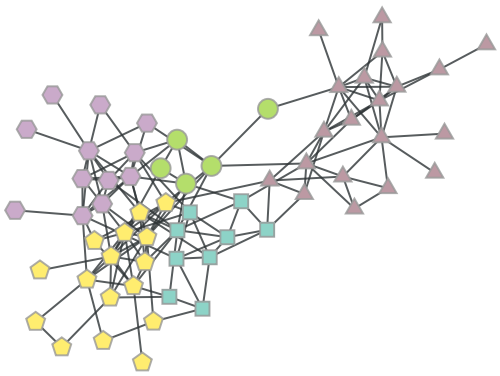


Figure A-4: Inferred community structures in the American college football network [11] with $B = 10, Q = 0.605$. Start from the partition given by the modularity maximisation approach, there is no room to refine according to the posterior probability of the uniform PP model.

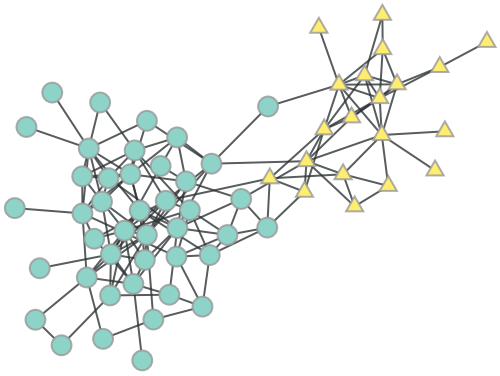
dividing into small communities. Using the PP model refinement leads to a partition which is almost identical to the underlying truth. We also test the Louvain algorithm with the uniform PP model refinement in empirical networks. In the American college football network, the maximum modularity approach identifies a partition with $B = 10$ communities, as shown in Fig. A-4. This partition cannot be further refined according to the the posterior probability of the uniform PP model. This is because the community structure is rather significant and each community has relatively small sizes. However, in general, refining the results given by the modularity maximisation according to the uniform PP model will lead to more conservative results. For example, in the bottlenose dolphins social network, modularity maximisation finds a partition with 5 communities and modularity value $Q = 0.529$, as given in Fig A-5. Refining this partition according to the posterior probability of the uniform PP model leads to the partition where four communities in the left are merged together.

Modularity maximisation



$B = 5, Q = 0.529$

PP (uniform)



$B = 2, Q = 0.373$

Figure A-5: Inferred community structures in the social network of the bottlenose dolphins [165]. The modularity maximisation approach identify 5 communities. Running the refinement according to the posterior probability of the uniform PP model reduce the number of communities to 2.

Appendix B

Supplementary materials for Chapter 4

B.1 Derivation of the logarithm of the joint probability distribution for DC-SBM in the clique network

Here we provide the derivation of the logarithm of the joint probability of DC-SBM for the clique network

$$\ln P(\boldsymbol{A}, \boldsymbol{k}, \boldsymbol{e}, \boldsymbol{b}) \approx (E - N) \ln B^* - (E + {B^*}^2/2) g\left(\frac{E}{E + {B^*}^2/2}\right), \tag{B.1}$$

where $g(x) = -x \ln x - (1-x)\ln(1-x)$. Firstly, because of the microcanonical nature of the model, we can decompose the joint probability as follows

$$\ln P(\boldsymbol{A}, \boldsymbol{k}, \boldsymbol{e}, \boldsymbol{b}) = \ln P(\boldsymbol{A}|\boldsymbol{k}, \boldsymbol{e}, \boldsymbol{b}^*) + \ln P(\boldsymbol{k}|\boldsymbol{e}, \boldsymbol{b}) + \ln P(\boldsymbol{e}|\boldsymbol{b}^*) + \ln P(\boldsymbol{b}^*), \tag{B.2}$$

with each term in the last equation as given below

$$P(\boldsymbol{A}|\boldsymbol{k}, \boldsymbol{e}, \boldsymbol{b}^*) = \frac{(2E/B^*)!!^{B^*}}{(2E/B^*)!^{B^*}} \times \frac{\prod_u k_u!}{\prod_{u<v} A_{uv}! \prod_u A_{uu}!!}, \tag{B.3}$$

$$P(\boldsymbol{b}^*) = \frac{(N/B^*)!^{B^*}}{N!} \times \binom{N-1}{B^*-1}^{-1} \frac{1}{N}, \tag{B.4}$$

$$P(\boldsymbol{e}|\boldsymbol{b}^*) = \left(\!\!\binom{B^*(B^*+1)/2}{E}\!\!\right)^{-1}, \tag{B.5}$$

$$P(\boldsymbol{k}|\boldsymbol{e}, \boldsymbol{b}^*) = \left(\!\!\binom{N/B^*}{2E/B^*}\!\!\right)^{-B^*}. \tag{B.6}$$

Now we just need to take the logarithm of these terms and make use of the Stirling's approximation $\ln x! \approx x \ln x - x$. Also note that we will drop all the constants that are independent of B^* because they do not affect the result of optimising the joint prob-

ability (therefore the description length) with respect to the number of communities B.

$$\begin{aligned}\ln P(\boldsymbol{A}|\boldsymbol{k},\boldsymbol{e},\boldsymbol{b}^*) &= B^* \ln\left((E/B^*)!2^{E/B^*}\right) - B^* \ln(2E/B^*)! \\ &= B^*\left[\frac{E}{B^*}\ln\frac{E}{B^*} - \frac{E}{B^*} + \frac{E}{B^*}\ln 2 - \frac{2E}{B^*}\ln\frac{2E}{B^*} + \frac{2E}{B^*}\right] \\ &= E\ln B^*. \end{aligned} \tag{B.7}$$

$$\begin{aligned}&\ln P(\boldsymbol{b}^*) = B^*\ln(N/B^*)! - \ln\frac{(N-1)!}{(N-B^*)!(B^*-1)!} \\ &= B^*\left[\frac{N}{B^*}\ln\frac{N}{B^*} - \frac{N}{B^*}\right] - [(N-1)\ln(N-1) - (N-B^*)\ln(N-B^*) - (B^*-1)\ln(B^*-1)] \\ &= -N\ln B^* - (N-1)g\left(\frac{N-B^*}{N-1}\right) \\ &= -N\ln B^*. \end{aligned} \tag{B.8}$$

Note that from the second to the last line to the last line above, we make use of the fact that $g(x) = -x\ln x - (1-x)\ln(1-x)$ is the binary entropy function which takes values at the order of $\mathcal{O}(1)$ [1].

$$\begin{aligned}\ln P(\boldsymbol{e}|\boldsymbol{b}^*) &= -\ln\frac{(E+B^*(B^*+1)/2-1)!}{(B^*(B^*+1)/2-1)!E!} \\ &= -(E+B^*(B^*+1)/2-1)\ln(E+B^*(B^*+1)/2-1)+ \\ &\qquad (B^*(B^*+1)/2-1)\ln(B^*(B^*+1)/2-1) - E\ln E \\ &= (E+{B^*}^2/2)g\left(\frac{E}{E+B^*/2}\right). \end{aligned} \tag{B.9}$$

In the last equation, we assume that $E, B^* \gg 1$ and only keep leading terms depending on B^*. Finally,

$$\begin{aligned}\ln P(\boldsymbol{k}|\boldsymbol{e},\boldsymbol{b}^*) &= -B^*\ln\frac{(N/B^*+2E/B^*-1)!}{(N/B^*-1)!(2E/B^*)!} \\ &= -B^*\Big[(N/B^*+2E/B^*-1)\ln(N/B^*+2E/B^*-1) - (N/B^*-1)\ln(N/B^*-1) \\ &\qquad - 2E/B^*\ln 2E/B^*\Big] \\ &= (N+2E-B^*)g\left(\frac{N/B^*}{N/B^*+2E/B^*-1}\right). \end{aligned} \tag{B.10}$$

[1] The binary entropy function must be non-negative and it reaches its maximum when $x = 1-x = 1/2$, which leads to $g(x) = \ln 2$, therefore $g(x) = \mathcal{O}(1)$ since $0 \le g(x) \le \ln 2$ for any values of B^*.

Again, since $g(x) = \mathcal{O}(1)$, the leading terms in $\ln P(\boldsymbol{k}|\boldsymbol{e}, \boldsymbol{b}^*)$ are constants independent of B^*. Combining equations (B.7)-(B.10), we reach the desired equation in equation (B.1).

B.2 Numerical estimate of the resolution limit of DC-SBM

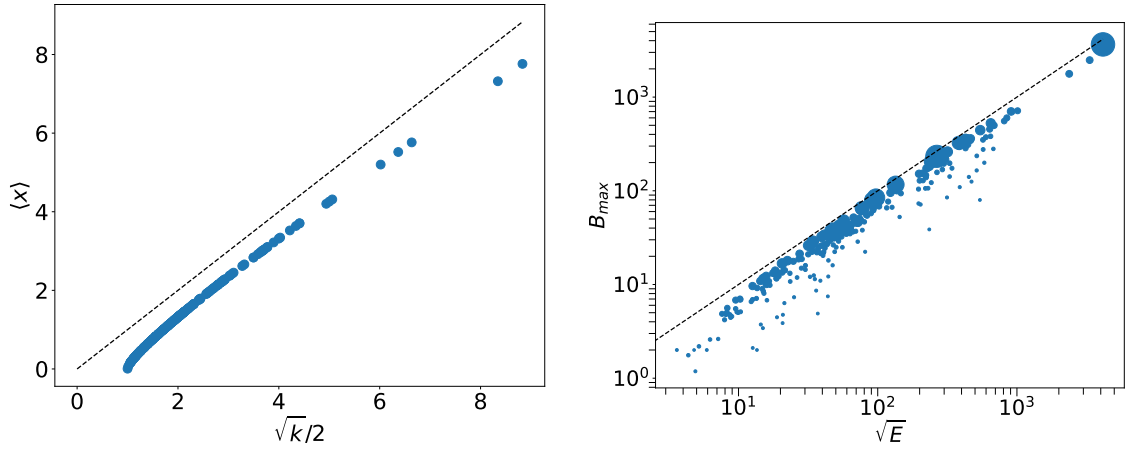


Figure B-1: We consider the empirical network corpus as detailed in Section B.6. The resolution limit of DC-SBM has the expression $B_{\max} = x(\langle k \rangle)\sqrt{N}$. (a) Numerical estimate of x (b) Numerical estimate of $B_{\max}$.

As explained in Section 4.1, the resolution limit of DC-SBM has the expression

$$B_{\max} = x(\langle k \rangle)\sqrt{N}, \tag{B.11}$$

where $x(\langle k \rangle)$ is the solution of the following equation

$$\langle k \rangle - 2 = 2x^2 \ln \frac{\langle k \rangle + x^2}{x^2}. \tag{B.12}$$

In the last equation, $\langle k \rangle$ is the average degree $\langle k \rangle = 2E/N$. Generally, $x(\langle k \rangle)$ does not permit analytical solution. However, we can compute numerical approximaten of $x(\langle k \rangle)$, thereby obtaining the resolution limit of DC-SBM. We do so for datasets in the empirical network corpus as detailed in Section B.6, using the Newton–Raphson method [136]. As can be seen from Fig B-1(a), $x(\langle k \rangle)$ rougly scales as the function $\sqrt{\langle k \rangle}/2$. In Fig B-1(b), the estimate of $B_{\max}$ seems to have the scaling $\mathcal{O}(\sqrt{E})$, which is consistent with the results in literature [63, 106]. The size of points in Fig B-1(b) is proportional to the average degree $\langle k \rangle$. It seems that the larger $\langle k \rangle$ is, the closer $B_{\max}$ is compared to the $\sqrt{E}$.

B.3 Compare the non-uniform PP model to DC-SBM in an empirical network corpus

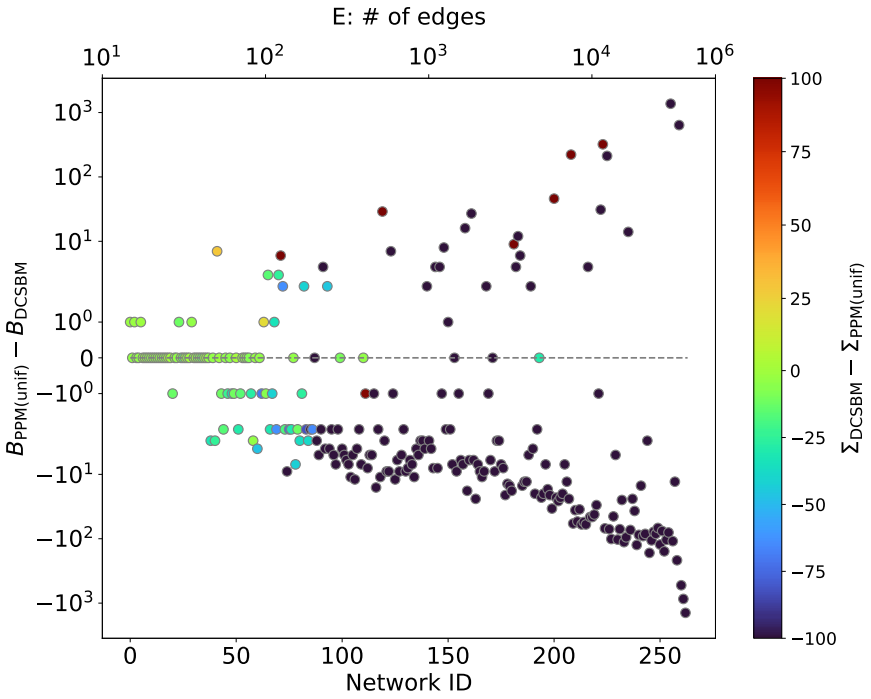


Figure B-2: Inferred number of communities given by the uniform PP model subtracted from that given by the single layer DC-SBM. Networks indices are ordered in the increasing order of network sizes (number of edges) and the colouring of points indicates the description length difference per edges (nats): $(\Sigma_{\mathrm{DCSBM}} - \Sigma_{\mathrm{PPM(unif)}})/E$.

B.4 Samples from the posterior distribution of the uniform PP and DC-SBM

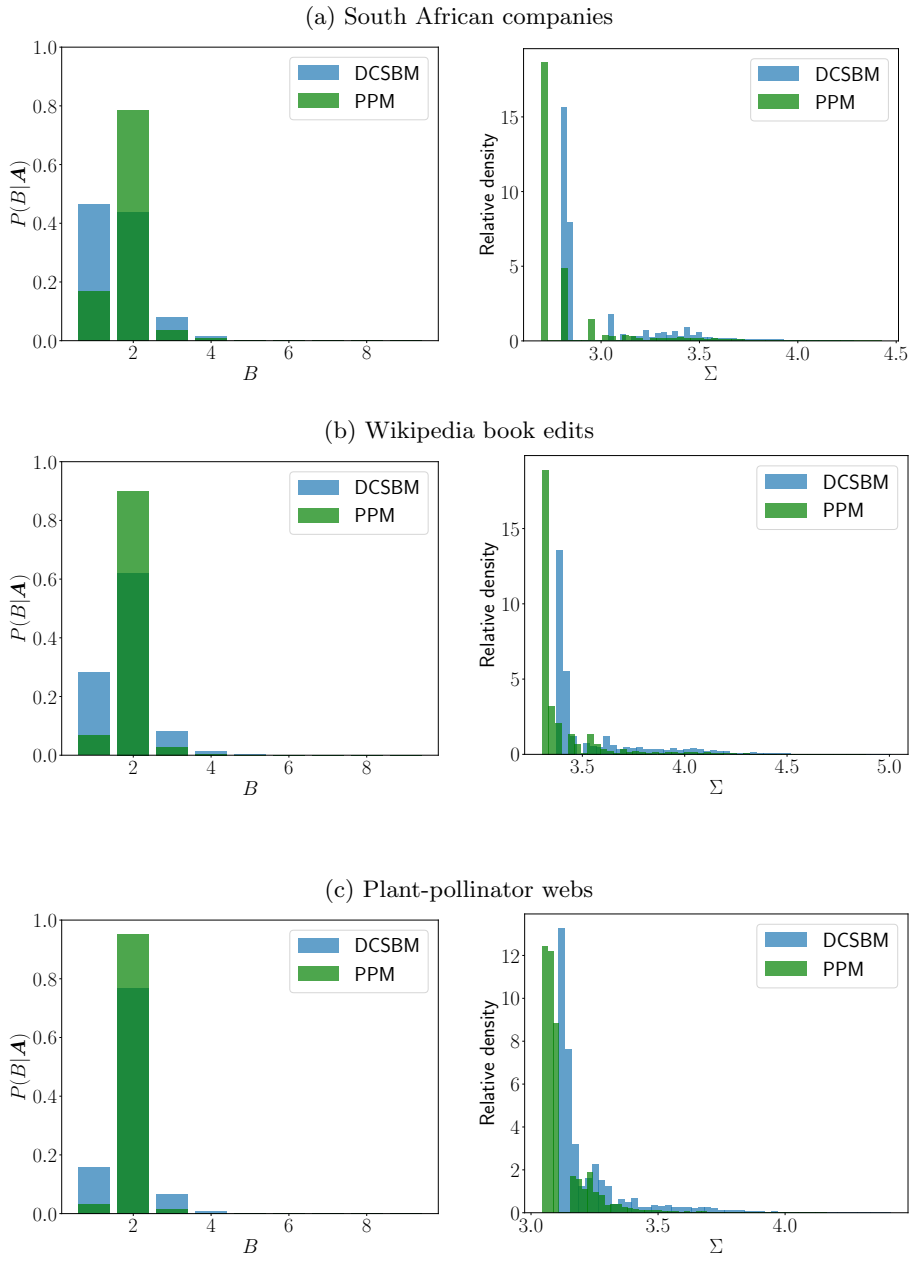


Figure B-3: Posterior distribution of the number of communities and description length obtained with the uniform PP model and DC-SBM for three bipartite networks (a) South African companies [140] (b) Wikipedia book edits [141] (c) Plant-pollinator web in Safariland [142].

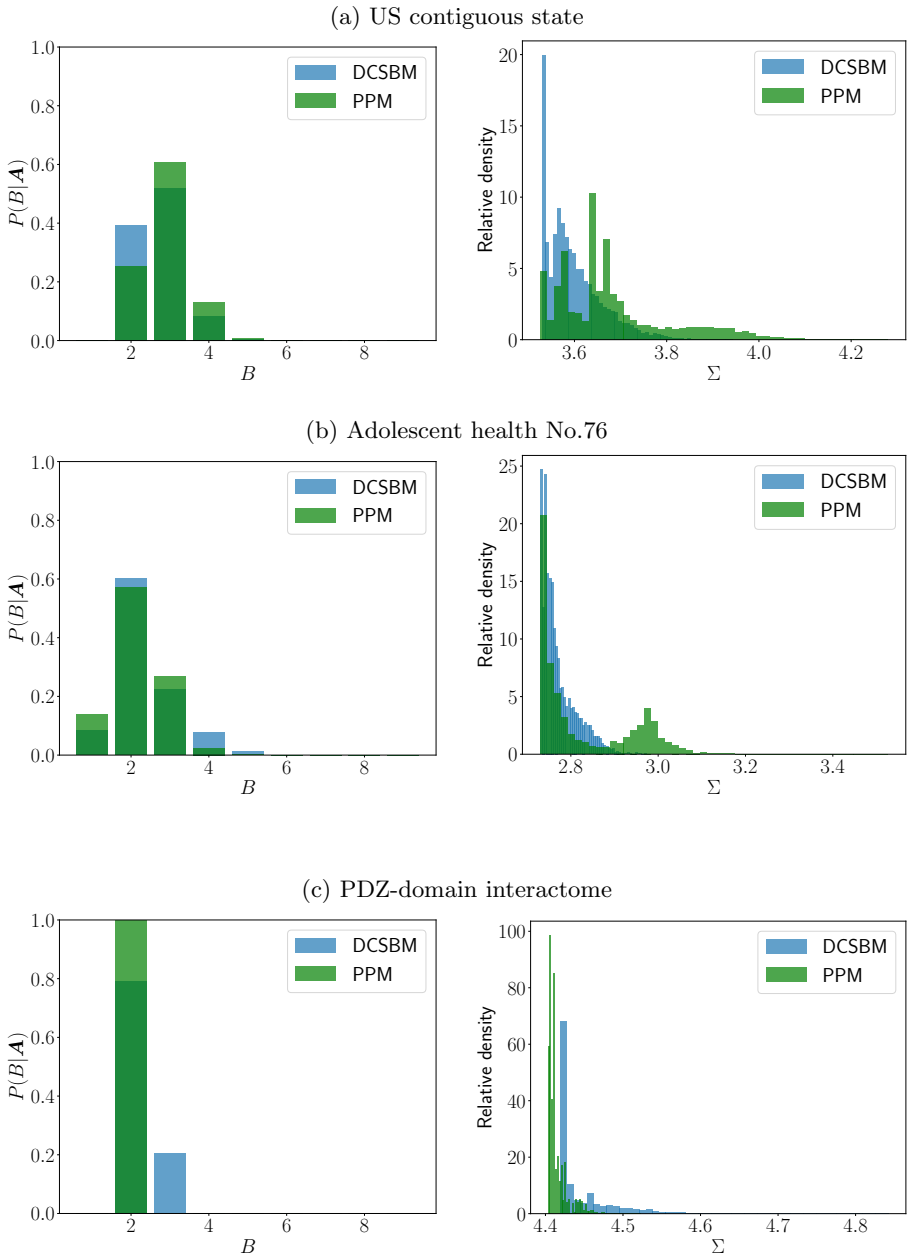


Figure B-4: Posterior distribution of the number of communities and description length obtained with the uniform PP model and DC-SBM for the network of (a) US contiguous [144] (b) No.76 dataset from the Adolescent health dataset [137] (c) PDZ-domain interactive [146].

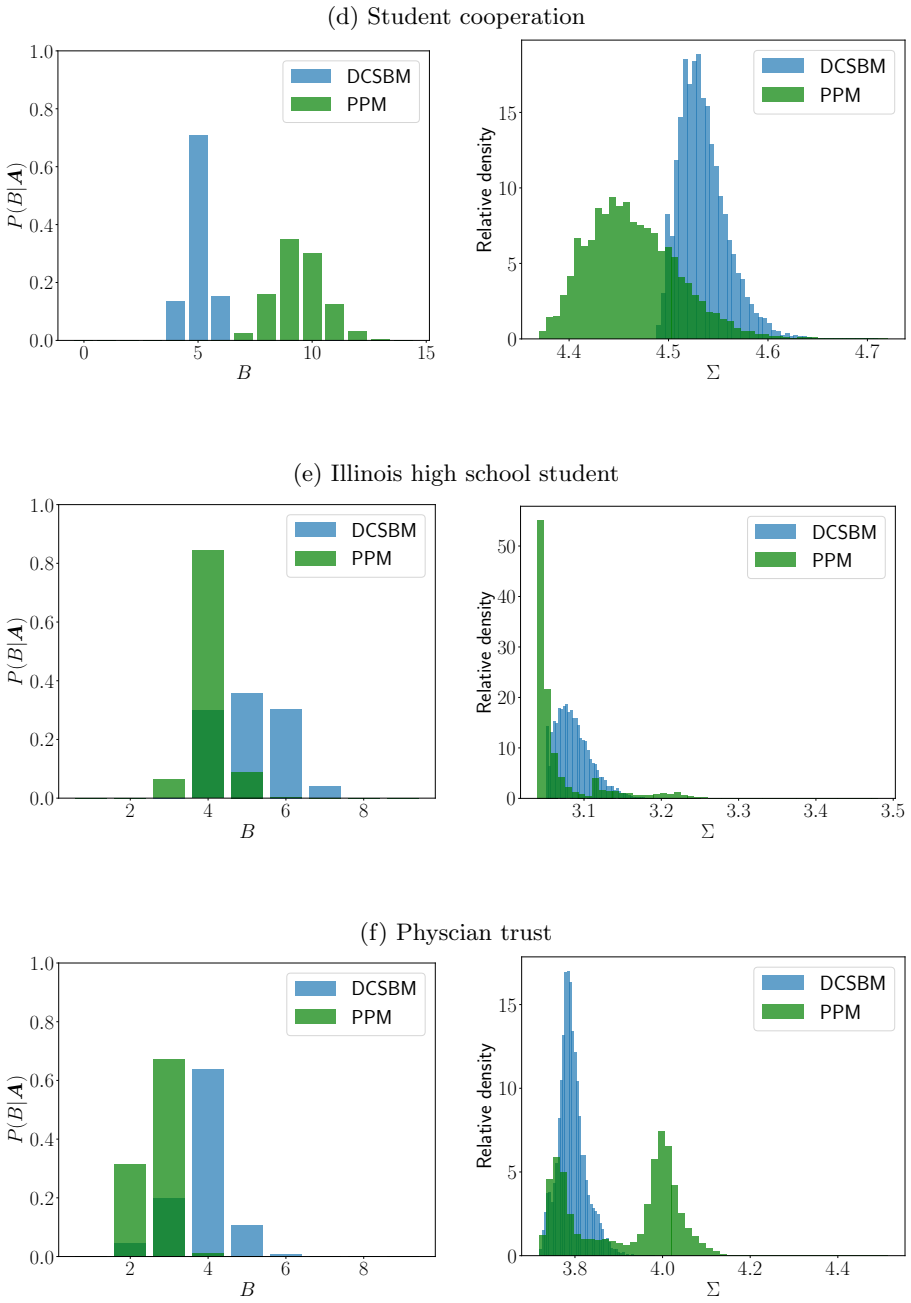


Figure B-4: Posterior distribution of the number of communities and description length obtained with the uniform PP model and DC-SBM for the network of (d) student cooperation [143](e) high school [166] (f) physician trust [167].

(g) American college football

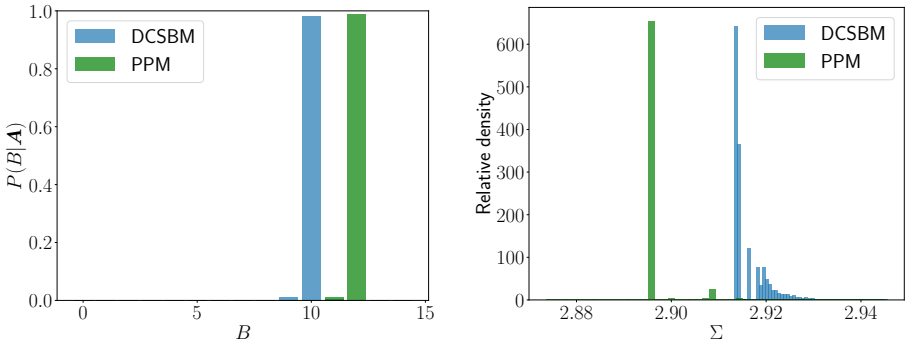


Figure B-4: Posterior distribution of the number of communities and description length obtained with the uniform PP model and Nested DC-SBM for the network of (g) American college football [11]

B.5 Samples from the posterior distribution of the uniform PP model and Nested DC-SBM

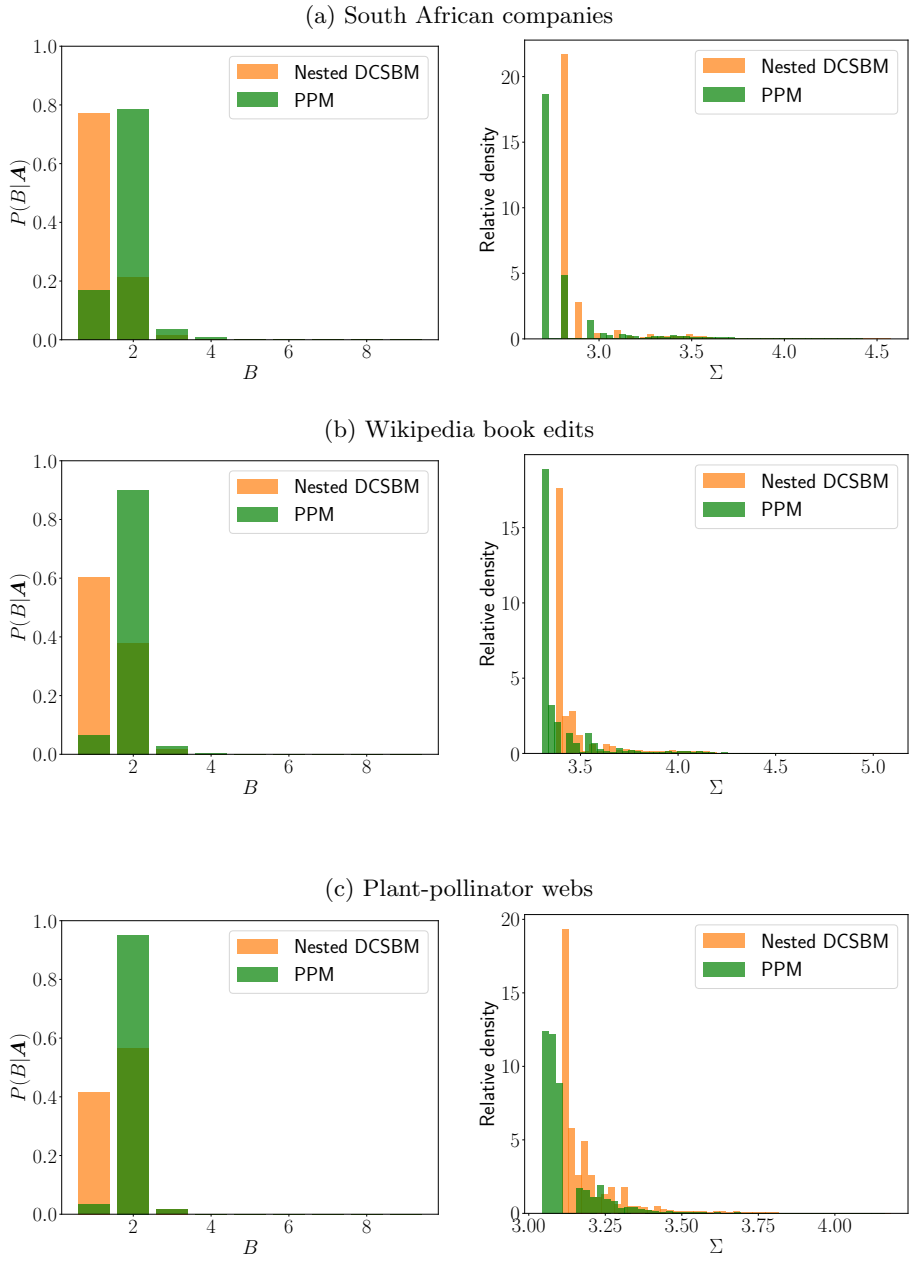


Figure B-5: Posterior distribution of the number of communities and description length obtained with the uniform PP model and Nested DC-SBM for three bipartite networks (a) South African companies [140] (b) Wikipedia book edits [141] (c) Plant-pollinator web in Safariland [142].

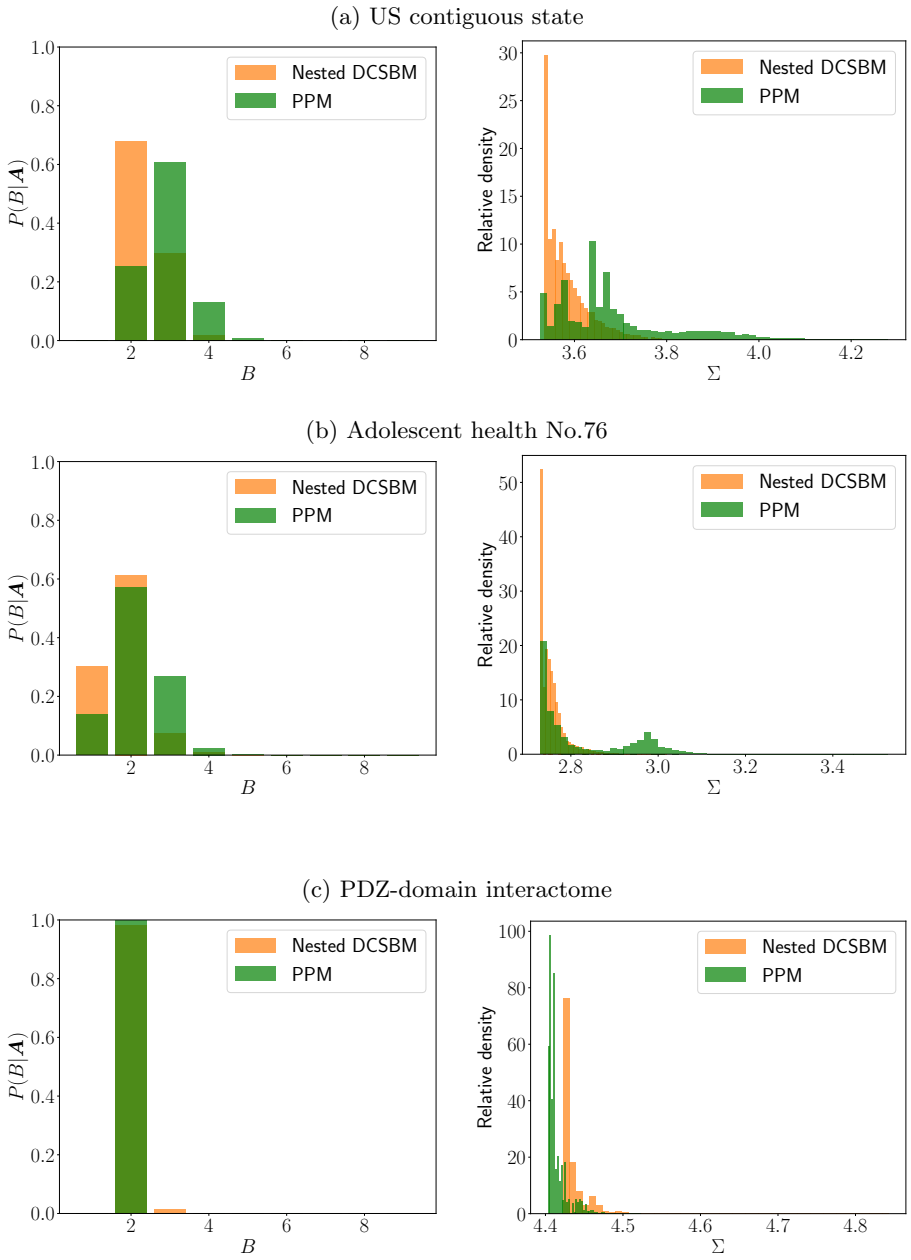
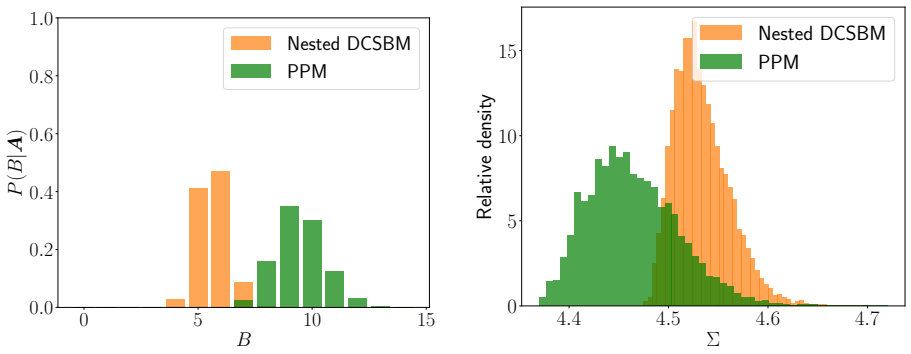
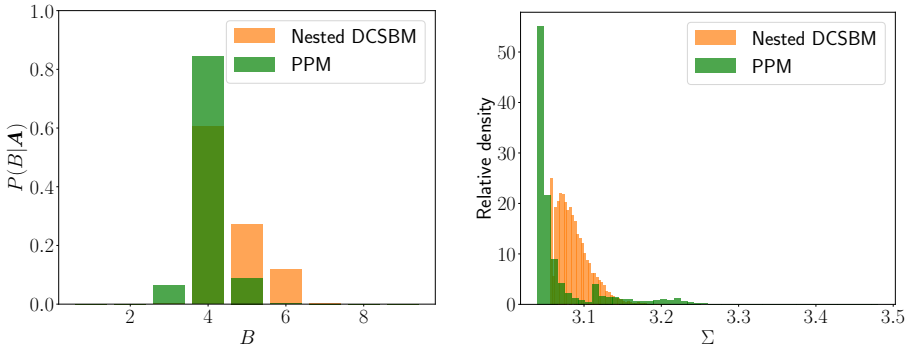


Figure B-6: Posterior distribution of the number of communities and description length obtained with the uniform PP model and DC-SBM for the network of (a) US contiguous [144] (b) No.76 dataset from the Adolescent health dataset [137] (c) PDZ-domain interactive [146].

(d) Student cooperation



(e) Illinois high school student



(f) Physcian trust

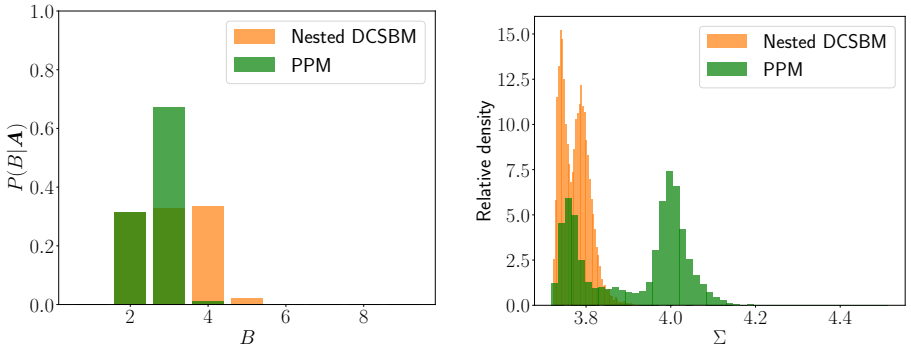


Figure B-6: Posterior distribution of the number of communities and description length obtained with the uniform PP model and Nested DC-SBM for the network of (d) student cooperation [143](e) high school [166] (f) physician trust [167].

(g) American college football

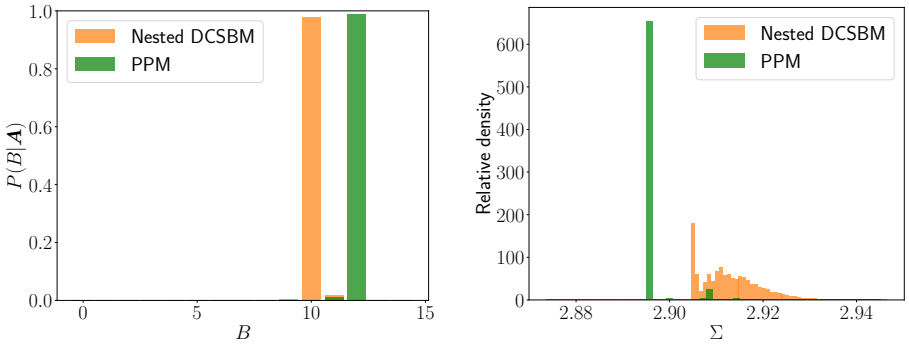


Figure B-6: Posterior distribution of the number of communities and description length obtained with the uniform PP model and Nested DC-SBM for the network of (g) American college football [11]

B.6 Details of the network corpus

Name	Description	N	E
sa_companies	A bipartite network of the affiliations between a small group of individuals and five important companies in South African finance, around 1919 [140].	13	11
ambassador (1985_1989)	A temporal network representing snapshots of relationships among individuals directly or indirectly associated with Philippines Ambassador Residence Bombing 2000, Jakarta [168].	19	11
edit_wikibooks (aa)	Two bipartite user-page networks extracted from Wikipedia, about books [141].	23	23
internet_top_pop (Aarnet)	Assorted snapshots of internet graph at the Point of Presence (PoP) level (which lies between the IP and AS levels), collected from around the world and at various times [169].	24	19
florentine_families	Multiplex network with 2 edge types representing marriage alliances and business relationships between Florentine families during the Italian Renaissance [170].	27	15
plant_pol_vazquez (Safariland)	Eight bipartite networks of plants and pollinators, from the Nahuel Huapi National Park and surrounding areas in Rio Negro, Argentina, from September 1999 to Feburary 2000 [142].	35	31
moreno_taro	A network of gift-giving relationships (taro exchange) among households in a Papuan village [171].	39	22
edit_wiktionary (aa)	Three bipartite user-page networks extracted from Wiktionary, for French, German, and English [141].	51	32
new_guinea_tribes	A network of friendships among tribes of Gahuku-Gama alliance structure of the Eastern Central Highlands region in New Guinea [172].	58	16
dutch_school (klas12b-net-1)	A series of snapshots of the friendships among freshmen at secondary school in The Netherlands, in 2003-2004 [173].	63	26
november17	A network representing connections among members of the November 17 (N17) Greek terrorist group [174].	66	22
moviegalaxies (1)	Social graphs for over 700 movies from the moviegalaxies [175].	68	21
rhesus_monkey	Grooming interactions among a group of wild adult rhesus monkeys (Macaca mulatta) in Cayo Santiago, during a two month period in 1963 [176].	69	16
montreal	Network representing relationships between gangs, obtained from Montreal Police's central intelligence database, spanning 2004 to 2007 [177].	75	29
karate (77)	Network of friendships among members of a university karate club [178].	77	34
dutch_criticism	A network of criticisms among Dutch literary authors in 1976 [179].	80	35
add_health (comm3)	A directed network of friendships obtained through a social survey of high school students in 1994 [124].	91	32
kangaroo	Dominance relations among a group of free-ranging grey kangaroos (Macropus giganteus) [180].	91	17
ceo_club	A bipartite network of the memberships of chief executive officers and the social organizations (clubs) to which they belong, from the Minneapolis-St. Paul area [181].	95	40

elite	A small bipartite network of the affiliations among elite individuals and the corporate, museum, university boards, or social clubs to which they belonged, from 1962 [182].	99	44
zebras	Social interactions among a group of wild Grevy's zebras (Equus grevyi), observed in Mpala Ranch in Kenya in 2002 [82].	105	23
add_health (comm77)	A directed network of friendships obtained through a social survey of high school students in 1994 [137].	106	25
contiguous_usa	A network of contiguous states in the USA, in which each state is a node and two nodes are connected if they share a land-based geographic border [144].	107	49
terrorists_911	Network of individuals and their known social associations, centered around the hijackers that carried out the September 11th, 2001 terrorist attacks [183].	152	62
high_tech_company	Multiplex network of 3 edge types representing relationships (advice, friendship, and "reports to") between managers of a high-tech company [184].	159	21
dolphins	An undirected social network of frequent associations observed among 62 dolphins (Tursiops) in a community living off Doubtful Sound, New Zealand, from 1994-2001 [123].	159	62
revolution	A bipartite network of the memberships of notable people and organizations, from the American Revolution (1765-1783) between users and groups on YouTube, extracted from a larger YouTube network in 2007 [5].	160	141
blumenau_drug	A network of drug-drug interactions, extracted from 18 months of electronic health records (EHRs) from the city of Blumenau in Southern Brazil [185].	181	75
board_directors (net2m_2002-05-01)	224 networks of the affiliations among board directors due to sitting on common boards of Norwegian public limited companies (as of 5 August 2009), from May 2002 onward, in monthly snapshots through August 2011 [186].	184	179
add_health (comm76)	A directed network of friendships obtained through a social survey of high school students in 1994 [137].	185	43
cattle	Dominance interactions among a group of dairy cattle at the Iberia Livestock Experiment Station in Jenerette, Louisiana [187]	205	28
interactome_pdz	A network of PDZ-domain-mediated protein–protein binding interactions, extracted from the PDZBase database [146].	209	161
add_health (comm1)	A directed network of friendships obtained through a social survey of high school students in 1994 [137].	220	69
bison	Dominance relations among a group of American bisons in the National Bison Range in Moiese Montana, observed in 1972 [188].	222	26
marvel_partnerships	A network of partnerships among characters in the Marvel comic book universe [189].	224	181
fresh_webs (AkatoreA)	A set of 26 networks of trophic-level species interactions in streams in New Zealand, Maine and North Carolina [190].	227	84
swingers	A bipartite sexual affiliation network representing "swing unit" couples (one node per couple) and the parties they attended [191]	232	96
moreno_sheep	Dominance interactions among a group of female bighorn sheep (Ovis canadensis) from the National Bison Range in western Montana USA, over a 27 month period ending in 1984 [192].	235	28

train_terrorists	A network of associations among the terrorists involved in the 2004 Madrid train bombing, as reconstructed from press stories after-the-fact [193].	243	64
7th_graders	A small multiplex network of friendships among 29 seventh grade students in Victoria, Australia [194].	250	29
lesmis	The network of scene coappearances of characters in Victor Hugo's novel "Les Miserables [144].	254	77
student_cooperation	Network of cooperation among students in the "Computer and Network Security" course at Ben-Gurion University, in 2012 [143].	256	141
highschool	A network of friendships among male students in a small high school in Illinois from 1958 [195].	274	70
add_health (comm63)	A directed network of friendships obtained through a social survey of high school students in 1994 [137].	282	96
macaque_neural	A network of cortical regions in the Macaque cortex [196].	313	47
windsurfers	A network of interpersonal contacts among windsurfers in southern California during the Fall of 1986 [197].	336	43
add_health (comm70)	A directed network of friendships obtained through a social survey of high school students in 1994 [137].	344	76
add_health (comm2)	A directed network of friendships obtained through a social survey of high school students in 1994 [137].	348	103
game_thrones	Network of coappearances of characters in the Game of Thrones series, by George R. R. Martin, and in particular coappearances in the book "A Storm of Swords [198].	352	107
cs_department	Multiplex network consisting of 5 edge types corresponding to online and offline relationships (Facebook, leisure, work, co-authorship, lunch) between employees of the Computer Science department at Aarhus [199].	353	61
kidnappings	Bipartite network of members of the Abu Sayyaf Group in the Philippines, and the kidnapping events they were involved in [200].	357	285
add_health (comm71)	A directed network of friendships obtained through a social survey of high school students in 1994 [137].	358	74
add_health (comm6)	A directed network of friendships obtained through a social survey of high school students in 1994 [137].	378	108
college_freshmen	A small network of friendships among freshmen at Dutch College in 1994-1995 [201].	422	32
adjnoun	A network of word adjacencies of common adjectives and nouns in the novel "David Copperfield" by Charles Dickens [202].	425	112
london_transport	Multiplex network with 3 edge types representing links within the three layers of London train stations: Underground, Overground and DLR [203].	430	369
openstreetmap (01-AL-cities-street_networks:0100124_Abbeville)	The road network for the entire United States, as extracted from the OpenStreetMap project in c. 2018. [147].	434	351
polbooks	A network of books about U.S. politics published close to the 2004 U.S. presidential election, and sold by Amazon.com [121].	441	105
ecoli_transcription (v1.0)	Network of operons and their pairwise interactions, via transcription factor-based regulation, within the bacteria Escherichia coli [138].	456	329

physician_trust	A network of trust relationships among physicians in four midwestern (USA) cities in 1966 [167].	465	117
sp_kenyan_households	A network of proximity contacts measured between members of 5 households of rural Kenya, between April 24 and May 12, 2012 [204].	504	47
ugandan_village (friendship-1)	Complete friendship and health advice social networks among households in 17 rural villages bordering Lake Victoria in Mayuge District, Uganda in 2013 [205]	547	202
webkb (webkb_wisconsin_link1)	Web graphs crawled from four Computer Science departments in 1998, with each page manually classified into one of 7 categories: course, department, faculty, project, staff, student, or other [206].	553	280
football	A network of American football games between Division IA colleges during regular season Fall 2000 [11].	613	115
copenhagen (sms)	A network of social interactions among university students within the Copenhagen Networks Study, over a period of four weeks, sampled every 5 minutes [207].	628	457
add_health (comm37)	A directed network of friendships obtained through a social survey of high school students in 1994 [137]	694	358
add_health (comm5)	A directed network of friendships obtained through a social survey of high school students in 1994 [137]	730	157
sp_office	A temporal network of contacts between individuals, measured in an office building in France, from June 24 to July 3, 2013 [208].	755	92
add_health (comm55)	A directed network of friendships obtained through a social survey of high school students in 1994 [137]	800	331
add_health (comm8)	A directed network of friendships obtained through a social survey of high school students in 1994 [137]	809	204
add_health (comm67)	A directed network of friendships obtained through a social survey of high school students in 1994 [137]	913	439
netscience	A coauthorship network among scientists working on network science, from 2006 [202].	914	379
add_health (comm9)	A directed network of friendships obtained through a social survey of high school students in 1994 [137]	1004	248
law_firm	Multiplex network with 3 edge types representing relationships (coworkers, friendship, advice) between partners and associates of a corporate law firm [209].	1008	71
yeast_transcription	Network of operons and their pairwise interactions, via transcription factor-based regulation, within the yeast Saccharomyces cerevisiae [210].	1065	664
add_health (comm4)	A directed network of friendships obtained through a social survey of high school students in 1994 [137]	1136	281
sp_hospital	This dataset contains the temporal network of contacts between patients, patients and health-care workers (HCWs) and among HCWs in a hospital ward in Lyon, France, from Monday, December 6, 2010 at 1:00 pm to Friday, December 10, 2010 at 2:00 pm [211].	1139	75
macaques	Dominance interactions among a group of adult female Japanese macaques (Macaca fuscata fuscata), observed during a non-mating season in 1976 [212].	1167	62
add_health (comm18)	A directed network of friendships obtained through a social survey of high school students in 1994 [137].	1189	284

plant_pol_kato	A bipartite network of plants and pollinators from Kyoto University Forest of Ashu, Japan, from 1984 to 1987 [213].	1205	768
unicodelang	A bipartite network of languages and the countries in which they are spoken, as estimated by Unicode [120]	1249	858
euroroad	A network of international "E-roads", mostly in Europe [149].	1305	1039
add_health (comm78)	A directed network of friendships obtained through a social survey of high school students in 1994 [137]	1333	430
crime	A network of associations among suspects, victims, and/or witnesses involved in crimes in St. Louis in the 1990s [214].	1377	1263
add_health (comm56)	A directed network of friendships obtained through a social survey of high school students in 1994 [137]	1393	444
add_health (comm72)	A directed network of friendships obtained through a social survey of high school students in 1994 [137]	1398	352
celegans_interactomes (wi2007)	Ten networks of protein-protein interactions in Caenorhabditis elegans (nematode), from yeast two-hybrid experiments, biological process maps, literature curation, orthologous interactions, and genetic interactions [215].	1500	1108
add_health (comm53)	A directed network of friendships obtained through a social survey of high school students in 1994 [137]	1510	579
add_health (comm21)	A directed network of friendships obtained through a social survey of high school students in 1994 [137]	1531	377
human_brains	Networks of neural interactions extracted from human patients using the Magnetic Resonance One-Click Pipeline (MROCP), where nodes are voxels of neural tissue and edges represent connections by single fibers [216].	1563	116
add_health (comm11)	A directed network of friendships obtained through a social survey of high school students in 1994 [137]	1590	411
add_health (comm31)	A directed network of friendships obtained through a social survey of high school students in 1994 [137]	1635	728
add_health (comm51)	A directed network of friendships obtained through a social survey of high school students in 1994 [137]	1665	676
add_health (comm74)	A directed network of friendships obtained through a social survey of high school students in 1994 [137]	1677	654
add_health (comm7)	A directed network of friendships obtained through a social survey of high school students in 1994 [137]	1698	437
sp_high_school_new (2011)	These datasets contain the temporal network of contacts between students in a high school in Marseilles, France [217].	1709	126
add_health (comm80)	A directed network of friendships obtained through a social survey of high school students in 1994 [137]	1745	594
residence_hall	A network of friendships among students living in a residence hall at Australian National University (date unknown) [218].	1839	217
add_health (comm65)	A directed network of friendships obtained through a social survey of high school students in 1994 [137]	1869	557
eu_procurements_alt (AT_2008)	These 234 networks represent the annual national public procurement markets of 26 European countries from 2008-2016, inclusive [218].	1921	1684

add_health (comm38)	A directed network of friendships obtained through a social survey of high school students in 1994 [137]	1925	521
interactome_yeast	A network of protein-protein binding interactions among yeast proteins [219].	1948	1458
facebook_friends	A small anonymized Facebook ego network, from April 2014 [220].	1954	329
edit_wikiquote (af)	A bipartite user-page network extracted from Wikiquotes [141].	1956	1119
celegans_metabolic	List of edges comprising the metabolic network of the nematode C. elegans [139].	2025	453
add_health (comm26)	A directed network of friendships obtained through a social survey of high school students in 1994 [137]	2066	551
foodweb_baywet	Networks of carbon exchanges among species in the cypress wetlands of South Florida [221].	2075	128
add_health (comm19)	A directed network of friendships obtained through a social survey of high school students in 1994 [137]	2096	492
contact	A network of human proximities, as measured by carried wireless devices [222]	2124	274
celegansneural	A network representing the neural connections of the Caenorhabditis elegans nematode [150].	2148	297
sp_hypertext (contacts)	The temporal network of contacts among attendees of the ACM Hypertext 2009 conference, which spanned 2.5 days of time [223].	2196	113
fullerene_structures (C1500)	Fifteen networks of carbon atoms and the atomic bonds that connect them within molecules of fullerenes, from 60 atoms up to 6000 atoms [224]	2250	1500
kegg_metabolic (aae)	109 metabolic networks of various species, as extracted from the Kyoto Encyclopedia of Genes and Genomes (KEGG) database in March 2006 [225].	2296	880
add_health (comm13)	A directed network of friendships obtained through a social survey of high school students in 1994 [137]	2350	652
facebook_organizations (S1)	Six networks of friendships among users on Facebook who indicated employment at one of the target corporation [226]	2369	320
faa_routes	A network of air traffic routes, from the FAA (Federal Aviation Administration) National Flight Data Center (NFDC) preferred routes database (www.fly.faa.gov) [227].	2408	1226
foodweb_little_rock	A food web among the species found in Little Rock Lake in Wisconsin [228].	2434	183
add_health (comm22)	A directed network of friendships obtained through a social survey of high school students in 1994 [137]	2449	612
ego_social (facebook_0)	Ego networks associated with a set of accounts of three social media platforms (Facebook, Google+, and Twitter) [229].	2514	324
add_health (comm27)	A directed network of friendships obtained through a social survey of high school students in 1994 [137]	2530	1152
add_health (comm29)	A directed network of friendships obtained through a social survey of high school students in 1994 [137]	2534	569
add_health (comm25)	A directed network of friendships obtained through a social survey of high school students in 1994 [137]	2567	790
add_health (comm14)	A directed network of friendships obtained through a social survey of high school students in 1994 [137]	2688	562
add_health (comm62)	A directed network of friendships obtained through a social survey of high school students in 1994 [137]	2737	1040

add_health (comm30)	A directed network of friendships obtained through a social survey of high school students in 1994 [137]	2740	718
jazz_collab	The network of collaborations among jazz musicians, and among jazz bands, extracted from The Red Hot Jazz Archive digital database, covering bands that performed between 1912 and 1940 [230].	2742	198
add_health (comm45)	A directed network of friendships obtained through a social survey of high school students in 1994 [137]	2745	921
sp_infectious	This dataset contains the daily dynamic contact networks collected during the Infectious SocioPatterns event that took place at the Science Gallery in Dublin, Ireland, during the artscience exhibition INFECTIOUS: STAY AWAY [223]	2765	410
add_health (comm10)	A directed network of friendships obtained through a social survey of high school students in 1994 [137]	2795	678
add_health (comm12)	A directed network of friendships obtained through a social survey of high school students in 1994 [137]	2805	581
malaria_genes (HVR_1)	Networks of recombinant antigen genes from the human malaria parasite P. falciparum [231].	2812	307
add_health (comm64)	A directed network of friendships obtained through a social survey of high school students in 1994 [137]	2861	694
add_health (comm66)	A directed network of friendships obtained through a social survey of high school students in 1994 [137]	2865	644
add_health (comm23)	A directed network of friendships obtained through a social survey of high school students in 1994 [137]	2897	667
eu_airlines	A multiplex network of airline routes among European airports, where each of the 37 edge types represents routes by a different airline [232].	2953	417
add_health (comm24)	A directed network of friendships obtained through a social survey of high school students in 1994 [137]	3003	849
add_health (comm54)	A directed network of friendships obtained through a social survey of high school students in 1994 [137]	3037	1035
interactome_stelzl	A network of human proteins and their binding interactions [233].	3106	1615
add_health (comm35)	A directed network of friendships obtained through a social survey of high school students in 1994 [137]	3151	851
email_company	A network of emails among employee email addresses at a mid-sized manufacturing company [234].	3250	167
add_health (comm59)	A directed network of friendships obtained through a social survey of high school students in 1994 [137]	3339	971
add_health (comm32)	A directed network of friendships obtained through a social survey of high school students in 1994 [137]	3440	853
add_health (comm16)	A directed network of friendships obtained through a social survey of high school students in 1994 [137]	3462	778
add_health (comm57)	A directed network of friendships obtained through a social survey of high school students in 1994 [137]	3480	1180
add_health (comm83)	A directed network of friendships obtained through a social survey of high school students in 1994 [137]	3583	1260

add_health (comm69)	A directed network of friendships obtained through a social survey of high school students in 1994 [137]	3678	891
add_health (comm60)	A directed network of friendships obtained through a social survey of high school students in 1994 [137]	3699	1131
add_health (comm81)	A directed network of friendships obtained through a social survey of high school students in 1994 [137]	3795	1290
add_health (comm68)	A directed network of friendships obtained through a social survey of high school students in 1994 [137]	3871	1385
add_health (comm39)	A directed network of friendships obtained through a social survey of high school students in 1994 [137]	3915	987
add_health (comm84)	A directed network of friendships obtained through a social survey of high school students in 1994 [137]	4003	1545
add_health (comm82)	A directed network of friendships obtained through a social survey of high school students in 1994 [137]	4020	921
add_health (comm20)	A directed network of friendships obtained through a social survey of high school students in 1994 [137]	4075	910
add_health (comm79)	A directed network of friendships obtained through a social survey of high school students in 1994 [137]	4154	1190
new_zealand_collab	A network of scientific collaborations among institutions in New Zealand	4246	1463
add_health (comm47)	A directed network of friendships obtained through a social survey of high school students in 1994 [137]	4378	985
add_health (comm15)	A directed network of friendships obtained through a social survey of high school students in 1994 [137]	4391	1062
add_health (comm75)	A directed network of friendships obtained through a social survey of high school students in 1994 [137]	4396	994
add_health (comm61)	A directed network of friendships obtained through a social survey of high school students in 1994 [137]	4399	1710
celegans_2019 (male_chemical)	Networks among neurons of both the adult male and adult hermaphrodite worms C. elegans, constructed from electron microscopy series, to include directed edges (chemical) and undirected (gap junction), and spanning including nodes for muscle and non-muscle end organs [235].	4500	559
add_health (comm28)	A directed network of friendships obtained through a social survey of high school students in 1994 [137]	4715	1136
add_health (comm33)	A directed network of friendships obtained through a social survey of high school students in 1994 [137]	4833	1974
add_health (comm42)	A directed network of friendships obtained through a social survey of high school students in 1994 [137]	4931	1405
us_agencies (alabama)	50 networks, one for each U.S. state, representing the web-based links between their associated government agencies websites [236].	4983	1115
add_health (comm48)	A directed network of friendships obtained through a social survey of high school students in 1994 [137]	4990	1171
add_health (comm44)	A directed network of friendships obtained through a social survey of high school students in 1994 [137]	5096	1127

route_views (19971108)	733 daily network snapshots denoting BGP traffic among autonomous systems (ASs) on the Internet, from the Oregon Route Views Project, spanning 8 November 1997 to 2 January 2000 [237].	5156	3015
add_health (comm43)	A directed network of friendships obtained through a social survey of high school students in 1994 [137]	5242	1638
add_health (comm17)	A directed network of friendships obtained through a social survey of high school students in 1994 [137]	5337	1218
add_health (comm52)	A directed network of friendships obtained through a social survey of high school students in 1994 [137]	5438	1719
uni_email	A network representing the exchange of emails among members of the Rovira i Virgili University in Spain, in 2003 [238].	5451	1133
add_health (comm58)	A directed network of friendships obtained through a social survey of high school students in 1994 [137]	5605	1703
add_health (comm34)	A directed network of friendships obtained through a social survey of high school students in 1994 [137]	5748	1605
add_health (comm46)	A directed network of friendships obtained through a social survey of high school students in 1994 [137]	5749	1519
sp_high_school (proximity)	These data sets correspond to the contacts and friendship relations between students in a high school in Marseilles, France, in December 2013, as measured through several techniques [239].	5818	327
sp_primary_school (day_1)	Two temporal networks of contacts among students and teachers at a primary school in Lyon, France, on consecutive days of in October 2009 [240].	5899	236
interactome_vidal	A network of human proteins and their binding interactions [241].	6007	2783
messal_shale	A network of feeling links among taxa based on the 48 million years old uppermost early Eocene Messel Shale [242].	6395	700
interactome_figeys	A network of human proteins and their binding interactions [243].	6418	2217
physics_collab (pierreAuger)	Two multiplex networks of coauthorships among the Pierre Auger Collaboration of physicists (2010-2012) and among researchers who have posted preprints on arXiv [244].	6426	475
wiki_science	A network of scientific fields, extracted from the English Wikipedia in early 2020 [245]	6517	677
power	A network representing the Western States Power Grid of the United States, in which nodes are transforms or power relay points and two nodes are connected if a power line runs between them [150].	6594	4941
add_health (comm73)	A directed network of friendships obtained through a social survey of high school students in 1994 [137].	6813	1630
add_health (comm49)	A directed network of friendships obtained through a social survey of high school students in 1994 [137].	7031	1877
add_health (comm36)	A directed network of friendships obtained through a social survey of high school students in 1994 [137].	7986	2152
collins_yeast	Network of protein-protein interactions in Saccharomyces cerevisiae (budding yeast), measured by co-complex associations identified by high-throughput affinity purification and mass spectrometry (AP/MS) [246].	8319	1004
un_migrations	A network of migration between countries, collected by the United Nations [247]	8405	232

add_health (comm40)	A directed network of friendships obtained through a social survey of high school students in 1994 [137].	8522	1996
add_health (comm41)	A directed network of friendships obtained through a social survey of high school students in 1994 [137].	8646	2064
bible_nouns	A network of noun phrases (places and names) in the King James Version of the Bible [248].	9059	1707
fao_trade	Multiplex network representing trade relationships between countries from the Food and Agricultural Organization of the United Nations [249].	9420	214
dnc	A network representing the exchange of emails among members of the Democratic National Committee, in the email data leak released by WikiLeaks in 2016 [250]	10384	849
add_health (comm50)	A directed network of friendships obtained through a social survey of high school students in 1994 [137]	10455	2539
foursquare (NYC_restaurant_checkin)	Two bipartite networks of users and restaurant locations in New York City on Foursquare, from 24 October 2011 to 20 February 2012 [251].	13457	4906
bitcoin_alpha	A network of who-trusts-whom relationships among users of the Bitcoin Alpha platform [252].	14120	3775
plant_pol_robertson	A bipartite network of plants and pollinators, from southwestern Illinois, USA [253].	15254	1882
genetic_multiplex (Arabidopsis)	Multiplex networks representing different types of genetic interactions, for different organisms [254].	16064	6692
polblogs	A directed network of hyperlinks among a large set of U.S. political weblogs from before the 2004 election. Includes blog political affiliation as metadata [127].	16714	1222
us_congress (H93)	Two temporal networks of bill co-sponsorship tendencies among US Congresspeople, from 1973 (93rd Congress) to 2016 (114th Congress) [255].	18083	446
openflights	A network of regularly occurring flights among airports worldwide, extracted from the openflights [256]	18833	3188
chicago_road	A transportation network of Chicago, USA, from an unknown date (probably late 20th century) [257].	20627	12979
bitcoin_trust	A network of who-trusts-whom relationships among users of the Bitcoin OTC platform [258].	21489	5875
escorts	A bipartite network of escort and individuals who buy sex from them in Brazil, extracted from a Brazilian online community for such ratings [259].	38540	15810
edit_wikinews (ar)	Two bipartite user-page networks extracted from Wikipedia, about news events [141].	38619	23975
advogato	A network of trust relationships among users on Advogato, an online community of open source software developers [260].	39227	5042
gnutella (04)	A sequence of 9 snapshots of the Gnutella peer-to-peer file sharing network from 5-31 August 2002 [261].	39994	10876
hiv_transmission	A set of networks of HIV transmissions between people through sexual, needle-sharing, or social connections, based on combining 8 datasets collected from 1988 to 2001 [262].	41206	26706
word_adjacency (darwin)	Directed Networks of word adjacency in texts of several languages including English, French, Spanish and Japanese [263].	44205	7377

arxiv_collab (cond-mat-1999)	Collaboration graphs for scientists, extracted from the Los Alamos e-Print arXiv (physics), for 1995-1999 for three categories, and additionally for 1995-2003 and 1995-2005 for one category [264].	44619	13861
internet_as	A symmetrized snapshot of the structure of the Internet at the level of Autonomous Systems (ASs), reconstructed from BGP tables posted by the University of Oregon Route Views Project [128].	48436	22963
anybeat	A snapshot of the Anybeat online social network from 2013, before it was shut down [265].	49132	12645
inploid	Inploid is a social question and answer website in Turkish. Users can follow others and see their questions and answers on the main page. Each user is associated with a reputability score which is influenced by feedback of others about questions and answers of the user. Each user can also specify interest in topics [266].	49334	14360
dblp_cite	Citations among papers contained in the DBLP computer science bibliography [267].	49579	12494
jung	A network of software class dependency within the JUNG 2.0.1 and javax 1.6.0.7 library namespaces edu.uci.ics.jung and java/javax. Nodes represent classes and a directed edge indicates a dependency of one class on another [268].	50290	6120
caida_as (20071112)	A sequence of 122 network snapshots denoting Autonomous System (AS) relationships on the Internet, from 2004-2007, inferred using the Serial-1 method from RouteViews BGP table snapshots and a set of heuristics [269].	52861	26389
jdk	A network of class dependencies within the JDK (Java SE Development Kit) 1.6.0.7 framework. Nodes represent classes and a directed edge indicates a dependency of one class on another [270].	53658	6434
us_roads (AK)	The road networks of the 50 US States and the District of Columbia based on UA Census 2000 TIGER/Line Files [148].	55014	48560
chess	A network among chess players (nodes) giving the chess match outcomes (edges), for game-by-game results among the world's top chess players [271].	55779	7115
us_air_traffic	Yearly snapshots of flights among all commercial airports in the United States from 1990 to today [272].	58226	2274
nematode_mammal	A global interaction web of interactions between nematodes and their host mammal species, extracted from the helminthR package and dataset [273].	58825	26197
budapest_connectome (all_20k)	A parameterizable consensus brain graph, derived from connectomes of 477 people, each computed from MRI datasets of the Human Connectome Project [274].	70654	1015
movielens_100k	Three bipartite networks that make up the MovieLens 100K Dataset, a stable benchmark dataset of 100,000 ratings from 1000 users on 1700 movies [275].	70946	23761
digg_reply	Network of replies among users of digg.com. Each node in the network is a digg user, and each directed edge indicates that user i replied to user j [276].	84781	29652
cora	Citations among papers indexed by CORA, from 1998, an early computer science research paper search engine [277].	89157	23166
foldoc	A network of hyperlinks among entries in the Free On-line Dictionary of Computing [278].	91471	13356
marvel_universe	The Marvel Universe collaboration network, where two Marvel characters are considered linked if they jointly appear in the same Marvel comic book [279].	95445	19182

dbpedia_writer	A bipartite network of writers and the written works they created, as extracted from Wikipedia by the DBpedia project [280].	98895	74775
elec	A network of votes on Request for Adminship (RfA) elections from a 2008 snapshot of Wikipedia [281].	100667	7066
topology	An integrated snapshot of the structure of the Internet at the level of Autonomous Systems (ASs), reconstructed from multiple sources, including the RouteViews and RIPE BGP trace collectors, route servers, looking glasses, and the Internet Routing Registry databases [282].	107720	34761
python_dependency	Python's package dependency networks [283].	107819	58302
slashdot_threads	A network of replies among users of the website Slashdot on the various discussion threads on the site [284].	116573	51083
reactome	A network of human proteins and their binding interactions, extracted from Reactome project [285].	145778	5973
google	A directed network of webpages from Google's own sites, and the hyperlinks among them [286].	148585	15763
dbpedia_producer	A bipartite network of producers and the works they created, as extracted from Wikipedia by the DBpedia project [280].	150714	111632
email_enron	The Enron email corpus, containing all the email communication from the Enron corporation, which was made public as a result of legal action [287].	180811	33696
wiki_rfa	The set of all votes on Requests for Adminships (RfA), from 2003 to May 2013, represented as a directed, signed network in which nodes represent Wikipedia members and edges represent votes [288].	181906	11381
facebook_wall	Friendship relationships and interactions (wall posts) for a subset of the Facebook social network in 2009, recorded over a 2 year period [289].	182384	43953
arxiv_authors (AstroPh)	Scientific collaborations between authors of papers submitted to arxiv [129].	196972	17903
pgp_strong	Strongly connected component of the Pretty-Good-Privacy (PGP) web of trust among users, circa November 2009 [290].	197150	39796
dbpedia_occupation	A bipartite network of the affiliations between notable people and occupations, as extracted from Wikipedia by the DBpedia project [280].	203971	143222
linux	A network of Linux (v3.16) source code file inclusion [291].	213208	30817
scotus_majority (2008)	Network of legal citations among majority opinions written by the Supreme Court of the United States (SCOTUS), from 1754-2002 (2008 version) and 1792-2006 (2007 version) [292].	216436	25389
dbpedia_recordlabel	Bipartite networks of the affiliations (contractual relations) between artists and the record labels under which they have performed, as extracted from Wikipedia by the DBpedia project [280].	222516	169463
dbpedia_location	A bipartite network of the affiliations between named entities from Wikipedia and particular notable locations, as extracted from Wikipedia by the DBpedia project [280].	263957	181951
dbpedia_starring	A bipartite network of movies and the actors that played in them, as extracted from Wikipedia by the DBpedia project [280].	264909	134016
corporate_directors	Bipartite network of directors and the companies on whose boards they sit, spanning 54 countries worldwide, constructed from data collected by the Financial Times [293].	293598	264699

word_assoc	A network of word associations showing the count of such associations as collected from subjects, from the Edinburgh Associative Thesaurus (EAT) [294].	297094	23132
twitter_events (NYClimateMarch2014)	Various multiplex networks of retweets, mentions, and replies among Twitter users during specific events or occasions in 2013 and 2014 [295].	325093	99666
douban	A friendship network among users on Douban [296].	327162	154908
email_eu	An email network (anonymized) from a large European research institution, collected from October 2003 to May 2005 (18 months) [129].	339925	224832
foursquare_friendships (old)	A network of user friendships on Foursquare, from April 2012 to September 2013 [297].	357921	105091
epinions_trust	A who-trusts-whom online social network of the general consumer review site Epinions [298]	405739	75877
github	The bipartite project-user membership network of the software development hosting site GitHub [299].	417361	139752
arxiv_citation (HepPh)	Citations among papers posted on arxiv.org under the hep-ph and hep-th categories, between 1993 and 2003. This time begins a few months after axiv was launched. If a paper i cites a paper j also in this data set, then a directed edge connects i to j. (Papers not in the data set are excluded.) These data were originally released as part of the 2003 KDD Cup [300].	420784	34401
dbpedia_genre	A bipartite network of the affiliations between artists and their works on one side and genre classifications on the other, as extracted from Wikipedia by the DBpedia project [280].	458324	259139
slashdot_zoo	A network of interactions among users on Slashdot (slashdot.org), a technology news website [301].	467731	79116
wordnet	A network of English words from the WordNet [302]	656230	145145
wiki_users	A network derived from interactions between editors of the English language Wikipedia, as derived from the edit histories of 563 wiki pages related to politics [303].	715334	137740
wiki_link_dyn	Six networks of the evolving hyperlink structure among wikipedia articles, for simple English (en), German (de), Dutch (nl), Polish (pl), Italian (it), French (fr), taken in August 2011 [304].	824581	99636
academia_edu	Snapshot of the follower relationships among users of academia [301].	1022440	200167
myspace_aminer	This network contains the social graph of MySpace, a social networking website which also has a strong music emphasis [305–308].	5635236	853360
as_skitter	An aggregate snapshot of the Internet Protocol (IP) graph, as measured by the traceroute tool on CAIDA's skitter infrastructure, in 2005 [309].	11094209	1694616
libimseti	A network of ratings given between users at Libimseti [310].	17233144	220970

B.7 Results of fitting SBMs to randomised networks

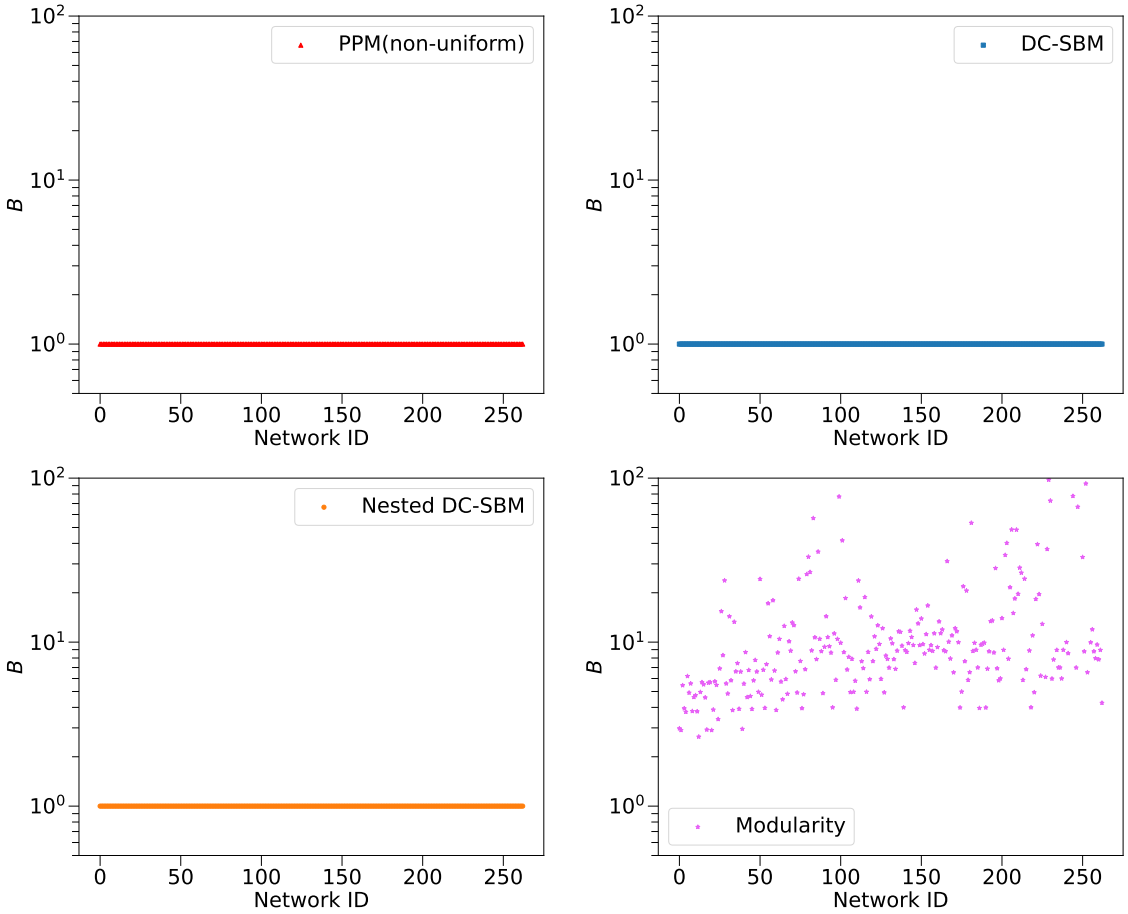


Figure B-7: Inferred number of communities in randomised networks, using the non-uniform PP model, DC-SBM, and Nested DC-SBM. For comparison, we also present the result of the modularity maximisation approach.

Appendix C

Supplementary materials for Chapter 5

C.1 Comparison of the BP running time

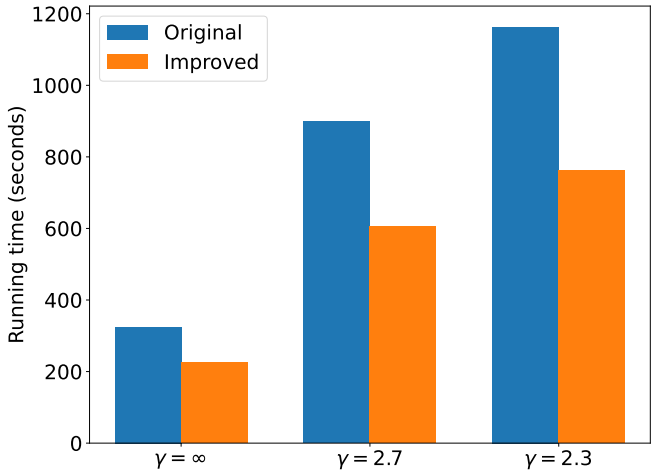


Figure C-1: Running time of one BP iteration in networks with different level of heterogeneity in degree distribution. Networks are sampled from the (degree-corrected) planted partition model with $N = 10^5, B = 2, \langle k \rangle = 10$. The assortativity structure is set according to the parameterisation defined in the equation (5.42), with $\epsilon = 0.8$.

The x-axis is the value of the shape parameter ζ of the Zipf's distribution, which was used to generate the degree propensity parameters θ_u. When ζ takes a value in $[2, 3]$, the Zipf's distribution has a finite mean but its variance diverges. Smaller values of ζ make the Zipf's distribution more heterogeneous, and $\zeta = \infty$ represents the homogenous case $\theta_u = (N/B)^{-1}$. To avoid arbitrarily large samples, we set a cut-off value at $x_{\max} = 50$. As can be seen from the figure, the running time of the BP iterations significantly increases as the degree distribution becomes more heterogenous. However, about a third of the running time is caused by unnecessarily repeated computation, which can be avoided by updating the BP messages in a modified way as we introduce in

Chapter 5. Although the modified updating scheme requires extra time for maintaining precomputed terms, we found that the BP implementation is more efficient when the modification is in place, even when the degree distribution is set to be homogeneous. The advantage of the modified updating scheme gets more clear as the heterogeneity in the degree distribution increases. A simple demonstration of the difference between the two different update schemes is available in author's Github repository.