Two statistical methods for graph model selection : Distance to the microcanonical ensemble and prequential inference on edge sequences

Louis Duvivier

N° d'ordre NNT :2021LYSEI093

Thèse de doctorat de l'Université de Lyon
opérée au sein de
L'INSA de Lyon

Ecole Doctorale N° 512
Mathématiques et Informatique (InfoMaths)

Spécialité / discipline de doctorat : Informatique

À soutenir publiquement par
Louis Duvivier

Two statistical methods for graph model selection:

Distance to the microcanonical ensemble test and prequential inference on edge sequences

Devant le jury composé de:

Pr. Pierre Borgnat	ENS-Lyon	Examinateur
Dr. Rémy Cazabet	Université Claude Bernard Lyon 1	Co-directeur de thèse
Pr. Catherine Matias	Sorbonne Université	Rapporteur
Pr. Tiago Peixoto	Université centrale européenne	Examinateur
Pr. Céline Robardet	INSA-Lyon	Directrice de thèse
Dr. Lionel Tabourier	Sorbonne Université	Rapporteur

Avant-propos

Une thèse, comme la plupart des écrits scientifiques, n'est pas une lecture très accessible. Pourtant, je veux profiter de ces quelques lignes pour souligner que ce travail n'est pas qu'une affaire de spécialistes.

D'abord parce qu'aucune recherche scientifique n'est un travail purement personnel. Durant ce doctorat, j'ai profité des conseils de mes encadrants, Rémy Cazabet et Céline Robardet, de l'équipe du LIRIS, mais aussi de toutes les recherches précédentes, dont une partie seulement est présente dans la bibliographie. Plus largement, j'ai pu profiter de vingt ans d'enseignement, de l'école à l'université, et du travail de tout ceux qui, à travers la société, contribuent à les faire fonctionner : professeurs bien sûr, mais aussi les secrétaires, agents d'entretien et les maçons qui ont construit les bureaux dans lesquels nous travaillons. Je ne connais évidemment pas tous leurs noms, mais je tiens à dire que sans eux, ce travail n'existerait pas.

Au-delà des remerciements, ces liens nous engagent. La science n'a de sens que si elle permet à l'humanité de comprendre son environnement et sa propre organisation sociale de manière à pouvoir s'y orienter. L'étude des réseaux en particulier doit nous permettre de prendre conscience des liens qui existent, dans la nature comme dans la société, que ce soient les liens entre protéines nécessaires au fonctionnement d'une cellule ou les interactions sociales qui permettent la vie collective. Pouvoir décrire, comprendre et anticiper l'évolution des besoins en terme d'échange d'information, d'énergie, de transport, dans une période ou ces échanges se développent à l'échelle mondiale, c'est un outil formidable que l'humanité a à sa disposition pour pouvoir organiser consciemment sa vie sociale.

Il est d'autant plus important de le rappeler que bien souvent, les résultats des travaux scientifiques sont utilisés non pour améliorer le sort de l'humanité, mais contre elle. Sans même parler des merveilles technologiques qui foisonnent dans l'industrie de l'armement, il est certain que les applications de l'étude des graphes pour l'espionnage automatique et la spéculation financière mobilisent plus de moyens que pour la recherche en chimie et en médecine.

Arracher la science à la dictature du profit pour la mettre au service des besoins de l'humanité entière, c'est une tâche qui dépasse évidemment le cadre strictement universitaire, et qui ne pourra être accomplie que par la masse des travailleurs eux-même. En ce sens, ce modeste travail leur appartient.

Foreword

A thesis, as any scientific writing, is not a very accessible reading. Yet, I want to take advantage of those few lines to underling that this work is not only a matter for specialists

First of all because no scientific research is a fully personal work. All along this PhD, I benefited of advice from my supervisors, Rémy Cazabet and Céline Robardet, from the LIRIS team, but also from all previous research only part which is mentioned in the bibliography. Beyond that, I benefited of twenty years of instruction, from school to university, and from the work of all those who, all across society, contribute to it : teachers of course, but also the secretaries, cleaners, and mason who built the office we work in everyday. I cannot know all their names, but I want to say that without them, this work would not exist.

Beyond gratitude, these links bind us. Science has no sense but to give humanity the ability to understand its environnement and its own social organization in order to direct itself. Studying networks in particular must help us to be conscious of the links that exist, in nature as well as in society, be there protein interactions necessary to the functionning of the cell, or social interactions necessary to collective living. Being able to describe, understand and predict the evolution of the needs in terms of communication and transportation, in a period where those exchanges develop at a global scale, is a wonderful tool humanity has at its disposal to organize consciously it social live.

It is even more important to recall this that, most often, scientific results are used not to improve humanity's fate but against it. Setting aside the technological wonders which abound in the armament industry, it is certain that graph theory applications in automatic spying and financial speculation attract more money than chemical and medical research.

Take science out of the dictatorship of profit to make it serve the needs of the whole humanity is a task that obviously outreaches the academic framework. It can only be achieved by the bulk of workers themselves. In that respect, this modest work belongs to them.

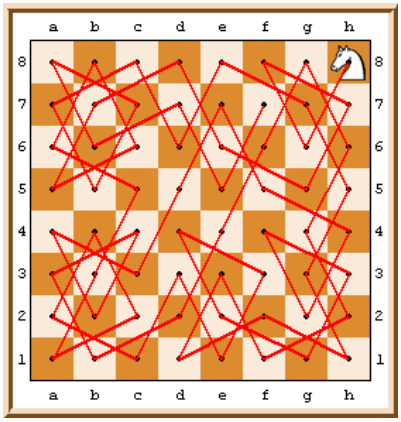
Contents

Chapter 1

Introduction

Why study graphs?

At first glance, graphs might seem a very abstract topic. Indeed, they are a mathematical construction, not something people deal with in their everyday life. What we do encounter are situations in which people, objects or places interact with each other. People are connected by social relationships, electric devices by wires, cities by roads, atoms by chemical bonds, etc. Graphs, defined as a set of n nodes V connected together by a set of edges $E \subset V^2$ are a fundamental object to reason about those interconnected systems.



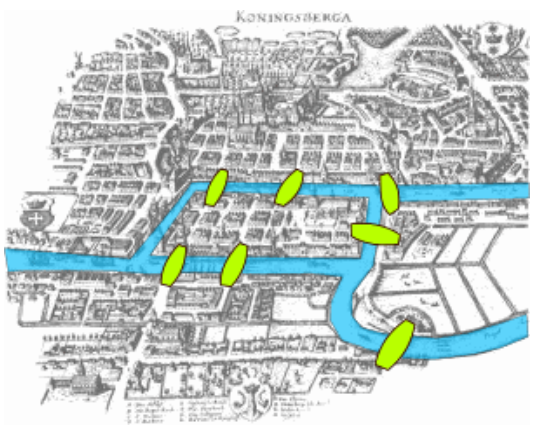


Figure 1.1 – Illustration of the knight and seven bridges of Königsberg problems (source `https://fr.wikipedia.org/wiki/Problème_du_cavalier` and `https://fr.wikipedia.org/wiki/Problème_des_sept_ponts_de_Königsberg`). Ces images sont disponibles sous license Creative Commons Attribution-Share Alike 3.0 Unported

One of the oldest known problem about graphs dates back to about 840. In his book *Kitab ash-shatranj*, the Arab mathematician Al-Adli studied chess. He wanted to determine whether it is possible for a knight to go through every 64 squares of a chessboard without passing two times on the same square. This problem can easily be formulated using graph formalism: V is defined as $[\![0,7]\!] \times [\![0,7]\!]$ (each node is associated to the coordinates (i, j) of a square on the chessboard) and two nodes are connected if and only if the knight can go from

one square to the other. Then the problem boils down to deciding whether it is possible to go through all nodes of the graphs exactly once by following edges, which is a classical route problem in graph theory. One similar problem was studied by Leonhard Euler in 1741, when he wanted to determine whether it was possible to travel through the city of Königsberg going exactly once over each of the seven bridges in the city Euler [1741]. In the XIX century, graph theory found applications in chemistry when Arthur Cayley Cayley [1857], followed later on by George Pölya Polya and Read [2012]; Pólya [1937] used it to enumerate the molecules that could be formed with a given set of atoms, knowing the number of bonds an atom can build.

These questions have in common that they do not depend on the nature of the interactions (knight move, bridge or chemical bond). Graphs allow to neglect this information when it is not relevant, and it provides an efficient summary of a set of constraint which must be satisfied. It is thus helpful when studying systems in which local constraints are easy to describe and one wonders how they affect the global behaviour of the system.

A good example of such a problem is map colouring: in 1852, Francis Guthrie conjectured that any map could be coloured with no more than four colours, in such a way that no two adjacent countries would be of the same colour. By representing each country by a node and connecting them if they share a border, this conjecture is equivalent to: any planar graph can be coloured with four different colours in such a way that any two adjacent nodes have different colours. This statement is easy to understand, and it is easily verified on any small graph simply by colouring it by hand. Yet, the formal proof of this theorem, known as the four colours theorem, was only found more than a century later by Kenneth Appel and Wolfgang Hakenin Appel and Haken [1989] and it required hours of computations to check thousands of special cases. One of the reason that makes this result hard to prove is that, as explained by Arthur Cayley in 1879 Cayley [1879], if one manages to colour a graph of n nodes with four colours, and then adds one node, there is no guarantee that this new node can be coloured without modifing others' nodes colour, even if they are very far in the graph.

This illustrates how local constraints may induce long range correlations between variables, which is at the heart of many problems in graph theory. Along the XX century, many types of graphes were studied, to model different kind of constraints: directed edges, when interactions are asymmetric, weighted edges, when the strength of interaction must be taken into account, etc. For a summary of classical results in graph theory, one may refer to Berge [1958].

Why study random graphs?

However rich these definitions of a graph may be, they all remain in a deterministic framework: given two nodes, one can always tell whether they are adjacent (*i.e.* connected by an edge) or not. This hypothesis is too strong in many real life situations, in which the graph of interactions cannot be entirely known. There might be various reasons to this impossibility: interactions may evolve quickly over time, they may be difficult or expansive to detect, etc. To take into account this uncertainty, one must study random graphs defined as a probability distribution over all possible graphs.

A typical random graph question is percolation: given the probabilities for each edge to be present, what is the probability for the overall graph to be connected (*i.e.* For any pair

of node (u, v), there exist a path $(u_0 = u, u_1, u_2, \ldots, u_{n-1}, u_n = v)$ such that for all i, u_i and u_{i+1} are adjacent) Bollobas et al. [2006]? For example, in the domain of telecommunications, when relay stations with a limited communication range might be added or deleted quickly, what is the required density of stations in order for a message to be able to go from one point to another Gilbert, E.N. [1959]?

One of the simplest, yet extremely rich, model of random graphs was described in 1959 by Paul Erdös and Alfréd Rényi Karoński and Ruciński [1997]. They define a random graph $G(n, m)$ as a set of n nodes V, which are connected by m edges drawn at random from V^2. This model allowed them to derive the probability that $G(n, m)$ is connected, depending on m Erdös, P. and Rényi, A. [1959, 1960]. Studying a relaxation of this model, where the number of edges m is replaced with a probability p for each edge to be drawn at random, they showed that there exists a critical probability p_c such that if $p < p_c$, then the probability for $G(n, p)$ to be connected is asymptotically null, while if $p > p_c$ it tends to 1.

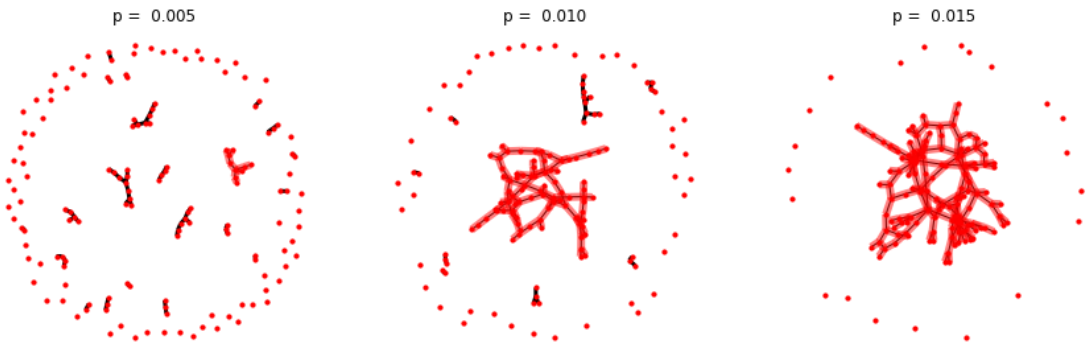


Figure 1.2 – The largest connected component of a random graph with 150 nodes and an edge probability ranging from 0.005 to 0.015.

This result is important because it shows how processes occuring on graphs may exhibit non-linear behaviour. For example, in the context of an epidemic which spreads following random contacts between people, the number of people infected is not proportional to the probability of transmitting the disease. As long as the contagion probability remains below the critical probability, the disease will remain limited to a small group of people, but as soon as it crosses this threshold, it will almost surely infect the whole population. Of course, modelling social interactions as random is simplistic, but such threshold appears also under stronger assumptions, even though its value may vary Sattenspiel and Simon [1988].

It should be stressed that this threshold was obtained as a result of combinatorics, independently from the physical mechanisms which determine the actual probability p for an interaction to occur between two nodes in the graph. This shows that a global property of the system (for example, the whole population is infected) might be determined as much by the structure of the interaction graph (do people encounter each other uniformly at random or not?) as it is by the nature of interactions (what is the probability of contagion when they encounter?).

Complex networks

The spread of computers in the 1990's induced a major shift in graph theory. Until then, it had essentially been a matter of interest for mathematicians, who dealt with hand-built networks. They had studied specific classes of graphs defined by a set of strict rules and displaying a lot of structure and symmetries, such as lattices, cliques, trees, etc., and on the other hand random graphs, which by definition have no structure at all. Yet, there was a lack of graphs obtained from observations with which theoretical results could be compared. Sociologists did use graph to model social interactions Sampson [1968]; Zachary [1977] but as observations had to be recorded manually, the size of the few graphs available was necessarily limited. To obtain a social network, for example, sociologists would need to count the number of interactions between each person in the group , as would do biologists observing a group of animals Lusseau et al. [2003]. These empirical networks would have at most a few dozens nodes. It was thus hard to infer general results about graphs from so little information.

Computers changed this situation by making many real-life interaction networks easier to record, and also giving the computation power necessary to automatically analyze networks with up to millions of nodes. In biology, it allowed to study networks of chemical reactions in the metabolism Hartwell et al. [1999]; Holme et al. [2003], neurons in the brain Eguiluz et al. [2005]; Rubinov and Sporns [2010], or protein-protein interaction networks Maslov and Sneppen [2002]. Infrastructure networks also became easier to study, be they power grid Amara et al. [2011], railways Latora and Marchiori [2002], or internet Faloutsos et al. [2011]. Communication networks, even with millions of nodes became available, like the world wide web Kleinberg et al. [1999], citation networks between authors Seglen [1992] and of course online social networks Adamic et al. [2001].

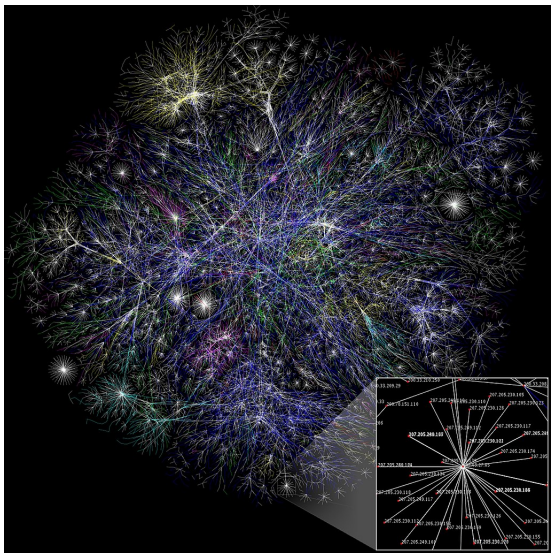


Figure 1.3 – A map of a portion of the internet network in 2005 (source `https://fr.wikipedia.org/wiki/Internet`, disponible sous licence Creative Commons Attribution 2.5 Générique). Nodes are IP adress, and edges indicate their connections.

This novel availability of real-life networks quickly led to observations that existing models were not able to explain about their degree distribution, clustering coefficient, path length, and so on, which has fuelled interest in graph theory to design new models. The wealth of structures exhibited by graphs at the local and global level implies that in many cases, various models can be applied to a given graph. This is problematic as those models may induce different explanations of a given phenomenon, and different previsions about futur interactions. For example, is the number of phone calls from one city to another determined by their population? The distance between them? The language spoken in each? If different factors must be taken into account, what should be their relative weights? Such questions require to evaluate the relevance of a model with respect to an observed graph.

In chapter 2, I start by reviewing some of the most common models and model design frameworks that have been developed in the last twenty years. Then, I present model selection techniques, and how they were adapted to the case of graph models. In chapter 3, I introduce Graph Spaces, an approach that combines probability distributions with graph distances to evaluate the probability that a given graph was generated by a candidate model. I then move on in chapter 4 from this statistical test approach to bayesian model selection, and more precisely to the case of entropy based model selection that was used in the case of stochastic blockmodels. After showing some of its limits, I develop another model selection framework in chapter 5, based on a reformulation of graph models as probability distributions on sequences of edges.

Chapter 2

State of the art

Definitions

Before entering in more details into the state of the art, I will first state some basic definition and notation that will be used in the rest of the thesis. A *labeled* graph is a graph whose nodes are labeled from 0 to $n-1$. Such a graph can be represented by its *adjacency matrix* $A \in \mathcal{M}_n(\{0,1\})$ defined as:

$$\forall i,j \in [\![0,n-1]\!]^2, A_{i,j} = \begin{cases} 1 & \text{if } (i,j) \in E \\ 0 & \text{else} \end{cases}$$

An example of a labelled graph and its adjacency matrix are displayed in figure 2.1.

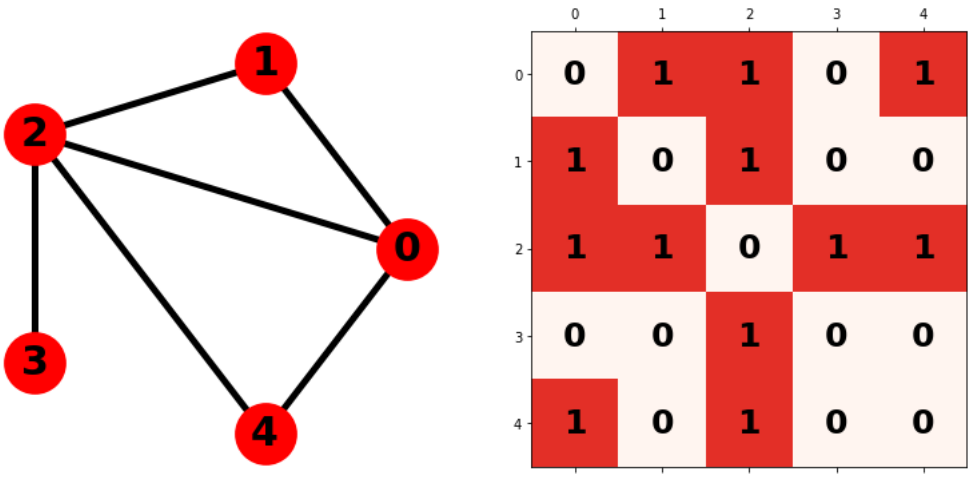


Figure 2.1 – Example of a simple undirected graph and its adjacency matrix

A graph is said to be *undirected* if its adjacency matrix is symmetric, else it is *directed*. It has no self-loop if $\forall i \in [\![0,n-1]\!], A_{ii} = 0$. Finally, it is *weighted* if there exists a function $w : E \to \mathbb{R}$ which associates a weight to each of its edges. In this case, one may consider the *weight matrix* $W \in \mathcal{M}_n(\mathbb{R})$ such that $W_{i,j} = w(i,j)$ rather than the adjacency matrix of the graph. A

graph with no weight is called a *simple* graph. In this thesis, unless otherwise specified, we will consider labelled and directed multigraphs with self loops, because, although they are not the most widely used in practice, they allow for simpler computations.

Definition 1. *A **graph model** is a probability distribution $\mathbb{P}$ over a set of graphs Ω. If this set is a singleton, the model is said to be **deterministic**, else it is **probabilistic**. A **model class** is a function that maps a set of parameter $\theta \in \Theta$ to a model*

$$\Phi : \theta \mapsto \mathbb{P}_\theta$$

For example, the class of complete simple graphs takes as parameter a number of nodes n and maps it to the deterministic model corresponding to the complete simple graph with n nodes K_n. Most classes of graphs in classical graph theory like n-ary trees, lattices, and so on, can be seen as classes of deterministic models of graphs (or *deterministic classes*). In the litterature, the word "model" is frequently used to designate both models and model classes. However, it is useful to distinguish them to reason about parameter inference and model selection.

In this thesis, I will focus on probabilistic model classes, the simplest of which, as already mentioned in introduction, is the random graph model studied by Erdös and Rényi Erdös, P. and Rényi, A. [1959, 1960]. It was initialy defined over the set of simple undirected graphs with no self-loops. It takes as parameters a number of nodes n and a number of edges m, and the model $ER(n, m)$ generates a graph by picking at random m edges among the $M = \frac{n(n-1)}{2}$ possible pairs of nodes. Each graph is thus generated with a probability

$$\mathbb{P}_{ER(n,m)}[G] = \frac{1}{\binom{M}{m}}$$

This formulation can be said "hard", in the sense that it assigns null probability to any graph on n nodes which does not have precisely m edges. In practice, a softer formulation is frequently used which takes as parameter a probability $p \in [0, 1]$. Then, for each pair of node $u < v \in V$, the corresponding edge is selected with probability p and rejected with probability $(1 - p)$. The probability to generate a graph G with m edges is then

$$\mathbb{P}_{ER(n,p)}[G] = p^m(1 - p)^{M-m}$$

The possibility to derive exact formulas for the probability distribution allows to compute expected values for the degree distribution, the average path length, the size of greatest connected component, etc. This model is rarely useful in modeling real data because the hypothesis that edges are placed at random is most of the time too strong an assumption. Yet, it provides a null model against which real networks can be compared in order to understand what makes them different from random.

The exploration of a large amount of networks coming from different fields of research has highlighted non-trivial common properties such as the "small-world" property, long-tailed degree distributions, or community structure. All those properties do not match with deterministic classes of graphs, like lattices or trees, which have high average path length, constant degree distribution, and exhibit no cluster. They do not match either with random graph models whose clustering coefficient are low, whose degree distribution are binomial

and which also have no cluster. Indeed, these so-called complex networks seem to lie somewhere between perfectly symmetric and random graphs, which imposes to propose new models to explain these observations.

In section 2.2, I will present a selection of the most widely used graph model classes. The interested reader can find more complete reviews in Barthelemy [2011]; Goldenberg et al. [2009]; Goyal and Ferrara [2018]; Newman [2003]. In section 2.3, I present the statistical methods that were developed to infer the parameter of a model from an observed graph. Finaly, in section 2.4, I conclude the state of the art with a review of model selection techniques, and how they have been adapted to the study of graphs.

Empirical models

The development of new graph models steemed from the observation of real world networks whose structure could not be explained by graph theory. It thus started on an empirical basis: as new properties were observed, new models were proposed to explain the processes through which they could have emerged. Following this line of thought, I will present in this section some of the most famous ones, along with the properties they were designed to explain.

Degree distribution

Random graphs assume that all nodes are equally likely to interact with any other node, which is not a realistic assumption in many situations. All chemical elements do not establish the same number of bonds, every person do not establish the same number of social ties, etc. The simpler indicator of a node propensity to interact is the number of other nodes it is adjacent to, measured by its *degree*:

$$k_i = \sum_{j=0}^{n-1} A_{i,j}$$

When studying directed graphs, one can distinguish incoming and outgoing edges to obtain the corresponding *indegree* and *outdegree*.

The distribution of degrees $(k_i)_{i \in V}$ in social networks was studied as early as the 1960's, with works by Rapoport Rapoport and Horvath [1961] or Price Price [2011], who showed that it was significantly different from the limit Poisson distribution expected for large Erdös-Rényi random graphs. This sparked interest toward the study of graphs with prescribed degree sequence: sociologist and biologist in particular were interested to know whether the strong interactions they observed between some people or species should be considered the result of their higher interaction propensity or whether additional reasons should be seeked Connor and Simberloff [1979]; Harper jr [1978]; Strauss [1982]. To do so, they needed to generate random networks having the same degree distribution as the observed network in order to have a comparison point. This class of random networks with a prescribed degree distribution is known as the configuration model. I will present its undirected version, as the directed one can easily be derived from it.

The theoretical analysis of the configuration model is harder than the one of the Erdös-Rényi class because of the difficulties to compute the probability distribution associated to a degree sequence $(k_i)_{i \in [\![0,n-1]\!]}$. Computing the number of different simple undirected

graphs which correspond to a given sequence is an open question, and only asymptotical results were found in the 1970's Bender [1974]; Bender and Canfield [1978]. Analytical results have been obtained for the average path length, the size of the giant component, and other characteristic graph properties Chung and Lu [2002]; Molloy and Reed [1998]; Molloy et al. [2011]; Newman et al. [2001], but the lack of a closed formula for the probability distribution implies that any new question about the configuration model requires an ad hoc analytical approach.

When the analytical study is not possible, one may turn to a more experimental approach. Algorithms exist that generate graphs with prescribed degree sequence. In particular, the *switching algorithm* starts from a graph G with the desired degree sequence (typically, the observed network), and generates a random graph with the same sequence by randomly switching pairs of edges. At each swith, four nodes i,j,k,l are chosen such that $(i,j),(k,l) \in E$ and $(i,k),(j,l) \notin E$, then edges (i,j) and (k,l) are removed and replaced by (i,k) and (j,l). By definition, each switch preserves the degree sequence. This algorithm allows to check at each step that no multiedge or self-loop is formed, which is useful to obtain a simple graph, but the number of switches necessary to shuffle edges can be very high. What is more, this algorithm does not sample uniformly from the set of all possible simple graphs. This is problematic if one wants to use it, for example, to estimate the probability of an edge between two nodes from a randomly generated sample. In Milo et al. [2003], authors present an algorithm that does sample uniformly, but at the cost of an even longer running time. The uniform generation of graphs, especially simple graphs, with prescribed degree sequence is still an active topic of research, information about recent advances can be found in Arman et al. [2019].

The Molloy-Reed algorithm allows to quickly generate graphs with an arbitrary degree sequence $(k_i)_i$. It starts with each node i having k_i half-edges attached to it and sequentially selects pairs of nodes with free half-edges to connect them together. Because it may select pairs of nodes that have already been selected, or the pairs of the form (i,i), it generates multigraphs with self-loops. It can be used to generate simple graphs by relaunching the algorithm until no self-loop or multiedge is obtained, but there is no guarantee that the process terminates, because some degree sequences cannot be realized as a simple graph Hakimi [1962]; Havel [1955]. Also, in this case too the algorithm does not sample uniformly from the set of simple graphs.

These theoretical issues with the configuration model are one of the main reason why the analytical study of graph models are easier on multigraphs with self-loops. In this case, the Molloy-Reed algorithm shows that the expected number of edges between two nodes of degree k_i and k_j is

$$\mathbb{E}[W_{i,j}] = \begin{cases} \dfrac{k_i k_j}{2m-1} & \text{if } i \neq j \\ \dfrac{k_i(k_i-1)}{2m-1} & \text{if } i = j \end{cases}$$

Yet, the probability for two nodes i and j to be connected is not easy to compute. In the litterature, the following approximation is frequently made

$$\mathbb{P}[W_{i,j} = 1] \approx \mathbb{E}[W_{i,j}] = \frac{k_i k_j}{2m-1}$$

However, it is important to bare in mind that this formula does not correspond to a proper probability, which can be seen by the simple fact that it may be greater than 1, and that this approximation is valid only if the probability that i and j are connected by more than one edge is negligible.

To obtain a proper probability distribution to work with, one can relax the degree constraints. In Giona Casiraghi and Nanumyan [2018] authors propose a soft configuration model (SCM) defined as an hypergeometric probability distribution

$$\mathbb{P}_{SCM((k_i)_i)}[G] = \frac{\prod_{i<j\in V} \binom{2\Xi_{i,j}}{G_{i,j}} \prod_{i\in V} \binom{\Xi_{i,i}}{G_{i,i}/2}}{\binom{M}{m}}$$

where $\forall i, j, \Xi_{i,j} = k_i k_j$ and $M = \sum_{i,j} \Xi_{i,j}$. This distribution is defined over the set of all graphs with n nodes and m edges, and it assigns non-null probability to graphs whose degree distribution is not equal to $(k_i)_{i\in V}$. The degree constraints are verified only on average, in the sense that

$$\forall i \in V, \mathbb{E}[\deg_G(i)] = k_i$$

This is why this version of the configuration model is said to be soft, with respect to the microcanonical version that assigns null probability to any graph whose degree sequence does not match the prescribed objective values.

Scale-free degree distribution

A particular family of degree distributions which has been observed in many different context are the so-called long-tailed degree distributions Albert and Barabási [2002]; Faloutsos et al. [2011]; Price [2011], and more specifically power laws of the form

$$\mathbb{P}[k] = C \times k^{-\alpha}, \alpha > 1$$

where C is a normalization constant defined as $C = \int_0^{+\infty} x^{-\alpha} \mathrm{d}x$. These distributions have been particularly studied because of their scale invariance property, in the sense that if the variable k is scaled by a factor λ, the distribution is identical up to a constant $\mathbb{P}[\lambda k] = \lambda^{-\alpha} \cdot \mathbb{P}[k]$. The fact that degrees distribute, at least approximately, according to such distributions would suggest that there exists no characteristic number of neighbours by nodes, no characteristic scale of study. This has been shown to have several consequences on the properties of the graph Albert et al. [2000]; Bollobás and Riordan [2004]; Cohen et al. [2011]; Pastor-Satorras and Vespignani [2001].

In Barabasi and Albert [1999], Barabasi and Albert develop a preferential attachment model to explain the emergence of such a scale-invariant degree distribution. Instead of a fixed number of nodes n, they consider a dynamic model, starting with a small number of nodes n_0, and then adding nodes one at a time. Each time a new node is added, it links to m other nodes. The probability that it links to node i is given by

$$\mathbb{P}[i] = \frac{k_i}{\sum_{j\in V} k_j}$$

They show that this simple preferential attachment mechanism leads to a power law distribution of degrees.

In practice, it is often hard to tell whether the power law is a better explanation to networks degree distribution than other long-tailed distribution (*i.e.* distributions decaying more slowly than exponential) Broido and Clauset [2019]; Clauset et al. [2009], which has fulled a long debate on the topic Holme [2019]. However, what is clear is that most networks exhibit significantly more high-degree nodes than would be expected for a random graphs and that this property has important consequences regarding the structure of the graph.

Small-world property

The particularities of real networks are not limited to their degree distribution. It has been shown that many networks exhibit a transitive structure, in the sense that if there is an edge (i, j) and an edge (j, k) in the graph, there is a high probability that edge (i, k) is present too Holland and Leinhardt [1976]. Let's denote $N(i) = \{j \in V \mid (i, j) \in E\}$ the set of neighbours of a node i. By definition the number of neighbours is $|N(i)| = k_i$ and, in a simple undirected graph, there can be at most $\frac{k_i(k_i-1)}{2}$ edges between them. Then, the transitivity of interactions around a node i can be measured through its clustering coefficient

$$CC_G(i) = \frac{2}{k_i(k_i - 1)} \sum_{j,k \in N(i)} G_{j,k}$$

At the level of the graph, one can consider its mean value $CC_G = \frac{1}{n} \sum_{i \in V} CC_G(i)$. A high value of this clustering coefficient means that nodes tend to interact locally and form small clusters of densely interconnected nodes.

On the other hand, it is also known that distances between nodes in a network tend to be surprisingly small, as was illustrated by a famous experiment by Milgram in 1967 Milgram [1967]. In random graphs, average distances are known to be proporitional to the logarithm of the number of nodes Chung and Lu [2002]. However, one would expect average distances in networks with a high clustering coefficient to be much higher, as nodes tend to connect with the neighbours of their neighbours. Indeed, on regular lattices, the average distance grows proportionally to the number of nodes.

In their paper Watts and Strogatz [1998], Watts and Strogatz show that various networks exhibit at the same time a high clustering coefficient and short average distances. They reproduce these features by randomly rewiring edges in a regular ring of nodes. As more and more randomness is added, average distances drop while the clustering coefficient remains high until a significant portion of edges has been rewired. They do not claim that this model actually replicate the real mechanisms of networks formation, but it shows how both properties can be obtained as a result of a combination of randomness and structure in the network.

Community structure

The degree distribution and clustering coefficient focus on local correlations in edge distribution. In Girvan and Newman [2002], authors argue that there exists groups of nodes, which they call communities, which tend to connect more densely with each others, beyond their imediate neighbourhood and that such communities can be found in several real world networks.

This idea can be formalized into a class of graph models using stochastic blockmodels (SBM), which were introduced in Holland et al. [1983]. This class of models takes as parameters a partition of the n nodes in p blocks $\mathcal{B} = (b_1, \dots, b_p)$ and a block adjacency matrix $M \in \mathcal{M}_p(\mathbb{N})$. It generates an undirected simple loop-free graph G by picking at random $M_{i,j}$ edges among the $K_{i,j} = |b_i||b_j|$ different pairs of nodes between block b_i and b_j (within a given block, $K_i = \frac{|b_i|(|b_i|-1)}{2}$). A graph G is therefore generated with probability

$$\mathbb{P}_{SBM(\mathcal{B},M)}[G] = \frac{1}{\left(\prod_{1 \leqslant i < j \leqslant p} \binom{K_{i,j}}{M_{i,j}} \times \prod_{i=1}^{p} \binom{K_i}{M_{i,i}}\right)}$$

Just as the Erdös Rényi model, it can be relaxed by using a block probability matrix $P \in \mathcal{M}_n([0,1])$, rather than a block adjacency matrix. In this case, the graph G is generated with probability

$$\mathbb{P}_{SBM(\mathcal{B},P)}[G] = \prod_{1 \leqslant i \leqslant j \leqslant p} P_{i,j}^{m_{i,j}} (1 - P_{i,j})^{K_{i,j} - m_{i,j}}$$

where $m_{i,j}$ is the number of edges between block b_i and b_j in G.

Many variations over the stochastic blockmodel have been introduced. In particular, in Karrer and Newman [2010], authors propose a degree-corrected version to model simultaneously the influence of the degree distribution and of the community structure over the distribution of edges. In Airoldi et al. [2008], a mixed-membership stochastic blockmodel is introduced, in which nodes are allowed to belong to multiple blocks. A review of the variations and developments over the stochastic blockmodel can be found in Lee and Wilkinson [2019].

Spatial models

Many real-world network such as transportation, communication and infrastructure networks are typically embedded in a geometric space, most of the time in 2 dimensions. In networks where the interaction cost between entities depends on the distance between them, it may be a crucial element to explain edge distribution Barthelemy [2011]. In particular, the gravity model has been used on various spatial network dataset Bhattacharya et al. [2008]; Jung et al. [2008]; Lambiotte et al. [2008]; Levy [2010]. It is based on the same principle as the configuration model, but it adds a deterrence function $f(d)$ to ponder the expected weight of an edge (i, j) based on the distance between the nodes d_{ij}

$$\forall i, j \in V, \mathbb{E}[W_{i,j}] = f(d_{ij}) k_i k_j$$

In Expert et al. [2011] and Cazabet et al. [2017], authors looked after communities by opitimizing an alternative modularity formula where the expected number of edges between two nodes in the configuration model had been replaced by the same quantity in the gravity model. It allows them to detect deviations from a null model which takes into account the effect of space, and thus detect communities beyond groups of nodes which are geometricaly close.

The deterrence function can be learned on the data, but it adds many parameters to the model, which can lead to overfitting. In Simini et al. [2012], authors propose a simpler model to capture the effect of distance on interactions. Inspired from the radiation process, this

model takes as parameter for each node i its outdegree k_i^{out} and its population n_i. For each pair of node i, j separated by a distance d_{ij}, they derive the expected weight of the edge $i \to j$ as

$$\mathbb{E}[W_{i,j}] = k_i \frac{n_i n_j}{(n_i + s_{ij})(n_i + n_j + s_{ij})}$$

where $s_{ij} = \sum_{k|0<d_{ik}<d_{ij}} n_k$. This model has no free parameters, which ensures that the correspondence between its prediction and observed data cannot be an artifact of the learning procedure. However, it also makes the model less flexible and it seems not to adapt well to all scales of study Barbosa et al. [2018]; Liang et al. [2013]; Masucci et al. [2013].

Graph embedding Spatial models can be extended to graphs whose nodes have no spatial coordinates by embedding their nodes in a low-dimension space. The random dot product model presented in Nickel [2008] is a class of models in which nodes are represented by vectors $(x_i)_{i\in V}$ in $\mathbb{R}^d$, where d is supposed to be small, and the probability of an edge between vertices i and j is defined as a function of the dot product of their associated vectors:

$$\mathbb{P}_{DP((x_i)_i)}[A_{i,j} = 1] = \mathrm{f}\left(x_i^T \cdot x_j\right)$$

Authors study various embedding and probability function to show that those models are able to reproduce observed properties of real networks such as clustering and heavy-tailed degree distribution.

Relying on the dot product implies that only undirected graphs can be modeled, as it assigns symmetric roles to the vectors x_i and x_j. However, more general functions can be considered as well. For example, stochastic blockmodels can be considered as a special case of graph embedding. The SBM defined by the partition $b_1, \ldots, b_p$ and the probability matrix P can be described using the canonical base $(e_i)_{i\in[\![0,p-1]\!]}$ of $\mathbb{R}^d$, by associating to each node $u \in b_i$ the vector e_i and defining

$$\mathrm{f}(e_i, e_j) = P_{i,j}$$

Embedding node in an euclidian space allows to use generic tools on vectorial data, for example to perform clustering Lyzinski et al. [2016]. A review of graph embedding techniques and usage can be found in Cai et al. [2018].

As we see, the wealth of structure in networks has sparcked the development of a large variety of probabilistic classes of graph models. Those classes have a direct practical interest in the sense that, given a model class and a set of parameter θ, one can use the probability distribution $\mathbb{P}_\theta$ to generate synthetic networks mimicking real networks structure. These can then serve as benchmarks to test network algorithms, or as null models to be compared with real networks. In practice though, one often needs to generate graphs that reproduce the structure of an observed network G. This means that the parameter set θ cannot be chosen arbitrarily by the user, but should be infered on observed data.

Parameter inference

Parameter inference is the inverse problem of graph generation. Given a model class Φ and a set of parameter $\theta \in \Theta$, one can generate a graph G by randomly picking it from Ω following the probability distribution $\mathbb{P}_\theta$. On the other hand, the problem of parameter inference is:

given an observed graph G and a model class Φ, to find the parameter set $\theta^\star \in \Theta$ such that the model $\mathbb{P}_{\theta^\star}$ best fits (in some sense that will be detailed later on) the graph G. For certain classes of models, this inference is obvious. In the case of the hard configuration model, for example, the only degree sequence $(k_i^\star)_{i\in V}$ such that $\mathbb{P}_{CM((k_i^\star)_i)}[G] \neq 0$ is given by

$$\forall i \in V, k_i^\star = \deg_G(i)$$

For other classes, there exist several parameters θ such that $\mathbb{P}_\theta[G] \neq 0$, each of the corresponding model $\mathbb{P}_\theta$ is a *candidate model* for G. For example, in the stochastic blockmodel class, each partition $\mathcal{B} = (b_1, \dots, b_p)$ of V corresponds to a candidate model M_B.

Many methods exist to select the best parameter set $\theta^\star$. In this section, after illustrating why a statistical approach to parameter inference is necessary, I review the statistical methods that have been used to rigorously compute the likelihood that an observed network was generated by a candidate model with a specific parameter set.

Non statistical parameter inference

To select the optimal parameter $\theta^\star$ given an observed graph G, one must define an objective function $\Psi : (\Theta, \Omega) \to \mathbb{R}$ such that

$$\theta^\star = \underset{\theta\in\Theta}{\operatorname{argmax}}\ \Psi(\theta, G)$$

Various objective function can be defined to infer the parameters of a model class. A good example of this is community detection: since the fundamental article Girvan and Newman [2002], many methods to partition the nodes of a graph have been proposed, relying on various definitions of a community and as many algorithms to detect them, which do not lead to the same partition for a given graph. A review can be found in Fortunato and Hric [2016].

What is more, even when the graph is generated with known block structure, they do not necessarily retrieve it. A typical example is the famous modularity function, which was introduced by Girvan and Newman Newman [2004, 2006]; Newman and Girvan [2004]. It relies on the fact that the expected number of edges between two nodes u and v with degrees k_u and k_v is, according to the configuration model, $\mathbb{E}[W_{u,v}] = \frac{k_u k_v}{2m}$. Consequently, they define the modularity of a partition of the nodes $B = b_1, \dots, b_p$ as

$$Q(\mathcal{B}, G) = \sum_{i=1}^{p} \sum_{u,v\in b_i} A_{u,v} - \frac{k_u k_v}{2m}$$

They argue that maximizing the modularity of the partition corresponds to groups of nodes which are more densely connected than one would expect if edges were randomly distributed, and therefore select the partition $\mathcal{B}^\star$ as

$$\mathcal{B}^\star = \underset{\mathcal{B}}{\operatorname{argmax}}\ Q(\mathcal{B}, G)$$

However, communities found this way are not necessarly meaningful. Guimerà, Sales-Padro and Amaral showed that modular partitions can be found even in random graphs

Guimera et al. [2004], as illustrated in figure 2.2. This is due to the fact that density heterogeneities in the distribution of edges arise from random fluctuation, an issue which was further investigated in Reichardt and Bornholdt [2006]. What is more, modularity appears to have a resolution limit: it is biased toward communities of a specific size, therefore splitting larger communities and merging smaller ones Fortunato and Barthelemy [2007]; Lancichinetti and Fortunato [2011]. Indeed, finding communities in graphs is not just about discovering densely interconnected groups of nodes. One also has to determine whether these heterogeneities are stronger than expected from random fluctuations. Without considering this statistical significance, the node partition found cannot be rigorously interpreted.

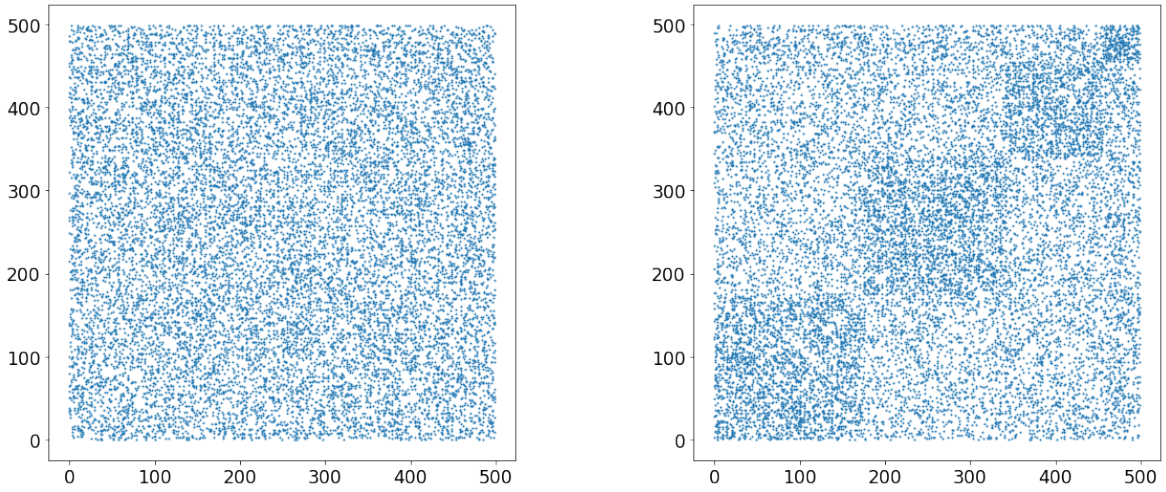


Figure 2.2 – Adjacency matrix of a random graph (left), and the same matrix with nodes reordered according to the maximum modularity partition (right). Modularity optimization detects variation of densities that are present in the graph, but they are due to random fluctuations and not to a block structure .

Graph embedding also often relies on the optimization of ad hoc objective functions Goyal and Ferrara [2018]. As the feature vectors are learned using machine learning techniques, the corresponding model has no direct interpretation and must be used as a black box. What is more, as underlined in Tang et al. [2015, 2017], different embeddings can lead to the same probability distributions, which makes even harder the inference, interpretation and comparison of models learned from different observations.

Canonical and microcanonical ensembles

These issues illustrate the need for a principled parameter inference methodology. In particular, the use of probabilistic classes of models makes it natural to use the likelihood of a parameter set $\mathcal{L}(\theta, G)$ as an objective function

$$\begin{aligned}\theta^\star &= \underset{\theta\in\Theta}{\operatorname{argmax}}\ \mathcal{L}(\theta, G)\\ &= \underset{\theta\in\Theta}{\operatorname{argmax}}\ \mathbb{P}_\theta[G]\end{aligned}$$

This way, the selected set of parameter $\theta^\star$ has a natural interpretation as the parameter set whose associated model generates the observed network with the highest probability.

However, by itself likelihood maximisation is not sufficient to ensure that the infered parameter set can be properly interpreted. Indeed, we have seen in the first section that the definition of a model class $\phi : \theta \mapsto \mathbb{P}_\theta$ may introduce bias in the probability distribution associated with a parameter set. For example, if one defines the configuration model based on the switching algorithm, the associated probability distribution $\mathbb{P}_{CM((k_i)_i)}$ is not uniform on all possible graphs with the prescribed degree sequence. Therefore, the fact that an observed graph G is generated with high probability does not necessarily mean that its structure is determined by its degree distribution. It may be an artefact of the model class definition.

To avoid such bias, statistical methods for graph modeling have been developed Cimini et al. [2018]; Park and Newman [2004a]. They fundamentaly shift the modeling perspective: empirical modeling starts by defining an algorithm to generate graphs with prescribed properties, and then studies the probability distribution induced by this algorithm; on the other hand, statistical modeling considers those properties as a constraint imposed on the graph and define the associated probability distribution as the maximaly random distribution under these constraints, independently of the mechanisms that could lead to it.

It is analogous to the statistical mechanics treatment of a gas made of a large number n of particles. Just as the state of such a physical system is characterized by a large number of quantities (at least the position p_i and velocity v_i of each particle i), the state of a graph is characterized by the set of edges E which connect its nodes. In both cases, determining the exact state s_t of the system from an initial state s_0 and some intrinsic laws is intractable. Therefore, rather than trying to compute this exact state, statistical modeling enumerates all the possible states and considers the probability for the system to be in each of these states. In Jaynes [1957], the author defends such a modeling procedure in statistical physics by arguing that even though we have no guarantee that the obtained model is the right one, it is the best possible hypothesis that can be made given available observations: the same argument is valid as well for graphs.

To define the probability distribution, statistical modeling relies on a set of constraints. In statistical physics, it can typically be macroscopic quantities such as the temperature, pressure, etc. which can be measured on a real system. In graph modeling, these measureable quantities are connectivity measures which will be denoted $(\epsilon_i)_{i\in[\![0,d-1]\!]}$, along with a set of objective values $(\eta_i)_{i\in[\![0,d-1]\!]}$ (typically measured on the real network which is studied). Examples of common statistical graph classes and their associated connectivity measures are presented in table 2.1.

Table 2.1 – Common statistical graph models and their associated properties

Model M	Connectivy measure $(\epsilon_i)_i$
Erdös-Renyi	number of nodes and number of edges
Configuration model	degree distribution
Stochastic block model	block to block density
Gravity model	node position, strength, and deterrence function
Radiation model	node position and strength

Given the connectivity measures and the objective values, there exists two methods to

define the associated statistical model.

Microcanonical ensemble The first is to define the distribution on the so-called microcanonical ensemble

$$\Omega_{(\eta_i)_i} = \{H = (V, E_H) \mid \forall i, \epsilon_i(H) = \eta_i\}$$

In this case, the constraints are said to be hard, because any state of the system which does not fit the constraints has a null probability. For the states H which are in $\Omega_{(\eta_i)_i}$, there is no reason to favour one of them, so the probability distribution is uniform

$$\forall H \in \Omega_{(\eta_i)_i}, \mathbb{P}[H] = \frac{1}{|\Omega_{(\eta_i)_i}|}$$

Let's remark that this probability distribution is the one which maximizes Shannon's entropy $\mathbb{S}[\mathbb{P}] = \sum_{H \in \Omega_{(\eta_i)_i}} \mathbb{P}[H] \log_2(\mathbb{P}[H])$, under the constraint that $\forall H \notin \Omega_{(\eta_i)_i}, \mathbb{P}[H] = 0$, which ensures that no undesired bias is added to the model.

This modeling methodology is simple, but computing the probability distribution requires to compute the cardinal of the microcanonical ensemble, also called its entropy. For classes as simple as the configuration model, this issue is known to be difficult. Asymptotic values for the number of graphs with specified degree sequence were found as soon as 1974 Bender [1974]; Bender and Canfield [1978], but there still is no exact general formula. Exact and approximate values for other microcanonical ensembles' entropy where computed in Bianconi [2008, 2009]; Coon et al. [2018]; Peixoto [2012].

Canonical ensemble To circumvent this obstacle, one can relax the constraints and impose that the objective values are matched only on average. In this case, the probability distribution is defined on the canonical ensemble, *i.e.* the set of all possible graphs on nodes V

$$\Omega_V = \{H = (V, E)\}$$

Depending on the graph studied, this ensemble can be restricted to undirected graphs, simple graphs, to forbid self-loops, etc. The probability distribution is then defined as the maximum entropy distribution which satisfies

$$\forall i, \mathbb{E}[\epsilon_i(H)] = \eta_i$$

As developed in Barndorff-Nielsen [2014], such a distribution can always be written

$$\mathbb{P}[G] = \frac{1}{Z} \exp\left(\sum_{i=0}^{d-1} \theta_i(\bar{\eta}) \epsilon_i(G)\right)$$

where Z is the normalization constant (also called partition function) defined as

$$Z = \sum_G \exp\left(\sum_{i=0}^{d-1} \theta_i(\bar{\eta}) \epsilon_i(G)\right)$$

To fit such a model, one must compute the values of the parameters $(\theta_i)_i$. Theoretically, the derivative of $\ln(Z)$ according to a given θ_i gives the expected value of the corresponding observable

$$\begin{aligned}\frac{\partial \ln(Z)}{\partial \theta_i} &= \frac{1}{Z}\sum_G \epsilon_i(G)\exp\left(\sum_{i=0}^{d-1}\theta_i(\bar{\eta})\epsilon_i(G)\right)\\ &= \mathbb{E}[\epsilon_i(G)]\end{aligned}$$

Thus, if these derivatives can be computed explicitly with respect to $(\theta_i)_i$, imposing that $\forall i \in [\![1, d]\!], \mathbb{E}[\epsilon_i(G)] = \eta_i$ gives a set of d equations with d unknowns to solve, whose solution corresponds to the more likely set of parameters $(\hat{\theta}_i)_i$. However, in practice, these derivatives may be hard to compute or give rise to intricated systems of equations.

Both methods have been used to define statistical models, so we will briefly review these applications.

Exponential random graphs

The first attempt to apply statistical physics tools to graph modeling dates back to the 1980's with what would then be called exponential random graphs. They were designed to go beyond the configuration model and capture higher order correlations between edges Anderson et al. [1999]; Wasserman and Pattison [1996].

To do so, these classes rely on the canonical ensemble. In Holland and Leinhardt [1981], Holland and Leinhardt consider the number of node pairs (i, j) in a directed graph which are mutually connected, in order to evaluate the tendency of interactions to be reciprocal. In Park and Newman [2004b], Park and Newman study the case of 2-stars. However, to account for more complex correlations between edges, one has to introduce more complex patterns like triangles, stars, etc. which involve a higher number of nodes and edges. Doing so, the explicit formula for the partition function Z implies to enumerate those patterns on all possible graphs in Ω, which is impossible as soon as the number of nodes grows above 10.

In Franck O. and Strauss D. [1986], Franck and Strauss considered the number of triangles and stars to account for correlations between edges sharing at least one node (a model they call Markov graphs). To overcome the partition function computation issue, they introduce a pseudolikelihood estimator to estimate values for the parameters $(\theta_i)_i$. This methodology was then adapted by Wasserman and Pattison in Wasserman and Pattison [1996] and Wasserman, Anderson and Crouch in Anderson et al. [1999] to the more general case of an arbitrary set of connectivity measures $\bar{\epsilon}$. However, the pseudolikelihood estimator neglects the fact that pattern count are not independent from one another. For example, triangles contains 2-stars, and thus increasing the number of triangles in a graph implies that the number of 2-stars will increase too. What is more, the larger the patterns considered, the more likely they are to overlap and the more hazardous it is to neglect those dependencies. In practice, it causes models to be easily degenerate Handcock et al. [2003]; Newman [2003] and very sensitive Van Duijn et al. [2009].

Stochastic blockmodels. In Hastings [2006], Hastings uses the exponential random graphs formalism to determine the most likely node affiliation in p blocks, with a connection prob-

ability inside blocks p_{in} and outside blocks p_{out}. Apart from its statistical rigor, this method is more flexible than modularity maximization as it allows to seek groups of nodes with similar connectivity patterns beyond assortativity (*i.e.* nodes belonging to the same group connects more densely with each other). In Newman and Leicht [2007], authors extend it to an arbitrary number p of blocks with no prescribed connectivity pattern, and in Karrer and Newman [2010], to the case of a degree-corrected stochastic blockmodel, which takes into account both the degree distribution and a decomposition of the graph in blocks, and show that it is able to outperform classical stochastic blockmodels in detecting communities in real networks. This formalism also provided a framework to study the detectability of communities and allowed to identify density threshold beyond which even perfectly defined communities cannot be recovered Abbe and Sandon [2015]; Decelle et al. [2011a,b]; Hu et al. [2012].

Microcanonical stochastic blockmodel inference

Statistical stochastic blockmodeling highlighted another pitfall for parameter inference procedure. Apart from the bias toward specific values, any inference procedure should incorporate mechanisms to prevent overfitting. This is particularly obvious in the case of stochastic blockmodel, because there is one trivial node partition which always leads to the best possible fit: the total partition where $\mathcal{B}_{tot} = (b_i = \{i\})_{i \in V}$, whose corresponding stochastic blockmodel generates the observed graph with probability 1. This model perfectly fits observed data, but it is useless as it does not extract any structure from it. Indeed, a good model should both fit the observations and simplify them. Previously mentioned papers solved this problem by imposing the number of blocks p to seek in the graph. However, most of the time this information is not available beforehand and one needs to infer it too.

In Peixoto [2013], the author presents a methodology based on the maximum description length (MDL)Grunwald [2004] to do so. The model takes as parameters a node partition $\mathcal{B}$ and a block adjacency matrix M whose entries M_{rs} corresponds to the number of edges going from block r to block s. The model is based on the microcanonical ensemble

$$\Omega_{SBM(\mathcal{B},M)} = \left\{ G \mid \forall r,s \in [\![0, p-1]\!]^2, \sum_{i \in b_r, j \in b_s} G_{ij} = M_{rs} \right\}$$

To obtain the probability that a given graph G was generated using parameters $\mathcal{B}$ and M, the formula is reversed using Baye's theorem.

$$\mathbb{P}[\mathcal{B}, M|G] = \frac{\mathbb{P}[G|\mathcal{B}, M] \times \mathbb{P}[\mathcal{B}, M]}{\mathbb{P}[G]}$$

where $\mathbb{P}[\mathcal{B}, M]$ is a prior probability distribution on the parameter set. As $\mathbb{P}[G]$ does not depend on $\mathcal{B}$ and M, finding the most probable parameters given G amounts to maximizing $\mathbb{P}[G|\mathcal{B}, M] \times \mathbb{P}[\mathcal{B}, M]$ To avoid overfitting, $\mathbb{P}[\mathcal{B}, M]$ imposes a lower probability to finer partitions. This way, refining the partition reduces $\mathbb{P}[G|\mathcal{B}, M] = \frac{1}{|\Omega_{SBM(\mathcal{B},M)}|}$ but at the same time increases $\mathbb{P}[\mathcal{B}, M]$. This is made possible thanks to the unification of the parameter set of all stochastic blockmodels, which allows to define a unique prior distribution on all possible parameters. The difficulty then shifts to the definition of the prior distribution, which must be designed not to introduce new biases.

This method has been applied to degree-corrected stochastic blockmodels, nested stochastic blockmodels Peixoto [2019], and weighted stochastic blockmodels Peixoto [2018]. All these contributions remain within the scope of finding a block structure, but statistical modeling paves the way for a wider comparison methodology between graph classes. Indeed, most classes of models are designed to capture one aspect of the graph structure, and neglect the other ones. Even the stochastic blockmodels, which are flexible enough to describe both assortative or dissasortative, core-periphery, and nested structure cannot capture, for instance, the spatial structure of a network. Yet, it is very unlikely that any real network's structure can be explained as the result of a single mechanism: nodes' interaction intensity, transitivity, communities, distances, and other factors all have an impact on the resulting distribution of edges. Thus, nearly any class of model can be fitted on a dataset and give some information about the network. In order to determine whether one of these factors dominates the other, it is necessary to be able to compare them both in terms of accuracy and simplicity.

Model selection

Generaly speaking, the problem of model selection is, given a set of candidate models $(M_i)_{i\in Q}$ to find the best one with respect to an observed dataset. It is a classical problem in data analysis: linear or polynomial regression, clustering in $\mathbb{R}^d$, or fitting a normal distribution are typical example of model selection Ding et al. [2018]. Parameter inference, as described in the previous section, is a particular case of model selection in which the candidate models belong to the same model class. They can thus be indexed by their associated parameters, which all lie in the same parameter space Θ. For example, to fit a stochastic blockmodel, one explores the space of all node partitions to find the most appropriate one. In this section, we consider a more general problem in which models do not belong to the same class, which imply that their parameters do not necessarily lie in the same parameter space.

A central issue in selecting a model, especially parametrical models, is to avoid both overfitting and underfitting. Generally speaking, by increasing the number of parameters of a model, it can be made arbitrarily close to the data. For example, when fitting a polynomial curve of unknown degree on a set of n points, increasing the degree of the polynomial increases the number of degrees of freedom of the model, up to the point where a polynomial of degree $n-1$ can pass exactly through each point in the set. Yet, such an overfitted model is useless as it lacks generalizability and may not fit new data produced by the same mechanism. On the other hand, an underfitted model will lack accuracy and neglect important piece of information present in the observations, as illustrated on figure 2.3

Performing model selection on graphs brings about new challenges as classical model selection technique were designed for numerical of vectorial datasets. Of course, through its adjacency matrix a graph can be considered as a vector and statistical models can be seen as probability distributions on subset of $\mathbb{R}^{n^2}$. Model selection thus boils down to finding the probability distribution that best fits the observed dataset, which is a common task in statistical inference. The main problem is that in the vast majority of cases, the dataset to be modeled consists of a single network while statistical inference requires many observations to be sound.

This issue questions the very definition of overfitted and underfitted model: as there is a single observation, there is no clue to the intrinsic variability of the dataset with which to

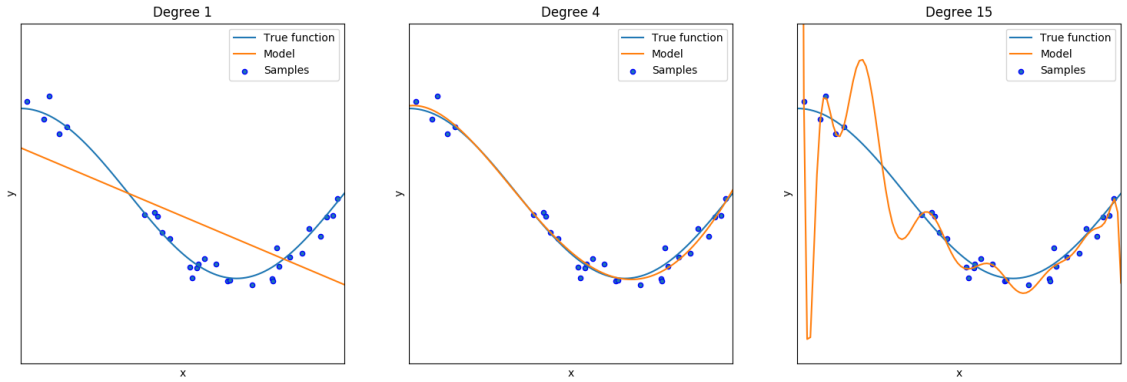


Figure 2.3 – Examples of underfitted (left) and overfitted (right) models

compare the model. In Ghasemian et al. [2019], authors tackle this question, showing that there is a trade-off between a model's capacity to describe the edges present in the observed graph and its capacity to predict new edges produced by the same mechanism. They identify this trade-off as a way to determine whether a model is overfited or underfited. Doing so, they slightly change the modeling perspective by considering the edges and not the whole graph as the fundamental observations. We will see in the thesis the possible consequences of such a perspective shift.

In the following, we present some classical model selection techniques and how they were adapted to the specific case of graph dataset.

Frequentist inference and statistical tests

Frequentist inference relies on convergence theorems like the central limit theorem to design statistical tests. Assuming that a sequence of random variables $(X_i)_{i\in[\![0,n-1]\!]}$ were independently generated from the same distribution $\mathbb{P}_0$, it states that the distribution $\bar{\mathbb{P}}$ of their mean $\bar{X}_n = \frac{1}{n}\sum_{i=0}^{n-1} X_i$ converges toward a normal law $\mathbb{P}_{\mathcal{N}(\mu,\sigma)}$.

Statistical tests Based on this theorem, various tests have been designed to check whether a sample $(x_i)_{i\in[\![0,n-1]\!]}$ was generated from a distribution $\mathbb{P}_0$, whether two samples were generated from the same distribution, and so on. As these tests rely on asymptotic results, their application requires the sample size n to be large enough to justify the use of the limit distribution. These tests were originaly designed for real-valued random variable so they need to be adapted to graphs. In practice, efforts in this domain have focused on determining whether two graphs (or two graph samples) were generated by the same model Fraiman and Fraiman [2018]; Fujita et al. [2020]; Ghoshdastidar et al. [2017]; Tang et al. [2014, 2017]. This allows to perform statistical tests on models' parameters, which are real numbers, but it does not provide a statistical procedure to determine which model is the more relevant given an observed graph.

In Takahashi et al. [2012], authors develop a fitting procedure based on the comparison of the observed graph spectrum with the average spectrum of graphs generated by a model.

However, this procedure is based on information theoretic criterion rather than statistical inference, and only serves as a preliminary phase for the statistical comparison of graph samples. In Cerqueira et al. [2017], a methodology is presented to evaluate the statistical relevance of a model with respect to a graph sample, based on the distance between the average graph of the sample and the mean of the model distribution. This method can indeed be used to statisticaly infer the most fitted model, but it relies on a graph sample which must be sufficiently large, while in most real situation only a single graph is observed.

Graph distances This last work also brings forward the question of the distance between graphs. Statistical tests on real numbers rely on an intuitive notion of distance, for example to compute confidence interval. When considering graphs, the distance which is used is in itself an important question as it may change the appreciation of whether a model generates graph similar to those observed, and thus its relevance.

In Cerqueira et al. [2017], authors rely on euclidian space distances. Given a labeling function $\phi : V \to [\![0, n-1]\!]$, there is a one-to-one correspondance between graphs and adjacency matrices so any distance over $\mathcal{M}_n(\mathbb{R})$ can be applied to graphs. In particular, the $\mathcal{L}_1$ distance

$$\mathrm{d}_1(G, H) = \sum_{i,j} |G_{i,j} - H_{i,j}|$$

has a natural interpretation as it counts the number of differences between edges in two graphs G and H over the same set of vertices. In this context, it is called the *edit distance*. However, this interpretation is correct only under the assumption that node labeling is the same between graph G and H. In our case, this means that models must keep node labels unchanged when generating random graphs.

Apart from the labeling question, the edit distance as well as other purely algebraic distances do not take into account the topological significance of different edges when computing the distance. For example in figure 2.4, there is a single edge of difference both between graph G_1 and G_2 and between graph G_1 and G_3. However, deleting the bridge edge has much more impact on the overall topology of the graph, which is not accounted for by the edit distance. It is to take into account those topological properties that graph-specific distances where designed, such as DeltaCon Koutra et al. [2013], the Resistance-perturbation distance Monnig and Meyer [2016] or the Network portrait divergence Bagrow and Bollt [2019]. A review of existing graph distances can be found in Wills and Meyer [2019].

Bayesian inference

In situations where the central limit theorem hardly applies, statistical inference may be performed using bayesian techniques. Theoretically, it is grounded on representation theorems which originated with De Finetti's work on exchangeable random sequences Finetti [1937]. A sequence of random variables $(X_i)_{i\in\mathbb{N}}$ taking value in a space Ω is said to be *exchangeable*, if for any permutation σ of $\mathbb{N}$ and any sequence of subset of Ω, $(A_i)_{i\in\mathbb{N}}$

$$\mathbb{P}[(X_i \in A_i)_{i\in\mathbb{N}}] = \mathbb{P}[(X_{\sigma(i)} \in A_i)_{i\in\mathbb{N}}]$$

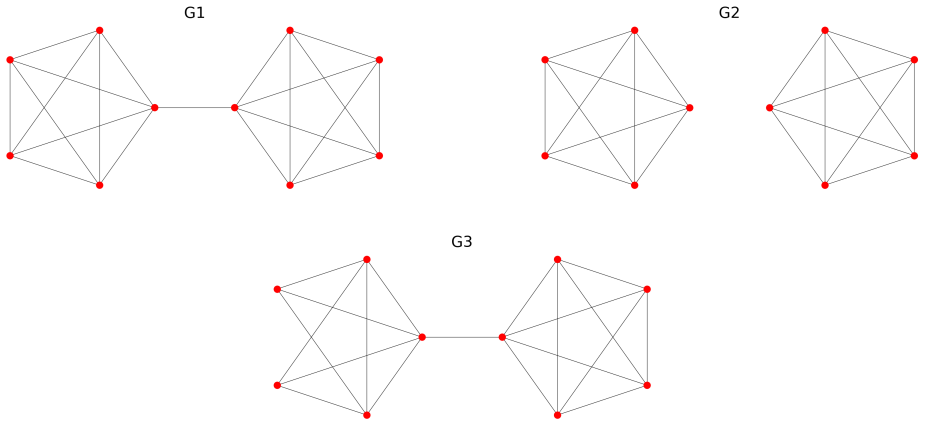


Figure 2.4 – Example of the non equivalence of edges in computing distances between graphs

De Finetti's theorem states that for such a sequence, there always exists a probability measure μ on the set $\mathrm{Prob}(\Omega)$ of probability distributions on Ω such that

$$\mathbb{P}[(X_i \in A_i)_{i\in\mathbb{N}}] = \int_{\mathbb{Q}\in\mathrm{Prob}(\Omega)} \prod_{i\in\mathbb{N}} \mathbb{Q}(A_i)\, \mathrm{d}\mu(\mathbb{Q})$$

This result can and is frequently understood as a two stage procedure to generate the random sequence $(X_i)_i$:

- first, a probability distribution $\mathbb{Q}$ (the model), is drawn at random, following the *prior distribution* μ.
- second, the X_i are generated independently, following the distribution $\mathbb{Q}$.

In particular, the joint distribution for the model and the observations can be written

$$\begin{aligned}\mathbb{P}[(X_i)_i, \mathbb{Q}] &= \mathbb{P}[(X_i)_i|\mathbb{Q}] \times \mathrm{d}\mu(\mathbb{Q}) \\ &= \prod_i \mathbb{Q}[X_i] \times \mathrm{d}\mu(\mathbb{Q})\end{aligned}$$

In order to select the model that most likely generated a sequence of observations $(x_i)_{i\in[\![0,n-1]\!]}$ given a prior distribution μ, this formula can be reversed using Bayes' theorem

$$\begin{aligned}\hat{Q} &= \operatorname*{argmin}_{\mathbb{Q}\in\mathrm{Prob}(\Omega)} \mathbb{P}[\mathbb{Q}|(x_i)_{i\in[\![0,n-1]\!]}] \\ &= \operatorname*{argmin}_{\mathbb{Q}\in\mathrm{Prob}(\Omega)} \frac{\mathbb{P}[(X_i)_{i\in[\![0,n-1]\!]}|\mathbb{Q}] \times \mathrm{d}\mu(\mathbb{Q})}{\mathbb{P}[(x_i)_{i\in[\![0,n-1]\!]}]} \\ &= \operatorname*{argmin}_{\mathbb{Q}\in\mathrm{Prob}(\Omega)} \prod_{i=0}^{n-1} \mathbb{Q}[x_i] \times \mathrm{d}\mu[\mathbb{Q}] \text{ as } \mathbb{P}[(x_i)_{i\in[\![0,n-1]\!]}] \text{ does not depend on } \mathbb{Q}\end{aligned}$$

In practice, two difficulties arise to use such techniques in statistical inference. First, given a dataset, there might be different possibilities to define the exchangeable sequence on which the representation theorem is applied, and they lead to different model. Most modelisation methods based on this bayesian framework have relied on the hypothesis that vertices are exchangeable. It allows to use the random arrays representation theorem by Aldous Aldous [1981] and Hoover Hoover [1979]. These results have been widely applied to graph modelisation, and have shown nice connections with the theory of graph limits Lovász and Szegedy [2006], giving rise to the theory of graphons Bollobás and Riordan [2011]; Diaconis and Janson [2007]; Orbanz and Roy [2015]. The major limit this theory suffers from to model real network is that it can only model dense graphs (*i.e.* the number of edges in the graph scales proportionaly to $|V|^2$). Indeed, the only graphon which produces sparse graph is the zero-graphon which generates graphs with no edges.

The stochastic blockmodel inference presented in section 2.3.4 can also be seen as an application of this methodology, in which the exchangeable element is implicitly considered to be the whole graph and the model space is restricted to stochastic blockmodels. However, it implies that when a single graph is observed, the sequence of observations contains a single element, which is problematic to perform statistical inference. Finally, in two articles, Crane and Dempsey advocated for the use of edge exchangeable models Crane and Dempsey [2016, 2018], showing that they can be used to model sparse networks.

Minimum description length

Once the exchangeable elements are chosen, they define a value space Ω and thus a model space $\text{Prob}(\Omega)$, but one still has to define the prior distribution μ in order not to bias the model selection. To do so, the connection between bayesian inference and the minimum description length (MDL) principle for model selection Barron et al. [1998]; Grunwald [2004]; Grünwald and Roos [2019] prooves useful. This principle steams from the idea that a good model needs to satisfy two antagonistic requirement: it must fit to the observations, but at the same time simplify them in order to extract their structure and remove noise. The best model is thus defined as the one which compress the most the observations. The description length of the observations $(x_i)_{i\in[\![0,n-1]\!]}$ using a model $\mathbb{Q}$ is defined as

$$\text{DL}((x_i)_{i\in[\![0,n-1]\!]}, \mathbb{Q}) = \text{DL}((x_i)_{i\in[\![0,n-1]\!]}|\mathbb{Q}) + \text{DL}(\mathbb{Q})$$

where $\text{DL}(\mathbb{Q})$ is the description length of the model and $\text{DL}((x_i)_{i\in[\![0,n-1]\!]}|\mathbb{Q})$ the description length of the dataset, given the model. The first term corresponds to the complexity of the model: the more parameters it has, the longer it is to encode. The second term corresponds to the accuracy of the description: the closer the model fits to the dataset, the less additional information is required to describe $(x_i)_{i\in[\![0,n-1]\!]}$ when $\mathbb{Q}$ is known.

To compute these description length, one needs a quantitative theory of information. It can be defined in different ways depending on the type of encoding Kolmogorov [1968], but the most common one is the one defined by Shannon in 1948 Shannon [1948] which relies on statistical regularities to encode messages. Given a set of messages Ω, and a source which picks independently at random messages following a probability distribution $\mathbb{P}$ and sends them over a binary channel, it can be shown that the best average compression is obtained when the code length of a message $x \in \Omega$ is $\text{DL}(x) = -\log_2(\mathbb{P}[x])$ Grunwald [2004]. Therefor,

with the notations introduced above, the total description length can be written as

$$\begin{aligned}\mathrm{DL}((x_i)_{i\in[\![0,n-1]\!]},\mathbb{Q}) &= -\log_2(\mathbb{P}[(x_i)_{i\in[\![0,n-1]\!]},\mathbb{Q}]) \\ &= -\log_2\left(\prod_{i=0}^{n-1}\mathbb{Q}[x_i]\times \mathrm{d}\mu(\mathbb{Q})\right) \\ &= -\sum_{i=0}^{n-1}\log_2(\mathbb{Q}[x_i]) - \log_2(\mathrm{d}\mu(\mathbb{Q}))\end{aligned}$$

This expression highlights that the prior distribution μ captures the complexity of the model, by assigning lower probabilities (and thus longer description length) to more complex models. This way it counterbalances the fact that a model with more parameters will always be able to fit more tightly to observations and thus predict them with higher accuracy (leading to shorter description length). What is more, this expression also implies that, as the size of the sample grows, the relative weight of the prior distribution decreases, which is coherent with the fact that the risk of overfitting decreases as the sample grows.

Conclusion

As we have seen, there is a wide spectrum of graph models and modeling frameworks, many of them being parametric and thus susceptible to be fitted to any particular observed network. To capture the fundamental structure of a network, one must be able to sort out the most relevant ones which implies to compare their relative performance. This is necessary as well for models whose parameters can range over several orders of granularity, such as the node partition of blockmodels or the deterence function of the gravity model, as a too low level of granularity may lead to underfitting and a too high level to overfitting.

The main difficulty in performing such performance comparison is to find meaningful criterions which can be applied to models whose parameters do not belong to the same parameter space. While one may compare community detection algorithms based on their ability to recover a known planted partition, There is no standard procedure to compare the relevance of a stochastic blockmodel and a configuration model in modeling an observed network. Being probabilistic models, they all incorporate probability distributions that can be used to evaluate their prediction power, but those probability distributions are not always defined on the same sets, and sometimes we do not even have an explicit formula for them (for example in the case of the configuration model). What is more, several models' distributions are defined on sets of graphs (microcanonical and canonical ensembles), which implies that their fitting on a single network amounts to statistical inference on a single observation. This undermines any attempt to interpret rigorously the results of such an hazardous inference process.

In this thesis, I first present in chapter 3 a statistical test methodology to evaluate the probability that an observed graph was generated by a candidate model. To do so, I introduce and study the properties of graph spaces, the set of graphs that a given model can generate, whose probabilitic structure is enriched with a geometric structure. This approach is inspired by the frequentist inference methodology, while the rest of the thesis focuses on bayesian inference model selection techniques. In chapter 4, I study bayesian stochastic

blockmodel inference based on the microcanonical ensemble. I show that using the entropy of the microcanonical ensemble to measure the complexity of a model leads to a bias toward overfitted model, which can only be mitigated by the introduction of some specific prior distribution over candidates models. Finally, in chapter 5, I introduce edge statistical models, a reformulation of statistical models as probability distributions on sequence of edges to avoid the single observation inference issue. I then show how bayesian inference can be applied to this new formulation and illustrate its results both on stochastic blockmodel inference and on stochastic blockmodel and configuration model comparison.

Some of these contributions were published in the Complex Networks conferences in 2019 and 2020 Duvivier et al. [2019, 2020]. Two others have been submitted to the IEEE TNSE journal Duvivier et al. [2021b] and to the Journal of Complex Networks Duvivier et al. [2021a].

Chapter 3

Statistical test over a metric microcanonical ensemble

Statistical modeling provides a common formulation for various graph models, which is useful to perform model selection. By formalizing them as probability distributions over sets of graphs, it allows to evaluate the likelihood that an observed graph was generated by a candidate model. This is particularly interesting in the case of the microcanonical ensemble, as the probability distribution is fully characterized by the entropy of the associated ensemble, *i.e.* the logarithm of its cardinal. Therefore, many papers have focused on computing the size of various graph ensemble Bianconi [2009], Peixoto [2012], Zingg et al. [2019]. A whole methodology for community detection based on those principles has been developed in the case of stochastic blockmodels Peixoto [2019]. The issue is that, by considering models as unstructured sets, this approach neglects graphs topology.

Indeed, graphs are also geometrical objects, in the sense that one can define distances between them. Such a distance induces a structure on a model's ensemble. Much work has been devoted to quantifying how similar two graphs are Wills and Meyer [2019], especially from a topological point of view Monnig and Meyer [2016], Koutra et al. [2013]. These distances between two graphs can be generalized to evaluate the quality of a model by computing a distance between an observed graph and the graph ensemble associated to a model. For example, the widely used measure for community detection known as modularity Newman [2006] evaluate the quality of a partition by comparing the edge weight in the observed graph with the expected edge weight of the graph in the configuration model. In this case, the problem is to evaluate the statistical significance of the results, in order not to mistake noise for structure, as was pointed out in Guimera et al. [2004].

In this chapter, I first review in section 3.1.1 and 3.1.2 existing techniques to measure the relevance of a model with respect to a graph based on the microcanonical ensemble. Then, I introduce in section 3.2 the edit distance expected value, a measure which takes into account both the geometric and the probabilistic structure of the graph ensemble. Finally, I show how this measure can be used to evaluate a model relevance with respect to a given graph in section 3.3.

The microcanonical ensemble

As explained in section 2.3 of the state of the art, the microcanonical ensemble describes the structure of a graph G based on connectivity measures $(\epsilon_i)_i$ and objective values $(\nu_i)_i$ Cimini et al. [2018], which define the set of graphs

$$\Omega_M = \{H \mid \forall i, \epsilon_i(H) = \nu_i\}$$

In the rest of the chapter, we will consider labelled directed multigraphs with self-loops. Although they are not the most widely used in practice, it makes probability derivations easier, especially for the configuration model.

Entropy

The probability distribution $\mathbb{P}$ associated to the microcanonical ensemble Ω_M is the uniform one:

$$\forall H \in \Omega_M, \mathbb{P}(H) = \frac{1}{|\Omega_M|}$$

Thus, computing the probability to choose G among all possible graphs in Ω_M boils down to counting the number of graphs it contains. The Shannon entropy of this distribution is $\mathbb{S}(\mathbb{P}) = \log(|\Omega_M|)$, and it is directly related to the likelihood of a given model. If we observe a graph G and consider a set of models $\mathcal{M} = \{M_1, \ldots, M_p\}$, we can find which model G has most likely been sampled from by maximising its likelihood

$$M^* = \underset{M_i \in \mathcal{M}}{\operatorname{argmax}}\, \mathbb{P}[M_i|G]$$

which can be done using Bayes theorem

$$\mathbb{P}[M_i|G] = \frac{\mathbb{P}[G|M_i] \times \mathbb{P}[M_i]}{\mathbb{P}[G]}$$

If we assume a non-informative uniform prior distribution on all models $\mathbb{P}[M_i] = \frac{1}{p}$, as $\mathbb{P}[G]$ is a constant, maximising the likelihood is equivalent to maximising $\mathbb{P}[G|M_i] = \frac{1}{|\Omega_{M_i}|}$ or minimising the entropy $\log(|\Omega_{M_i}|)$. This idea is developed and applied to the case of stochastic blockmodels in Peixoto [2019].

However, considering the common situation in which one is interested in modeling the global topology of G rather than its precise edge list, likelihood maximisation appears insufficient as a model selection criterion. For example, if we consider the three graphs on figure 3.1, with G_1 as a reference, both G_2 and G_3 are different from G_1, but the topology of G_1 and G_2 is almost the same. Therefore, a model which produces mostly G_3-like graphs cannot be considered as good a model for G_1 as one which produces G_2-like ones. This shows that a purely probabilistic approach to model selection is not fully adequate, at least in the microcanonical framework.

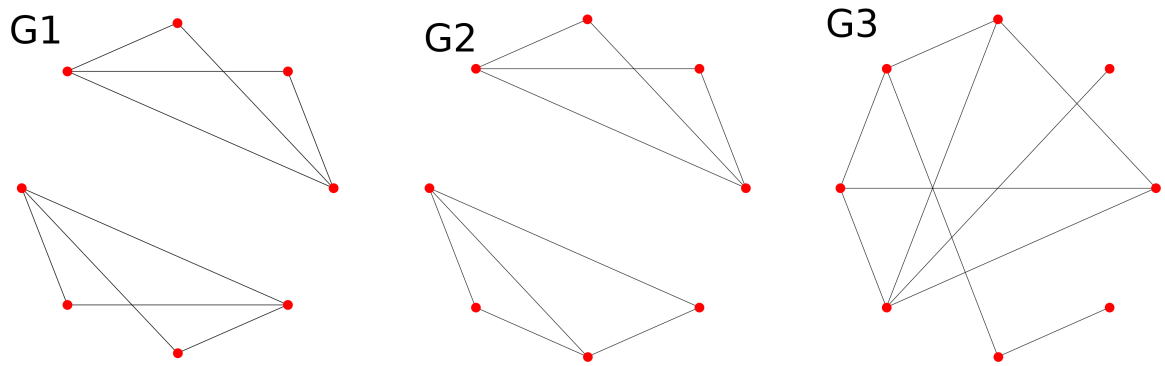


Figure 3.1 – **Three graphs with 8 nodes and 10 edges.** G_1 and G_2 (leftmost) share the same topology even if they are not strictly equal, which is not the case of G_3 (rightmost).

Distance to the barycenter

On the other hand, one may measure how typical G is with respect to Ω_M by comparing it with an appropriate representative of this ensemble, for example its barycenter:

$$G_M = \sum_{H \in \Omega_M} \mathbb{P}(H) H$$

If we denote W_G the weight matrix of graph G, it can easily be derived that

$$\forall (i,j) \in V^2, W_{G_M}(i,j) = \mathbb{E}[W_H(i,j)] \tag{3.1}$$

Remark. *G_M does not necessarily belong to Ω_M. In particular, even if all graphs in Ω_M have whole weight, it does not imply that G_M's edge weights are integers. Examples of barycenter weights for common models are derived in appendix 7.1.1 and given in table 3.1*

Table 3.1 – Statistical graph models' barycenter weight

Model M	Parameters	$G_M(i,j)$	m_{G_M}
Erdös-Renyi	n, m	$\frac{m}{n^2}$	m
Configuration model	$(k_i^{in}, k_i^{out})_{i \in V}$	$\frac{k_i^{out} k_j^{in}}{m}$	m
Stochastic block model	$(m_{r,s})_{r,s}$	$\frac{m_{c_i,c_j}}{\lvert c_i \rvert \lvert c_j \rvert}$	m
Gravity model	$(k_i)_{i \in V}, f$	$k_i k_j f(d(i,j))$	m
Radiation model	$(k_i^{in}, k_i^{out})_{i \in V}$	$\frac{k_i^{out} k_i^{in} k_j^{in}}{(k_i^{in}+s_{ij})(k_i^{in}+k_j^{in}+s_{ij})}$	m

The famous modularity function to evaluate the quality of a node partition $B = (b_1, \ldots, b_p)$ on a graph $G = (V, E)$ with weight matrix W_G is defined as the difference between the

number of edges inside each cluster and the expected number for a random graph with the same degree distribution (*i.e.* following the configuration model). It can be understood as a comparison with the barycenter G_M of the corresponding configuration model.

$$\begin{aligned}\mathrm{Q}(G,B) &= \frac{1}{2m}\sum_{i=1}^{p}\sum_{u,v\in b_i}\left(W_G(u,v) - \frac{k_u^{out}k_v^{in}}{m}\right)\\ &= \frac{1}{2m}\sum_{i=1}^{p}\sum_{u,v\in b_i}\left(W_G(u,v) - W_{G_M}(i,j)\right)\\ &= \mathrm{d}(G,G_M)\end{aligned}$$

A problem is that G is compared with a single graph G_M which is supposed to account for the whole graph ensemble Ω_M. In particular, all information about the dispersion around the barycenter is lost, which undermines any attempt to interpret statistically the results. This in turn shows how a purely geometrical approach to model selection also fails to account for the whole structure of Ω_M.

Graph space and the edit distance expected value

As we have seen, existing techniques to compare a graph G and a model M exploit in different ways the ensemble Ω_M. Entropy based techniques described in section 3.1.1 focus on its cardinal, but they neglect the topological similarities of graphs inside the ensemble. On the other hand, as described in section 3.1.2, an objective function such as the modularity accounts for these similarities, but it does so with a single graph which is supposed to represent the whole set. Reality is more complex: Ω_M is a set of graphs with a probability distribution, and it can be further structured with a metric. Both aspects, probabilistic and geometric, should be taken into account in order to understand the structure of Ω_M, and the plausibility that a graph G was generated by the associated model M.

Definition 2. *A graph space is a triplet* $(\Omega_M, \mathbb{P}, \mathrm{d})$ *where* Ω_M *is a set of graph,* $\mathbb{P} : \Omega_M \to [0,1]$ *is a probability distribution on this set, and* $\mathrm{d} : \Omega_M^2 \to \mathbb{R}^+$ *is a distance on* Ω_M.

Many different measures exist to compute a similarity score between two graphs G and H Wills and Meyer [2019]. One of the simplest is the edit distance. For two graphs on the same vertex sets $G_1 = (V, E_1)$ and $G_2 = (V, E_2)$, it counts the number of differences between their respective sets of edges.

$$\mathrm{ed}(G_1,G_2) = \sum_{(i,j)\in V^2} |W_{G_1}(i,j) - W_{G_2}(i,j)|$$

As expected from its name, edit distance is a distance between graphs. Indeed, if we consider the weight matrix W_G of a graph G as a point in $\mathbb{R}^{n^2}$, the edit distance corresponds to the $\mathcal{L}_1$ distance and for any model M, Ω_M is a subset of $\mathbb{R}^{n^2}$. The dimension prevents any direct drawing of it for graphs with more than 2 nodes, but it is possible to obtain some intuition about its shape.

In the following, we will use a normalized version of the edit distance which can be interpreted as the fraction of different edges between G_1 and G_2.

$$\text{ned}(G_1, G_2) = \frac{1}{2m} \sum_{(i,j)\in V^2} |W_{G_1}(i,j) - W_{G_2}(i,j)|$$

This normalized edit distance is no longer a distance on $\mathbb{R}^{n^2}$. Yet, for all models M considered here, the number of edges m is constant over the set Ω_M. Thus, the normalized edit distance is equivalent to edit distance inside Ω_M and it allows to compare more easily results between various models, because whatever the model M, the distance between any two graphs G_1 and G_2 in Ω_M is at most 1.

Edit distance to the barycenter

Ω_M barycenter has already been introduced in section 3.1.2, where it was used as a proxy for the whole space. Using normalized edit distance, it is possible to check how much graphs in Ω_M are similar to the barycenter G_M. We consider six different models:

1. EM: Erdös-Renyi with 50 nodes and 1000 edges
2. CFM cst: configuration model with 50 nodes and a constant degree distribution ($k_i^{in} = k_i^{out} = 20$)
3. CFM arith: configuration model with 50 nodes and an arithmetic degree distribution ($k_i^{in} = k_i^{out} = i + 1$)
4. SBM hom: stochastic block model with 50 nodes and 5 communities, each having internal density 1.2, and external density 0.2.
5. SBM het: stochastic block model with 50 nodes and 5 communities, with internal density 0.4, 0.8, 1.2, 1.6, 2, and external density 0.2.

For each model M, we pick a random sample $\mathcal{S}_M$ of 100 graphs in Ω_M and for all $G \in \mathcal{S}_M$ we compute the normalized edit distance to the barycenter $\text{ned}(G, G_M)$. Results are shown in figure 3.2.

The first thing to underline is that whatever the model, $\text{ned}(G, G_M)$ is greater than 0.5, which means that most graphs in Ω_M have at most half of their edges in common with G_M. This observation shows that for those models, the graph space is not concentrated around its barycenter. On the contrary, most graphs in Ω_M seem to be at a specific distance from its barycenter, as would happen for a sphere with a radius depending on the model: 0.67 for ER and CFM cst, 0.55 for CFM arith, 0.69 for SBM hom and 0.71 for SBM het.

All models were chosen to have similar entropy, as shown in table 3.2, yet their characteristic distance to the barycenter vary greatly. Furthermore, we observe that these quantities are not positively correlated: CFM arith, which is the model with the larger entropy is also the one which is the most concentrated around its barycenter. This means that even if this model can generate a higher number of different graphs, the graphs it produces tend to be more similar one to the other than for other models. This is logical as this model preserves a degree

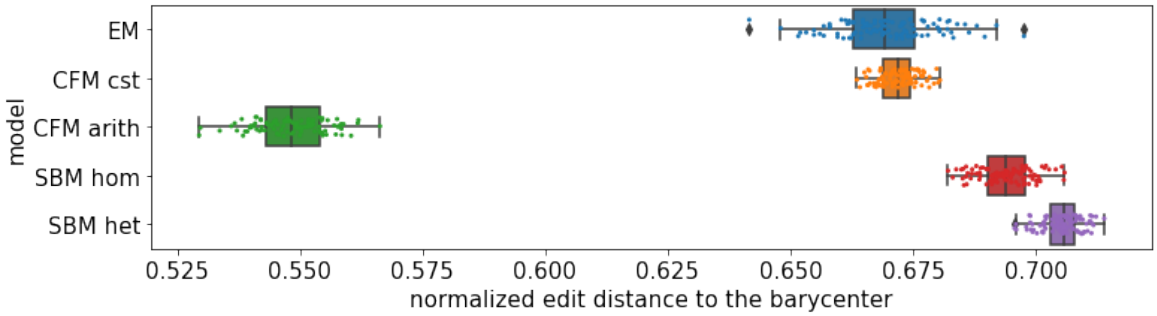


Figure 3.2 – **Edit distance to the barycenter for 6 different models**. For each model M in ordinate, we draw 100 graphs G at random from Ω_M and compute for each of them $\mathrm{ned}(G, G_M)$. The distribution of results is then plotted as a boxplot.

Model	Characteristic distance	Entropy
ER	0.67	2050
CFM cst	0.67	2500
CFM arith	0.55	3300
SBM hom	0.69	1840
SBM het	0.71	1840

Table 3.2 – **Edit distance to the barycenter and entropy.**

distribution, which enforces more constraints on edges' distribution than an Erdös-Renyi or a stochastic block model.

This concentration of graphs at a specific distance from the barycenter is a consequence of the dimensionality of the vector space. Let's denote $\mathcal{B}(G, r)$ the ball of center G and radius r in $(\mathbb{R}^{n^2}, \mathrm{ed})$. We consider the set

$$\begin{aligned}\Omega_M(r) &= \{G \in \Omega_M | \, \mathrm{ned}(G, G_M) \leqslant r\} \\ &= \{G \in \Omega_M | \, \mathrm{ed}(G, G_M) \leqslant 2mr\} \\ &= \Omega_M \cap \mathcal{B}(G_M, 2mr)\end{aligned}$$

The volume $V_n(r)$ of $\mathcal{B}(G_M, 2mr)$ is proportional to r^{n^2}, therefore

$$\forall r < 1, \frac{V_n(r)}{V_n(1)} \underset{n\to\infty}{\longrightarrow} 0 \tag{3.2}$$

The volume of the ball concentrates quickly at its periphery as the dimension increases, and so does the volume of Ω_M. The additional constraints on Ω_M modify its shape in such a way that graphs too far away from the barycenter are rare, which explains why the concentration does not happen at distance 1 from the barycenter. Still, this phenomenon is strong enough to imply that even graphs generated according to a model M will share only a relatively small fraction of their edges with the barycenter of the model.

Edit distance expected value

The previous observations on the structure of graph spaces show that in order to compare a graph G with a model M, one should consider more than the mere cardinal of Ω_M. One possibility to evaluate how similar to G are the graphs in Ω_M is to compute the expected value of the normalized edit distance:

$$\mathrm{EDEV}(G, M) = \mathop{\mathbb{E}}_{H \in \Omega_M} \left[\frac{1}{2m} \sum_{(i,j) \in V^2} |W_G(i,j) - W_H(i,j)| \right]$$

To illustrate how EDEV provides further information on the place of G within the graph space, we compare it with entropy for different synthetic graphs. A low value indicates that G is close to other graphs in Ω_M, and thus that it is typical of the model, while a high value shows that it is an outlier. As a case study, we consider the Erdös-Renyi model. Let's recall that we consider multigraphs, which implies that we allow for densities rising above 1. The extension of Erdös-Renyi model to multigraphs is straigthforward, $\Omega_{ER(n,m)}$ contains all multigraphs with n nodes and m edges and each multigraph is generated with the same probability $\frac{1}{|\Omega_{ER(n,m)}|}$. In practice, we consider models with $n = 100$ nodes and a number of edges m ranging from 100 to 500000. For each, we consider three graphs:

- $G_1(m)$, picked uniformly at random inside $\Omega_{\mathrm{ER}(n,m)}$
- $G_2(m)$, a graph made of two equal communities, each with $\frac{n}{2}$ nodes and $\frac{m}{2}$ edges, perfectly separated.
- $G_3(m)$, the graph where all edges are between nodes 0 and 1.

Results are shown on figure 3.3.

For each value of m, all three graphs belong to the graph ensemble $\Omega_{\mathrm{ER}(n,m)}$. We observe that as density increases $|\Omega_{ER(n,m)}|$ grows exponentially, which implies that the probability to pick at random $G_1(m)$, $G_2(m)$ or $G_3(m)$ becomes even less probable. Yet, in the case of the random graph $G_1(m)$ this is counter-intuitive: as density grows and becomes higher than 1, most graph in $\Omega_{\mathrm{ER}(n,m)}$ become complete graphs with each edge having weight about $\frac{m}{n^2}$. This is the case of G_1 too with a high probability, so $\mathrm{ER}(n,m)$ is very likely to produce graphs similar to $G_1(m)$, even if it is very unlikely to produce $G_1(m)$ itself.

On the other hand, edit distance expected value is able to capture this phenomenon. While it is close to 1 for all three types of graphs when density is low because in this situation a random model can hardly predict correctly which edge is present in any graph, it decreases quickly towards 0 when density rises above 0.1 for $G_1(m)$. For $G_2(m)$ we have an intermediate situation: edit distance decreases too, but it reaches its minimum around 0.5, indicating that even when it is densely populated, the model is only able to reproduce correctly half of its edges. This is normal as $G_2(m)$ concentrates them inside the communities, which means on half of all possible node pairs. These observations are actually a particular case of a more general result, which can be stated as:

Theorem 1. *Let B_1 and B_2 be two partition on $[\![1, n]\!]$, with p_1 and p_2 blocks respectively. Let $M_1 \in \mathcal{M}_{p_1}(\mathbb{N})$ and $M_2 \in \mathcal{M}_{p_2}(\mathbb{N})$ be two block adjacency matrices such that*

$$\sum_{i,j\in[1,p_1]^2} M_1[i,j] = \sum_{k,l\in[1,p_2]^2} M_2[k,l] = m$$

Let's consider two series of stochastic blockmodels defined as $S_1(k) = (B_1, k \cdot M_1)$ and $S_2(k) = (B_2, k \cdot M_2)$, whose barycenters are denoted $G_1(k)$ and $G_2(k)$. We have that

1. *There exists $d \in \mathbb{R}, \forall k \in \mathbb{N}, \mathrm{ed}(G_1(k), G_2(k)) = d$*
2. *Let $(G_k)_{k\in\mathbb{N}}$ be a series of random graphs, each drawn following model $S_1(k)$.*

$$\mathrm{EDEV}(G_k, S_2(k)) \xrightarrow[k\to\infty]{\mathbb{P}} d \tag{3.3}$$

a proof of this result can be found in appendix 7.1.2

Remark. *In particular, if $M_1 = M_2$, this theorem means that the normalized edit distance expected value between a graph picked at random and the barycenter of the stochastic blockmodel converges toward 0 as the density grows to infinity: $\Omega_{S(k)}$ shrinks around $G_{S(k)}$. This is what we observe with $G_1(m)$. Yet, we also observe on figure 3.3 that the normalized edit distance converges toward 0 only as density rises above 1. Thus, in practice, the vast majority of real world networks are too sparse for this assumption to hold true and most graphs in $\Omega_{S(k)}$ are far from $G_{S(k)}$, as developed in section 3.2.1*

Model likelihood

As the distance to the barycenter, the expected value of the normalized edit distance is characteristical of a model. For a model M, the values of $\mathrm{EDEV}(H, M)$ for graphs H in Ω_M are concentrated around a specific value d_M. We can use this fact to rule out models which fit badly on an observed graph G.

For example, let's consider the configuration model $CFMD(n, k_i^{out}, k_i^{in})$

$$n = 50$$
$$\forall i \in [\![0, n-1]\!], k_i^{out} = k_i^{in} = i$$

We use this model to generate a graph G_i. $CFMD(n, k_i^{out}, k_i^{in})$ will be called the *generative* model, and G_i the *observed* graph. On this observed graph, we test the stochastic blockmodel SBM_i obtained by partitioning its nodes in two blocks: B_0 contains even nodes and B_1 odd nodes (this way we avoid to put all high-degree nodes in the same block) and learning the block adjacency matrix on G_i. We call SBM_i the *candidate* model. We generate a sample $\mathcal{S}_i$ of 100 test graphs with the candidate model SBM_i and compare the normalized edit distance $\mathrm{EDEV}(G_i, \Omega_{SBM_i})$ of the observed graph to the candidate model with $\mathrm{EDEV}(H, \Omega_{SBM_i})$ for all test graphs $H \in \mathcal{S}_i$. This experiment is performed 5 times, and results are shown on figure 3.4.

We observe that for all five experiments, the normalized edit distance expected value to the candidate model SBM_i for the observed graph G_i is around 0.74, while for the test

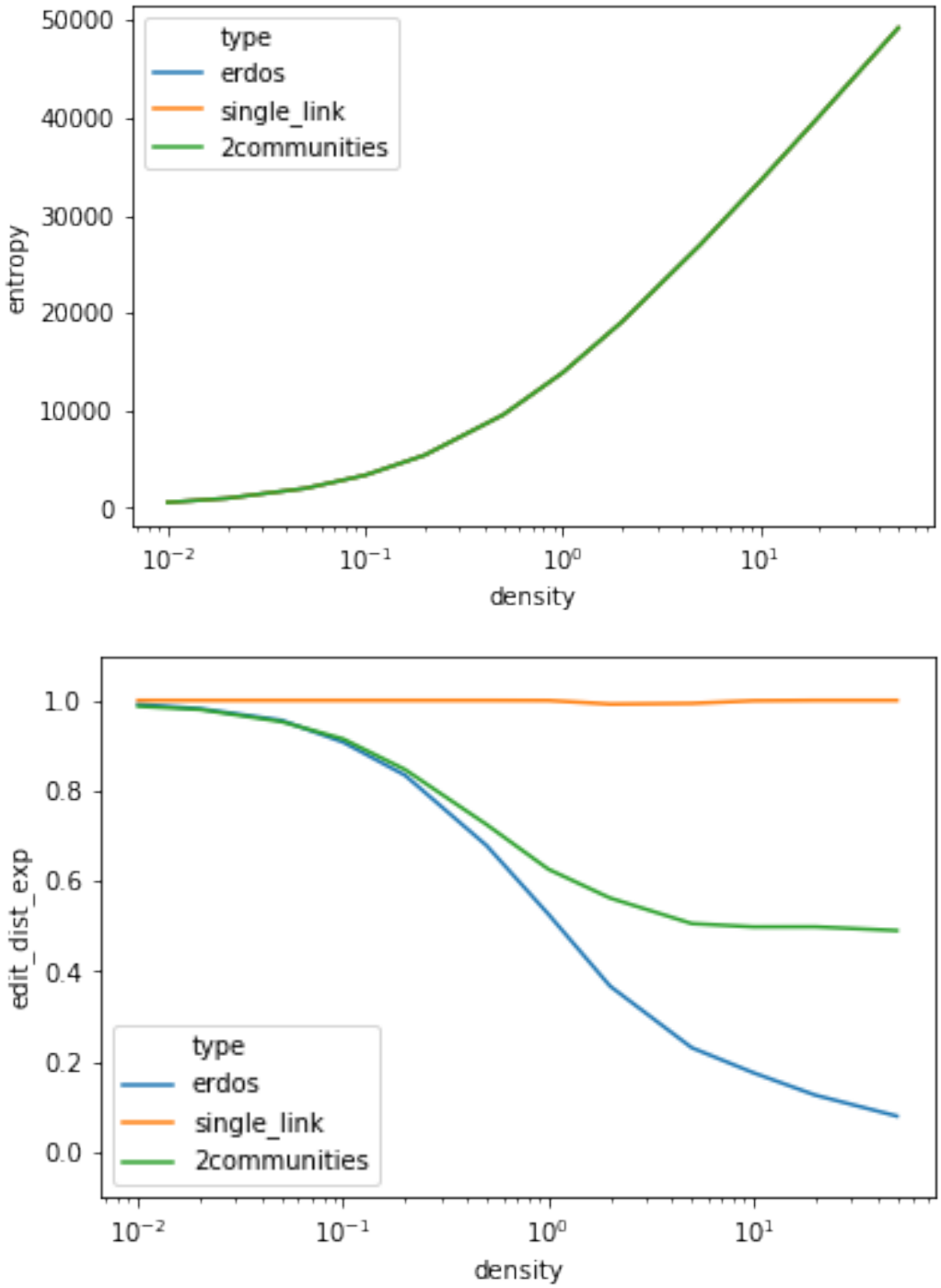


Figure 3.3 – **Entropy and edit distance expected value against density.** For each density, three graphs are generated. $G_1(m)$ is random, $G_2(m)$ is made of two random communities and $G_3(m)$ has its edges concentrated on a single pair of nodes of weight m. On the top plot, the entropy $\log(|\Omega_{ER(n,m)}|)$ is plotted against the density $\frac{m}{n^2}$. As all graphs belong to the same graph ensemble $\Omega_{ER(n,m)}$, the three curves are the same. On the bottom plot the edit distance expected value $EDEV(G(m), \Omega_{ER(n,m)})$ is plotted against density.

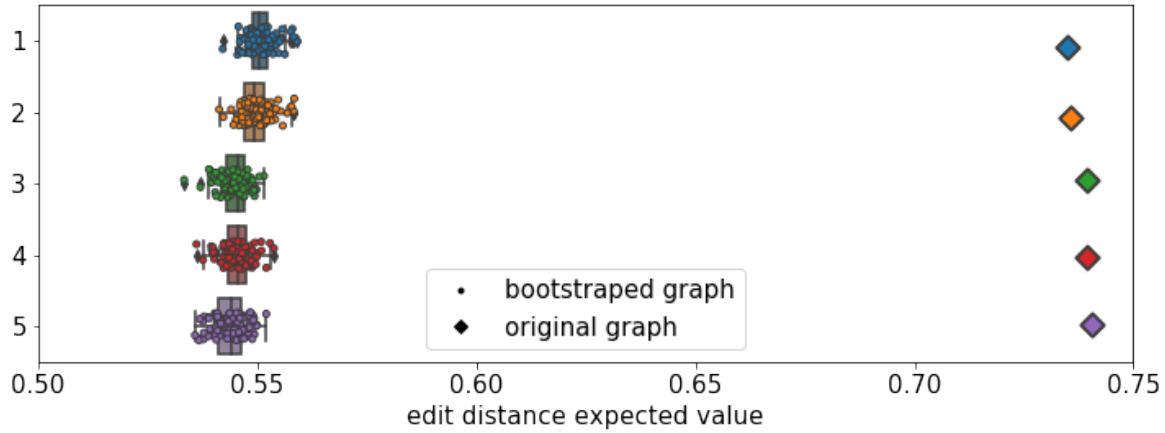


Figure 3.4 – 5 graphs G_i were generated with a generative model $CFMD(n, k_i^{out}, k_i^{in})$. Their normalized edit distance expected value with respect to a candidate model SBM_i is plotted as diamond. As a comparison point, the distribution of the normalized edit distance expected value for 100 test graphs generated with SBM_i is plotted as dots and boxplot.

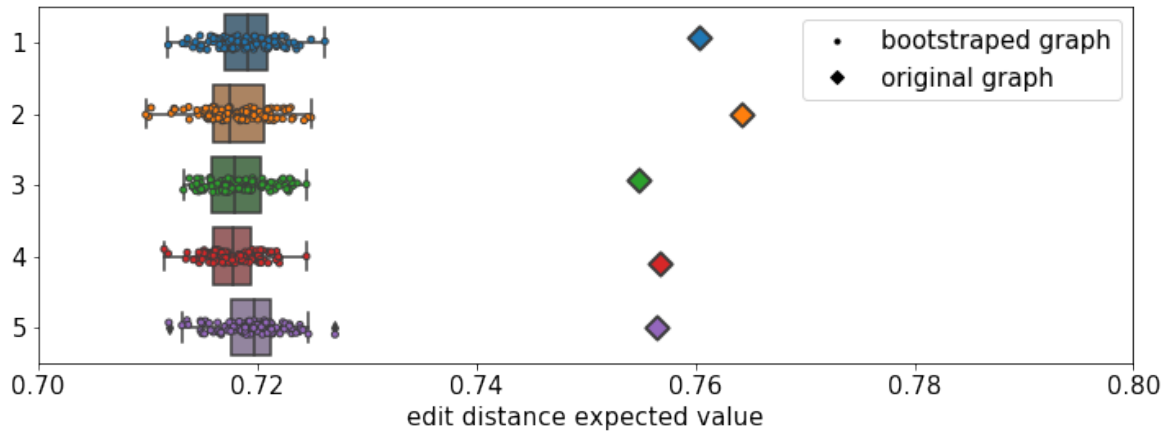


Figure 3.5 – 5 graphs G_i were generated with a generative model $SBM(n, B, M)$. Their normalized edit distance expected value with respect to a test model $CFMD_i$ is plotted as diamond. As a comparison point, the distribution of the normalized edit distance expected value for 100 test graphs generated with $CFMD_i$ is plotted as dots and boxplot.

graphs generated by SBM_i it is between 0.53 and 0.56. This shows that the normalized edit distance expected value to SBM_i is significantly different for the observed graphs, which were generated by the configuration model $CFMD(n, k_i^{out}, k_i^{in})$, and for the test graphs generated by the stochastic blockmodel SBM_i. It is thus very unlikely that the observed graph G_i was generated by the candidate model SBM_i.

We then perform the same experiment the other way round, by considering as generative model a stochastic blockmodel $SBM_0(n, B, M)$ defined by

$$n = 50$$
$$B = [\![0, 24]\!], [\![25, 49]\!]$$
$$M = \begin{bmatrix} 500 & 0 \\ 0 & 500 \end{bmatrix}$$

5 graphs G_i are generated using this stochastic blockmodel. As candidate model, we consider a configuration model $CFMD_i$ obtained by learning the degree sequence of G_i. A sample $\mathcal{S}'$ of 100 test graphs is randomly picked in Ω_{CMFD_i} and we compare $\text{EDEV}(G_i, \Omega_{CFMD_i})$ with $\text{EDEV}(H, \Omega_{CFMD_i})$ for all $H \in \mathcal{S}'$. Results are shown on figure 3.5. Once again, we observe that the normalized edit distance expected value to the candidate model $CFMD_i$ is significantly different for the observed graphs, which were generated by $SBM(n, B, M)$, and for the test graphs generated by $CFMD_i$. This allows us to reject the hypothesis that the observed graph G_i was generated by the candidate model $CFMD_i$.

Statistical hypothesis testing

This methodology can be formalized using statistical hypothesis testing. Let's say we have a graph G and a model M (possibly obtained by fitting some parameters on G). We want to test the hypothesis $\mathcal{H}$: "the observed graph G has been generated by the candidate model M". We do so in the following way:

1. Choose a confidence level δ, for example 0.01.
2. Generate a random sample from the candidate model $\mathcal{S} \subset \Omega_M$, of n_b test graphs.
3. Infer the probability distribution $\mathbb{P}_{ed}$ of $\text{EDEV}(H, \Omega_M)$ for $H \in \Omega_M$ from the sample $\mathcal{S}$.
4. Compute the probability
$$p = \mathbb{P}_{ed}[|\text{EDEV}(H, \Omega_M) - d_M| \geqslant |\text{EDEV}(G, \Omega_M) - d_M|]$$
5. If $p < \delta$, it means that the probability that G was generated by M is inferior to δ and the hypothesis $\mathcal{H}$ can be rejected.

In practice, to infer the probability distribution $\mathbb{P}_{ed}$, we assume that this distribution is normal, so we only need to infer the mean d_M and the standard deviation σ_M of the distribution. This assumption is verified on the models described in the previous subsection, by comparing the cumulative distribution function of the sample with the one of the corresponding normal distribution. An example is shown in figure 3.6. The shapiro-wilk test returned a p-value below 0.05 (around 0.01) for 2 of the 10 models. This means that the

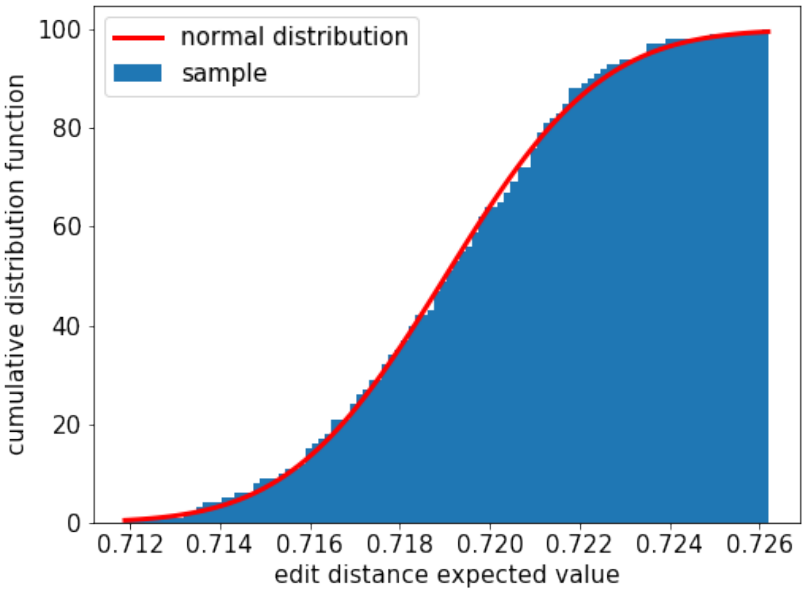


Figure 3.6 – Comparison between the empirical distribution of the edit distance expected value for graphs generated with a configuration model ($CFMD_0$) and the normal distribution with the same mean and standard deviation.

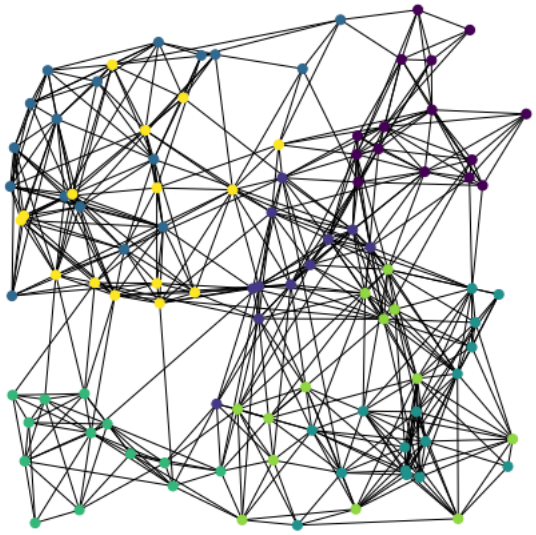


Figure 3.7 – Spatial graph generated using the waxman random geometric model. 100 nodes are randomly distributed in a $[0,1] \times [0,1]$ square. They are then connected with a probability depending on the distance d between nodes: $p(d) = 10 \exp\left(\frac{d}{0.05L}\right)$, with L the maximum distance between two nodes. Communities are computed using graph tools[1].

normal distribution can only be considered a rough estimate in general, but we found it good enough as a first approximation.

Let's consider a situation in which one wishes to evaluate the relevance of the block structure computed on a graph G, with a weight matrix W_G. Whatever the graph, and the partition $B = (b_1, \dots, b_p)$ of its n nodes, it is always possible to define $M \in \mathcal{M}_p(\mathbb{N})$ as

$$\forall i, j \in [\![1, p]\!], M[i, j] = \sum_{u \in b_i, v \in b_j} W_G[u, v]$$

such that $G \in \Omega_{SBM(n,B,M)}$. One may then wonder whether this stochastic blockmodel is a relevant model of G. An even trickier question is to evaluate whether any stochastic blockmodel can be a relevant model. In particular, spatial models can generate graphs with groups of nodes densely connected due to their position rather than to block membership. An example of such a graph is shown in figure 3.7. It may then be hard to tell whether the blocks found are indeed a legitimate model of the observed graph or should be considered as artefacts, consequences of the underlying spatial structure.

To illustrate how statistical hypothesis testing allows to adress this issue, we consider eight models: four stochastic blockmodels and four waxman model for random geometric graphs[2] with different sets of parameters. The waxman model for spatial graphs allows to easily control the strength of the spatial structure, by tuning the speed at which edge probability decays as the distance between nodes rises. The number of nodes is fixed to $n = 100$ and the parameters are fixed such as to ensure a density d around 0.036. All stochastic blockmodels use a node partition in four blocks of 25 nodes, with a block adjacency matrix of the form:

$$\begin{bmatrix} m_{int} & m_{ext} & m_{ext} & m_{ext} \\ m_{ext} & m_{int} & m_{ext} & m_{ext} \\ m_{ext} & m_{ext} & m_{int} & m_{ext} \\ m_{ext} & m_{ext} & m_{ext} & m_{int} \end{bmatrix}$$

The four stochastic blockmodels are then defined by:

1. M_0: $m_{int} = 90$, $m_{ext} = 0$
2. M_1: $m_{int} = 75$, $m_{ext} = 5$
3. M_2: $m_{int} = 60$, $m_{ext} = 10$
4. M_3: $m_{int} = 45$, $m_{ext} = 15$

This way, the graphs generated using SBM_0 are made of perfectly separated blocks of nodes, while those generated by SBM_3 have blocks with as many internal and external edges.

For the waxman models, we also consider four parameter sets:

1. M_4: $\alpha = 0.1$, $\beta = 1$
2. M_5: $\alpha = 0.08$, $\beta = 1.6$

[1] `https://graph-tool.skewed.de/`

[2] `https://networkx.org/documentation/stable/reference/generated/networkx.generators.geometric.waxman_graph.html`

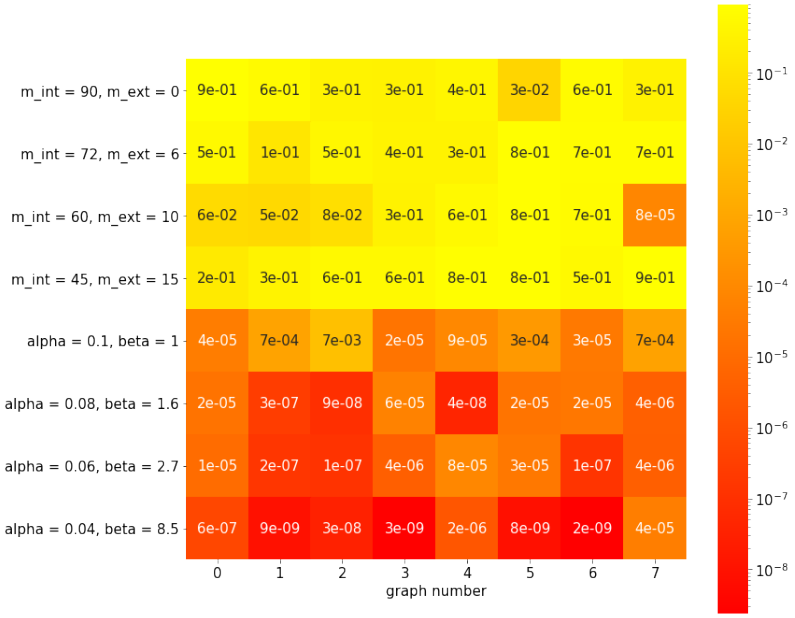


Figure 3.8 – For each model M_i in ordinate, and each graph number i in absciss, we compute the probability $p_{i,j}$ that a graph generated using $SBM_{i,j}$ has a normalized edit distance expected value to $SBM_{i,j}$ further away from the mean value $d_{i,j}$ than $G_{i,j}$. The probability is plotted in heatmap.

3. M_6: $\alpha = 0.06, \beta = 2.7$

4. M_7: $\alpha = 0.04, \beta = 8.5$

The lower the value of α, the stronger the spatial structure.

With each model M_i, we generate 8 graphs $(G_{i,j})_{j\in[\![0,7]\!]}$. On each of those observed graph, we find the minimum entropy node partition $B_{i,j}$ using graph tools, and fit a candidate stochastic blockmodel $SBM_{i,j}$ on $G_{i,j}$ based on this node partition. We then evaluate the relevance of this stochastic blockmodel using the previously described methodology. We use a confidence level δ of 0.01 and a sample size $n_b = 200$. The probabilities $p_{i,j}$ obtained are plotted in figure 3.8.

We observe that for all graphs generated by stochastic block models but one, $p_{i,j} > 10^{-2} = \delta$. On the other hand, for all spatial graphs $p_{i,j} < \delta$. This means that the hypothesis "$G_{i,j}$ has been generated by the candidate stochastic block model $SBM_{i,j}$" is rejected for all 32 spatial graphs, and for the 8^{th} graph generated by SBM_2. Let's stress again that a probability $p_{i,j}$ superior to δ does not mean that $SBM_{i,j}$ is the right model for $G_{i,j}$. It only means that there is not enough statistical evidence to reject it.

These results show that, on 32 spatial graphs generated with various sets of parameters, the statistical hypothesis testing methodology is able to correctly identify that the block structure found is not a relevant model, despite being the solution maximizing the likelihood. This result is not trivial, as illustrated by the graph depicted on figure 3.7. What is more, this methodology manages to reject the block structure for all spatial graphs while spuriously

rejecting it for only one out of 32 graphs originally generated with a stochastic blockmodel. In other words, there is no false negative, and only one false positive.

Strictly speaking, these results only allow to rule out one node partition. As the block structure tested were fitted on the observed graphs using minimum entropy, one could argue that ruling out this partition implies that no other node partition can lead to a relevant model. However, for most real graphs, there is more than one plausible node partition and minimizing the entropy of a partition is a stochastic process. Therefore, the experiment should be performed more than once to conclude that the observed graph has no block structure.

Conclusion

Widely used quality measures for graph models rely either on the number of different graphs they can produce, which neglects the geometric structure of the graph space, or on a direct comparison with the barycenter of those graphs, which discards information about the distribution around this barycenter. Because of these restrictions, they are unable to distinguish between graphs which have a typical structure of a model and graphs which may be generated by this model but as outliers.

This chapter shows how graph spaces can provide additional information on the structure of graphs generated by a model, which is captured neither by the entropy nor the barycenter. By computing the expected value of the normalized edit distance for a given graph, we obtain a criterion which can be used to evaluate the model quality with respect to this graph. Finally, we incorporate this criterion to a statistical hypothesis testing methodology to perform model selection.

This theoretical framework can be used for any statistical model, and particularly spatial models. It allows to compare them with SBM or configuration model, and perform model selection between models of different nature. What is more, statistical hypothesis testing provides a rigorous methodology to evaluate the relevance of a candidate model to an observed graph.

Any graph property could be used for such a test: average path length, clustering coefficient, etc. The main requirement for such a property to be used in statistical testing is that the distribution of its possible values for a given model is concentrated around its mean. If it is not, no value measured on the observed graph will allow to reject the candidate model.

Apart from its simplicity to compute and interpret, one important advantage of the normalized edit distance expected value is that for most models its values concentrate quickly around the mean. It should be underlined that this fundamental property is a consequence of a geometrical result (the volume of a ball in n dimensions), which highlights the benefits of considering the geometric structure of graph ensembles. However, the edit distance is not the only distance that can be used. Considering other metrics which are more sensitive to the global topology of the network, like the perturbation-resistance distance or spectral distances, could provide additional insight on the structure of the graph space.

Chapter 4

The limits of entropy

Statistical tests, as described in the previous section, are a powerful and well-established method to confront the predictions of a probabilistic model with observations. Yet, in the case of automatic model selection, it suffers from an import drawback: by design, it can only reject a candidate model if the probability it assigns to the observation is too low. When comparing different models, it may reject some of them and not others, but it does not provide an automatic way to rank them in order to select the best one. On the other hand, the minimum description length principle does provide such a ranking, as illustrated by the stochastic blockmodel inference methodology presented in section 2.3 of the state of the art.

This method relies on the hypothesis that there exists an original partition of the nodes, and that the graph under study was generated by picking edges at random with a probability that depends only on the communities to which its extremities belong. The idea is to find the most likely original partition for a SBM with respect to a graph by maximizing simultaneously the probability to choose this partition and the probability to generate this graph, given the partition. To perform the second maximization, it assumes that all graphs are generated with the same probability and it thus searches a partition of minimal entropy, in the sense that the cardinal of its microcanonical ensemble (*i.e.* the number of graphs the corresponding SBM can theoretically generate Cimini et al. [2018]) is minimal, which is equivalent to maximizing its likelihoodPeixoto [2012].

Contrarily to other ad hoc community detection methods, such as modularity maximization, the minimization of the microcanonical ensemble entropy relies on a statistical reasoning to select the node partition. It thus claims to select the most likely node partition with respect to the evidence present in the observed graph. However, communities were first defined empiricaly as groups of nodes defined by a characteristic connection density. In particular, it is commonly accepted that complete graphs with no edges between them should correspond to different communities.

In this chapter, we show that in practice the minimum entropy partition does not always correspond to this intuitive definition of communities. Even when the number and the size of the communities are prescribed beforehand, the node partition which corresponds to the sharper communities is not always the one with the lower entropy, even asymptoticaly. We demonstrate that when community sizes and edge distribution are heterogeneous enough, a node partition which places small communities where there are the most edges will always have a lower entropy. Finally, we illustrate how this issue implies that such heterogeneous

stochastic block models cannot be identified correctly by this model selection method. As the minimization of the entropy of the microcanonical ensemble is equivalent to the maximization of the likelihood of the associated node partition, these results question the underlying statistical reasoning. We thus conclude by discussing the relevance of assuming an equal probability for all graphs in this context.

Entropy based stochastic block model selection

The stochastic block model is a generative model for random graphs. It takes as parameters a set of nodes $V = [1; n]$ partitioned in p blocks (or communities) $C = (c_i)_{i\in[1;p]}$ and a block-to-block adjacency matrix M whose entries correspond to the number of edges between two blocks. The corresponding microcanonical ensemble is defined as

$$\Omega_{C,M} = \left\{ G \mid \forall c_1, c_2 \in C, \sum_{i\in c_i, j\in c_j} W_{(i,j)} = M_{(c_1,c_2)} \right\}$$

Where W is the weight matrix of graph G. Generating a graph with the stochastic block model associated to C, M amounts to drawing at random $G \in \Omega_{C,M}$. The probability distribution $\mathbb{P}[G|C, M]$ on this ensemble is defined as the uniform one:

$$\mathbb{P}[G|C, M] = \frac{1}{|\Omega_{C,M}|}$$

whose entropy equals $S = \ln(|\Omega_{C,M}|)$. It has been computed for different SBM flavours in Peixoto [2012]. It measures the number of different graphs a SBM can generate with a given set of parameters. The lower it is, the higher the probability to generate any specific graph G.

On the other hand, given a graph $G = (V, E)$, with a weight matrix W, it may have been generated by many different stochastic block models. For any partition $C = (c_i)_{i\in[1;p]}$ of V, there exists one and only one matrix M such that $G \in \Omega_{C,M}$, and it is defined as:

$$\forall c_1, c_2 \in C, M_{(c_1,c_2)} = \sum_{i\in c_1, j\in c_2} W_{(i,j)}$$

Therefore, when there is no ambiguity about the graph G, we will consider indifferently a partition and the associated SBM in the following.

The objective of stochastic block model inference is to find the partition C that best describes G. To do so, bayesian inference relies on the Bayes theorem which stands that:

$$\mathbb{P}[C, M|G] = \frac{\mathbb{P}[G|C, M] \times \mathbb{P}[C, M]}{\mathbb{P}[G]} \tag{4.1}$$

As $\mathbb{P}[G]$ is the same whatever C, it is sufficient to maximize $\mathbb{P}[G|C, M] \times \mathbb{P}[C, M]$. The naive approach which consists in using a maximum-entropy uniform prior distribution for $\mathbb{P}[C, M]$ simplifies the computation to maximizing directly $\mathbb{P}[G|C]$ (the so called likelihood function) but it will always lead to the trivial partition $\forall i \in V, c_i = \{i\}$, which is of no use

because the corresponding SBM reproduces G exactly: $M = W$ and $\mathbb{P}[G|C] = 1$. To overcome this overfitting problem, another prior distribution was proposed in Peixoto [2019], which assigns lower probabilities to the partitions with many communities. Yet, when comparing two models C_1, M_1 and C_2, M_2 with equal prior probability, the one which is chosen is still the one minimizing $|\Omega_{C,M}|$ or equivalently the entropy $S = \ln(|\Omega_{C,M}|)$, as logarithm is a monotonous function.

The issue with heavily populated graph regions

In this chapter, we focus on the consequence of minimizing the entropy to discriminate between node partitions. To do so, we need to work on a domain of partitions on which the prior distribution is uniform. As suggested by Peixoto [2019], we restrict ourselves to finding the best partition when the number p and the sizes $(s_i)_{i\in[1;p]}$ of communities are fixed because in this case, both $P[C]$ and $P[M|C]$ are constant. This is a problem of node classification, and in this situation the maximization of equation 7.12 boils down to minimizing the entropy of $\Omega_{C,M}$, which can be written as:

$$S = \sum_{i,j\in[1;p]} \ln\left[\binom{s_i s_j + M_{(i,j)} - 1}{M_{(i,j)}}\right]$$

as shown in Peixoto [2012].

Yet, even within this restricted domain (p and $(s_i)_i$ are fixed), the lower entropy partition for a given graph G is not always the one which corresponds to the sharper communities. To illustrate this phenomena, let's consider the stochastic block models whose matrices M are shown on figure 4.1, and a multigraph $G \in \Omega_{SBM_1} \cap \Omega_{SBM_2}$.

- SBM_1 corresponds to $C_1 = \{c_1^a : \{0,1,2,3,4,5\}, c_1^b : \{6,7,8\}, c_1^c : \{9,10,11\}\}$
- SBM_2 corresponds to $C_2 = \{c_2^a : \{0,1,2\}, c_2^b : \{3,4,5\}, c_2^c : \{6,7,8,9,10,11\}\}$.

As $G \in \Omega_{SBM_1} \cap \Omega_{SBM_2}$, it could have been generated using SBM_1 or SBM_2. Yet, the point of inferring a stochastic block model to understand the structure of a graph is that it is supposed to identify groups of nodes (blocks) such that the edge distribution between any two of them is homogeneous and characterized by a specific density. From this point of view C_1 seems a better partition than C_2:

- The density of edges inside and between c_2^a and c_2^b is the same (10), so there is no justification for dividing c_1^a in two.
- On the other hand, c_1^b and c_1^c have an internal density of 1 and there is no edge between them, so it is logical to separate them rather than merge them into c_2^c.

Yet, if we compute the entropy of SBM_1 and SBM_2:

$$S_1 = \ln\left[\binom{395}{360}\right] + 2 \times \ln\left[\binom{17}{9}\right] = 136$$

$$S_2 = \ln\left[\binom{53}{18}\right] + 4 \times \ln\left[\binom{98}{90}\right] = 135$$

The entropy of SBM_2 is lower and thus partition C_2 will be the one selected. Of course, as $|\Omega_{SBM_2}| < |\Omega_{SBM_1}|$, the probability to generate G with SBM_2 is higher than the probability to generate it with SBM_1. But this increased probability is not due to a better identification of the edge distribution heterogeneity, it is a mechanical effect of imposing smaller communities in the groups of nodes which contain the more edges, even if their distribution is homogeneous. Doing so reduces the number of possible positions for each edge and thus the number of different graphs the model can generate.

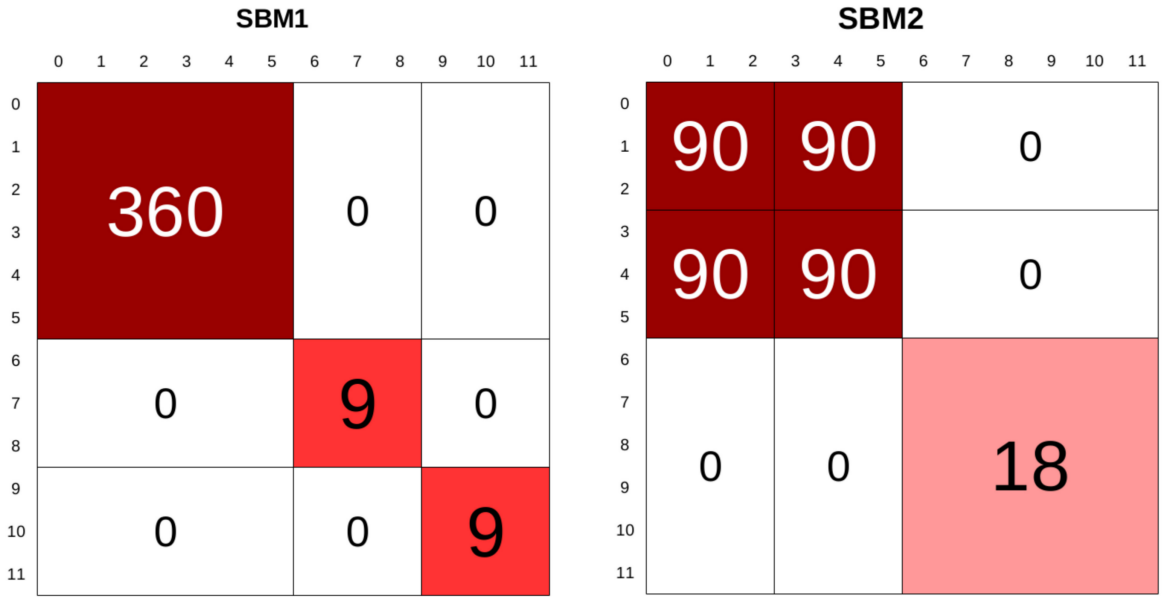


Figure 4.1 – **Block-to-block adjacency matrices of two overlapping stochastic block models.** Even though the communities of SBM_1 are better defined, SBM_2 can generate less different graphs and thus generates them with higher probability.

This problem can also occur with smaller densities, as illustrated by the stochastic block models whose block-to-block adjacency matrices are shown on figure 4.2. SBM_3, defined as one community of 128 nodes and density 0.6 and 32 communities of 4 nodes and density 0.4 has an entropy of 17851. SBM_4 which merges all small communities into one big and splits the big one into 32 small ones has an entropy of 16403.

The density threshold

More generally, let's consider a SBM (C_1, M_1) with one big community of size s, containing $c \times m_0$ edges and q small communities of size $\frac{s}{q}$ containing $(m_i)_{i\in[1;q]}$ edges each, as illustrated on figure 4.3. Its entropy is equal to:

$$S_1(c) = \ln\left[\binom{s^2 + c \times m_0 - 1}{c \times m_0}\right] + \sum_{i=1}^{q} \ln\left[\binom{\frac{s^2}{q^2} + m_i - 1}{m_i}\right]$$

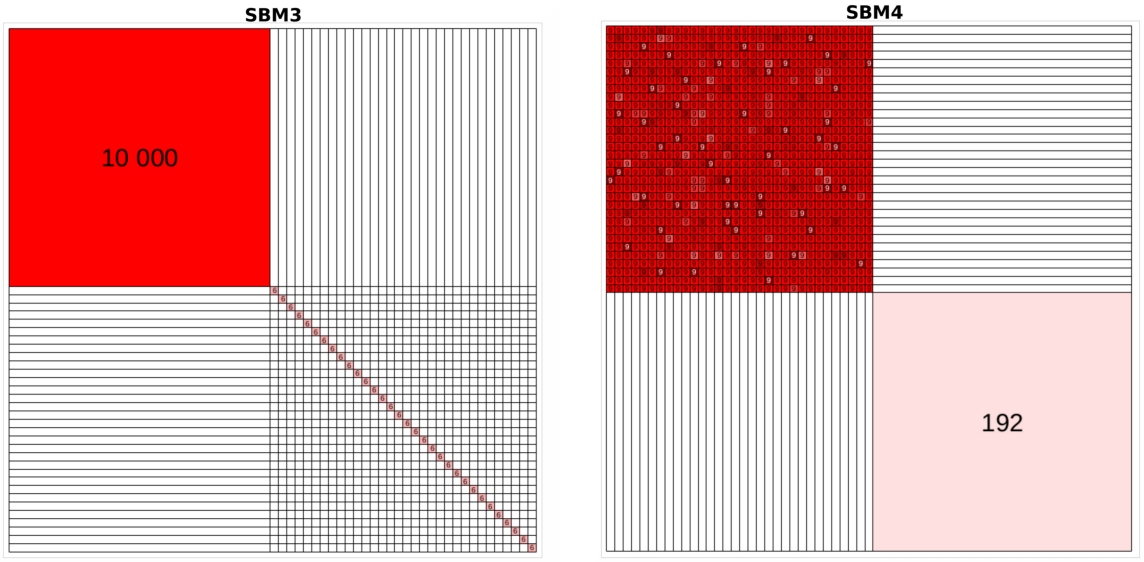


Figure 4.2 – **Block-to-block adjacency matrices of two overlapping stochastic block models with lower densities.** Once again, even though SBM_3 has better defined communities, SBM_4 is more likely a model for graphs $G \in \Omega_{SBM_3} \cap \Omega_{SBM_4}$

On the other hand, the entropy of the SBM (C_2, M_2) which splits the big community into q small ones of size $\frac{s}{q}$ and merges the q small communities into one big is:

$$S_2(c) = \ln\left[\binom{s^2 + \sum_{i=1}^{q} m_i - 1}{\sum_{i=1}^{q} m_i}\right] + q^2 \ln\left[\binom{\frac{s^2 + c \times m_0}{q^2} - 1}{\frac{c \times m_0}{q^2}}\right]$$

So, with $C_1 = \sum_{i=1}^{q} \ln\left[\binom{\frac{s^2}{q^2} + m_i - 1}{m_i}\right]$ and $C_2 = \ln\left[\binom{s^2 + \sum_{i=1}^{q} m_i - 1}{\sum_{i=1}^{q} m_i}\right]$, which are constants with

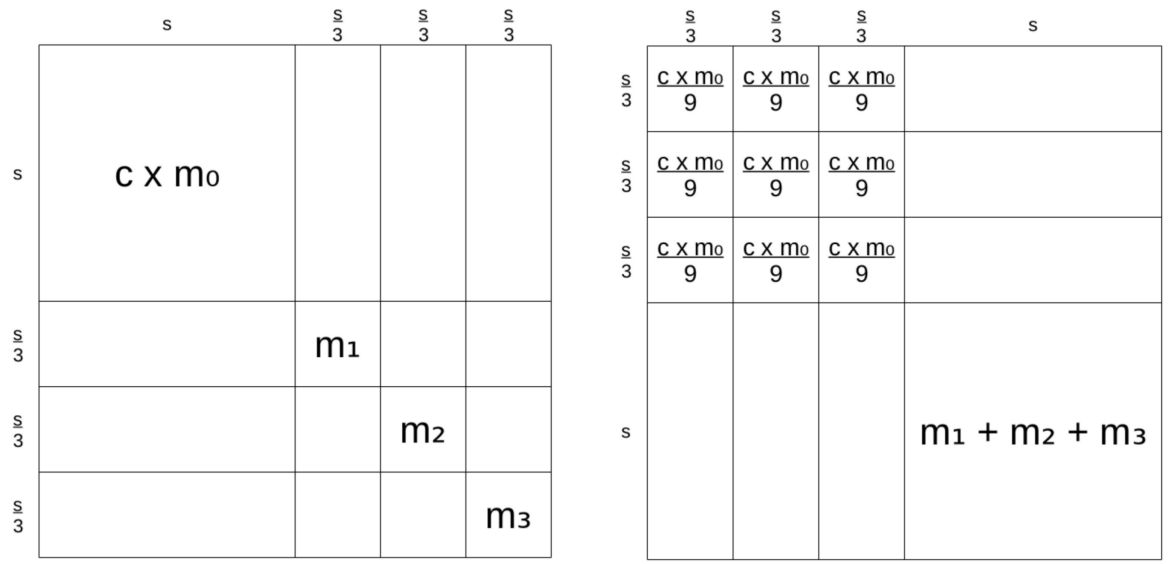


Figure 4.3 – **Theoretical pair of stochastic block models.** The right-side partition splits the big community in $q = 3$ small ones and merges the small communities in one big.

respect to c:

$$
\begin{aligned}
S_1(c) - S_2(c) &= \ln\left[\binom{s^2 + c \times m_0 - 1}{c \times m_0}\right] - q^2 \ln\left[\binom{\frac{s^2 + c\times m_0}{q^2} - 1}{\frac{c\times m_0}{q^2}}\right] + C_1 - C_2 \\
&= \ln\left[\prod_{k=1}^{c\times m_0} \frac{k + s^2 - 1}{k}\right] - \ln\left[\left(\prod_{k=1}^{\frac{c\times m_0}{q^2}} \frac{k + \frac{s^2}{q^2} - 1}{k}\right)^{q^2}\right] + C_1 - C_2 \\
&= \ln\left[\prod_{k=1}^{\frac{c\times m_0}{q^2}} \frac{\prod_{i=0}^{q^2-1}(k + s^2 - 1 + i \times \frac{c\times m_0}{q^2})}{(k + \frac{s^2}{q^2} - 1)^{q^2}}\right] + C_1 - C_2 \\
&> \ln\left[\prod_{k=1}^{\frac{c\times m_0}{q^2}} \left(\frac{k + s^2 - 1}{k + \frac{s^2}{q^2} - 1}\right)^{q^2}\right] + C_1 - C_2 \\
&> q^2 \sum_{k=1}^{\frac{c\times m_0}{q^2}} \ln\left[1 + \frac{(q^2-1)s^2}{q^2 k + s^2 - q^2}\right] + C_1 - C_2 \qquad (4.2)
\end{aligned}
$$

Now, as

$$
\ln\left[1 + \frac{(q^2-1)s^2}{q^2 k + s^2 - q^2}\right] \underset{k\to\infty}{\sim} \frac{(q^2-1)s^2}{q^2 k + s^2 - q^2}
$$

and

$$\sum_{k=1}^{\frac{c\times m_0}{q^2}} \frac{(q^2-1)s^2}{q^2k+s^2-q^2} \underset{c\to\infty}{\to} \infty$$

we have that

$$q^2 \sum_{k=1}^{\frac{c\times m_0}{q^2}} \ln\left[1+\frac{(q^2-1)s^2}{q^2k+s^2-q^2}\right] \underset{c\to\infty}{\to} \infty \tag{4.3}$$

and thus, by injecting equation 4.3 inside 4.2, $\exists c, \forall c' > c, S_2(c') < S_1(c')$. Which means that for any such pair of stochastic block models, there exists some density threshold for the big community in C_1 above which (C_2, M_2) will be identified as the most likely model for all graphs $G \in \Omega_{(C_1,M_1)} \cap \Omega_{(C_2,M_2)}$.

Consequences on model selection

In practice, this phenomena implies that a model selection technique based on the minimization of entropy will not be able to identify correctly some SBM when they are used as generative models for synthetic graphs. To illustrate this, we generate graphs and try to recover the original partition. The experiment is conducted on two series of stochastic block models, one with relatively large communities and another one with smaller but more sharply defined communities:

- $SBM_7(d)$ is made of 5 blocks (1 of 40 nodes, and 4 of 10 nodes). Its density matrix D is given on figure 4.4 (left) (one can deduce the block adjacency matrix by $M_{(c_i,c_j)} = |c_i||c_j| \times D_{(c_i,c_j)}$).
- $SBM_8(d)$ is made of 11 blocks (1 of 100 nodes, and 10 of 10 nodes). The internal density of the big community is d, it is 0.15 for the small ones and 0.01 between communities.

For each of those two models, and for various internal densities d of the largest community, we generate 1000 random graphs. For each of these graphs, we compute the entropy of the original partition (correct partition) and the entropy of the partition obtained by inverting the big community with the small ones (incorrect partition). Then, we compute the percentage of graphs for which the correct partition has a lower entropy than the incorrect one and plot it against the density d. Results are shown on figure 4.4 and 4.5.

We observe that as soon as d reaches a given density threshold (about 0.08 for $SBM_7(d)$ and 0.18 for $SBM_8(d)$), the percentage of correct match drops quickly to 0. As d rises over 0.25, the correct partition is never the one selected. It should be highlighted that in these experiments we only compared two partitions among the B_n possible, so the percentage of correct match is actually an upper bound on the percentage of graphs for which the correct partition is identified. This means that if $SBM_7(d)$ or $SBM_8(d)$ are used as generative models for random graphs, with $d > 0.25$, and one wants to use bayesian inference for determining the original partition, it will almost never return the correct one. What is more, the results of section 4.3 show that this will occur for any SBM of the form described in figure 4.3, as soon as the big community contains enough edges.

d	0.05	0.05	0.05	0.05
0.05	**0.2**	0.1	0.1	0.1
0.05	0.1	**0.2**	0.1	0.1
0.05	0.1	0.1	**0.2**	0.1
0.05	0.1	0.1	0.1	**0.2**

SBM7

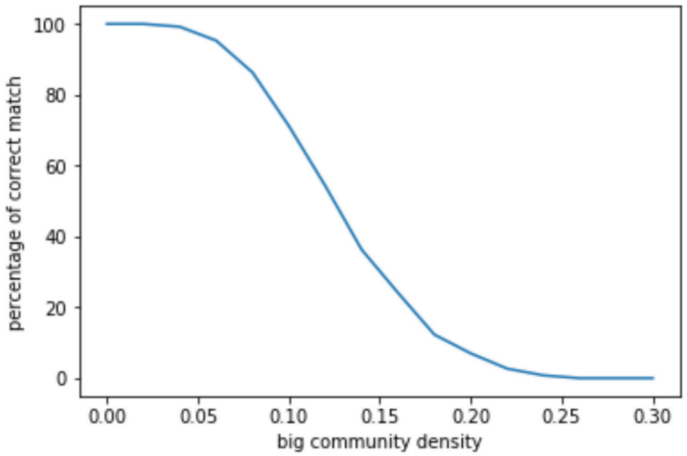


Figure 4.4 – Block-to-block adjacency matrix of $SBM_7(d)$ (left) and percentage of graphs generated using $SBM_7(d)$ for which the original partition has a lower entropy than the inverted one against the density d of the big community (right).

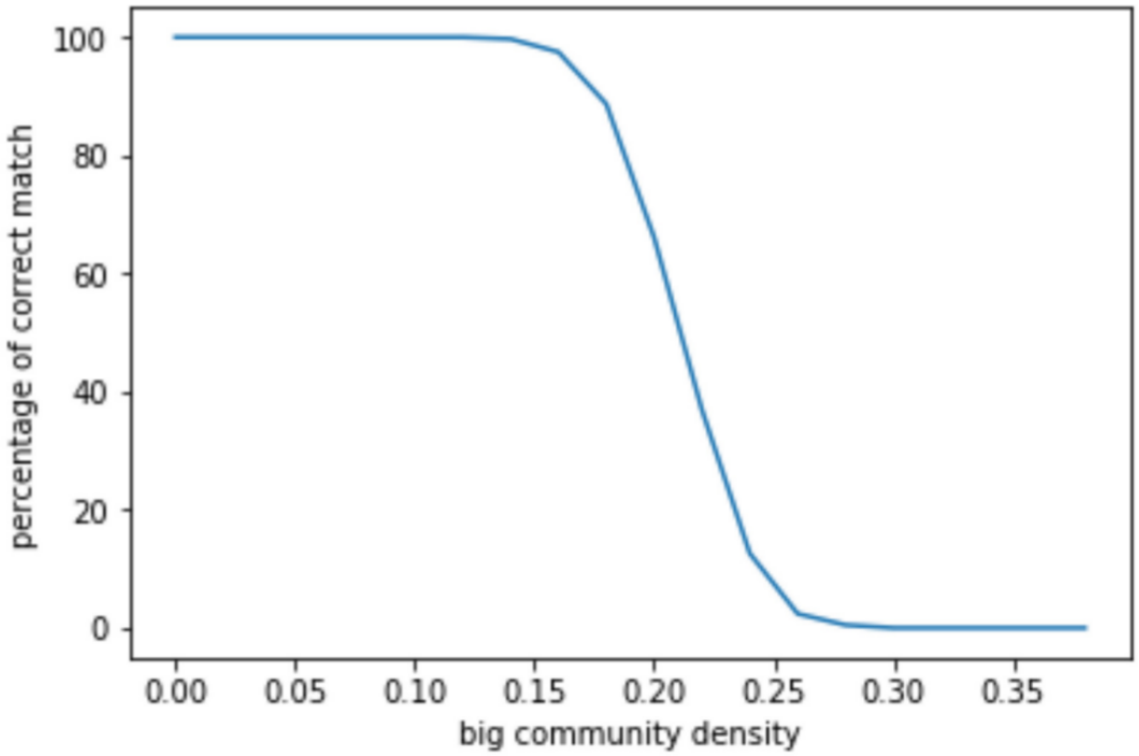


Figure 4.5 – Percentage of graphs generated using $SBM_8(d)$ for which the original partition has a lower entropy than the inverted one against the density d of the big community.

Discussion

We have seen in section 4.1 that model selection techniques that rely on the maximization of the likelihood function to find the best node partition given an observed graph boils down to the minimization of the entropy of the corresponding ensemble of generable graphs in the microcanonical framework. Even in the case of bayesian inference, when a non-uniform prior distribution is defined on the set of possible partitions, entropy remains the criterion of choice between equiprobable partitions. Yet, as shown in section 4.2 and 4.3, entropy behaves counter intuitively when a large part of the edges are concentrated inside one big community. In this situation, a partition that splits this community in small ones will have a lower entropy, even though the edge density is homogeneous. Furthermore, this happens even when the number and sizes of communities are known. Practically, as explained in section 4.4, this phenomena implies that stochastic block models of this form cannot be recovered using model selection techniques based on the mere minimization of the cardinal of the associated microcanonical ensemble.

Let's stress that contrarily to the resolution limit described in Fortunato and Barthelemy [2007] or Peixoto [2013], the problem is not about being able or not to detect small communities with no prior knowledge about the graph, it occurs even though the number and sizes of communities are known. It is also different from the phase transition issue that has been investigated in Abbe and Sandon [2015]; Decelle et al. [2011a,b]; Hu et al. [2012] for communities detection or recovery because it happens even when communities are dense and perfectly separated. Entropy minimization fails at classifying correctly the nodes between communities because it only aims at identifying the SBM that can generate the lowest number of different graphs. Splitting dense groups of nodes into small blocks enforces more constraints on edge positions and thus mechanically reduces the size of the microcanonical ensemble. This is a form of overfitting, in the sense that the higher probability to generate the observed graph is not due to a better identification of the heterogeneity in the observed edge distribution, but is an artifact due to the model selection technique.

The results presented in this chapter were obtained for a particular class of stochastic block models. First of all, they were obtained for the multigraph flavour of stochastic block models. As the node classification issue occurs also for densities below 1, they can probably be extended to simple graphs, but this would need to be checked, as well as the case of degree-corrected stochastic block models. Furthermore, the reason why the log-likelihood of a stochastic block model C, M for a graph G is equal to the entropy of $\Omega_{C,M}$ is that we consider the microcanonical ensemble, in which all graphs have an equal probability to be generated. It would be interesting to check if similar results can be obtained when computing $\mathbb{P}[G|C, M]$ in the canonical ensemble Peixoto [2012]. Finally, we assumed that for a graph G and two partitions C_1 and C_2 with the same number and sizes of blocks, the associated block-to-block adjacency matrices M_1 and M_2 have the same probability to be generated, and this assumption too could be questioned.

Yet, within this specific class of SBM, our results illustrate a fundamental issue with the stochastic block model statistical inference process. Since the random variable whose distribution we are trying to infer is the whole graph itself, we are performing statistical inference on a single observation. This makes frequentist inference impossible, but bayesian inference also has strong limitations in this context. In particular, the only tool to counterbalance the observation and avoid overfitting is to specify the kind of communities we are looking for

beforehand, through the prior distribution. If it is agnostic about the distribution of edge densities among these communities, the mere minimization of the entropy of the posterior distribution fails to identify the heterogeneity in the edge distribution. Beside refining even more the prior distribution, another approach could be to consider a graph as the aggregated result of a series of edge positioning. If the considered random variable is the position of an edge, a single graph observation contains information about many of its realizations, which reduces the risk of overfitting.

Chapter 5

Edge sequence statistical models prequential inference

As recalled in the state of the art, by defining models as probability distributions, statistical models offer a natural measure of their complexity, the entropy, which makes them comparable. What is more, this probabilistic definition allows to rigorously compute the most likely set of parameters used to generate a given observation, thanks to bayesian inference. This has been applied to community detection in Prokhorenkova and Tikhonov [2019], by defining a stochastic blockmodel with a given set of parameters as a probability distribution on a set of graphs, and applying Bayes' theorem to compute the most likely set of parameters given an observed graph. In Peixoto [2019], the author leverages the fact that in the microcanonical ensemble, the maximisation of the likelihood of a set of parameters is equivalent to the minimization of the entropy of its associated probability distribution to perform inference.

However, as developed in chapter 4, these works relies on probability distributions defined on sets of graphs, which means that the observation of a single graph (which in practice is the most common situation) corresponds to a single realization of the random variable. Even though bayesian inference requires less observations than frequentist inference to be sound, a single realization induces a high risk of overfitting. It is also not trivial to adapt this methodology to compare not only different sets of parameters (such as node partitions in the stochastic blockmodel), but models of a different nature (such as a stochastic blockmodel and a configuration model).

In this chapter, we introduce an alternative point of view on graph statistical models, which relies on probability distributions defined on sets of edges. Because a single graph contains many edges, it implies that the same observed graph corresponds to several realizations of the random variable rather than a single one. As a consequence, the inference of the underlying probability distribution can be made more rigorous using prequential inference Dawid [1984]. It allows to control the number of parameters of the model in order to ensure that it remains below the number of observations, which is a necessary condition to avoid overfitting. Moreover, as it formulates all statistical models in terms of probability distributions on the same set of edges, it provides a natural framework to compare them and find the most relevant one with respect to a given graph.

The chapter is organized as follows. In section 5.1, we introduce edge sequence statistical models and explain how they differ from usual graph statistical models. In section 5.2, we

develop sequential edge probability inference, a theoretical framework to perform inference using probability distribution on sets of edges. We then illustrate in section 5.3 how it can be used both to infer the parameters of a statistical model (subsection 5.3.1) and to compare stochastic blockmodel and configuration model with respect to a given graph (subsection 5.3.2).

Edge sequence statistical model

Definition

Statistical models aim at describing the distribution of edges in a graph as the result of a random process subject to some constraints. As developed in the state of the art, statistical graph models are usually defined as a set of graphs Ω_M and a probability distribution $\mathbb{P}_M$ on this set. There exists two main ways to define such models, inspired from statistical physics: microcanonical and canonical ensembles Cimini et al. [2018], whose definitions can be found in section 2.3 of the state of the art. In both cases, the random variable whose probability distribution is studied is a graph. As in practice we almost always study a single graph, the problem with such a model definition is that statistical inference involves to fit a probability distribution on a single realization of the random variable, which implies a high risk of overfitting.

To overcome this issue, we consider graphs as the aggregated trace of a sequence of edges $E = (e_1, \dots, e_m)$, and define models as probability distributions $\mathbb{P}$ over the set of all edge sequences

$$\mathcal{E} = \bigcup_{m \geqslant 0} \left\{ (e_1, \dots, e_m) \in \left([\![0, n-1]\!]^2 \right)^m \right\}$$

For the simplicity of computations, we consider directed graphs and authorize self-loops but the methodology could easily be adapted for undirected edges and forbidden self-loops by restricting this set of possible edges. We will call this type of models **edge sequence statistical models**.

Edge sequence statistical models naturally generate temporal multigraphs, in which edges are ordered and each edge may appear multiple times. Indeed, even if a given edge (u, v) has already been sampled, its probability to be sampled again is a priori not null. This is a natural way to model many real life interactions, even though this type of graphs is not the most widely used in practice. Fortunately, edge sequence statistical model adapts easily for static and simple graphs, since a static graph can be considered as the trace of a temporal one, in which edge ordering has been dropped.

Definition 3. *We say that an edge sequence* $E = (e_1, \dots, e_m)$ collapses *to a static multigraph* G*, described by its weight matrix* W_G *iff:*

$$\forall u, v \in [\![1, n]\!], W_G[u, v] = |\{k \in [1, m] \mid e_k = (u, v)\}|$$

We denote this $E \downarrow G$*, and for any static multigraph* G *we define the set of edge sequences which collapse to it by*

$$\mathcal{E}_G^{\downarrow} = \{E \mid E \downarrow G\}$$

The probability to generate a static graph G is thus defined as

$$\mathbb{P}^{\downarrow}[G] = \sum_{E \in \mathcal{E}_G^{\downarrow}} \mathbb{P}[E]$$

This definition is not very practical as it implies to compute the probability of all edge sequences in $\mathcal{E}_G^{\downarrow}$ to compute the probability of G. It is very demanding as the size of the set is the multinomial coefficient $\binom{m}{w_1,\dots,w_{n^2}}$. In the rest of the chapter, we restrict ourselves to edge sequence models in which edges are generated independently from one another, from a fixed probability distribution $\mathbb{P}_M$ on $[\![0, n-1]\!]^2$. The probability of a sequence $E = (e_1, \dots e_m)$ is thus

$$\mathbb{P}[E] = \prod_{i=1}^{m} \mathbb{P}_M[e_i]$$

which does not depend on the order of edges in the sequence. Examples are given below of statistical models which verify this hypothesis.

Example. *Let's take some examples to illustrate how frequently used statistical models can be formulated as probability distribution on edges. The simplest model is the fully random Erdos-Reyni. It corresponds to the uniform distribution on $[\![1, n]\!]^2$:*

$$\forall u, v \in [1, n], \mathbb{P}_{ER(n)}[u, v] = \frac{1}{n^2}$$

Then, the configuration model: instead of a degree sequence, it takes as parameter a probability distribution $(p_i)_{i \in [1,n]}$ corresponding for each node to its probability of being picked at random as an extremity of the generated edge:

$$\forall u, v \in [1, n], \mathbb{P}_{CFM((p_i)_i)}[u, v] = p_u \times p_v$$

It's directed version is straightforward, considering the probability distributions $(p_i^{out})_i$ and $(p_i^{in})_i$.
Finally, the stochastic blockmodel takes as parameter a partition $B = (b_1, \dots, b_p)$ and a block probability matrix $P \in M_p([0; 1])$ such that $\sum_{i,j} |b_i||b_j|P_{i,j} = 1$. If $u \in b_i$ and $v \in b_j$, the edge (u, v) is generated with probability:

$$\mathbb{P}_{SBM(B,P)}[u, v] = P_{i,j}$$

These models fit well within the assumption of independent edge generation. On the contrary, this is not the case, for instance, of the preferential attachment model of Barabasì and Albert Barabasi and Albert [1999]. It is naturally described as an edge sequence probability distribution, but at each step the probability to generate an edge depends on the previously generated ones.

Under this assumption, as all edge sequences in $\mathcal{E}_G^{\downarrow}$ contain the same edges with the same

multiplicity, by definition we have:

$$\begin{aligned}\forall E_0 \in \mathcal{E}_G^{\downarrow},\ \mathbb{P}_M[G] &= \sum_{E \in \mathcal{E}_G^{\downarrow}} \mathbb{P}_M[E] \\ &= |\mathcal{E}_G^{\downarrow}| \times \mathbb{P}_M[E_0]\end{aligned}$$

This means that, from a probabilistic point of view, any sequence $E_0 \in \mathcal{E}_G^{\downarrow}$ can equivalently be chosen as a representative of G.

Let's stress that the assumption of independent edge generation is made only for the sake of tractability of the study of static graphs. The edge sequence framework is of course particularly natural if one wants to study temporal graphs, in which case there is a priori no reason to make such an assumption.

Beyond edge ordering, considering a simple graph means that we also discard edge multiplicity.

Definition 4. *We say that an edge sequence* simplifies *to a static simple graph G described by its adjacency matrix A_G iff:*

$$\forall u, v \in [\![1, n]\!], A_G[u, v] = \mathbb{1}_{(u,v)\in E}$$

We denote this $E \Downarrow G$, and for any static simple graph G with m edges, we define the set of edge sequences which simplify to it, by:

$$\begin{aligned}\mathcal{E}_G^{\Downarrow k} &= \{|E| = m + k \mid E \Downarrow G\} \\ \mathcal{E}_G^{\Downarrow} &= \bigcup_{k \geqslant 0} \mathcal{E}_G^{\Downarrow k}\end{aligned}$$

The number of edge sequences of length $(m+k)$ which simplify to G grows exponentially with k as $m!m^k \leqslant |\mathcal{E}_G^{\Downarrow k}| \leqslant m^{m+k}$. On the other hand, the probability to sample longer sequences decreases exponentially with k

$$\forall M, \forall E \in \mathcal{E}_G^{\Downarrow k}, \mathbb{P}_M[E] \leqslant \prod_{e \in G} \mathbb{P}_M[e] \times p_0^k \text{ with } p_0 = \max_{e \in G} \mathbb{P}_M[e]$$

Therefore, as long as we consider models M such that $\exists K, \max_{e\in G} \mathbb{P}_M[e] \leqslant \frac{K}{n^2}$ and $\frac{m}{n^2} \ll \frac{1}{K}$, the weight of $\mathcal{E}_G^{\Downarrow k}$ decreases exponentially with k in $\mathcal{E}_G^{\Downarrow}$. Thus, we assume that the weight is concentrated on $\mathcal{E}_G^{\downarrow} = \mathcal{E}_G^{\Downarrow 0}$ and that we can choose a representative of G, $E_0 \in \mathcal{E}_G^{\downarrow}$.

At this point, it is worth stressing how the graph and edge sequence model formulation differ in the very definition of models. For the sake of simplicity, let's consider the Erdös-Rényi model for multigraphs with n nodes and m edges. In the microcanonical formulation, each multigraph in

$$\Omega_{ER(n,m)} = \left\{ G \mid \sum_{u,v \in V^2} W_G[u, v] = m \right\}$$

is generated with the same probability

$$\mathbb{P}_{ER(n,m)}[G] = \frac{1}{|\Omega_{ER(n,m)}|}$$

On the other hand, in the edge sequence based formulation it is the edges in the sequence which are generated uniformly with probability $\frac{1}{n^2}$, such that any sequence of length m is generated with probability $\frac{1}{n^{2m}}$. Therefore, a graph $G \in \Omega_{ER(n,m)}$ is generated with probability

$$\begin{aligned}\mathbb{P}_M[G] &= |\mathcal{E}_G^{\downarrow}| \times \mathbb{P}_M[E_0] \\ &= \binom{m}{w_1, \dots, w_{n^2}} \times \frac{1}{n^{2m}}\end{aligned}$$

which is clearly not the uniform distribution on $\Omega_{ER(n,m)}$. A multigraph whose links are concentrated on a single pair of nodes (u, v) with weight m will be generated with probability $\frac{1}{n^{2m}}$ while a multigraph with m different edges of weight 1 (we suppose here that $m \leqslant n^2$) will be generated with probability $\frac{m!}{n^{2m}}$. This illustrates how the choice of the fundamental elements in generating a graph (vertex, edge or the whole graph itself) subsequently modifies the set on which maximum entropy probability distributions are computed, and therefore the probability distribution associated with a model.

Edge probability distribution statistical inference

As an edge statistical model is defined as a probability distribution on $[\![1, n]\!]^2$, an edge sequence E corresponds to m independent realizations of a random variable following the same unknown probability distribution $\mathbb{P}_0$. The objective of statistical inference is to make an estimation $\mathbb{Q}^*(E)$ of $\mathbb{P}_0$, avoiding both overfitting and underfitting, among the set of all possible models.

Definition 5. *Let $\mathcal{M}_n^{\bullet}([0,1])$ be the set of all probability distributions on $[\![1, n]\!]^2$:*

$$\mathcal{M}_n^{\bullet}([0,1]) = \left\{ \mathbb{Q} \in \mathcal{M}_n([0,1]) \mid \sum_{u,v \in [\![1,n]\!]^2} \mathbb{Q}[u,v] = 1 \right\}.$$

Its elements can be seen as $n \times n$ matrices or as probability distributions. In the following we will use both points of view.

We use the cross entropy $\mathbb{H}[\mathbb{P}, \mathbb{Q}] = -\sum_{u,v} \mathbb{P}[u,v] \log_2(\mathbb{Q}[u,v])$ as a measure of similarity on $\mathcal{M}_n^{\bullet}([0,1])$. It can be understood as the expected length of a message generated following $\mathbb{P}$ but encoded with a code optimal for $\mathbb{Q}$. It is minimal when $\mathbb{Q} = \mathbb{P}$, in which case it is equal to the entropy $\mathbb{S}[\mathbb{P}]$. In this paper, the sequence to encode will be E, therefore the best compression is achieved for a code based on the empirical distribution $\mathbb{P}_E$.

Definition 6. *Let $\mathbb{P}_E$ be the* empirical distribution

$$\forall (u,v) \in [\![1, n]\!]^2, \mathbb{P}_E[u,v] = \frac{\#\{k \mid e_k = u \to v\}}{m}.$$

We can observe that this naive estimation leads to overfitting, as the corresponding code would probably perform poorly for another sequence E' generated using the same original distribution $\mathbb{P}_0$. On the other hand, the most general code, which performs equally well on all possible edge sequences, is obtained based on the uniform distribution $\mathbb{P}_U$, but it is clearly underfitting as this code does not tell us anything about $\mathbb{P}_0$. This is illustrated on Figure 5.1.

Definition 7. *We say that an estimation* $\mathbb{Q}^*(E)$ *of* $\mathbb{P}_0$ *is* overfitting *if*

$$\mathbb{H}[\mathbb{P}_E, \mathbb{Q}^*(E)] < \mathbb{H}[\mathbb{P}_E, \mathbb{P}_0]$$

on the other hand, we say it is underfitting *if*

$$\mathbb{H}[\mathbb{P}_E, \mathbb{Q}^*(E)] > \mathbb{H}[\mathbb{P}_E, \mathbb{P}_0]$$

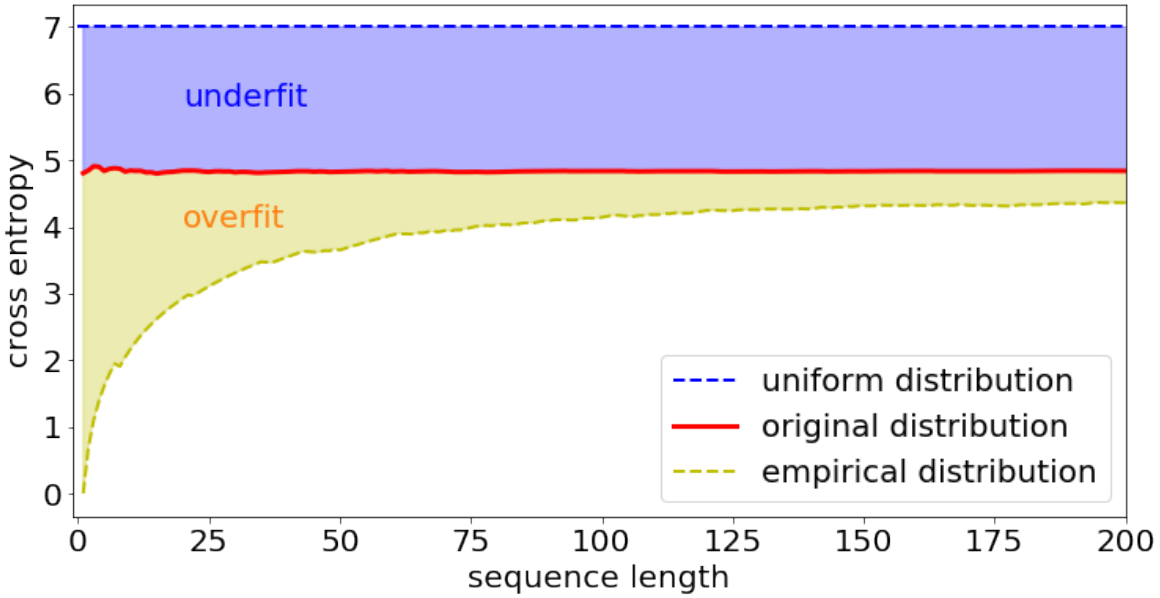


Figure 5.1 – Given an original probability distribution $\mathbb{P}_0$, we generate a sequence of edges $(e_1, \dots, e_m)$. For $k \in [\![1, m]\!]$, we plot the cross entropy of the empirical distribution $\mathbb{P}_{(e_1,\dots,e_k)}$ with the uniform distribution (blue line), original distribution (red line), and the empirical distribution itself (yellow line), against k. We say that an estimation $\mathbb{Q}^*(e_1, \dots, e_k)$ is overfitting if $\mathbb{H}(\mathbb{P}_{(e_1,\dots,e_k)}, \mathbb{Q}^*(e_1, \dots, e_k))$ lies in the yellow zone, and that it is underfitting if it lies in the blue zone.

For a given sequence E, it is very likely that our estimation $\mathbb{Q}^*(E)$ will be at least slightly overfitting or underfitting, but our objective is that

$$\frac{\mathop{\mathbb{E}}_{e_i \sim \mathbb{P}_0} \left[\mathbb{H}[\mathbb{P}_{(e_1,\dots,e_m)}, \mathbb{Q}^*(e_1, \dots, e_m)]\right]}{\mathop{\mathbb{E}}_{e_i \sim \mathbb{P}_0} \left[\mathbb{H}[\mathbb{P}_{(e_1,\dots,e_m)}, \mathbb{P}_0]\right]} \underset{m \to \infty}{\longrightarrow} 1$$

The main risk of overfitting comes from the fact that estimating $\mathbb{Q}^*(E)$ implies the inference of $n^2 - 1$ parameters: we infer $\mathbb{Q}^*(E)[u, v]$ for each $(u, v) \in [\![1, n]\!]^2$, under the constraint that $\phi_0(\mathbb{Q}^*(E)) = \sum_{u,v} \mathbb{Q}^*(E)[u, v] - 1 = 0$. As m is typically much smaller than n^2, such a large number of parameters induces a high risk of overfitting.

To avoid this phenomenon, we need to make assumptions about $\mathbb{P}_0$ in order to restrict the search space. We do so by introducing hyperparameters to control the number of degrees

of freedom of the model by adding constraints on the probability distribution. A hyperparameter can be described as a function

$$\begin{aligned}\phi\colon M_n(\mathbb{R}) &\to \mathbb{R}^{s+1}\\ \mathbb{Q} &\mapsto (\phi_0(\mathbb{Q}), \dots, \phi_s(\mathbb{Q})\end{aligned}$$

where $\phi_0(\mathbb{Q}) = \sum_{u,v} \mathbb{Q}[u,v] - 1$ is the basic constraint assuring that $\mathbb{Q}$ belongs to $\mathcal{M}_n^\bullet([0,1])$ and s is the number of additional constraints. The search space under these constraints is reduced to:

$$\mathcal{M}_n^\phi([0,1]) = \{\mathbb{Q} \in \mathcal{M}_n([0,1]) \mid \phi(\mathbb{Q}) = 0\}$$

We can suppose that the constraints are independent (if they are not, it means that the same search space could be obtained with less constraints). Thus, the number of parameters to infer boils down to $n^2 - s - 1$.

Example. *Let's assume that $\mathbb{P}_0$ is a stochastic blockmodel based on a partition $B = (b_1, \dots, b_p)$. According to the definition given above, it means that*

$$\exists M \in \mathcal{M}_p([0,1]), \forall u \in b_i, v \in b_j, \mathbb{P}_0[u,v] = M[i,j]$$

It is equivalent to say that

$$\forall i,j \in [\![1,p]\!], \forall u,u' \in b_i, \forall v,v' \in b_j, \mathbb{P}_0[u,v] - \mathbb{P}_0[u',v'] = 0$$

which corresponds to a system of $n^2 - p^2$ linearly independent constraints. Thus, under this assumption, we are left with only $p^2 - 1$ parameters to infer.

Therefore, edge statistical model selection involves two distinct issues:

1. For each possible hyperparameter ϕ, estimate the probability distribution $\mathbb{Q}_\phi^*(E)$ that most likely generated E in $\mathcal{M}_n^\phi([0,1])$.
2. Select the best model $\mathbb{Q}^*(E)$ among all possible estimate $(\mathbb{Q}_\phi^*(E))_\phi$.

These two main questions are discussed in the next section.

Edge sequence model selection

Parameter inference by minimum description length

Let's consider first the issue of estimating the probability distribution $\mathbb{Q}_\phi^*(E)$ that most likely generated E in $\mathcal{M}_n^\phi([0,1])$, given the hyperparameter ϕ. We rely on the minimum description length principle (a detailed tutorial can be found in Grunwald [2004]; Grünwald and Roos [2019]). It states that, as any regularity in a sequence of observations can be used to compress it, the best statistical model for the sequence E is the one which minimizes the description

length of the model $D_\phi(\mathbb{Q})$ plus the description length of the observations compressed using this model $D(E|\mathbb{Q})$:

$$\mathbb{Q}^*_\phi(E) = \underset{\mathbb{Q}\in\mathcal{M}^\phi_n([0,1])}{\operatorname{argmin}} \; D(E|\mathbb{Q}) + D_\phi(\mathbb{Q}) \tag{5.1}$$

The description length of the sequence can be computed as

$$D(E|\mathbb{Q}) = -\sum_{i=1}^{m} \log_2(\mathbb{Q}[e_i])$$

as detailed in Annex 7.2.4.

Then, to compute the description length of the model $D_\phi(\mathbb{Q})$, we need to define a probability distribution $\bar{\mathbb{P}}_\phi$ on $\mathcal{M}^\phi_n([0,1])$. This so-called prior distribution is used to encode the model with a length $D_\phi(\mathbb{Q}) = -\log_2(\bar{\mathbb{P}}_\phi[\mathbb{Q}])$. The goal of this term is to take into account the complexity of the model, in order to avoid overfitting. Therefore, simpler models should have shorter description length. To achieve this, we define the prior distribution such that the description length of a model is inversely proportional to its information content, measured by its entropy:

$$\bar{\mathbb{P}}_\phi[\mathbb{Q}] = \frac{1}{Z_\phi} \times 2^{\mathbb{S}[\mathbb{Q}]}$$

with $Z_\phi = \int_{\mathcal{M}^\phi_n([0,1])} 2^{\mathbb{S}[\mathbb{Q}]}\, \mathrm{d}\mathbb{Q}$ a normalization constant to ensure that $\bar{\mathbb{P}}_\phi[\mathbb{Q}]$ integrates to 1 over $\mathcal{M}^\phi_n([0,1])$.

Remark. *The expression prior distribution we used to refer to $\bar{\mathbb{P}}_\phi$ refers to the bayesian terminology. This is on purpose, as this approach is equivalent to bayesian statistical inference, as detailed in Annex 7.2.5.*

With this definition of $\bar{\mathbb{P}}_\phi$,

$$\begin{aligned} D_\phi[\mathbb{Q}] &= -\log_2(\bar{\mathbb{P}}_\phi[\mathbb{Q}]) \\ &= -\mathbb{S}[\mathbb{Q}] + \log_2(Z_\phi) \end{aligned}$$

As Z_ϕ is constant on $\mathcal{M}^\phi_n([0,1])$, we can neglect it in the minimization and equation 5.1 becomes

$$\mathbb{Q}^*_\phi(E) = \underset{\mathbb{Q}\in\mathcal{M}^\phi_n([0,1])}{\operatorname{argmin}} \; -\sum_{i=1}^{m} \log_2(\mathbb{Q}[e_i]) - \mathbb{S}[\mathbb{Q}] \tag{5.2}$$

In the following, we denote

$$\mathrm{f}(\mathbb{Q}, E) = -\sum_{i=1}^{m} \log_2(\mathbb{Q}[e_i]) + \sum_{u,v} \mathbb{Q}[u,v] \log_2(\mathbb{Q}[u,v])$$

and thus we can rewrite equation 5.2 as

$$\mathbb{Q}^*_\phi(E) = \underset{\mathbb{Q}\in\mathcal{M}^\phi_n([0,1])}{\operatorname{argmin}} \; \mathrm{f}(\mathbb{Q}, E) \tag{5.3}$$

We have the following property (see proof in Annex 7.2.1):

Property 1. *If $\mathcal{M}_n^\phi([0,1])$ is a convex set, then for any edge sequence E, $\mathfrak{f}$ has a unique minimum $\mathbb{Q}_\phi^*(E)$ over $\mathcal{M}_n^\phi([0,1])$.*

Remark. *In particular, if ϕ is an affine function, $\mathcal{M}_n^\phi([0,1])$ is the intersection of an affine subspace of $\mathcal{M}_n(\mathbb{R})$ with $[0,1]^{n^2}$. Consequently, it is convex and $\mathbb{Q}_\phi^*(E)$ exists and is unique.*

According to the Lagrange multiplier theorem, this minimum verifies

$$\exists(\lambda_j) \in \mathbb{R}^{s+1}, \vec{\nabla} f(\mathbb{Q}_\phi^*(E), E) + \sum_{j=1}^{s+1} \lambda_j \vec{\nabla} \phi_j(Q_\phi^*(E)) = 0$$

This is a set of $n^2 + s + 1$ equations with as many unknowns which we solve numerically using Newton's method.

Finally, we obtain the following result (see proof in Annex 7.2.3):

Theorem 2. *Let $(e_i)_{i\in\mathbb{N}}$ be a sequence of independent and identically distributed random variables following $\mathbb{P}_0 \in \mathcal{M}_n^\bullet([0,1])$.*

$$\forall \phi, \mathbb{Q}_\phi^*(e_1, \ldots, e_x) \underset{x\to\infty}{\longrightarrow} \underset{\mathbb{Q}\in\mathcal{M}_n^\phi([0,1])}{\operatorname{argmin}} \mathbb{H}(\mathbb{P}_0, \mathbb{Q})$$

Remark. *In particular, if $\mathbb{P}_0$ belongs to $\mathcal{M}_n^\phi([0,1])$, it means that $\mathbb{Q}_\phi^*(E)$ converges toward $\mathbb{P}_0$ as the number of observations grows.*

Hyperparameter selection by prequential inference

Now that we know how to infer $\mathbb{Q}_\phi^*(E)$ for any given ϕ, the second step for model selection consists in choosing the best estimation $\mathbb{Q}^*(E)$ among them. Let's consider a set of hyperparameters $\Phi = \{\phi_1, \ldots, \phi_q\}$. To select the best hyperparameter $\phi^*(E) \in \Phi$, we keep using the minimum description length principle.

However, applying it straightforwardly leads to the same risk of overfitting as if we had considered the whole graph to be our random variable in the first place. It would annihilate the advantage of considering each edge as an independent observation, and thus the very reason to use edge statistical model rather than graph statistical models. Indeed, the naive approach would be to select ϕ^* by minimizing

$$\phi^* = \underset{\phi\in\Phi}{\operatorname{argmin}}\, D(E|\phi) + D(\phi)$$

where the description length of the hyperparameter is defined based ona a prior distribution $D(\phi) = -\log_2(\bar{\mathbb{P}}[\phi])$ and the description length $D(E|\phi)$ of E given a hyperparameter ϕ is defined as its description length using the best model compatible with ϕ, $\mathbb{Q}_\phi^*(E)$.

$$\begin{aligned} D[E|\phi] &= -\sum_{i=1}^{m} \log_2(\mathbb{Q}_\phi^*(E)[e_i]) + D_\phi[\mathbb{Q}_\phi^*(E)] \\ &= m \times \mathbb{H}[\mathbb{P}_E, \mathbb{Q}_\phi^*(E)] + D_\phi[\mathbb{Q}_\phi^*(E)] \end{aligned}$$

In this case, it is clear that $D[E|\phi]$ would be minimum for the null hyperparameter $\phi_0 : \mathbb{Q} \to \sum_{u,v} \mathbb{Q}[u,v] - 1$ because in this case $\mathcal{M}_n^{\phi_0}([0,1]) = \mathcal{M}_n^{\bullet}([0,1])$, so $\mathbb{P}_E \in \mathcal{M}_n^{\phi_0}([0,1])$ and $\mathbb{Q}_{\phi_0}^*(E) = \mathbb{P}_E$, which by definition minimizes $\mathbb{H}[\mathbb{P}_E, \mathbb{Q}_\phi^*(E)$. Yet, this model is just a copy of the observations and this would be a total overfit. The only way to mitigate this overfitting would be to rely on an ad hoc prior distribution $\bar{\mathbb{P}}$ on Φ, which is exactly what is done in Peixoto [2019] in the case of graph statistical model and the microcanonical ensemble.

Our objective is to define a methodology which does not need this ad hoc prior distribution. We do not use the full sequence of edge E to optimize the model given the hyperparameter, but rather a training set of edges L to compute the optimal model $\mathbb{Q}_\phi^*(L)$ for each hyperparameter, and then use this model to define the description length of E given an hyperparameter ϕ and a learning set L

$$\begin{aligned} D_L[E|\phi] &= -\sum_{i=1}^{m} \log_2(\mathbb{Q}_\phi^*(L)[e_i]) + D_\phi[\mathbb{Q}_\phi^*(L)] \\ &= m \times \mathbb{H}[\mathbb{P}_E, \mathbb{Q}_\phi^*(L)] + D_\phi[\mathbb{Q}_\phi^*(L)] \end{aligned}$$

In practice, L is necessarily a subset of E, but the question is its size. The smaller it is, the more we risk underfitting: the extreme example is for $L = \varnothing$, because then $\forall \phi, \mathbb{Q}_\phi^*(L)$ is the uniform distribution, which is the extreme case of underfitting. On the other hand, the larger the size of the learning set, the more we favour hyperparameters with many degrees of freedom and risk overfitting: if $L = E$, we get back to the previously described issue.

Choosing a fixed size for the learning set (or a fixed proportion of the total number of edges) would still amount to arbitrarily decide where to put the limit between overfitting and underfitting. Instead, our aim is that this limit is discovered based on the dataself itself. To do so, we use prequential inference instead of a fixed learning set. This statistical inference methodology was introduced in Dawid [1984] and its connection with the minimum description length principle is developed in Barron et al. [1998]; Grünwald and Roos [2019]. To understand the difference, let's go back to the basis and consider the situation where E is a sequence of messages that a source (Alice) draws at random and transmits to a destination (Bob).

In the classical minimum description length setting, the prior distribution on all possible hyperparameters corresponds to code for each hyperparameter that Alice and Bob agree on beforehand. They also agree on the length m' of the subset to be used as learning set. When drawing at random a sequence of edges E to transmit, Alice reads the learning set L made of the first m' edges, she computes the optimal model $\mathbb{Q}_\phi^*(L)$ for each hyperparameter, and the optimal hyperparameter $\phi^* = \underset{\phi \in \Phi}{\operatorname{argmin}}\, D_L(E|\phi) + D(\phi)$. Finally, she sends to Bob the code of the optimal hyperparameter $\phi*$, the code of the optimal model given this hyperparameter $\mathbb{Q}_\phi^*(L)$, and the code for the sequence, given this model.

In prequential inference on the other hand, instead of using a fixed code $C^*(L)$, Alice updates the optimal model (and thus her code) for each hyperparameter as she observes more and more edges. At step k, Alice has observed edges $(e_1, \dots, e_{k-1})$ and she has transmitted them to Bob. Therefore, both of them can compute $\mathbb{Q}_\phi^*(e_1, \dots, e_{k-1})$ and the corresponding code $C_\phi^*(k-1)$. Alice draws the edge e_k and transmits it to Bob using this code. Then Alice

and Bob both update their code to $C^*_\phi(k)$, and so on. This way, the description length is

$$D[E|\phi] = -\sum_{k=1}^{m} \log_2(\mathbb{Q}^*_\phi(e_1,\dots,e_{k-1})[e_k])$$

At step k, the probability distribution $\mathbb{Q}^*_\phi(e_1,\dots,e_{k-1})$ is the model that best fits the first $k-1$ observations within $\mathcal{M}^\phi_n([0,1])$. Thus, $\mathbb{Q}^*_\phi(e_1,\dots,e_{k-1})[e_k]$ is the probability, given those observations and the hyperparameter, to correctly guess the k^{th} edge, and $-\log_2(\mathbb{Q}^*_\phi(e_1,\dots,e_{k-1})[e_k])$ can be interpreted as the quantity of information about e_k contained in the previous edges. This way, we do not need to rely on an ad hoc prior distribution to avoid overfitting. Hyperparameters which give too much degrees of freedom to the model induce models which are close to already observed edges but do not necessarily predict well the next ones. On the other hand, hyperparameters that do not have enough degrees of freedom induce models which are not able to capture the statistical reguarities present in observed edges to predict the next ones. Overall, both lead to poor description lengths.

As we do not rely on the prior distribution to counterbalance overfitting, we can use a non-informative uniform distribution on Φ as prior distribution, so $D(\Phi)$ is constant and the best hyperparameter is computed as

$$\phi^*(E) = \underset{\phi\in\Phi}{\operatorname{argmin}} \; -\sum_{k=1}^{m} \log_2(\mathbb{Q}^*_\phi(e_1,\dots,e_{k-1})[e_k])$$

Overall, the model selected is

$$\mathbb{Q}^*(E) = \mathbb{Q}^*_{\phi^*(E)}(E)$$

Prequential inference implies that the optimal model $\mathbb{Q}^*(E)$ is dependent on the order of edges in E. This means that if the studied graph G is static, the model selected depends on the ordering of edges we make when we choose a representative $E \in \mathcal{E}^{\downarrow}_G$. However, as we assumed in the beginning that all edges were indepently generated from the same probability distribution, the procedure will always converge toward the same distribution so if there are enough edges in the graph the results should not vary much with the order of edges. In practice, we observe that changing edge ordering have little impact on the results even for graphs with realistic densities.

Applications to model selection

Stochastic blockmodel partition selection

Finding the appropriate number of blocks of the partition. To test prequential edge probability inference, we start by using it to tackle the classical problem of partition selection in stochastic blockmodels. We consider an edge sequence $E = (e_1,\dots,e_m)$ which we assume was generated by a stochastic blockmodel $\mathbb{P}_0$ based on a partition B_0, as described in section 5.1.1. Our objective is to retrieve $\mathbb{P}_0$ and B_0 among the set of all possible stochastic blockmodels. Each partition $B = (b_1,\dots,b_p)$ of $[\![1,n]\!]$ corresponds to a hyperparameter ϕ^B

made of $n^2 - p^2 + 1$ constraints. If we designate inside each block b_i a representative u_i, this hyperparameter can be expressed as:

$$\phi_0^B(\mathbb{Q}) = 1 - \sum_{u,v} \mathbb{Q}[u,v]$$
$$\forall i,j, \forall u \in b_i \backslash \{u_i\}, \forall v \in b_j \backslash \{u_j\}, \phi_{u,v}^B(\mathbb{Q}) = \mathbb{Q}[u_i,u_j] - \mathbb{Q}[u,v]$$

The constraint $\phi^B(\mathbb{Q}) = 0$ expresses the fact that $\mathbb{Q}$ is a probability distribution and that edge generation probabilities are constant along the blocks defined by B. Thus, selecting the partition within a set $\{B_1, \dots, B_q\}$ that is more likely to be the original one boils down to the inference of the most likely hyperparameter in $\Phi = \{\phi^{B_1}, \dots, \phi^{B_q}\}$. In particular, it should be noted that all those hyperparameters are affine functions, so Remark 5.2.1 tells us that for each of them, $\mathbb{Q}_\phi^*(E)$ exists, is unique, and can be computed using Lagrange multipliers and Newton's method.

Exploring the full partition space is a challenge on its own, as this space grows exponentially with n. Therefore, to perform our test, we generate synthetic graphs with a stochastic blockmodel and observe how it behaves for a particular subset of the possible partitions of the nodes. Of course, this means that we cannot be sure that the minimum we find corresponds to the minimum over every possible partition. Yet, it allows us to test the robustness of prequential edge probability inference against common pitfalls, and in particular with respect to partitions which are a coarsening or a refinement of the original partition.

We consider a stochastic blockmodel $S_0 = (B_0, M_0)$ on 128 nodes divided into 4 blocks:

$$B = [\![1,32]\!], [\![33,64]\!], [\![65,96]\!], [\![97,128]\!]$$
$$M = \frac{1}{128^2} \cdot \begin{bmatrix} 4 & 0 & 0 & 0 \\ 0 & 4 & 0 & 0 \\ 0 & 0 & 4 & 0 \\ 0 & 0 & 0 & 4 \end{bmatrix}$$

Remark. *As the stochastic blockmodel defined here is an edge statistical model, the coefficients $M[i,j]$ should not be interpreted as the density between blocks i and j. They are the probability for each edge going from block i to block j to be generated:*

$$\forall u \in b_i, v \in b_j, \mathbb{P}_{S_0}[u,v] = M[i,j]$$

We generate 50 graphs with S_0 and test 8 hyperparameters corresponding to partitions refined from 1 block to 128. Each partition is obtained by dividing the blocks of the previous one in half. We plot the mean prediction probability $\frac{1}{m}\sum_{k=1}^{m} \mathbb{Q}_{\phi^B}^*(e_1, \dots, e_{k-1})[e_k]$ against the number of blocks in B. Results are shown in Figure 5.2. We observe that the mean prediction probability rises as the number of blocks of the partition grows from one to four, which corresponds to the original partition used to generate the graphs. Then, further refinement of the partition used as hyperparameter does not bring significant increase in the mean prediction probability.

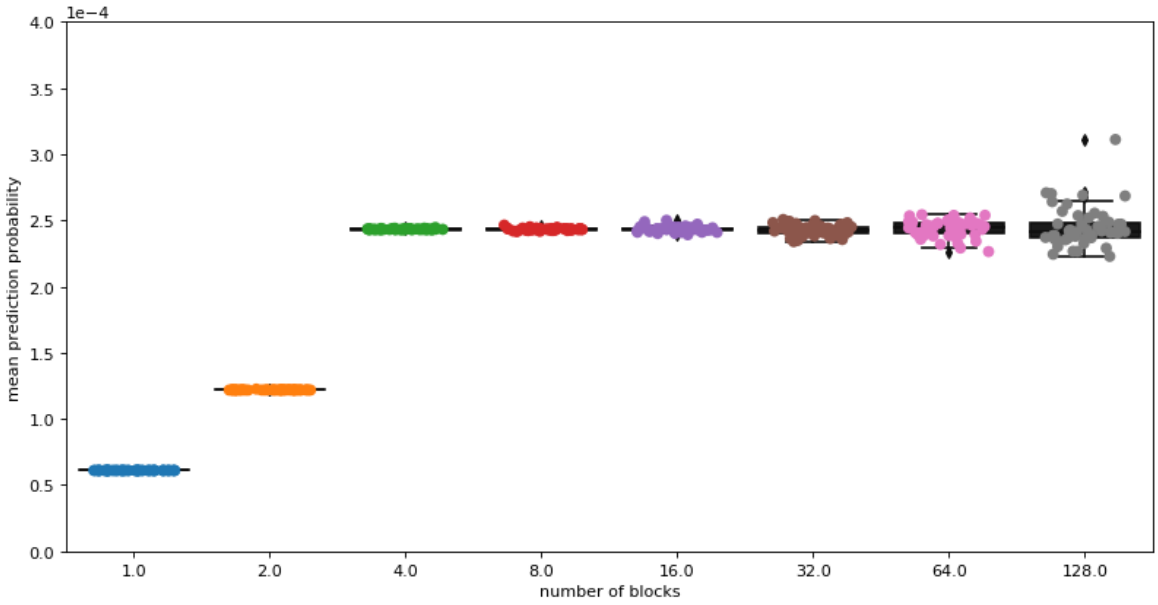


Figure 5.2 – For each of the 50 graphs generated with $S_0 = (B_0, M_0)$, we plot the mean prediction probability against the number of blocks of the hyperparameter for partitions ranging from 1 single block to 128 blocks containing a single node.

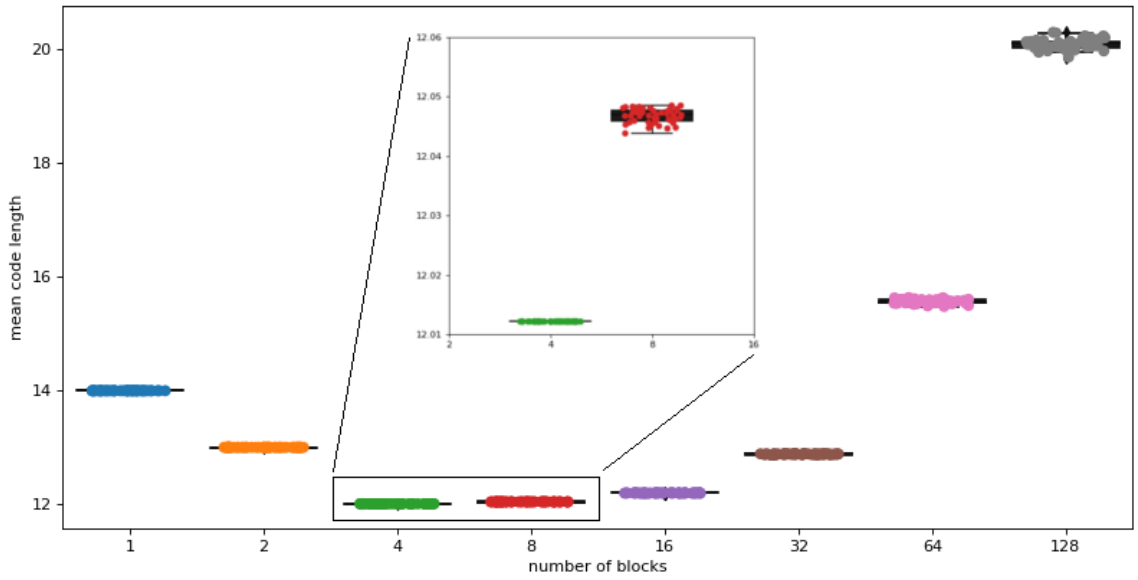
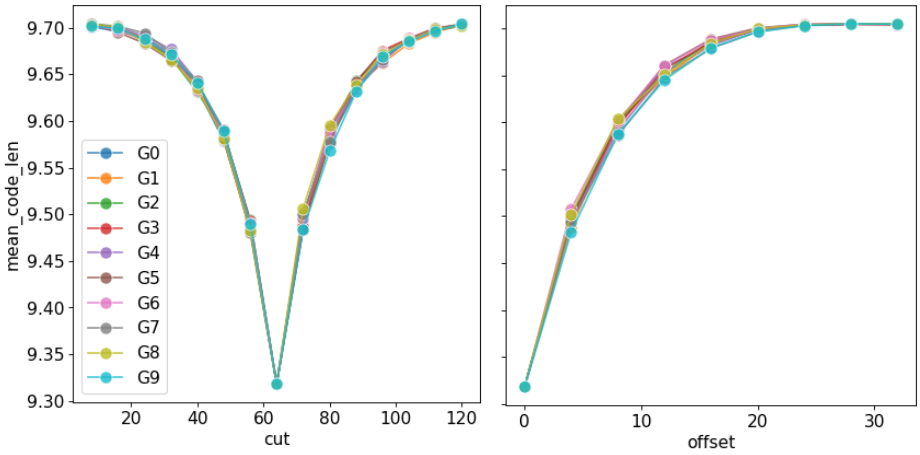


Figure 5.3 – For the 50 graphs generated with S_0, and the eight partitions obtained by coarsening / refining B_0, we plot the mean code length against the number of blocks in the partition.

Then, we plot the mean code length $-\frac{1}{m}\sum_{k=1}^{m}\log_2(\mathbb{Q}^*_{\phi^B}(e_1,\ldots,e_{k-1})[e_k])$ against number of blocks in B in Figure 5.3. The mean code length is proportional to the description length $D[E|\phi^B]$ so they have the same minimum, but it has the advantage of being insensitive to the

Figure 5.4 – Mean code length against cut (left) and offset (right)



length of the edge sequence. We observe that for all fifty graphs, it presents a clear minimum at the original four blocks partition B_0. For coarser partitions, the mean code length is higher because, as illustrated in Figure 5.2, the prediction probability is lower and

$$\begin{aligned} &\mathbb{Q}^*_{\phi^B}(e_1,\dots,e_{k-1})[e_k] < \mathbb{Q}^*_{\phi^{B_2}}(e_1,\dots,e_{k-1})[e_k] \\ \implies &-\log_2(\mathbb{Q}^*_{\phi^B}(e_1,\dots,e_{k-1})[e_k]) > -\log_2(\mathbb{Q}^*_{\phi^{B_2}}(e_1,\dots,e_{k-1})[e_k]) \end{aligned}$$

Then, for finer partitions, it is due to the slower convergence rate. Indeed, as logarithm is a concave function

$$-\log_2\left(\frac{1}{m}\sum_{k=1}^{m}\mathbb{Q}^*_{\phi^B}(e_1,\dots,e_{k-1})[e_k]\right) < -\frac{1}{m}\sum_{k=1}^{m}\log_2(\mathbb{Q}^*_{\phi^B}(e_1,\dots,e_{k-1})[e_k])$$

Therefore, the greater the fluctuations of the prediction probability, the higher the mean code length. More details about the convergence of the prediction probability depending on the hyperparameter can be found in Annex 7.2.6. In the end, the minimum description length makes it possible to retrieve the original partition B_0, avoiding both overfitting and underfitting, with no previous knowledge or assumption about the number of blocks.

Cutoff and offset We then considered the performance of the mean code length when modifing blocks' sizes or shifting blocks. To do so, we generated 10 graphs with 128 nodes and 2800 edges, made of two perfectly separated communities of equal size. Then, for each of these graphs, we computed the mean code length for two sequence of partitions.

- $S_{cut} = (B(c) = ([1,c],[c,128]))_{c\in\{0,8,16,24,\dots,128\}}$
- $S_{offset} = (B(o) = ([1+o,64+o],[1,o]\cup[65+o,128])_{o\in\{0,4,8,12,\dots,32\}})$

Results are plotted, respectively against c and o, on figure 5.4.

We observe that for all graphs, the minimum of mean code length is reached when $c = 64$ in the first sequence, and when $o = 0$ in the second, which both correspond to the partition B_1 used to generate them. This means that mean code length is robust against shifting blocks and modifying blocks' sizes.

Merge / split issue. As shown in chapter 4, stochastic blockmodel selection based on the minimization of the microcanonical ensemble entropy, even though it is statistically grounded, may be subject to overfitting in the sense that splitting large communities while merging small ones may lead to a lower entropy because it imposes more constraints on edges' position.

To illustrate how prequential edge probability inference helps solving this problem, let's consider a stochastic blockmodel $S_1 = (B, M)$ defined on a set of $n = 12$ nodes:

$$B = [\![0;5]\!], [\![6;8]\!], [\![9;11]\!]$$
$$M = \begin{bmatrix} 0.026 & 0 & 0 \\ 0 & 0.003 & 0 \\ 0 & 0 & 0.003 \end{bmatrix}$$

We test two different partitions: the original one, B, and the inverse partition in which the large communities is split and small ones are merged $B^\dagger = [\![0;2]\!], [\![3;5]\!], [\![6;11]\!]$. To do so, we generate 100 graphs G_i made of $m = 378$ edges with S_1 and for each graph, we compute the mean code length and the entropy (using graph-tool[1]) for both partitions. Then, for both quality functions, we compute the percentage of graphs for which the original partition is identified as better than the inverse one. Results are shown in Table 5.1. While the mean code length almost always correctly identifies the original partition, the entropy of the microcanonical ensemble never does so. The graphs considered here have a very high density, which makes them not very realistic, but same results can be obtained with lower density graphs. Let's consider a stochastic blockmodel S_2 with $n = 256$ nodes, partitioned in 33 communities, one of size 128, and 32 of size 4:

$$B = [\![1, 128]\!], [\![129, 132]\!], [\![133, 136]\!], \ldots, [\![253, 256]\!]$$

The internal probability of the big community is 6×10^{-5}, the one of the small communities is 7.6×10^{-4}, and the probability between communities is null. We compare this original partition with the inverse one:

$$B^\dagger = [\![1, 4]\!], [\![5, 8]\!], \ldots, [\![125, 128]\!], [\![129, 256]\!]$$

We generate 100 graphs with S_2 and compute for each of them the entropy of both partitions and the mean code length with ϕ_B and $\phi_{B^\dagger}$. Results are shown in Table 5.1. We see that in this case too, the mean code length always identifies the original partition as the best one, while the entropy does not.

[1] `https://graph-tool.skewed.de`

Table 5.1 – Percentage of correct match for heterogeneous graphs.

SBM	Mean code length	Entropy
S_1	96%	0%
S_2	100%	0%

Zachary Karate Club Finally, we test the mean code length quality function on the zachary karate club network. We study three different partitions of it. First of all, the sociological partition, B_{100}, which is the partition described in the original paper as corresponding to the sociological ground truth about communities in the karate club. B_{200} is the partition obtained by minimizing the modularity using the louvain algorithm, and B_{300} the partition obtained by minimizing the entropy using the graph-tool library. Those partitions are illustrated on figure 5.5.

For each of these partitions, we compute the mean code length. We also do so for 100 random partitions of the graph, with 1 to 5 blocks, and for each of these partitions, we compute the mean code length for 99 random refinement of them, obtained by randomly dividing each block in two. Results are plotted on figure 5.6.

We observe that the mean code length is minimum for the minimum entropy partition. All studied partitions perform better than the random ones, so the mean code length captures the fact that they reproduce part of the structure of the network. Yet, for B_{100} and B_{200} many of their random refinements improve the compression, sometimes by a large amount, indicating that they are not optimal. This is not the case for the minimum entropy partition B_{300}. There are only 2 refinements out of 99 which perform a little better, an issue we have seen may happen due to random fluctuations. These results are coherent with previous work showing that B_{100} is actually not fully supported by statistical evidence in the network. In the case of B_{200}, modularity is defined based on nodes' degree, so the selected partition compensate for node degrees, which are not considered here. Finally, minimizing the entropy without correcting for the degree leads to the identification of two blocks of hubs, at the center of each sociological communities, and two blocks corresponding to their periphery. This is not necessarily what we expect, because we are used to communities defined with an implicit or explicit degree correction, but as we have not imposed such constraints so far, this result corresponds to the statistical evidence present in the network.

Stochastic blockmodel and configuration model

The main benefit of edge statistical models is that it provides a common framework to compare models whose parameters lie in different parameter space. To illustrate this, let's consider two widespread models: the stochastic blockmodel and the configuration model. The first one has been introduced in the previous section, so we start by describing the edge-version of the configuration model, and then show how both models can be compared using prequential inference.

We consider the directed version of the configuration model. The classical version of this model takes as parameters the sequences of node in $(k_u^{out})_{u\in V}$ and out $(k_u^{in})_{u\in V}$ degrees. For the edge version, we keep the idea that the probability of generating an edge $u \to v$ is determined by two probability distributions p^{out} and p^{in} over $[\![1, n]\!]$. p_u^{out} is the probability to

Figure 5.5 – Three different partitions of the zachary karate club network. Sociological (upper left), minimum modularity (upper right), minimum entropy (lower)

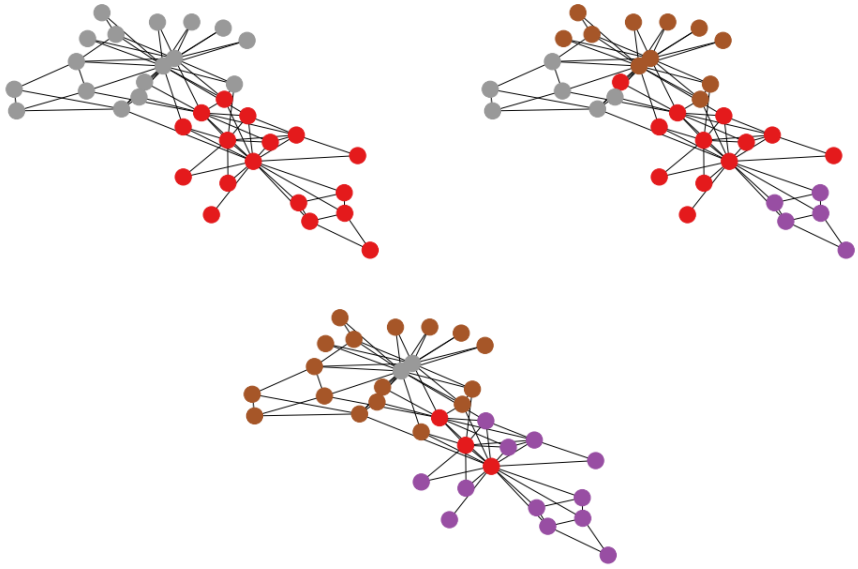
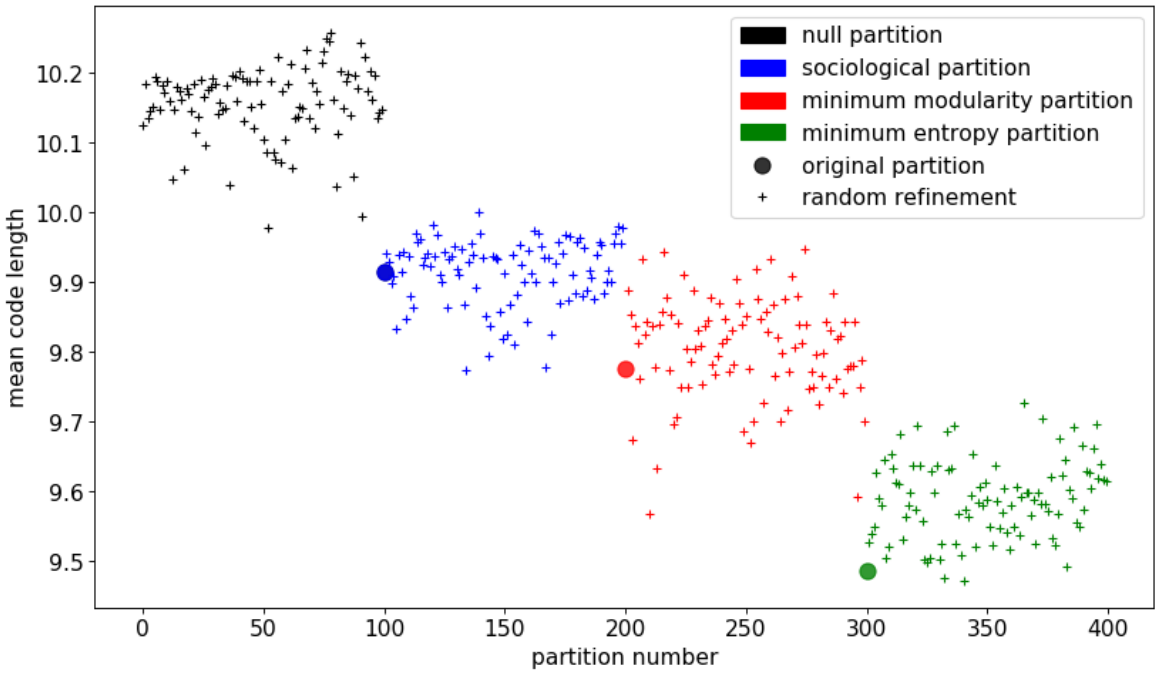


Figure 5.6 – Mean code length for different partitions of the zachary karate club network



pick node u as the source of the edge and p_v^{in} the probability to pick v as its destination:

$$\forall u, v, \mathbb{Q}_{CM}[u, v] = p_u^{out} \times p_v^{in}$$

Therefore, a probability distribution $\mathbb{Q} \in \mathcal{M}_n^{\bullet}([0,1])$ corresponds to a directed configuration model if and only if:

$$\forall u, v, \mathbb{Q}[u, v] \times \mathbb{Q}[1, 1] - \mathbb{Q}[u, 1] \times \mathbb{Q}[1, v] = 0$$

In this case, $p_u^{out} = \sum_v \mathbb{Q}[u, v]$ and $p_v^{in} = \sum_u \mathbb{Q}[u, v]$. This gives us a system of $(n-1)^2$ independent constraints to use as hyperparameter ϕ^{CM}. It is worth noting that this hyperparameter is not an affine function, so Remark 5.2.1 does not apply. However, we have the following result (see proof in Annex 7.2.2):

Property 2. *For any edge sequence E, f has a unique minimum $\mathbb{Q}^*_{\phi^{CM}}(E)$ over $\mathcal{M}_n^{\phi^{CM}}([0,1])$.*

Yet, this still leaves $2n-2$ parameters to infer, which remains high in comparison with the number of observations m and thus induces a risk of overfitting. To overcome this problem, we consider a block version of the configuration model. It means that, given two partitions of $[\![1, n]\!]$, B^{in} and B^{out}, $(p_u^{in})_{u \in [\![1,n]\!]}$ is constant over the blocks of B^{in} and $(p_u^{out})_{u \in [\![1,n]\!]}$ is constant over the blocks of B^{out}. Thus, if B^{in} is made of q^{in} blocks and B^{out} of q^{out} blocks, there are only $q^{out} + q^{in} - 2$ parameters left to infer.

At this point, the benefit of prequential inference becomes even clearer. If we had relied on the classical minimum description length formulation, we would have had to define a prior distribution on all the possible hyperparameters, both for the stochastic blockmodels' partitions and for the configuration models' partitions. It is not clear at all how such a distribution could be defined without introducing bias in the model selection. On the other hand, prequential inference allows us to neglect this issue and let the inference process itself select the hyperparameter which best manages to predict succesive edges based on the previously observed ones.

In practice, we consider two models on $n = 128$ nodes: the stochastic blockmodel $S_1 = (B_1, M_1)$ (and its associated hyperparameter ϕ_{B_1}) defined as

$$B_1 = [\![1, 64]\!], [\![65, 128]\!]$$
$$M_1 = \frac{1}{n^2} \cdot \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$$

and the block configuration model CM defined by

$$B^{out} = [\![1, 96]\!], [\![97, 120]\!], [\![121, 126]\!], [\![127, 128]\!]$$
$$p^{out} = [0.0054; 0.0109; 0.0217; 0.0435]$$
$$B^{in} = ([\![1, 2]\!], [\![3, 8]\!], [\![9, 32]\!], [\![33, 128]\!])$$
$$p^{in} = [0.0435; 0.0217; 0.0109; 0.0054]$$

which corresponds to a hyperparameter $\phi_{B^{out}, B^{in}}$.

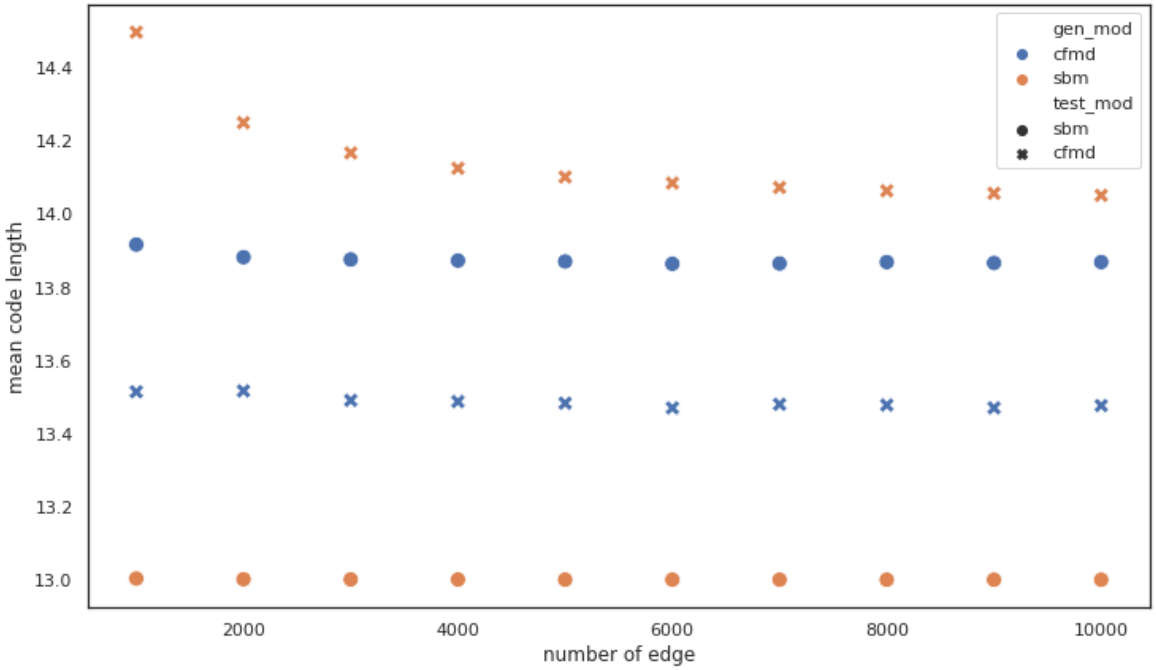


Figure 5.7 – Mean code length of two families of edges sequences, encoded using stochastic blockmodel hyperparameter and configuration model hyperparameter.

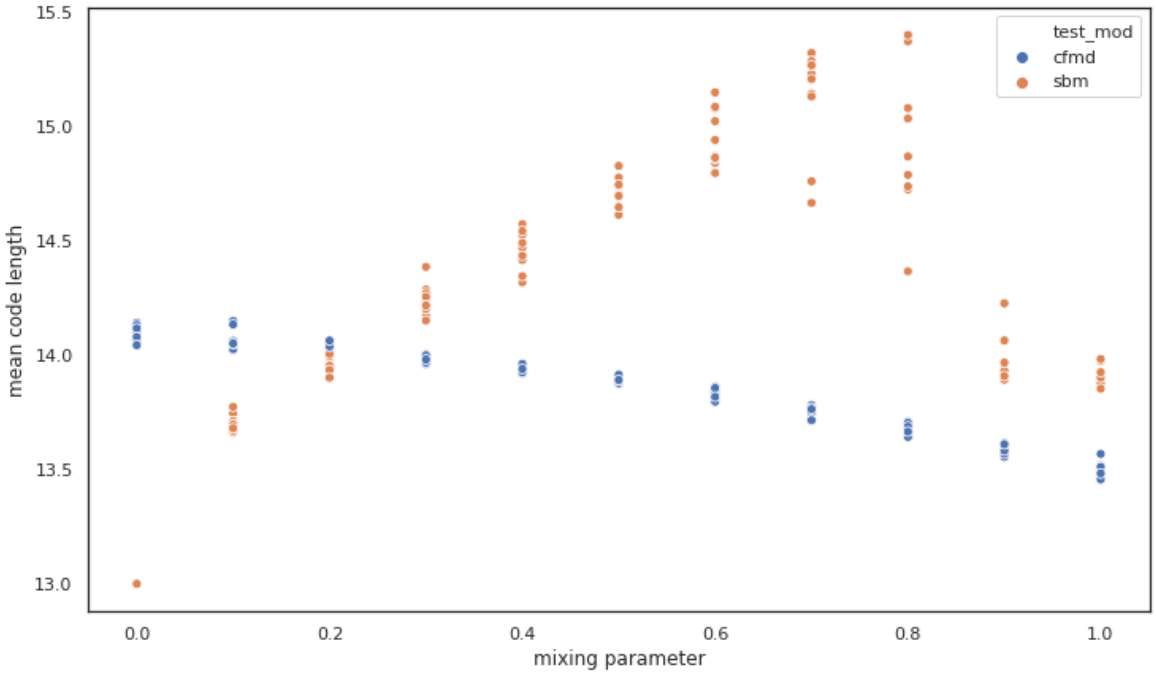


Figure 5.8 – Mean code length against mixing parameter.

For each probability distribution $\mathbb{P}_{S_1}$ and $\mathbb{P}_{CM}$, we generate 10 edge sequences of length 1000 to 10000. Then, for each edge sequence, we compute its mean code length using hyperparameters ϕ_{B_1} and $\phi_{B^{out},B^{in}}$. Results are shown in Figure 5.7.

We observe that, for the sequences of edges which are generated using $\mathbb{P}_{CM}$ (blue dots and crosses), the mean code length is lower when using the configuration model hyperparameter $\phi_{B^{out},B^{in}}$. On the other hand, for the sequences generated using $\mathbb{P}_{S_1}$ (yellow dots and crosses), the mean code length is lower when using the stochastic blockmodel hyperparameter ϕ_{B_1}. Thus, the best compression actually corresponds to the correct hyperparameter.

What is even more interested, is that we can also use sequential edge probability inference to identify the most significant property when the edge distribution is the result of a combination of factors. Continuing with models $\mathbb{P}_{S_1}$ and $\mathbb{P}_C$, let's define the mixed model:

$$\mathbb{P}(\lambda) = \lambda \cdot \mathbb{P}_C + (1 - \lambda) \cdot \mathbb{P}_{S_1}$$

We consider 11 values of λ between 0 and 1, and for each, we generate 10 edge sequences of length 2800. Then, for each edge sequence, we compute its mean code length using $\phi_{B^{out},B^{in}}$ and ϕ_{B_1}. Results are shown in Figure 5.8

We observe that as λ rises from 0 to 1, the mean code length using the block hyperparameter ϕ_{B_1} rises from 13 to 14, with a pick up to 15.5. On the other hand, the mean code length using the configuration structure decreases from a little more than 14 down to 13.5. It shows that the mean code length is able to capture the increasing influence of the block structure and the decreasing influence of the configuration structure in the distribution of edges. When one model clearly dominates the other (*i.e.* $\lambda \leqslant 0.2$ or $\lambda \geqslant 0.8$) the corresponding hyperparameter leads to a better compression.

Conclusion

In conclusion, we have shown how prequential inference can be used in graph model selection, by considering probability distributions on edge sequences rather than static graphs. Describing models of various nature as probability distributions on edge allows to easily compare their performance thanks to minimum description length (or equivalently bayesian inference). Moreover, by introducing additional constraints as hyperparameters, we are able to lower the number of parameters of the model below the number of observations on which inference is performed, which is necessary to avoid overfitting.

We have illustrated how this framework can be used to select the most significant node partition according to information present in edge distribution. Because it relies on statistical inference, it provides a simple way to discriminate automatically between too fine and too coarse partitions with no a priori information.

The main advantage of prequential edge probability inference is that it provides a common formulation of models of different nature in order to compare them. It is thus able, for example, to automatically detect whether the distribution of edges is determined rather by nodes' block membership (block structure) or by their potential to emit or receive edges (configurational structure), even in cases where both structures are mixed.

We believe these results to be a foretaste of the potential of this approach. Because it has firm theoretical grounds, we are convinced that it can provide fruitful applications in many domains where interactions are the results of entangled mechanisms whose effect

on the overall graph topology can only be told apart by rigorous statistical analysis. It therefore provides a reliable criterion which, combined with a methodology to explore the hyperparameter search space, can lead to the automatic selection of the best model for a given graph.

Chapter 6

Conclusion

The study of real-world networks for the last twenty years has revealed their wealth of structure at various scales and in turn has fostered the development of a wide variety of models to explain the emergence of these structures. Yet, this very wealth of structure implies that any modeling attempt which is not firmly grounded on statistical basis risks to mistake random fluctuations for significant patterns or vice-versa while fitting the model. What is more, the existence of various candidate models for an observed network stresses the lack of a principled methodology to compare their relative relevance.

Statistical modeling provides a firm basis to tackle this model evaluation issue. Fitting probability distributions on observations, and comparing their relevance with respect to a set of observations are common tasks in statistical analysis. Formulating graph models as probability distributions allows to mobilize its results to perform rigorous model fitting and selection. However, graphs have peculiarities that must be accounted for in order for these results to be interpretable. In this thesis, we explored different ways of adapting statistical analysis tools to graph model selection.

In chapter 3, we focus on the formulation of a statistical test to evaluate the probability that a candidate model was used to generate an observed network. Our main contribution is to study the structure of the microcanonical ensemble associated with a model not only from a combinatorial point of view (to compute its entropy), but also from a geometric point of view thanks to the normalized edit distance. We show that the obtained graph space's shape is such that the distance to the barycenter and the normalized edit distance expected value concentrate around values which are characterics of the model. As a consequence, we are able to statisticaly test the hypothesis that an observed network was generated by a candidate model. Contrarily to the mere likelihood, this test does not only measure the probability to generate the observed network itself, but also networks which are close to it.

In chapter 4 and 5, we move on to a bayesian inference approach. In chapter 4, we investigate the asymptotic properties of bayesian inference based on the microcanonical ensemble's entropy. We observe that, under certain circumstances, increasing the density of edges within a planted partition generative model makes it not easier but harder to retrieve, even though the number and sizes of blocks are known. We show that this is due to the fact that imposing small blocks in dense regions imposes more constraints on the model and thus lowers entropy independently of the statistical evidence present in the graph, which is a form of overfitting. At a more fundamental level, it is a consequence of the fact that defining the

probability distribution associated with a model on the microcanonical ensemble implies to consider the whole graph as the random variable, therefore performing statistical inference on a single observation.

To overcome this issue, we propose in chapter 5 a reformulation of graph statistical models in terms of probability distribution on the set of edges. This way, even a single graph contains several realization of the random variable. This formulation allows us to use prequential inference, a statistical inference methodology which, while still based on the minimum description length principle, does not require to rely on an ad hoc prior distribution to avoid overfitting. The parameters of the model are updated sequentially as edges are observed one after the other, and at each step it measures the ability of the candidate model to take advantage of the statistical regularities in the first observed edges to predict the next one. This way, it is able to discard both the models which do not have enough degrees of freedom to fit the observations (and thus underfit them) and those which have too much degrees of freedom and fail to predit yet unobserved edges (thus overfitting). We also show how prequential inference can be used to compare models that do not share the same parameter space, namely the configuration model and the stochastic block model. As it does not rely on a prior distribution on the parameter space to avoid overfitting, it removes the need to assign beforehand a probability to each possible parameter set, which is hard to do without introducing bias in the model selection.

Perspectives and future work

Statistical graph model selection is a recent topic of research, especially if we consider not only the selection of the right set of parameters but also the comparison of models whose parameters do not belong to the same parameter space. As always, these contributions raise as much new questions than they answer. Graph space geometric structure and properties under other metric would deserve to be investigated. The edit distance we considered is simple, but it is not the most adapted to the study of graphs as it neglects the topological role of edges. Graph designed metrics such as DeltaCon, the perturbation-resistance metric, or others could provide more meaningful tests. Apart from the metric, it would also be interesting to investigate the change to the geometric structure of the graph space induced when restricting the model to simple and undirected graphs, which are more used in practice.

Considering edge statistical models, we believe our theoretical results to give some hints about the nature of overfit and underfit in graph model selection. Edge statistical model and prequential inference show that a model selection procedure relying only on the statistical evidence present in the data is possible. External specification such as a prior distribution, which are necessary to other procedure to avoid overfitting, induce the risk to introduce bias in the selection procedure. This is even truer as the models to compare are based on parameters belonging to different parameter space (such as a stochastic blockmodel and a configuration model). Removing these specifications thus opens the way to unbiased model comparison.

In practice, we tested the procedure almost exclusively on small synthetic graphs. Although the results are promising, it is clearly not sufficient to state definitive conclusions about it. One important obstacle is the computational complexity of prequential inference, as it requires to refit the model several times for a single network. Testing the procedure on

real networks or exploring large parameter spaces like the node partition space for stochastic blockmodel inference would imply to drastically reduce this computation time.

It would also be interesting to investigate the performance of this methodology on temporal graphs which are naturaly described by edge sequences. In such a case, as the order of edges is known, the assumption of independence between edges' generation could probably be removed, allowing to consider a wider class of models. In particular, the Barabasi-Albert model of preferential attachment could be evaluated using this methodology.

Bibliography

Emmanuel Abbe and Colin Sandon. Community detection in general stochastic block models: fundamental limits and efficient recovery algorithms. *arXiv:1503.00609 [cs, math]*, March 2015. URL http://arxiv.org/abs/1503.00609. arXiv: 1503.00609.

Lada A Adamic, Rajan M Lukose, Amit R Puniyani, and Bernardo A Huberman. Search in power-law networks. *Physical review E*, 64(4):046135, 2001.

Edoardo Maria Airoldi, David M Blei, Stephen E Fienberg, and Eric P Xing. Mixed membership stochastic blockmodels. *Journal of machine learning research*, 2008.

Réka Albert and Albert-László Barabási. Statistical mechanics of complex networks. *Reviews of modern physics*, 74(1):47, 2002.

Réka Albert, Hawoong Jeong, and Albert-László Barabási. Error and attack tolerance of complex networks. *nature*, 406(6794):378–382, 2000.

David J. Aldous. Representations for partially exchangeable arrays of random variables. *Journal of Multivariate Analysis*, 11(4):581–598, December 1981. ISSN 0047259X. doi: 10.1016/0047-259X(81)90099-3. URL https://linkinghub.elsevier.com/retrieve/pii/0047259X81900993.

LAN Amara, Antonio Scala, Marc Barthelemy, and H Eugene Stanley. Classes of small-world networks. In *The Structure and Dynamics of Networks*, pages 207–210. Princeton University Press, 2011.

Carolyn J Anderson, Stanley Wasserman, and Bradley Crouch. A p* primer: Logit models for social networks. *Social networks*, 21(1):37–66, 1999.

Kenneth Appel and Wolfgang Haken. *Every Planar Map is Four Colorable*, volume 98 of *Contemporary Mathematics*. American Mathematical Society, Providence, Rhode Island, 1989. ISBN 978-0-8218-5103-6 978-0-8218-7686-2. doi: 10.1090/conm/098. URL http://www.ams.org/conm/098/.

Andrii Arman, Pu Gao, and Nicholas Wormald. Fast uniform generation of random graphs with given degree sequences. *arXiv:1905.03446 [cs, math]*, May 2019. URL http://arxiv.org/abs/1905.03446. arXiv: 1905.03446.

James P. Bagrow and Erik M. Bollt. An information-theoretic, all-scales approach to comparing networks. *Applied Network Science*, 4(1):45, December 2019. ISSN 2364-8228. doi: 10.1007/s41109-019-0156-x. URL https://appliednetsci.springeropen.com/articles/10.1007/s41109-019-0156-x.

Albert-Laszlo Barabasi and Reka Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, October 1999. ISSN 00368075, 10959203. doi: 10.1126/science.286.5439.509. URL http://arxiv.org/abs/cond-mat/9910332. arXiv: cond-mat/9910332.

Hugo Barbosa, Marc Barthelemy, Gourab Ghoshal, Charlotte R James, Maxime Lenormand, Thomas Louail, Ronaldo Menezes, José J Ramasco, Filippo Simini, and Marcello Tomasini. Human mobility: Models and applications. *Physics Reports*, 734:1–74, 2018.

Ole Barndorff-Nielsen. *Information and exponential families: in statistical theory*. John Wiley & Sons, 2014.

Andrew Barron, Jorma Rissanen, and Bin Yu. The minimum description length principle in coding and modeling. *IEEE transactions on information theory*, 44(6):2743–2760, 1998.

Marc Barthelemy. Spatial Networks. *Physics Reports*, 499(1-3):1–101, February 2011. ISSN 03701573. doi: 10.1016/j.physrep.2010.11.002. URL http://arxiv.org/abs/1010.0302. arXiv: 1010.0302.

Edward A. Bender. The asymptotic number of non-negative integer matrices with given row and column sums. *Discrete Mathematics*, 10(2):217–223, 1974. ISSN 0012365X. doi: 10.1016/0012-365X(74)90118-6. URL https://linkinghub.elsevier.com/retrieve/pii/0012365X74901186.

Edward A Bender and E.Rodney Canfield. The asymptotic number of labeled graphs with given degree sequences. *Journal of Combinatorial Theory, Series A*, 24(3):296–307, May 1978. ISSN 00973165. doi: 10.1016/0097-3165(78)90059-6. URL http://linkinghub.elsevier.com/retrieve/pii/0097316578900596.

Claude Berge. *Théorie des graphes et ses applications*. 1958.

Kunal Bhattacharya, Gautam Mukherjee, Jari Saramäki, Kimmo Kaski, and Subhrangshu S Manna. The international trade network: weighted network analysis and modelling. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(02):P02002, 2008.

Ginestra Bianconi. The entropy of randomized network ensembles. *EPL (Europhysics Letters)*, 81(2):28005, January 2008. ISSN 0295-5075, 1286-4854. doi: 10.1209/0295-5075/81/28005. URL http://arxiv.org/abs/0708.0153. arXiv: 0708.0153.

Ginestra Bianconi. The entropy of network ensembles. *Physical Review E*, 79(3), March 2009. ISSN 1539-3755, 1550-2376. doi: 10.1103/PhysRevE.79.036114. URL http://arxiv.org/abs/0802.2888. arXiv: 0802.2888.

Béla Bollobás and Oliver Riordan. Robustness and vulnerability of scale-free random graphs. *Internet Mathematics*, 1(1):1–35, 2004.

Béla Bollobás and Oliver Riordan. Sparse graphs: metrics and random models. *Random Structures & Algorithms*, 39(1):1–38, 2011.

Bela Bollobas, Bela Bollobás, Oliver Riordan, and O. Riordan. *Percolation*. Cambridge University Press, September 2006. ISBN 978-0-521-87232-4. Google-Books-ID: PMOAA2SacuYC.

Anna D Broido and Aaron Clauset. Scale-free networks are rare. *Nature communications*, 10 (1):1–10, 2019.

Hongyun Cai, Vincent W. Zheng, and Kevin Chen-Chuan Chang. A Comprehensive Survey of Graph Embedding: Problems, Techniques and Applications. *arXiv:1709.07604 [cs]*, February 2018. URL http://arxiv.org/abs/1709.07604. arXiv: 1709.07604.

A. Cayley Cayley, Arthur. XXVIII. On the theory of the analytical forms called trees. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 13 (85):172–176, March 1857. ISSN 1941-5982. doi: 10.1080/14786445708642275. URL https://doi.org/10.1080/14786445708642275. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/14786445708642275.

Professor Cayley. On the Colouring of Maps. *Proceedings of the Royal Geographical Society and Monthly Record of Geography*, 1(4):259, April 1879. ISSN 0266626X. doi: 10.2307/1799998. URL https://www.jstor.org/stable/1799998?origin=crossref.

Remy Cazabet, Pierre Borgnat, and Pablo Jensen. Enhancing Space-Aware Community Detection Using Degree Constrained Spatial Null Model. In Bruno Gonçalves, Ronaldo Menezes, Roberta Sinatra, and Vinko Zlatic, editors, *Complex Networks VIII*, pages 47–55. Springer International Publishing, Cham, 2017. ISBN 978-3-319-54240-9 978-3-319-54241-6. doi: 10.1007/978-3-319-54241-6_4. URL http://link.springer.com/10.1007/978-3-319-54241-6_4.

Andressa Cerqueira, Daniel Fraiman, Claudia D Vargas, and Florencia Leonardi. A test of hypotheses for random graph distributions built from eeg data. *IEEE Transactions on Network Science and Engineering*, 4(2):75–82, 2017.

F. Chung and L. Lu. The average distances in random graphs with given expected degrees. *Proceedings of the National Academy of Sciences*, 99(25):15879–15882, December 2002. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.252631999. URL http://www.pnas.org/cgi/doi/10.1073/pnas.252631999.

Giulio Cimini, Tiziano Squartini, Fabio Saracco, Diego Garlaschelli, Andrea Gabrielli, and Guido Caldarelli. The Statistical Physics of Real-World Networks. *arXiv:1810.05095 [cond-mat, physics:physics]*, October 2018. URL http://arxiv.org/abs/1810.05095. arXiv: 1810.05095.

Aaron Clauset, Cosma Rohilla Shalizi, and Mark EJ Newman. Power-law distributions in empirical data. *SIAM review*, 51(4):661–703, 2009.

Reuven Cohen, Keren Erez, Shlomo Havlinl, Mark Newman, Albert-László Barabási, Duncan J Watts, et al. Resilience of the internet to random breakdowns. In *The Structure and Dynamics of Networks*, pages 507–509. Princeton University Press, 2011.

Edward F Connor and Daniel Simberloff. The assembly of species communities: chance or competition? *Ecology*, 60(6):1132–1140, 1979.

Justin P. Coon, Carl P. Dettmann, and Orestis Georgiou. Entropy of Spatial Network Ensembles. *Physical Review E*, 97(4):042319, April 2018. ISSN 2470-0045, 2470-0053. doi: 10.1103/PhysRevE.97.042319. URL http://arxiv.org/abs/1707.01901. arXiv: 1707.01901.

Harry Crane and Walter Dempsey. Edge exchangeables models for network data. page 35, 2016.

Harry Crane and Walter Dempsey. Edge exchangeable models for interaction networks. *Journal of the American Statistical Association*, 113(523):1311–1326, 2018. ISSN 0162-1459. doi: 10.1080/01621459.2017.1341413. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6241523/.

A Philip Dawid. Present position and potential developments: Some personal views statistical theory the prequential approach. *Journal of the Royal Statistical Society: Series A (General)*, 147(2):278–290, 1984.

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*, 84(6):066106, December 2011a. ISSN 1539-3755, 1550-2376. doi: 10.1103/PhysRevE.84.066106. URL http://arxiv.org/abs/1109.3041. arXiv: 1109.3041.

Aurelien Decelle, Florent Krzakala, Cristopher Moore, and Lenka Zdeborová. Phase transition in the detection of modules in sparse networks. *Physical Review Letters*, 107(6):065701, August 2011b. ISSN 0031-9007, 1079-7114. doi: 10.1103/PhysRevLett.107.065701. URL http://arxiv.org/abs/1102.1182. arXiv: 1102.1182.

Persi Diaconis and Svante Janson. Graph limits and exchangeable random graphs. *arXiv:0712.2749 [math]*, December 2007. URL http://arxiv.org/abs/0712.2749. arXiv: 0712.2749.

Jie Ding, Vahid Tarokh, and Yuhong Yang. Model Selection Techniques – An Overview. *arXiv:1810.09583 [physics, stat]*, October 2018. doi: 10.1109/MSP.2018.2867638. URL http://arxiv.org/abs/1810.09583. arXiv: 1810.09583.

Louis Duvivier, Céline Robardet, and Rémy Cazabet. Minimum entropy stochastic block models neglect edge distribution heterogeneity. In *International Conference on Complex Networks and Their Applications*, pages 545–555. Springer, 2019.

Louis Duvivier, Rémy Cazabet, and Céline Robardet. Edge based stochastic block model statistical inference. In *International Conference on Complex Networks and Their Applications*, pages 462–473. Springer, 2020.

Louis Duvivier, Rémy Cazabet, and Céline Robardet. Graph space: using both geometric and probabilistic structure to evaluate statistical graph models. *arXiv preprint arXiv:2106.13587*, 2021a.

Louis Duvivier, Rémy Cazabet, and Céline Robardet. Graph model selection by edge probability sequential inference. *arXiv preprint arXiv:2106.13579*, 2021b.

Victor M Eguiluz, Dante R Chialvo, Guillermo A Cecchi, Marwan Baliki, and A Vania Apkarian. Scale-free brain functional networks. *Physical review letters*, 94(1):018102, 2005.

Erdös, P. and Rényi, A. On random graphs. 1959. URL https://snap.stanford.edu/class/cs224w-readings/erdos59random.pdf.

Erdös, P. and Rényi, A. On the evolution of random graphs. 1960. URL http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.348.530&rep=rep1&type=pdf.

Leonhard Euler. Solutio problematis ad geometriam situs pertinentis. page 15, 1741.

P. Expert, T. S. Evans, V. D. Blondel, and R. Lambiotte. Uncovering space-independent communities in spatial networks. *Proceedings of the National Academy of Sciences*, 108(19): 7663–7668, May 2011. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1018962108. URL http://www.pnas.org/cgi/doi/10.1073/pnas.1018962108.

Michalis Faloutsos, Petros Faloutsos, and Christos Faloutsos. On power-law relationships of the internet topology. In *The Structure and Dynamics of Networks*, pages 195–206. Princeton University Press, 2011.

Bruno De Finetti. La prévision : ses lois logiques, ses sources subjectives. page 69, 1937.

S. Fortunato and M. Barthelemy. Resolution limit in community detection. *Proceedings of the National Academy of Sciences*, 104(1):36–41, January 2007. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.0605965104. URL http://www.pnas.org/cgi/doi/10.1073/pnas.0605965104.

Santo Fortunato and Darko Hric. Community detection in networks: A user guide. *Physics Reports*, 659:1–44, November 2016. ISSN 03701573. doi: 10.1016/j.physrep.2016.09.002. URL http://arxiv.org/abs/1608.00163. arXiv: 1608.00163.

Daniel Fraiman and Ricardo Fraiman. An anova approach for statistical comparisons of brain networks. *Scientific reports*, 8(1):1–14, 2018.

Franck O. and Strauss D. Markov Graphs. *Journal of the American Statistical Association*, 1986. URL http://www.uvm.edu/pdodds/files/papers/others/everything/frank1986a.pdf.

Andre Fujita, Eduardo Silva Lira, Suzana de Siqueira Santos, Silvia Yumi Bando, Gabriela Eleuterio Soares, and Daniel Yasumasa Takahashi. A semi-parametric statistical test to compare complex networks. *Journal of Complex Networks*, 8(2):cnz028, April 2020. ISSN 2051-1329. doi: 10.1093/comnet/cnz028. URL https://academic.oup.com/comnet/article/doi/10.1093/comnet/cnz028/5543003.

Amir Ghasemian, Homa Hosseinmardi, and Aaron Clauset. Evaluating Overfit and Underfit in Models of Network Community Structure. *arXiv:1802.10582 [physics, q-bio, stat]*, April 2019. URL http://arxiv.org/abs/1802.10582. arXiv: 1802.10582.

Debarghya Ghoshdastidar, Maurilio Gutzeit, Alexandra Carpentier, and Ulrike von Luxburg. Two-Sample Tests for Large Random Graphs Using Network Statistics. *arXiv:1705.06168 [stat]*, May 2017. URL http://arxiv.org/abs/1705.06168. arXiv: 1705.06168.

Gilbert, E.N. Random plane networks. 1959.

Giona Casiraghi and Vahan Nanumyan. Generalised hypergeometric ensembles of random graphs: the configuration model as an urn problem. *arXiv:1810.06495 [physics]*, October 2018. URL http://arxiv.org/abs/1810.06495. arXiv: 1810.06495.

M. Girvan and M. E. J. Newman. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, 99(12):7821–7826, June 2002. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.122653799. URL http://www.pnas.org/cgi/doi/10.1073/pnas.122653799.

Anna Goldenberg, Alice X. Zheng, Stephen E. Fienberg, and Edoardo M. Airoldi. A survey of statistical network models. *arXiv:0912.5410 [physics, q-bio, stat]*, December 2009. URL http://arxiv.org/abs/0912.5410. arXiv: 0912.5410.

Palash Goyal and Emilio Ferrara. Graph Embedding Techniques, Applications, and Performance: A Survey. *Knowledge-Based Systems*, 151:78–94, July 2018. ISSN 09507051. doi: 10.1016/j.knosys.2018.03.022. URL http://arxiv.org/abs/1705.02801. arXiv: 1705.02801.

Peter Grunwald. A tutorial introduction to the minimum description length principle. *arXiv:math/0406077*, June 2004. URL http://arxiv.org/abs/math/0406077. arXiv: math/0406077.

Peter Grünwald and Teemu Roos. Minimum description length revisited. *International journal of mathematics for industry*, 11(01):1930001, 2019.

Roger Guimera, Marta Sales-Pardo, and Luis A. N. Amaral. Modularity from Fluctuations in Random Graphs and Complex Networks. *Physical Review E*, 70(2):025101, August 2004. ISSN 1539-3755, 1550-2376. doi: 10.1103/PhysRevE.70.025101. URL http://arxiv.org/abs/cond-mat/0403660. arXiv: cond-mat/0403660.

S Louis Hakimi. On realizability of a set of integers as degrees of the vertices of a linear graph. i. *Journal of the Society for Industrial and Applied Mathematics*, 10(3):496–506, 1962.

Mark S Handcock, Garry Robins, Tom Snijders, Jim Moody, and Julian Besag. Assessing degeneracy in statistical models of social networks. Technical report, Citeseer, 2003.

Charles W. Harper jr. Groupings by locality in community ecology and paleoecology: tests of significance. *Lethaia*, 11(3):251–257, 1978.

Leland H Hartwell, John J Hopfield, Stanislas Leibler, and Andrew W Murray. From molecular to modular cell biology. *Nature*, 402(6761):C47–C52, 1999.

M. B. Hastings. Community Detection as an Inference Problem. *Physical Review E*, 74(3): 035102, September 2006. ISSN 1539-3755, 1550-2376. doi: 10.1103/PhysRevE.74.035102. URL http://arxiv.org/abs/cond-mat/0604429. arXiv: cond-mat/0604429.

Václav Havel. A remark on the existence of finite graphs. *Casopis Pest. Mat.*, 80:477–480, 1955.

Paul W Holland and Samuel Leinhardt. Local structure in social networks. *Sociological methodology*, 7:1–45, 1976.

Paul W. Holland and Samuel Leinhardt. An Exponential Family of Probability Distributions for Directed Graphs. *Journal of the American Statistical Association*, 76(373):33–50, March 1981. ISSN 0162-1459. doi: 10.1080/01621459.1981.10477598. URL https://www.tandfonline.com/doi/abs/10.1080/01621459.1981.10477598. Publisher: Taylor & Francis _eprint: https://www.tandfonline.com/doi/pdf/10.1080/01621459.1981.10477598.

Paul W. Holland, Kathryn Blackmond Laskey, and Samuel Leinhardt. Stochastic blockmodels: First steps. *Social Networks*, 5(2):109–137, June 1983. ISSN 03788733. doi: 10.1016/0378-8733(83)90021-7. URL https://linkinghub.elsevier.com/retrieve/pii/0378873383900217.

P. Holme, M. Huss, and H. Jeong. Subnetwork hierarchies of biochemical pathways. *Bioinformatics*, 19(4):532–538, March 2003. ISSN 1367-4803, 1460-2059. doi: 10.1093/bioinformatics/btg033. URL https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btg033.

Petter Holme. Rare and everywhere: Perspectives on scale-free networks. *Nature communications*, 10(1):1–3, 2019.

Douglas N Hoover. Relations on probability spaces and arrays of random variables. *Preprint, Institute for Advanced Study, Princeton, NJ*, 2:275, 1979.

Dandan Hu, Peter Ronhovde, and Zohar Nussinov. Phase transitions in random Potts systems and the community detection problem: spin-glass type and dynamic perspectives. *Philosophical Magazine*, 92(4):406–445, February 2012. ISSN 1478-6435, 1478-6443. doi: 10.1080/14786435.2011.616547. URL http://arxiv.org/abs/1008.2699. arXiv: 1008.2699.

E. T. Jaynes. Information Theory and Statistical Mechanics. *Physical Review*, 106(4):620–630, May 1957. ISSN 0031-899X. doi: 10.1103/PhysRev.106.620. URL https://link.aps.org/doi/10.1103/PhysRev.106.620.

Woo-Sung Jung, Fengzhong Wang, and H Eugene Stanley. Gravity model in the korean highway. *EPL (Europhysics Letters)*, 81(4):48005, 2008.

Micha Karoński and Andrzej Ruciński. The origins of the theory of random graphs. In *The mathematics of Paul Erdös I*, pages 311–336. Springer, 1997.

Brian Karrer and M. E. J. Newman. Stochastic blockmodels and community structure in networks. August 2010. doi: 10.1103/PhysRevE.83.016107. URL https://arxiv.org/abs/1008.3926.

Jon M Kleinberg, Ravi Kumar, Prabhakar Raghavan, Sridhar Rajagopalan, and Andrew S Tomkins. The web as a graph: Measurements, models, and methods. In *International Computing and Combinatorics Conference*, pages 1–17. Springer, 1999.

A. N. Kolmogorov. Three approaches to the quantitative definition of information. *International Journal of Computer Mathematics*, 2(1-4):157–168, January 1968. ISSN 0020-7160, 1029-0265. doi: 10.1080/00207166808803030. URL `http://www.tandfonline.com/doi/abs/10.1080/00207166808803030`.

Danai Koutra, Joshua T. Vogelstein, and Christos Faloutsos. DELTACON: A Principled Massive-Graph Similarity Function. *arXiv:1304.4657 [physics]*, April 2013. URL `http://arxiv.org/abs/1304.4657`. arXiv: 1304.4657.

Renaud Lambiotte, Vincent D Blondel, Cristobald De Kerchove, Etienne Huens, Christophe Prieur, Zbigniew Smoreda, and Paul Van Dooren. Geographical dispersal of mobile communication networks. *Physica A: Statistical Mechanics and its Applications*, 387(21):5317–5325, 2008.

Andrea Lancichinetti and Santo Fortunato. Limits of modularity maximization in community detection. *Physical Review E*, 84(6):066122, December 2011. ISSN 1539-3755, 1550-2376. doi: 10.1103/PhysRevE.84.066122. URL `http://arxiv.org/abs/1107.1155`. arXiv: 1107.1155.

Vito Latora and Massimo Marchiori. Is the boston subway a small-world network? *Physica A: Statistical Mechanics and its Applications*, 314(1-4):109–113, 2002.

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.

Moshe Levy. Scale-free human migration and the geography of social networks. *Physica A: Statistical Mechanics and its Applications*, 389(21):4913–4917, 2010.

Xiao Liang, Jichang Zhao, Li Dong, and Ke Xu. Unraveling the origin of exponential law in intra-urban human mobility. *Scientific reports*, 3(1):1–7, 2013.

László Lovász and Balázs Szegedy. Limits of dense graph sequences. *Journal of Combinatorial Theory, Series B*, 96(6):933–957, 2006.

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.

Vince Lyzinski, Minh Tang, Avanti Athreya, Youngser Park, and Carey E. Priebe. Community Detection and Classification in Hierarchical Stochastic Blockmodels. *arXiv:1503.02115 [stat]*, August 2016. URL `http://arxiv.org/abs/1503.02115`. arXiv: 1503.02115.

Sergei Maslov and Kim Sneppen. Specificity and stability in topology of protein networks. *Science*, 296(5569):910–913, 2002.

A Paolo Masucci, Joan Serras, Anders Johansson, and Michael Batty. Gravity versus radiation models: On the importance of scale and heterogeneity in commuting flows. *Physical Review E*, 88(2):022812, 2013.

Stanley Milgram. The small world problem. *Psychology today*, 2(1):60–67, 1967.

R. Milo, N. Kashtan, S. Itzkovitz, M. E. J. Newman, and U. Alon. On the uniform generation of random graphs with prescribed degree sequences. *arXiv:cond-mat/0312028*, December 2003. URL http://arxiv.org/abs/cond-mat/0312028. arXiv: cond-mat/0312028.

Michael Molloy and Bruce Reed. The size of the giant component of a random graph with a given degree sequence. *Combinatorics probability and computing*, 7(3):295–305, 1998.

Michael Molloy, Bruce Reed, Mark Newman, Albert-László Barabási, and Duncan J Watts. A critical point for random graphs with a given degree sequence. In *The Structure and Dynamics of Networks*, pages 240–258. Princeton University Press, 2011.

Nathan D. Monnig and Francois G. Meyer. The Resistance Perturbation Distance: A Metric for the Analysis of Dynamic Networks. *arXiv:1605.01091 [physics]*, May 2016. URL http://arxiv.org/abs/1605.01091. arXiv: 1605.01091.

M. E. J. Newman. The structure and function of complex networks. *SIAM Review*, 45(2): 167–256, January 2003. ISSN 0036-1445, 1095-7200. doi: 10.1137/S003614450342480. URL http://arxiv.org/abs/cond-mat/0303516. arXiv: cond-mat/0303516.

M. E. J. Newman. Fast algorithm for detecting community structure in networks. *Physical Review E*, 69(6):066133, June 2004. ISSN 1539-3755, 1550-2376. doi: 10.1103/PhysRevE.69.066133. URL http://arxiv.org/abs/cond-mat/0309508. arXiv: cond-mat/0309508.

M. E. J. Newman. Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, 103(23):8577–8582, June 2006. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.0601602103. URL http://www.pnas.org/cgi/doi/10.1073/pnas.0601602103.

M. E. J. Newman and M. Girvan. Finding and evaluating community structure in networks. *Physical Review E*, 69(2):026113, February 2004. ISSN 1539-3755, 1550-2376. doi: 10.1103/PhysRevE.69.026113. URL http://arxiv.org/abs/cond-mat/0308217. arXiv: cond-mat/0308217.

M. E. J. Newman and E. A. Leicht. Mixture models and exploratory analysis in networks. *Proceedings of the National Academy of Sciences*, 104(23):9564–9569, June 2007. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.0610537104. URL http://www.pnas.org/cgi/doi/10.1073/pnas.0610537104.

M. E. J. Newman, S. H. Strogatz, and D. J. Watts. Random graphs with arbitrary degree distributions and their applications. *Physical Review E*, 64(2), July 2001. ISSN 1063-651X, 1095-3787. doi: 10.1103/PhysRevE.64.026118. URL https://link.aps.org/doi/10.1103/PhysRevE.64.026118.

Christine Leigh Myers Nickel. Random dot product graphs, a model for social networks. page 259, 2008.

Peter Orbanz and Daniel M. Roy. Bayesian Models of Graphs, Arrays and Other Exchangeable Random Structures. *arXiv:1312.7857 [math, stat]*, February 2015. URL http://arxiv.org/abs/1312.7857. arXiv: 1312.7857.

Juyong Park and M. E. J. Newman. The statistical mechanics of networks. *Physical Review E*, 70(6), December 2004a. ISSN 1539-3755, 1550-2376. doi: 10.1103/PhysRevE.70.066117. URL http://arxiv.org/abs/cond-mat/0405566. arXiv: cond-mat/0405566.

Juyong Park and Mark EJ Newman. Solution of the two-star model of a network. *Physical Review E*, 70(6):066146, 2004b.

Romualdo Pastor-Satorras and Alessandro Vespignani. Epidemic spreading in scale-free networks. *Physical review letters*, 86(14):3200, 2001.

Tiago P. Peixoto. Entropy of stochastic blockmodel ensembles. *Physical Review E*, 85(5), May 2012. ISSN 1539-3755, 1550-2376. doi: 10.1103/PhysRevE.85.056122. URL http://arxiv.org/abs/1112.6028. arXiv: 1112.6028.

Tiago P. Peixoto. Parsimonious module inference in large networks. *Physical Review Letters*, 110(14):148701, April 2013. ISSN 0031-9007, 1079-7114. doi: 10.1103/PhysRevLett.110.148701. URL http://arxiv.org/abs/1212.4794. arXiv: 1212.4794.

Tiago P. Peixoto. Nonparametric weighted stochastic block models. *Physical Review E*, 97 (1), January 2018. ISSN 2470-0045, 2470-0053. doi: 10.1103/PhysRevE.97.012306. URL http://arxiv.org/abs/1708.01432. arXiv: 1708.01432.

Tiago P Peixoto. Bayesian stochastic blockmodeling. *Advances in network clustering and blockmodeling*, pages 289–332, 2019.

Georg Polya and R. C. Read. *Combinatorial Enumeration of Groups, Graphs, and Chemical Compounds*. Springer Science & Business Media, December 2012. ISBN 978-1-4612-4664-0. Google-Books-ID: QyjUBwAAQBAJ.

DJ de S Price. *Networks of scientific papers*. Princeton University Press, 2011.

Liudmila Prokhorenkova and Alexey Tikhonov. Community Detection through Likelihood Optimization: In Search of a Sound Model. In *The World Wide Web Conference on - WWW '19*, pages 1498–1508, San Francisco, CA, USA, 2019. ACM Press. ISBN 978-1-4503-6674-8. doi: 10.1145/3308558.3313429. URL http://dl.acm.org/citation.cfm?doid=3308558.3313429.

G. Pólya. Kombinatorische Anzahlbestimmungen für Gruppen, Graphen und chemische Verbindungen. *Acta Mathematica*, 68(0):145–254, 1937. ISSN 0001-5962. doi: 10.1007/BF02546665. URL http://projecteuclid.org/euclid.acta/1485888172.

Anatol Rapoport and William J Horvath. A study of a large sociogram. *Behavioral science*, 6 (4):279–291, 1961.

Joerg Reichardt and Stefan Bornholdt. When are networks truly modular? *Physica D: Nonlinear Phenomena*, 224(1-2):20–26, December 2006. ISSN 01672789. doi: 10.1016/j.physd.2006.09.009. URL http://arxiv.org/abs/cond-mat/0606220. arXiv: cond-mat/0606220.

Mikail Rubinov and Olaf Sporns. Complex network measures of brain connectivity: Uses and interpretations. *NeuroImage*, 52(3):1059–1069, September 2010. ISSN 10538119. doi: 10.1016/j.neuroimage.2009.10.003. URL https://linkinghub.elsevier.com/retrieve/pii/S105381190901074X.

Samuel Franklin Sampson. *A novitiate in a period of change: An experimental and case study of social relationships*. Cornell University, 1968.

Lisa Sattenspiel and Carl P. Simon. The spread and persistence of infectious diseases in structured populations. *Mathematical Biosciences*, 90(1-2):341–366, July 1988. ISSN 00255564. doi: 10.1016/0025-5564(88)90074-0. URL https://linkinghub.elsevier.com/retrieve/pii/0025556488900740.

Per O Seglen. The skewness of science. *Journal of the American society for information science*, 43(9):628–638, 1992.

C. E. Shannon. A Mathematical Theory of Communication. *Bell System Technical Journal*, 27(3):379–423, July 1948. ISSN 00058580. doi: 10.1002/j.1538-7305.1948.tb01338.x. URL https://ieeexplore.ieee.org/document/6773024.

Filippo Simini, Marta C. González, Amos Maritan, and Albert-László Barabási. A universal model for mobility and migration patterns. *Nature*, 484(7392):96–100, February 2012. ISSN 0028-0836, 1476-4687. doi: 10.1038/nature10856. URL http://arxiv.org/abs/1111.0586. arXiv: 1111.0586.

Richard E Strauss. Statistical significance of species clusters in association analysis. *Ecology*, 63(3):634–639, 1982.

Daniel Yasumasa Takahashi, João Ricardo Sato, Carlos Eduardo Ferreira, and André Fujita. Discriminating Different Classes of Biological Networks by Analyzing the Graphs Spectra Distribution. *PLoS ONE*, 7(12):e49949, December 2012. ISSN 1932-6203. doi: 10.1371/journal.pone.0049949. URL https://dx.plos.org/10.1371/journal.pone.0049949.

Minh Tang, Avanti Athreya, Daniel L Sussman, Vince Lyzinski, and Carey E Priebe. A nonparametric two-sample hypothesis testing problem for random dot product graphs. *arXiv preprint arXiv:1409.2344*, 2014.

Minh Tang, Avanti Athreya, Daniel L. Sussman, Vince Lyzinski, and Carey E. Priebe. A nonparametric two-sample hypothesis testing problem for random dot product graphs. *arXiv:1409.2344 [math, stat]*, November 2015. URL http://arxiv.org/abs/1409.2344. arXiv: 1409.2344.

Minh Tang, Avanti Athreya, Daniel L Sussman, Vince Lyzinski, Youngser Park, and Carey E Priebe. A semiparametric two-sample hypothesis testing problem for random graphs. *Journal of Computational and Graphical Statistics*, 26(2):344–354, 2017.

Marijtje AJ Van Duijn, Krista J Gile, and Mark S Handcock. A framework for the comparison of maximum pseudo-likelihood and maximum likelihood estimation of exponential family random graph models. *Social networks*, 31(1):52–62, 2009.

Stanley Wasserman and Philippa Pattison. Logit models and logistic regressions for social networks: I. an introduction to markov graphs andp. *Psychometrika*, 61(3):401–425, 1996.

Duncan J Watts and Steven H Strogatz. Collective dynamics of 'small-world' networks. 393: 3, 1998.

Peter Wills and Francois G. Meyer. Metrics for Graph Comparison: A Practitioner's Guide. *arXiv:1904.07414 [physics, q-bio, stat]*, April 2019. URL `http://arxiv.org/abs/1904.07414`. arXiv: 1904.07414.

Wayne W. Zachary. An Information Flow Model for Conflict and Fission in Small Groups. *Journal of Anthropological Research*, 33(4):452–473, 1977. URL `http://www.jstor.org/stable/3629752`.

Christian Zingg, Giona Casiraghi, Giacomo Vaccario, and Frank Schweitzer. What is the Entropy of a Social Organization? *arXiv:1905.09772 [physics]*, May 2019. URL `http://arxiv.org/abs/1905.09772`. arXiv: 1905.09772.

Chapter 7

Appendix

Graph space

Barycenter graph weight of various statistical models

We have defined the barycenter of a graph model as

$$G_M = \sum_{H \in \Omega_M} \mathbb{P}(H) \cdot H$$

Which means that

$$\begin{aligned} \forall (i,j) \in V^2, W_{G_M}(i,j) &= \sum_{H \in \Omega_M} \mathbb{P}(H) \times W_H(i,j) \\ &= \mathbb{E}[W_H(i,j)] \end{aligned}$$

Let's illustrate how this can be computed for some classical models.

Erdös-Rényi model The simplest graph model is the Erdös-Rényi model for random graphs. It's associated microcanonical ensemble can be defined as:

$$\Omega_{ER(n,m)} = \{H = (V,E) \mid |V| = n \wedge |E| = m\}$$

Let's recall that for the sake of simplicity, we chose to consider multigraphs with self loops. Thus, the computation of $\mathbb{E}[W_H(i,j)]$ is particularly simple. Indeed, if for each pair of node $(i,j) \in V^2$ and each $k \in [1,m]$ we define the random variable $X_{i,j,k}$ which is equal to 1 if the k^{th} edge is $i \to j$ and to 0 else, then we have that $W_H(i,j) = \sum_{k=1}^{m} X_{i,j,k}$. It is a sum of independent Bernouillis' random variable so it follows a binomial law of parameters m and $\frac{1}{n^2}$, and thus

$$W_{G_{ER(n,m)}}(i,j) = \mathop{\mathbb{E}}_{\Omega_{ER}}[W_H(i,j)] = \frac{m}{n^2} \tag{7.1}$$

Configuration Model For the configuration model, all graphs in the microcanonical ensemble must have the same degree distribution. Let's consider the directed version.

$$\Omega_{CFMD} = \{G \mid \forall i \in V, \deg_G^{out}(i) = k_i^{out} \wedge \deg_G^{in}(i) = k_i^{in}\}$$

To compute the weight of the barycenter graph's edges, we consider that each node i has k_i^{out} outgoing stubs and k_i^{in} ingoing stubs. Any graph in Ω_{CFMD} is characterized by a configuration of connections of outgoing stubs with ingoing stubs. For every pair of nodes $i, j \in V^2$ and any pair of stub $k \in [1, k_i^{out}]$, $l \in [1, k_j^{in}]$, we define the random variable $X_{i,j,k,l}$ which is equal to 1 if the k^{th} outgoing stub of i is connected to the l^{th} ingoing stub of j, and to 0 otherwise. Then

$$W_H(i,j) = \sum_{k=1}^{k_i^{out}} \sum_{l=1}^{k_j^{in}} X_{i,j,k,l}$$

As each outgoing stub of i has the same probability to be connected to any of the m ingoing stubs

$$\forall i, j, k, l, \mathbb{P}[X_{i,j,k,l} = 1] = \frac{1}{m}$$

Thus, $W_H(i,j)$ follows a binomial law of parameters $\frac{1}{m}$ and $k_i^{out} \times k_j^{in}$. Finally

$$W_{G_{CFMD}}(i,j) = \mathop{\mathbb{E}}_{\Omega_{CFMD}}[W_H(i,j)] = \frac{k_i^{out} \times k_j^{in}}{m} \tag{7.2}$$

Stochastic blockmodel The case of the stochastic blockmodel can be treated in the same way as erdös-rényi. It is defined, considering a partition of the nodes $B = (b_1, \ldots, b_q)$ and a block adjacency matrix $M \in \mathcal{M}_q(\mathbb{N})$ by

$$\Omega_{SBM} = \left\{ H \mid \forall b_k, b_l, \sum_{i \in b_k} \sum_{j \in b_l} W_H(i,j) = M(k,l) \right\}$$

So, for any pair of nodes $i \in b_k$, $j \in b_l$, $W_H(i,j)$ follows a binomial law of parameters $(M(k,l), |b_k||b_l|)$. Thus

$$W_{G_{SBM}}(i,j) = \mathop{\mathbb{E}}_{\Omega_{SBM}}[W_H(i,j)] = \frac{M(k,l)}{|b_k||b_l|} \tag{7.3}$$

Spatial models References for the gravitational model and the radiation model can be found in Barthelemy [2011] and Simini et al. [2012]. In both cases, they are constructed in such a way that edges weight have a given expected value. In the case of the gravitational model, it is

$$W_{G_{\text{grav}}}(i,j) = \mathrm{f}(\mathrm{d}(i,j)) \times k_i^{out} \times k_j^{in} \tag{7.4}$$

where $\mathrm{d}(i,j)$ is the distance from node i to node j, and f is a deterence function.

Finally, in the case of the radiation model, it is

$$W_{G_{\text{rad}}}(i,j) = \frac{k_i^{out} \times k_i^{in} \times k_j^{in}}{(k_i^{in} + s_{ij}) \times (k_i^{in} + k_j^{in} + s_{ij})} \tag{7.5}$$

with $s_{ij} = \sum_{u \in C(i,j)} k_u^{in}$ and $C(i,j) = \{u \in V \mid 0 < \mathrm{d}(i,u) < \mathrm{d}(i,j)\}$.

Convergence proof for the edit distance expected value

First of all, let's proove the following lemma

Lemma 1. *Let B be a partition of $[\![1, n]\!]$ with p blocks. Let $M \in \mathcal{M}_p(\mathbb{N})$ be a block adjacency matrix. For all $k \in \mathbb{N}$, we define the stochastic blockmodel $S(k) = (B, k \cdot M)$, and its barycenter $G_{S(k)}$. We consider a sequence of random graphs $(G_k)_{k\in\mathbb{N}}$, each drawn from $S(k)$. We have that*

$$\mathrm{ed}(G_k, G_{S(k)}) \xrightarrow[k\to\infty]{\mathbb{P}} 0$$

Given the notation above, we want to proove that:

$$\forall \alpha > 0, \mathbb{P}[\mathrm{ed}(G_k, G_{S(k)}) > \alpha] \underset{k\to\infty}{\longrightarrow} 0$$

Let $\alpha > 0$. Let's denote $m = \sum_{i,j} M_{i,j}$ the number of edges of graphs in $\Omega_{S(1)}$. By definition,

$$\mathrm{ed}(G_k, G_{S(k)}) = \frac{1}{2km} \sum_{u,v} |W_{G_k}(u, v) - W_{G_{S(k)}}(u, v)|$$

Thus,

$$\mathrm{ed}(G_k, G_{S(k)}) > \alpha \Rightarrow \exists (u, v), \left| \frac{W_{G_k}(u, v) - W_{G_{S(k)}}(u, v)}{2km} \right| > \frac{\alpha}{n^2}$$

and

$$\mathbb{P}[\mathrm{ed}(G_k, G_{S(k)}) > \alpha] \leqslant \sum_{u,v} \mathbb{P}\left[\left| \frac{W_{G_k}(u, v) - W_{G_{S(k)}}(u, v)}{2km} \right| > \frac{\alpha}{n^2} \right]$$

Let's consider two blocks b_i and b_j in B. We know that $\forall u \in b_i, v \in b_j, W_{G_{S(k)}}(u, v) = k \cdot \frac{M_{i,j}}{|b_i||b_j|}$ and $W_{G_k}(u, v) \sim \mathcal{B}(k \cdot M_{i,j}, p_{i,j})$ with $p_{i,j} = \frac{1}{|b_i||b_j|}$. Therefore, according to the Bienaymé-Tchebychev inequality:

$$\begin{aligned} \mathbb{P}\left[\left| \frac{W_{G_k}(u, v) - W_{G_{S(k)}}}{2km} \right| > \frac{\alpha}{n^2} \right] &\leqslant \frac{k \times M_{i,j} \times p_{i,j} \times (1 - p_{i,j}) \times n^2}{4 \times k^2 \times m^2 \times \alpha} \\ &\leqslant \frac{M_{i,j} \times p_{i,j} \times (1 - p_{i,j}) \times n^2}{4 \times k \times m^2 \times \alpha} \\ &\underset{k\to\infty}{\longrightarrow} 0 \end{aligned}$$

Thus,

$$\mathbb{P}[\mathrm{ed}(G_k, G_{S(k)}) > \alpha] \underset{k\to\infty}{\longrightarrow} 0 \tag{7.6}$$

Which prooves the lemma.

We can now proove the theorem

Theorem 3. *Let B_1 and B_2 be two partition on $[\![1,n]\!]$, with p_1 and p_2 blocks respectively. Let $M_1 \in \mathcal{M}_{p_1}(\mathbb{N})$ and $M_2 \in \mathcal{M}_{p_2}(\mathbb{N})$ be two block adjacency matrices such that*

$$\sum_{i,j\in[1,p_1]^2} M_1[i,j] = \sum_{k,l\in[1,p_2]^2} M_2[k,l] = m$$

Let's consider two series of stochastic blockmodels defined as $S_1(k) = (B_1, k \cdot M_1)$ and $S_2(k) = (B_2, k \cdot M_2)$, whose barycenters are denoted $G_1(k)$ and $G_2(k)$. We have that

1. *There exists $d \in \mathbb{R}, \forall k \in \mathbb{N}, \mathrm{ed}(G_1(k), G_2(k)) = d$*
2. *Let $(G_k)_{k\in\mathbb{N}}$ be a serie of random graph each drawn following model $S_1(k)$.*

$$\mathrm{EDEV}(G_k, S_2(k)) \xrightarrow[k\to\infty]{\mathbb{P}} d$$

For any pair of nodes i, j, belonging to blocks $b(i)$ and $b(j)$ in B_1 (resp. B_2), the weight of the edge $i \to j$ in $G_1(k)$ (resp. $G_2(k)$) is given by:

$$W_{G_1(k)}[i,j] = k \cdot \frac{M[b(i),b(j)]}{|b(i)||b(j)|}$$

Therefore, the edit distance between $G_1(k)$ and $G_2(k)$ is

$$\begin{aligned}
\mathrm{ed}(G_1(k), G_2(k)) &= \frac{1}{2km} \sum_{i,j\in[1,n]^2} \left|W_{G_1(k)}[i,j] - W_{G_2(k)}[i,j]\right| \\
&= \frac{1}{2km} \sum_{i,j\in[1,n]^2} \left|k \cdot \frac{M_1[b_1(i),b_1(j)]}{|b_1(i)||b_1(j)|} - k \cdot \frac{M_2[b_2(i),b_2(j)]}{|b_2(i)||b_2(j)|}\right| \\
&= \frac{1}{2m} \sum_{i,j\in[1,n]^2} \left|\frac{M_1[b_1(i),b_1(j)]}{|b_1(i)||b_1(j)|} - \frac{M_2[b_2(i),b_2(j)]}{|b_2(i)||b_2(j)|}\right|
\end{aligned}$$

which is constant with respect to k. In the following we will denote this distance d for the sake of conciseness. We want to show that

$$\mathrm{EDEV}(G_k, S_2(k)) \xrightarrow[k\to\infty]{\mathbb{P}} d$$

We start by noticing that

$$\begin{aligned}
\mathrm{EDEV}(G_k, S_2(k)) - d &= \mathop{\mathbb{E}}_{H\in S_2(k)}[\mathrm{ed}(G_k, H)] - d \\
&\leqslant \mathop{\mathbb{E}}_{H\in S_2(k)}[\mathrm{ed}(G_k, G_1(k)) + \mathrm{ed}(G_1(k), G_2(k)) + \mathrm{ed}(G_2(k), H)] - d \\
&\leqslant \mathrm{ed}(G_k, G_1(k)) + \mathop{\mathbb{E}}_{H\in S_2(k)}[\mathrm{ed}(G_2(k), H)]
\end{aligned}$$

On the other hand,

$$\begin{aligned} d - \mathrm{EDEV}(G_k, S_2(k)) &= \mathop{\mathbb{E}}_{H \in S_2(k)} [\mathrm{ed}(G_1(k), G_2(k)) - \mathrm{ed}(G_k, H)] \\ &\leqslant \mathop{\mathbb{E}}_{H \in S_2(k)} [\mathrm{ed}(G_1(k), G_k) + \mathrm{ed}(G_k, H) + \mathrm{ed}(H, G_2(k)) - \mathrm{ed}(G_k, H)] \\ &\leqslant \mathrm{ed}(G_k, G_1(k)) + \mathop{\mathbb{E}}_{H \in S_2(k)} [\mathrm{ed}(G_2(k), H)] \end{aligned}$$

Thus

$$|\mathrm{EDEV}(G_k, S_2(k)) - d| \leqslant \mathrm{ed}(G_k, G_1(k)) + \mathop{\mathbb{E}}_{H \in \Omega_{S_2(k)}} [\mathrm{ed}(G_2(k), H)] \tag{7.7}$$

Because G_k is generated following $S_1(k)$, a direct application of lemma 1 is that

$$\mathrm{ed}(G_k, G_1(k)) \xrightarrow[k \to \infty]{\mathbb{P}} 0$$

What is more, if H is generated following $S_2(k)$, we also have that

$$\mathrm{ed}(H, G_2(k)) \xrightarrow[k \to \infty]{\mathbb{P}} 0$$

which implies that $\mathrm{ed}(H, G_2(k)) \xrightarrow[k \to \infty]{\mathcal{L}} 0$ and in particular

$$\mathop{\mathbb{E}}_{H \in \Omega_{S_2(k)}} [H, G_2(k)] \longrightarrow_{k \to \infty} 0$$

Finally, we obtain that

$$\mathrm{ed}(G_k, G_1(k)) + \mathop{\mathbb{E}}_{H \in \Omega_{S_2(k)}} [H, G_2(k)] \xrightarrow[k \to \infty]{\mathbb{P}} 0$$

And thanks to equation 7.7:

$$\mathrm{EDEV}(G_k, S_2(k)) \xrightarrow[k \to \infty]{\mathbb{P}} d \tag{7.8}$$

Edge statistical model sequential inference

Proof of existence and unicity of the minimum

We prove the following result

> If $\mathcal{M}_n^\phi([0,1])$ is a convex set, then for any edge sequence E, f has a unique minimum $\mathbb{Q}_\phi^*(E)$ over $\mathcal{M}_n^\phi([0,1])$.

Let's consider ϕ such that $\mathcal{M}_n^\phi([0,1])$ is a convex set, and let E be an edge sequence. Let's denote

$$\mathcal{M}_n^\phi(]0,1]) = \{\mathbb{Q} \in \mathcal{M}_n^\phi([0,1]) \mid \forall u,v, \mathbb{Q}[u,v] > 0\}$$

All $\mathbb{Q}$ thus removed from $\mathcal{M}_n^\phi([0,1])$ lies on its boundary, so as it is supposed to be convex, $\mathcal{M}_n^\phi(]0,1])$ is convex too. We consider the function

$$\begin{aligned} \mathrm{f}_E : \mathcal{M}_n^\phi(]0,1]) &\to \mathbb{R} \\ \mathbb{Q} &\mapsto -\sum_{i=0}^{m-1} \log_2(\mathbb{Q}[e_i]) + \sum_{u,v} \mathbb{Q}[u,v] \log_2(\mathbb{Q}[u,v]) \end{aligned}$$

For all pairs of nodes (u,v), we denote

$$K_{u,v} = \#\{k \in [\![0, m-1]\!] \mid e_k = u \to v\}$$

Then, f_E can be rewriten

$$\forall \mathbb{Q}, \mathrm{f}_E(\mathbb{Q}) = \sum_{u,v \in [\![0,n-1]\!]} (\mathbb{Q}[u,v] - K_{u,v}) \log_2(\mathbb{Q}[u,v])$$

f_E is C_2 on $\mathcal{M}_n^\phi(]0,1])$ and it's Hessian matrix is

$$\begin{bmatrix} \frac{K_{0,0}}{\mathbb{Q}[0,0]^2} + \frac{1}{\mathbb{Q}[0,0]} & 0 & \dots & 0 \\ 0 & \frac{K_{1,0}}{\mathbb{Q}[1,0]^2} + \frac{1}{\mathbb{Q}[1,0]} & \dots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & 0 & \frac{K_{n-1,n-1}}{\mathbb{Q}[n-1,n-1]^2} + \frac{1}{\mathbb{Q}[n-1,n-1]} \end{bmatrix}$$

which is positive definite on $\mathcal{M}_n^\phi(]0,1])$, so f_E is strictly convex on this set. As $\mathcal{M}_n^\phi(]0,1])$ is a convex set, we obtain that f_E has a unique minimum over it, which we can denote $\mathbb{Q}_\phi^*(E)$. It remains to be proven that $\mathbb{Q}_\phi^*(E)$ is the minimum of f_E over $\mathcal{M}_n^\phi([0,1])$. $\mathcal{M}_n^\phi([0,1])$ is the closure (in the topological sense) of $\mathcal{M}_n^\phi(]0,1])$, so f_E can be continuously extended to it, provided that we extend it's codomain to $\bar{\mathbb{R}} = \mathbb{R} \cup \infty$. Let's consider $\mathbb{Q} \in \mathcal{M}_n^\phi([0,1])$ such that $\exists u,v, \mathbb{Q}[u,v] = 0$ and a sequence $(\mathbb{Q}_i)_{i \in \mathbb{N}} \in \mathcal{M}_n^\phi(]0,1])$ that converges toward $\mathbb{Q}$. There are two different situations.

1. If $\exists u_0, v_0, \mathbb{Q}[u_0, v_0] = 0 \wedge K_{u_0,v_0} > 0$. Then,

$$\begin{aligned}\forall i, \mathrm{f}_E(\mathbb{Q}_i) &= \sum_{u,v\in[\![0,n-1]\!]} (\mathbb{Q}_i[u,v] - K_{u,v}) \log_2(\mathbb{Q}_i[u,v]) \\ &= \sum_{u,v\neq u_0,v_0} (\mathbb{Q}_i[u,v] - K_{u,v}) \log_2(\mathbb{Q}_i[u,v]) + \\ &\quad (\mathbb{Q}_i[u_0,v_0] - K_{u_0,v_0}) \log_2(\mathbb{Q}_i[u_0,v_0])\end{aligned}$$

Thus,

$$\mathrm{f}_E(\mathbb{Q}_i) \underset{i\to\infty}{\longrightarrow} \infty$$

So we define $\mathrm{f}_E(\mathbb{Q}) = \infty$ and in particular $\mathrm{f}_E(\mathbb{Q}) > \mathrm{f}_E(\mathbb{Q}^*_\phi(E))$.

2. If $\forall u, v, \mathbb{Q}[u,v] = 0 \Rightarrow K_{u,v} = 0$. Then,

$$\begin{aligned}\forall i, \mathrm{f}_E(\mathbb{Q}_i) &= \sum_{u,v\in[\![0,n-1]\!]} (\mathbb{Q}_i[u,v] - K_{u,v}) \log_2(\mathbb{Q}_i[u,v]) \\ &= \sum_{u,v|\mathbb{Q}[u,v]>0} (\mathbb{Q}_i[u,v] - K_{u,v}) \log_2(\mathbb{Q}_i[u,v]) + \\ &\quad \sum_{u,v|\mathbb{Q}[u,v]=0} \mathbb{Q}_i[u,v] \log_2(\mathbb{Q}_i[u,v])\end{aligned}$$

Thus,

$$\mathrm{f}_E(\mathbb{Q}_i) \underset{i\to\infty}{\longrightarrow} \sum_{u,v|\mathbb{Q}[u,v]>0} (\mathbb{Q}[u,v] - K_{u,v}) \log_2(\mathbb{Q}[u,v])$$

and we define

$$\mathrm{f}_E(\mathbb{Q}) = \sum_{u,v|\mathbb{Q}[u,v]>0} (\mathbb{Q}[u,v] - K_{u,v}) \log_2(\mathbb{Q}[u,v])$$

By continuity of f_E, we know that $\mathrm{f}_E(\mathbb{Q}) \geqslant \mathrm{f}_E(\mathbb{Q}^*_\phi(E))$. Let's show that this inequality is strict. We consider the restriction of f_E to the interval

$$I = \{\lambda \cdot \mathbb{Q}^*_\phi(E) + (1-\lambda)\cdot\mathbb{Q}, \lambda \in [0,1[\} \subset \mathcal{M}^\phi_n(]0,1])$$

Because of the strict convexity of f_E on $\mathcal{M}^\phi_n(]0,1])$, $\mathrm{f}_E\big|_I$ is a strictly increasing function of λ. As a consequence,

$$\mathrm{f}_E(\mathbb{Q}^*_\phi(E)) < \lim_{\lambda\to 1} \mathrm{f}_E\big|_I(\lambda\cdot\mathbb{Q}^*_\phi(E) + (1-\lambda)\cdot\mathbb{Q}) = \mathrm{f}_E(\mathbb{Q})$$

Which proves that in both cases, $\mathbb{Q}^*_\phi(E)$ is the only minimum of f over $\mathcal{M}^\phi_n([0,1])$.

Proof of existence and unicity of the minimum (configuration model)

We prove the following result

For any edge sequence E, f has a unique minimum $\mathbb{Q}^*_{\phi^{CM}}(E)$ over $\mathcal{M}_n^{\phi^{CM}}([0,1])$.

Let's define the set of probability distributions on $[\![1,n]\!]$:

Definition 8. *We denote $\mathcal{V}_n([0,1])$ the set*

$$\mathcal{V}_n([0,1]) = \left\{ p \in [0,1]^n \mid \sum_{u=0}^{n-1} p[u] = 1 \right\}$$

By definition, we have a bijection

$$\begin{aligned} \psi : \mathcal{V}_n([0,1])^2 &\to \mathcal{M}_n^{\phi^{CM}}([0,1]) \\ (p^{out}, p^{in}) &\mapsto \mathbb{Q} = p^{out} \cdot (p^{in})^T \end{aligned}$$

Let's consider a probability distribution $\mathbb{Q} \in \mathcal{M}_n^{\phi^{CM}}([0,1])$, and $p^{out}, p^{in} \in \mathcal{V}_n([0,1])^2$ such that $\forall u, v, \mathbb{Q}[u,v] = p^{out}[u] \cdot p^{in}[v]$, then

$$\begin{aligned} \mathrm{f}(\mathbb{Q}, E) &= -\sum_{i=1}^{m} \log_2(\mathbb{Q}[e_i]) + \sum_{u,v} \mathbb{Q}[u,v] \log_2(\mathbb{Q}[u,v]) \\ &= -\sum_{i=1}^{m} \log_2(p^{out}[u_i] \cdot p^{in}[v_i]) + \sum_{u,v} (p^{out}[u] \cdot p^{in}[v]) \log_2(p^{out}[u] \cdot p^{in}[v]) \\ &= -\sum_{i=1}^{m} \log_2(p^{out}[u_i]) + \sum_{u} \sum_{v} p^{out}[u] \cdot p^{in}[v] \cdot \log_2(p^{out}[u]) + \\ &\quad -\sum_{i=1}^{m} \log_2(p^{in}[v_i]) + \sum_{u} \sum_{v} p^{out}[u] \cdot p^{in}[v] \cdot \log_2(p^{in}[v]) \\ &= -\sum_{i=1}^{m} \log_2(p^{out}[u_i]) + \sum_{u} p^{out}[u] \cdot \log_2(p^{out}[u]) + \\ &\quad -\sum_{i=1}^{m} \log_2(p^{in}[v_i]) + \sum_{v} p^{in}[v] \cdot \log_2(p^{in}[v]) \end{aligned} \tag{7.9}$$

Hence, if we introduce

$$\begin{aligned} K_u &= \#\{k \in [\![1,m]\!], u_k = u\} \\ K_v &= \#\{k \in [\![1,m]\!], v_k = v\} \end{aligned}$$

following the same reasoning as in Annex 7.2.1, we can define

$$\begin{aligned} \mathrm{g}_E^{out} : \mathcal{V}_n([0,1]) &\to \mathbb{R} \\ p &\mapsto \sum_u (p[u] - K_u) \cdot \log_2(p[u]) \end{aligned}$$

$$\begin{aligned} g_E^{in} : \mathcal{V}_n([0,1]) &\to \mathbb{R} \\ p &\mapsto \sum_v (p[v] - K_v) \cdot \log_2(p[v]) \end{aligned}$$

They both have a unique minimum which we denote respectively $p^{out*}(E)$ and $p^{in*}(E)$. Then, we define

$$\mathbb{Q}^*_{\phi^{CM}}(E) = \psi(p^{out*}(E), p^{in*}(E))$$

Let's show that $\mathbb{Q}^*_{\phi^{CM}}(E)$ is the unique minimum of f over $\mathcal{M}_n^{\phi^{CM}}([0,1])$. Let $\mathbb{Q} \in \mathcal{M}_n^{\phi^{CM}}([0,1])$ such that $f(\mathbb{Q}, E) \leqslant f(\mathbb{Q}^*_{\phi^{CM}}(E), E)$. Let $p^{out} \in \mathcal{V}_n([0,1])$ and $p^{in} \in \mathcal{V}_n([0,1])$ such that $\mathbb{Q} = \psi(p^{out}, p^{in})$. According to equation 7.9,

$$f(\mathbb{Q}, E) = g_E^{out}(p^{out}) + g_E^{in}(p^{in})$$

So, by definition of $\mathbb{Q}$,

$$g_E^{out}(p^{out}) + g_E^{in}(p^{in}) \leqslant g_E^{out}(p^{out*}) + g_E^{in}(p^{in*})$$

Which implies that $p^{out} = p^{out*}$ and $p^{in} = p^{in*}$, and thus that $\mathbb{Q} = \mathbb{Q}^*_{\phi^{CM}}(E)$. So $\mathbb{Q}^*_{\phi^{CM}}(E)$ is the unique minimum of $f(\mathbb{Q}, E)$ over $\mathcal{M}_n^{\phi^{CM}}([0,1])$.

Proof of convergence

We prove the following result:

Let $(e_i)_{i\in\mathbb{N}}$ be a sequence of independent and identically distributed random variables following $\mathbb{P}_0 \in \mathcal{M}_n^\bullet([0,1])$.

$$\forall \phi, \mathbb{Q}_\phi^*(e_1,\dots,e_x) \underset{x\to\infty}{\longrightarrow} \underset{\mathbb{Q}\in Prob\_mat_\phi}{\operatorname{argmin}} \mathbb{H}(\mathbb{P}_0,\mathbb{Q})$$

Let $\mathbb{P}_0 \in \mathcal{M}_n^\bullet([0,1])$.

Let $(e_i)_{i\in\mathbb{N}}$ be a sequence of independent and identically distributed random variables following $\mathbb{P}_0$.

Let's consider the function

$$f(\mathbb{Q},x) = -\sum_{i=1}^{x} \log_2(\mathbb{Q}[e_i]) + \sum_{u,v} \mathbb{Q}[u,v]\log_2(\mathbb{Q}[u,v])$$

We want to show that:

$$\forall \phi, \underset{\mathbb{Q}\in\mathcal{M}_n^\phi([0,1])}{\operatorname{argmin}} f(\mathbb{Q},x) \underset{x\to\infty}{\longrightarrow} \underset{\mathbb{Q}\in\mathcal{M}_n^\phi([0,1])}{\operatorname{argmin}} \mathbb{H}(\mathbb{P}_0,\mathbb{Q})$$

Let ϕ be an hyperparameter and $\mathbb{Q} \in \mathcal{M}_n^\phi([0,1])$. Following the weak law of large numbers

$$-\frac{1}{x}\sum_{i=1}^{x} \log_2(\mathbb{Q}[e_i]) \underset{x\to\infty}{\longrightarrow} \mathbb{H}(\mathbb{P}_0,\mathbb{Q})$$

Hence

$$\frac{1}{x}f(\mathbb{Q},x) - \mathbb{H}(\mathbb{P}_0,\mathbb{Q}) \underset{x\to\infty}{\longrightarrow} 0$$

So if we consider the sequence of functions

$$\begin{aligned} g_x \colon \mathcal{M}_n^\phi([0,1]) &\to \mathbb{R} \\ \mathbb{Q} &\mapsto \frac{1}{x}f(\mathbb{Q},x) - \mathbb{H}(\mathbb{P}_0,\mathbb{Q}) \end{aligned}$$

it converges point-wise toward 0. As it is an equicontinuous family of functions defined on a compact set of R^n, it converges uniformly toward 0. This means that

$$\forall \delta > 0, \exists A \in \mathbb{R}^+, \forall \mathbb{Q} \in \mathcal{M}_n^\phi([0,1]), \forall x \geqslant A, \left|\frac{1}{x}f(\mathbb{Q},x) - \mathbb{H}(\mathbb{P}_0,\mathbb{Q})\right| < \delta \tag{7.10}$$

What is more, if we let $\mathbb{P}' = \underset{\mathbb{Q}\in\mathcal{M}_n^\phi([0,1])}{\operatorname{argmin}} \mathbb{H}(\mathbb{P}_0,\mathbb{Q})$. $\mathbb{H}$ is a strictly convex function of $\mathbb{Q}$ so

$$\forall \epsilon > 0, \exists \delta > 0, \forall \mathbb{Q} \in \mathcal{M}_n^\phi([0,1]), |\mathbb{H}(\mathbb{P}_0,\mathbb{Q}) - \mathbb{H}(\mathbb{P}_0,\mathbb{P}')| < \delta \Rightarrow |\mathbb{Q} - \mathbb{P}'| < \epsilon \tag{7.11}$$

With those two inequalities, we can proceed to the convergence demonstration. Let $\epsilon > 0$, δ such as in equation 7.11, A such as in equation 7.10 with $\frac{\delta}{3}$, and $x \geqslant A$. Let $\mathbb{Q}(x) = \underset{\mathbb{Q} \in \mathcal{M}_n^\phi([0,1])}{\operatorname{argmin}} \frac{1}{x} f(\mathbb{Q}, x)$. Because of equation 7.10, we have that

$$\left| \frac{1}{x} f(\mathbb{Q}(x), x) - \mathbb{H}(\mathbb{P}_0, \mathbb{Q}(x)) \right| < \frac{\delta}{3}$$
$$\left| \frac{1}{x} f(\mathbb{P}', x) - \mathbb{H}(\mathbb{P}_0, \mathbb{P}') \right| < \frac{\delta}{3}$$

Thus, if $|\mathbb{H}(\mathbb{P}_0, \mathbb{Q}(x)) - \mathbb{H}(\mathbb{P}_0, \mathbb{P}')| \geqslant \delta$:

$$\begin{aligned} \frac{1}{x} f(\mathbb{Q}(x), x) &\geqslant \mathbb{H}(\mathbb{P}_0, \mathbb{Q}(x)) - \frac{\delta}{3} \\ &\geqslant \mathbb{H}(\mathbb{P}_0, \mathbb{P}') + \frac{2\delta}{3} \\ &> \mathbb{H}(\mathbb{P}_0, \mathbb{P}') + \frac{\delta}{3} \\ &> \frac{1}{x} f(\mathbb{P}', x) \end{aligned}$$

Which contradicts the definition of $\mathbb{Q}(x)$. Thus $|\mathbb{H}(\mathbb{P}_0, \mathbb{Q}(x)) - \mathbb{H}(\mathbb{P}_0, \mathbb{P}')| < \delta$, and because of equation 7.11

$$|\mathbb{Q}(x) - \mathbb{P}'| < \epsilon$$

Which proves that:

$$\underset{\mathbb{Q} \in \mathcal{M}_n^\phi([0,1])}{\operatorname{argmin}} f(\mathbb{Q}, x) \underset{x \to \infty}{\longrightarrow} \underset{\mathbb{Q} \in \mathcal{M}_n^\phi([0,1])}{\operatorname{argmin}} \mathbb{H}(\mathbb{P}, \mathbb{Q})$$

And as this is true for any hyperparameter ϕ, the result is proved.

Description length computation

To compute the description length of a sequence E, a fundamental result in information theory states that, if a source (let's call her Alice) draws messages independently at random from a set Ω following a probability distribution $\mathbb{Q}$ and then transmit them to a destination (Bob) over a binary channel, then the code $C_{\mathbb{Q}} : \Omega \to [0,1]^*$ which minimizes the expected length of the total message $\mathbb{E}_{x\in\Omega}[|C(x)|]$ will be such that:

$$\forall x \in \Omega, |C(x)| = -\log_2(\mathbb{Q}[x])$$

Therefore, if we suppose that all edges $e_i \in E$ were generated independently following a probability distribution $\mathbb{Q}$, we obtain that:

$$\begin{aligned} D(E|\mathbb{Q}) &= -\log_2(\mathbb{Q}[E]) \\ &= -\log_2\left(\prod_{k=1}^{m}\mathbb{Q}[e_k]\right) \\ &= -\sum_{i=1}^{m}\log_2(\mathbb{Q}[e_k]) \end{aligned}$$

Bayesian inference

We have defined the estimation $\mathbb{Q}^*_\phi(E)$ as the model which allows for the best compression of E. Yet, if we consider $\mathcal{M}^\phi_n([0,1])$ as the set of models which could have been used to generate E, $\mathbb{Q}^*_\phi(E)$ can also be interpreted as the most likely hypothesis among them.

According to Bayes' theorem, the probability that a model $\mathbb{Q} \in \mathcal{M}^\phi_n([0,1])$ was the one used to generate the edge sequence E is

$$\mathbb{P}_\phi[\mathbb{Q}|E] = \frac{\mathbb{P}[E|\mathbb{Q}] \times \bar{\mathbb{P}}_\phi[\mathbb{Q}]}{\mathbb{P}[E]}$$

Therefore, as $\mathbb{P}[E]$ does not depend on $\mathbb{Q}$,

$$\mathbb{Q}^*_\phi(E) = \underset{\mathbb{Q}\in\mathcal{M}^\phi_n([0,1])}{\operatorname{argmax}} \mathbb{P}[E|\mathbb{Q}] \times \bar{\mathbb{P}}_\phi[\mathbb{Q}] \tag{7.12}$$

In practice, it means that if we infer the most likely model for an empty sequence, $\mathbb{Q}^*_\phi[\varnothing]$ will be the highest entropy model within $\mathcal{M}^\phi_n([0,1])$. On the other hand, as we have more and more observations, the sequence E becomes longer and the influence of the prior distribution $\bar{\mathbb{P}}_\phi[\mathbb{Q}]$ becomes negligible. As the probability to generate an edge (u,v) with a model $\mathbb{Q}$ is simply $\mathbb{Q}[u,v]$ and edges are assumed to be independent, this equation becomes

$$\mathbb{Q}^*_\phi(E) = \underset{\mathbb{Q}\in\mathcal{M}^\phi_n([0,1])}{\operatorname{argmax}} \prod_{i=1}^{m} \mathbb{Q}[e_i] \times \frac{1}{Z_\phi} \times 2^{\$[\mathbb{Q}]}$$

To perform the maximization, it is simpler to consider the logarithm of this expression. As $\log_2$ is a monotonous function, it does not change the value of $\mathbb{Q}^*_\phi(E)$.

$$\begin{aligned}
\mathbb{Q}^*_\phi(E) &= \underset{\mathbb{Q}\in\mathcal{M}^\phi_n([0,1])}{\operatorname{argmax}} \log_2\left(\prod_{i=1}^{m} \mathbb{Q}[e_i] \times \frac{1}{Z_\phi} \times 2^{\$[\mathbb{Q}]}\right) \\
&= \underset{\mathbb{Q}\in\mathcal{M}^\phi_n([0,1])}{\operatorname{argmax}} \sum_{i=1}^{m} \log_2(\mathbb{Q}[e_i]) + \$[\mathbb{Q}] \\
&= \underset{\mathbb{Q}\in\mathcal{M}^\phi_n([0,1])}{\operatorname{argmin}} -\sum_{i=1}^{m} \log_2(\mathbb{Q}[e_i]) - \$[\mathbb{Q}]
\end{aligned}$$

Model	Partition	Block probability matrix
S_0	$B_0 = [\![1, 128]\!]$	$M_0 = \frac{1}{n^2} \cdot \begin{bmatrix} 1 \end{bmatrix}$
S_1	$B_1 = [\![1, 64]\!], [\![65, 128]\!]$	$M_1 = \frac{1}{n^2} \cdot \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$
S_2	$B_2 = [\![1, 32]\!], [\![33, 64]\!], [\![65, 96]\!], [\![97, 128]\!]$	$M_2 = \frac{1}{n^2} \cdot \begin{bmatrix} 4 & 0 & 0 & 0 \\ 0 & 4 & 0 & 0 \\ 0 & 0 & 4 & 0 \\ 0 & 0 & 0 & 4 \end{bmatrix}$

Table 7.1 – Three stochastic blockmodels defined as edge probability distributions.

Edge prediction probability

We investigate how the prediction probability of the next edge evolves as Alice draws more and more edges. We consider three stochastic blockmodels on $n = 128$ nodes based on three partitions B_0, B_1 and B_2. B_0 is made of a single block of size 128, B_1 of two blocks of size 64 and B_2 of four blocks of size 32, obtained by dividing B_1's block in half. The three SBMs are fully described in Table 7.1. For each SBM, we randomly sample $m = 2800$ edges and thus obtain three graphs: G_0, G_1 and G_2. We want to study the edge prediction probability evolution depending on the constraints used to learn the model. Thus, for each of the three graphs, and each of the three hyperparameters ϕ^{B_0}, ϕ^{B_1}, ϕ^{B_2}, we plot the evolution of the prediction probability $\mathbb{Q}^*_\phi(e_1, \ldots, e_{k-1})[e_k]$ against k in Figure 7.1.

This simple example shows how the level of constraints imposed by the hyperparameter acts on the probability prediction of the next edge. For all three graphs, whatever k, the prediction probability based on the null partition B_0 is constant at 0.00006 (black dots). This is logical, as the only probability matrix in $\mathcal{M}_n^{\phi^{B_0}}([0, 1])$ is the uniform distribution. Therefore,

$$\forall k, \mathbb{Q}^*_{\phi^{B_0}}(e_1, \ldots, e_{k-1})[e_k] = \frac{1}{n^2} = \frac{1}{128^2} \approx 0.00006$$

For other hyperparameters (red and yellow dots), the results depend on the graph. On G_0, generated with B_0 and thus presenting no block structure, models based on more refined partitions do not lead on average to better prediction probabilities than the one based on B_0. For some edges their prediction probability is better, but as often it is worse. On average, they have the same prediction power, but the convergence toward the generative probability distribution is slowed down by random fluctuations due to the additional degree of freedom allowed.

On the other hand, for G_1, generated with B_1 (two blocks), we observe that refining the partition from one block to two allows the prediction probability to increase quickly. While it remains $\frac{1}{n^2}$ for the hyperparameter ϕ^{B_0}, it converges to $\frac{2}{n^2}$ for the hyperparameter ϕ^{B_1} (red dots). Yet, refining even more the partition is worthless, as illustrated by the B_2 partition (yellow dots), with 4 blocks, which does not bring any improvement on average. Finally, considering G_2, we observe that refining the partition brings more and more improvement to the prediction probability. With B_0 it remains stable at $\frac{1}{n^2}$, with B_1 it rises up to $\frac{2}{n^2}$, and with

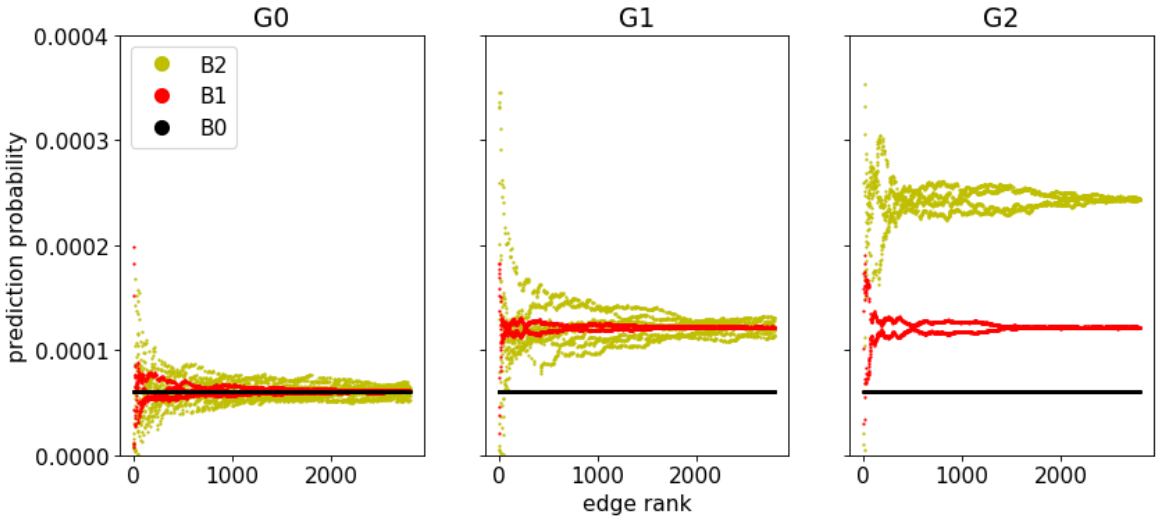


Figure 7.1 – For each graph, we plot the edge prediction probability $\mathbb{Q}^*_{\phi^B}(e_1, \ldots, e_{k-1})[e_k]$ against k. Black dots corresponds to the model learned with partition B_0, red dots with partition B_1 and yellow dots with partition B_2. As the number of observed edges grows, the prediction converges to a value which depends on G and B. When the learning partition is coarser than the original partition, the prediction probability converges to a lower value. When it is finer, it converges toward the same value, but more slowly.

B_2 up to $\frac{4}{n^2}$. This shows that increasing the number of degrees of freedom of the model (*i.e.* reducing the number of constraints of the hyperparameter) is a double-edged sword. As long as it allows the model to better fit correlations that are present in the observations, it leads to better prediction performance. Yet, this comes at the price of a slower convergence of the model. It is the combination of those two effects which allows us to detect both overfitting and underfitting models.